Modelling mutations and homologous proteins

Andrej Šali

The Rockefeller University, New York, USA

A protein sequence with at least 40% identity to a known structure can now be modelled automatically, with an accuracy approaching that of a low-resolution X-ray structure or a medium-resolution nuclear magnetic resonance structure. In general, these models have good stereochemistry and an overall structural accuracy that is as high as the similarity between the template and the actual structure being predicted. As a result, the number of sequences that can be modelled is an order of magnitude larger than the number of experimentally determined protein structures. In addition, evaluation techniques are available that can estimate errors in different regions of the model. Thus, the number of applications where homology modelling is proving useful is growing rapidly.

Current Opinion in Biotechnology 1995, **6**:437–451

Introduction

Comparative or homology protein modelling uses experimentally determined protein structures (templates) to predict the conformation of another protein with a similar amino acid sequence (target) [1–5,6••,7•–9•]. Comparative modelling is possible because a small change in the sequence usually results in a small change in the three-dimensional structure [10–14,15••,16•]. Although considerable progress has been made in *ab initio* structure prediction [17•,18•,19,20•,21•], comparative modelling remains the only modelling method that can provide models with a root mean square (rms) error lower than 2 Å. In general, the best comparative techniques can produce models with good stereochemistry and overall structural accuracy that is as high as the similarity between the template and the actual target structure. Thus, the comparative method can result in models with a main-chain rms error as low as 1 Å for 90% of the main-chain residues, if a sequence is at least 40% identical to one or more of the templates [22•,23]. In this range of sequence similarity, the alignment is mostly straightforward to construct, few gaps exist 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 ~1.5 Å for ~80% of the residues. The rest of the residues are modelled with large errors because the methods generally cannot model structural distortions and rigid body shifts, and they cannot recover from misalignments. In such situations, model evaluation methods can be used to identify the inaccurately modelled regions of a protein. When sequence identity drops below 30%, the main problem becomes the identification of related templates and their alignment with the sequence to be modelled.

Despite these limitations, comparative modelling is useful because about one-third of known sequences appear to be related to at least one known structure [24,25••]. Because only ~2000 of the about 100 000 known protein sequences have had their structures determined experimentally, the number of sequences that can be modelled relatively accurately 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.

In the early eighties, manual comparative modelling [26,27] was facilitated by the manipulations of protein molecules on the graphics terminal [3,28], which was made possible by computer programs such as FRODO [29]. This approach was later improved by the introduction of largely automated modelling algorithms that could use several known structures to model the unknown member of the family [30,31]. This group of methods is based on the assembly of the model from a few core regions, and loops and side chains, which are obtained from dissected related structures [1,3,28]. Another group of comparative methods relies on the approximate positions of conserved atoms from the templates to calculate the coordinates of other atoms, using a database of short segments of protein structure, energy or geometry rules, or some combination of these criteria [32–35]. A third group of comparative methods is based on the satisfaction of spatial restraints obtained from the alignment of the target sequence with homologous templates of known structure [2,36,37].

Abbreviations

NMR—nuclear magnetic resonance; **rms**—root mean square; **WWW**—World-Wide Web.

 ISSN 0958-1669

And a final group of methods, which is not covered in this review, consists of recognition of the native fold by threading a target sequence through each fold in a database of all known folds [38,39]. This can be seen as a first step towards modelling sequences that are only distantly related to the known protein structures [40–50,51•,52•,53]. In addition to methods for modelling the whole fold, numerous other techniques for prediction of 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 modelling techniques.

This review is organized in terms of the main stages that are shared by all comparative modelling methods. The first step is always to align the target sequence with all the related proteins whose three-dimensional structures are known. In the second step, the alignment and the structures are used to build a model for the target sequence. The main difference between the comparative methods is in how the alignment is used to get the three-dimensional model. In the third step, the model is evaluated and, if necessary, the alignment and model building are repeated until a satisfactory model is obtained. For each of the three steps, I first provide a brief historical overview and then describe in more detail the latest developments published since 1993.

Finding and aligning template structures with the target sequence

The first task in comparative modelling is to identify all protein structures related to the target sequence, some of which will be used as templates. This is greatly facilitated by databases of protein sequences and structures and software for scanning those databases (for reviews, see [6••,9•,54•,55–57]). At present, the probability is ~30% that a sequence picked randomly from a sequence database has at least 25% sequence identity to at least one known structure [25••].

The target sequence can be searched against sequence databases, such as Protein Identification Resource (PIR) [58], GENBANK [59], SWISS-PROT [60], or EMBO nucleotide sequences database [61], and/or structure databases such as the Brookhaven Protein Databank [62] and SCOP [63•]. The most popular programs, including FASTA [64] and BLAST [65], compare the target sequence with each sequence in a database. Program MODELLER (see below), which implements all the stages in comparative modelling [5], can also automatically search for proteins with known three-dimensional structure that are related to a given sequence. The sensitivity of the search can be improved if the target sequence is aligned against sequence templates constructed from multiply aligned sequences [66,67•,68,69].

Additional sensitivity in detecting remote relationships is gained when structural information about potential homologues is used. Typically, the target sequence is matched against a library of three-dimensional profiles or threaded through a library of three-dimensional folds [45–47,55,70,71•]. These more sensitive fold identification techniques are especially useful for finding significant structural relationships when sequence identity drops below 30%.

Once all the structures related to the target sequence are identified, the second task is to prepare a multiple alignment of the target sequence with all the potential template structures. The alignment can frequently be improved if other sequences from the same family are also aligned at the same time. This additional effort is often useful because the quality of the alignment is the single most important factor determining the accuracy of the three-dimensional model. In principle, most sequence-alignment and structure-comparison methods can be used, but in practice it is frequently necessary to edit manually the positions of insertions and deletions to ensure that they occur in a reasonable structural context (e.g. not in the middle of a helix). Comparison methods are not reviewed here (for reviews, see [6••,54•,56,72]). Although profile matching and threading techniques are relatively successful in identifying related folds, they appear to be somewhat less successful in generating correct alignments. This limits the use of alignments from threading because comparative modelling cannot, at present, recover from an incorrect alignment. At 30% sequence identity, the fraction of correctly aligned residues is ~80%, but this number drops sharply with further decrease in sequence similarity [73]. This implies that reasonable homology models can be obtained only for sequences that have more than 30% identity to at least one known structure. With such a high similarity, the potential template structures can be almost always identified and aligned using the simplest sequence based searches and alignment techniques. Sequence identity of at least 30% almost guarantees that two chains longer than 50 residues will have related three-dimensional structures [12].

The power of the databases to address various questions is greatly enhanced when relationships between the proteins are established. Several collections of alignments of protein structures have been published that facilitate both the development and the use of comparative modelling techniques [74–78,79•].

Once a multiple alignment is constructed, a matrix of pairwise sequence identities is usually calculated and employed to construct a phyletic tree that expresses the relationships among the proteins in the family [80]. All significantly different structures in the cluster that contains the target sequence are usually used as templates in subsequent model building [81]. Some methods allow short segments of known structure, such as loops [32], to be added to the alignment at this stage [5].

Model building

Modelling by assembly of rigid bodies

The first approach used for comparative modelling was to assemble a model from a small number of rigid

bodies obtained from the aligned protein structures [1,3,26–28,30,31,82–85]. For example, in the computer program COMPOSER [30,31], three types of rigid body are used to build the model. Each individual rigid body of the model is selected as the best rigid body from the corresponding set of all possible rigid bodies. These sets are the following: first, for a conserved core region, the equivalent segments of contiguous main-chain atoms from homologous structures; second, for a loop, the equivalent loops from homologous proteins and loops satisfying certain geometric criteria from other structures; and finally, for a side chain, the equivalent side chains from homologous structures, as well as the most likely side-chain conformations found in proteins in general. These rigid bodies are assembled on the framework, which is defined as the average C_α atoms in the conserved regions of the fold.

Recently, Srinivasan and Blundell [23] have extensively evaluated comparative modelling by rigid body assembly. They found that the accuracy of a model can be somewhat increased when more than one template structure is used to construct the framework and when the templates are averaged into the framework using weights corresponding to their sequence similarities to the target sequence. For example, differences between the modelled and X-ray structures of the modelled protein may be slightly smaller than the differences between the X-ray structures of the modelled protein and the homologues used to build the model. Possible future improvements of modelling by rigid body assembly include incorporation of rigid body shifts, such as the relative shifts in the packing of α-helices [86].

Peitsch and Jongeneel [87] described an automated approach to homology modelling, similar to that of Blundell and co-workers [30,86]. They applied their approach to model the CD40 ligand [87].

Kajihara *et al.* [85] constructed a three-dimensional model of bovine pancreatic β-trypsin from four parts corresponding to each of its exons. These four building blocks were obtained as the most similar regions found in four other serine proteases with known three-dimensional structure. The model was then refined by molecular dynamics simulation. In agreement with [23], it was shown that this 'chimaera' approach is better than using only a single template structure.

