7

Comparative Protein Structure Modeling

András Fiser and Andrej Sali
The Rockefeller University, New York, New York, U.S.A.

1 INTRODUCTION

Functional characterization of a protein sequence is one of the most frequent problems in biology. This task is usually facilitated by an accurate three-dimensional (3D) structure of the studied protein. A three-dimensional structure of natural proteins is guided by two distinct sets of principles operating on vastly different time scales: the laws of physics and the theory of evolution. Each of the two sets of principles that apply to the natural protein sequences gave rise to a class of protein structure prediction methods (Fig. 1) (Baker and Sali, 2001; Fiser et al., 2002).

The first approach, de novo or ab initio methods, predict the structure from sequence alone, without relying on similarity at the fold level between the modeled sequence and any of the known structures (Bonneau, Baker, 2001). The de novo methods assume that the native structure corresponds to the global free-energy minimum accessible during the lifespan of the protein and attempt to find this minimum by an exploration of many conceivable protein conformations. The two key components of de novo methods are the procedure for efficiently carrying out the conformational search and the free-energy function used for evaluating possible conformations.

The second class of methods, including threading (Domingues et al., 2000) and comparative modeling (Blundell et al., 1987; Marti-Renom et al.,

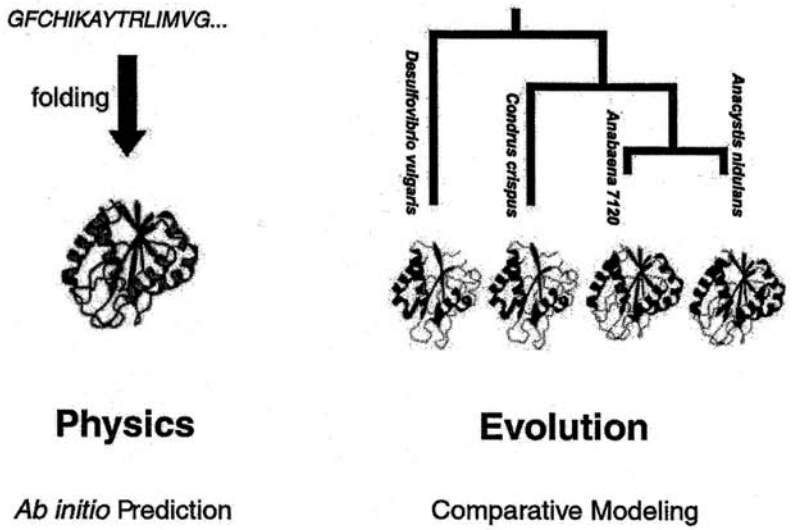

Figure 1 De novo structure prediction and comparative protein structure modeling. Proteins obey two distinct sets of principles, the laws of physics and the theory of evolution, each giving rise to the corresponding variety of protein structure prediction methods. (From Fiser et al., 2002.)

2000), rely on detectable similarity spanning most of the modeled sequence and at least one known structure. When the structure of one protein in the family has been determined by experiment, the other members of the family can be modeled based on their alignment to the known structure.

Comparative, or homology, protein structure modeling builds a three-dimensional model for a protein of unknown structure (the target) based on one or more related proteins of known structure (the templates) (Blundell et al., 1987; Greer, 1981; Johnson et al., 1994; Sali, Blundell, 1993; Sali, 1995; Sanchez, Sali, 1997a; Marti-Renom et al., 2000; Fiser et al., 2001; Fiser et al., 2002; Sanchez, Sali, 2000; Fiser, Sali, 2002). The necessary conditions for calculating a useful model are (1) detectable similarity between the target sequence and the template structures and (2) availability of a correct alignment between them. The comparative approach to protein structure prediction is possible because a small change in the protein sequence usually results in a small change in its 3D structure (Chothia, Lesk, 1986). It is also facilitated by the fact that 3D structure of proteins from the same family is more conserved than their primary sequences (Lesk, Chothia, 1980). Therefore, if similarity between two proteins is detectable at the sequence level, structural similarity can usually be assumed. Moreover, pro-

teins that share low or even nondetectable sequence similarity many times also have similar structures. Despite progress in ab initio protein structure prediction (Bonneau, Baker, 2001), comparative modeling remains the only method that can reliably predict the 3D structure of a protein with an accuracy comparable to a low-resolution experimentally determined structure (Marti-Renom et al., 2000).

All current comparative modeling methods consist of five sequential steps (Fig. 2). The first step is to search for proteins with known 3D structures that are related to the target sequence. The second step is to pick those structures that will be used as templates. The third step is to align their sequences with the target sequence. The fourth step is to build the model for the target sequence given its alignment with the template structures. The last step is to evaluate the model using a variety of criteria. If necessary, template selection, alignment, and model building can be repeated until a satisfactory model is obtained.

Currently, the probability of finding related proteins of known structure for a sequence picked randomly from an organism's genome ranges approximately from 30% to 65%, depending on which genome is examined (Kelley et al., 2000; Sanchez, Sali, 1998; Teichmann et al., 1999; Pieper et al., 2002). Approximately 57% of all known sequences have at least one domain that is detectably related to at least one protein of known structure (Pieper et al., 2002). Since the number of known protein sequences is approximately 1,200,000 (Benson et al., 2002; Bairoch, Apweiler, 2000), comparative modeling can be applied to domains in approximately 600,000 proteins. This number is an order of magnitude larger than the number of experimentally determined protein structures deposited in the Protein Data Bank (PDB) ($\sim$ 15,000) (Westbrook et al., 2002). Furthermore, the usefulness of comparative modeling is steadily increasing because the number of different structural folds that proteins adopt is limited (Chothia, 1992; Lo et al., 2000; Holm, Sander, 1997; Bray et al., 2000) and because the number of experimentally determined novel structures is increasing. This trend is accentuated by the recently initiated structural genomics project that aims to determine at least one structure for most protein families (Burley et al., 1999). It is conceivable that this aim will be substantially achieved in less than 10 years, making comparative modeling applicable to most protein sequences (Vitkup et al., 2001).

There are several computer programs and Web servers that automate the comparative modeling process. The first Web server for automated comparative modeling was the Swiss-Model server (http://www.expasy.ch/swissmod/), followed by CPHModels (http://www.cbs.dtu.dk/services/CPHmodels/), SDSC1 (http://cl.sdsc.edu/hm.html), FAMS (http://physchem.pharm.kitasato-u.ac.jp/FAMS/fams.html), and ModWeb

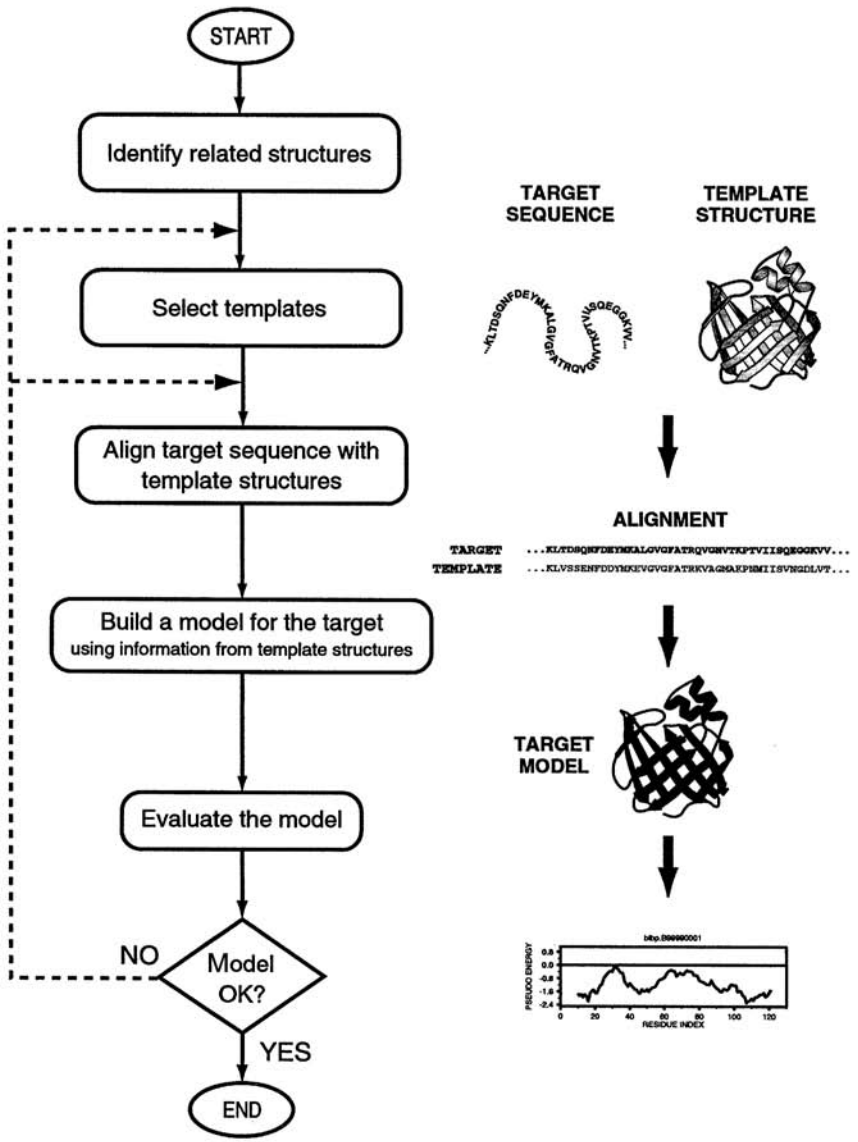

Figure 2 Steps in comparative protein structure modeling. See text for a description of each step. (From Fiser et al., 2001.)

(http://guitar.rockefeller.edu/modweb). These servers accept a sequence from a user and return an all-atom comparative model when possible. In addition to modeling a given sequence, ModWeb is capable of returning comparative models for all sequences in the TrEMBL database that are detectably related to an input, user-provided structure. While the Web servers are convenient and useful, the best results in the difficult or unusual modeling cases, such as problematic alignments, modeling of loops, existence of multiple conformational states, and modeling of ligand binding, are still obtained by nonautomated, expert use of the various modeling tools. A number of resources useful in comparative modeling are listed in Table 1.

This chapter begins with a description of all the steps in comparative modeling, fold assignment, template selection, sequence–structure alignment, model building, and model assessment. We conclude by describing errors in comparative models and sample applications of comparative modeling to individual proteins and to whole genomes. We emphasize our own work and experience, although we have profited greatly from the contributions of many others, cited in the list of references.

2 STEPS IN COMPARATIVE MODELING

2.1 Searching for Structures Related to the Target Sequence

Comparative modeling usually starts by searching the PDB (Westbrook et al., 2002) of known protein structures using the target sequence as the query. This search is generally done by comparing the target sequence with the sequence of each of the structures in the database.

