Annu. Rev. Biophys. Biomol. Struct. 2000. 29:291–325

Comparative Protein Structure Modeling of Genes and Genomes

Marc A. Martí-Renom, Ashley C. Stuart, András Fiser, Roberto Sánchez, Francisco Melo, and Andrej Šali
Laboratories of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, Rockefeller University, 1230 York Ave, New York, NY 10021; e-mail: sali@rockefeller.edu

Key Words protein structure prediction, fold assignment, alignment, homology modeling, model evaluation, fully automated modeling, structural genomics

■ **Abstract** Comparative modeling predicts the three-dimensional structure of a given protein sequence (target) based primarily on its alignment to one or more proteins of known structure (templates). The prediction process consists of fold assignment, target–template alignment, model building, and model evaluation. The number of protein sequences that can be modeled and the accuracy of the predictions are increasing steadily because of the growth in the number of known protein structures and because of the improvements in the modeling software. Further advances are necessary in recognizing weak sequence–structure similarities, aligning sequences with structures, modeling of rigid body shifts, distortions, loops and side chains, as well as detecting errors in a model. Despite these problems, it is currently possible to model with useful accuracy significant parts of approximately one third of all known protein sequences. The use of individual comparative models in biology is already rewarding and increasingly widespread. A major new challenge for comparative modeling is the integration of it with the torrents of data from genome sequencing projects as well as from functional and structural genomics. In particular, there is a need to develop an automated, rapid, robust, sensitive, and accurate comparative modeling pipeline applicable to whole genomes. Such large-scale modeling is likely to encourage new kinds of applications for the many resulting models, based on their large number and completeness at the level of the family, organism, or functional network.

CONTENTS

INTRODUCTION

The aim of comparative or homology protein structure modeling is to build a three-dimensional (3D) model for a protein of unknown structure (the target) on the basis of sequence similarity to proteins of known structure (the templates) (10, 17, 80, 145, 155). Two conditions must be met to build a useful model. First, the similarity between the target sequence and the template structure must be detectable. Second, a substantially correct alignment between the target sequence and the template structures must be calculated. Comparative modeling is possible because small changes in the protein sequence usually result in small changes in its 3D structure (34). Although considerable progress has been made in ab initio protein structure prediction (92), comparative protein structure modeling remains the most accurate prediction method. The overall accuracy of comparative models spans a wide range, from low resolution models with only a correct fold to more accurate models comparable to medium resolution structures determined by crystallography or nuclear magnetic resonance (NMR) spectroscopy (155). Even low resolution models can be useful in biology because some aspects of function can sometimes be predicted only from the coarse structural features of a model.

The 3D structures of proteins in a family are more conserved than their sequences (101). Therefore, if similarity between two proteins is detectable at the sequence level, structural similarity can usually be assumed. Moreover, even proteins that have nondetectable sequence similarity can have similar structures. It has been estimated that approximately one third of all sequences are recognizably related to at least one known protein structure (54, 77, 81, 144, 157). Because the number of known protein sequences is approximately 500,000 (9), comparative modeling could in principle be applied to more than 150,000 proteins. This number can be compared to approximately 10,000 protein structures determined by experiment (1, 211). The usefulness of comparative modeling is steadily increasing because the number of unique structural folds that proteins adopt is limited (204) and because the number of experimentally determined new structures is increasing exponentially (70). It is possible that in less than 10 years at least one example of most structural folds will be known, making comparative modeling applicable to most protein sequences (70, 155).

All current comparative modeling methods consist of four sequential steps (Figure 1) (155): fold assignment and template selection, template–target alignment, model building, and model evaluation. If the model is not satisfactory, template selection, alignment, and model building can be repeated until a satisfactory

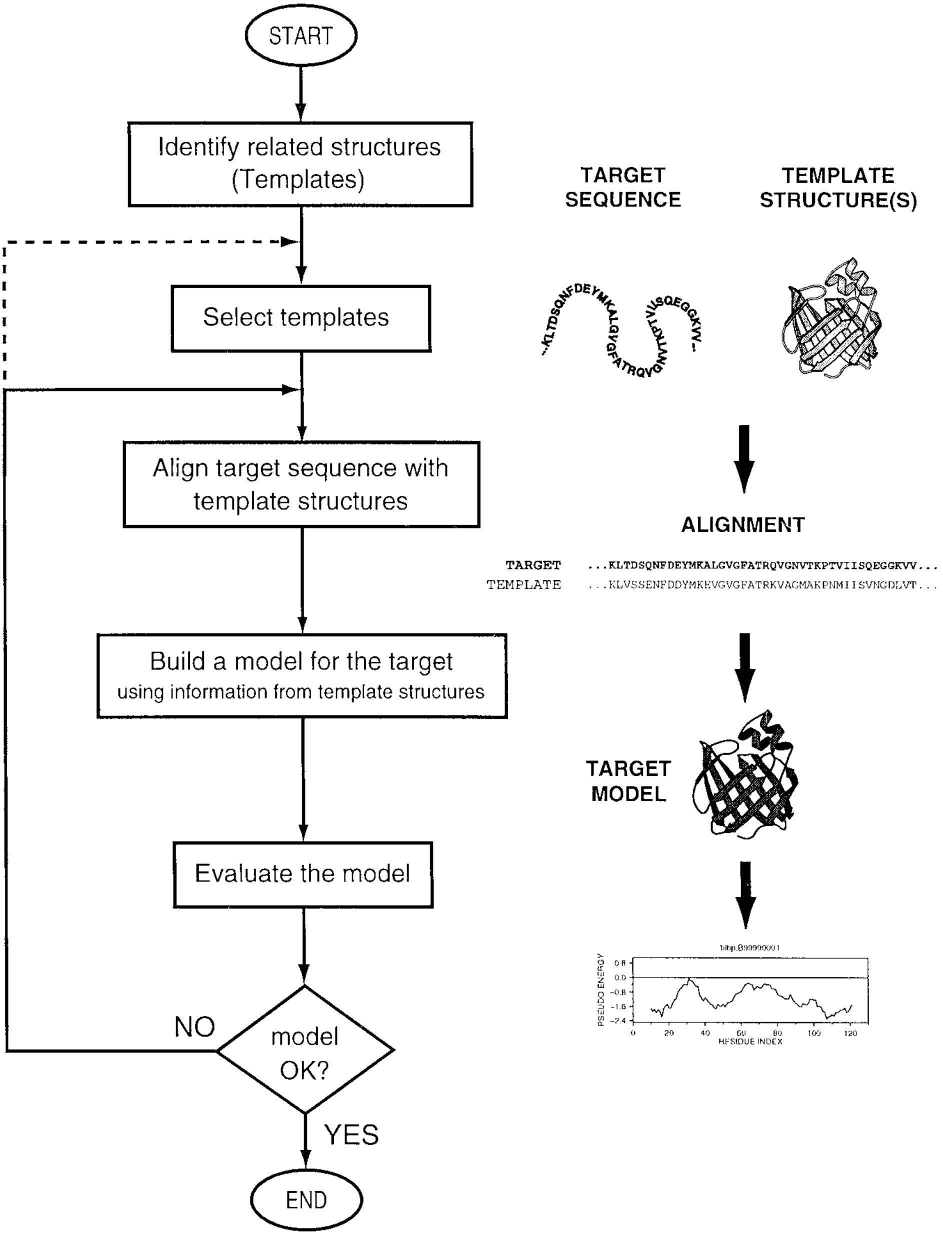
START
Identify related structures (Templates)
Select templates
Align target sequence with template structures
Build a model for the target using information from template structures
Evaluate the model
model OK?
NO
YES
END
TARGET SEQUENCE
TEMPLATE STRUCTURE(S)
...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
ALIGNMENT
TARGET ...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
TEMPLATE ...KLVSSENFDDYMKEVGVGFATRKVAGMAKPNMIISVNGDLVT...
TARGET MODEL
blbp.B99990001
PSEUDO ENERGY
0.8
0.0
−0.8
−1.6
−2.4
0
20
40
60
80
100
120
RESIDUE INDEX

Figure 1 Steps in comparative protein structure modeling. See text for description.

model is obtained. For each of the steps in the modeling process, there are many different methods, programs, and World Wide Web servers (Table 1).

We begin this review by describing the techniques used for all the steps in comparative modeling. We continue by discussing the errors in model structures and methods for detecting these errors, and we conclude by outlining the applications of comparative modeling to individual proteins and to whole genomes. The review focuses on using the methods and tools for comparative modeling, rather than on the physical principles on which the methods are based. The bibliography is not exhaustive, but an attempt was made to quote the latest papers or reviews in the relevant fields.

STEPS IN COMPARATIVE MODELING

Fold Assignment and Template Selection

The starting point in comparative modeling is to identify all protein structures related to the target sequence, and then to select those that will be used as templates. This step is facilitated by numerous protein sequence and structure databases, and database scanning software available on the World Wide Web (4, 11, 70, 163) (Table 1). Templates can be found using the target sequence as a query for searching structure databases such as the Protein Data Bank (1, 211), SCOP (76), DALI (71), and CATH (124). Depending on a genome, the probability of finding a related protein of known structure for a sequence randomly picked from a genome ranges from 20% to 50% (54, 77, 81, 144, 157).

There are three main classes of protein comparison methods that are useful in fold identification. The first class includes the methods that compare the target sequence with each of the database sequences independently, using pairwise sequence–sequence comparison (7). The performance of these methods in searching for related protein sequences (128) and structures (22) has been evaluated exhaustively. Frequently used programs in this class include FASTA (129) and BLAST (5).