Modelling by segment matching or coordinate reconstruction

The build-up procedure constructs the three-dimensional model by assembling short segments of the structure. The segments were originally generated and assembled according to the energetic criteria [88]. The use of this idea in comparative modelling was facilitated by the finding that only ~100 different hexamers can be joined together to cover 99% of the residues in proteins [33]. This paved the way to a new approach to comparative modelling, in which a subset of atomic positions in the template is used to identify short segments in all known protein structures that fit on the guiding positions. The short segments are then assembled into the complete model. For example, Claessens *et al.* [34] developed a method for modelling the backbone with 'spare parts', short segments of varying length from other structures that were identified by matching the guiding C_α positions. Other similar backbone reconstruction procedures have been described [89–91]. A more general segment match modelling by Levitt [35] is guided by the positions of some atoms (usually C_α atoms) to find the matching segments in the representative database of all known protein structures. This method can construct both side-chain and main-chain atoms, and it can also model insertions and deletions.

Many methods for constructing coordinates of missing atoms from the positions of guiding atoms rely on geometric or energetic criteria and possibly on a conformational search, instead of depending on a database of segments [92–96,97•]. Usually, the guiding positions are C_α atoms of a subset of residues, and either main-chain or full-atom models are constructed. These methods can be applied to comparative modelling when homologous structures are used as the source of the guiding positions and when combined with the loop and side-chain construction algorithms [89,90,98].

Even the class of loop construction methods based on finding suitable fragments in the database of known structures [32] can be seen as a segment matching or coordinate reconstruction method. The same is true for some side-chain modelling methods [99•].

Payne [96] used C_α coordinates to reconstruct complete backbone coordinates and side-chain directions. A potential of mean force, derived from a database of protein structures was employed to orient the peptide groups around axes connecting successive C_α atoms. Because terms of the scoring function were local in nature, a dynamic programming procedure could be used for optimization.

Van Gelder *et al.* [97•] have presented a new method to build a complete protein structure from C_α coordinates. The first step in this approach is to generate an approximate backbone using geometrical criteria only. In the second step, the backbone is refined and side chains are positioned using exhaustive molecular dynamics simulation. These authors used the method to generate full-atom models of two proteins from their low-resolution C_α traces.

Modelling by satisfaction of spatial restraints

It is important to distinguish between constraints and restraints. Constraints restrict a spatial feature, such as a distance between two atoms, to a particular single value, whereas restraints allow a wider range of values, possibly with varying probabilities.

Srinivasan *et al.* [36] described a three-dimensional model of bungarotoxin that was obtained through the

use of distance geometry to satisfy main-chain distance constraints extracted from the cobratoxin structure and a low-resolution structure of bungarotoxin. A general method for modelling by optimization of spatial restraints obtained from the alignment of the target sequence with homologous templates of known structure was proposed by Šali and Blundell [2,100]. An elegant distance geometry approach for constructing all-atom models from lower and upper bounds on distances and dihedral angles was described in detail by Havel and Snow [37]. Other methods based on satisfaction of main-chain distance restraints by molecular dynamics were reported by Fujiyoshi-Yoneda *et al.* [101] and Engh *et al.* [102]. A protein backbone has also been modelled by satisfying C_α–C_α contacts predicted by a neural network that relied on an alignment between the target sequence and a template structure [103].

Recently, Havel [104] has extended the earlier approach of Havel and Snow [37]. Lower and upper bounds on C_α–C_α distances, 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 permitted 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 approximate three-dimensional models, which were then exhaustively refined by restrained molecular dynamics with simulated annealing in water.

Comparative modelling by optimization of a potential function constructed from a sequence alignment with related structures was described by Snow [105]. His model consists of C_α atoms that are restrained by a form of a Lennard–Jones potential. The position of the minimum of each Lennard–Jones term corresponds to a weighted average of equivalent distances in homologous structures and the depth of the minimum is inversely proportional to the variability among these distances. The 'energy' is minimized by a simulated-annealing procedure in the angle and dihedral angle space, followed by a conjugate gradients refinement in the Cartesian space. The method is tested by modelling rubredoxin on the basis of four other rubredoxin structures.

The method developed by Srinivasan *et al.* [106] uses a single template structure to obtain distance constraints on the target sequence. As in [37], constraints are derived for all pairs of atoms that have equivalent pairs in the template structure. Distance constraints are satisfied by a distance geometry program and a subsequent energy refinement. When the template and target sequences are similar, the target structure is also very similar to the template structure. Subsequently, the method has been improved by relaxing distance constraints on the target sequence outside the manually delineated structurally conserved regions [107•]. This relaxation facilitates three-dimensional embedding and energy minimization, and increases the rms 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 [107•].

Brocklehurst and Perham [108] have described an automated method for constructing a three-dimensional model of a sequence that is aligned with related structures. 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 act upon main-chain hydrogen bonds, attractive van der Waals contacts, and main-chain and side-chain dihedral angles. The optimization relies on the program X-PLOR [109] and consists of molecular dynamics with simulated annealing. The method has been applied to two domains from the dehydrogenase family.

I now focus on the modelling approach of Šali and Blundell [2,5,22•,79•,100,110•]. The question addressed is 'What is the most probable structure for a certain sequence given its alignment with related structures?' Our approach follows from the method for comparison of protein structures implemented in the program COMPARER [100,111] and was developed to use as many different types of data about the target sequence as possible. It is implemented in the computer program MODELLER (which is available by anonymous ftp from tammy.harvard.edu:pub/modeller and also as part of QUANTA [MSI, Burlington, Massachusetts, USA; E-mail: jcollins@msi.com]). The input to the program is an alignment of the target sequence with related known three-dimensional structures. The output, obtained without any user intervention, is a three-dimensional 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 three-dimensional 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 three-dimensional structures, was constructed [79•] 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 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 a 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 [112] are combined into an objective function. Finally, the model is obtained by

optimizing the objective function in Cartesian space. This optimization is carried out using the variable target function method [113], employing methods of conjugate gradients and molecular dynamics with simulated annealing. Several slightly different models can be calculated by varying the initial structure.

One of the strengths of modelling by satisfaction of spatial restraints is that constraints or restraints derived from a number of different sources could easily be added to the homology-derived restraints. For example, restraints can be provided by rules for secondary-structure packing [86,114–116], analyses of hydrophobicity [117,118•] and correlated mutations [119,120], empirical potentials of mean force [121], nuclear magnetic resonance (NMR) experiments [122,123•], cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, site-directed mutagenesis [124], intuition, *et cetera*. In this way, a homology 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.

Modelling of loops

Loops can be calculated by searching a structure database for segments that fit on fixed endpoints [32], by conformational search with an optional energy minimization [125–127], or by a combination of these approaches [128,129]. Many different implementations of the basic techniques have been described [32,125–137,138••,139–144,145•,146,147•,148,149].

Collura and colleagues [141,142] used Monte Carlo and simulated-annealing algorithms to optimize a united atom energy function for a loop that spans given anchor regions. The energy function consists of non-bonded atomic interactions and a harmonic potential applied to terminal residues to force the loop closure. The degrees of freedom include only the main-chain and side-chain dihedral angles, excluding the ω dihedral angle. The optimization started from a completely extended conformation. For loops seven residues long, the average rms error was 1 Å for main-chain atoms and 2.3 Å for all heavy atoms.

Rao and Teeter [139], who also relied on a united atom model, optimized the energy of a single turn by a molecular dynamics procedure, as implemented in both AMBER [150] and X-PLOR [139]. An incorrect starting conformation changed into approximately correct conformation, as seen in the refined X-ray structure. It was shown that, in contrast to the original model, the predicted turn conformation refined with the X-ray data in fewer cycles, without any manual intervention, and with better refinement statistics.

Zheng *et al.* [144] described a new method for loop closure that started with all bonds scaled so that a random starting confirmation fitted on the anchor regions. The loop was then relaxed to its standard geometry. The predictions were enhanced by taking into account the protein environment of the loop. The method has been combined with multiple copy sampling to increase its efficiency by up to a factor of five [145•]. It has also been demonstrated that the variability in the predicted loop conformations can be used to estimate the accuracy of the models. In further development, the technique has been applied to model more than one loop at the same time [151•]. As a result, more accurate predictions were invariably obtained. The rms errors for 5–7 residue loops ranged from 0.6–1.7 Å for backbone atoms.

Srinivasan and Blundell [23] have described a collar extension idea for modelling loops. This relies on an equivalent loop from a homologue that differs by one or two residues in length. The equivalent part is copied from the template to the target and the remaining short gap is modelled by the database search approach, as described above [32].

Topham *et al.* [136] improved selection of the correct loop from an ensemble of candidate loops that already fit relatively well on the two anchor regions. This was achieved using three-dimensional profiles; candidate loops were ranked by a scoring function based on three-dimensional profiles that evaluated the compatibility between each residue in the target sequence and the environment implied by the structure of a candidate loop. The criteria included in the three-dimensional profiles were main-chain conformation, solvent accessibility, hydrogen bonding, disulphide bonding, and *cis*-peptide conformation.

Fidelis *et al.* [138••] have compared database and conformational search methods for loop modelling. They show that little correlation exists between the similarity in the anchor and loop regions of two segments and that the database of segments is sparse for segments longer than eight residues. The systematic search procedure can generate almost all structures of short segments in proteins and is thus the preferred method for modelling loops.