There are three main classes of protein comparison methods that are useful in fold identification. The first class compares the target sequence with each of the database sequences independently, using pairwise sequence–sequence comparison (Apostolico, Giancarlo, 1998). The performance of these methods in sequence searching (Pearson, 2000; Pearson, 1995) and fold assignments has been evaluated exhaustively (Brenner et al., 1998). The most popular programs in the class include FASTA (Pearson, 2000) and BLAST (Altschul et al., 1997).

The second class of methods relies on multiple sequence comparisons to improve greatly the sensitivity of the search (Henikoff et al., 2000; Krogh et al., 1994; Gribskov, Veretnik, 1996; Altschul et al., 1997; Jaroszewski et al., 1998). The most well known program in this class is PSI-BLAST (Altschul et al., 1997). Another, similar approach that appears to perform even slightly better than PSI-BLAST has been implemented in the program PDB-BLAST (Jaroszewski et al., 1998). PDB-BLAST begins by finding all

Table 1 Programs and Servers Useful for Comparative Protein Structure Modeling

Databases	
NCBI	www.ncbi.nlm.nih.gov/
PDB	www.rcsb.org/
MSD	www.rcsb.org/databases.html
CATH	www.biochem.ucl.ac.uk/bsm/cath/
TrEMBL	srs.ebi.ac.uk/
Scop	scop.mrc-lmb.cam.ac.uk/scop/
Presage	presage.stanford.edu
ModBase	guitar.rockefeller.edu/modbase/
GeneCensus	bioinfo.mbb.yale.edu/genome
GeneBank	www.ncbi.nlm.nih.gov.Genbank/GenbankSearch.html
PSI	www.structuralgenomics.org
Template Search, fold assignment	
PDB-Blast	bioinformatics.burnham-inst.org/pdb_blast
BLAST	www.ncbi.nlm.nih.gov/BLAST/
FastA	www.ebi.ac.uk/fasta33
DALI	www2.ebi.ac.uk/dali/
PhD, TOPITS	www.embl-heidelberg.de/predictprotein/predictprotein/html
THREADER	bioinf.cs.ucl.ac.uk/threader/
123D	123d.ncifcrf.gov
UCLA-DOE	fold.doe-mbi.ucla.edu
PROFIT	lore.came.sbg.ac.at/
MATCHMAKER	www.tripos.com/software/mm.html
3D-PSSM	www.sbg.bio.ic.ac.uk/~3dppsm
BIOINBGU	www.cs.bgu.ac.il/~bioinbgu/
FUGUE	www.cryst.bioc.cam.ac.uk/~fugue
LOOPP	ser-loopp.tc.cornell.edu/loopp.html
FASS	bioinformatics.burnham-inst.org/FFAS/index.html
SAM-T99/T98	www.cse.ucsc.edu/research/compbio/sam.html
Comparative modeling	
3D-JIGSAW	www.bmm.icnet.uk/servers/3digsaw/
CPH-Models	www.cbs.dtu.dk/services/CPHmodels/
COMPOSER	www-cryst.bioc.cam.ac.uk/
FAMS	physchem.pharm.kitasato-u.ac.jp/FAMS/fams.html
Modeller	guitar.rockefeller.edu/modeller/modeller/html
PrISM	honiglab.cpmc.columbia.edu/
SWISS-MODEL	www.expasy.ch/swissmod/SWISS-MODEL.html
SDSC1	cl.sdsc.edu.hm.html
WHAT IF	www.combi.kun.nl/whatif/
ICM	www.molsoft.com/
SCWRL	www.fccc.edu/research/labs/dunbrack/scwrl/

Table 1 *Continued*

InsightII	www.accelrys.com
SYBYL	www.tripos.com
Model evaluation	
PROCHECK	www.biochem.ucl.ac.uk/~roman/procheck/procheck.html
WHATCHECK	www.cmbi.kun.nl/gv/servers/WIWWWI/
Prosall	www.came.sbg.ac.at
BIOTECH	biotech.embl-ebi.ac.uk:8400/
VERIFY3D	www.doe.mbi.ucla.edu/Services.Verify_3D/
ERRAT	www.doe-mbi.ucla.edu/Services/Errat.html
ANOLEA	guitar.rockefeller.edu/~fmelo/anolea/anolea.html
AQUA	urchin.bmrb.wisc.edu/~jurgen/Aqua/server/
SQUID	www.yorvic.york.ac.uk/~oldfield/squid
PROVE	www.ucmb.ulb.ac.be/UCMB/PROVE/

An up-to-date version of this table can be found on the Web at http://guitar.rockefeller.edu/bioinformatics_resources.shtml.

sequences in a sequence database that are clearly related to the target and easily aligned with it. The multiple alignment of these sequences is the target sequence profile. Similar profiles are also constructed for all potential template structures. The templates are then found by comparing the target sequence profile with each of the template sequence profiles, using a local dynamic programming method that relies on the common BLOSUM62 residue substitution matrix (Henikoff, Henikoff, 2000). These more sensitive fold identification techniques based on sequence profiles are especially useful for finding structural relationships when sequence identity between the target and the template drops below 25%.

The third class of methods relies on pairwise comparison of a protein sequence and a protein structure; that is, structural information is used for one of the two proteins that are being compared, and the target sequence is matched against a library of 3D profiles or threaded through a library of 3D folds. These methods are also called fold assignment, threading, or 3D template matching (Johnson, Overington, 1993; Bowie et al., 1991; D.T Jones et al., 1992; Godzik, 1996; Sippl, 1995). They are reviewed in D.T. Jones, 1997; Smith et al., 1997; and Torda, 1997, and evaluated in Domingues et al., 2000. These methods are especially useful when sequence profiles are not possible to construct because there are not enough known sequences that are clearly related to the target or potential templates.

What similarity between the target and template sequences is needed to have a chance of obtaining a useful comparative model? The answer

depends on the question that is asked of a model. In general, the usefulness of a template should be assessed by evaluation of the corresponding 3D model based on a given template, using methods described later. This approach is optimal because the evaluation of a 3D model is generally more sensitive and robust than the evaluation of an alignment (Sanchez, Sali, 1998). A good starting point for template searches are the many database search servers on the Web (Table 1).

2.2 Selecting Templates

Once a list of potential templates is obtained using searching methods, it is necessary to select one or more templates that are appropriate for the particular modeling problem. Several factors need to be taken into account when selecting a template.

The quality of a template increases with its overall sequence similarity to the target and decreases with the number and length of gaps in the alignment. The simplest template selection rule is to select the structure with the highest sequence similarity to the modeled sequence.

The family of proteins that includes the target and the templates can frequently be organized into subfamilies. The construction of a multiple alignment and a phylogenetic tree (Retief, 2000; Felsenstein, 1981) can help in selecting the template from the subfamily that is closest to the target sequence.

The similarity between the "environment" of the template and the environment in which the target needs to be modeled should also be considered. The term *environment* is used here in a broad sense, including everything that is not the protein itself (e.g., solvent, pH, ligands, quaternary interactions). If possible, a template bound to the same or similar ligands as the modeled sequence should generally be used.

The quality of the experimentally determined structure is another important factor in template selection. The resolution and R-factor of a crystallographic structure and the number of restraints per residue for an NMR structure are indicative of the accuracy of the structure. For instance, if two templates have comparable sequence similarity to the target, the one determined at the highest resolution should generally be used.

The criteria for selecting templates also depend on the purpose of a comparative model. For example, if a protein–ligand model is to be constructed, the choice of the template that contains a similar ligand is probably more important than the resolution of the template. On the other hand, if the model is to be used to analyze the geometry of the active site of an enzyme, it may be preferable to use a high-resolution template structure.

It is not necessary to select only one template. In fact, the use of several templates generally increases the model accuracy. One strength of the comparative modeling program MODELLER (Sali, Blundell, 1993) is that it can combine information from multiple template structures, in two ways. First, multiple template structures may be aligned with different domains of the target, with little overlap between them, in which case the modeling procedure can construct a homology-based model of the whole target sequence. Second, the template structures may be aligned with the same part of the target, in which case the modeling procedure is likely to automatically build the model on the locally best template (Sanchez, Sali, 1997b; Sali et al., 1995). In general, it is frequently beneficial to include in the modeling process all the templates that differ substantially from each other, if they share approximately the same overall similarity to the target sequence.

An elaborate way to select suitable templates is to generate and evaluate models for each candidate template structure and/or their combinations. The optimized all-atom models are evaluated by an energy or scoring function, such as the Z-score of PROSA (Sippl, 1993). The PROSA Z-score of a model is a measure of compatibility between its sequence and its structure. Ideally, the Z-score of the model should be comparable to the Z-score of the template. The PROSA Z-score is frequently sufficiently accurate to identify one of the most accurate of the generated models (Wu et al., 2000). This trial-and-error approach can be viewed as limited threading (i.e., the target sequence is threaded through similar template structures).

2.3 Aligning the Target Sequence with One or More Structures

To build a model, all comparative modeling programs depend on a list of assumed structural equivalences between the target and template residues. This list is defined by an alignment of the target and template sequences. Although many template search methods will produce such an alignment, it is usually not the optimal target–template alignment in the more difficult alignment cases (e.g., at less than 30% sequence identity). Search methods tend to be tuned for the detection of remote relationships, not for optimal alignment. Therefore, once the templates are selected, an alignment method should be used to align them with the target sequence.

The alignment is relatively simple to obtain when the target–template sequence identity is above 40%. In most such cases, an accurate alignment can be calculated automatically using standard sequence–sequence alignment methods. If the target–template sequence identity is lower than 40%, the alignment generally has gaps and needs manual intervention to

minimize the number of misaligned residues. In these low-sequence identity cases, the alignment accuracy is the most important factor affecting the quality of the resulting model. Alignments can be improved by including structural information from the template. For example, gaps should be avoided in secondary structure elements, in buried regions, or between two residues that are far apart in space. Some alignment methods take such criteria into account (Sanchez, Sali, 1998; Jennings et al., 2001; Blake, Cohen, 2001; Shi et al., 2001). It is important to inspect and edit the alignment in view of the template structure, especially if the target–template sequence identity is low. A misalignment by only one residue position will result in an error of approximately 4Å in the model because the current modeling methods generally cannot recover from errors in the alignment.

When multiple templates are selected, a good strategy is to superpose them with each other first, to obtain a multiple structure–based alignment. In the next step, the target sequence is aligned with this multiple structure–based alignment. Another improvement is to calculate the target and template sequence profiles, by aligning them with all sequences from a nonredundant sequence database that are sufficiently similar to the target and template sequences, respectively, so that they can be aligned without significant errors (e.g., better than 40% sequence identity). The final target–template alignment is then obtained by aligning the two profiles, not the template and target sequences alone. The use of multiple structures and multiple sequences benefits from the evolutionary and structural information about the templates as well as evolutionary information about the target sequence, and often produces a better alignment for modeling than the pairwise sequence alignment methods (Sauder et al., 2000; Jaroszewski et al., 2000).