The second set of methods relies on multiple sequence comparisons to improve the sensitivity of the search (6, 63, 67, 94, 144). A widely used program in this class is PSI-BLAST (6), which iteratively expands the set of homologs of the target sequence. For a given sequence, an initial set of homologs from a sequence database is collected, a weighted multiple alignment is made from the query sequence and its homologs, a position-specific scoring matrix is constructed from the alignment, and the matrix is used to search the database for new homologs. These steps are repeated until no new homologs are found. In comparison to BLAST, PSI-BLAST finds homologs of known structure for approximately twice as many sequences (125, 176). A related approach (144) also begins by finding all sequences clearly related to the target sequence to obtain the target sequence profile. In addition, similar profiles are constructed for all known protein structures. The potential templates are then found by comparing the target sequence profile with each of the sequence profiles for known structures. Another variation uses multiple sequence

TABLE 1 Programs and World Wide Web servers useful in comparative modeling

Name	Type[a]	World Wide Web address[b]	Reference[c]
Databases			
CATH	S	www.biochem.ucl.ac.uk/bsm/cath/	124
GenBank	S	www.ncbi.nlm.nih.gov/GenBank	15
GeneCensus	S	bioinfo.mbb.yale.edu/genome	58
ModBase	S	guitar.rockefeller.edu/modbase/	159
PDB	S	www.rcsb.org/pdb/	16
Presage	S	presage.stanford.edu	21
Scop	S	scop.mrc-lmb.cam.ac.uk/scop/	76
SWISSPROT+TrEMBL	S	www.ebi.ac.uk/swissprot	9
Template search			
123D	S	www.lmmb.ncifcrf.gov/∼nicka/123D.html	2
BLAST	S	www.ncbi.nlm.nih.gov/BLAST/	5
DALI	S	www2.ebi.ac.uk/dali/	71
FastA	S	www2.ebi.ac.uk/fasta3	127
MATCHMAKER	P	bioinformatics.burnham-inst.org	59
PHD, TOPITS	S	www.embl-heidelberg.de/predictprotein/predictprotein.html	139
PROFIT	P	www.came.sbg.ac.at	57
THREADER	P	globin.bio.warwick.ac.uk/∼jones/threader.html	82
UCLA-DOE FRSVR	S	www.doe-mbi.ucla.edu/people/frsvr/frsvr.html	53
Sequence alignment			
BCM SERVER	S	dot.imgen.bcm.tmc.edu:9331/	170
BLAST	S	www.ncbi.nlm.nih.gov/BLAST	6
BLOCK MAKER	S	blocks.fhcrc.org/blocks/blockmkr/make_blocks.html	68
CLUSTAL	S	www2.ebi.ac.uk/clustalw/	78
FASTA3	S	www2.ebi.ac.uk/fasta3/	127
MULTALIN	S	pbil.ibcp.fr/	41
Modeling			
COMPOSER	P	www-cryst.bioc.cam.ac.uk	179
CONGEN	P	www.congenomics.com/congen/congen.html	29
CPH models	S	www.cbs.dtu.dk/services/CPHmodels/	206
DRAGON	P	www.nimr.mrc.ac.uk/∼mathbio/a-aszodi/dragon.html	8
ICM	P	www.molsoft.com	(a)
InsightII	P	www.msi.com	(b)
MODELLER	P	guitar.rockefeller.edu/modeller/modeller.html	148
LOOK	P	www.mag.com	102
QUANTA	P	www.msi.com	(b)
SYBYL	P	www.tripos.com	(c)
SCWRL	P	www.cmpharm.ucsf.edu/∼bower/scrwl/scrwl.html	19
SWISS-MOD	S	www.expasy.ch/swissmod	131
WHAT IF	P	www.sander.embl-heidelberg.de/whatif/	194

(*continued*)

TABLE 1 (*Continued*)

Name	Type[a]	World Wide Web address[b]	Reference[c]
Model evaluation			
ANOLEA	S	www.fundp.ac.be/pub/ANOLEA.html	113
AQUA	P	www-nmr.chem.ruu.nl/users/rull/aqua.html	98
BIOTECH[d]	S	biotech.embl-ebi.ac.uk:8400/	73, 96
ERRAT	S	www.doe-mbi.ucla.edu/errat_server.html	40
PROCHECK	P	www.biochem.ucl.ac.uk/∼roman/procheck/procheck.html	96
ProCeryon[e]	P	www.proceryon.com/	(d)
ProsaII[e]	P	www.came.sbg.ac.at	169
PROVE	S	www.ucmb.ulb.ac.be/UCMB/PROVE	134
SQUID	P	www.yorvic.york.ac.uk/∼oldfield/squid	121
VERIFY3D	S	www.doe-mbi.ucla.edu/verify3d.html	105
WHATCHECK	P	www.sander.embl-heidelberg.de/whatcheck/	73

[a] S, server; P, program.

[b] Some of the sites are mirrored on additional computers.

[c] (a) MolSoft Inc., San Diego. (b) Molecular Simulations Inc., San Diego. (c) Tripos Inc., St Louis. (d) ProCeryon Biosciences Inc. New York.

[d] The BIOTECH server uses PROCHECK and WHATCHECK for structure evaluation.

[e] ProCyon is a new software package that includes PeeP, PROFIT and PROSUP, a structure comparison program.

alignments combined with structural information predicted from the sequence of the target (54). The multiple sequence methods for fold identification are especially useful for finding significant structural relationships when the sequence identify between the target and the template drops below 25%. This class of methods appears to be one of the most sensitive fully-automated approaches for detecting remote sequence–structure relationships (6, 77, 203, 213).

The third class of methods are the so-called threading or 3D template matching methods (20, 59, 82), reviewed in (83, 103, 173, 185). These methods rely on pairwise comparison of a protein sequence and a protein of known structure. Whether or not a given target sequence adopts any one of the many known 3D folds is predicted by an optimization of the alignment with respect to a structure-dependent scoring function, independently for each sequence–structure pair. That is, the target sequence is threaded through a library of 3D folds. These methods are especially useful when there are no sequences clearly related to the modeling target, and thus the search cannot benefit from the increased sensitivity of the sequence profile methods.

A useful fold assignment approach is to accept an uncertain assignment provided by any of the methods, build a full-atom comparative model of the target sequence based on this match, and make the final decision about whether or not the match is real by evaluating the resulting comparative model (64, 115, 156).

Once a list of all related protein structures has been obtained, it is necessary to select those templates that are appropriate for the given modeling problem.

Usually, a higher overall sequence similarity between the target and the template sequence yields a better template. In any case, several other factors should be taken into account when selecting the templates:

- The family of proteins, which includes the target and the templates, can frequently be organized in subfamilies. The construction of a multiple alignment and a phylogenetic tree (46) can help in selecting the template from the subfamily that is closest to the target sequence.
- The template "environment" should be compared to the required environment for the model. The term environment is used in a broad sense and includes all factors that determine protein structure except its sequence (e.g., solvent, *p*H, ligands, and quaternary interactions).
- The quality of the experimental template structure is another important factor in template selection. The resolution and the R-factor of a crystallographic structure and the number of restraints per residue for an NMR structure are indicative of its accuracy.

The priority of the criteria for template selection depends on the purpose of the comparative model. For instance, 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 is preferable to use a high resolution template. It is not necessary to select only one template. In fact, the use of several templates approximately equidistant from the target sequence generally increases the model accuracy (156, 174).

Target–Template Alignment

Most fold assignment methods produce an alignment between the target sequence and template structures. However, this is often not the optimal target–template alignment for comparative modeling. Searching methods are usually tuned for detection of remote relationships, not for optimal alignments. Therefore, once templates have been selected, a specialized method should be used to align the target sequence with the template structures (14, 23, 70, 171, 180). For closely related protein sequences with identity over 40%, the alignment is almost always correct. Regions of low local sequence similarity become common when the overall sequence identity is under 40% (160). The alignment becomes difficult in the "twilight zone" of less than 30% sequence identity (140). As the sequence similarity decreases, alignments contain an increasingly large number of gaps and alignment errors, regardless of whether they are prepared automatically or manually. For example, only 20% of the residues are likely to be correctly aligned when two proteins share 30% sequence identity (79). Maximal effort to obtain the most accurate alignment possible is needed because no current comparative modeling method can recover from an incorrect alignment. There is a great variety of protein sequence alignment methods, many of which are based on dynamic programming techniques (120, 172). A frequently used program for multiple sequence alignment is CLUSTAL (78), which is also available as a World Wide Web server (Table 1).

In the more difficult alignment cases, it is frequently beneficial to rely on multiple structure and sequence information (12, 181). First, the alignment of the potential templates is prepared by superposing their structures. Next, the sequences that are clearly related to the templates and are easily aligned with them are added to the alignment. The same is done for the target sequence. Finally, the two profiles are aligned with each other, taking structural information into account as much as possible (78, 93, 152). In principle, most sequence alignment and structure comparison methods can be used for these tasks (11, 70, 104, 171). The information from structures helps to avoid gaps in secondary structure elements, in buried regions, or between two residues that are far in space. It is generally necessary to check and edit the alignment by inspecting the template structures visually, especially if the target–template sequence identity is low. Secondary structure predictions for the target sequence and its profile are also frequently useful in obtaining a more accurate alignment (3, 141). Because evaluation of a model is more reliable than an evaluation of an alignment, a good way to proceed in the difficult cases is to generate 3D models for all alternative alignments, evaluate the corresponding models, and pick the best model according to the 3D model evaluation rather than the alignment score (64, 115, 156).

Model Building

Once an initial target–template alignment has been built, a variety of methods can be used to construct a 3D model for the target protein. The original and still widely used method is modeling by rigid-body assembly (17, 27, 61). Another family of methods, modeling by segment matching, relies on the approximate positions of conserved atoms in the templates (38, 85, 102, 187). The third group of methods, modeling by satisfaction of spatial restraints, uses either distance geometry or optimization techniques to satisfy spatial restraints obtained from the alignment (8, 24, 66, 148, 175). Accuracies of the various model building methods are relatively similar when used optimally. 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 allows a degree of flexibility and automation, making it easier and faster to obtain better models. For example, a method should permit 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 the tools to incorporate prior knowledge about the target (e.g., experimental data or predicted features such as secondary structure). There are many reviews of comparative model building methods (10, 17, 80, 145, 155). Several programs and World Wide Web servers for comparative modeling are listed in Table 1.

Modeling by Assembly of Rigid Bodies This method assembles a model from a small number of rigid bodies obtained from aligned protein structures. The approach is based on the natural dissection of the protein folds into conserved core regions, variable loops, and side chains. For example, the following semiautomated procedure is implemented in the computer program COMPOSER (179). First, the template structures are selected and superposed. Second, the frame-work is calculated by averaging the coordinates of the C_α atoms of structurally conserved regions in the template structures. Third, the main chain atoms of each core region in the target model are generated by superposing the core segment from the template with the highest sequence similarity to the target on the framework. 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. 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. And finally, the stereochemistry of the model is improved either by a restrained energy minimization or a molecular dynamics refinement. The accuracy of a model can be somewhat increased if more than one template structure is used to construct the framework and if the templates are averaged into the framework using weights corresponding to their sequence similarities to the target sequence (174). For example, differences between the model and X-ray structures may be slightly smaller than the differences between the X-ray structure of the modeled protein and the best template used to build the model. Future improvements in modeling by rigid body assembly may include the incorporation of rigid body shifts, such as the relative shifts in the packing of α-helices.

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 approximately 100 structural classes (187). 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. Usually the C_α atoms of the segments, which are conserved in the alignment between the template structure and the target sequence, are taken as the guiding positions. 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 (38, 69), or by a conformational search restrained by an energy function (13, 190). For example, a general method for modeling by segment matching is guided by the positions of some atoms (usually C_α atoms) to find the matching segments in a representative database of all known protein structures (102). This method can construct both main chain and side chain atoms and can also model insertions and deletions. It is implemented in the program SEGMOD (Table 1). Some side chain modeling methods (195) and loop construction methods based on finding suitable fragments in the database of known structures (85) can be seen as segment matching or coordinate reconstruction methods.