A new type of method, based on the self-consistent field approximation that was previously applied to side-chain modelling [152••], has been described by Koehl and Delarue [153••]. The method uses a database search scheme to generate possible main-chain fragments for modelling loops [32] and a rotamer library to define possible side-chain conformations [154]. It then iteratively refines the probabilities that the backbone and side chain of each residue correspond to database fragment j and rotamer k, respectively. Each residue experiences the average of all possible environments. The energy function includes only van der Waals terms, but can clearly be extended to include other terms, such as hydrogen bonds and solvation. The method usually converges to a single answer very close to one of the template structures. The self-consistent field method can be seen as the way to select one of the segments with which the target sequence is aligned. In principle, the method could be used to model whole structures.

Sudarsanam *et al.* [147•] have described a method for modelling loops on a given framework. Of the order

of 10 000 loop conformations are generated for each loop. Starting from the amino terminus of the loop, conformations are generated by assigning randomly selected pairs of Φ_{i+1}, Ψ_i for each dimer, using standard geometry and *trans* peptide configuration. The predicted loop is the conformation that maximally satisfies the loop closure condition and does not have any atom–atom overlaps. Additional filters, such as disulphide bonds, can easily be imposed on the construction of the loops.

The conformational properties of tight two-residue β-turns have been examined by analyzing 3899 examples in 205 protein chains [155•] and by empirical energy function calculations [156•]. It was shown that the conformation of such turns is determined by the twist of the β-sheet and a local electrostatic effect, and that the conformation can be modelled well when the rest of the protein and crystal water molecules are included in the calculations [156•]. Borchert *et al.* [149] modelled a seven-residue loop in a monomeric triosephosphate isomerase fold using program ICM, a general tool for conformational search in the dihedral angle space guided by a detailed energy function [157••]. The predicted loop had an rms of only 1.2 Å for the 28 main-chain atoms. These successes are in agreement with the analysis of Fidelis *et al.* [138••], who showed that loops shorter than seven residues can often be modelled correctly. This is probably the result of a relatively small number of loop conformations consistent with given anchor regions. Database search methods do not work well for loops longer than eight residues because the database is not likely to contain an example of a loop being modelled and because the correlation between the similarity of the anchor and loop regions is very weak [140], so that even if the correct conformation were in the database, it could not be easily identified. Possible reasons for failure of conformational search methods to model loops longer than eight residues include insufficient accuracy of the energy function, inadequate sampling of the phase space, and failure to include enough of the loop environment in the optimization. Fortunately, few insertions in a family of homologous proteins are longer than eight residues [14,158,159].

Modelling of side chains

As for loops, side-chain conformation has been predicted from similar structures, from proteins in general, and from steric or energy considerations [31,83,98,99•,154, 160–169,170••,171,172,173•–176•,177,178,179••,180•–182•]. The geometry of disulphide bridges has been modelled from disulphide bridges in protein structures in general [183–187,188•] and from equivalent disulphide bridges in related structures [79•]. For information on modelling the stability and conformation of point mutations by free energy perturbation simulations, the reader is referred elsewhere [189–192].

Dunbrack and Karplus [169] have developed a backbone-dependent rotamer library for amino acid side chains, using it to construct side-chain conformations from main-chain coordinates. They found significant correlations between side-chain dihedral angle probabilities and backbone Φ, Ψ values. These correlations go beyond the dependence of side-chain conformation on the secondary structure [164]. This automated method first places the side chains according to the rotamer library, and then removes steric clashes by energy minimization. It is also demonstrated that simple arguments based on conformational analysis can account for many features of the observed dependence of the side-chain rotamers on the backbone [170••].

Eisenmenger *et al.* [193] used program ICM [157••] to model side chains. Each side chain was configured in the environment of only the backbone atoms by a systematic search procedure combined with extensive local energy minimization of van der Waals, hydrogen-bond, torsional, and tether terms. Tests using main-chain atoms or both the main-chain and remaining side-chain atoms in the energy evaluations established the dominance of the main-chain contribution. The final model is obtained by a full energy refinement of the structure.

Tanimura *et al.* [181•] predicted side-chain conformations on a given backbone by a conformational search procedure relying either on side-chain–side-chain interactions or side-chain–main-chain interactions. In agreement with [193], removal of side-chain–side-chain interactions did not cause a large decrease in the prediction accuracy. Even so, the model having only side-chain–side-chain interactions still retained a significant level of accuracy. These results suggest that the two classes of interaction are consistent with each other and work in harmony to stabilize native conformations [194].

Wilson *et al.* [171] randomly picked local sites of adjacent side chains throughout the protein and evaluated all combinations of side-chain rotamers within each site using a molecular mechanics force field enhanced by the inclusion of a solvation term. At each site, the lowest energy combination of side chains is identified and added onto the fixed backbone. The procedure is repeated until side-chain conformations converge. The robustness of the method is evaluated by perturbing the backbone coordinates. The rms of predicted side chains rose from 1.3 Å in a test case with the correct backbone to 2.7 Å in one with <35% identity.

Vriend *et al.* [99•] have introduced SCAN3D, a new database system for integrated sequence–structure analysis. Site-dependent, side-chain rotamer distributions are obtained by extracting short segments with a given main-chain conformation from a database of protein structures. These rotamer distributions are then used in side-chain modelling. In a separate analysis, a set of predictive rules was derived that relied on the site-dependent rotamers and hydrogen-bonding criterion to explain 85% of point mutations currently available [173•].

Cregut *et al.* [174•] have tested three methods for side-chain prediction. The methods included molecular mechanics conformational search, the use of a rotamer database, and a combination of these two methods. It was

shown that the rotamer-based method is more efficient and that energy minimization before rotamer selection does not afford clearly improved predictions. It was also demonstrated that implicit solvation terms improve the predictions and that most errors can be identified by a combination of evaluation criteria, including solvation energy, rms deviations, χ_1 angle distributions, and hydrogen bonds.

Laughton [175•] has compared the local environment of each residue whose side chain is to be predicted with a database of local environments for the same residue type constructed from an analysis of high-resolution protein structures. Local environments are described in terms of the residue type and location of neighbouring residues that interact with the given side chain. The best few matches are inputted into a Monte Carlo procedure, which gives the final model by removing the steric clashes in the structure. In further development, the AMBER program has been used to explore a variety of simulated-annealing protocols and modifications of the united-atom force field for side-chain modelling [176•]. In this work, the modelling problems are generated by defining a percentage of side chains in a given X-ray structure as undefined.

The use of the dead-end elimination theorem in side-chain modelling (which was discussed previously [167]) has been re-examined [168]. The original theorem was meant to identify rotamers and pairs of rotamers that could not be part of the global energy minimum. This would allow a large reduction in the search space, thus leading to a possibility of a deterministic search for the global energy minimum. It was shown that the dead-end elimination theorem was not correct for rotamer pairs. Yet, a 'fuzzy' version of the theorem was proposed that still reduced the size of the search for the best conformations. It was used to help in modelling side-chain conformations given known backbone coordinates and a library of side-chain rotamers.

Koehl and Delarue [152••] have described a side-chain modelling method based on a rotamer library and a given backbone. It employs the self-consistent mean field approach. The method iteratively refines a conformational matrix of each side chain in a protein such that its current element *i, j* at each cycle gives the probability that the corresponding side-chain *i* adopts the conformation of its possible rotamer *j*. Each residue experiences the average of all possible environments, weighted by their respective probabilities. The final prediction corresponds to the rotamers with the highest probabilities. Estimates of the conformational entropy of side chains in the folded proteins are also given.

Lee [179••] has improved an earlier side-chain modelling method [177,178] by developing a new powerful optimization technique based on the self-consistent mean field approximation. Side-chains are built on a fixed backbone. The energetics of the system are described by Lennard–Jones and simple dihedral angle terms; no electrostatic or hydrogen-bonding terms are used. Side-chain dihedral angles are allowed to assume 33 different values. All possible conformations are initially set to have the same probability. The optimization then proceeds in the space of these probabilities, rather than in conformational space. The energy of each state for each dihedral angle is calculated as the mean field energy, given the current probabilities for all the other possible conformations. The probabilities are then recalculated using the Boltzmann distribution. These two steps are repeated many times at successively lower temperatures until the convergence in energy is achieved. To speed up the calculations, the mean field is not calculated exhaustively, but by Monte Carlo sampling of possible conformations. The method shares many features with that of Koehl and Delarue [152••], and they are both related to an approach to threading [43]. The robustness, speed, and size dependence of the new optimizer compare favourably with an earlier simulated-annealing method [178]. The method was used for calculating both protein thermostability differences and side-chain conformations. The calculated thermostability of the hydrophobic core mutants of λ repressor compared very well with experimental data. Similarly, side-chain conformation was predicted reliably; for example, flavodoxin core side chains were modelled with an rms error of 1 Å when the crystallographically determined backbone of flavodoxin was used.

Although the solvation term is irrelevant for the modelling of core side chains, it is important for the modelling of exposed side chains [83,171,174•]. It has also been demonstrated that treating hydrogen bonds explicitly can significantly improve side-chain prediction [169,173•]. It appears that relatively fine sampling of the dihedral angle space is necessary to model some side chains; for example, Schrauber *et al.* [195] reported that 5–30% of the side chains, depending on the residue type, are substantially different (by >20°) from common rotameric states in highly resolved structures. Another point relevant for homology modelling is that the best side-chain conformation depends relatively strongly on the backbone conformation of the residue [99•,169,181•,195]. For example, the preferred rotamers can vary within the same secondary structure, with the changes in the Φ, Ψ dihedral angles as small as 20° [169]. Because these changes are smaller than the differences between closely related homologues, the prediction of the side-chain conformation generally cannot be uncoupled from backbone prediction. This partly explains why conformation of equivalent side chains in homologous structures is useful in side-chain modelling [5]. This is consistent with the X-ray structure of a variant of λ repressor which reveals that the protein accommodates the potentially disruptive residues with shifts in its α-helical arrangement and with only limited changes in side-chain orientations [196•].