2.4 Model Building

2.4.1 Modeling by Assembly of Rigid Bodies

The first and still widely used approach in comparative modeling is to assemble a model from a small number of rigid bodies obtained from the aligned protein structures (Greer, 1990; Blundell et al., 1987; Browne et al., 1969). The approach is based on the natural dissection of the protein structure into conserved core regions, variable loops that connect them, and side chains that decorate the backbone. For example, the following semiautomated procedure is implemented in the computer program COMPOSER (Sutcliffe et al., 1987). First, the template structures are selected and superposed. Second, the "framework" is calculated by averaging the coordinates of the C_α atoms of structurally conserved regions in the template structures. Third, the mainchain atoms of each core region in the target model are obtained by superposing on the framework the core segment from the tem-

plate whose sequence is closest to the target. Fourth, the loops are generated by scanning a database of all known protein structures to identify the structurally variable regions that fit the anchor core regions and have a compatible sequence (Topham et al., 1993). Fifth, the side chains are modeled based on their intrinsic conformational preferences and on the conformation of the equivalent side chains in the template structures (Sutcliffe et al., 1987). And finally, the stereochemistry of the model is improved by either a restrained energy minimization or a molecular dynamics refinement. 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 (Srinivasan, Blundell, 1993). For example, differences between the model and X-ray structures may be slightly smaller than the differences between the X-ray structures of the modeled protein and the homologs used to build the model. Possible future improvements of modeling by rigid-body assembly include incorporation of rigid-body shifts, such as the relative shifts in the packing of α-helices and β-sheets (Reddy, Blundell, 1993; Nagarajaram et al., 1999).

2.4.2 Modeling by Segment Matching or Coordinate Reconstruction

The basis of modeling by coordinate reconstruction is the finding that most hexapeptide segments of protein structure can be clustered into only 100 structurally different classes (Unger et al., 1989; Bystroff, Baker, 1998). Thus, comparative models can be constructed by using a subset of atomic positions from template structures as "guiding" positions and by identifying and assembling short, all-atom segments that fit these guiding positions. The guiding positions usually correspond to the Cα atoms of the segments that are conserved in the alignment between the template structure and the target sequence. The all-atom segments that fit the guiding positions can be obtained either by scanning all the known protein structures, including those that are not related to the sequence being modeled (Claessens et al., 1989; Holm, Sander, 1991), or by a conformational search restrained by an energy function (Bruccoleri, Karplus, 1990; van Gelder et al., 1994). For example, a general method for modeling by segment matching is guided by the positions of some atoms (usually Cα atoms) to find the matching segments in the representative database of all known protein structures (Levitt, 1992). This method can construct both main-chain and side chain atoms and can also model gaps. It is implemented in the program SegMod. Even some side-chain modeling methods (Chinea et al., 1995) and the class of loop construction methods based on finding suitable fragments in the database of known structures

(Jones, Thirup, 1986) can be seen as segment matching or coordinate reconstruction methods.

2.4.3 Modeling by Satisfaction of Spatial Restraints

The methods in this class begin by generating many constraints or restraints on the structure of the target sequence, using its alignment to related protein structures as a guide. The procedure is conceptually similar to that used in determination of protein structures from NMR-derived restraints. The restraints are generally obtained by assuming that the corresponding distances between aligned residues in the template and the target structures are similar. These homology-derived restraints are usually supplemented by stereochemical restraints on bond lengths, bond angles, dihedral angles, and nonbonded atom–atom contacts that are obtained from a molecular mechanics force field. The model is then derived by minimizing the violations of all the restraints. This can be achieved by either distance geometry or real-space optimization. For example, an elegant distance geometry approach constructs all-atom models from lower and upper bounds on distances and dihedral angles (Havel, Snow, 1991). Lower and upper bounds on Cα–Cα and main-chain–side-chain distances, hydrogen bonds, and conserved dihedral angles were derived for *E. coli* flavodoxin from four other flavodoxins; bounds were calculated for all distances and dihedral angles that had equivalent atoms in the template structures. The allowed range of values of a distance or a dihedral angle depended on the degree of structural variability at the corresponding position in the template structures. Distance geometry was used to obtain an ensemble of approximate 3D models, which were then exhaustively refined by restrained molecular dynamics with simulated annealing in water.

2.4.4 Modeling by Satisfaction of Spatial Restraints in the MODELLER Program

We now describe our own approach in more detail (Sali et al., 1990; Sali, Blundell, 1993; Fiser et al., 2000; Sali, Overington, 1994) (Fig. 3). The approach was developed to use as many different types of data about the target sequence as possible. It is implemented in the computer program MODELLER (Table 1). The comparative modeling procedure begins with an alignment of the target sequence with related known 3D structures. The output, obtained without any user intervention, is a 3D model for the target sequence containing all main-chain and side-chain non-hydrogen atoms.

In the first step of model building, distance and dihedral angle restraints on the target sequence are derived from its alignment with tem-

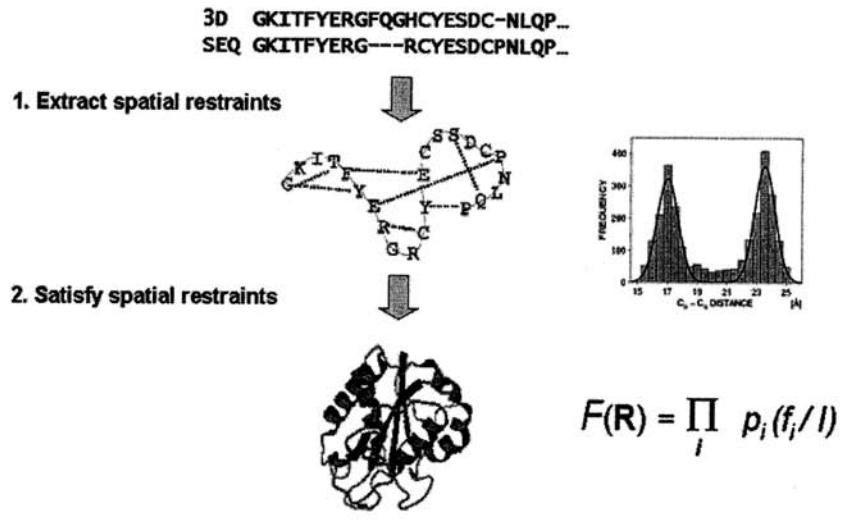

Figure 3 Comparative protein structure modeling by satisfaction of spatial restraints as implemented in MODELLER. First, spatial restraints are extracted from the input alignment, general spatial preferences found in known protein structures, and a molecular mechanics force field. Second, all the restraints are combined into an objective function that is optimized to obtain the final model.

plate 3D structures. The form of these restraints was obtained from a statistical analysis of the relationships between similar protein structures. The analysis relied on a database of 105 family alignments that included 416 proteins of known 3D structure (Sali, Overington, 1994). By scanning the database of alignments, tables quantifying various correlations were obtained, such as the correlations between two equivalent Cα–Cα distances or between equivalent main-chain dihedral angles from two related proteins (Sali, Blundell, 1993). These relationships are expressed as conditional probability density functions (pdf's) 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. Another example is the pdf for a certain Cα–Cα distance given equivalent distances in two related protein structures. An important feature of the method is that the forms of spatial restraints were obtained empirically, from a database of protein structure alignments.

In the second step, the spatial restraints and the CHARMM22 force-field terms enforcing proper stereochemistry (MacKerell Jr. et al., 1998; Brooks III et al., 1983) are combined into an objective function. The general form of the objective function is similar to that in molecular dynamics programs, such as CHARMM22 (Brooks III et al., 1983). The objective function depends on the Cartesian coordinates of the atoms (3D points) that form the modeled molecules. For a 10,000-atom system, there can be on the order of 200,000 restraints. The functional form of each term is simple; it includes a quadratic function, harmonic lower and upper bounds, cosine, a weighted sum of a few Gaussian functions, Coulomb's law, Lennard–Jones potential, and cubic splines. The geometric features presently include a distance, an angle, a dihedral angle, a pair of dihedral angles between two, three, four, and eight atoms, respectively, the shortest distance in the set of distances, solvent accessibility in Å^2, and atom density, expressed as the number of atoms around the central atom. Some restraints can be used to restrain pseudo-atoms, such as the gravity center of several atoms.

Finally, the model is obtained by optimizing the objective function in Cartesian space. The optimization is carried out by the use of the variable target function method (Braun, Go, 1985), employing methods of conjugate gradients and molecular dynamics with simulated annealing (Clore et al., 1986). Several slightly different models can be calculated by varying the initial structure, and the variability among these models can be used to estimate the lower bound on the errors in the corresponding regions of the fold.

Because the modeling by satisfaction of spatial restraints can use many different types of information about the target sequence, it is perhaps the most promising of all comparative modeling techniques. One of the strengths of modeling by satisfaction of spatial restraints is that constraints or restraints derived from a number of different sources can easily be added to the homology-derived restraints. For example, restraints could be provided by rules for secondary structure packing (Cohen, Kuntz, 1989), analyses of hydrophobicity (Aszodi, Taylor, 1994) and correlated mutations (Taylor, Hatrick, 1994), empirical potentials of mean force (Sippl, 1990), nuclear magnetic resonance (NMR) experiments (Sutcliffe et al., 1992), cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, site-directed mutagenesis (Boissel et al., 1993), intuition, *etc.* In this way, a comparative model, especially in the difficult cases, could be improved by making it consistent with available experimental data and/or with more general knowledge about protein structure.

Accuracies of the various model building methods are relatively similar when used optimally (Marti-Renom et al., 2002). Other factors, such as template selection and alignment accuracy, usually have a larger impact on

the model accuracy, especially for models based on less than 40% sequence identity to the templates. However, it is important that a modeling method allow a degree of flexibility and automation to obtain better models more easily and rapidly. For example, a method should allow for an easy recalculation of a model when a change is made in the alignment; it should be straightforward to calculate models based on several templates; and the method should provide tools for incorporation of prior knowledge about the target (e.g., cross-linking restraints, predicted secondary structure) and allow *ab initio* modeling of insertions (e.g., loops), which can be crucial for annotation of function. Loop modeling is an especially important aspect of comparative modeling in the range from 30% to 50% sequence identity. In this range of overall similarity, loops among the homologs vary, while the core regions are still relatively conserved and aligned accurately. Next, we review loop modeling.

2.4.5 Loop Modeling

In comparative modeling, target sequences often have inserted residues relative to the template structures or have regions that are structurally different from the corresponding regions in the templates. Thus, no structural information about these inserted segments can be extracted from the template structures. These regions frequently correspond to surface loops. Loops often play an important role in defining the functional specificity of a given protein framework, forming the active and binding sites. The accuracy of loop modeling is a major factor determining the usefulness of comparative models in applications such as ligand docking. Loop modeling can be seen as a mini-protein-folding problem because the correct conformation of a given segment of a polypeptide chain has to be calculated mainly from the sequence of the segment itself. However, loops are generally too short to provide sufficient information about their local fold. Even identical decapeptides in different proteins do not always have the same conformation (Kabsch, Sander, 1984; Mezei, 1998). Some additional restraints are provided by the core anchor regions that span the loop and by the structure of the rest of a protein that cradles the loop. Although many loop-modeling methods have been described, it is still not possible to model correctly and with high confidence loops longer than approximately eight residues (Fiser et al., 2000).

