66

Comparative Protein Structure Modeling

M. S. Madhusudhan, Marc A. Marti-Renom, Narayanan Eswar, Bino John, Ursula Pieper, Rachel Karchin, Min-Yi Shen, and Andrej Sali

1. Introduction

Three-dimensional protein structures are invaluable sources of information for the functional annotation of protein molecules. These structures are best determined by experimental methods such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. However, the experimental methods cannot always be applied. In such cases, prediction of the protein structure by computational methods can frequently result in a useful model.

Protein structures can be modeled either *ab initio* from sequence alone or by comparative methods that rely on a database of known protein structures ***(1,2)***. *Ab initio* methods are largely based on the laws of physics, while comparative methods, including comparative (or homology) modeling and threading, are based primarily on statistical learning. Although there have been significant improvements in the *ab initio* ***(3)*** and threading methods ***(4)***, comparative modeling gives the most accurate results if a known protein structure that is sufficiently similar to the modeled sequence is available ***(1)***.

To predict protein structure by comparative modeling, two conditions have to be met ***(5,6)***. First, the sequence to be modeled (i.e., the target sequence) must have detectable similarity to another sequence of known structure (i.e., the template). Second, it must be possible to compute an accurate alignment between the target sequence and the template structure. The whole prediction process consists of fold assignment, target–template alignment, model building, and model evaluation (**Fig. 1**).

A simple predictor of the overall model accuracy is the degree of sequence similarity between the target and the template (**Fig. 2**). The higher is the sequence similarity to the template, the more accurate is the model. Although high-accuracy models are most informative, low-accuracy models may also provide coarse structural and functional annotation (**Fig. 3**) ***(1)***.

Comparative models can currently be built for domains in approx 57% of the approx 1.5 million protein sequences in the TrEMBL database ***(7)***. However, approximately two-thirds of the models are likely to contain significant errors because they are based on less than 30% sequence identity to the closest known protein structure. The primary sources of geometrical errors in the final models based on less than 30% sequence identity are the mistakes in the target–template alignment. Other errors include incor-

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

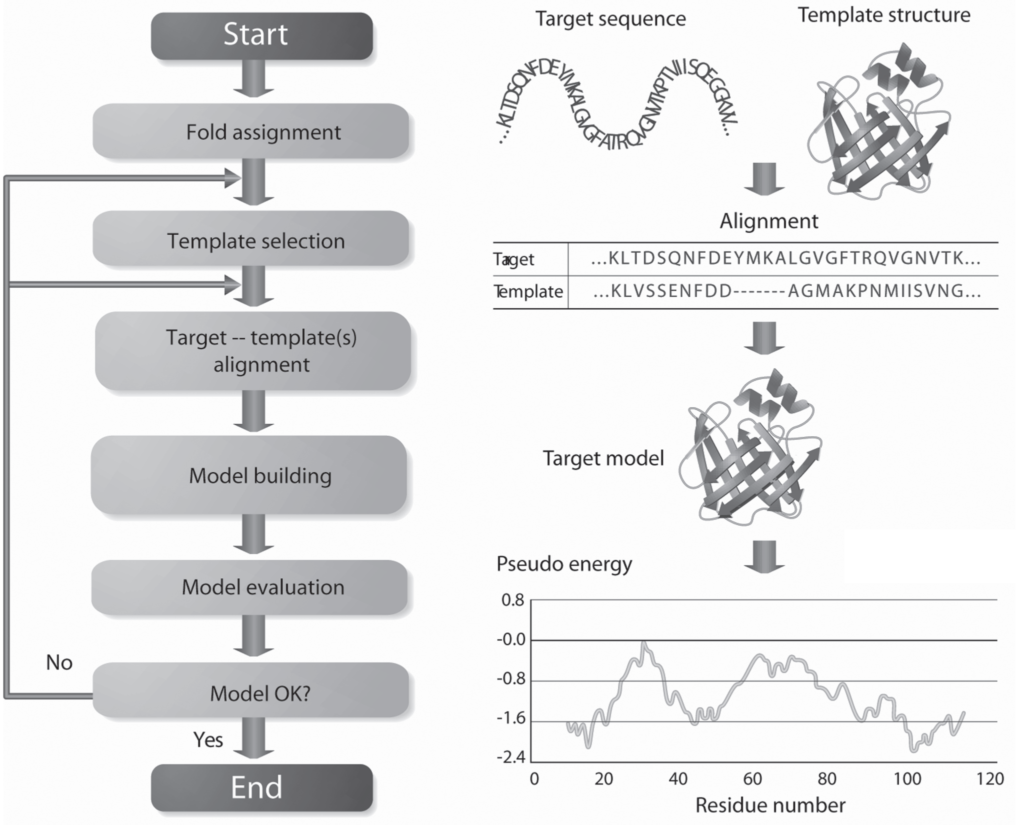

Fig. 1. A flow chart of the steps involved in comparative protein structure modeling.

rect fold assignments as well as incorrect modeling of loops, correctly aligned core segments, and side chains. No current modeling program can generally recover from an incorrect starting alignment. Therefore, one of the priorities for methods developers is to improve the accuracy of sequence-structure alignment and/or to minimize the dependence of the modeling methods on the input sequence-structure alignment.

The importance of comparative modeling derives partly from its role in structural genomics ***(8–10)***. Structural genomics aims to structurally characterize most protein sequences by an efficient combination of experiment and prediction ***(9,11–14)***. This aim will be achieved by careful selection of target proteins and their structure determination by X-ray crystallography or NMR spectroscopy. There are a variety of target selection schemes ***(15)***, ranging from focusing on only novel folds to selecting all proteins in a model genome. A model-centric view requires that targets be selected such that most of the remaining sequences can be modeled with useful accuracy by comparative modeling. Even with structural genomics, the structure of most of the proteins will be modeled, not determined by experiment. As mentioned above, the accuracy of comparative models and correspondingly the variety of their applications decrease sharply below the 30% sequence-identity cutoff, mainly as a result of a rapid increase in alignment errors. Thus, structural genomics should determine protein structures such that most of the remaining sequences are related to at least one known structure at

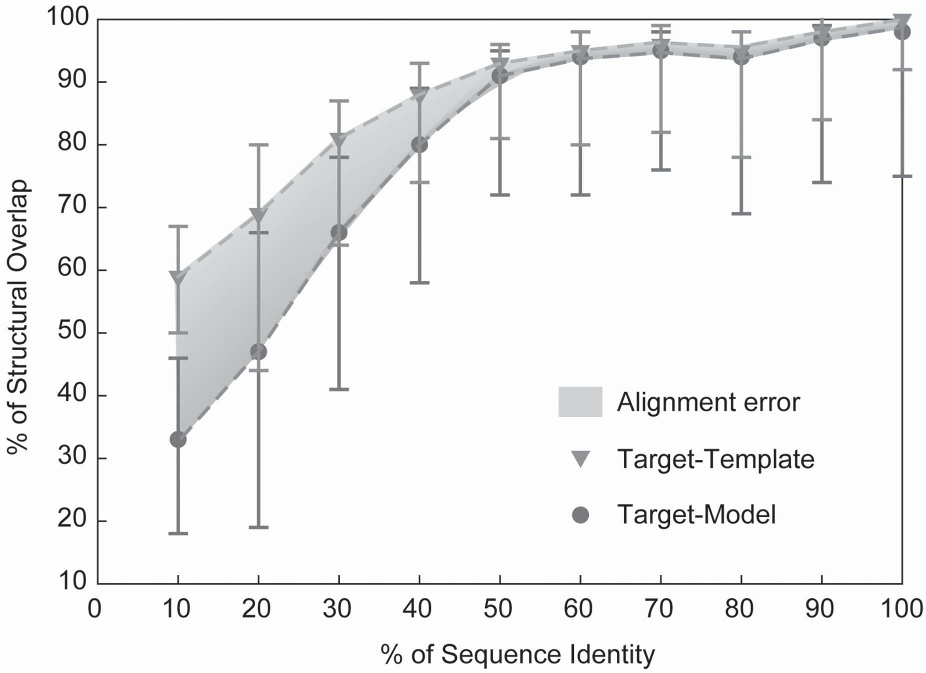

Fig. 2. Average accuracy of models calculated by ModPipe ***(218)*** with respect to the percentage sequence identity to the template. The average overlap of the experimentally determined protein structure with its calculated model (lower dashed line) and with the template on which the model was based (upper dashed line) are shown as a function of the target–template sequence identity. This sequence identity is calculated from the modeling alignment. The structure overlap is defined as the fraction of the equivalent C_α atoms after rigid superimposition of the two structures. Two C_α atoms are considered equivalent if they are within 3.5Å of each other. The points in the curves correspond to the median values, and the error bars in the positive and negative directions correspond to the average positive and negative differences from the median, respectively. The shaded area between the two curves corresponds approximately to the model error that arises from the alignment error.

higher than 30% sequence identity ***(15,16)***. It was estimated that this cutoff requires a minimum of 16,000 targets to cover 90% of all protein domain families, including those of membrane proteins ***(16)***. These 16,000 structures will allow the modeling of a very much larger number of proteins. For example, New York Structural Genomics Research Consortium measured the impact of its structures by documenting the number and accuracy of the corresponding models for detectably related proteins in the non-redundant sequence database. For each new structure, on the average approx 100 protein sequences without any prior structural characterization could be modeled at least at the fold level (http://www.nysgxrc.org/). This large leverage of structure determination by protein structure modeling illustrates and justifies the premise of structural genomics.

This chapter describes methods and computer programs used in all the steps of comparative modeling (**Table 1**). We conclude by reviewing several sample applications of the models.