Model evaluation

It is clearly necessary for a good model to have a low energy according to a molecular mechanics force

field, such as that of CHARMM [112]. Correspondingly, stereochemical tests have been incorporated in PROCHECK [197], a program that is based on the analysis of known protein structures [198]. The local criteria checked by PROCHECK include the distribution of main-chain and side-chain dihedral angles, and the geometry of bonds, angles, improper dihedral angles, planes and chiral centres. The seminal work by Novotny *et al.* [199,200] showed, however, that low molecular mechanics energy does not ensure that the model is correct. Thus, distributions of many spatial features have been compiled from high-resolution protein structures, and any large deviations from the most likely values have been interpreted as strong indicators of errors in the model. Such features include packing [201], creation of a hydrophobic core [202], residue and atomic solvent accessibilities [203–206,207•], spatial distribution of charged groups [208], distribution of atom–atom distances [209], and main-chain hydrogen bonding [197].

Another group of methods for testing three-dimensional models, which implicitly takes into account many of the criteria listed above, involves three-dimensional profiles or threading. These methods evaluate the environment of each residue in a model with respect to the expected environment, as found in the high-resolution X-ray structures. The programs implementing this approach include 3DPROFILE [210], PROSAII [211,212], THREADER [46], MATCHMAKER [45], and HARMONY [213•]. In principle, an improvement in the accuracy of a model is possible by incorporating the quality criteria into a scoring function being optimized to derive the model in the first place.

Recently, protein modellers were challenged to model sequences without available three-dimensional structures and to submit them to the first 'Meeting on Critical Assessment of Techniques for Protein Structure Prediction' in Asilomar in December of 1994. At the same time, the three-dimensional structures were being determined by X-ray crystallography and NMR methods. Because these structures were only released at the meeting, it was possible to test the modelling methods objectively. The evaluation of comparative modelling can be summarized as follows [22•]. In general, the best comparative techniques can produce models with good stereochemistry and overall structural accuracy that is as high as the similarity between the template and the actual target structures. The errors can be divided into four categories: first, errors in side-chain packing; second, distortions or shifts of a region that is aligned correctly with the templates (e.g. loops, helices and strands); third, distortions or shifts of a region that does not have an equivalent segment in any of the templates (e.g. inserted loops); and fourth, distortions or shifts of a region that is aligned incorrectly with the templates (e.g. loops and larger segments with low sequence identity to the templates). The last three of these errors are relatively infrequent when sequences with >40% identity to the templates are modelled. For example, in such a case, approximately 90% of the main-chain atoms are likely to be modelled with an rms error of ~1 Å [22•]. Below 40% sequence identity, misalignments and insertions in the target sequence become the major problems. Insertions longer than about eight residues cannot be modelled accurately at this time, even when the alignment of the stem regions delimiting the insertion is correct. Most of the insertions shorter than eight residues also cannot be modelled successfully, primarily because the alignment of the inserted and neighbouring residues is frequently incorrect. If the length of an insertion can be extended enough to make the alignment of the delimiting stem regions reliable (but not too much, so that less than eight residues are inserted) the insertions can frequently be modelled sucessfully [138••,156•,157••]. In general, it can be expected that ~20% of the residues will be misaligned, and consequently incorrectly modelled, when the level of sequence identity between the target and templates is 30% [73].

To put errors into perspective, the differences among experimentally determined structures of the same protein can be compared. The 1 Å accuracy of main-chain atom positions corresponds to X-ray structures defined at a resolution of ~2.5 Å and with an R-factor of ~25% [214•] as well as to NMR structures determined from 10 inter-proton distance restraints per residue [215,216•]. Similarly, differences between the highly refined X-ray and NMR structures of the same protein also tend to be ~1 Å [215]. Changes in the environment (e.g. crystal packing, solvent and ligands) can also have a significant effect on the structure [217]. Overall, homology modelling based on templates with >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 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 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 [218].

We need a standardized, centralized, and comprehensive suite of model tests in order to bench-mark existing methods and to aid in the development of new methods. Alignment as well as modelling of side chains, loops, and whole structures should be tested in an automated way. The first step in this direction is the 'Biotech Validation Suite for Protein Structures' accessible at the World-Wide Web (WWW) address http://www.embl-heidelberg.de:8400.

Conclusions

The existing comparative modelling techniques can be used in an automated way and without any subjective decisions, provided templates with at least 40% sequence identity are known; no significant improvement of such

models is achieved by subjective interventions. On the other hand, for sequences with sequence identity <40%, large errors in the alignment can sometimes be prevented by examining and editing the alignment manually. In general, models have good stereochemistry and overall structural accuracy that is as high as the similarity between the template and the actual structure being predicted. As a result, the number of sequences that can be modelled is an order of magnitude larger than the number of experimentally determined protein structures, and the accuracy of a large fraction of these models is in many ways comparable to the accuracy of low-resolution X-ray structures and medium-resolution NMR structures. That is, >90% of main-chain atoms can be modelled with an accuracy of ~1 Å, provided a template structure with at least 40% sequence identity is available. The errors in different regions of the model can be estimated by a variety of evaluation techniques.

Future improvements of comparative modelling should aim to model proteins with lower similarities to known structures, to increase the accuracy of the models, and to make modelling fully automated. The improvements are likely to include the simultaneous optimization of side-chain and backbone conformations in side-chain modelling, and simultaneous optimization of a loop and its environment in loop modelling. At the same time, better potential functions and possibly better optimizers are needed. The potential function should guide the model away from the templates 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 [5], 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 [45,46], are needed.

Even though comparative modelling needs significant improvements, it is already a mature technique that can be used to address many practical problems. Some successful predictions include identification of the heparin-binding site in the mouse mast cell tryptases [219•], design of micromolar inhibitors of the malarial cysteine protease [220], prediction and conversion of substrate specificity of granzyme B [221], and solution of a molecular replacement problem in X-ray crystallography [222]. With the increase in the number of protein sequences and in the fraction of all folds that are known, comparative modelling will become even more useful in the future.

Acknowledgements

I am grateful to John Overington, Tom Blundell, Martin Karplus, and Mark Johnson for discussions concerning protein modelling. I also thank Daša Šali for her comments on this manuscript.

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. 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.

2. Šali A, Overington JP, Johnson MS, Blundell TL: **From comparisons of protein sequences and structures to protein modelling and design.** *Trends Biochem Sci* 1990, **15**:235–240.

3. Greer J: **Comparative modelling methods: application to the family of the mammalian serine proteases.** *Proteins* 1990, **7**:317–334.

4. Swindells MB, Thornton JM: **Modelling by homology.** *Curr Opin Struct Biol* 1991, **1**:219–223.

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

6. •• Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL: **Knowledge-based protein modelling.** *CRC Crit Rev Biochem Mol Biol* 1994, **29**:1–68.
A comprehensive review of homology modelling and some other relevant techniques.

7. • Bajorath J, Stenkamp R, Aruffo A: **Knowledge-based model building of proteins: concepts and examples.** *Protein Sci* 1994, **2**:1798–1810.
Comparative modelling by assembly of rigid bodies is reviewed.

8. • May ACW, Blundell TL: **Automated comparative protein modelling of protein structures.** *Curr Opin Biotechnol* 1994, **5**:355–360.
A short review of comparative modelling.

9. • Holm L, Rost B, Sander C, Schneider R, Vriend G: **Data based modeling of proteins.** In *Statistical Mechanics, Protein Structure, and Protein Substrate Interactions.* Edited by Doniach S. New York: Plenum Press; 1994:277–296.
Databases of protein structures and comparative modelling are reviewed.

10. Lesk AM, Chothia CH: **The response of protein structures to amino-acid sequence changes.** *Phil Trans Roy Soc* 1986, **317**:345–356.

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

12. Sander C, Schneider R: **Database of homology-driven protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9**:56–68.

13. Hilbert M, Bohm G, Jaenicke R: **Structural relationships of homologous proteins as a fundamental principle in homology modelling.** *Proteins* 1993, **17**:138–151.

14. Flores TP, Orengo CA, Moss DS, Thornton JM: **Comparison of conformational characteristics in structurally similar protein pairs.** *Protein Sci* 1993, **2**:1811–1826.

15. •• 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.
Side-chain–side-chain contacts, accessibility, secondary structure, and rms deviations are compared within 607 pairs of proteins having similar three-dimensional structures. This should be useful in determining the invariants in protein structure and thus facilitate the design of better comparative modelling methods.

16. • Chelvanayagam G, Roy G, Argos P: **Easy adaptation of protein structure to sequence.** *Protein Eng* 1994, **7**:173–184.
Analysis of conservation of coarse, medium, and fine-grain structural properties is performed over a data set of 175 protein tertiary structures in 34 different families.

