Modeling Protein Structure from Its Sequence

Functional characterization of a protein sequence is one of the most frequent problems in biology. This task is usually facilitated by accurate three-dimensional (3-D) structure of the studied protein. In the absence of an experimentally determined structure, comparative or homology modeling can sometimes provide a useful 3-D model for a protein that is related to at least one known protein structure. Comparative modeling predicts the 3-D structure of a given protein sequence (target) based primarily on its alignment to one or more proteins of known structure (templates). The prediction process consists of fold assignment, target-template alignment, model building, and model evaluation. The number of protein sequences that can be modeled, as well as the accuracy of the predictions, is increasing steadily because of growth in the number of known protein sequences and structures as well as improvements in the modeling software. It is currently possible to model, with useful accuracy, significant parts of approximately one-half of all known protein sequences (Pieper et al., 2002).

Despite progress in ab initio protein structure prediction (Baker, 2000; Bonneau and Baker, 2001), comparative modeling remains the only method that can reliably predict the 3-D structure of a protein with an accuracy comparable to a low-resolution experimentally determined structure (Marti-Renom et al., 2000). Even models with errors may be useful, because some aspects of function can be predicted from only coarse structural features (Marti-Renom et al., 2000; Baker and Sali, 2001).

There are several computer programs and Web servers that automate the comparative modeling process (see Table 5.1.1). Several of these servers are being evaluated in an automated, continuous and large-scale fashion by EVA-cm (*http://cubic.bioc.columbia.edu/eva*; Eyrich et al., 2001) and LiveBench (*http://bioinfo.pl/LiveBench/*; Bujnicki et al., 2001).

While the Web servers are convenient and useful, the best results in the difficult or unusual modeling cases, such as problematic alignments, modeling of loops, existence of multiple conformational states, and modeling of ligand binding, are still obtained by nonautomated, expert use of the various modeling tools. A number of resources useful in comparative modeling are listed in Table 5.1.1.

In this unit, generic considerations in all four steps of comparative modeling (Fig. 5.1.1) are first described. These considerations are then illustrated by a detailed discussion of modeling of lactate dehydrogenase from *Trichomonas vaginalis* using the program MODELLER, which was developed by the authors of this unit (Sali and Blundell, 1993; Sali and Overington, 1994; Fiser et al., 2000; also see Internet Resources). Finally, the authors outline typical applications of comparative modeling of genes and genomes (Fig. 5.1.2).

STEPS IN COMPARATIVE MODELING

Fold Assignment and Template Selection

The starting point in comparative modeling is to identify all protein structures related to the target sequence, and then select those structures that will be used as templates. This step is facilitated by numerous protein sequence and structure databases and by the use of database-scanning software available on the Web (Altschul et al., 1994; Barton, 1998; Holm and Sander, 1996; also see Table 5.1.1). Templates can be found using the target sequence as a query for searching structure databases such as the Protein Data Bank (Westbrook et al., 2002), SCOP (Lo Conte et al., 2002), DALI (Holm and Sander, 1999),

Contributed by Marc A. Marti-Renom, Andras Fiser, M.S. Madhusudhan, Bino John, Ashley Stuart, Narayanan Eswar, Ursula Pieper, Min-yi Shen, and Andrej Sali