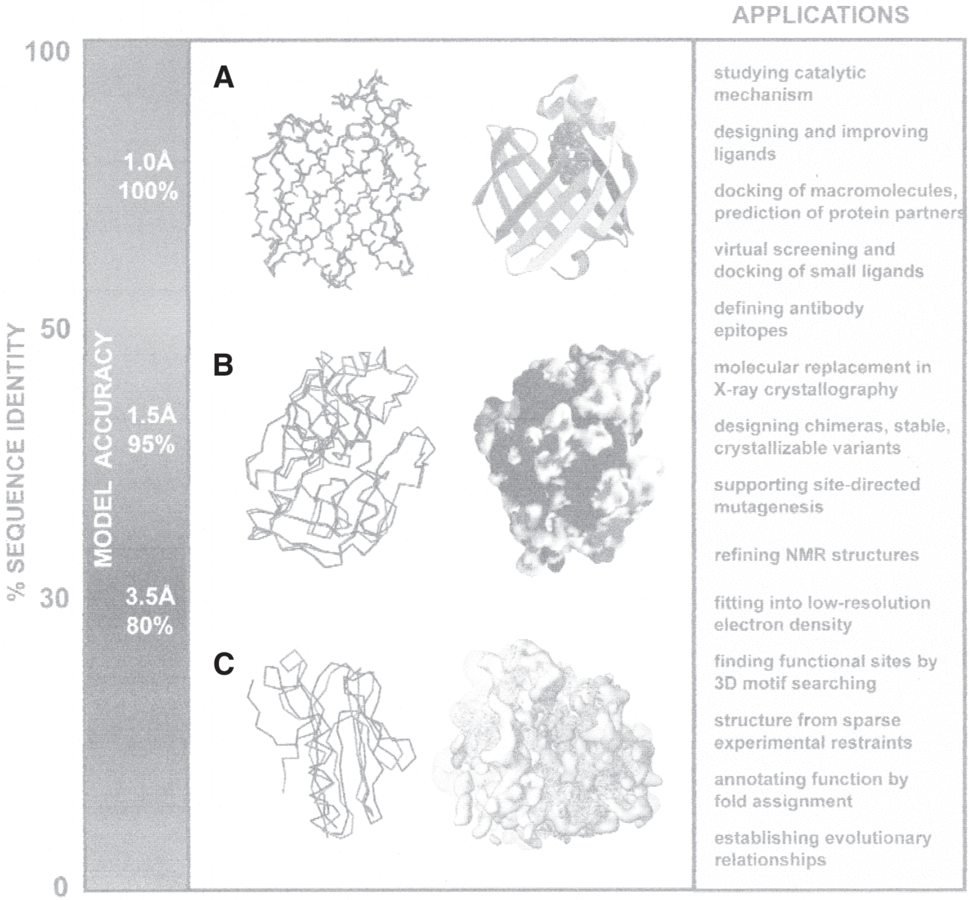
APPLICATIONS
100
50
30
0
% SEQUENCE IDENTITY
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
A
B
C
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screening and docking of small ligands
defining antibody epitopes
molecular replacement in X-ray crystallography
designing chimeras, stable, crystallizable variants
supporting site-directed mutagenesis
refining NMR structures
fitting into low-resolution electron density
finding functional sites by 3D motif searching
structure from sparse experimental restraints
annotating function by fold assignment
establishing evolutionary relationships

Fig. 3. Accuracy and applications of protein structure models. The vertical axis indicates the different ranges of applicability of comparative protein structure modeling, the corresponding accuracy of protein structure models, and their sample applications. **(A)** The docosahexaenoic fatty acid ligand was docked into a high accuracy comparative model of brain lipid-binding protein (right), modeled based on its 62% sequence identity to the crystallographic structure of adipocyte lipid-binding protein (PDB code *1adl*). A number of fatty acids were ranked for their affinity to brain lipid-binding protein consistently with site-directed mutagenesis and affinity chromatography experiments ***(194)***, even though the ligand specificity profile of this protein is different from that of the template structure (left). **(B)** A putative proteoglycan binding patch was identified on a medium accuracy comparative model of mouse mast cell protease 7 (right), modeled based on its 39% sequence identity to the crystallographic structure of bovine pancreatic trypsin (*2ptn*) that does not bind proteoglycans. The prediction was confirmed by site-directed mutagenesis and heparin-affinity chromatography experiments ***(193)***. Typical accuracy of a comparative model in this range of sequence similarity is indicated by a comparison of a trypsin model with the actual structure. **(C)** A molecular model of the whole yeast ribosome (right) was calculated by fitting atomic rRNA and protein models into the electron density of the 80S ribosomal particle, obtained by electron microscopy at 15Å resolution ***(229)***. Most of the models for 40 out of the 75 ribosomal proteins were based on template structures that were approx 30% sequentially identical. Typical accuracy of a comparative model in this range of sequence similarity is indicated by a comparison of a model for a domain in L2 Protein from *Bacillus Stearothermophilus* with the actual structure (*1rl2*).

Table 1
Programs and Web Servers Useful in Comparative Protein Structure Modeling

Name	World-Wide Web address[b]	***Reference***[c]
	Databases	
BALIBASE	http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE/	***196***
CATH	http://www.biochem.ucl.ac.uk/bsm/cath/	***197***
GENBANK	http://www.ncbi.nlm.nih.gov/Genbank/	***198***
GENECENSUS	http://bioinfo.mbb.yale.edu/genome/	***199***
MODBASE	http://www.salilab.org/modbase/	***7***
PDB	http://www.pdb.org	***200***
PRESAGE	http://presage.berkeley.edu	***201***
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/	***202***
SWISSPROT-TREMBL	http://www.expasy.org	***203***
	Template search	
123D	http://123d.ncifcrf.gov/	***204***
3D PSSM	http://www.sbg.bio.ic.ac.uk/~3dpssm	***77***
BLAST	http://www.ncbi.nlm.nih.gov/BLAST/	***22***
DALI	http://www2.ebi.ac.uk/dali/	***19***
FASTA	http://www.ebi.ac.uk/fasta33/	***23***
MATCHMAKER	http://bioinformatics.burnham-inst.org	***205***
PREDICTPROTEIN	http://cubic.bioc.columbia.edu/predictprotein/	***206***
PROFIT	http://www.bioinfo.org.uk/software	***207***
THREADER	http://bioinf.cs.ucl.ac.uk/threader/threader.html	***70***
UCLA-DOE FOLD SERVER	http://fold.doe-mbi.ucla.edu	***208***
SUPERFAMILY	http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/	***209***
	Target–template alignment	
BCM SERVERF	http://searchlauncher.bcm.tmc.edu	***210***
BLAST2	http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html	***211***
BLOCK MAKERF	http://blocks.fhcrc.org/	***212***
CLUSTALW	http://www2.ebi.ac.uk/clustalw/	***62***
MULTALIN	http://prodes.toulouse.inra.fr/multalin/	***213***
SEA	http://ffas.ljcrf.edu/sea/	***214***
FFAS03	http://ffas.ljcrf.edu/	***26***,***64***
SAM-T02	http://www.soe.ucsc.edu/research/compbio/HMM-apps/	***215***
FUGUE	http://www-cryst.bioc.cam.ac.uk/fugue	***75***
TCOFFEE	http://www.ch.embnet.org/software/TCoffee.html	***216***
COMPASS	ftp://iole.swmed.edu/pub/compass/	***27***
MUSCLE	http://www.drive5.com/muscle	***217***
SALIGN	http://www.salilab.org/modeller	***218***
USC SEQALN	http://www-hto.usc.edu/software/seqaln	***219***
	Modeling	
COMPOSER	http://www.tripos.com/sciTech/inSilicoDisc/	***87***
CONGEN	http://www.congenomics.com/congen/congen_toc.html	***94***
ICM	http://www.molsoft.com/bioinfomatics/	[a]***220***
DISCOVERY STUDIO	http://www.accelrys.com/composer.html	[b]

(continued)

Table 1 (*Continued*)
Programs and Web Servers Useful in Comparative Protein Structure Modeling

Name	World-Wide Web address[b]	*Reference*[c]
MODELLER	http://www.salilab.org/modeller/	***101***
SYBYL	http://www.tripos.com	*c*
SCWRL	http://dunbrack.fccc.edu/SCWRL3.php	***157***
SNPWEB	http://salilab.org/snpweb	***218***
SWISS-MODEL	http://www.expasy.org/swissmod	***221***
WHAT IF	http://www.cmbi.kun.nl/whatif/	***222***
	Model evaluation	
ANOLEA	http://protein.bio.puc.cl/cardex/servers/	***188***
AQUA	http://nmr.chem.uu.nl/users/jurgen/Aqua/server	***184***
BIOTECH	http://biotech.embl-heidelberg.de:8400	***183***
ERRAT	http://www.doe-mbi.ucla.edu/Services/ERRAT/	***223***
PROCHECK	http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html	***178***
PROSAII	http://www.came.sbg.ac.at	***181***
PROVE	http://www.ucmb.ulb.ac.be/UCMB/PROVE	***224***
SQUID	http://www.ysbl.york.ac.uk/~oldfield/squid/	***185***
VERIFY3D	http://www.doe-mbi.ucla.edu/Services/Verify_3D/	***74***
WHATCHECK	http://www.cmbi.kun.nl/gv/whatcheck/	***225***
	Methods evaluation	
CASP	http://predictioncenter.llnl.gov	***226***
CAFASP	http://bioinfo.pl/cafasp.html	***170***
EVA	http://cubic.bioc.columbia.edu/eva/	***173***
LIVEBENCH	http://bioinfo.pl/LiveBench/	***171***
CASA	http://capb.dbi.udel.edu/casa	***227***
AMAS	http://www.compbio.dundee.ac.uk/	***228***

Some of the sites are mirrored on additional computers.
[a]MolSoft Inc., San Diego.
[b]Accelrys Inc., San Diego.
[c]Tripos Inc., St. Louis.
The BIOTECH server uses PROCHECK and WHATCHECK for structure evaluation.

2. Steps in Comparative Modeling

Comparative modeling consists of four sequential steps: fold assignment, target–template alignment, model building, and model assessment (**Fig. 1**). If an assessment of the model is not positive, the model can be rebuilt by selecting different templates, refining the target–template alignment, or changing model-building parameters. The sections below deal with each one of the four steps in the modeling protocol.

2.1. Fold Assignment and Template Selection

The initial step in comparative modeling is to assign the likely fold of the target sequence. Template identification can be achieved using any one of the many programs that scan sequence and structure databases, such as Protein Data Bank (PDB) ***(17)***, structural classifcation of proteins (SCOP) ***(18)***, distance-matrix alignment

(DALI) ***(19)***, and Class, Architecture, Topology, and Homology (CATH) ***(20,21)*** (**Table 1**). Template search methods can be categorized into three different classes:

First, pairwise comparison methods, which include the popular programs Basic Local Alignment Search Tool (BLAST) ***(22)*** and FASTA ***(23)***, align the target sequence with all the sequences in the database of known structures. The performance and efficiency of this class of methods has been studied extensively ***(24)***. Second, sequence profile methods, such as position specific iterative (PSI)-BLAST ***(25)*** and HMMER (http://hmmer.wustl.edu), rely on profiles derived from multiple sequence alignments to increase the sensitivity and accuracy of the template search. The profile enhances the sensitivity of the search ***(26–29)***. Profiles are also utilized by the intermediate sequence search algorithms that establish a homology between two remotely related sequences through an intermediary sequence ***(30–36)***. Third, the so-called threading methods use a combination of sequence and structure considerations to detect similarities between sequences and structures ***(37–41)***. In these methods, the target sequence is threaded through a library of 3-D profiles or folds, and each threading is assessed based on a certain scoring function. Commonly used methods and servers in this category include Superfamily ***(42)*** and GenThreader ***(43)***. The threading methods are more effective in detecting homology at low sequence similarity than the methods relying on sequence information alone ***(44)***.