17. Kolinski A, Skolnick J: **Monte Carlo simulations of protein folding. I Lattice model and interaction scheme.** *Proteins* 1994, **18**:338–352.
• Describes the latest version of these authors' lattice Monte Carlo folding algorithm based on a hierarchy of models of increasing detail. The potential of mean force, predominantly of statistical origin, contains several novel terms that facilitate the cooperative assembly of secondary structure elements and the cooperative packing of the side chains.

18. Vasquez M, Nemethy G, Scheraga HA: **Conformational energy calculations on polypeptides and proteins.** *Chem Rev* 1994, **94**:2183–2239.
• A comprehensive review of techniques for optimization of energy functions.

19. Ring CS, Cohen FE: **Modeling protein structures: construction and their applications.** *FASEB J* 1993, **7**:783–790.

20. Benner SA, Gerloff DL, Jenny TF: **Predicting protein crystal structures.** *Science* 1994, **265**:1642–1644.
• A short summary of one approach to *ab initio* protein structure prediction.

21. Dandekar T, Argos P: **Folding the main chain of small proteins with the genetic algorithm.** *J Mol Biol* 1994, **236**:844–861.
• Folding of four-helix bundles is simulated using a genetic algorithm.

22. Šali A, Potterton L, Yuan F, Van Vlijmen H, Karplus M: **Evaluation of comparative protein modelling by MODELLER.** *Proteins* 1995, in press.
• An evaluation of three homology models submitted to the Asilomar conference on 'The Critical Assessment of Techniques for Protein Structure Prediction'. The models span a range of difficulty, from easy (human nucleoside diphosphate kinase), through medium (mouse cellular retinoic acid and binding protein I), to difficult (human eosinophil neurotoxin).

23. Srinivasan N, Blundell TL: **An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure.** *Protein Eng* 1993, **6**:501–512.

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

25. Orengo CA, Jones DT, Thornton JM: **Protein superfamilies and domain superfolds.** *Nature* 1994, **372**:631–634.
•• The relationships between proteins in sequence and structure databases are described. The number of different families is estimated and the growth in sequence and structure data is analyzed.

26. 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.

27. Warme PK, Momany FA, Rumball SV, Tuttle RW, Scheraga HA: **Computation of structures of homologous proteins: α-lactalbumin from lysozyme.** *Biochemistry* 1974, **13**:768–782.

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

29. Jones TA: **A graphics model building and refinement system for macromolecules.** *J Appl Crystallogr* 1978, **11**:268–272.

30. Sutcliffe MJ, Haneef I, Carney D, 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.

31. Sutcliffe MJ, Hayes FRF, Blundell TL: **Knowledge based modeling of homologous proteins, part II: rules for the conformation of substituted side-chains.** *Protein Eng* 1987, **1**:385–392.

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

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

34. 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.

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

36. Srinivasan S, Shibata M, Rein R: **Multistep modeling of protein structure: application to bungarotoxin.** *Int J Quantum Chem Quantum Biol Symp* 1986, **13**:167–174.

37. 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.

38. Wodak SJ, Rooman MJ: **Generating and testing protein folds.** *Curr Opin Struct Biol* 1993, **3**:247–259.

39. Sippl MJ: **Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures.** *J Comput Aid Mol Design* 1993, **7**:473–501.

40. Johnson MS, Overington JP, Blundell TL: **Alignment and searching for common protein folds using a data bank of structural templates.** *J Mol Biol* 1993, **231**:735–752.

41. Thornton JM, Flores TP, Jones DT, Swindells MB: **Prediction of progress at last.** *Nature* 1991, **354**:105–106.

42. 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.

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

44. Friedrichs MS, Goldstein RA, Wolynes PG: **Generalized protein tertiary structure recognition using associative memory Hamiltonians.** *J Mol Biol* 1991, **222**:1013–1034.

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

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

47. Sippl MJ, Weitckus S: **Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations.** *Proteins* 1992, **13**:258–271.

48. Maiorov VN, Crippen GM: **Contact potential that recognizes the correct folding of globular proteins.** *J Mol Biol* 1992, **227**:876–668.

49. Bryant SH, Lawrence CE: **An empirical energy function for threading protein sequence through the folding motif.** *Proteins* 1993, **16**:92–112.

50. Ouzounis C, Sander C, Scharf M, Schneider R: **Prediction of protein structure by evaluation of sequence-structure fitness: aligning sequences to contact profiles derived from three-dimensional structures.** *J Mol Biol* 1993, **232**:805–825.

51. Lathrop RH: **The protein threading problem with sequence amino acid interaction preferences is NP-complete.** *Protein Eng* 1994, **7**:1059–1068.
• Computational complexity of the threading problem is investigated.

52. Abagyan R, Frishman D, Argos P: **Recognition of distantly related proteins through energy calculations.** *Proteins* 1994, **19**:132–140.
• Describes a new template matching method that does not rely on a scoring function determined by the use of a database of known structures. Instead, direct energy calculations are performed.

53. Matsuo Y, Nishikawa K: **Protein structural similarities predicted by a sequence-structure compatibility method.** *Protein Sci* 1994, **3**:2055–2063.

54. Holm L, Sander C: **Searching protein structure databases has come of age.** *Proteins* 1994, **19**:165–173.
• Structure comparison methods are reviewed.

55. Blundell TL, Johnson MS: **Catching a common fold.** *Protein Sci* 1993, **2**:877–883.

56. Barton G: **Protein sequence alignment and database scanning.** In *Protein Structure Prediction: A Practical Approach.* Edited by

Steinberg MJE. Oxford: IRL Press at Oxford University Press; 1995:in press.

57. Altschul SF, Boguski MS, Gish W, Wootton JC: **Issues in searching molecular sequence databases.** *Nature Genet* 1994, **6**:119–129.

58. George DG, Barker WC, Hunt LT: **The protein identification resource.** *Nucleic Acids Res* 1986, **14**:11–15.

59. Burks HS, Burks C: **The Genbank sequence data bank.** *Nucleic Acids Res* 1988, **15**:1861–1864.

60. Bairoch A, Boeckmann B: **The SWISS-PROT protein sequence data bank.** *Nucleic Acids Res* 1991, **19**:2247–2249.

61. Hamm G, Cameron G: **The EMBL data library.** *Nucleic Acids Res* 1986, **14**:5–9.

62. 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, Bergerkoff G, Sievers R. Bonn: Data Commission of the International Union of Crystallography; 1987:107–132.

63. • Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247**:536–540.
This 'Structural Classification of Proteins' database allows homology and keyword searching. It is accessible on Internet as a WWW service at http://scop.mrc-lmb.cam.ac.uk/scop/.

64. Pearson WR: **Rapid and sensitive comparison with FASTA and FASTP.** *Methods Enzymol* 1990, **183**:63–98.

65. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–410.

66. Frishman D, Argos P: **Recognition of distantly related protein sequences using conserved motifs and neural networks.** *J Mol Biol* 1992, **228**:951–962.

67. • Henikoff S, Henikoff JG: **Protein family classification based on searching a database of blocks.** *Genomics* 1994, **19**:97–107.
The BLOCKS template searches are available as an Internet service at the WWW address http://www.blocks.fhcrc.org. This service can be used to find sequences related to a given sequence.

68. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology: applications to protein modeling.** *J Mol Biol* 1994, **235**:1501–1531.

69. Gribskov M: **Profile analysis.** *Methods Mol Biol* 1994, **25**:247–266.

70. Gracy J, Chiche L, Sallantin J: **Improved alignment of weakly homologous protein sequences using structural information.** *Protein Eng* 1993, **6**:821–829.

71. • Zhang KYJ, Eisenberg D: **The three-dimensional profile method using residue preference as a continuous function of residue environment.** *Protein Sci* 1994, **3**:687–695.
Describes an improvement of the original three-dimensional profile method that involves an expression of residue preferences as a continuous function of environmental variables. This leads to a more sensitive method for detecting remote sequence–structure relationships.

72. Orengo CA: **A review of methods for protein structure comparison.** In *Patterns in Protein Sequence and Structure, Springer Series in Biophysics,* vol 7. Edited by Taylor WR. Heidelberg: Springer-Verlag; 1992:155–188.

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

74. Pascarella S, Argos P: **A data bank merging related protein structures and sequences.** *Protein Eng* 1992, **5**:121–137.

75. Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G: **A database of protein structure families with common folding motifs.** *Protein Sci* 1992, **1**:1691–1698.

76. Orengo CA, Flores TP, Taylor WR, Thornton JM: **Identification and classification of protein fold families.** *Protein Eng* 1993, **6**:485–500.

77. Subbiah S, Laurents DV, Levitt M: **Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core.** *Curr Biol* 1993, **3**:141–148.

78. Overington JP, Zhu Z-Y, Šali A, Johnson MS, Sowdhamin R, Louie GV, Blundell TL: **Molecular recognition in protein families: a database of aligned three-dimensional structures of related proteins.** *Biochem Soc Trans* 1993, **21**:597–604.

79. • Šali A, Overington JP: **Derivation of rules for comparative protein modeling from a database of protein structure alignments.** *Protein Sci* 1994, **3**:1582–1596.
A database of protein structure alignments and tools for its use are described. The database is applied to derive spatial restraints on disulphide bridges and *cis/trans* isomerism of proline residues that are used for comparative modelling by MODELLER.

80. J Felsenstein: **Confidence limits on phylogenics: an approach using the bootstrap.** *Evolution* 1985, **39**:783–791.