Modeling by Satisfaction of Spatial Restraints The methods in this class generate many constraints or restraints on the structure of the target sequence, using its alignment to related protein structures as a guide. The restraints are generally obtained by assuming that the corresponding distances and angles 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 obtained from a molecular mechanics force field. The model is then derived by minimizing the violations of all the restraints. This can be achieved either by distance geometry or real–space optimization. A distance geometry approach constructs all-atom models from lower and upper bounds on distances and dihedral angles (66). A real–space optimization method, such as that implemented in the computer program MODELLER (148), starts by building the model using the distance and dihedral angle restraints on the target sequence derived from its alignment with template 3D structures. Then, the spatial restraints and the CHARMM22 force field terms, which enforce proper stereochemistry (106), are combined into an objective function. Finally, the model is generated by optimizing the objective function in Cartesian space. Because 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 might be obtained from NMR experiments, cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, site-directed mutagenesis, etc. In this way, a comparative model, especially in the difficult cases, could be improved by making it consistent with available experimental data.

Loop Modeling

In a given fold family, structural variability is a result of substitutions, insertions, and deletions of residues between members of the family. Such changes frequently correspond to exposed loop regions that connect elements of secondary structure in the protein fold. Thus, loops often determine the functional specificity of a given protein framework. They contribute to active and binding sites. Examples include binding of metal ions by metal-binding proteins, small protein toxins by their receptors, antigens by immunoglobulins, nucleotides by a variety of proteins, protein substrates by serine proteases, and DNA by DNA-binding proteins. Consequently, the accuracy of loop modeling is a major factor determining the usefulness of comparative models in studying interactions between the protein and its ligands. This includes the use of models for recognizing ligand binding sites (47, 84) and for ligand docking computations (87). Unfortunately, no generally reliable method is available for constructing loops longer than 5 residues (108), although recently some progress has been made and occasionally longer loops can be modeled correctly (122, 135, 142, 153, 191). An exhaustive set of references for loop modeling papers can be found in (55).

Loop modeling can be seen as a mini–protein folding problem. 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. Segments of up to 9 residues sometimes have entirely unrelated conformations in different proteins (114). Thus, the conformation of a given segment is also influenced by the core stem regions that span the loop and by the structure of the rest of a protein that cradles the loop.

Many loop modeling procedures have been described. Similarly to the prediction of whole protein structures, there are ab initio methods (30, 50, 119), database search techniques (35, 60, 85), and procedures that combine these two basic approaches (107, 191).

The ab initio loop prediction is based on a conformational search or enumeration of conformations in a given environment, guided by a scoring or energy function. There are many such methods, exploiting different protein representations, energy function terms, and optimization or enumeration algorithms. The search algorithms include sampling of main chain dihedral angles biased by their distributions in known protein structures (119), minimum perturbation random tweak method (165), systematic conformational search (26, 29), global energy minimization by mapping a trajectory of local minima (43), importance-sampling by local minimization of randomly generated conformations (95), local energy minimization (110), molecular dynamics simulations (28, 55), genetic algorithms (111, 136), biased probability Monte Carlo search (45, 183), Monte Carlo with simulated annealing (33, 39, 192), Monte Carlo and molecular dynamics (135), extended-scaled-collective-variable Monte Carlo (88), scaling relaxation and multiple copy sampling (138, 205), searching through discrete conformations by dynamic programming (51, 188), random sampling of conformations relying on dimers from known protein structures (177), self-consistent field optimization (91), and an enumeration based on the graph theory (153). A variety of representations have been used, such as unified atoms, all nonhydrogen atoms, nonhydrogen and “polar” hydrogen atoms, all atoms, as well as implicit and explicit solvent models. The optimized degrees of freedom include Cartesian coordinates and internal coordinates, such as dihedral angles, optimized in continuous or discrete spaces. Loop prediction by optimization is in principle applicable to simultaneous modeling of several loops and loops interacting with ligands, which is not straightforward for the database search approaches.

The database approach to loop prediction consists of finding a segment of main chain that fits the two stem regions of a loop (36, 60, 85, 122, 142, 166, 178, 198). The stems are defined as the main chain atoms that precede and follow the loop but are not part of it. They span the loop and are part of the core of the fold. The search is performed through a database of many known protein structures, not only homologs of the modeled protein. Usually, many different alternative segments that fit the stem residues are obtained and possibly sorted according to geometric criteria or sequence similarity between the template and target loop sequences. The selected segments are then superposed and annealed on the stem regions. These initial crude models are often refined by optimization of some energy function. The database search approach to loop modeling is accurate and efficient when a specific set of loops is created to address the modeling of that class of loops, such as β-hairpins (167) and the hypervariable regions in immunoglobulins (36). For immunoglobulins, an analysis of the hypervariable regions in known immunoglobulin structures resulted in rules with high prediction accuracy for other members of the family. These rules are possible because of the relatively small number of conformations for each loop and because of the dependence of loop conformation on loop length and certain key residues. There are attempts to classify loop conformations into more general categories, thus extending the impressive performance of the key residues approach to more cases (122, 142, 198). However, the database methods are limited by the exponential increase in the number of geometrically possible conformations as a function of loop length. Consequently, only segments of 7 residues or less have most of their conceivable conformations present in the database of known protein structures (48).

The problem of database completeness has recently been ameliorated by restrained energy minimization of the candidate loops obtained from a database search (191). Both the internal conformation and the global orientation relative to the rest of the protein have been optimized. It was concluded that the candidate segments from a database were suitable starting points for modeling loops up to 9 residues long, but extensive optimization was required for loops longer than 4 residues.

Sidechain Modeling

As with loops, side chain conformations are predicted from similar structures and from steric or energetic considerations (145, 193). Disulphide bridges can be treated as a special case in side-chain modeling. They are modeled using structural information from proteins in general (86) and from equivalent disulfide bridges in related structures (150).

Vásquez reviewed and commented on various approaches to side-chain modeling (193). The importance of two effects on sidechain conformation was emphasized: The first effect was the coupling between the main chain and side chains, and the second effect was the continuous nature of the distributions of side-chain dihedral angles; for example, 5–30% of side chains in crystal structures are significantly different from their rotamer conformations (162) and 6% of the χ_1 or χ_2 values are not within $\pm 40^\circ$ of any rotamer conformation (19). Both effects appear to be important when correlating packing energies and stability (100). The correct energetics may be obtained for the incorrect reasons; i.e., the side chains may adopt distorted conformations to compensate for the rigidity of the backbone.

Correspondingly, the backbone shifts may hinder the use of these methods when the template structures are related at less than 50% sequence identity (37). Some attempts to include backbone flexibility in side-chain modeling have been described (65, 75, 91) but are not yet generally applicable.

Significant correlations were found between side-chain dihedral angle probabilities and backbone Φ, Ψ values (19). These correlations go beyond the dependence of side chain conformation on the secondary structure (112). For example, the preferred rotamers can vary within the same secondary structure, even when Φ, Ψ dihedral angles change by as little as 20° (19). Because these changes are smaller than the differences between closely related homologs, the prediction of the side-chain conformations generally cannot be uncoupled from the backbone prediction. This partly explains why the conformation of equivalent side chains in homologous structures is useful in side-chain modeling (148).

Chung & Subbiah gave an elegant structural explanation for the rapid decrease in the conservation of sidechain packing as the sequence identity falls below 30% (37). Although the fold is maintained, the pattern of side-chain interactions is generally lost in this range of sequence similarity (143). Two sets of computations were done for two sample protein sequences: The side-chain conformation was predicted by maximizing packing on the fixed native backbone and on a fixed backbone with approximately 2 Å RMSD from the native backbone; the 2 Å RMSD generally corresponds to the differences between the conserved cores of two proteins related at 25–30% sequence identity. The side-chain predictions on the two backbones turned out to be unrelated. Thus, inasmuch as packing reflects the true laws determining side-chain conformation, a backbone with less than 30% sequence identity to the sequence being modeled is not sufficient to produce the correct packing of the buried side chains.

The solvation term is important for the modeling of exposed sidechains (42, 133, 161, 196), many of which are expected to be highly flexible without a single dominant conformation. It was also demonstrated that treating hydrogen bonds explicitly could significantly improve side-chain predictions (44, 49). Calculations that do not account for the solvent, either implicitly or explicitly, may introduce errors in the hydrogen-bonding patterns even in the core regions of a protein (133).

A recent survey analyzed the accuracy of three different side-chain prediction methods (75). These methods were tested by predicting side-chain conformations on near-native protein backbones with <4 Å RMSD to the native structures. The three methods include the packing of backbone-dependent rotamers (19), the self-consistent mean-field approach to positioning rotamers based on their van der Waals interactions (89), and the segment-matching method of Levitt (102). The accuracies of the methods were surprisingly similar. All were able to correctly predict approximately 50% of χ_1 angles and 35% of both χ_1 and χ_2 angles. In typical comparative modeling applications where the backbone is closer to the native structures (<2 Å RMSD), these numbers increase by approximately 20% (151).

ERRORS IN COMPARATIVE MODELS

As the similarity between the target and the templates decreases, the errors in the model increase. Errors in comparative models can be divided into five categories (151, 156) (Figure 2):

- Errors in side-chain packing. As the sequences diverge, the packing of side chains in the protein core changes. Sometimes even the conformation of identical side chains is not conserved, a pitfall for many comparative modeling methods. Side-chain errors are critical if they occur in regions that are involved in protein function, such as active sites and ligand-binding sites.
- Distortions and shifts in correctly aligned regions. As a consequence of sequence divergence, the main chain conformation changes, even if the overall fold remains the same. Therefore, it is possible that in some correctly aligned segments of a model, the template is locally different (<3 Å) from the target, resulting in errors in that region. The structural differences are sometimes not due to differences in sequence but are a consequence of artifacts in structure determination or structure determination in different environments (e.g., packing of subunits in a crystal). The simultaneous use of several templates can minimize this kind of error (156, 174).
- Errors in regions without a template. Segments of the target sequence that have no equivalent region in the template structure (i.e., insertions or loops) are the most difficult regions to model. If the insertion is relatively short, less than 9 residues long, some methods can correctly predict the conformation of the backbone (191). Conditions for successful prediction are the correct alignment and an accurately modeled environment surrounding the insertion.
- Errors due to misalignments. The largest source of errors in comparative modeling are misalignments, especially when the target–template sequence identity decreases below 30%. However, alignment errors can be minimized in two ways. First, it is usually possible to use a large number of sequences to construct a multiple alignment, even if most of these sequences do not have known structures. Multiple alignments are generally more reliable than pairwise alignments (12, 181). The second way of improving the alignment is to iteratively modify those regions in the alignment that correspond to predicted errors in the model (156).
- Incorrect templates. This is a potential problem when distantly related proteins are used as templates (i.e., less than 25% sequence identity). Distinguishing between a model based on an incorrect template and a model based on an incorrect alignment with a correct template is difficult. In both cases, the evaluation methods will predict an unreliable model. The