The three different classes of methods are best suited for identifying templates in different regimes of the sequence-identity spectrum. The pairwise sequence comparison methods are the least sensitive and are best used to detect close homologs. The profile-based methods are usually capable of recognizing homologs sharing only approx 25% sequence identity. Threading methods can sometimes recognize common folds even in the absence of any statistically significant sequence similarity. Because most of the fold assignment methods involve sequence alignment, some of them are discussed in more detail in the following section about sequence-structure alignment.

While a correct fold assignment can be used to build a useful model, an incorrect fold assignment renders the resulting model useless. Thus, when using a fold-recognition method, it is crucial to be aware of the accuracy of the method. In an assessment of different fold-recognition methods, the best method detected 75% of the closest structures correctly for a set of sequences related at the "family" level in the SCOP database ***(18)***. However, at the superfamily and fold levels, the accuracy dropped to 29 and 15%, respectively ***(44)***.

Once a list of all related protein structures is obtained, templates that are appropriate for the given modeling problem have to be selected. Usually, a higher overall sequence identity between the target and the template sequence yields a better template. Several other factors should also be considered in selecting templates.

Constructing a phylogenetic tree for the whole family can frequently help in selecting a template from the subfamily that is closest in structure to the target sequence. Databases of structure-based phylogenies, such as the database of Phylogeny and Alignment (PALI) ***(45)***, are useful in making a distinction between the sequence and structure similarity, which can be a key consideration for template identification.

Accuracy of the 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 structure accuracy.

It is also crucial to compare the environment of the template 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, pH, ligands, and quaternary interactions). For example, if the objective of the model-building exercise is to dock ligands in the model, it is usually best to use a template that is itself bound to an identical or similar ligand. In general, prior biological information about the target sequence can be valuable in identifying an appropriate template ***(46,47)***.

Prioritization 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 a 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 ***(48,49)***.

2.2. Target–Template Alignment

After identifying the template(s), the next crucial step in comparative modeling is to accurately align the target sequence to the template(s). Although most template-recognition methods produce a target–template alignment, there is frequently a need to use a specialized alignment method to realign the sequences because the template-identification step is often optimized to identify distant relationships, sometimes at the expense of alignment accuracy. The sequence-structure alignment is a vital step in the model-building process, and an erroneous alignment will almost certainly lead to the construction of an incorrect model.

An alignment between two sequences of residues is usually calculated by optimizing an alignment scoring function ***(50)***. The two common ingredients of the scoring function are a gap penalty function and a matrix of substitution scores for matching every residue in one sequence to every residue in the other sequence. The alignment score is usually a sum of the gap penalties, which depend linearly on the gap lengths, and the pairwise substitution scores, which depend on the matched residue types. The original and still widely used optimization method for sequence alignment is based on dynamic programming ***(51–53)***. Since its inception, the scoring function and its optimization by dynamic programming have been improved for alignment accuracy and speed, and applied to a variety of alignment problems.

In the next few paragraphs, we examine different methods to obtain substitution score matrices and gap penalties that optimize the accuracy of the output alignments. We examine the use of information from related multiple sequences and structures to enhance alignment accuracy and coverage, especially when target–template sequence identity decreases below 30%.

2.2.1. Using Multiple Sequence Information

The accuracy of a pairwise alignment method that uses dynamic programming greatly depends on the matrix of substitution scores and the gap penalties. Matrices with values for each of the possible residue type substitutions, such as Block Substitution Matrix (BLOSUM) ***(54)*** and point accepted mutation (PAM) ***(55)***, are useful only when sequence similarity is readily recognizable (e.g., above 30% sequence identity).

To increase the accuracy of the alignment between more divergent sequences, some methods construct the substitution scores by relying on substitution patterns revealed in a multiple sequence alignment (MSA) of many members of the corresponding protein family. A multiple sequence alignment is converted into a sequence profile that lists the likelihood of the 20 standard amino acid residue types at every position in a given MSA. Alignments based on sequence profiles rather than single sequences have been shown to be significantly more accurate ***(56–58)*** (**Table 2**). This improvement is reflected in the accuracy and extent of the resulting homology models.

Two popular profile alignment methods are PSI-BLAST ***(25)*** and SAM-T98 ***(59)***. Both methods take a single sequence as input and produce a sequence profile or a hidden Markov model (HMM) as output. PSI-BLAST relies on the BLAST algorithm ***(22)*** to collect homologs of a query sequence and to construct its profile by iteratively scanning a sequence database ***(25,32)***. SAM-T98 first uses BLAST to prefilter a large sequence database. It then constructs a multiple alignment and a HMM in parallel through several rounds of database searching and HMM building. The HMM is derived only from the sequences that score better than a specified threshold.

The latest generation of alignment methods extends sequence profile or MSA building to both sequences of interest, and aligns the two profiles or MSAs, rather than the individual sequences. These methods have been shown to be more sensitive than sequence-profile methods ***(26–28,60)***. The CLUSTALW program compares two MSAs by using a substitution matrix for all pairs of positions from the two alignments ***(61,62)***; each single value in this matrix is an average of residue-residue substitution scores over two matched alignment positions. The LAMA program aligns two MSAs by first transforming them into position-specific scoring matrices (PSSMs) and then comparing the two PSSMs to each other by the Pearson correlation coefficient ***(63)***. The FFAS program aligns two sequence profiles with each other using a dot product ***(26,64)***. A related approach, using mutual entropy, has been used by Yona and Levitt ***(65,66)*** to construct the ProtoMap database of protein sequence families ***(66–68)***. Most recently, the COMPASS program was developed to locally align two MSAs with assessment of statistical significance ***(27)***. The SALIGN command in the program MODELLER constructs a scoring matrix by comparing two profiles with mutual entropy and correlation coefficient measures ***(60)***. These methods compare two profiles by matching every position in one profile to each position in the other profile, followed by either local or global dynamic programming to calculate the optimal alignment.

2.2.2. Using Structural Information

Alignment accuracy can be significantly improved by incorporating information about protein structure. Threading and 3-D template-matching methods consider protein structure information for one of the sequences in a pairwise comparison ***(69–71)***. For a review of this class of methods, see ***(38–41,72)***. A combination of threading and sequence alignment scoring functions can also be used ***(43,73)***.

Another approach is to incorporate structural information into profile methods, by making substitution scores dependent on solvent exposure, secondary structure type, hydrogen bonding properties, and so on ***(74)***. Some methods in this category are FUGUE ***(75)***, 3D-PSSM ***(76,77)***, and SAM-T02 multitrack HMMs ***(78)***. These methods lie between traditional sequence-based algorithms and threading methods. The use of structural data is not restricted to the structure side of the aligned sequence-structure

Table 2
Different Programs for Aligning Two Protein Sequences, or a Protein Sequence and a Structure, Tested on a Benchmark of 200 Pairs of Related Known Structures*

		Average alignment accuracy					
Method	Type	1Å	2Å	3Å	4Å	5A	Average
CE	structure/structure	18.81	49.09	68.02	78.77	84.54	59.85
BLAST	sequence/sequence	7.60	17.07	22.72	26.41	29.29	20.62
ALIGN	sequence/sequence	6.86	18.11	27.19	34.79	41.44	25.68
PSI-BLAST	sequence/profile	9.07	23.50	33.16	40.28	45.63	30.33
SAM	sequence/profile	7.76	21.60	31.40	38.72	45.26	28.95
LOBSTER	sequence/profile	8.81	23.32	33.82	41.51	48.17	31.13
SEA	profile/profile	9.02	24.15	34.90	43.27	50.43	32.36
CLUSTALW	profile/profile	7.41	19.31	28.02	35.36	41.87	26.40
COMPASS	profile/profile	10.37	26.06	36.08	42.35	46.65	32.30
SALIGN	profile/profile	9.63	27.05	39.81	49.64	57.55	36.74

*An alignment is assessed here by a degree of structure similarity that it implies. This criterion was calculated by first superposing the two compared structures according to the tested alignment, and then calculating the percentage of the C_α positions that were within the specified cutoff of 1, 2, 3, 4, or 5 Å; in addition, the average of these percentages at all cutoffs was also calculated. For comparison, the actual structure similarity calculated from the structure-based alignments produced by the CE program is also given in the first row.

pair. For example, SAM-T02 and HMAP ***(79)*** make use of the predicted local structure to enhance homolog detection and alignment accuracy.

To improve the alignment accuracy, gap penalties can be adjusted according to the local environment in which they occur ***(80)***. For example, the SALIGN command in MODELLER ***(81)*** scales the gap insertion penalty depending on the structural environment of the gap; the cost of opening a gap in a region of regular secondary structure is greater than opening a gap in a random coil region. The SALIGN command of MODELLER ***(82)*** can also use structure-dependent gap penalties in conjunction with a sequence profile, similar to FUGUE ***(75)***.

Even when algorithms are enriched with structural and multiple sequence information, it remains difficult to align distantly homologous proteins in the "twilight zone" of sequence identity below 30% sequence identity ***(83)***. In a comparative modeling setting, the MOULDER algorithm ***(84)*** uses an iterative approach to build better alignments between distant homologs. The method relies on a genetic algorithm to iteratively (1) build target–template alignments; (2) build structural models based on the alignments; (3) assess the models; and (4) select the alignments that produce the best models as seeds to generate further alignments ***(84)***. The method was shown to improve significantly the alignment accuracy of alignments that fell within the twilight zone.

2.3. Model Building

The target–template alignment, maps the sequence of the target on the template structure. This mapping is utilized in constructing the 3-D model of the target protein. There are several methods of constructing the model, and some of these approaches are reviewed below. The various model-building procedures lead to the construction of models of similar accuracy when used optimally ***(85)***. In addition to the different schemes for building whole models, this review also examines techniques for constructing inserted loop segments of the target that have no corresponding template and for packing the side chains on a given backbone scaffold.

2.3.1. Modeling by Assembly of Rigid Bodies

The first and still widely used approach in comparative modeling is to assemble a model from a small number of rigid bodies obtained from the aligned protein structures ***(6,86)***. The approach is based on the natural dissection of the protein structure into conserved core regions, variable loops that connect them, and side chains that decorate the backbone. For example, the following semiautomated procedure is implemented in the computer program COMPOSER ***(87)***. First, the template structures are selected and superposed. Second, the "framework" is calculated by averaging the coordinates of the C_α atoms of structurally conserved regions in the template structures. Third, the main-chain atoms of each core region in the target model are obtained by superposing on the framework the core segment from the template whose sequence is closest to the target. Fourth, the loops are generated by scanning a database of all known protein structures to identify the structurally variable regions that fit the anchor core regions and have a compatible sequence ***(88)***. 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 ***(87)***. 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 when more than one template structure is used to construct the framework and when the templates are averaged into the framework using weights corresponding to their sequence similarities to the target sequence ***(48)***. Possible future improvements of modeling by rigid-body assembly include incorporation of rigid-body shifts, such as the relative shifts in the packing of α-helices and β-sheets ***(89)***.