81. Johnson MS, Overington JP, Šali A: **Knowledge-based protein modelling: human plasma kallikrein and human neutrophil defensin.** In *Current Research in Protein Chemistry: Techniques, Structure and Function.* Edited by Villafranca JJ. San Diego: Academic Press, Inc; 1990:567–574.

82. Robson B, Platt E, Fishleigh RV, Marsden A, Millard P: **Expert system for protein engineering: its application in the study of chloramphenicol acetyltransferase and avian pancreatic polypeptide.** *J Mol Graph* 1987, **5**:5–17.

83. Schiffer CA, Caldwell JW, Kollman PA, Stroud RM: **Prediction of homologous protein structures based on conformational searches and energetics.** *Proteins* 1990, **8**:30–43.

84. Stewart DE, Weiner PK, Wampler JE: **Prediction of the structure of proteins using related structures, energy minimisation and computer graphics.** *J Mol Graph* 1987, **5**:133–140.

85. Kajihara A, Komooka H, Kamiya K, Umeyama H: **Protein modelling using a chimera reference protein derived from exons.** *Protein Eng* 1993, **6**:615–620.

86. Reddy BVB, Blundell TL: **Packing of secondary structural elements in proteins analysis and prediction of inter-helix distances.** *J Mol Biol* 1993, **233**:464–479.

87. 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.

88. Vasquez M, Scheraga HA: **Calculation of protein conformation by the build-up procedure. Application to bovine pancreatic trypsin inhibitor using limited simulated nuclear magnetic resonance data.** *J Biomol Struct Dynam* 1988, **5**:705–755.

89. Reid LS, Thornton JM: **Rebuilding flavodoxin from C_α coordinates: a test study.** *Protein* 1989, **5**:170–182.

90. Holm L, Sander C: **Database algorithm for generating protein backbone and side-chain co-ordinates from C_α trace: application to model building and detection of co-ordinate errors.** *J Mol Biol* 1991, **218**:183–194.

91. Wendoloski JJ, Salemme FR: **PROBIT: a statistical approach to modeling proteins from partial coordinate data using substructure libraries.** *J Mol Graph* 1992, **10**:124–126.

92. Bassolino-Klimas D, Bruccoleri RE: **Application of a directed conformational serach for generating 3-D coordinates for protein structures from α-carbon coordinates.** *Proteins* 1992, **14**:465–474.

93. Correa PE: **The building of protein structures from α-carbon coordinates.** *Proteins* 1990, **7**:366–377.

94. Luo Y, Jiang X, Lai L, Qu C, Xu X, Tang Y: **Building protein backbones from C_α coordinates.** *Protein Eng* 1992, **5**:147–150.

95. Rey A, Skolnick J: **Efficient algorithm for the reconstruction of a protein backbone from the α-carbon coordinates.** *J Comput Chem* 1992, **13**:443–456.

96. Payne PW: **Reconstruction of protein conformations from estimated positions of the C_α coordinates.** *Protein Sci* 1993, **2**:315–324.

97. • Van Gelder CWG, Leusen FJJ, Leunissen JAM, Noordik JH: **A molecular dynamics approach for the generation of complete protein structures from limited coordinate data.** *Proteins* 1994, **18**:174–185.

A new method to build a complete protein structure from C_α coordinates is presented. An approximate backbone is first generated using geometrical criteria only. Then, the backbone is refined and side chains are placed through the use of exhaustive molecular dynamics simulation. The method is used to generate full-atom models of two proteins from their low-resolution C_α traces.

98. Holm L, Sander C: **Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology.** *Proteins* 1992, **14**:213–223.

99. • Vriend G, Sander C, Stouten PFW: **A novel search method for protein sequence-structure relations using property profiles.** *Protein Eng* 1994, **7**:23–29.

SCAN3D, a new database system for integrated sequence-structure analysis is introduced. Site-dependent side-chain rotamer distributions can be easily obtained by extracting short segments with a given main-chain conformation. These rotamer distributions can then be used in side-chain modelling.

100. Šali A, Blundell TL: **Definition of general topological equivalence in protein structures: a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming.** *J Mol Biol* 1990, **212**:403–428.

101. Fujiyoshi-Yoneda T, Yoneda S, Kitamura K, Amisaki T, Ikeda K, Inoue M, Ishida T: **Adaptability of restrained molecular dynamics for tertiary structure prediction: application to *Crotalus atrox* venom phospholipase A_2.** *Protein Eng* 1991, **4**:443–450.

102. Engh RA, Wright HT, Huber R: **Modeling the intact form of the α-proteinase inhibitor.** *Protein Eng* 1990, **3**:469–477.

103. Bohr H, Bohr J, Brunak S, Cotterill RMJ, Fredholm H, Lautrup B, Petersen SB: **A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks.** *FEBS Lett* 1990, **261**:43–46.

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

105. Snow ME: **A novel parameterization scheme for energy equations and its use to calculate the structure of protein molecules.** *Proteins* 1993, **15**:183–193.

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

107. • Sudarsanam S, March CJ, Srinivasan S: **Homology modeling of divergent proteins.** *J Mol Biol* 1994, **241**:143–149.

The authors improve an earlier method by relaxing distance constraints on the target sequence that are transferred from the template structure. This facilitates three-dimensional embedding and energy minimization, and increases the rms between the template and the model; however, it does not appear to increase the accuracy of the models relative to the template structure.

108. 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 method for the prediction of protein tertiary structure.** *Protein Sci* 1993, **2**:626–639.

109. Brünger AT: *XPLOR Manual Version 2.1.* New Haven, Connecticut: Yale University; 1990.

110. • Šali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints.** In *Protein Structure by Distance Analysis.* Edited by Bohr H, Brunak S. Amsterdam: IOS Press; 1994:64–86.

An update on MODELLER that describes the use of the CHARMM force field (see [112]) to restrain the stereochemistry of the comparative models.

111. Zhu Z-Y, Šali A, Blundell TL: **A variable gap penalty function and feature weights for protein 3-D structure comparisons.** *Protein Eng* 1992, 5:43–51.

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

113. Braun W, Gō N: **Calculation of protein conformations by proton–proton distance constraints: a new efficient algorithm.** *J Mol Biol* 1985, **186**:611–626.

114. Cohen FE, Kuntz ID: **Tertiary structure prediction.** In *Prediction of Protein Structure and the Principles of Protein Conformation.* Edited by Fasman GD. New York: Plenum Press; 1989:647–705.

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

116. Tuffery P, Lavery R: **Packing and recognition of protein structural elements: a new approach applied to the 4-helix bundle of myoheherythrin.** *Proteins* 1993, **15**:413–425.

117. Saitoh S, Nakai T, Nishikawa K: **A geometrical constraint approach for reproducing the native backbone conformation of a protein.** *Proteins* 1993, **15**:191–204.

118. • Aszodi A, Taylor WR: **Secodary structure formation in model polypeptide chains.** *Protein Eng* 1994, **7**:633–644.

Inter-residue distances are predicted from hydrophobicity of residues and are used with the distance geometry technique to fold polypeptide chains into compact conformations.

119. Taylor WR, Hatrick K: **Compensating changes in protein multiple sequence alignments.** *Protein Eng* 1994, **7**:341–348.

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

121. 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.

122. Sutcliffe MJ, Dobson CM, Oswald RE: **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* 1992, **31**:2962–2970.

123. • Haqq CM, King CY, Ekiyama E, Falsafi S, Haqq TN, Donahoe PK, Weiss MA: **Molecular basis of mammalian sexual determination: activation of mullerian inhibiting substance gene expression by SRY.** *Science* 1994, **266**:1494–1500.

Demonstrates the use of homology-derived restraints in NMR structure refinement.

124. 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

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

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

127. 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.

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

129. 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.

130. Summers NL, Karplus M: **Modeling of globular proteins: a distance-based search procedure for the construction of**

insertion/deletion regions and pro→non-pro mutations. *J Mol Biol* 1990, **216**:991–1016.

131. Bruccoleri RE, Haber E, Novotny J: **Structure of antibody hypervariable loops reproduced by a conformational search algorithm.** *Nature* 1988, **335**:564–568.

132. Sibanda BL, Blundell TL, Thornton JM: **Conformation of β-hairpins in protein structures: a systematic classification with applications to modelling by homology, electron deosity fitting and protein engineering.** *J Mol Biol* 1989, **206**:759–777.

133. Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR *et al.*: **Conformation of immunoglobulin hypervariable regions.** *Nature* 1989, **342**:877–883.

134. Dudek MJ, Scheraga HA: **Protein structure prediction using a combination of sequence homology and global energy minimization. I. Global energy minimization of surface loops.** *J Comput Chem* 1990, **11**:121–151.

135. Mas MT, Smith KC, Yarmush DL, Aisaka K, Fine RM: **Modeling the anti-CEA antibody combining site by homology and conformational search.** *Proteins* 1992, **14**:483–498.

136. Topham CM, McLeod A, Eisenmenger F, Overington JP, Johnson MS, Blundell TL: **Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables.** *J Mol Biol* 1993, **229**:194–220.

137. Pedersen J, Searle S, Henry A, Rees AR: **Antibody modeling: beyond homology.** *Immunomethods* 1992, **1**:126–136.