There are two main classes of loop-modeling methods: (1) the database search approaches, where a segment that fits on the anchor core regions is found in a database of all known protein structures (T.A. Jones, Thirup, 1986; Chothia, Lesk, 1987); (2) the conformational search approaches (Moult, James, 1986; Bruccoleri, Karplus, 1987; Shenkin et al., 1987).

There are also methods that combine these two approaches (van Vlijmen, Karplus, 1997; Deane, Blundell, 2001).

The database search approach to loop modeling is accurate and efficient when a database of specific loops is created to address the modeling of the same class of loops, such as β-hairpins (Sibanda et al., 1989), or loops on a specific fold, such as the hypervariable regions in the immunoglobulin fold (Chothia et al., 1989; Chothia, Lesk, 1987). For example, an analysis of the hypervariable immunoglobulin regions resulted in a series of rules that allowed a very high accuracy of loop prediction in other members of the family. These rules were based on the small number of conformations for each loop and on the dependence of the loop conformation on its length and certain key residues. There are attempts to classify loop conformations into more general categories, thus extending the applicability of the database search approach to more cases (Rufino et al., 1997; Oliva et al., 1997; Ring et al., 1992). However, the database methods are limited by the fact that the number of possible conformations increases exponentially with the length of a loop. As a result, only loops up to four to seven residues long have most of their conceivable conformations present in the database of known protein structures (Fidelis et al., 1994; Lessel, Schomburg, 1994). Even according to the more optimistic estimate, approximately 30% and 60% of all the possible eight- and nine-residue loop conformations, respectively, are missing from the database (Fidelis et al., 1994). This is made even worse by the requirement for an overlap of at least one residue between the database fragment and the anchor core regions, which means that the modeling of a five-residue insertion requires at least a seven-residue fragment from the database (Claessens et al., 1989). Despite the rapid growth of the database of known structures, there is no possibility to cover most of the conformations of a nine-residue segment in the foreseeable future. On the other hand, most of the insertions in a family of homologous proteins are shorter than 10–12 residues (Fiser et al., 2000).

To overcome the limitations of the database search methods, conformational search methods were developed (Moult, James, 1986; Bruccoleri, Karplus, 1987). There are many such methods, exploiting different protein representations, objective function terms, and optimization or enumeration algorithms. The search algorithms include the minimum perturbation method (Fine et al., 1986), molecular dynamics simulations (Bruccoleri, Karplus, 1990; van Vlijmen, Karplus, 1997), genetic algorithms (Ring, Cohen, 1993), Monte Carlo and simulated annealing (Abagyan, Totrov, 1994; Collura et al., 1993; Higo et al., 1992), multiple-copy simultaneous search (Zheng et al., 1993; Zheng et al., 1994), self-consistent field optimization (Koehl, Delarue, 1995), and an enumeration based on graph theory (Samudrala, Moult, 1998).

The loop-modeling module in MODELLER implements the optimization-based approach (Fiser et al., 2000; Fiser et al., 2002). The main reasons are the generality and conceptual simplicity of energy minimization, as well as the limitations on the database approach imposed by a relatively small number of known protein structures (Fidelis et al., 1994). Loop prediction by optimization is applicable to simultaneous modeling of several loops and loops interacting with ligands, which is not straightforward for the database search approaches. Loop optimization in MODELLER relies on conjugate gradients and molecular dynamics with simulated annealing. The pseudo-energy function is a sum of many terms, including some terms from the CHARMM-22 molecular mechanics force field (MacKerell Jr. et al., 1998) and spatial restraints based on distributions of distances (Sippl, 1990) and dihedral angles in known protein structures. The method was tested on a large number of loops of known structure, both in the native and near-native environments. In the case of five-residue loops in the correct environments, the average error was 0.6 Å, as measured by local superposition of the loop main-chain atoms alone. For eight-residue loops in the correct environments, 90% of the loops had less than 2-Å main-chain RMS error, with an average of less than 1.2 Å. Even 12-residue loops are modeled with useful accuracy in 30% of the cases. To simulate comparative modeling problems, the loop-modeling procedure was evaluated by predicting loops of known structure in only approximately correct environments. Such environments were obtained by distorting the anchor regions, corresponding to the three residues at either end of the loop, and all the atoms within 10 Å of the native loop conformation for up to 2–3 Å by molecular dynamics simulations. When the RMSD distortion of the environment atoms is 2.5 Å, the average loop prediction error increases by 180%, 25% and 3% for 4-, 8- and 12-residue loops, respectively. It is no longer too optimistic to expect useful models for loops as long as 12 residues, if the environment of the loop is at least approximately correct. It is possible to estimate whether or not a given loop prediction is correct, based on the structural variability of the independently derived lowest-energy loop conformations. Typically, the loop prediction corresponds to the lowest-energy conformation out of the 500 independent optimizations. The algorithm allows straightforward incorporation of additional spatial restraints, including those provided by template fragments, disulfide bonds, and ligand binding sites.

2.5 Evaluating a Model

After a model is built, it is important to check it for possible errors. The quality of a model can be predicted approximately from the sequence simi-

larity between the target and the template (Figs. 2, 5, 7). Sequence identity above 30% is a relatively good predictor of the expected accuracy of a model. However, other factors, including the environment, can strongly influence the accuracy of a model. For instance, some calcium-binding proteins undergo large conformational changes when bound to calcium. If a calcium-free template is used to model the calcium-bound state of a target, it is likely that the model will be incorrect irrespective of the target–template similarity. This estimate also applies to determination of protein structure by experiment; a structure must be determined in the functionally meaningful environment. If the target–template sequence identity falls below 30%, the sequence identity becomes significantly less reliable as a measure of expected accuracy of a single model. The reason is that below 30% sequence identity, models are often obtained that deviate significantly, in both directions, from the average accuracy. It is in such cases that model evaluation methods are most informative.

Two types of evaluation can be carried out. "Internal" evaluation of self-consistency checks whether or not a model satisfies the restraints used to calculate it. "External" evaluation relies on information that was not used in the calculation of the model (Luthy et al., 1992; Sippl, 1993).

Assessment of a model's stereochemistry (e.g., bonds, bond angles, dihedral angles, and nonbonded atom–atom distances) with programs such as PROCHECK (Laskowski et al., 1993) and WHATCHECK (Hooft et al., 1996) is an example of internal evaluation. Although errors in stereochemistry are rare and less informative than errors detected by methods for external evaluation, a cluster of stereochemical errors may indicate that the corresponding region also contains other larger errors (e.g., alignment errors).

When the model is based on less than $\approx$30% sequence identity to the template, the first purpose of the external evaluation is to test whether or not a correct template was used. This test is especially important when the alignment is only marginally significant or several alternative templates with different folds are to be evaluated. A complication is that at low similarities the alignment generally contains many errors, making it difficult to distinguish between an incorrect template on one hand and an incorrect alignment with a correct template on the other hand. It is generally possible to recognize a correct template only if the alignment is at least approximately correct. This complication can sometimes be overcome by testing models from several alternative alignments for each template. One way to predict whether or not a template is correct is to compare the PROSA Z-score (Sippl, 1993) for the model and the template structure(s). Since the Z-score of a model is a measure of compatibility between its sequence and structure, the model Z-score should be comparable to that of the template.

However, this evaluation does not always work. For example, a well-modeled part of a domain is likely to have a bad *Z*-score because some interactions that stabilize the fold are not present in the model. Correct models for some membrane proteins and small disulfide-rich proteins also tend to be evaluated incorrectly, apparently because these structures have distributions of residue accessibility and residue–residue distances that are different from those for the larger globular domains, which were the source of the PROSA statistical potential function.

The second, more detailed kind of external evaluation is the prediction of unreliable regions in the model. One way to approach this problem is to calculate a "pseudo-energy" profile of a model, such as that produced by PROSA (Sippl, 1995). The profile reports the energy for each position in the model. Peaks in the profile frequently correspond to errors in the model. There are several pitfalls in the use of energy profiles for local error detection. For example, a region can be identified as unreliable only because it interacts with an incorrectly modeled region; there are also more fundamental problems (Fiser et al., 2000).

Finally, a model should be consistent with experimental observations, such as site-directed mutagenesis, cross-linking data, and ligand binding.

It is frequently difficult to select best templates or to calculate a good alignment. One way of improving a comparative model in such cases is to proceed with an iteration consisting of template selection, alignment, and model building, guided by model assessment. This iteration can be repeated until no improvement in the model is detected (Guenther et al., 1997; Sanchez, Sali, 1997b).

3 ERRORS IN COMPARATIVE MODELS

The overall accuracy of comparative models spans a wide range (Figs. 5, 7). At the low end of the spectrum are the low-resolution models, whose only essentially correct feature is their fold. At the high end of the spectrum are the models with an accuracy comparable to medium resolution crystallographic structures (Baker, Sali, 2001; Marti-Renom et al., 2000). Even low-resolution models are often useful for addressing biological questions, because function can many times be predicted from only coarse structural features of a model.

The errors in comparative models can be divided into five categories (Fig. 4): (1) errors in sidechain packing; (2) distortions or shifts of a region that is aligned correctly with the template structures; (3) distortions or shifts of a region that does not have an equivalent segment in any of the template structures; (4) distortions or shifts of a region that is aligned incorrectly with the template structures; (5) a misfolded structure resulting from using an