2.3.2. Modeling by Segment Matching or Coordinate Reconstruction

The basis of modeling by coordinate reconstruction is the finding that most hexapeptide segments of protein structure can be clustered into only 100 structurally different classes ***(90,91)***. Thus, comparative models can be constructed by using a subset of atomic positions from template structures as "guiding" positions, and by identifying and assembling short, all-atom segments that fit these guiding positions. The guiding positions usually correspond to the C_α atoms of the segments that are conserved in the alignment between the template structure and the target sequence. The all-atom segments that fit the guiding positions can be obtained either by scanning all the known protein structures, including those that are not related to the sequence being modeled ***(92,93)***, or by a conformational search restrained by an energy function ***(94,95)***. 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 the representative database of all known protein structures ***(96)***. This method can construct both main-chain and side-chain atoms, and can also model unaligned regions (gaps). It is implemented in the program SegMod. Even some side-chain modeling methods ***(97)*** and the class of loop construction methods based on finding suitable fragments in the database of known structures ***(98)*** can be seen as segment-matching or coordinate-reconstruction methods.

2.3.3. Modeling by Satisfaction of Spatial Restraints

The methods in this class begin by generating many constraints or restraints on the structure of the target sequence, using its alignment to related protein structures as a guide. The procedure is conceptually similar to that used in determination of protein structures from NMR-derived restraints. The restraints are generally obtained by assuming that the corresponding distances between aligned residues in the template and the target structures are similar. These homology-derived restraints are usually supplemented by stereochemical restraints on bond lengths, bond angles, dihedral angles, and nonbonded atom–atom contacts that are obtained from a molecular-mechanics force field. The model is then derived by minimizing the violations of all the restraints. This optimization can be achieved either by distance geometry or real-space optimization. For example, an elegant distance-geometry approach constructs all-atom models from lower and upper bounds on distances and dihedral angles ***(99)***.

We now describe our own approach to comparative modeling by satisfaction of special restrains in more detail ***(100–103)***. The approach was developed to use as many different types of data about the target sequence as possible. It is implemented in the computer program MODELLER ***(101)***. The comparative modeling procedure begins with an alignment of the target sequence with related known 3-D structures. The output, obtained without any user intervention, is a 3-D model for the target sequence containing all main-chain and side-chain nonhydrogen atoms.

In the first step of model building, distance and dihedral angle restraints on the target sequence are derived from its alignment with template 3-D structures. The form of these restraints was obtained from a statistical analysis of the relationships between similar protein structures. The analysis relied on a database of 105 family alignments that included 416 proteins of known 3-D structure ***(103)***. By scanning the database of alignments, tables quantifying various correlations were obtained, such as the correlations between two equivalent C_α-C_α distances, or between equivalent main-chain dihedral angles from two related proteins ***(101)***. These relationships are expressed as conditional probability density functions (PDFs) and can be used directly as spatial restraints. For example, probabilities for different values of the main-chain dihedral angles are calculated from the type of a residue considered, from main-chain conformation of an equivalent residue, and from sequence similarity between the two proteins. Another example is the PDF for a certain C_α-C_α distance given equivalent distances in two related protein structures. An important feature of the method is that the forms of spatial restraints were obtained empirically, from a database of protein structure alignments.

In the second step, the spatial restraints and the CHARMM22 force-field terms enforcing proper stereochemistry ***(104,105)*** are combined into an objective function. The general form of the objective function is similar to that in molecular dynamics programs, such as CHARMM22 ***(105)***. The objective function depends on the Cartesian coordinates of approx 10,000 atoms (3-D points) that form the modeled molecules. For a 10,000-atom system, there can be on the order of 200,000 restraints. The functional form of each term is simple; it includes a quadratic function, harmonic lower and upper bounds, cosine, a weighted sum of a few Gaussian functions, Coulomb's law, Lennard–Jones potential, and cubic splines. The geometric features presently include a distance; an angle; a dihedral angle; a pair of dihedral angles between two, three, four atoms and eight atoms, respectively; the shortest distance in the set of distances; solvent accessibility in Å$^{2;}$ and atom density, expressed as the number of atoms around the central atom. Some restraints can be used to restrain pseudo-atoms such as the gravity center of several atoms.

Finally, the model is obtained by optimizing the objective function in Cartesian space. The optimization is carried out by the use of the variable target function method ***(106)*** employing methods of conjugate gradients and molecular dynamics with simulated annealing ***(107)***. Several slightly different models can be calculated by varying the initial structure, and the variability among these models can be used to estimate the lower bound on the errors in the corresponding regions of the fold.

Because the modeling by satisfaction of spatial restraints can use many different types of information about the target sequence, it is perhaps the most promising of all comparative modeling techniques. One of the strengths of modeling by satisfaction of spatial restraints is that constraints or restraints derived from a number of different sources can easily be added to the homology-derived restraints. For example, restraints could be provided by rules for secondary structure packing ***(108)***, analyses of hydrophobicity ***(109)*** and correlated mutations ***(110)***, empirical potentials of mean force ***(111)***, NMR experiments ***(112)***, cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, site-directed mutagenesis ***(113)***, intuition, and so on. In this way, a comparative model, especially in the difficult cases, could be improved by making it consistent with available experimental data and/or with more general knowledge about protein structure.

Accuracies of the various model-building methods are relatively similar when used optimally ***(85)***. Other factors such as template selection and alignment accuracy usually have a larger impact on the model accuracy, especially for models based on less than 40% sequence identity to the templates. However, it is important that a modeling method allow a degree of flexibility and automation to obtain better models more easily and rapidly. For example, a method should allow for an easy recalculation of a model when a change is made in the alignment; it should be straightforward to calculate models based on several templates; and the method should provide tools for incorporation of prior knowledge about the target (e.g., cross-linking restraints, predicted secondary structure) and allow *ab initio* modeling of insertions (e.g., loops), which can be crucial for annotation of function. Loop modeling is an especially important aspect of comparative modeling in the range from 30 to 50% sequence identity. In this range of overall similarity, loops among the homologs vary while the core regions are still relatively conserved and aligned accurately.

2.3.4. Loop Modeling

In comparative modeling, target sequences often have residues inserted relative to the template structures, or have regions that are structurally different from the corresponding regions in the templates. Thus, no structural information about these inserted segments can be extracted from the template structures. These regions frequently correspond to surface loops. Loops often play an important role in defining the functional specificity of a given protein framework, forming the active and binding sites. The accuracy of loop modeling is a major factor determining the usefulness of comparative models in applications such as ligand docking. Loop modeling can be seen as a mini-protein folding problem, because the correct conformation of a given segment of a polypeptide chain has to be calculated mainly from the sequence of the segment itself. However, loops are generally too short to provide sufficient information about their local fold. Even identical decapeptides in different proteins do not always have the same conformation ***(114**,**115)***. Some additional restraints are provided by the core anchor regions that span the loop, and by the structure of the rest of a protein that cradles the loop. Although many loop-modeling methods have been described, it is still not possible to model correctly and confidently loops longer than approximately eight residues ***(102)***.

There are two main classes of loop-modeling methods: (1) the database search approaches that scan a database of all known protein structures to find segments fitting the anchor core regions ***(98**,**116)***; and (2) the conformational search approaches that rely on an optimization of a scoring function ***(117**,**118)***. There are also methods that combine these two approaches ***(119**,**120)***.

The database search approach to loop modeling is accurate and efficient when a database of specific loops is created to address the modeling of the same class of loops, such as β-hairpins ***(121)***, or loops on a specific fold, such as the hyper-variable regions in the immunoglobulin fold ***(116**,**122)***. There are attempts to classify loop conformations into more general categories, thus extending the applicability of the database search approach to more cases ***(123–125)***. However, the database methods are limited by the fact that the number of possible conformations increases exponentially with the length of a loop. As a result, only loops up to four to seven residues long have most of their conceivable conformations present in the database of known protein structures ***(126**,**127)***. Even according to the more optimistic estimate, approx 30% and 60% of all the possible eight- and nine-residue loop conformations, respectively, are missing from the database ***(126)***. This limitation is made even worse by the requirement for an overlap of at least one residue between the database fragment and the anchor core regions, which means that the modeling of a 5-residue insertion requires at least a 7-residue fragment from the database ***(92)***. Despite the rapid growth of the database of known structures, it does not seem possible to cover most of the conformations of a 9-residue segment in the foreseeable future. On the other hand, most of the insertions in a family of homologous proteins are shorter than 10–12 residues ***(102)***.

To overcome the limitations of the database search methods, conformational search methods were developed ***(117**,**128)***. There are many such methods, exploiting different protein representations, objective function terms, and optimization or enumeration algorithms. The search algorithms include the minimum perturbation method ***(129)***, molecular dynamics simulations ***(94**,**119)***, genetic algorithms ***(130)***, Monte Carlo and simulated annealing ***(131–133)***, multiple-copy simultaneous search ***(134)***, self-consistent field optimization ***(135)***, and an enumeration based on the graph theory ***(136)***. The accuracy of loop predictions can be further improved by clustering the sampled loop conformations and therefore partially accounting for the entropic contribution to the free energy ***(137)***. Another way to improve the accuracy of loop predictions is to consider the solvent effects. Improvements in implicit solvation models, such as the generalized Born solvation model (GB) ***(138)*** and surface-generalized Born with nonpolar correction (SGB/NP) ***(139)***, motivated their use in loop modeling. The solvent contribution to the free energy can be added to the scoring function for optimization, or it can be used to rank the sampled loop conformations after they are generated with a scoring function that does not include the solvent terms ***(2,140–143)***.

The loop modeling module in MODELLER implements the optimization-based approach ***(2,102)***. The main reasons are the generality and conceptual simplicity of scoring function minimization, as well as the limitations on the database approach imposed by a relatively small number of known protein structures ***(126)***. Loop prediction by optimization is applicable to simultaneous modeling of several loops and loops interacting with ligands, which is not straightforward for the database search approaches. Loop optimization in MODELLER relies on conjugate gradients and molecular dynamics with simulated annealing. The pseudo-energy function is a sum of many terms, including some terms from the CHARMM22 molecular mechanics force field ***(104)*** and spatial restraints based on distributions of distances ***(111,144)*** and dihedral angles in known protein structures. The method was tested on a large number of loops of known structure, both in the native and near-native environments ***(102)***.