138. •• 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.
Database and conformational search methods for loop modelling are compared. It is demonstrated that little correlation exists between the similarity in the anchor and loop regions of two segments and that the database of segments is sparse for segments longer than eight residues. The systematic search procedure can generate almost all structures of short segments in proteins and is thus the preferred method for modeling loops.

139. Rao U, Teeter MM: **Improvement of turn structure prediction by molecular dynamics: a case study of α-purithionin.** *Protein Eng* 1993, **6**:837–847.

140. Tramontano A, Lesk AM: **Common features of the conformations of antigen-binding loops in immunoglobulins and application to modeling loop conformations.** *Proteins* 1992, **13**:231–245.

141. Collura V, Higo J, Garnier J: **Modeling of protein loops by simulated annealing.** *Protein Sci* 1993, **2**:1502–1510.

142. Higo J, Collura V, Garnier J: **Development of an extended simulated annealing method: application to the modeling of complementary determining regions of immunoglobulins.** *Biopolymers* 1992, **32**:33–43.

143. Bassolino-Klimas D, Bruccoleri RE, Subramaniam S: **Modeling the antigen combining site of anti-dinitrophenyl antibody, ANO2.** *Protein Sci* 1992, **1**:1465–1476.

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

145. • Zheng Q, Rosenfeld R, DeLisi C, Kyle DJ: **Multiple copy sampling in protein loop modeling: computational efficiency and sensitivity to dihedral angle perturbations.** *Protein Sci* 1994, **3**:493–506.
An original method for loop modelling by the same authors is combined with multiple copy sampling to increase its efficiency by up to a factor of five. It is also shown that the variability in the predicted loop conformations can be used to estimate the accuracy of the models.

146. Shenkin PS, Yarmush DL, Fine RM, Wang H, Levinthal C: **Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ring-like structures.** *Biopolymers* 1987, **26**:2053–2085.

147. • Sudarsanam S, DuBose RF, March CJ, Srinivasan S: **Modeling protein loops using a ϕ_{i+1}, ψ_i dimer database.** *Protein Sci* 1995, in press.
A method for modelling loops on a given framework is described. Possible loop conformations are generated by randomly combining dimers from a database of dimers. The predicted loop is the conformation that maximally satisfies the loop closure condition and does not have any atom–atom overlaps. Additional filters, such as disulphide bonds, can be easily imposed on the construction of loops.

148. Bruccoleri RE: **Application of systematic conformational search to protein modeling.** *Mol Simulat* 1993, **10**:151–174.

149. Borchert TV, Abagyan RA, Kishan KVR, Zeelen JP, Wierenga RK: **The crystal structure of an engineered monomeric triosephosphate isomerase, monotim: the correct modelling of an eight residue loop.** *Structure* 1993, **1**:205–213.

150. Weiner SJ, Kollman PA, Case DA, Singh VC, Ghio C, Alagona G, Profeta S, Weiner PA: **AMBER: assisted model building with energy refinement. A general program for modelling molecules and their interactions.** *J Comput Chem* 1981, **2**:287–303.

151. • Rosenbach D, Rosenfeld R: **Simultaneous modeling of multiple loops in proteins.** *Protein Sci* 1995, **4**:496–505.
The bond-scaling-relaxation algorithm by Zheng *et al.* [144,145•] is used to model more than one loop at the same time. More accurate predictions are invariably obtained.

152. •• 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.
This method is based on a rotamer library and refines iteratively a conformational matrix of the side chain of a protein such that its current element *i, j,* at each cycle, gives the probability that the corresponding side-chain *i* adopts the conformation of its possible rotamer *j*. Each residue is influenced by the average of all possible environments, weighted by their respective probabilities. The final prediction corresponds to the rotamers with the highest probabilities. Estimates of the conformational entropy of side chain in the folded proteins are also given.

153. •• Koehl P, Delarue M: **A self consistent mean field approach to simultaneous gap closure and side-chain positioning in protein homology modelling.** *Nature Struct Biol* 1995, **2**:163–170.
The method described uses a database scheme to generate possible fragments for modelling gaps, and a rotamer library to define possible side-chain conformations. It then iteratively refines the probabilities that the backbone correponds to database fragment *j* and side chain corresponds to rotamer for each residue, which experiences the average of all possible environments. The energy function includes only van der Waals terms.

154. Tuffery P, Etchebest C, Hazout S, Lavery RA: **A new approach to the rapid determination of protein side chain conformations.** *J Biomol Struct Dynam* 1991, **8**:1267–1289.

155. • Hutchinson EG, Thornton JM: **A revised set of potentials for β-turn formation in proteins.** *Protein Sci* 1994, **3**:2207–2216.
3899 β-turns from 205 protein structures are used to derive β-turn position potentials. Many positional preferences can be rationalized in terms of hydrogen bonds, preferences for amino acids to adopt a particular conformation in ϕ, ψ space, and the involvement of some turns in β-hairpins.

156. • Mattos C, Petsko GA, Karplus M: **Analysis of two-residue turns in proteins.** *J Mol Biol* 1994, **238**:733–747.
The conformational properties of tight two-residue β-turns are examined by empirical energy function calculations. It is shown that the conformation of such turns is determined by the twist of the β-sheet and a local electrostatic effect.

157. •• Abagyan R, Totrov M: **Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins.** *J Mol Biol* 1994, **235**:983–1002.
Describes a general method for calculating protein three-dimensional structures using a detailed energy function and external restraints. The method can be used for comparative modelling when atomic positions are restrained to those in template structures, loop modelling when regions delimiting short loops are defined, and side-chain modelling when the backbone is kept fixed.

158. Pascarella S, Argos P: **Analysis of insertions/deletions in protein structures.** *J Mol Biol* 1992, **224**:461–471.

159. Benner SA, Gonnet GH, Cohen MA: **Empirical and structural models for insertions and deletions in the divergent evolution of proteins.** *J Mol Biol* 1993, **229**:1065–1082.

160. 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.

161. Summers NL, Carson WD, Karplus M: **Analysis of side chain orientations in homologous proteins.** *J Mol Biol* 1987, **196**:175–198.

162. Summers NL, Karplus M: **Construction of side-chains in homology modelling: application to the C-terminal lobe of rhizopuspepsin.** *J Mol Biol* 1989, **210**:785–811.

163. Snow ME, Amzel LM: **Calculating three-dimensional changes in protein structure due to amino-acid substitutions: the variable region of the immunoglobulins.** *Proteins* 1986, **1**:267–279.

164. McGregor MJ, Islam SA, Sternberg MJE: **Analysis of the relationship between side-chain conformation and secondary structure in globular proteins.** *J Mol Biol* 1987, **198**:295–310.

165. Singh J, Thornton JM: **SIRIUS. An automated method for the analysis of the preferred packing arrangements between protein groups.** *J Mol Biol* 1990, **17**:195–225.

166. Tuffery P, Etchebest C, Hazout S, Lavery R: **A critical comparison of search algorithms applied to the optimization of protein side-chains conformations.** *J Comput Chem* 1993, **14**:790–798.

167. Desmet J, De Maeyer M, Hazes B, Lasters I: **The dead-end elimination theorem and its use in protein side-chain positioning.** *Nature* 1992, **356**:539–542.

168. Lasters I, Desmet J: **The fuzzy-end elimination theorem: correctly implementing the side-chain placement algorithm based on the dead-end elimination theorem.** *Protein Eng* 1993, **6**:717–722.

169. Dunbrack RL, Karplus M: **Prediction of protein side-chain conformations from a back-bone conformation dependent rotamer library.** *J Mol Biol* 1993, **230**:543–571.

170. •• Dunbrack RL, Karplus M: **Conformational analysis of the backbone-dependent rotamer preferences of protein side-chains.** *Nature Struct Biol* 1994, **1**:334–340.
It is demonstrated that simple arguments based on conformational analysis can account for many of the features of the observed backbone dependence of the side-chain rotamers.

171. Wilson C, Gregoret LM, Agard DA: **Modeling side-chain conformation for homologous proteins using an energy-based rotamer search.** *J Mol Biol* 1993, **229**:996–1006.

172. Roitberg A, Elber R: **Modeling side-chains in peptides and proteins: application of the locally enhanced sampling and the simulated annealing method to find minimum energy conformations.** *J Chem Phys* 1991, **95**:9277–9287.

173. • De Filippis V, Sander C, Vriend G: **Predicting local structural changes that result from point mutations.** *Protein Eng* 1994, **7**:1203–1208.
A set of predictive rules is derived that relies on the site-dependent rotamers and a hydrogen-bonding criterion to explain 85% of point mutations currently available.

174. • Cregut D, Liautard J-P, Chiche L: **Homology modeling of annexin I: implicit solvation improves side-chain prediction and combination of evaluation criteria allows recognition of different types of conformational error.** *Protein Eng* 1994, **7**:1333–1344.
Three methods for side-chain prediction are tested. They are based on a molecular mechanism conformational search, the use of a rotamer database, or a combination of these two methods. It is shown that implicit solvation terms improve the predictions and that most errors can be identified by a combination of evalutaion criteria, including solvation energy, rms deviations, χ_1 angles, and hydrogen bonds.