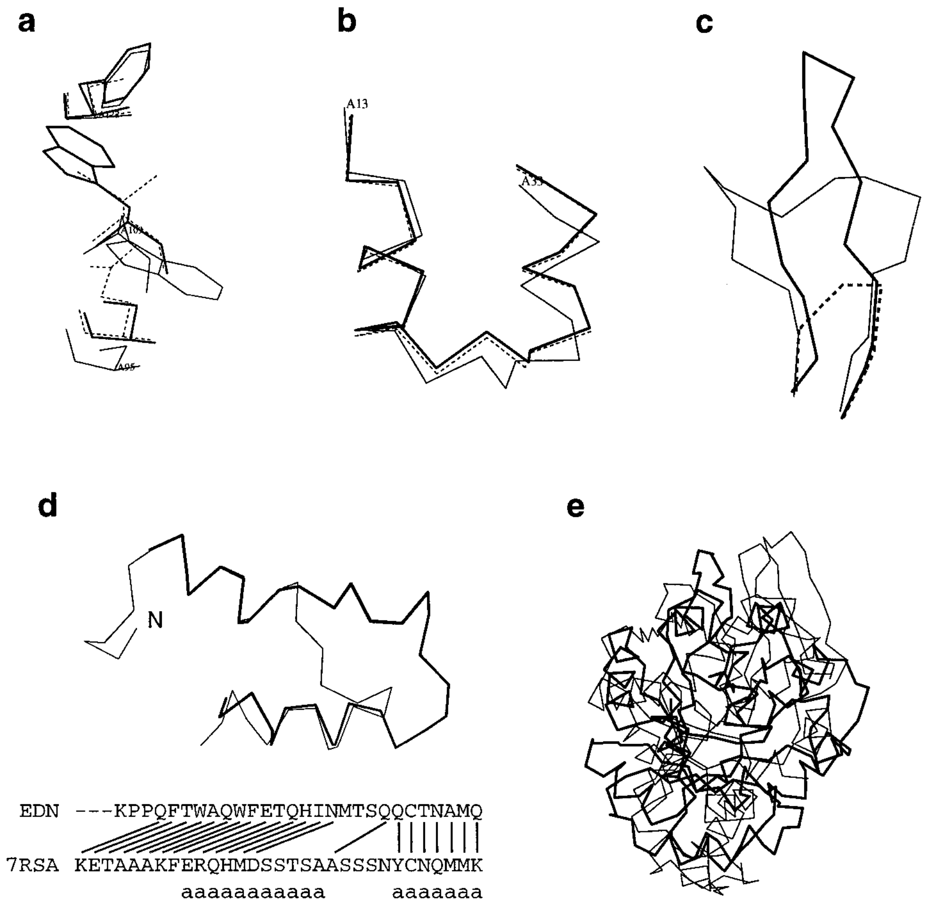
a
A109
A95
b
A13
A33
c
d
N
EDN ---KPPQFTWAQWFETQHINMTSQQCTNAMQ
7RSA KETAAAKFERQHMDSSTSAASSSNYCNQMMK
aaaaaaaaaaa aaaaaaa
e

Figure 2 Typical errors in comparative modeling (151, 156). (*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 C_α 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 C_α 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.

conservation of the key functional or structural residues in the target sequence increases the confidence in a given fold assignment.

Comparative modeling has been criticized for its inability to provide a final model closer to the target-experimental structure than the template used to generate the model (108). This is only the case when there are errors in the template–target alignment used for modeling. When the evaluation of the template–target similarity is based on the template–target alignment used for modeling the model is generally closer to the target structure than is any of the templates (156).

An informative way to test protein structure modeling methods as well as the modelers using them is provided by the biannual meetings on Critical Assessment of Techniques for Protein Structure Prediction (CASP) (92, 118). The last meeting was held in December of 1998 and is summarized in the special issue of *Proteins* Suppl. 3, 1999 (212). 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 only become available after the models are calculated and submitted. Thus, a bona fide evaluation of protein structure modeling methods is possible. An important extension of the CASP meetings is a completely automated and online evaluation of protein structure modeling servers on the World Wide Web. The idea has so far been implemented in the threading category only (52). It is likely to be extended to other kinds of structural prediction.

EVALUATION OF MODELS

The quality of the predicted model determines the information that can be extracted from it. Thus, estimating the accuracy of 3D protein models is essential for interpreting them. The model can be evaluated as a whole as well as in individual regions. There are many model evaluation programs and servers (97, 197) (Table 1).

The first step in model evaluation is to assess if the model has the correct fold (157). A model will have the correct fold if the correct template is picked and if that template is aligned at least approximately correctly with the target sequence. The fold of a model can be assessed by a high sequence similarity with the closest template, an energy based Z-score (157, 169), or by conservation of the key functional or structural residues in the target sequence.

Once the fold of a model is assessed, a more detailed evaluation of the overall model accuracy can be obtained based on the similarity between the target and template sequences (Figure 3) (157). Sequence identity above 30% is a relatively good predictor of the expected accuracy. The reasons are the well-known relationship between structural and sequence similarities of two proteins (34), the ‘geometrical’ nature of modeling that forces the model to be as close to the template as possible

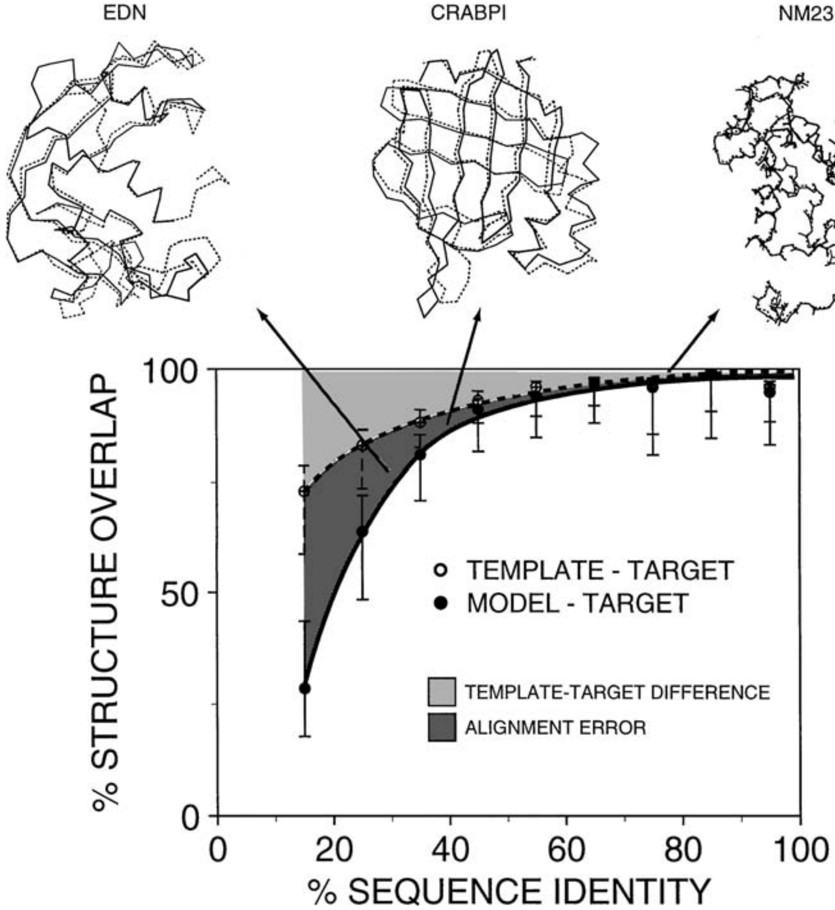

Figure 3 Average model accuracy as a function of the template–target sequence similarity. At the top, sample models (solid line) at three typical accuracy levels are compared with corresponding actual structures (dotted line). The models were calculated with MODELLER in a completely automated fashion before the experimental structures were available (151). When multiple sequence and structure information is used, and the alignments are edited by hand, the models can be significantly more accurate than shown here (156). At the bottom, the models of known protein structures used to determine the dependence of the overall model accuracy as a function of template–target sequence identity were calculated entirely automatically, based on single templates (157). Percentage structure overlap is defined as the fraction of equivalent residues. Two residues are equivalent when their C_α atoms are within 3.5 Å of each other upon rigid-body, least-squares superposition of the two structures.

(148), and the inability of any current modeling procedure to recover from an incorrect alignment (156). The dispersion of the model–target structural overlap increases with the decrease in sequence identity. If the target–template sequence identity falls below 30%, the sequence identity becomes unreliable as a measure of expected accuracy of a single model. Models that deviate significantly from the average accuracy are frequent. It is in such cases that model evaluation methods are particularly useful.

In addition to the target–template sequence similarity, 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 the target, it is likely that the model will be incorrect irrespective of the target–template similarity or accuracy of the template structure (126). This also applies to the experimental determination of protein structure; a structure must be determined in the functionally meaningful environment.

A basic requirement for a model is to have good stereochemistry. Some useful programs for evaluating stereochemistry are PROCHECK (96), PROCHECK-NMR (98), AQUA (98), SQUID (121), and WHATCHECK (72). The features of a model that are checked by these programs include bond lengths, bond angles, peptide bond and side-chain ring planarities, chirality, main-chain and side-chain torsion angles, and clashes between nonbonded pairs of atoms.

Distributions of many spatial features have been compiled from high resolution protein structures, and large deviations from the most likely values have been interpreted as strong indicators of errors in the model. Such features include packing (62), formation of a hydrophobic core (31), residue and atomic solvent accessibilities (90), spatial distribution of charged groups (32), distribution of atom–atom distances (40), atomic volumes (134), and main-chain hydrogen bonding (96).

There are also methods for testing 3D models that implicitly take into account many of the criteria listed above. These methods are based on 3D profiles and statistical potentials of mean force (105, 168). Programs implementing this approach include VERIFY3D (105), PROSAII (169), HARMONY (184), and ANOLEA (113). The programs evaluate the environment of each residue in a model with respect to the expected environment as found in the high-resolution X-ray structures. There is a concern about the theoretical validity of the energy profiles for detecting regional errors in models. It is likely that the contributions of the individual residues to the overall free energy of folding vary widely, even when normalized by the number of atoms or interactions made. If this is correct, the correlation between the prediction errors and energy peaks is greatly weakened, resulting in the loss of predictive power of the energy profile. Despite these concerns, error profiles have been useful in some applications (115).

Recently, a physics-based approach to deriving energy functions has been tested for use in protein structure evaluation. Lazaridis & Karplus (99) used an effective energy function that combined the CHARMM (25) vacuum potential with a Gaussian model for the solvation free energy. The results showed that the native state was always more stable than grossly misfolded conformations. Moreover, the authors concluded that molecular mechanics energy functions, complemented by a simple model for the solvation free energy, can perform as well as statistical functions in discriminating correct and misfolded models.