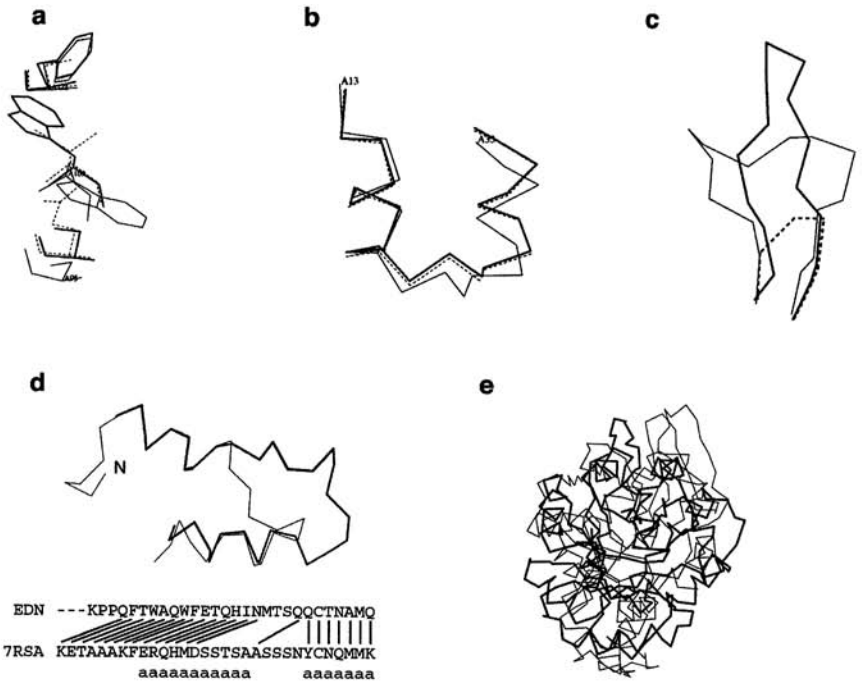

Figure 4 Errors in comparative protein structure modeling. (a) Errors in side-chain packing. The Trp 109 residue in the crystal structure of mouse cellular retinoic acid–binding protein I (thin line) is compared with its model (thick line) and with the template mouse adipocyte lipid–binding protein (broken line). (b) Distortions and shifts in correctly aligned regions. A region in the crystal structure of mouse cellular retinoic acid–binding protein I is compared with its model and with the template fatty acid–binding protein using the same representation as in panel a. (c) Errors in regions without a template. The Cα trace of the 112–117 loop is shown for the X-ray structure of human eosinophil neurotoxin (thin line), its model (thick line), and the template ribonuclease A structure (residues 111–117; broken line). (d) Errors due to misalignments. The N-terminal region in the crystal structure of human eosinophil neurotoxin (thin line) is compared with its model (thick line). The corresponding region of the alignment with the template ribonuclease A is shown. The black lines show correct equivalences, that is, residues whose Cα atoms are within 5 Å of each other in the optimal least squares superposition of the two X-ray structures. The "a" characters in the bottom line indicate helical residues. (e) Errors due to an incorrect template. The X-ray structure of α-trichosanthin (thin line) is compared with its model (thick line), which was calculated using indole-3-glycerophosphate synthase as the template. (From Fiser et al., 2001.)

incorrect template. Significant methodological improvements are needed to address all of these errors.

Errors 3–5 are relatively infrequent when sequences with more than 40% identity to the templates are modeled. For example, in such a case, approximately 90% of the main-chain atoms are likely to be modeled with an rms error of about 1 Å. In this range of sequence similarity, the alignment is mostly straightforward to construct, there are not many gaps, and structural differences between the proteins are usually limited to loops and side chains. When sequence identity is between 30% and 40%, the structural differences become larger, and the gaps in the alignment are more frequent and longer. As a result, the mainchain RMS error rises to about 1.5 Å for about 80% of residues. The rest of the residues are modeled with large errors because the methods generally fail to model structural distortions and rigid-body shifts and are unable to recover from misalignments. Below 40% sequence identity, misalignments and insertions in the target sequence become the major problems. When sequence identity drops below 30%, the main problem becomes the identification of related templates and their alignment with the sequence to be modeled (Figs. 5, 7). In general, it can be expected that about 20% of residues will be misaligned, and consequently incorrectly modeled with an error larger than 3 Å, at this level of sequence similarity (Johnson, Overington, 1993). These misalignments are a serious impediment to comparative modeling because it appears that presently at least one-half of all related protein pairs are related at less than 30% sequence identity (Rost, 1999; Sanchez, Sali, 1998).

It has been pointed out that a comparative model is frequently more distant from the actual target structure than the closest template structure used to calculate the model (Martin et al., 1997). However, at least for some modeling methods, this is only the case when there are errors in the template–target alignment used for modeling, and when the correct structure-based template–target alignment is used for comparing the template with the actual target structure (Sanchez, Sali, 1997b). In contrast, the model is generally closer to the target structure than any of the templates if the modeling target–template alignment is used in evaluating the similarity between the actual target structure and the template (Sanchez, Sali, 1997b). When more than one template is used for modeling, it is sometimes possible to obtain a model that is significantly closer to the target structure than any of the templates (Sanchez, Sali, 1997b). This improvement occurs because the model tends to inherit the best regions from each template. Therefore, using a model is generally better than using the template structure, even when the alignment is incorrect, because the actual target structure and, therefore, the correct template–target alignment are not available in practical modeling applications (Fig. 5).

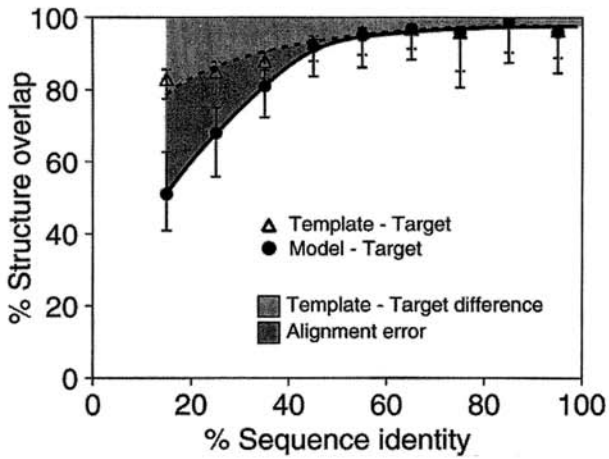

Figure 5 Average model accuracy as a function of sequence identity. As the sequence identity between the target sequence and the template structure decreases, the average structural similarity between the template and the target also decreases (dotted line, triangles). Structural overlap is defined as the fraction of equivalent Cα atoms. For the comparison of the model with the actual structure (filled circles), two C_α atoms were considered equivalent if they belonged to the same residue and were within 3.5 Å of each other after least squares superposition of all Cα atoms by the ALIGN3D command in MODELLER. For comparison of the template structure with the actual target structure (triangles), two Cα atoms were considered equivalent if they were within 3.5 Å of each other after alignment and rigid-body superposition. At high sequence identities, the models are close to the templates and therefore also close to the experimental target structure (solid line, filled circles). At low sequence identities, errors in the target–template alignment become more frequent and the structural similarity of the model with the experimental target structure falls below the target–template structural similarity. The difference between the model and the actual target structure is a combination of the target–template differences (light area) and the alignment errors (dark area). The figure was constructed by calculating 3993 comparative models based on single templates of varying similarity to the targets. All targets had known (experimentally determined) structures, and therefore the comparison of the models and templates with the experimental structures was possible. (From Sanchez, Sali, 1998.)

To put the errors in comparative models into perspective, we list the differences among structures of the same protein that have been determined experimentally (Fig. 6). The 1-Å accuracy of main-chain atom positions corresponds to X-ray structures defined at a low resolution of about 2.5 Å and with an R-factor of about 25% (Ohlendorf, 1994), as well as to medium-resolution NMR structures determined from

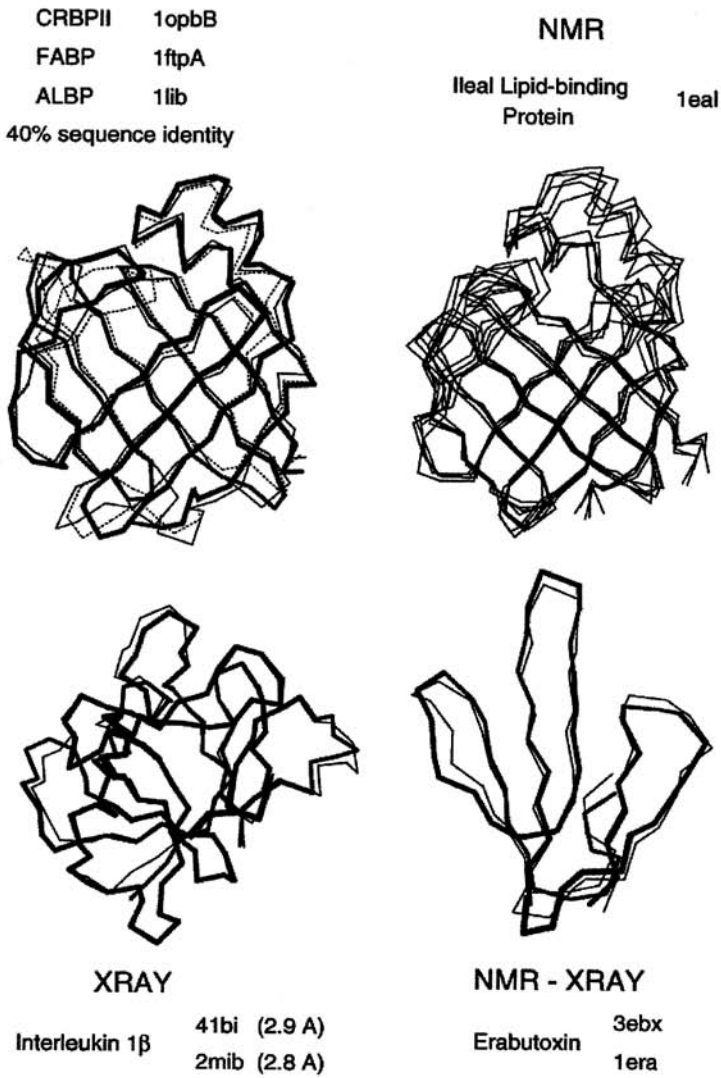

Figure 6 Accuracy of comparative models as compared to low-resolution crystallographic structure determination and medium-resolution NMR structure determination. *Upper left panel*: Comparison of homologous structures that share ~ 40% sequence identity. *Upper right panel*: 20 conformations of ileal lipid–binding protein that satisfy the NMR restraints equally well. *Lower left panel*: Comparison of two independently determined X-ray structures of interleukin 1β. *Lower right panel*: Comparison of the X-ray and NMR structures of erabutoxin. (From Fiser et al., 2001.)

10 interproton distance restraints per residue (Clore et al., 1993). Similarly, differences between the highly refined X-ray and NMR structures of the same protein also tend to be about 1 Å (Clore et al., 1993). Changes in the environment (e.g., oligomeric state, crystal packing, solvent, ligands) can also have a significant effect on the structure (Faber, Matthews, 1990). Overall, comparative models based on templates with more than 40% identity are almost as good as medium-resolution experimental structures, simply because the proteins at this level of similarity are likely to be as similar to each other as are the structures for the same protein determined by different experimental techniques under different conditions. However, the caveat in comparative protein modeling is that some regions, mainly loops and side chains, may have larger errors.

A way to test protein structure modeling methods, including comparative modeling, is provided by the biannual meetings on critical assessment of techniques for protein structure prediction (CASP) (Moult et al., 1995; Zemla et al., 2001; Marti-Renom et al., 2002). Protein modelers are challenged to model sequences with unknown 3D structure and to submit their models to the organizers before the meeting. At the same time, the 3D structures of the prediction targets are being determined by X-ray crystallography or NMR methods. They become available only after the models are calculated and submitted. Thus, a bona fide evaluation of protein structure modeling methods is possible. Large-scale, continuous, and automated complements to this experiment are implemented in two web servers, Live Bench (Bujnicki et al., 2001) and EVA (Eyrich et al., 2001).