2.3.5. Side-Chain Modeling

Two simplifications are frequently applied in the modeling of side-chain conformations. First, amino acid replacements often leave the backbone conformation almost unchanged ***(145)***, allowing us to fix the backbone during the search for the best side-chain conformations. Second, most side chains in high-resolution crystallographic structures can be represented by a limited number of conformers that comply with stereochemical and energetic constraints ***(146)***. This observation motivated Ponder and Richards to develop the first library of side-chain rotamers for the 17 types of residues with dihedral angle degrees of freedom in their side chains, based on 10 high-resolution protein structures determined by X-ray crystallography ***(147)***. Subsequently, a number of additional libraries have been derived ***(148–152)***.

Rotamers on a fixed backbone are often used when all the side chains need to be modeled on a given backbone. This approach overcomes the combinatorial explosion associated with a full conformational search of all the side chains, and is applied by some comparative modeling ***(6)*** and protein design approaches ***(153)***. However, approx 15% of the side chains cannot be represented well by these libraries ***(154)***. In addition, it has been shown that the accuracy of side-chain modeling on a fixed backbone decreases rapidly when the backbone errors are larger than only 0.5 Å ***(155)***. Fortunately, these two approximations may be unnecessary in the modeling of a single-point mutation that in general does not trigger changes in many dihedral angles ***(152)***.

Earlier methods for side-chain modeling often put less emphasis on the energy or scoring function. The function was usually greatly simplified, and consisted of the empirical rotamer preferences and simple repulsion terms for non-bonded contacts ***(151)***. Nevertheless, these approaches have been justified by their performance. For example, a method based on a rotamer library compared favorably with that based on a molecular-mechanics force field ***(156)*** , and more recently all the new and most efficient methods are also based on rotamer library ***(152,157)***. In contrast, a lot of attention has been paid to the optimization procedure. The various approaches include a Monte Carlo simulation ***(158)***, simulated annealing ***(159)***, a combination of Monte Carlo and simulated annealing ***(160)***, the dead-end elimination theorem ***(161,162)***, genetic algorithms ***(148)***, neural network with simulated annealing ***(163)***, mean field optimization ***(164)***, and combinatorial searches ***(151,165,166)***. It was suggested that the modeling accuracy for up to 10-residue segments is currently limited by the accuracy of the scoring function, not by the thoroughness of the search algorithms ***(102)***. Several recent papers focused on the testing of more sophisticated potential functions for conformational search ***(166,167)*** and development of new scoring functions for side-chain modeling ***(168)***, report favorable performance compared to earlier studies.

3. Errors in Comparative Modeling

It is crucial for method developers and users alike to assess the accuracy of their methods. An attempt to address this problem has been made by the Critical Assessment of Techniques for Proteins Structure Prediction (CASP) ***(169)*** and the Critical Assessment of Fully Automated Structure Prediction (CAFASP) experiments ***(170)***. However, both CASP and CAFASP assess methods only over a limited number of target protein sequences ***(85,171)***. To overcome this limitation, two additional evaluation experiments have been described, LiveBench ***(171)*** and EVA ***(172,173)***. EVA is a large-scale and continuously running Web server that automatically assesses protein structure prediction servers in the categories of secondary structure prediction, residue-residue contact prediction, fold assignment, and comparative modeling. The aims of EVA are (1) to evaluate continuously and automatically blind predictions by prediction servers, based on identical and sufficiently large data sets; (2) to provide weekly updates of the method assessments on the Web; and (3) to enable developers, non-expert users, and reviewers to determine the performance of the tested prediction servers.

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 as follows ***(49)*** (**Fig. 4**).

1. First, errors in side-chain packing. As the sequences diverge, the packing of the atoms 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.
2. Second, 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 an error ***(49,174)***.
3. Third, 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. As mentioned in the section on loop modeling, this problem is akin to *ab initio* fold prediction. If the insertion is relatively short, less than nine residues long, some methods can correctly predict the conformation of the backbone ***(102,119,143,175)***. Conditions for successful prediction are the correct alignment and an accurately modeled environment surrounding the insertion.

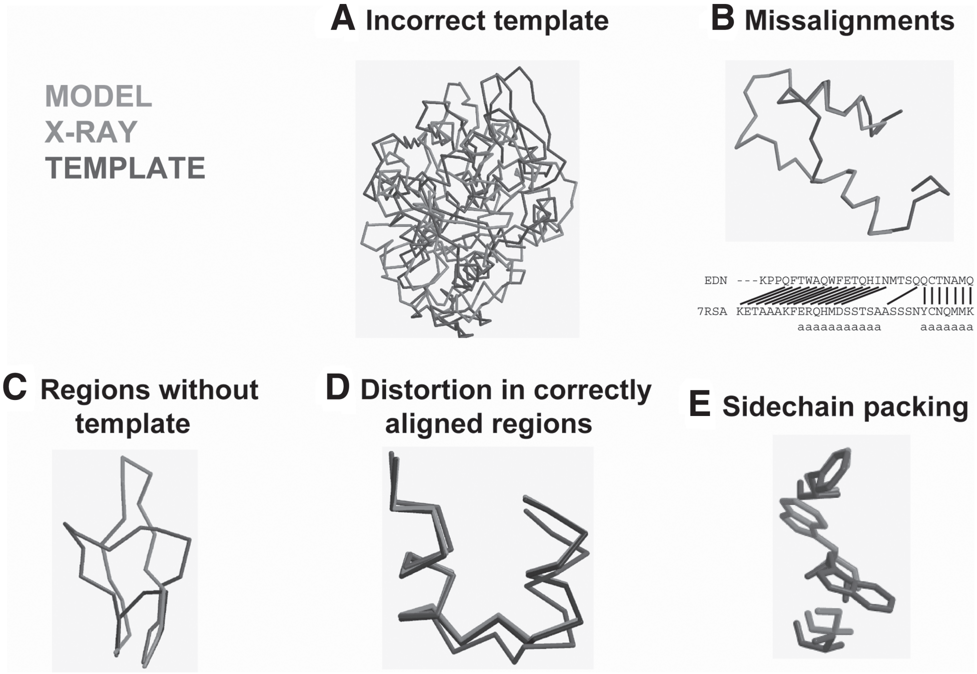
MODEL
X-RAY
TEMPLATE
A Incorrect template
B Missalignments
EDN ---KPPQFTWAQWFETQHINMTSQQCTNAMQ
7RSA KETAAAKFERQHMDSSTSAASSSNYCNQMMK
aaaaaaaaaaa aaaaaaa
C Regions without template
D Distortion in correctly aligned regions
E Sidechain packing

Fig. 4. Typical errors in comparative modeling.

4. Fourth, errors resulting from misalignments. The largest source of errors in comparative modeling is 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 ***(176,177)***. A second way of improving the alignment is to iteratively modify those regions in the alignment that correspond to predicted errors in the model ***(49)***.
5. Fifth, selection of incorrect templates. This error 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 conservation of the key functional or structural residues in the target sequence increases the confidence in a given fold assignment.

4. Predicting Model Accuracy

The accuracy and extent of the predicted structure determines the information that can be extracted from it. Thus, estimating the accuracy of 3-D protein models in the absence of the known structures is essential for interpreting them. The model can be evaluated as a whole as well as in the individual regions. There are many model evaluation programs and servers ***(178,179)*** (**Table 1**).

The first step in model evaluation is to determine whether the model has the correct fold ***(180)***. 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 confidence in the fold of a model is generally increased by a high sequence similarity with the closest template, an energy-based Z-score ***(180**,**181)***, or by conservation of the key functional or structural residues in the target sequence.

Once the fold of a model is accepted, a more detailed evaluation of the overall model accuracy can be obtained, based on the similarity between the target and template sequences ***(180)***. Sequence identity above 30% is a relatively good predictor of the expected accuracy, because the deviation from the least-squares curve relating sequence identity to the accuracy is relatively small. The reasons are the well-known relationship between structure and sequence similarities of two proteins ***(145)***, the “geometrical” nature of modeling (which forces the model to be as close to the template as possible) ***(101)***, and the inability of any current modeling procedure to recover from an incorrect alignment ***(49)***. The dispersion of the model-target structure overlap increases with the decrease in sequence identity. If the target–template sequence identity falls below 30%, the sequence identity becomes unreliable as a predictor of the model accuracy. 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 ***(182)***. This observation 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 ***(183)***, PROCHECK-NMR ***(184)***, AQUA ***(184)***, SQUID ***(185)***, and WHATCHECK ***(186)***. 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 non-bonded pairs of atoms.

There are also methods for testing 3-D models that implicitly take into account many spatial features compiled from high-resolution protein structures. These methods are based on 3-D profiles and statistical potentials of mean force ***(74**,**111**,**144)***. Programs implementing this approach include VERIFY3D ***(74)***, PROSAII ***(181)***, HARMONY ***(187)***, ANOLEA ***(188)***, and DFIRE ***(189)***. These 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 ***(102)***. 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 expectation 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 ***(190)***.

5. Applications of Comparative Modeling

Comparative models have been used in a myriad of applications ***(1**,**191)***. The applicability of a model depends on its accuracy (**Fig. 3**). We now list typical applications of comparative models.

Models that are built using as templates protein structures with which they share less than approx 25% in sequence identity are usually used for fold assignment. Such models often have less than 50% of their C_α positions within 3.5 Å of the actual structure. Nevertheless, fold assignment is frequently sufficient to assign coarse protein function ***(20**,**192)***. At this level of target–template similarity, model evaluation can be used as a discriminator between correct and incorrect fold assignment ***(49**,**144**,**180)***.

Models built on approx 35% sequence identity to the templates, on the average cover about 85% of the residues to within 3.5 Å of their correct positions. Since the active and binding sites of proteins are frequently more conserved than the rest of the fold, they tend to be modeled more accurately than the rest of the fold ***(180)***. In general, medium-resolution models frequently allow a refinement of the functional prediction based on sequence alone, because ligand binding is most directly determined by the structure of the binding site rather than its sequence. It may be 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 ***(193)*** , and the size of a ligand may be predicted from the volume of the binding-site cleft ***(194)***. Medium-resolution models can also be used to construct site-directed mutants with altered binding capacity, which in turn could test hypotheses about the sequence-structure-function relationships. Other problems that can be addressed with medium-resolution comparative models include designing proteins that have compact structures—without long tails, loops, and exposed hydrophobic residues—for better crystallization; or designing proteins with added disulfide bonds for extra stability.

The high end of the accuracy spectrum corresponds to models based on 50% sequence identity or more. The average accuracy of these models approaches that of low-resolution X-ray structures (3Å resolution) or medium-resolution NMR structures (10 distance restraints per residue) ***(49)***. The alignments on which these models are based generally contain almost no errors. In addition to the already listed applications, high-accuracy models can be used for docking of small ligands ***(130)*** or whole proteins onto the given protein ***(195)***. For an overall view of the scope of applicability of computational models, *see* **refs. *1*,*191***.