APPLICATIONS OF COMPARATIVE MODELING

Comparative modeling is an increasingly 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 a protein's function (18, 200), identifying active and binding sites (164), identifying, designing and improving ligands for a given binding site (137), modeling substrate specificity (201), predicting antigenic epitopes (149), simulating protein–protein docking (189), inferring function from a calculated electrostatic potential around the protein (109), facilitating molecular replacement in X-ray structure determination (74), refining models based on NMR constraints (116), testing and improving a sequence–structure alignment (199), confirming a remote structural relationship (64, 115), and rationalizing known experimental observations. For an exhaustive review of comparative modeling applications see (80).

Fortunately, a 3D model does not have to be absolutely perfect to be helpful in biology, as demonstrated by the applications listed above. However, the type of question that can be addressed with a particular model does depend on its accuracy (Figure 4). Three levels of model accuracy and some of the corresponding applications are as follows.

- At the low end of the spectrum, there are models based on less than 30% sequence identity and have sometimes less than 50% of their C_α atoms within 3.5 Å of their correct positions. Such models still have the correct fold, which is frequently sufficient to predict 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 (123). Models in this low range of accuracy combined with model evaluation can be used to confirm or reject a match between remotely related proteins (156, 157).
- In the middle of the accuracy spectrum are the models based on approximately 30–50% sequence identity, corresponding to 85% of the C_α 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 (157). In general, medium

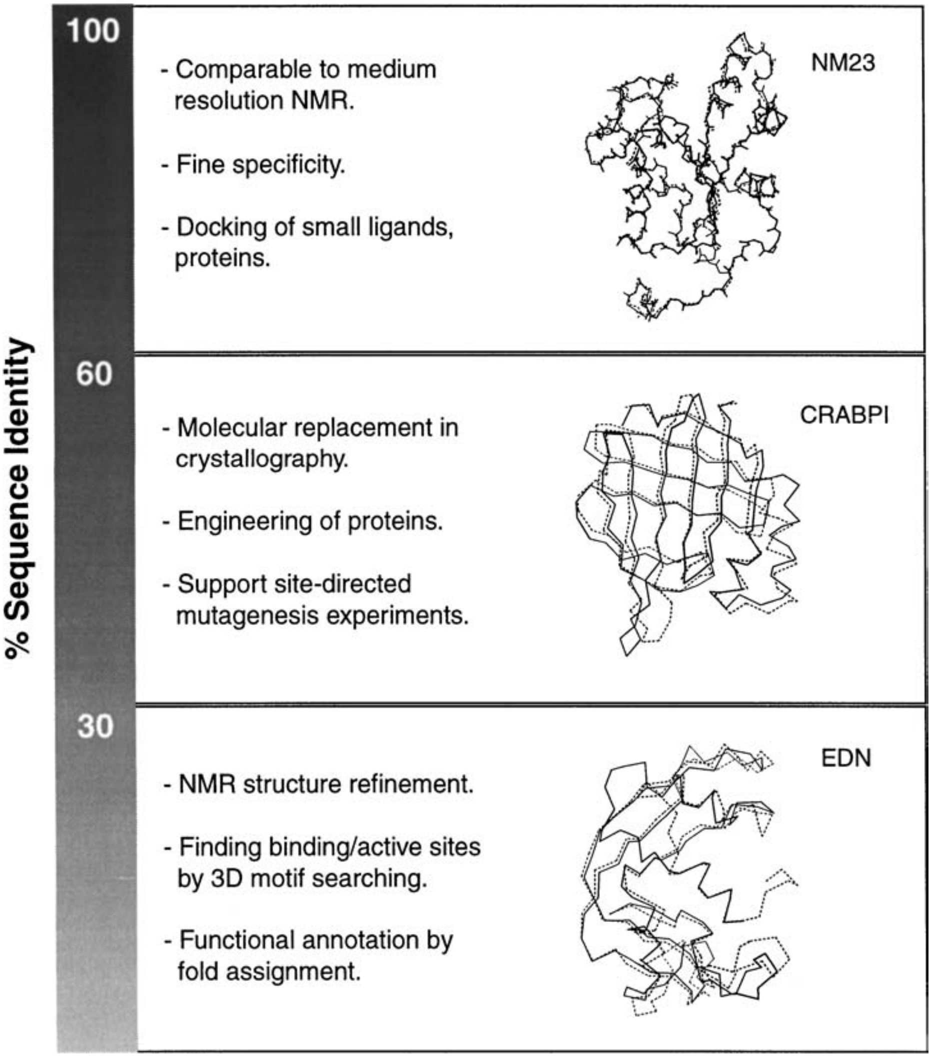
% Sequence Identity
100
- Comparable to medium resolution NMR.
- Fine specificity.
- Docking of small ligands, proteins.
NM23
60
- Molecular replacement in crystallography.
- Engineering of proteins.
- Support site-directed mutagenesis experiments.
CRABPI
30
- NMR structure refinement.
- Finding binding/active sites by 3D motif searching.
- Functional annotation by fold assignment.
EDN

Figure 4 Applications of comparative modeling. The potential uses of a comparative model depend on its accuracy. This in turn depends significantly on the sequence identity between the target and the template structure on which the model was based. Sample models from Figure 3 are shown on the right.

resolution models frequently allow 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 by 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 (109), and the size of a ligand can be predicted from the volume of the binding site cleft (201). 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 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; or designing proteins with added disulfide bonds for extra stability.

- The high end of the accuracy spectrum corresponds to models based on more than 50% sequence identity. The average accuracy of these models approaches that of low resolution X-ray structures (3 Å resolution) or medium resolution NMR structures (10 distance restraints per residue) (156). 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 (137) or whole proteins onto a given protein (186, 189).

COMPARATIVE MODELING IN STRUCTURAL GENOMICS

The aim of structural genomics is to determine or accurately predict the 3D structure of all the proteins encoded in the genomes (117, 147, 182, 202, 207–209). This aim will be achieved by a focused, large-scale determination of protein structures by X-ray crystallography and NMR spectroscopy, combined efficiently with accurate protein structure modeling techniques. The structural genomics project will deliver improved methods and a process for high-throughput protein structure determination. The process involves (*a*) selection of the target proteins or domains, (*b*) cloning, expression, and purification of the targets, (*c*) crystallization and structure determination by X-ray crystallography or by NMR spectroscopy, and (*d*) archiving and annotation of the new structures. Given the current state of comparative modeling, a target sequence should have at least 30% sequence identity to a structural template. This corresponds to one experimentally determined structure per sequence family, rather than fold family. Because there are approximately five times more sequence families than fold families (70), it is likely that structural genomics will have to determine the structure of approximately 10,000 protein domains. Experimental structure determination of approximately 10,000 properly chosen proteins should result in useful 3D models for domains in hundreds of thousands of other protein sequences.

For comparative modeling to contribute to structural genomics, automation of all the steps in the modeling process is essential. There are at least five good reasons for automation. First, modeling of hundreds of thousands of protein sequences is obviously feasible only when it is completely automated. Second, automation makes it efficient for both the experts and non-experts to use comparative models, allowing them to spend more time designing experiments. Third, it is important that the best possible models are easily accessible to the non-experts. The results of the CASP meetings have shown that human expertise is usually a critical component of modeling success (92). However, modeling experts will not be on hand for every template selection or alignment question, just as structural biologists cannot solve every protein structure of biological interest. Fourth, automation encourages development of better methods. And finally, automated modeling removes any human bias, thus making the models more objective. Model evaluation by computational means is an essential component of large-scale modeling.

The automation of large-scale comparative modeling involves assembling a software pipeline that consists of modules for fold assignment, template selection, target–template alignment, model generation, and model evaluation. Computer programs for these individual operations already exist, and it may seem trivial to combine them. Yet, manual intervention is prevalent in setting many modeling parameters, especially during template selection and alignment. This enables an expert to implement various sources of information that can be difficult to exploit in an automated fashion. Thus, the primary challenge in large-scale comparative modeling is to build an automated, rapid, robust, sensitive, and accurate comparative modeling pipeline applicable to whole genomes; such a pipeline should perform at least as well as a human expert on individual proteins.

Two examples of large-scale comparative modeling for complete genomes have been described (132, 157). The sequences encoded in the *E. coli* genome have been used to build models for 10–15% of the proteins using the SWISS-MODEL web server (130, 132). Peitsch et al have also recently modeled many proteins in the SWISS-PROT database and made the models available on their web site (Table 1). The second large-scale modeling pipeline, MODPIPE, produced models for five procaryotic and eukaryotic genomes (157, 159). This calculation resulted in models for substantial segments of 17.2%, 18.1%, 19.2%, 20.4%, and 15.7% of all proteins in the genomes of *Saccharomyces cerevisiae* (6218 proteins in the genome); *Escherichia coli* (4290 proteins), *Mycoplasma genitalium* (468 proteins), *Caenorhabditis elegans* (7299 proteins, incomplete), and *Methanococcus janaschii* (1735 proteins), respectively. An important element in this study was the evaluation of the model's reliability. This is important because most of the related protein pairs share less than 30% sequence identity, resulting in significant errors in many models. The models were assigned to a reliable or unreliable class by a procedure (157) that depends on the statistical potential function from PROSAII (169). This allowed the identification of those models that were likely to be based on correct templates and approximately correct alignments. As a result, 236 yeast proteins lacking any prior structural information were assigned to a fold family; 40 of these proteins had no prior functional annotations. A more precise evaluation was used to calibrate the relationship between model accuracy and the percentage of sequence identity on which the model was based (157). Almost half of the 1071 reliably modeled proteins in the yeast genome share more than approximately 35% sequence identity with their templates. All the alignments, models, and model evaluations are available in the MODBASE database of comparative protein structure models (159, 210). Most recently, the MODPIPE pipeline software has been improved by using PSI-BLAST (6) for fold assignment, multiple templates and sequences for target–template alignment, and a complex statistical potential of mean force for model evaluation. This resulted in models for approximately 17,000 proteins, covering substantial segments of 18–45% of the proteins in 12 complete genomes (210).

Large-scale comparative modeling will extend opportunities to tackle a myriad of problems by providing many protein models for many genomes. A large database of experimental structures leveraged by comparative models will arouse questions about protein evolution, such as the physical origins of protein structure stability and protein activity, regulatory differences among similar enzymes, and the specificity and plasticity of ligand binding sites. Structural genomics will also aid in the process of drug design. A collection of experimentally determined complexes of proteins with their ligands, aligned with comparative models for the rest of the family members, will permit a facile comparison of ligand binding requirements and also reveal permitted substitutions in and around important residues. Structural genomics will provide an obvious resource for many questions and, hopefully, will provoke new ones.