4 APPLICATIONS OF COMPARATIVE MODELING

4.1 Modeling of Individual Proteins

Comparative modeling is often an efficient way to obtain useful information about the proteins of interest. For example, comparative models can be helpful in designing mutants to test hypotheses about the protein's function (Vernal et al., 2002; Wu et al., 1999a), identifying active and binding sites (Sheng et al., 1996), searching for, designing, and improving ligands for a given binding site (Ring et al., 1993), modeling substrate specificity (Xu et al., 1996), predicting antigenic epitopes (Sali et al., 1993), simulating protein–protein docking (Vakser, 1995), inferring function from calculated electrostatic potential around the protein (Matsumoto et al., 1995), facilitating molecular replacement in X-ray structure determination (Howell et al., 1992), refining models based on NMR constraints (Modi et al., 1996; Barrientos et al., 2001), testing and improving a sequence–structure align-

ment (Wolf et al., 1998), confirming a remote structural relationship (Guenther et al., 1997; Wu et al., 2000), and rationalizing known experimental observations (Fig. 7). For a lengthy review of comparative modeling applications see Johnson et al., 1994.

Fortunately, a 3D model does not have to be absolutely perfect to be helpful in biology, as demonstrated by the applications just listed. The type of a question that can be addressed with a particular model does depend on its accuracy.

At the low end of the accuracy spectrum, there are models that are based on less than 25% sequence identity and have sometimes less than 50% of their Cα atoms within 3.5 Å of their correct positions. However, such models still have the correct fold, and even knowing only the fold of a protein is frequently sufficient to predict its approximate biochemical function. More specifically, only nine out of 80 fold families known in 1994 contained proteins (domains) that were not in the same functional class, although 32% of all protein structures belonged to one of the nine superfolds (Orengo et al., 1997). Models in this low range of accuracy combined with model evaluation can be used for confirming or rejecting a match between remotely related proteins (Sanchez, Sali, 1997b; Sanchez, Sali, 1998).

In the middle of the accuracy spectrum are the models based on approximately 35% sequence identity, corresponding to 85% of the C atoms modeled within 3.5 Å of their correct positions. Fortunately, the active and binding sites are frequently more conserved than the rest of the fold and are thus modeled more accurately (Sanchez, Sali, 1998). In general, medium-resolution models frequently allow a refinement of the functional prediction based on sequence alone because ligand binding is most directly determined by the structure of the binding site rather than its sequence. It is frequently possible to correctly predict important features of the target protein that do not occur in the template structure. For example, the location of a binding site can be predicted from clusters of charged residues (Matsumoto et al., 1995), and the size of a ligand may be predicted from the volume of the binding site cleft (Xu et al., 1996). Medium-resolution models can also be used to construct site-directed mutants with altered or destroyed binding capacity, which in turn could test hypotheses about the sequence–structure–function relationships. Other problems that can be addressed with medium-resolution comparative models include designing proteins that have compact structures without long tails, loops, and exposed hydrophobic residues for better crystallization and designing proteins with added disulfide bonds for extra stability.

The high end of the accuracy spectrum corresponds to models based on 50% sequence identity or more. The average accuracy of these models approaches that of low-resolution X-ray structures (3Å resolution) or med-

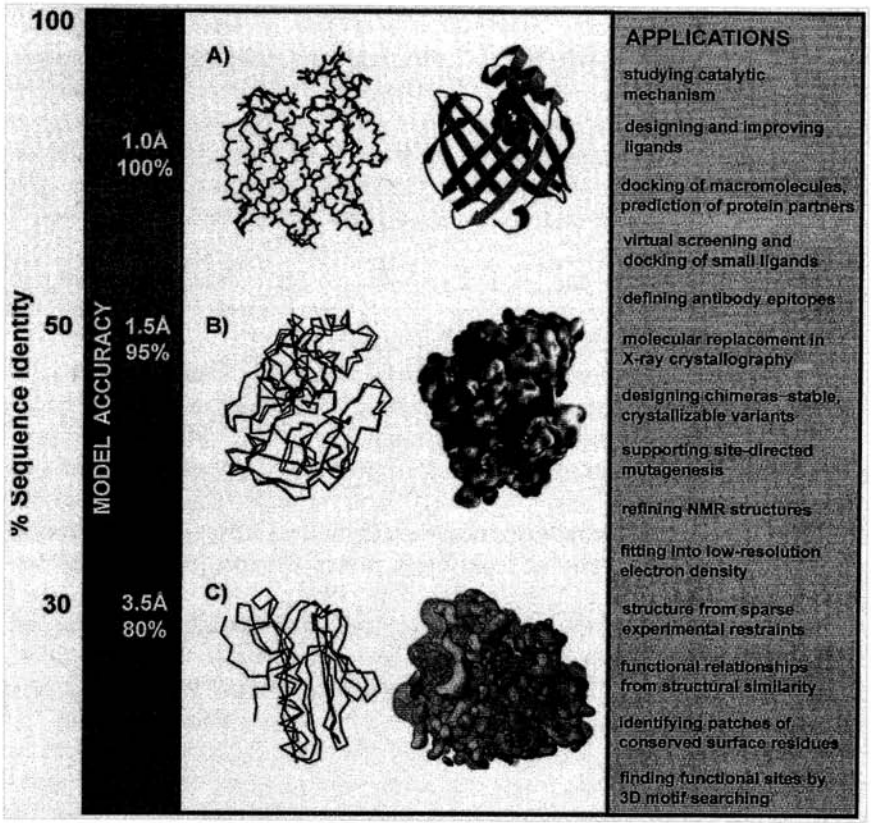

Figure 7 Accuracy of comparative models and their applications. The vertical axis indicates the different ranges of applicability of comparative protein structure modeling, the corresponding accuracy of protein structure models, and their sample applications. In panels A–C, typical overall accuracy of a comparative model (right) is indicated by a comparison of a model with an actual structure (left). (A) The complex between docosahexaenoic fatty acid (violet) and brain lipid–binding protein (right), modeled based on its 62% sequence identity to the crystallographic structure of adipocyte lipid–binding protein (PDB code 1ADL). (From Xu et al., 1996.) A number of fatty acids were ranked for their affinity to brain lipid–binding protein consistently with site-directed mutagenesis and affinity chromatography experiments, even though the ligand specificity profile of this protein is different from that of the template structure. (B) A putative proteoglycan–binding patch was identified on a medium-accuracy comparative model of mouse mast cell protease 7 (right), modeled based on its 39% sequence identity to the crystallographic structure of bovine pancreatic trypsin (2PTN) that does not bind proteoglycans. (From Matsumoto et al., 1995.) The prediction was confirmed by site-directed mutagenesis and heparin-affinity chromatography experiments. (C) A molecular model of the

ium-resolution NMR structures (10 distance restraints per residue) (Sanchez, Sali, 1997b). The alignments on which these models are based generally contain almost no errors. In addition to the already listed applications, high-quality models can be used for docking of small ligands (Ring et al., 1993) or whole proteins onto a given protein (Totrov, Abagyan, 1994; Vakser, 1995).

We now describe two applications of comparative modeling in more detail: (1) modeling of substrate specificity aided by a high-accuracy model and (2) substantiating a remote relationship between two proteins aided by a low-accuracy model.

Example 1: Modeling of Substrate Specificity. Brain lipid–binding protein (BLBP) is a member of the family of fatty acid–binding proteins that was isolated from brain (Xu et al., 1996). The problem was to find out which one of the many fatty acids known to bind to fatty acid–binding proteins in general is the likely physiological ligand of BLBP. To address this problem, comparative models of BLBP complexed with many fatty acids were calculated by relying on the structures of the adipocyte lipid–binding protein and muscle fatty acid–binding protein, in complex with their ligands. The models were evaluated by binding and site-directed mutagenesis experiments (Xu et al., 1996). The model of BLBP indicated that its binding cavity was just large enough to accommodate docosahexaenoic acid (DHA) (Fig. 8). Because DHA filled the BLBP–binding cavity completely, it was unlikely that BLBP would bind a larger ligand. Thus, DHA was the ligand predicted to have the highest affinity for BLBP. The prediction was confirmed by the measurement of binding affinities for many fatty acids. It turned out that the BLBP-DHA interaction was the strongest fatty acid–protein interaction known to date. The binding affinities of the ligands correlated with the surface areas buried by the protein–ligand interactions, as calculated from the corresponding models, and explained why DHA had the highest affinity. This case illustrates how a comparative model provides new information that cannot be deduced directly from the template structures despite their high, 60% sequence identity to BLBP. The two templates have smaller binding sites and consequently different patterns of binding affi-

---

whole yeast ribosome (right) was calculated by fitting atomic rRNA and protein models into the electron density of the 80S ribosomal particle, obtained by electron microscopy at 15-Å resolution. (From Spahn et al., 2001.) Most of the models for 40 out of the 75 ribosomal proteins were based on approximately 30% sequence identity to their template structures.

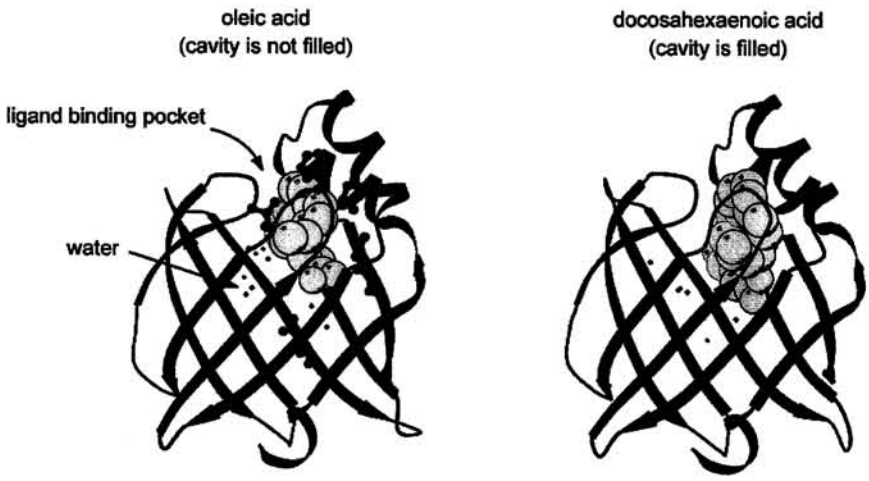

Figure 8 Modeling of the substrate specificity of the brain lipid–binding protein. The fatty acid ligand is shown in the CPK representation. The small spheres in the ligand–binding cavity are water molecules. *Left panel*, the model of the BLBP–oleic acid complex. *Right panel*, the model of the BLBP–docosahexaenoic acid complex. (From Xu et al., 1996.)

nities for the same set of ligands. The study also illustrated how new information is obtained relative to the target–template alignment even when the similarity between the target and the template sequences is high. The volumes and contact surfaces can be calculated only from a 3D model.