Acknowledgments

Our research has been supported by Sandler Family Supporting Foundation, NIH/NIGMS R01 GM 54762, NIH/NIGMS P50 GM62529, NIH/NCI R33 CA84699, Sun Academic Equipment Grant EDUD-7824-020257-US, an IBM SUR grant, and an Intel computer hardware gift.

References

1. Baker, D. and Sali, A. (2001) Protein structure prediction and structural genomics. *Science* **294,** 93–96.
2. Fiser, A., Feig, M., Brooks, C. L. III, and Sali, A. (2002) Evolution and physics in comparative protein structure modeling. *Acc. Chem. Res.* **35,** 413–421.
3. Bradley, P., Chivian, D., Meiler, J., et al. (2003) Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins* **53 Suppl. 6,** 457–468.
4. Kinch, L. N., Wrabl, J. O., Krishna, S. S., et al. (2003) CASP5 assessment of fold recognition target predictions. *Proteins* **53 Suppl. 6,** 395–409.
5. Marti-Renom, M. A., Stuart, A., Fiser, A., et al. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291–325.
6. Blundell, T. L., Sibanda, B. L., Sternberg, M. J., and Thornton, J. M. (1987) Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **326,** 347–352.
7. Pieper, U., Eswar, N., Braberg, H., et al. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* **32,** D217–D222.
8. Chance, M. R., Bresnick, A. R., Burley, S. K., et al. (2002) Structural Genomics: A pipeline for providing structures for the biologist. *Protein Sci.* **11,** 723–738.
9. Burley, S. K., Almo, S. C., Bonanno, J. B., et al. (1999) Structural genomics: beyond the human genome project. *Nat. Genet.* **23,** 151–157.
10. Sanchez, R., Pieper, U., Mirkovic, N., et al. (2000) MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.* **28,** 250–253 .
11. Sali, A. and Kuriyan, J. (1999) Challenges at the frontiers of structural biology. *Trends Cell Biol.* **9,** M20–M24.
12. Montelione, G. T. and Anderson, S. (1999) Structural genomics: keystone for a Human Proteome Project. *Nat. Struct. Biol.* **6,** 11–12.
13. Sali, A. (1998) 100,000 protein structures for the biologist. *Nat. Struct. Biol.* **5,** 1029–1032.
14. Gerstein, M., Edwards, A., Arrowsmith, C. H., and Montelione, G. T. (2003) Structural genomics: current progress. *Science* **299,** 1663.
15. Brenner, S. E. (2000) Target selection for structural genomics. *Nat. Struct. Biol.* **7 Suppl.,** 967–969.
16. Vitkup, D., Melamud, E., Moult, J., and Sander, C. (2001) Completeness in structural genomics. *Nat. Struct. Biol.* **8,** 559–566.
17. Westbrook, J., Feng, Z., Jain, S., et al. (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.* **30,** 245–248.
18. Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* **30,** 264–267.
19. Holm, L. and Sander, C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.* **27,** 244–247.
20. Orengo, C. A., Bray, J. E., Buchan, D. W., et al. (2002) The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics* **2,** 11–21.
21. Pearl, F. M., Lee, D., Bray, J. E., et al. (2002) The CATH extended protein-family database: Providing structural annotations for genome sequences. *Protein Sci.* **11,** 233–244.
22. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.
23. Pearson, W. R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183,** 63–98.

24. Brenner, S. E., Chothia, C., and Hubbard, T. J. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA* **95,** 6073–6078.
25. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402.
26. Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **9,** 232–241.
27. Sadreyev, R. and Grishin, N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.* **326,** 317–336.
28. Panchenko, A. R. (2003) Finding weak similarities between proteins by sequence profile comparison. *Nucleic Acids Res.* **31,** 683–689.
29. Wallner, B., Fang, H., Ohlson, T., Frey-Skott, J., and Elofsson, A. (2004) Using evolutionary information for the query and target improves fold recognition. *Proteins* **54,** 342–50.
30. Teichmann, S. A., Chothia, C., Church, G. M., and Park, J. (2000) Fast assignment of protein structures to sequences using the intermediate sequence library PDB-ISL. *Bioinformatics* **16,** 117–124.
31. Li, W., Pio, F., Pawlowski, K., and Godzik, A. (2000) Saturated BLAST: an automated multiple intermediate sequence search used to detect distant homology. *Bioinformatics* **16,** 1105–1110.
32. Park, J., Karplus, K., Barrett, C., et al. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284,** 1201–1210.
33. Gerstein, M. (1998) Measurement of the effectiveness of transitive sequence comparison, through a third ‘intermediate’ sequence. *Bioinformatics* **14,** 707–714.
34. Pipenbacher, P., Schliep, A., Schneckener, S., et al. (2002) ProClust: improved clustering of protein sequences with an extended graph-based approach. *Bioinformatics* **18 Suppl. 2,** S182–S191.
35. Salamov, A. A., Suwa, M., Orengo, C. A., and Swindells, M. B. (1999) Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng.* **12,** 95–100.
36. John, B. and Sali, A. (2004) Detection of homologous proteins by an intermediate sequence search. *Protein Sci.* **13,** 54–62.
37. Jones, D. T. (1997) Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins* **Suppl. 1,** 185–191.
38. Smith, T. F., Lo Conte, L., Bienkowska, J., et al. (1997) Current limitations to protein threading approaches. *J. Comput. Biol.* **4,** 217–225.
39. Torda, A. E. (1997) Perspectives in protein-fold recognition. *Curr. Opin. Struct. Biol.* **7,** 200–205.
40. Levitt, M. (1997) Competitive assessment of protein fold recognition and alignment accuracy. *Proteins* **Suppl. 1,** 92–104.
41. David, R., Korenberg, M. J., and Hunter, I. W. (2000) 3D-1D threading methods for protein fold recognition. *Pharmacogenomics* **1,** 445–455.
42. Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313,** 903–919.
43. Jones, D. T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287,** 797–815.

44. Lindahl, E. and Elofsson, A. (2000) Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* **295,** 613–625.
45. Balaji, S., Sujatha, S., Kumar, S. S., and Srinivasan, N. (2001) PALI-a database of Phylogeny and ALIgnment of homologous protein structures. *Nucleic Acids Res.* **29,** 61–65.
46. Navaratnam, N., Fujino, T., Bayliss, J., et al. (1998) *Escherichia coli* cytidine deaminase provides a molecular model for ApoB RNA editing and a mechanism for RNA substrate recognition. *J. Mol. Biol.* **275,** 695–714.
47. Reva, B., Finkelstein, A., and Topiol, S. (2002) Threading with chemostructural restrictions method for predicting fold and functionally significant residues: application to dipeptidylpeptidase IV (DPP-IV). *Proteins* **47,** 180–193.
48. Srinivasan, S., March, C. J., and Sudarsanam, S. (1993) An automated method for modeling proteins on known templates using distance geometry. *Protein Sci.* **2,** 277–289.
49. Sanchez, R. and Sali, A. (1997) Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins* **Suppl. 1,** 50–58.
50. Barton, G. J. (1996) Protein sequence alignment and database scanning. In: (Sternberg, M. J. E., ed.) *Protein Structure Prediction: A Practical Approach*, IRL Press at Oxford University Press, Oxford, UK.
51. Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443–453.
52. Sellers, P. H. (1974) Theory and computation of evolutionary distances. *Siam Journal on Applied Mathematics* **26,** 787–793.
53. Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197.
54. Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89,** 10,915–10,919.
55. Dayhoff, M., Schwartz, R., and BC, O. (1978) A model of evolutionary change in proteins, 345–352, National Biomedical Research Foundation, Washington, DC.
56. Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* **84,** 4355–4358.
57. Gribskov, M. (1994) Profile analysis. *Methods Mol. Biol.* **25,** 247–266.
58. Gribskov, M., Luthy, R., and Eisenberg, D. (1990) Profile analysis. *Methods Enzymol.* **183,** 146–159.
59. Karplus, K., Barrett, C., and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14,** 846–856.
60. Marti-Renom, M. A., Madhusudhan, M. S., and Sali, A. (2004) Alignment of protein sequences by their profiles. *Protein Sci.* **13(4),** 1071–1087.
61. Higgins, D. G. and Sharp, P. M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73,** 237–244.
62. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680.
63. Pietrokovski, S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.* **24,** 3836–3845.
64. Jaroszewski, L., Rychlewski, L., and Godzik, A. (2000) Improving the quality of twilight-zone alignments. *Protein Sci.* **9,** 1487–1496.
65. Yona, G., Linial, N., and Linial, M. (2000) ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.* **28,** 49–55.
66. Yona, G. and Levitt, M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.* **315,** 1257–1275.

67. Yona, G., Linial, N., and Linial, M. (1999) ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins* **37,** 360–378.
68. Yona, G. and Levitt, M. (2000) Towards a complete map of the protein space based on a unified sequence and structure analysis of all known proteins. *ISMB* **8,** 395–406.
69. Bowie, J. U., Luthy, R., and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253,** 164–170.
70. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) A new approach to protein fold recognition. *Nature* **358,** 86–89.
71. Godzik, A. and Skolnick, J. (1992) Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. *Proc. Natl. Acad. Sci. USA* **89,** 12,098–12,102.
72. Jones, D. T. (1997) Progress in protein structure prediction. *Curr. Opin. Struct. Biol.* **7,** 377–387.
73. Teodorescu, O., Galor, T., Pillardy, J., and Elber, R. (2004) Enriching the sequence substitution matrix by structural information. *Proteins* **54,** 41–8.
74. Luthy, R., Bowie, J. U., and Eisenberg, D. (1992) Assessment of protein models with three-dimensional profiles. *Nature* **356,** 83–85.
75. Shi, J., Blundell, T. L., and Mizuguchi, K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310,** 243–257.
76. Bates, P. A., Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins* **Suppl. 5,** 39–46.
77. Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299,** 499–520.
78. Karchin, R., Cline, M., Mandel-Gutfreund, Y., and Karplus, K. (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins* **51,** 504–514.
79. Tang, K. S., Fersht, A. R., and Itzhaki, L. S. (2003) Sequential unfolding of ankyrin repeats in tumor suppressor p16. *Structure (Camb.)* **11,** 67–73.
80. Zhu, Z. Y., Sali, A., and Blundell, T. L. (1992) A variable gap penalty function and feature weights for protein 3-D structure comparisons. *Protein Eng.* **5,** 43–51.
81. Madhusudhan, M. S., Marti-Renom, M. A., Sanchez, R., and Sali, A. (2004) Variable gap penalty function for protein sequence—structure alignment. in preparation.
82. Madhusudhan, M. S., Marti-Renom, M. A., Eswar, N., and Sali, A. (2004) SALIGN: a comprehensive sequence/structure alignment algorithm. in preparation.
83. Venclovas, C. (2003) Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance. *Proteins* **53 Suppl. 6,** 380–388.
84. John, B. and Sali, A. (2003) Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res.* **31,** 3982–3992.
85. Marti-Renom, M. A., Madhusudhan, M. S., Fiser, A., Rost, B., and Sali, A. (2002) Reliability of assessment of protein structure prediction methods. *Structure* **10,** 435–440.
86. Browne, W. J., North, A. C. T., Phillips, D. C., et al. (1969) A possible three-dimensional structure of bovine lactalbumin based onthat of hen's egg-white lysosyme. *J. Mol. Biol.* **42,** 65–86.
87. Sutcliffe, M. J., Haneef, I., Carney, D., and Blundell, T. L. (1987) Knowledge based modelling of homologous proteins, Part I: Three- dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.* **1,** 377–384.