Current Protocols in Bioinformatics (2003) 5.1.1-5.1.32

Table 5.1.1 Programs and Web Servers Useful in Comparative Modeling

Name	Type[a]	WWW address[b]	Reference
Databases			
CATH	S	*http://www.biochem.ucl.ac.uk/bsm/cath/*	Orengo et al. (2002)
GenBank	S	*http://www.ncbi.nlm.nih.gov/Genbank/*	Blundell et al. (1987)
GeneCensus	S	*http://bioinfo.mbb.yale.edu/genome/*	Gerstein and Levitt (1997)
MODBASE	S	*http://www.salilab.org/modbase/*	Pieper et al. (2002)
PALI	S	*http://pauling.mbu.iisc.ernet.in/~pali/*	Gowri et al. (2003)
PDB	S	*http://www.rcsb.org/pdb/*	Westbrook et al. (2002)
PRESAGE	S	*http://presage.berkeley.edu*	Brenner et al. (1999)
SCOP	S	*http://scop.mrc-lmb.cam.ac.uk/scop/*	Lo Conte et al. (2002)
TrEMBL	S	*http://srs.ebi.ac.uk*	Bairoch and Apweiler (2000)
Template search			
123D	S	*http://123d.ncifcrf.gov/123D+.html*	Alexandrov et al. (1996)
BLAST	S	*http://www.ncbi.nlm.nih.gov/BLAST/*	Altschul et al. (1990)
DALI	S	*http://www2.ebi.ac.uk/dali/*	Holm and Sander (1999)
FastA	S	*http://www.ebi.ac.uk/fasta33/*	Pearson (1990)
MATCHMAKER	P	*http://bioinformatics.burnham-inst.org*	Godzik et al. (1992)
PHD, TOPITS	S	*http://cubic.bioc.columbia.edu/predictprotein/*	Rost (1995); Rost and Sander (1995)
PROFIT	P	*http://www.came.sbg.ac.at*	Flockner et al. (1995)
THREADER	P	*http://bioinf.cs.ucl.ac.uk/threader/threader.html*	Jones et al. (1992)
FRSVR	S	*http://fold.doe-mbi.ucla.edu*	Fischer and Eisenberg (1996)
Sequence alignment			
BCM SERVER	S	*http://searchlauncher.bcm.tmc.edu*	Smith et al. (1996)
BLAST2	S	*http://www.ncbi.nlm.nih.gov/gorf/bl2.html*	Altschul et al. (1997)
BLOCK MAKER	S	*http://blocks.fhcrc.org/blocks/blockmkr/make_blocks.html*	Henikoff et al. (1995)
CLUSTAL	S	*http://www2.ebi.ac.uk/clustalw/*	Thompson et al. (1994)
FASTA3	S	*http://www2.ebi.ac.uk/fasta3/*	Pearson (1990)
MULTALIN	S	*http://pbil.ibcp.fr*	Corpet (1988)
Modeling			
COMPOSER	P	*http://www.tripos.com/*	Sutcliffe et al. (1987)
CONGEN	P	*http://www.congenomics.com/congen/congen.html*	Bruccoleri and Karplus (1990)
ICM	P	*http://www.molsoft.com*	—[c]
InsightII	P	*http://www.accelrys.com*	—[d]
MODELLER	P	*http://www.salilab.org/modeller/*	Sali and Blundell (1993)[e]
QUANTA	P	*http://www.accelrys.com*	—[d]
SYBYL	P	*http://www.tripos.com*	—[f]
SCWRL	P	*http://www.fccc.edu/research/labs/dunbrack/scwrl/*	Bower et al. (1997)
SNPWEB	S	*http://salilab.org/snpweb-cgi/snpweb.cgi*	Mirkovic et al. (2003)
SWISS-MOD	S	*http://www.expasy.org/swissmod/SWISS-MODEL.html*	Peitsch and Jongeneel (1993)
WHAT IF	P	*http://www.cmbi.kun.nl/whatif/*	Vriend (1990)

continued

and CATH (Orengo et al., 2002). The probability of finding a related protein of known structure for a sequence picked randomly from a genome ranges from 20% to 70% (Fischer and Eisenberg, 1997; Huynen et al., 1998; Rychlewski et al., 1998; Sanchez and Sali, 1998; Jones, 1999; Pieper et al., 2002).

There are three main classes of protein comparison methods that are useful in fold identification. The first class includes the methods that compare the target sequence with each of the database sequences independently, using pairwise sequence-sequence comparison (Apostolico and Giancarlo, 1998). The performance of these methods in searching