175. • Laughton CA: **Prediction of protein side-chain conformations from local three-dimensional homology relationships.** *J Mol Biol* 1994, **235**:1088–1097.
The method described involves the comparison of the local environment of each residue whose side chain is to be predicted with a database of local environments for the same residue type constructed from an analysis of high-resolution protein structures. Local environments are described in terms of the residue type and location of residues that interact with the given side chain. The best few matches are inputted into a Monte Carlo procedure, which gives the final model by removing the steric clashes in the structure.

176. • Laughton CA: **A study of simulated annealing protocols for use with molecular dynamics in protein structure protein.** *Protein Eng* 1994, **7**:235–241.
The AMBER program is used to explore a variety of simulated-annealing protocols and modifications of the united-atom force field for side-chain modelling. The modelling problems are generated by defining a percentage of side chains in a given X-ray structure as undefined.

177. Lee C, Levitt M: **Accurate prediction of the stability and activity effects of site directed mutagenesis on a protein core.** *Nature* 1991, **352**:448–451.

178. Lee C, Subbiah S: **Prediction of protein side-chain conformation by packing optimization.** *J Mol Biol* 1991, **217**:373–388.

179. •• Lee C: **Predicting protein mutant energetics by self consistent ensemble optimisation.** *J Mol Biol* 1994, **236**:918–939.
A self-consistent ensemble optimization is applied to predicting the conformation of side chains in the core of a protein and the effect of mutations on protein stability. A simple model based on steric interactions is used and a fixed backbone is assumed. The optimization method is superior to simulated annealing.

180. • Zheng Q, Kyle DJ: **Multiple copy sampling: rigid versus flexible protein.** *Proteins* 1994, **19**:324–329.
The effects of protein flexibility on multiple copy conformational sampling are systematically evaluated by studying placement of one side chain. The technique performs better when a flexible, but restrained, protein is used instead of a rigid protein.

181. • Tanimura R, Kidera A, Nakamura H: **Determinants of protein side-chain packing.** *Protein Sci* 1994, **3**:2358–2365.
The problem of side-chain packing for a given backbone is investigated using three different prediction models: rotamer search, minimizations of side-chain–side-chain interactions or side-chain–main-chain interactions. Conclusions about the factors determining side-chain conformation are drawn.

182. • Kono H, Doi J: **Energy minimization method using automata network for sequence and side-chain conformation prediction from given backbone geometry.** *Proteins* 1994, **19**:244–255.
A method is described for prediction of sequences and side-chain conformations for a given backbone. The method can be used for automated sequence generation in the *de novo* design of proteins.

183. Thornton JM: **Disulphide bridges in globular proteins.** *J Mol Biol* 1981, **151**:261–287.

184. Sowdhamini R, Srinivasan N, Shoichet B, Santi DV, Ramakrishnan C, Balaram P: **Stereochemical modeling of disulphide bridges. Criteria for introduction into proteins by site-directed mutagenesis.** *Protein Eng* 1989, **3**:95–103.

185. Sowdhamini R, Ramakrishnan C, Balaram P: **Modelling multiple disulphide loop containing polypeptides by random conformation generation. The test cases of α-conotoxin GI and endothelin.** *Protein Eng* 1993, **6**:873–882.

186. Pabo CO, Suchanek EG: **Computer aided model building strategies for protein design.** *Biochemistry* 1986, **25**:5987–5991.

187. Harrison PM, Sternberg MJ: **Analysis and classification of disulphide connectivity in proteins. The entropic effect of cross-linkage.** *J Mol Biol* 1994, **244**:448–463.

188. • Jung S-H, Pastan I, Lee B: **Design of interchain disulfide bonds in the framework region of the Fv fragment of the monoclonal antibody B3.** *Proteins* 1994, **19**:35–47.
Possible disulphide sites in an Fv fragment are identified and modelled. A method for modelling disulphide bridges on a given backbone is described.

189. Boresch S, Archontis G, Karplus M: **Free energy simulations: the meaning of the individual contributions from a component analysis.** *Proteins* 1994, **20**:25–33.

190. McCammon JA, Straatsma TP: **Alchemical free-energy simulation.** *Annu Rev Phys Chem* 1992, **43**:407.

191. Kollman PA: **Free energy calculations: applications to chemical and biochemical phenomena.** *Chem Rev* 1992, **93**:2395–2417.

192. Shi Y-Y, Mark AE, Wang C, Huang F, Berendsen HJC, Gunsteren WF: **Can the stability of protein mutants be predicted by free energy calculations?** *Protein Eng* 1993, **6**:289–295.

193. Eisenmenger F, Argos P, Abagyan R: **A method to configure protein side-chains from the main-chain trace in homology modelling.** *J Mol Biol* 1993, **231**:849–860.

194. Gō N, Abe H: **The consistency principle in protein structure and pathways of folding.** *Adv Biophysics* 1984, **18**:149–164.

195. 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.

196. • 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.
The X-ray structure of a variant of λ repressor reveals that the protein accommodates the potentially disruptive residues with shifts in its α-helical arrangement and with only limited changes in side-chain orientations.

197. Laskowski RA, McArthur MW, Moss DS, Thornton JM: **PROCHECK: A program to check the stereochemical quality of protein structures.** *J Appl Crystallogr* 1993, **26**:283–291.

198. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM: **Stereochemical quality of protein structure coordinates.** *Proteins* 1992, **12**:345–364.

199. Novotny J, Bruccoleri R, Karplus M: **An analysis of incorrectly folded protein models: implications for structural predictions.** *J Mol Biol* 1984, **177**:787–818.

200. Novotny J, Rashin AA, Bruccoleri RE: **Criteria that discriminate between native proteins and incorrectly folded models.** *Proteins* 1988, **4**:19–30.

201. Gregoret LM, Cohen FE: **Effect of packing density on chain conformation.** *J Mol Biol* 1991, **219**:109–122.

202. Bryant SH, Amzel LM: **Correctly folded proteins make twice as many hydrophobic contacts.** *Int J Pept Protein Res* 1987, **29**:46–52.

203. Chiche L, Gregoret LM, Cohen FE, Kollman PA: **Protein model structure evaluation using the solvation free energy of folding.** *Proc Natl Acad Sci USA* 1990, **87**:3240–3244.

204. Holm L, Sander C: **Evaluation of protein models by atomic solvation preference.** *J Mol Biol* 1992, **225**:93–105.

205. Baumann G, Frömmel C, Sander C: **Polarity as a criterion in protein design.** *Protein Eng* 1989, **2**:329–334.

206. Vila J, Williams RL, Vasquez M, Scheraga HA: **Empirical solvation models can be used to differentiate from near-native conformations of bovine pancreatic trypsin inhibitor.** *Proteins* 1991, **10**:199–218.

207. • Koehl P, Delarue M: **Polar and nonpolar atomic environments in the protein core: implications for folding and binding.** *Proteins* 1994, **20**:264–278.
Environment-dependent free energy of atoms is defined using contacts with solvent, polar and non-polar atoms. It is derived from a database of known structures and can be used to discriminate misfolded from correct structural models.

208. Bryant SH, Lawrence CE: **The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: a statistical model for nonbonded interactions.** *Proteins* 1991, **9**:108–119.

209. Colovos C, Yeates TO: **Verification of protein structures: patterns of non-bonded atomic interactions.** *Protein Sci* 1993, **2**:1511–1519.

210. Lüthy R, Bowie JU, Eisenberg D: **Assessment of protein models with three-dimensional profiles.** *Nature* 1992, **356**:83–85.

211. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ: **Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force.** *J Mol Biol* 1990, **216**:167–180.

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

213. • Topham CM, Srinivasan N, Thorpe CJ, Overington JP, Kalsheker NA: **Comparative modelling of major house dust mite allergen *Der p* I: structure validation using an extended environmental amino acid propensity table.** *Protein Eng* 1994, **7**:869–894.
Comparative modelling by assembly of rigid bodies is used to derive a three-dimensional model of the major house dust mite allergen *Der p* I. The structure is evaluated by program HARMONY, which implements environment-dependent amino acid residue substitution tables.

214. • 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.
To assess the accuracy of refined crystallographic structures, a comparison is made of independently determined structures of the same protein in the same crystal form. The positional error is estimated to be close to 1 Å.

215. 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.

216. • Zhao D, Jardetzky O: **An assessment of the precision and accuracy of protein structures determined by NMR.** *J Mol Biol* 1994, **239**:601–607.
The accuracy and precision of NMR structures are explored. The conclusion is that the accuracy of NMR structures is, at best, of the order of 1–2 Å.

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

218. Šali A, Matsumoto R, McNeil HP, Karplus M, Stevens RL: **Three-dimensional models of four mouse mast cell chymases, identification of proteoglycan-binding regions and protease-specific antigenic epitopes.** *J Biol Chem* 1993, **268**:9023–9034.

219. • 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, in press.
Mouse mast cell tryptase 7 is modelled using MODELLER. The heparin-binding site is predicted from the electrostatic potential around the model. The location of the site is confirmed by site-directed mutagenesis experiments.

220. 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.

221. Caputo A, James MNG, Powers JC, Hudig D, Bleackley RC: **Conversion of the substrate specificity of mouse proteinase granzyme B.** *Nature Struct Biol* 1994, **1**:364–367.

222. 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.

A Šali, Box 270, The Rockefeller University, 1230 York Avenue, New York, New York 10021-6399, USA.
E-mail: sali@rockvax.rockefeller.edu