88. Topham, C. M., McLeod, A., Eisenmenger, F., et al. (1993) Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. *J. Mol. Biol.* **229,** 194–220.
89. Nagarajaram, H. A., Reddy, B. V., and Blundell, T. L. (1999) Analysis and prediction of inter-strand packing distances between beta-sheets of globular proteins. *Protein Eng.* **12,** 1055–1062.
90. Unger, R., Harel, D., Wherland, S., and Sussman, J. L. (1989) A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins* **5,** 355–373.
91. Bystroff, C. and Baker, D. (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* **281,** 565–577.
92. Claessens, M., Van Cutsem, E., Lasters, I., and Wodak, S. (1989) Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng.* **2,** 335–345.
93. Holm, L. and Sander, C. (1991) Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *J. Mol. Biol.* **218,** 183–194.
94. Bruccoleri, R. E. and Karplus, M. (1990) Conformational sampling using high-temperature molecular dynamics. *Biopolymers* **29,** 1847–1862.
95. van Gelder, C. W., Leusen, F. J., Leunissen, J. A., and Noordik, J. H. (1994) A molecular dynamics approach for the generation of complete protein structures from limited coordinate data. *Proteins* **18,** 174–185.
96. Levitt, M. (1992) Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226,** 507–533.
97. Chinea, G., Padron, G., Hooft, R. W., Sander, C., and Vriend, G. (1995) The use of position-specific rotamers in model building by homology. *Proteins* **23,** 415–421.
98. Jones, T. A. and Thirup, S. (1986) Using known substructures in protein model building and crystallography. *EMBO J.* **5,** 819–822.
99. Havel, T. F. and Snow, M. E. (1991) A new method for building protein conformations from sequence alignments with homologues of known structure. *J. Mol. Biol.* **217,** 1–7.
100. Sali, A., Overington, J. P., Johnson, M. S., and Blundell, T. L. (1990) From comparisons of protein sequences and structures to protein modelling and design. *Trends Biochem. Sci.* **15,** 235–240.
101. Sali, A. and Blundell, T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234,** 779–815.
102. Fiser, A., Do, R. K., and Sali, A. (2000) Modeling of loops in protein structures. *Protein Sci.* **9,** 1753–1773.
103. Sali, A. and Overington, J. P. (1994) Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.* **3,** 1582–1596.
104. MacKerell, A. D., Jr., Bashford, D., Bellott, M., et al. (1998) All-atom empirical potential for molecular modleing and dynamics studies of proteins. *J. Phys. Chem. B* **102,** 3586–3616.
105. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., et al. (1983) CHARMM: A program for macromolecular energy minimization and dynamics calculations. *J. Comp. Chem.* **4,** 187–217.
106. Braun, W. and Go, N. (1985) Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. *J. Mol. Biol.* **186,** 611–626.
107. Clore, G. M., Brunger, A. T., Karplus, M., and Gronenborn, A. M. (1986) Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination. A model study of crambin. *J. Mol. Biol.* **191,** 523–551.
108. Cohen, F. E. and Kuntz, I. D. (1989) Tertiary structure prediction. In: (Fasman, G. D., ed.) *Prediction of Protein Structure and the Principles of Protein Conformations,* Plenum, New York, NY: 647–705.

109. Aszodi, A. and Taylor, W. R. (1994) Secondary structure formation in model polypeptide chains. *Protein Eng.* **7,** 633–644.
110. Taylor, W. R. and Hatrick, K. (1994) Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7,** 341–348.
111. Sippl, M. J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213,** 859–883.
112. Sutcliffe, M. J., Dobson, C. M., and Oswald, R. E. (1992) Solution structure of neuronal bungarotoxin determined by two-dimensional NMR spectroscopy: calculation of tertiary structure using systematic homologous model building, dynamical simulated annealing, and restrained molecular dynamics. *Biochemistry* **31,** 2962–2970.
113. Boissel, J. P., Lee, W. R., Presnell, S. R., Cohen, F. E., and Bunn, H. F. (1993) Erythropoietin structure-function relationships. Mutant proteins that test a model of tertiary structure. *J. Biol. Chem.* **268,** 15,983–15,993.
114. Kabsch, W. and Sander, C. (1984) On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc. Natl. Acad. Sci. USA* **81,** 1075–1078.
115. Mezei, M. (1998) Chameleon sequences in the PDB. *Protein Eng.* **11,** 411–414.
116. Chothia, C. and Lesk, A. M. (1987) Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196,** 901–917.
117. Bruccoleri, R. E. and Karplus, M. (1987) Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* **26,** 137–168.
118. Shenkin, P. S., Yarmush, D. L., Fine, R. M., Wang, H. J., and Levinthal, C. (1987) Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. *Biopolymers* **26,** 2053–2085.
119. van Vlijmen, H. W. and Karplus, M. (1997) PDB-based protein loop prediction: parameters for selection and methods for optimization. *J. Mol. Biol.* **267,** 975–1001.
120. Deane, C. M. and Blundell, T. L. (2001) CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci.* **10,** 599–612.
121. Sibanda, B. L., Blundell, T. L., and Thornton, J. M. (1989) Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J. Mol. Biol.* **206,** 759–777.
122. Chothia, C., Lesk, A. M., Tramontano, A., et al. (1989) Conformations of immunoglobulin hypervariable regions. *Nature* **342,** 877–883.
123. Rufino, S. D., Donate, L. E., Canard, L. H., and Blundell, T. L. (1997) Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling. *J. Mol. Biol.* **267,** 352–367.
124. Oliva, B., Bates, P. A., Querol, E., Aviles, F. X., and Sternberg, M. J. (1997) An automated classification of the structure of protein loops. *J. Mol. Biol.* **266,** 814–830.
125. Ring, C. S., Kneller, D. G., Langridge, R., and Cohen, F. E. (1992) Taxonomy and conformational analysis of loops in proteins. *J. Mol. Biol.* **224,** 685–699.
126. Fidelis, K., Stern, P. S., Bacon, D., and Moult, J. (1994) Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng.* **7,** 953–960.
127. Lessel, U. and Schomburg, D. (1994) Similarities between protein 3-D structures. *Protein Eng.* **7,** 1175–1187.
128. Moult, J. and James, M. N. (1986) An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins* **1,** 146–163.
129. Fine, R. M., Wang, H., Shenkin, P. S., Yarmush, D. L., and Levinthal, C. (1986) Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynam-

ics studies of MCPC603 from many randomly generated loop conformations. *Proteins* **1,** 342–362.

130. Ring, C. S., Sun, E., McKerrow, J. H., et al. (1993) Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc. Natl. Acad. Sci. USA* **90,** 3583–3587.
131. Abagyan, R. and Totrov, M. (1994) Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **235,** 983–1002.
132. Collura, V., Higo, J., and Garnier, J. (1993) Modeling of protein loops by simulated annealing. *Protein Sci.* **2,** 1502–1510.
133. Higo, J., Collura, V., and Garnier, J. (1992) Development of an extended simulated annealing method: application to the modeling of complementary determining regions of immunoglobulins. *Biopolymers* **32,** 33–43.
134. Zheng, Q., Rosenfeld, R., Vajda, S., and DeLisi, C. (1993) Determining protein loop conformation using scaling-relaxation techniques. *Protein Sci.* **2,** 1242–1248.
135. Koehl, P. and Delarue, M. (1995) A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. *Nat. Struct. Biol.* **2,** 163–170.
136. Samudrala, R. and Moult, J. (1998) A graph-theoretic algorithm for comparative modeling of protein structure. *J. Mol. Biol.* **279,** 287–302.
137. Xiang, Z., Soto, C. S., and Honig, B. (2002) Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc. Natl. Acad. Sci. USA* **99,** 7432–7437.
138. Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112,** 6127–6129.
139. Ghosh, A., Rapp, C.S., and Friesner, R.A. (1998) Generalized born model based on a surface integral formulation. *J. Phys. Chem. B* **102,** 10,983–10,990.
140. de Bakker, P. I., DePristo, M. A., Burke, D. F., and Blundell, T. L. (2003) *Ab initio* construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. *Proteins* **51,** 21–40.
141. DePristo, M. A., de Bakker, P. I., Lovell, S. C., and Blundell, T. L. (2003) *Ab initio* construction of polypeptide fragments: efficient generation of accurate, representative ensembles. *Proteins* **51,** 41–55.
142. Felts, A. K., Gallicchio, E., Wallqvist, A., and Levy, R. M. (2002) Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the Surface Generalized Born solvent model. *Proteins* **48,** 404–422.
143. Jacobson, M., Pincus, D., Rapp, C. S., et al. (2004) A hierarchical approach to all-atom loop prediction. *Proteins* **55,** 351–367.
144. Melo, F., Sanchez, R., and Sali, A. (2002) Statistical potentials for fold assessment. Protein Sci. **11,** 430–448.
145. Chothia, C. and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5,** 823–826.
146. Janin, J. and Chothia, C. (1978) Role of hydrophobicity in the binding of coenzymes. Appendix. Translational and rotational contribution to the free energy of dissociation. *Biochemistry* **17,** 2943–2948.
147. Ponder, J. W. and Richards, F. M. (1987) Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193,** 775–791.