Table 5.1.1 Programs and Web Servers Useful in Comparative Modeling, *continued*

Name	Type[a]	WWW address[b]	Reference
Model evaluation			
ANOLEA	S	*http://protein.bio.puc.cl/cardex/servers*	Melo and Feytmans (1998)
AQUA	P	*http://urchin.bmrb.wisc.edu/~jurgen/aqua/*	Laskowski et al. (1996)
BIOTECH[g]	S	*http://biotech.embl-heidelberg.de:8400*	Laskowski et al. (1998)
ERRAT	S	*http://www.doe-mbi.ucla.edu/Services/ERRAT/*	Colovos and Yeates (1993)
PROCHECK	P	*http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html*	Laskowski et al. (1998)
ProsaII	P	*http://www.came.sbg.ac.at*	Sippl (1993)
PROVE	S	*http://www.ucmb.ulb.ac.be/UCMB/PROVE*	Pontius et al. (1996)
SQUID	P	*http://www.ysbl.york.ac.uk/~oldfield/squid/*	Oldfield (1992)
VERIFY3D	S	*http://www.doe-mbi.ucla.edu/Services/Verify_3D/*	Luthy et al. (1992)
WHATCHECK	P	*http://www.sander.embl-heidelberg.de/whatcheck/*	Hooft et al. (1996a)
Methods evaluation			
CASP	S	*http://predictioncenter.llnl.gov*	Moult et al. (2001)
CAFASP	S	*http://cafasp.bioinfo.pl*	Fischer et al. (2001)
EVA	S	*http://cubic.bioc.columbia.edu/eva/*	Eyrich et al. (2001)
LiveBench	S	*http://bioinfo.pl/LiveBench/*	Bujnicki et al. (2001)

[a] S, server; P, program.
[b] Some of the sites are mirrored on additional computers.
[c] MolSoft Inc., San Diego, Calif.
[d] Accelrys Inc., San Diego, Calif.
[e] See Internet Resources for additional information.
[f] Tripos Inc., St Louis, Mo.
[g] The BIOTECH server uses PROCHECK and WHATCHECK for structure evaluation.

for related protein sequences and structures has been evaluated exhaustively (Thompson et al., 1999; Sauder et al., 2000). Frequently used programs in this class include FASTA (Pearson, 1995; Pearson and Lipman, 1988) and BLAST (*UNITS 3.3 & 3.4*; Altschul et al., 1990). Studies indicate that pairwise sequence comparison methods can detect only one half of the evolutionary relationships in the range of 20% to 30% sequence identity (Brenner et al., 1998).

In the second class of methods, sequence comparison benefits from multiple sequence information using profile analysis (Gribskov, 1994), profile-profile comparisons (Yona and Levitt, 2002; Rychlewski et al., 2000), Hidden Markov Models (Eddy, 1998; Karplus et al., 1998; Martelli et al., 2002), and intermediate sequence search (Park et al., 1997; Gerstein, 1998; Teichmann et al., 2000). Popularly used fold recognition programs include PSI-BLAST (Altschul et al., 1997) and SAM (Karplus et al., 1998). These approaches are especially useful for finding important structural relationships when the sequence identity between the target and the template drops below 25% (Muller et al., 1999). In general, multiple sequence comparisons detect three times more evolutionary relationships than pairwise sequence comparisons when sequence identity drops below 30% (Park et al., 1998).

The third class of methods are the so-called threading or 3-D template matching methods (Bowie et al., 1991; Godzik et al., 1992; Jones et al., 1992; reviewed in Jones, 1997; Levitt, 1997; Smith et al., 1997; Torda, 1997; David et al., 2000). These methods rely on pairwise comparison of a protein sequence and a protein of known structure. Whether or not a given target sequence adopts any one of the many known 3-D folds is predicted by an optimization of the alignment with respect to a structure-dependent scoring function, independently for each sequence-structure pair—i.e., the target sequence is threaded through a library of 3-D folds. Examples of widely used programs in this class include THREADER (Jones et al., 1992) and 3D-PSSM (Kelley et al., 2000). These methods are