Example 2: Detection of Remote Relationships. Genes coding for the core histones H2a, H2b, H3, and H4 of *Giardia lamblia* were sequenced (Wu et al., 1999b). The derived amino acid sequences of all four histones were similar to their homologs in other eukaryotes, although they were among the most divergent members of this protein family. Comparative protein structure modeling combined with energy evaluation (Sippl, 1993) of the resulting models indicated that the *G. lamblia* core histones individually and together can assume the same three-dimensional structures that were established by X-ray crystallography for *Xenopus laevis* histones and the nucleosome core particle (Wu et al., 2000) (Fig. 9). Since *G. lamblia* represents one of the earliest eukaryotes in many different molecular trees, the structure of its histones is potentially of relevance to understanding histone evolution. Our studies concluded that the *G. lamblia* histones do not represent an intermediate stage between archaeal and eukaryotic histones.

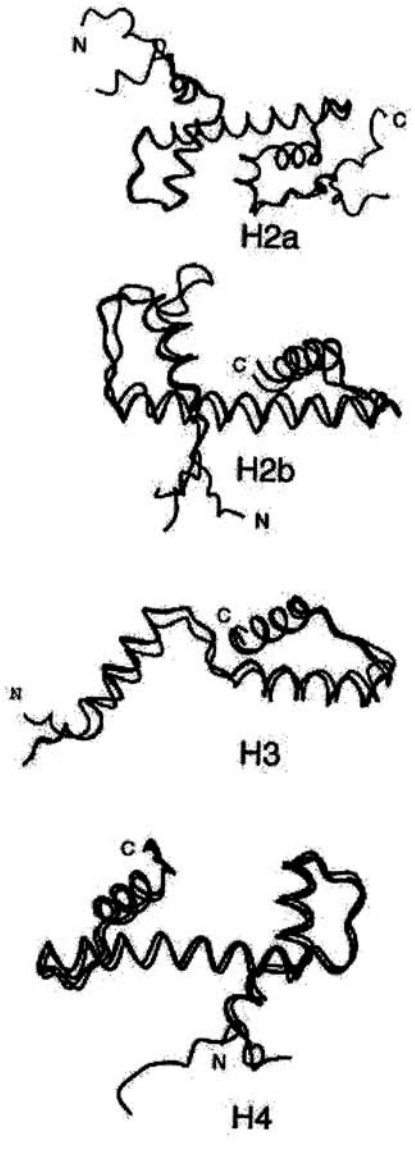

Figure 9 Substantiating the fold similarity of remotely related proteins by comparative modeling and assessment of the model energy. Comparative protein structure models of the *Giardia lamblia* core histones, based on the known structures of the *Xenopus histones*. The models and their evaluations indicate that the sequences of the *G. lamblia* histones are consistent with the structure of the corresponding *Xenopus* histones, with the exception of their terminal extension. (From Wu et al., 2000.)

4.2 Automated, Large-Scale Comparative Modeling

In a few years, the genome projects will have provided us with the amino acid sequences of millions of proteins—the catalysts, inhibitors, messengers, receptors, transporters, and building blocks of the living organisms. The full potential of the genome projects will be realized only once we assign and understand the function of these new proteins. This understanding will be facilitated by structural information for all or almost all proteins. Much of the structural information will be provided by structural genomics (Sali, 1998; Burley et al., 1999; Vitkup et al., 2001), a large-scale determination of protein structures by X-ray crystallography and nuclear magnetic resonance spectroscopy, combined efficiently with accurate, automated, and large-scale comparative protein structure modeling techniques (Sanchez et al., 2000a). Given limitations of the current modeling techniques, it seems reasonable to require models based on at least 30% sequence identity, corresponding to one experimentally determined structure per sequence family rather than fold family. It was estimated that the structures of representatives of approximately 16,000 sequence domain families need to be determined to provide comparative models based on at least 30% sequence identity for 90% of the protein sequences (Vitkup et al., 2001).

To enable large-scale comparative modeling needed for structural genomics, the steps of comparative modeling are being assembled into a completely automated pipeline (Sanchez, Sali, 1998). Since many computer programs for performing each of the operations in comparative modeling already exist, it may seem trivial to construct a pipeline that completely automates the whole process. In fact, it is not easy to do so in a robust manner. For a good reasons, most of the tasks in modeling of individual proteins, including template selection, alignment, and model evaluation, are typically performed with significant human intervention. This semiautomated modeling allows the use of the best tool for a particular problem at hand and consideration of many different sources of information that are difficult to take into account entirely automatically. Because large-scale modeling can be performed only in a completely automated manner, the main challenge is to build an automated and robust pipeline that approaches the performance of a human expert as much as possible.

Domains in approximately 57% of the 1,200,000 known protein sequences were modeled with MODELLER and deposited into a comprehensive database of comparative models, ModBase (http://guitar.rockefeller.edu/modbase/) (Sanchez et al., 2000b; Pieper et al., 2002; Sanchez, Sali, 1998). While the current number of modeled proteins may look impressive, usually only one domain per protein is modeled (on the average, proteins have slightly more than two domains), and two-thirds of the models are

based on less than 30% sequence identity to the closest template. The Web interface to ModBase allows flexible querying for fold assignments, sequence–structure alignments, models, and model assessments of interest. An integrated sequence/structure viewer, ModView, allows inspection and analysis of the query results (Ilyin, Sali, 2002). ModBase will be increasingly interlinked with other applications and databases such that structures and other types of information can easily be used for functional annotation.

Large-scale comparative modeling opens new opportunities for tackling existing problems by virtue of providing many protein models from many genomes. One example is the selection of a target protein for which a drug needs to be developed. A good choice is a protein that is likely to have high ligand specificity; specificity is important because specific drugs are less likely to be toxic. Large-scale modeling facilitates imposing the specificity filter in target selection by enabling a structural comparison of the ligand binding sites of many proteins, either human or from other organisms. Such comparisons may make it possible to select rationally the target whose binding site is structurally most different from the binding sites of all the other proteins that may potentially react with the same drug. For example, when a human pathogenic organism needs to be inhibited, it may be possible to select as the target that pathogen's protein that is structurally most different from all the human homologs. Alternatively, when a human metabolic pathway needs to be regulated, the target identification could focus on that particular protein in the pathway that has the binding site most dissimilar from its human homologs.

ACKNOWLEDGEMENTS

We are grateful to the members of our group for many discussions about protein structure prediction. Research was supported by NIH/NIGMS R01 GM54762, NIH/NIGMS P50 GM62529, Merck Genome Research Award, and the Mathers Fund Award. András Fiser was a Burroughs Wellcome Fund Postdoctoral Fellow and is a Charles Revson Foundation Postdoctoral Fellow. Andrej Sali is an Irma T. Hirschl Trust Career Scientist. This perspective is based partly on previous papers (Fiser et al., 2001; Fiser, Sali, 2002; Fiser et al., 2002; Baker, Sali, 2001; Marti-Renom et al., 2000).

REFERENCES

Abagyan R, Totrov M. 1994. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 235:983–1002.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Apostolico A, Giancarlo R. 1998. Sequence alignment in molecular biology. J Comput Biol 5:173–196.

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

Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45–48.

Baker D, Sali A. 2001. Protein structure prediction and structural genomics. Science, in press.

Barrientos LG, Campos-Olivas R, Louis JM, Fiser A, Sali A, Gronenborn AM. 2001. 1H, 13C, 15N resonance assignments and fold verification of a circular permuted variant of the potent HIV-inactivating protein cyanovirin-N. J Biomol NMR 19:289–290.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. 2002. GenBank. Nucleic Acids Res 30:17–20.

Blake JD, Cohen FE. 2001. Pairwise sequence alignment below the twilight zone. J Mol Biol 307:721–735.

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

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

Bonneau R, Baker D. 2001. Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct 30:173–189.

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

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

Bray JE, Todd AE, Pearl FM, Thornton JM, Orengo CA. 2000. The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. Protein Eng 13:153–165.

Brenner SE, Chothia C, Hubbard TJ. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA 95:6073–6078.

Brooks CL III, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. 1983. CHARMM: A program for macromolecular energy minimization and dynamics calculations. J Comp Chem 4:187–217.

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

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

Bruccoleri RE, Karplus M. 1990. Conformational sampling using high-temperature molecular dynamics. Biopolymers 29:1847–1862.

Bujnicki JM, Elofsson, Fischer D, Rychlewski L. 2001. Livebench-1: Continuous benchmarking of protein structure prediction servers. Protein Sci 10:352–361.

Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S. 1999. Structural genomics: beyond the human genome project. Nat Genet 23:151–157.

Bystroff C, Baker D. 1998. Prediction of local structure in proteins using a library of sequence–structure motifs. J Mol Biol 281:565–577.

Chinea G, Padron G, Hooft RW, Sander C, Vriend G. 1995. The use of position-specific rotamers in model building by homology. Proteins 23:415–421.

Chothia C. 1992. One thousand families for the molecular biologist. Nature 357:543–544.

Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826.

Chothia C, Lesk AM. 1987. Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol 196:901–917.

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

Claessens M, Van Cutsem E, Lasters I, Wodak S. 1989. Modelling the polypeptide backbone with "spare parts" from known protein structures. Protein Eng 2:335–345.

Clore GM, Brunger AT, Karplus M, Gronenborn AM. 1986. Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination. A model study of crambin. J Mol Biol 191:523–551.

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

Cohen FE, Kuntz ID. 1989. Tertiary structure prediction. In: Fasman GD, ed. Prediction of Protein Structure and the Principles of Protein Conformations. New York: Plenum Press, pp 647–705.

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

Deane CM, Blundell TL. 2001. CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 10:599–612.

Domingues FS, Lackner P, Andreeva A, Sippl MJ. 2000. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. J Mol Biol 297:1003–1013.

Eyrich V, Marti-Renom MA, Przybylski D, Fiser A, Pazos F, Valencia A, Sali A, Rost B. 2000. EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics, 17:1242–1243.

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

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376.

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

Fine RM, Wang H, Shenkin PS, Yarmush DL, Levinthal C. 1986. Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. Proteins 1:342–362.

Fiser A., Sali, A. Modeller: generation and refinement of homology models. Methods Enzymol, in press 2002.

Fiser A, Do RK, Sali A. 2000. Modeling of loops in protein structures. Protein Sci 9:1753–1773.

Fiser A, Sanchez R, Melo F, Sali A. 2001. Comparative protein structure modeling. In: Watanabe M, Roux B, MacKerell AD, Jr, Becker O, eds. Computational Biochemistry and Biophysics. New York: Marcel Dekker. pp 275–312.

Fiser A, Feig M, Brooks CL, III, Sali A. 2002. Evolution and Physics in Comparative Protein Structure Modeling. Acc Chem Res 35:413–421.

Godzik A. 1996. Knowledge-based potentials for protein folding: what can we learn from known protein structures? Structure 4:363-366.

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

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

Gribskov M, Veretnik S. 1996. Identification of sequence pattern with profile analysis. Methods Enzymol 266:198–212.

Guenther B, Onrust R, Sali A, O'Donnell M, Kuriyan J. 1997. Crystal structure of the δ-subunit of the clamp-loader complex of *E. coli* DNA polymerase III. Cell 91:335–345.

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

Henikoff S, Henikoff JG. 2000. Amino acid substitution matrices. Adv Protein Chem 54:73–97.