A specific example of a new opportunity for tackling existing problems by virtue of providing many protein models from many genomes is the selection of a target protein for drug development. A protein that is likely to have high ligand specificity is a good choice; 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, from human or other organisms. Such comparisons may make it possible to rationally select a 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 a pathogen target 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 the particular protein in the pathway that has the binding site most dissimilar from its human homologs.

CONCLUSION

Over the past few years, there has been a gradual increase in both the accuracy of comparative models and the fraction of protein sequences that can be modeled with useful accuracy. The magnitude of errors in fold assignment, alignment, and the modeling of sidechains, loops, distortions, and rigid body shifts has decreased measurably. This is a consequence of both better techniques and a larger number of known protein sequences and structures. Nevertheless, all the errors remain significant and demand future methodological improvements. In addition, there is a great need for more accurate detection of errors in a given protein structure model. Error detection is useful both for refinement and interpretation of the models.

It is now possible to predict by comparative modeling significant segments of approximately one third of all known protein sequences. One half of these models are in the least accurate class, based on less than 30% sequence identity to known protein structures. The remaining 35 and 15% of the models are in the medium (<50% sequence identity) and high (>50% identity) accuracy classes. The fraction of protein sequences that can be modeled by comparative modeling is currently increasing by approximately 4% per year (158). It has been estimated that globular protein domains cluster in only a few thousand fold families, approximately 900 of which have already been structurally defined (70, 76). Assuming the current growth rate in the number of known protein structures, the structure of at least one member of most of the globular folds will be determined in less than 10 years (70). According to this argument, comparative modeling will be applicable to most of the globular protein domains soon after the expected completion of the human genome project. However, there are some classes of proteins, such as membrane proteins, that will not be amenable to modeling without improvements in structure determination and modeling techniques. To maximize the number of proteins that can be modeled reliably, a concerted effort toward structure determination of new folds by X-ray crystallography and nuclear magnetic resonance spectroscopy is in order, as envisioned by structural genomics (117, 147, 182, 202, 207–209). The full potential of the genome sequencing projects will only be realized once all protein functions are assigned and understood. This will be facilitated by integrating genomic sequence information with databases arising from functional and structural genomics. Comparative modeling will play an important bridging role in these efforts.

ACKNOWLEDGMENTS

We are grateful to Dr. Azat Badretdinov and Mr. Eric Feyfant for many discussions about comparative protein structure modeling. MAM-R and AF are Burroughs Wellcome Fellows. RS is a Howard Hughes Medical Institute predoctoral fellow. FM is a Norman and Rosita Winston Biomedical Research Foundation Fellow. AŠ is a Sinsheimer Scholar and an Alfred P. Sloan Research Fellow. The investigations have also been aided by grants from NIH (GM 54762) and NSF (BIR-9601845). This review is based on (55, 56, 145, 146, 154, 156, 158).

LITERATURE CITED

1. Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein data bank. In *Crystallographic Databases—Information, Content, Software Systems, Scientific Applications,* ed. FH Allen, G Bergerhoff, R Sievers, pp. 107–132. Bonn/Cambridge/Chester. Data Commission Int. Union of Crystallography.
2. Alexandrov NN, Nussinov R, Zimmer RM. 1995. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. In *Pacific Symposium on Biocomputing '96,* ed. L Hunter, TE Klein, pp. 53–72, Singapore: World Sci. Pub.
3. Aloy P, Mas JM, Martí-Renom MA, Querol E, Avilés FX, Oliva B. 2000. Human a2 pro-carboxypeptidase model: secondary structure prediction as a powerful tool for homology modelling improvement. *J. Computer-Aided Molec. Design.* 14:83–92
4. Altschul SF, Boguski MS, Gish W, Wootton JC. 1994. Issues in searching molecular sequence databases. *Nat. Genet.* 6:119–29
5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–10
6. Altschul SF, Madden TL, Schaffer AA, Zhang J Zhang, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–402
7. Apostolico A, Giancarlo R. 1998. Sequence alignment in molecular biology. *J. Comput. Biol.* 5:173–96
8. Aszódi A, Taylor WR. 1996. Homology modelling by distance geometry. *Folding Design* 1:325–34
9. Bairoch A, Apweiler R. 1999. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* 27:49–54
10. Bajorath J, Stenkamp R, Aruffo A. 1994. Knowledge-based model building of proteins: Concepts and examples. *Protein Sci.* 2:1798–810
11. Barton GJ. 1998. Protein sequence alignment and database scanning. In *Protein Structure Prediction: A Practical Approach,* ed. MJE Sternberg. Oxford, UK: Oxford Univ. Press
12. Barton GJ, Sternberg MJE. 1987. A strategy for the rapid multiple alignment of protein sequences; confidence levels from tertiary structure comparisons. *J. Mol. Biol.* 198:327–37
13. Bassolino-Klimas D, Bruccoleri RE. 1992. Application of a directed conformational search for generating 3-D coordinates for protein structures from α-carbon coordinates. *Proteins* 14:465–74
14. Baxevanis AD. 1998. Practical aspects of multiple sequence alignment. *Methods Biochem. Anal.* 39:172–88
15. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. 2000. GenBank. *Nucleic Acid Res.* 28:15–18
16. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–42
17. Blundell TL, Sibanda BL, Sternberg MJE, Thornton JM. 1987. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 326:347–52
18. 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.* 268:15983–93

19. Bower MJ, Cohen FE, Dunbrack RL Jr. 1997. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J. Mol. Biol.* 267:1268–82
20. Bowie JU, Lüthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253:164–70
21. Brenner SE, Barken D, Levitt M. 1999. The PRESAGE database for structural genomics. *Nucleic Acids Res.* 27:251–53
22. Brenner SE, Chothia C, Hubbard TJ. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA* 95:6073–78
23. Briffeuil P, Baudoux G, Lambert C, Bolle X De, Vinals C, Feytmans E, Depiereux E. 1998. Comparative analysis of seven multiple protein sequence alignment servers: clues to enhance reliability of predictions. *Bioinformatics* 14:357–66
24. Brocklehurst SM, Perham RN. 1993. Prediction of the three-dimensional structures of the biotinylated domain from yeast pyruvate carboxylase and of the lipolyated H-protein from the pea leaf glycine cleavage system: a new automated methods for the prediction of protein tertiary structure. *Protein Sci.* 2:626–39
25. Brooks BR, 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
26. Brower RC, Vasmatzis G, Silverman M, DeLisi C. 1993. Exhaustive conformational search and simulated annealing for models of lattice peptides. *Biopolymers* 33:329–34
27. Browne WJ, North ACT, Phillips DC, Brew K, Vanaman TC, Hill RC. 1969. A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol.* 42:65–86
28. Bruccoleri BR, Karplus M. 1990. Conformational sampling using high temperature molecular dynamics. *Biopolymers* 29:1847–62
29. Bruccoleri RE. 1993. Application of systematic conformational search to protein modeling. *Molec. Simulat.* 10:151–74
30. Bruccoleri RE, Karplus M. 1987. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* 26:137–68
31. Bryant SH, Amzel LM. 1987. Correctly folded proteins make twice as many hydrophobic contacts. *Int. J. Peptide Protein Res.* 29:46–52
32. Bryant SH, Lawrence CE. 1991. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: a statistical model for nonbonded interactions. *Proteins* 9:108–19
33. Carlacci L, Englander SW. 1996. Loop problem in proteins: developments on the Monte Carlo simulated annealing approach. *Comp. Chem.* 17:1002–12
34. Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5:823–26
35. Chothia C, Lesk AM. 1987. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* 196:901–17
36. Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR, Colman PM, Spinelli S, Alzari PM, Poljak RJ. 1989. Conformation of immunoglobulin hypervariable regions. *Nature* 342:877–83
37. Chung SY, Subbiah S. 1996. A structural explanation for the twilight zone of protein sequence homology. *Structure* 4:1123–27
38. Claessens M, Cutsem EV, Lasters I, Wodak S. 1989. Modelling the polypeptide back-

bone with ‘spare parts’ from known protein structures. *Protein Eng.* 4:335–45

39. Collura V, Higo J, Garnier J. 1993. Modeling of protein loops by simulated annealing. *Protein Sci.* 2:1502–10
40. Colovos C, Yeates TO. 1993. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.* 2:1511–19
41. Corpet F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* 16:10881–90
42. Cregut D, Liautard J-P, Chiche L. 1994. Homology modeling of annexin I: implicit solvation improves side-chain prediction and combination of evaluation criteria allows recognition of different types of conformational error. *Protein Eng.* 7:1333–44
43. Dudek MJ, Ramnarayan K, Ponder JW. 1998. Protein structure prediction using a combination of sequence homology and global energy minimization: II. Energy functions. *J. Comp. Chem.* 19:548–73
44. Dunbrack RL, Karplus M. 1993. Prediction of protein side-chain conformations from a backbone conformation dependent rotamer library. *J. Mol. Biol.* 230:543–71
45. Evans JS, Mathiowetz AM, Chan SI, Goddard WA III. 1995. De novo prediction of polypeptide conformations using dihedral probability grid Monte Carlo methodology. *Protein Sci.* 4:1203–16
46. Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–91
47. Fetrow JS, Godzik A, Skolnick J. 1998. Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* 282:703–11
48. Fidelis K, Stern PS, Bacon D, Moult J. 1994. Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng.* 7:953–60
49. Filippis V De, Sander C, Vriend G. 1994. Predicting local structural changes that result from point mutations. *Protein Eng.* 7:1203–8
50. Fine RM, Wang H, Shenkin PS, Yarmush DL, Levinthal C. 1986. Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCP603 from many randomly generated loop conformations. *Proteins* 1:342–62
51. Finkelstein AV, Reva BA. 1992. Search for the stable state of a short chain in a molecular field. *Protein Eng.* 5:617–24
52. Fischer D, Barret C, Bryson K, Elofsson A, Godzik A, Jones D, Karplus KJ, Kelley LA, MacCallum RM, Pawowski K, Rost B, Rychlewski L, Sternberg MJE. 1999. CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins.* Suppl. 3, pp. 209–17
53. Fischer D, Eisenberg D. 1996. Fold recognition using sequence-derived predictions. *Protein Sci.* 5:947–55
54. Fischer D, Eisenberg D. 1997. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc. Natl. Acad. Sci. USA* 94:11929–34
55. Fiser A, Do RKG, Šali A. 2000. Modeling of loops in protein structures. Submitted
56. Fiser A, Sánchez R, Melo F, Šali A. 2000. Comparative protien structure modeling. In *Computational Biochemistry and Biophysics*, ed. M Watanabe, B Roux, A MacKerell, O Backer. New York: Marcel Dekker. In press
57. Flockner H, Braxenthaler M, Lackner P, Jaritz M, Ortner M, Sippl MJ. 1995. Progress in fold recognition. *Proteins* 23:376–86
58. Gerstein M, Levitt M. 1997. A structural census of the current population of protein sequences. *Proc. Natl. Acad. Sci. USA* 94:11911–16
59. Godzik A, Kolinski A, Skolnick J. 1992. Topology fingerprint approach to the