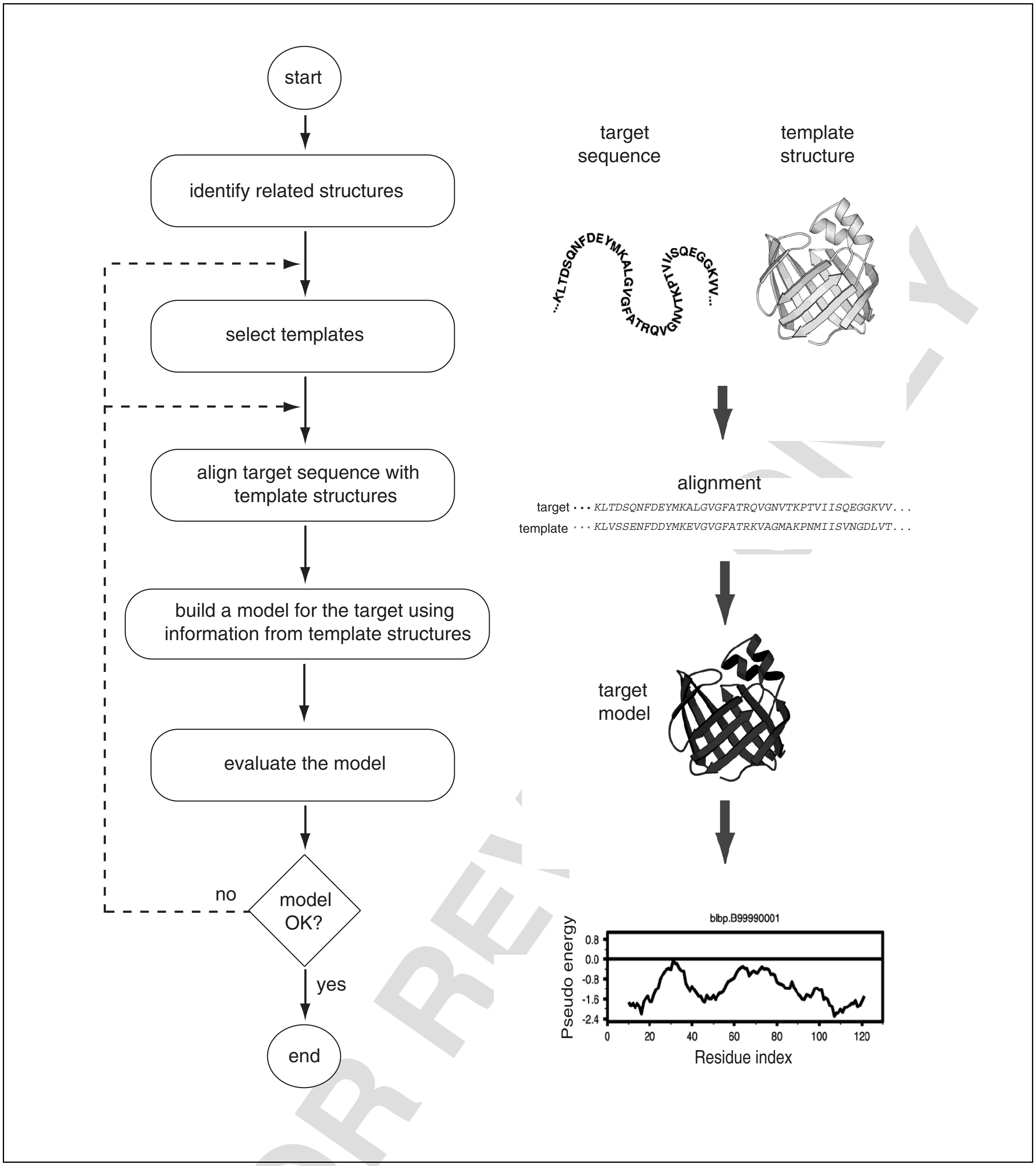
start
identify related structures
select templates
align target sequence with template structures
build a model for the target using information from template structures
evaluate the model
model OK?
no
yes
end
target sequence
template structure
...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
alignment
target ··· KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
template ··· KLVSSENFDDYMKEVGVGFATRKVAGMAKPNMIISVNGDLVT...
target model
blbp.B99990001
Pseudo energy
0.8
0.0
-0.8
-1.6
-2.4
0
20
40
60
80
100
120
Residue index

Figure 5.1.1 Steps in comparative protein structure modeling. See text for details.