148. Tuffery, P., Etchebest, C., Hazout, S., and Lavery, R. (1991) A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn.* **8,** 1267–1289.
149. Mendes, J., Baptista, A. M., Carrondo, M. A., and Soares, C. M. (1999) Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. *Proteins* **37,** 530–543.
150. Dunbrack, R. L., Jr. and Cohen, F. E. (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6,** 1661–1681.
151. Dunbrack, R. L. and Karplus, M. (1993) Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **230,** 543–574.
152. Xiang, Z. and Honig, B. (2001) Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.* **311,** 421–430.
153. Desjarlais, J. R. and Handel, T. M. (1999) Side-chain and backbone flexibility in protein core design. *J. Mol. Biol.* **290,** 305–318.
154. De Filippis, V., Sander, C., and Vriend, G. (1994) Predicting local structural changes that result from point mutations. *Protein Eng.* **7,** 1203–1208.
155. Chung, S. Y. and Subbiah, S. (1996) How similar must a template protein be for homology modeling by side-chain packing methods? *Pac. Symp. Biocomput.* 126–141.
156. Cregut, D., Liautard, J. P., and Chiche, L. (1994) Homology modelling 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,** 1333–1344.
157. Canutescu, A. A., Shelenkov, A. A., and Dunbrack, R. L., Jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12,** 2001–2014.
158. Eisenmenger, F., Argos, P., and Abagyan, R. (1993) A method to configure protein side-chains from the main-chain trace in homology modelling. *J. Mol. Biol.* **231,** 849–860.
159. Lee, G. M., Varma, A., and Palsson, B. O. (1991) Application of population balance model to the loss of hybridoma antibody productivity. *Biotechnol. Prog.* **7,** 72–75.
160. Holm, L. and Sander, C. (1992) Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. *Proteins* **14,** 213–223.
161. Lasters, I. and Desmet, J. (1993) The fuzzy-end elimination theorem: correctly implementing the side chain placement algorithm based on the dead-end elimination theorem. *Protein Eng.* **6,** 717–722.
162. Looger, L. L. and Hellinga, H. W. (2001) Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J. Mol. Biol.* **307,** 429–445.
163. Hwang, J. K. and Liao, W. F. (1995) Side-chain prediction by neural networks and simulated annealing optimization. *Protein Eng.* **8,** 363–370.
164. Koehl, P. and 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,** 249–275.
165. Bower, M. J., Cohen, F. E., and Dunbrack, R. L., Jr. (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J. Mol. Biol.* **267,** 1268–1282.
166. Petrella, R. J., Lazaridis, T., and Karplus, M. (1998) Protein sidechain conformer prediction: a test of the energy function. *Fold Des.* **3,** 353–377.
167. Jacobson, M. P., Kaminski, G. A., Friesner, R. A., and Rapp, C. S. (2002) Force field validation using protein side chain prediction. *J. Phys. Chem. B* **106,** 11,673–11,680.
168. Liang, S. and Grishin, N. V. (2002) Side-chain modeling with an optimized scoring function. *Protein Sci.* **11,** 322–331.
169. Zemla, A., Venclovas, Moult, J., and Fidelis, K. (2001) Processing and evaluation of predictions in CASP4. *Proteins* **45 Suppl. 5,** 13–21.

170. Fischer, D., Elofsson, A., Rychlewski, L., et al. (2001) CAFASP2: The second critical assessment of fully automated structure prediction methods. *Proteins* **45 Suppl. 5,** 171–183.

171. Bujnicki, J. M., Elofsson, A., Fischer, D., and Rychlewski, L. (2001) LiveBench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci.* **10,** 352–361.

172. Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., et al. (2001) EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics* **17,** 1242–1243.

173. Koh, I. Y., Eyrich, V. A., Marti-Renom, M. A., et al. (2003) EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res.* **31,** 3311–3315.

174. Srinivasan, N. and Blundell, T. L. (1993) An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. *Protein Eng.* **6,** 501–512.

175. Coutsias, E. A., Seok, C., Jacobson, M. P., and Dill, K. A. (2004) A kinematic view of loop closure. *J. Comput. Chem.* **25,** 510–528.

176. Barton, G. J. and Sternberg, M. J. (1987) A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J. Mol. Biol.* **198,** 327–337.

177. Taylor, W. R., Flores, T. P., and Orengo, C. A. (1994) Multiple protein structure alignment. *Protein Sci.* **3,** 1858–1870.

178. Laskowski, R. A., McArthur, A. G., and Thornton, J. M. (1998) PROCHECK: a program to check the stereochemical quality of portein structures. *J. Appl. Crystallog.* **26,** 283–291.

179. Wilson, C., Gregoret, L. M., and Agard, D. A. (1993) Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. *J. Mol. Biol.* **229,** 996–1006.

180. Sanchez, R. and Sali, A. (1998) Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc. Natl. Acad. Sci. USA* **95,** 13,597–13,602.

181. Sippl, M. J. (1993) Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aided Mol. Des.* **7,** 473–501.

182. Pawlowski, K., Bierzynski, A., and Godzik, A. (1996) Structural diversity in a family of homologous proteins. *J. Mol. Biol.* **258,** 349–366.

183. Laskowski, R. A., MacArthur, M. W., and Thornton, J. M. (1998) Validation of protein models derived from experiment. *Curr. Opin. Struct. Biol.* **8,** 631–639.

184. Laskowski, R. A., Rullmann, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8,** 477–486.

185. Oldfield, T. J. (1992) SQUID: a program for the analysis and display of data from crystallography and molecular dynamics. *J. Mol. Graph.* **10,** 247–252.

186. Hooft, R. W., Vriend, G., Sander, C., and Abola, E. E. (1996) Errors in protein structures. *Nature* **381,** 272.

187. Topham, C. M., Srinivasan, N., Thorpe, C. J., Overington, J. P., and Kalsheker, N. A. (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,** 869–894.

188. Melo, F. and Feytmans, E. (1998) Assessing protein structures with a non-local atomic interaction energy. *J. Mol. Biol.* **277,** 1141–1152.

189. Zhou, H. and Zhou, Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11,** 2714–2726.

190. Miwa, J. M., Ibanez-Tallon, I., Crabtree, G. W., et al. (1999) lynx1, an endogenous toxin-like modulator of nicotinic acetylcholine receptors in the mammalian CNS. *Neuron* **23,** 105–114.

191. Johnson, M. S., Srinivasan, N., Sowdhamini, R., and Blundell, T. L. (1994) Knowledge-based protein modelling. *CRC Crit. Rev. Biochem. Mol. Biol.* **29,** 1–68.
192. Orengo, C. A., Michie, A. D., Jones, S., et al. (1997) CATH—a hierarchic classification of protein domain structures. *Structure* **5,** 1093–1108.
193. Matsumoto, R., Sali, A., Ghildyal, N., Karplus, M., and Stevens, R. L. (1995) Packaging of proteases and proteoglycans in the granules of mast cells and other hematopoietic cells. A cluster of histidines on mouse mast cell protease 7 regulates its binding to heparin serglycin proteoglycans. *J. Biol. Chem.* **270,** 19,524–19,531.
194. Xu, L. Z., Sanchez, R., Sali, A., and Heintz, N. (1996) Ligand specificity of brain lipid-binding protein. *J. Biol. Chem.* **271,** 24,711–24,719.
195. Vakser, I. A. (1995) Protein docking for low-resolution structures. *Protein Eng.* **8,** 371–377.
196. Thompson, J. D., Plewniak, F., and Poch, O. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* **15,** 87–88.
197. Pearl, F. M., Bennett, C. F., Bray, J. E., et al. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.* **31,** 452–455.
198. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2004) GenBank: update. *Nucleic Acids Res.* **32 Database issue,** D23–D26.
199. Lin, J., Qian, J., Greenbaum, D., et al. (2002) GeneCensus: genome comparisons in terms of metabolic pathway activity and protein family sharing. *Nucleic Acids Res.* **30,** 4574–4582.
200. Bourne, P. E., Addess, K. J., Bluhm, W. F., et al. (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res.* **32 Database issue,** D223–D225.
201. Brenner, S. E., Barken, D., and Levitt, M. (1999) The PRESAGE database for structural genomics. *Nucleic Acids Res.* **27,** 251–253.
202. Andreeva, A., Howorth, D., Brenner, S. E., et al. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* **32 Database issue,** D226–D229.
203. Boeckmann, B., Bairoch, A., Apweiler, R., et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31,** 365–370.
204. Alexandrov, N. N., Nussinov, R., and Zimmer, R. M. (1996) Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. *Pac. Symp. Biocomput.* 53–72.
205. Godzik, A., Kolinski, A., and Skolnick, J. (1992) Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.* **227,** 227–238.
206. Rost, B. and Liu, J. (2003) The PredictProtein server. *Nucleic Acids Res.* **31,** 3300–3304.
207. Flockner, H., Braxenthaler, M., Lackner, P., et al. (1995) Progress in fold recognition. *Proteins* **23,** 376–386.
208. Mallick, P., Weiss, R., and Eisenberg, D. (2002) The directional atomic solvation energy: an atom-based potential for the assignment of protein sequences to known folds. *Proc. Natl. Acad. Sci. USA* **99,** 16,041–16,046.
209. Gough, J. and Chothia, C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* **30,** 268–272.
210. Worley, K. C., Culpepper, P., Wiese, B. A., and Smith, R. F. (1998) BEAUTY-X: enhanced BLAST searches for DNA queries. *Bioinformatics* **14,** 890–891.
211. Tatusova, T. A. and Madden, T. L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174,** 247–250.
212. Henikoff, J. G., Pietrokovski, S., McCallum, C. M., and Henikoff, S. (2000) Blocks-based methods for detecting protein homology. *Electrophoresis* **21,** 1700–1706.

213. Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16,** 10,881–10,890.

214. Ye, Y., Jaroszewski, L., Li, W., and Godzik, A. (2003) A segment alignment approach to protein comparison. *Bioinformatics* **19,** 742.

215. Karplus, K., Karchin, R., Draper, J., et al. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins* **53 Suppl. 6,** 491–496.

216. Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302,** 205–217.

217. Edgar, R. C. and Sjolander, K. (2004) COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics* **20(8),** 1309–1318.

218. Eswar, N., John, B., Mirkovic, N., et al. (2003) Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.* **31,** 3375–3380.

219. Pearson, W. R. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci.* **4,** 1145–1160.

220. Abagyan, R., Frishman, D., and Argos, P. (1994) Recognition of distantly related proteins through energy calculations. *Proteins* **19,** 132–140.

221. Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* **31,** 3381–3385.

222. Vriend, G. (1990) WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8,** 56.

223. Colovos, C. and Yeates, T. O. (1993) Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.* **2,** 1511–1519.

224. Pontius, J., Richelle, J., and Wodak, S. J. (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.* **264,** 121–136.

225. Hooft, R. W., Sander, C., and Vriend, G. (1996) Verification of protein structure: sidechain planarity. *J. Appl. Crystallog.* 714–716.

226. Moult, J., Fidelis, K., Zemla, A., and Hubbard, T. (2001) Critical assessment of methods of protein structure prediction (CASP): Round IV. *Proteins* **45 Suppl. 5,** 2–7.

227. Kahsay, R. Y., Wang, G., Dongre, N., Gao, G., and Dunbrack, R. L., Jr. (2002) CASA: a server for the critical assessment of protein sequence alignment accuracy. *Bioinformatics* **18,** 496–497.

228. Livingstone, C. D. and Barton, G. J. (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9,** 745–756.

229. Beckmann, R., Spahn, C. M., Eswar, N., et al. (2001) Architecture of the protein-conducting channel associated with the translating 80S ribosome. *Cell* **107,** 361–372.