inverse protein folding problem. *J. Mol. Biol.* 227:227–38

60. Greer J. 1980. Model for haptoglobin heavy chain based upon structural homology. *Proc. Natl. Acad. Sci. USA* 77:3393–97
61. Greer J. 1990. Comparative modelling methods: application to the family of the mammalian serine proteases. *Proteins* 7: 317–34
62. Gregoret LM, Cohen FE. 1991. Effect of packing density on chain conformation. *J. Mol. Biol.* 219:109–22
63. Gribskov M. 1994. Profile analysis. *Meth. Mol. Biol.* 25:247–66
64. Guenther B, Onrust R, Šali 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* 91:335–45
65. Harbury PB, Tidor B, Kim PS. 1995. Repacking proteins cores with backbone freedom: Structure prediction for coiled coils. *Proc. Natl. Acad. Sci. USA* 92:8408–12
66. 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
67. Henikoff S, Henikoff JG. 1994. Protein family classification based on searching a database of blocks. *Genomics* 19:97–107
68. Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S. 1995. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene* 163:17–26
69. Holm L, Sander C. 1991. Database algorithm for generating protein backbone and side-chain co-ordinates from C_α trace: application to model building and detection of co-ordinate errors. *J. Mol. Biol.* 218:183–94
70. Holm L, Sander C. 1996. Mapping the protein universe. *Science* 273:595–602
71. Holm L, Sander C. 1999. Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.* 27:244–47
72. Hooft RWW, Sander C, Vriend G. 1996. Verification of protein structures: side-chain planarity. *J. Appl. Crystallogr.* 29: 714–16
73. Hooft RWW, Vriend G, Sander C, Abola EE. 1996. Errors in protein structures. *Nature* 381:272
74. 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* 48:200–7
75. Huang ES, Koehl P, Levitt M, Pappu RV, Ponder JW. 1998. Accuracy of side-chain prediction upon near-native protein backbones generated by ab initio folding methods. *Proteins* 33:204–17
76. Hubbard TJP, Ailey B, Brenner SE, Murzin AG, Chothia C. 1999. SCOP: A structural classification of proteins database. *Nucleic Acids Res.* 27:254–56
77. Huynen M, Doerks T, Eisenhaber F, Orengo C, Sunyaev S, Yuan Y, Bork P. 1998. Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.* 280:323–26
78. Jeanmougin F, Thompson JD, Gouy M, Gibson DG, Higgins TJ. 1998. Multiple sequence alignment with CLUSTAL X. *Trends Biochem. Sci.* 23:403–5
79. Johnson MS, Overington JP. 1993. A structural basis for sequence comparisons: an evaluation of scoring methodologies. *J. Mol. Biol.* 233:716–38
80. Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL. 1994. Knowledge-based protein modelling. *CRC Crit. Rev. Biochem. Mol. Biol.* 29:1–68
81. Jones DT. 1999. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* 287:797–815
82. Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature* 358:86–89
83. Jones DT. 1997. Progress in protein struc-

ture prediction. *Curr. Opin. Struct. Biol.* 7:377–87

84. Jones S, Thornton JM. 1997. Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.* 272:133–43
85. Jones TH, Thirup S. 1986. Using known substructures in protein model building and crystallography. *EMBO J.* 5:819–22
86. Jung S-H, Pastan I, Lee B. 1994. Design of interchain disulfide bonds in the framework region of the Fv fragment of the monoclonal antibody B3. *Proteins* 19:35–47
87. Kick EK, Roe DC, Skillman AG, Liu G, Ewing TJ, Sun Y, Kuntz ID, Ellman JA. 1997. Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D. *Chem. Biol.* 4:297–307
88. Kidera A. 1995. Enhanced conformational sampling in Monte Carlo simulations of proteins: applications to a constrained peptide. *Proc. Natl. Acad. Sci. USA* 92:9886–89
89. Koehl P, Delarue M. 1994. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* 239:249–75
90. Koehl P, Delarue M. 1994. Polar and nonpolar atomic environments in the protein core: implication for folding and binding. *Proteins* 20:264–78
91. Koehl P, Delarue M. 1995. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in protein homology modelling. *Nat. Struct. Biol.* 2:163–70
92. Koehl P, Levitt M. 1999. A brighter future for protein structure prediction. *Nat. Struct. Biol.* 6:108–11
93. Koretke KK, Luthey-Schulten Z, Wolynes PG. 1996. Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. *Protein Sci.* 5:1043–59
94. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. 1994. Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* 235:1501–31
95. Lambert MH, Scheraga HA. 1989. Pattern recognition in the prediction of protein structure. I. Tripeptide conformational probabilities calculated from the amino acid sequence. *J. Comp. Chem.* 10:770–97
96. Laskowski RA, McArthur MW, Moss DS, Thornton JM. 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystalogr.* 26:283–91
97. Laskowski RA, MacArthur MW, Thornton JM. 1998. Validation of protein models derived from experiment. *Curr. Opin. Struct. Biol.* 5:631–39
98. Laskowski RA, Rullmann JAC, MacArthur MW, Kaptein R, Thornton JM. 1996. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* 8:477–86
99. Lazaridis T, Karplus M. 1999. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* 288:477–87
100. Lee C. 1995. Testing homology modeling on mutant proteins: predicting structural and thermodynamic effects in the Ala98 → Val mutants of T4 lysozyme. *Folding Design* 1:1–12
101. 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.* 136:225–70
102. Levitt M. 1992. Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* 226:507–33
103. Levitt M. 1997. Competitive assessment of protein fold recognition and alignment accuracy. *Proteins* (Suppl.) 1:92–104
104. Levitt M, Gerstein M. 1998. A unified statistical framework for sequence

comparison and structure comparison. *Proc. Natl. Acad. Sci. USA* 95:5913–20

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

106. MacKerell AD, Bashford D, Bellott M, Dunbrack RL Jr., Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux M., Smith JC, Stote J, Watanabe M., Wiorkiewicz-Kuczera J, Yin D, Karplus M. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* 102:3586–16

107. Martin ACR, Cheetham JC, Rees AR. 1989. Modeling antibody hypervariable loops: a combined algorithm. *Proc. Natl. Acad. Sci. USA* 86:9268–72

108. Martin ACR, MacArthur MW, Thornton JM. 1997. Assessment of comparative modeling in CASP2. *Proteins* (Suppl.) 1:14–28

109. Matsumoto R, Šali 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 in mouse mast cell protease-7 regulates its binding to heparin serglycin proteoglycan. *J. Biol. Chem.* 270:19524–31

110. Mattos C, Petsko GA, Karplus M. 1994. Analysis of two-residue turns in proteins. *J. Mol. Biol.* 238:733–47

111. McGarrah DB, Judson RS. 1993. Analysis of the genetic algorithm method of molecular conformation determination. *J. Comp. Chem.* 14:1385–95

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

113. Melo F, Feytmans E. 1998. Assessing protein structures with a non-local atomic interaction energy. *J. Mol. Biol.* 277:1141–52

114. Mezei M. 1998. Chameleon sequences in the PDB. *Protein Eng.* 11:411–14

115. Miwa JM, Ibanez-Tallon I, Crabtree GW, Sánchez R, Šali A, Role LW, Heintz N. 1999. lynx1, an endogenous toxin-like modulator of nicotinic acetylcholine receptors in the mammalian CNS. *Neuron* 23:105–14

116. Modi S, Paine MJ, Sutcliffe MJ, Lian L-Y, Primrose WU, Wolfe CR, Roberts GCK. 1996. A model for human cytochrome P_{450} 2D6 based on homology modeling and NMR studies of substrate binding. *Biochemistry* 35:4540–50

117. Montelione GT, Anderson S. 1999. Structural genomics: keystone for a human proteome project. *Nat. Struct. Biol.* 6:11–12

118. Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT. 1997. Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins* (Suppl.) 1:2–6

119. Moult J, James MNG. 1986. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins* 1:146–63

120. Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443–53

121. Oldfield TJ. 1992. Squid: a program for the analysis and display of data from crystallography and molecular dynamics. *J. Mol. Graphics* 10:247–52

122. Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJE. 1997. An automated classification of the structure of protein loops. *J. Mol. Biol.* 266:814–30

123. Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature* 372:631–34

124. Orengo CA, Pearl FMG, Bray JE, Todd AE, Martin AC, Conte L. Lo, Thornton JM. 1999. The CATH database provides insights into protein structure/function re-

lationship. *Nucleic Acids Res.* 27:275–79

125. Park J, Karplus K, Barret C, Hughey R, Haussler D, Hubbard T, Chothia C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* 284:1201–10
126. Pawlowski K, Bierzyński A, Godzik A. 1996. Structural diversity in a family of homologous proteins. *J. Mol. Biol.* 258:349–66
127. Pearson WR. 1990. Rapid and sensitive comparison with FASTA and FASTP. *Methods Enzymol.* 183:63–98
128. Pearson WR. 1995. Comparison of methods for searching protein sequence databases. *Protein Sci.* 4:1145–60
129. Pearson WR. 1998. Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* 276:71–84
130. Peitsch MC. 1996. PROMOD and SWISS-MODEL—Internet-based tools for automated comparative protein modeling. *Biochem. Soc. Trans* 24:274–79
131. Peitsch MC, Jongeneel CV. 1993. A 3-D model for the CD40 ligand predicts that it is a compact trimer similar to the tumor necrosis factors. *Int. Immunol.* 5:233–38
132. Peitsch MC, Wilkins MR, Tonella L, Sánchez JC, Appel RD, Hochstrasser DF. 1997. Large-scale protein modelling and integration with the SWISS-PROT and SWISS-2DPAGE databases: the example of *Escherichia coli*. *Electrophoresis* 18:498–501
133. Petrella RJ, Lazardis T, Karplus M. 1998. Protein side chain conformer prediction: a test of the energy function. *Folding Design* 3:353–77
134. Pontius J, Richelle J, Wodak SJ. 1996. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.* 264:121–36
135. Rapp CS, Friesner RA. 1999. Prediction of loop geometries using a generalized Born model of solvation effect. *Proteins* 35:173–83
136. Ring CS, Cohen FE. 1994. Conformational sampling of loop structures using genetic algorithm. *Isr. J. Chem.* 34:245–52
137. 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* 90:3583–87
138. Rosenbach D, Rosenfeld R. 1995. Simultaneous modeling of multiple loops in proteins. *Protein Sci.* 4:496–505
139. Rost B. 1995. Topits: Threading one-dimensional predictions into three-dimensional structures. In *The Third Int. Conf. on Intelligent Systems for Molecular Biology (ISMB)*, ed. C Rawlings, D Clark, R Altman, L Hunter, T Lengauer, S Wodak, pp. 314–21. Menlo Park, CA; AAAI Press
140. Rost B. 1999. Twilight zone of protein sequence alignments. *Protein Eng.* 12:85–94
141. Rost B, Sander C. 1993. Prediction of protein structure at better than 70% accuracy. *J. Mol. Biol.* 232:584–99
142. Rufino SD, Donate LE, Canard LHJ, Blundell TL. 1997. Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modeling. *J. Mol. Biol.* 267:352–67
143. Russell RB, Barton GJ. 1994. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J. Mol. Biol.* 244:332–50
144. Rychlewski L, Zhang B, Godzik A. 1998. Fold and function predictions for *Mycoplasma genitalium* proteins. *Folding Design* 3:229–38
145. Šali A. 1995. Modelling mutations and homologous proteins. *Curr. Opin. Biotech.* 6:437–51
146. Šali A. 1995. Protein modeling by

satisfaction of spatial restraints. *Mol. Med. Today* 1:270–77

147. Šali A. 1998. 100,000 protein structures for the biologist. *Nat. Struct. Biol.* 5: 1029–32