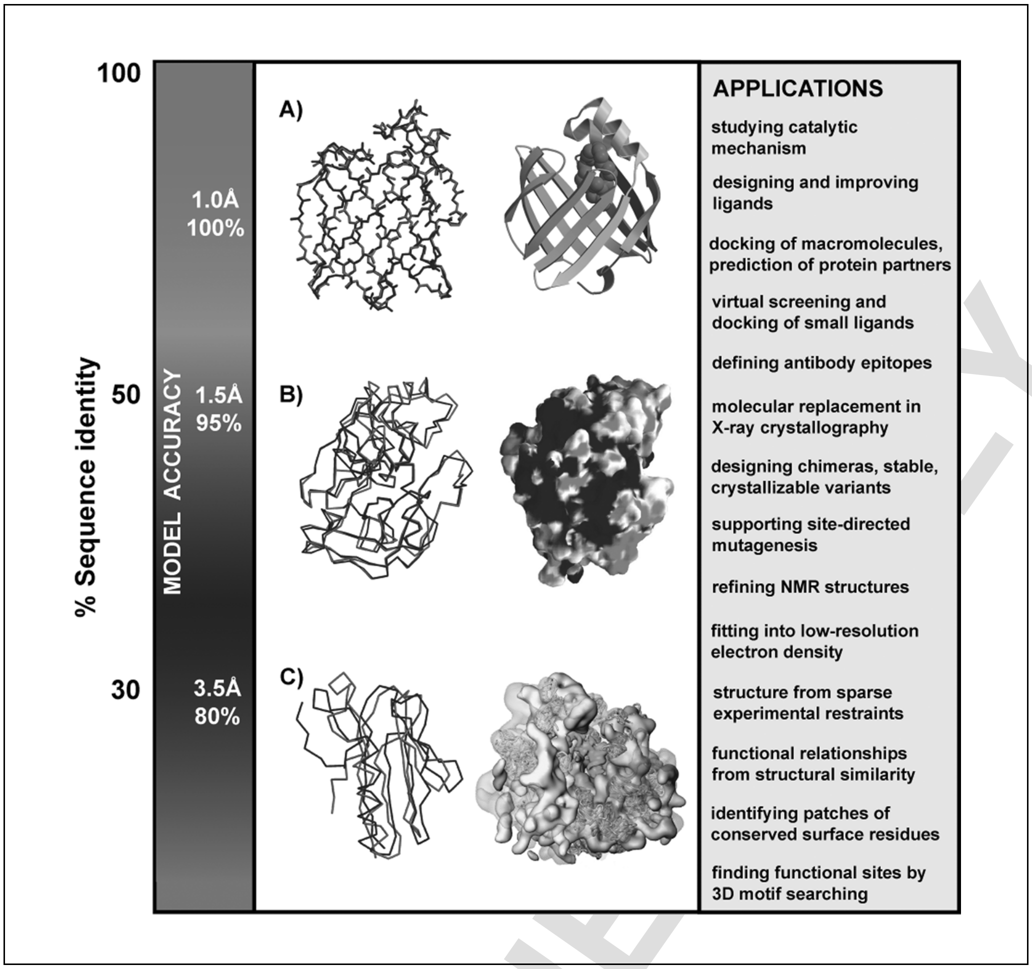
100
50
30
% Sequence identity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
A)
B)
C)
APPLICATIONS
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screening and docking of small ligands
defining antibody epitopes
molecular replacement in X-ray crystallography
designing chimeras, stable, crystallizable variants
supporting site-directed mutagenesis
refining NMR structures
fitting into low-resolution electron density
structure from sparse experimental restraints
functional relationships from structural similarity
identifying patches of conserved surface residues
finding functional sites by 3D motif searching

Figure 5.1.2 Accuracy and application 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 (van der Waals sphere model) 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 (Xu et al., 1996), even though the ligand-specificity profile of this protein is different from that of the template structure. Typical overall accuracy of a comparative model in this range of sequence similarity is indicated by a comparison of a model for adipocyte fatty acid binding protein with its actual 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 (PDB code, 2ptn) that does not bind proteoglycans. The prediction was confirmed by site-directed mutagenesis and heparin-affinity chromatography experiments (Matsumoto et al., 1995). 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 15 Å resolution (Beckmann et al., 2001). Most of the models for 40 out of the 75 ribosomal proteins were based on approximately 30% sequence identity to their template structures. 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 *B. Stearothermophilus* with the actual structure (PDB code, 1rl2).

especially useful when the target sequence is related to only a few other sequences, so the search cannot benefit from the increased sensitivity of the sequence profile methods. Threading-based methods have been shown to outperform other approaches based on sequence alone (Lindahl and Elofsson, 2000).