Henikoff JG, Pietrokovski S, McCallum CM, Henikoff S. 2000. Blocks-based methods for detecting protein homology. Electrophoresis 21:1700–1706.

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

Holm L, Sander C. 1991. Database algorithm for generating protein backbone and side-chain coordinates from a C alpha trace application to model building and detection of coordinate errors. J Mol Biol 218:183–194.

Holm L, Sander C. 1997. Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res 25:231–234.

Hooft RW, Vriend G, Sander C, Abola EE. 1996. Errors in protein structures. Nature 381:272.

Howell PL, Almo SC, Parsons MR, Hajdu J, Petsko GA. 1992. Structure determination of turkey egg-white lysozyme using Laue diffraction data. Acta Crystallogr B 48 (Pt 2):200–207.

Ilyin V, Sali, A. Modview. Bioinformatics, in press, 2002.

Jaroszewski L, Rychlewski L, Zhang B, Godzik A. 1998. Fold prediction by a hierarchy of sequence, threading, and modeling methods. Protein Sci 7:1431–1440.

Jaroszewski L, Rychlewski L, Godzik A. 2000. Improving the quality of twilight-zone alignments. Protein Sci 9:1487–1496.

Jennings AJ, Edge CM, Sternberg MJ. 2001. An approach to improving multiple alignments of protein sequences using predicted secondary structure. Protein Eng 14:227–231.

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

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

Jones DT. 1997. Progress in protein structure prediction. Curr Opin Struct Biol 7:377–387.

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

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

Kabsch W, Sander C. 1984. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. Proc Natl Acad Sci USA 81:1075–1078.

Kelley LA, MacCallum RM, Sternberg MJ. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 299:499–520.

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

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

Laskowski RA, Moss DS, Thornton JM. 1993. Main-chain bond lengths and bond angles in protein structures. J Mol Biol 231:1049–1067.

Lesk AM, Chothia C. 1980. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol Biol 136:225–270.

Lessel U, Schomburg D. 1994. Similarities between protein 3-D structures. Protein Eng 7:1175–1187.

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

Lo CL, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. 2000. SCOP: a structural classification of proteins database. Nucleic Acids Res 28:257–259.

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

MacKerell AD Jr., Bashford D, Bellott M, Dunbrack RL, Jr., Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Muczera K, Lau FTK, Mattos C, Michnik S, Nguyen DT, Ngo T, Prodhom B, reiher WE, III, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. 1998. All-atom empirical potential for molecular modleing and dynamics studies of proteins. J Phys Chem B 102:3586–3616.

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. 2000. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325.

Marti-Renom MA, Madhusudhan MS, Fiser A, Rost B, Sali A. 2002. Reliability of assessment of protein structure prediction methods. Structure 10:435–440.

Martin AC, MacArthur MW, Thornton JM. 1997. Assessment of comparative modeling in CASP2. *Proteins Suppl 1*:14-28.

Matsumoto R, Sali A, Ghildyal N, Karplus M, Stevens RL. 1995. Packaging of proteases and proteoglycans in the granules of mast cells and other hematopoietic cells. A cluster of histidines on mouse mast cell protease 7 regulates its binding to heparin serglycin proteoglycans. J Biol Chem 270:19524–19531.

Mezei M. 1998. Chameleon sequences in the PDB. Protein Eng 11:411–414.

Modi S, Paine MJ, Sutcliffe MJ, Lian LY, Primrose WU, Wolf CR, Roberts GC. 1996. A model for human cytochrome P450 2D6 based on homology modeling and NMR studies of substrate binding. Biochemistry 35:4540–4550.

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

Moult J, Pedersen JT, Judson R, Fidelis K. 1995. A large-scale experiment to assess protein structure prediction methods. Proteins 23:ii–iv.

Nagarajaram HA, Reddy BV, Blundell TL. 1999. Analysis and prediction of interstrand packing distances between beta-sheets of globular proteins. Protein Eng 12:1055–1062.

Ohlendorf DH. Accuracy of refined protein structures. Comparison of four independently refined models of human interleukin 1 beta. Acta Crystallogr D Biol Crystallogr D50:808–812. 1994.

Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJ. 1997. An automated classification of the structure of protein loops. J Mol Biol 266:814–830.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH—a hierarchic classification of protein domain structures. Structure 5:1093–1108.

Pearson WR. 1995. Comparison of methods for searching protein sequence databases. Protein Sci 4:1145–1160.

Pearson WR. 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132:185–219.

Pieper U, Eswar N, Ilyin VA, Stuart A, Sali A. 2002. ModBase, a database of annotated comparative protein structure models. Nucleic Acids Res 30:255–259.

Reddy BV, Blundell TL. 1993. Packing of secondary structural elements in proteins. Analysis and prediction of interhelix distances. J Mol Biol 233:464–479.

Retief JD. 2000. Phylogenetic analysis using PHYLIP. Methods Mol Biol 132:243–258.

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

Ring CS, Kneller DG, Langridge R, Cohen FE. 1992. Taxonomy and conformational analysis of loops in proteins. J Mol Biol 224:685–699.

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

Rost B. 1999. Twilight zone of protein sequence alignments. Protein Eng 12:85–94.

Rufino SD, Donate LE, Canard LH, Blundell TL. 1997. Predicting the conformational class of short and medium-size loops connecting regular secondary structures: application to comparative modelling. J Mol Biol 267:352–367.

Sali A. 1995. Modeling mutations and homologous proteins. Curr Opin Biotechnol 6:437–451.

Sali A. 1998. 100,000 protein structures for the biologist. Nat Struct Biol 5:1029–1032.

Sali A, Blundell TL. 1993. Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol 234:779–815.

Sali A, Overington JP. 1994. Derivation of rules for comparative protein modeling from a database of protein structure alignments. Protein Sci 3:1582–1596.

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

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

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

Samudrala R, Moult J. 1998. A graph-theoretic algorithm for comparative modeling of protein structure. J Mol Biol 279:287–302.

Sanchez R, Sali A. 1997a. Advances in comparative protein-structure modeling. Curr Opin Struct Biol 7:206–214.

Sanchez R, Sali A. 1997b. Evaluation of comparative protein structure modeling by MODELLER-3. Proteins Suppl 1:50–58.

Sanchez R, Sali A. 1998. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. Proc Natl Acad Sci USA 95:13597–13602.

Sanchez R, Sali A. 2000. Comparative protein structure modeling. Introduction and practical examples with modeller. Methods Mol Biol 143:97–129.

Sanchez R, Pieper U, Melo F, Eswar N, Marti-Renom MA, Madhusudhan MS, Mirkovic N, Sali A. 2000a. Protein structure modeling for structural genomics. Nat Struct Biol 7 Suppl:986–990.

Sanchez R, Pieper U, Mirkovic N, de Bakker PI, Wittenstein E, Sali A. 2000b. MODBASE, a database of annotated comparative protein structure models. Nucleic Acids Res 28:250–253.

Sauder JM, Arthur JW, Dunbrack RL, Jr. 2000. Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 40:6–22.

Sheng Y, Sali A, Herzog H, Lahnstein J, Krilis SA. 1996. Site-directed mutagenesis of recombinant human beta 2-glycoprotein I identifies a cluster of lysine residues that are critical for phospholipid binding and anti-cardiolipin antibody activity. J Immunol 157:3744–3751.

Shenkin PS, Yarmush DL, Fine RM, Wang HJ, Levinthal C. 1987. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. Biopolymers 26:2053–2085.

Shi J, Blundell TL, Mizuguchi K. 2001. FUGUE: sequence–structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310:243–257.

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

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

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

Sippl MJ. 1995. Knowledge-based potentials for proteins. Curr Opin Struct Biol 5:229–235.

Smith TF, Lo CL, Bienkowska J, Gaitatzes C, Rogers RG, Jr, Lathrop R. 1997. Current limitations to protein threading approaches. J Comput Biol 4:217–225.

Spahn CM, Beckmann R, Eswar N, Penczek PA, Sali A, Blobel G, Frank J. 2001. Structure of the 80S ribosome from *Saccharomyces cerevisiae*—tRNA-ribosome and subunit–subunit interactions. Cell 107:373–386.

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

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

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

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

Teichmann SA, Chothia C, Gerstein M. 1999. Advances in structural genomics. Curr Opin Struct Biol 9:390–399.

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

Torda AE. 1997. Perspectives in protein-fold recognition. Curr Opin Struct Biol 7:200–205.

Totrov M, Abagyan R. 1994. Detailed ab initio prediction of lysozyme–antibody complex with 1.6Å accuracy. Nat Struct Biol 1:259–263.

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

Vakser IA. 1995. Protein docking for low-resolution structures. Protein Eng 8:371–377.

van Gelder CW, Leusen FJ, Leunissen JA, Noordik JH. 1994. A molecular dynamics approach for the generation of complete protein structures from limited coordinate data. Proteins *8*:174–185.

van Vlijmen HW, Karplus M. 1997. PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 267:975–1001.

Vernal J, Fiser A, Sali A, Muller M, Jose CJ, Nowicki C. 2002. Probing the specificity of a trypanosomal aromatic alpha-hydroxy acid dehydrogenase by site-directed mutagenesis. Biochem Biophys Res Commun 293:633–639.

Vitkup D, Melamud E, Moult J, Sander C. 2001. Completeness in structural genomics. Nat Struct Biol 8:559–566.

Westbrook J, Feng Z, Jain S, Bhat TN, Thanki N, Ravichandran V, Gilliland GL, Bluhm W, Weissig H, Greer DS, Bourne PE, Berman HM. 2002. The Protein Data Bank: unifying the archive. Nucleic Acids Res 30:245–248.

Wolf E, Vassilev A, Makino Y, Sali A, Nakatani Y, Burley SK. 1998. Crystal structure of a GCN5-related *N*-acetyltransferase: *Serratia marcescens* aminoglycoside 3-*N*-acetyltransferase. Cell 94:439–449.

Wu G, Fiser A, ter Kuile B, Sali A, Muller M. 1999a. Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. Proc Natl Acad Sci USA 96:6285–6290.

Wu G, Fiser A, ter Kuile B, Sali A, Muller M. 1999b. Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. Proc Natl Acad Sci USA 96:6285–6290.

Wu G, McArthur AG, Fiser A, Sali A, Sogin ML, Müller M. 2000. Core histones of the amitochondriate protist, *Giardia lamblia*. Mol Biol Evol 17:1156–1163.

Xu LZ, Sanchez R, Sali A, Heintz N. 1996. Ligand specificity of brain lipid–binding protein. J Biol Chem 271:24711–24719.

Zemla A, Venclovas, Moult J, Fidelis K. 2001. Processing and evaluation of predictions in CASP4. Proteins 45 Suppl 5:13–21.

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

Zheng Q, Rosenfeld R, DeLisi C, Kyle DJ. 1994. Multiple copy sampling in protein loop modeling: computational efficiency and sensitivity to dihedral angle perturbations. Protein Sci 3:493–506.