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

149. Šali 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.* 268:9023–34

150. Šali A, Overington JP. 1994. Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.* 3:1582–96

151. Šali A, Potterton L, Yuan F, Vlijmen H, Karplus M. 1995. Evaluation of comparative protein structure modeling by MODELLER. *Proteins* 23:318–26

152. Šali A, Sánchez R, Badretdinov AY, Fiser A, Melo F, Overington JP, Feyfant E, Martí-Renom MA. 1999. *MODELLER, A Protein Structure Modeling Program, Release 5.* URL http://guitar. rockefeller. edu/

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

154. Sánchez R, Badretdinov A. Ya, Feyfant E, Šali A. 1997. Homology protein structure modeling. *Trans. Am. Cryst. Assoc.* 32:81–91

155. Sánchez R, Šali A. 1997. Advances in comparative protein-structure modeling. *Curr. Opin. Struct. Biol.* 7:206–14

156. Sánchez R, Šali A. 1997. Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins* (Suppl.) 1:50–58

157. Sánchez R, Šali A. 1998. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA* 95:13597–602

158. Sánchez R, Šali A. 1999. Comparative protein structure modeling in genomics. *J. Comp. Phys.* 151:388–401

159. Sánchez R, Šali A. 1999. The MODBASE database of comparative protein structure models. *Bioinformatics.* In press

160. Saqi MAS, Russell RB, Sternberg MJE. 1999. Misleading local sequence alignments: implications for comparative protein modelling. *Protein Eng.* 11:627–30

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

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

163. Schuler GD. 1998. Sequence alignment and database searching. *Meth. Biochem. Anal.* 39:145–71

164. Sheng Y, Šali A, Herzog H, Lahnstein J, Krilis S. 1996. Modelling, expression and site-directed mutagenesis of human β_2-glycoprotein I. Identification of the major phospholipid binding site. *J. Immunol.* 157:3744–51

165. Shenkin PS, Yarmush DL, Fine RM, Wang H, Levinthal C. 1987. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ring-like structures. *Biopolymers* 26:2053–85

166. Shepherd AJ, Gorse D, Thornton JM. 1999. Prediction of the location and type of β-turns in proteins using neural networks. *Protein Sci.* 8:1045–55

167. Sibanda BL, Blundell TL, Thornton JM. 1989. Conformation of β-hairpins in protein structures: a systematic classification with applications to modelling by homology, electron deosity fitting and protein engineering. *J. Mol. Biol.* 206:759–77

168. 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.* 213:859–83
169. Sippl MJ. 1993. Recognition of errors in three-dimensional structures of proteins. *Proteins* 17:355–62
170. Smith RF, Wiese BA, Wojzynski MK, Davison DB, Worley KC. 1996. Bcm search launcher–an integrated interface to molecular biology data base search and analysis services available on the world wide web. *Genome Res.* 6:454–62
171. Smith TF. 1999. The art of matchmaking: sequence alignment methods and their structural implications. *Structure* 7:R7–R12
172. Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195–97
173. Smith TF, Conte L. Lo, Bienkowska J, Gaitatzes C, Rogers RGJ, Lathrop R. 1997. Current limitations to protein threading approaches. *J. Comput. Biol.* 4:217–25
174. Srinivasan N, Blundell TL. 1993. An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. *Protein Eng.* 6:501–12
175. Srinivasan S, March CJ, Sudarsanam S. 1993. An automated method for modeling proteins on known templates using distance geometry. *Protein Sci.* 2:227–89
176. Sternberg MJE, Bates PA, Kelley LA, MacCallum RM. 1999. Progress in protein structure prediction: assessment of CASP3. *Curr. Opin. Struct. Biol.* 9:368–73
177. Sudarsanam S, DuBose RF, March CJ, Srinivasan S. 1995. Modeling protein loops using a ϕ_{i+1}, ψ_i dimer database. *Protein Sci.* 4:1412–20
178. Summers NL, Karplus M. 1990. Modeling of globular proteins: a distance-based search procedure for the construction of insertion/deletion regions and pro → non-pro mutations. *J. Mol. Biol.* 216:991–1016
179. 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.* 1:377–84
180. Taylor WR. 1996. Multiple protein sequence alignment: algorithms and gap insertion. *Methods Enzymol.* 266:343–67
181. Taylor WR, Flores TP, Orengo CA. 1994. Multiple protein structure alignment. *Protein Sci.* 3:1858–70
182. Terwilliger TC, Waldo G, Peat TS, Newman JM, Chu K, Berendzen J. 1998. Class-directed structure determination: foundation for a protein structure initiative. *Protein Sci.* 7:1851–56
183. Thanki N, Zeelen JP, Mathieu M, Jaenicke R, Abagyan RA, Wierenga RK, Schliebs W. 1997. Protein engineering with monomeric triosephosphate isomerase (monoTIM): the modelling and structure verification of a seven-residue loop. *Protein Eng.* 10:159–67
184. Topham CM, Srinivasan N, Thorpe CJ, Overington JP, Kalsheker NA. 1994. Comparative modelling of major house dust mite allergen *der p* I: structure validation using an extended environmental amino acid propensity table. *Protein Eng.* 7:869–94
185. Torda AE. 1997. Perspectives in protein-fold recognition. *Curr. Opin. Struct. Biol.* 7:200–5
186. Totrov M, Abagyan R. 1994. Detailed *ab initio* prediction of lysozyme-antibody complex with 1. 6 Å accuracy. *Nat. Struct. Biol.* 1:259–63
187. Unger R, Harel D, Wherland S, Sussman JL. 1989. A 3-D building blocks approach to analyzing and predicting structure of proteins. *Proteins* 5:355–73

188. Vajda S, DeLisi C. 1990. Determining minimum energy conformations of polypeptides by dynamic programming. *Biopolymers* 29:1755–72

189. Vakser IA. 1997. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins* (Suppl.) 1:226–30

190. Gelder CWG, Leusen FJJ, Leunissen JAM, Noordik JH. 1994. A molecular dynamics approach for the generation of complete protein structures from limited coordinate data. *Proteins* 18:174–85

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

192. Vasmatzis G, Brower RC, DeLisi C. 1994. Predicting immunoglobulin-like hypervariable loops. *Biopolymers* 34:1669–80

193. Vásquez M. 1996. Modeling side-chain conformation. *Curr. Opin. Struct. Biol.* 6:217–21

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

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

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

197. Wilson KS, Dauter Z, Lamsin VS, Walsh M, Wodak S, Richelle J, Pontius J, Vaguine A, Sander RWW Hooft, Vriend G, Thornton JM, Laskowski RA, MacArthur MW, Dodson EJ, Murshudov G, Oldfield TJ, Kaptein RR, Rullman JAC. 1998. Who checks the checkers? Four validation tools applied to eight atomic resolution structures. *J. Mol. Biol.* 276:417–36

198. Wojcik J, Mornon J-P, Chomilier J. 1999. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J. Mol. Biol.* 289:1469–90

199. Wolf E, Vassilev A, Makino Y, Šali A, Nakatani Y, Burley SK. 1998. Crystal structure of a GCN5-related N-acetyltransferase: *Serratia marcescens* aminoglycoside 3-N-acetyltransferase. *Cell* 94:51–61

200. Wu G, Fiser A, Kuile B, Šali A, Müller M. 1999. Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. *Proc. Natl. Acad. Sci. USA* 96:6285–90

201. Xu LZ, Sánchez R, Šali A, Heintz N. 1996. Ligand specificity of brain lipid binding protein. *J. Biol. Chem.* 271:24711–19

202. Zarembinski TI, Hung LW, Mueller-Dieckmann HJ, Kim KK, Yokota H, Kim R, Kim SH. 1998. Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc. Nat. Acad. Sci. USA* 95:15189–93

203. Zhang B, Jaroszewski L, Rychlewski L, Godzik A. 1998. Similarities and differences between nonhomologous proteins with similar folds: evaluation of threading strategies. *Folding Design* 2:307–17

204. Zhang ZT. 1997. Relations of the numbers of protein sequences, families and folds. *Protein Eng.* 10:757–61

205. Zheng Q, Kyle DJ. 1996. Accuracy and reliability of the scaling-relaxation method for loop closure: an evaluation based on extensive and multiple copy conformational samplings. *Proteins* 24:209–17

206. Lund O, Frimand K, Gorodkiu J, Bohr H, Bohr J, Brunak S. 1997. Protein distance constraints predicted by neural networks and probability density functions. *Prot. Eng.* 10:1241–48

207. Šali A, Kuriyan J. 1999. Challenges at the frontiers of structural biology. *Trends Biochem. Sci.* 22:M20–M24

208. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Šali A, Studier FW, Swaminathan S. 1999. Structural genomics: beyond the Human Genome Project. *Nat. Genet.* 23:151–57

209. Cort JR, Koonin EV, Bash PA, Kennedy MA. 1999. A phylogenetic approach to target selection for structural genomics: solution structure of YciH. *Nucl. Acids Res.* 27:4018–27

210. Sánchez R, Pieper U, Mirković N, de Bakker PIW, Wittenstein E, Šali A, 2000. MODBASE a database of annotated comparative protein structure models. *Nucl. Acids Res.* 28:250–53

211. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–42

212. Jones TA, Kleywegt GJ. 1999. CASP3. Comperative modeling evoluation. *Proteins: Struct., Funct. & Gen.* Suppl. 3 pp. 30–46

213. Muller A. MacCallum RM, Sternberg MJ. 1999. Benchmarking PSI-BLAST in genome anotation *J. Mol. Biol.* 293:1257–41

Annual Review of Biophysics and Biomolecular Structure
Volume 29, 2000

CONTENTS