A useful fold-assignment approach is to accept an uncertain assignment provided by any of the methods, build a full-atom comparative model of the target sequence based on this match, and make the final decision about whether or not the match is real by evaluating the resulting comparative model (Guenther et al., 1997; Sanchez and Sali, 1997a; Miwa et al., 1999).

Once a list of all related protein structures has been obtained, it is necessary to select those templates that are appropriate for the given modeling problem. Usually, a higher overall sequence similarity between the target and the template sequence yields a better model. In any case, several other factors should be taken into account when selecting the templates:

1. The family of proteins, which includes the target and the templates, can frequently be organized in subfamilies. The construction of a multiple alignment and a phylogenetic tree (Felsenstein, 1985; *UNITS 2.3, 3.6, & 6.3*; Chapter 6) can help in selecting the template from the subfamily that is closest to the target sequence.

2. The template environment should be compared to the required environment for the model. The term "environment" is used in a broad sense and includes all factors that determine protein structure, except its sequence (e.g., solvent, pH, ligands, and quaternary interactions). For instance, some calcium-binding proteins undergo large conformational changes when bound to calcium (Dainese et al., 2002). Thus, to model the calcium-bound state of the target, a calcium-bound template is preferred over a calcium-free template, irrespective of the target-template similarity or accuracy of the template structure (Pawlowski et al., 1996).

3. The quality of the experimental template structure is another important factor in template selection. The resolution and the R-factor of a crystallographic structure and the number of restraints per residue for an NMR structure are indicative of its accuracy.

Prior biological information of the target sequence can also be valuable in identifying an appropriate template (Navaratnam et al., 1998; Reva et al., 2002).

The priority of the criteria for template selection depends on the purpose of the comparative model. For instance, if a protein-ligand model is to be constructed, the choice of the template that contains a similar ligand is probably more important than the resolution of the template. On the other hand, if the model is to be used to analyze the geometry of the active site of an enzyme, it may be 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 (Srinivasan and Blundell, 1993; Sanchez and Sali, 1997).

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 (Lo Conte et al., 2002). However, at the superfamily and fold levels, the accuracy dropped to 29% and 15%, respectively (Lindahl and Elofsson, 2000).

Target-Template Alignment

Most fold-assignment methods produce an alignment between the target sequence and template structures. However, this is often not the optimal target-template alignment for comparative modeling. Searching methods are usually tuned for detection of remote relationships, not for optimal alignments. Therefore, once templates have been selected, a specialized method should be used to align the target sequence with the template structures (Holm and Sander, 1996; Taylor, 1996; Baxevanis, 1998; Briffeuil et al., 1998; Smith, 1999). For closely related protein sequences with identity higher than 40%, the alignment is almost always correct. Regions of low local sequence similarity become common when the overall sequence identity is below 40% (Saqi et al., 1998). The alignment becomes difficult in the "twilight zone" of less than 30% sequence identity (Rost, 1999). As the sequence similarity decreases, alignments contain an increasingly large number of gaps and alignment errors, regardless of whether they are prepared automatically or manually. For example, only 80% of the residues are likely to be correctly aligned when two proteins share 30% sequence identity (Johnson and Overington, 1993). Maximal effort to obtain the most accurate alignment possible is needed, because no current comparative modeling method can recover from an incorrect alignment (Sanchez and Sali, 1997).

Dynamic programming algorithms (Smith and Waterman 1981; Needleman and Wunsch, 1970; *UNIT 3.1*) that use standard substitution matrices, such as PAM (Dayhoff and Eck, 1968; *UNIT 3.5*) or BLOSUM (Henikoff and Henikoff, 1992; *UNIT 3.5*), were the initial methods of choice. Although dynamic programming guarantees the optimal solution, the insensitivity and generality of the substitution matrices limited the usefulness of such methods to cases of high sequence identity (> 30%). For alignments in the so called twilight zone (15% to 30% sequence identity), it is imperative that the alignment methods incorporate information that is specific to the family to which the proteins belong. Ideally, each sequence position in a sequence family should have its own residue-based substitution matrix. In the absence of sufficient observations to accomplish such a construct, several methods make approximations to this end.

Programs such as PSI-BLAST (Altschul et al., 1997) align the target sequence to a sequence profile constructed from a multiple alignment of members of a protein family. A further improvement of this class of methods is to align two sequence profiles. Examples of this approach include FFAS (Jaroszewski et al., 2000) and SALIGN. Alignment accuracy increases as one progresses from one generation of the profile methods to another. These methods improve alignments in the 20% to 30% sequence-identity range. A variation of the sequence profile scheme is to incorporate structural information into the alignment procedure. This information can be introduced during the profile building stage or it can be used to bias the insertion and extension of gaps in the alignment.

Another alignment strategy is to build models based on many alignments and then rank the alignments by the corresponding model assessment scores. This step can be iterated to improve the initial alignments. Although such a procedure can be time consuming, it can significantly improve the resulting comparative models (Xu et al., 1996; John and Sali, 2003).

Model Building

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 (Browne et al., 1969; Blundell et al., 1987; Greer, 1990). 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 semi-automated procedure is implemented in the computer program COMPOSER (Sutcliffe et al., 1987). First, the template structures are selected and superposed. Second, the "framework" is calculated by averaging the coordinates of the C_α atoms of structurally conserved regions in the template structures. Third, the 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 (Topham et al., 1993). Fifth, the side chains are modeled based on their intrinsic conformational preferences and on the conformation of the equivalent sidechains in the template structures (Sutcliffe et al., 1987). 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 (Srinivasan and Blundell, 1993). 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 (Nagarajaram et al., 1999).

Modeling by segment matching or coordinate reconstruction

The basis of modeling by coordinate reconstruction is the finding that most hexapeptide segments of protein structure can be clustered into only 100 structurally different classes (Unger et al., 1989; Bystroff and Baker, 1998). Thus, comparative models can be constructed by using a subset of atomic positions from template structures as "guiding" positions and 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 (Claessens et al., 1989; Holm and Sander, 1991), or by a conformational search restrained by an energy function (Bruccoleri and Karplus, 1990; van Gelder et al., 1994). For example, a general method for modeling by segment matching is guided by the positions of some atoms (usually C_α atoms) to find the matching segments in the representative database of all known protein structures (Levitt, 1992). This method can construct both main-chain and side-chain atoms, and can also model unaligned regions (gaps). It is implemented in the program SegMod. Even some side-chain modeling methods (Chinea et al., 1995) and the class of loop construction methods based on finding suitable fragments in the database of known structures (Jones and Thirup, 1986) can be seen as segment matching or coordinate reconstruction methods.

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 (Havel and Snow, 1991).

At this point the authors of this unit will describe their own approach to comparative modeling by satisfaction of spatial restrains in more detail (Sali et al., 1990; Sali and Blundell, 1993; Sali and Overington, 1994; Fiser et al., 2000). The approach was designed to use as many different types of data about the target sequence as possible. It is implemented in the computer program MODELLER. 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 non-hydrogen atoms.

In the first step of model building, distance and dihedral angle restraints on the target sequence are derived from its alignment with 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 (Sali and Overington, 1994). 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 (Sali and Blundell, 1993). These relationships are expressed as conditional probability density functions (pdf's), and can be used directly as spatial restraints. For example, probabilities for different values of the main-chain dihedral angles are calculated from the type of a residue considered, from main-chain conformation of an equivalent residue, and from sequence similarity between the two proteins. Another example is the pdf for a certain 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 (Brooks et al., 1983; MacKerell, Jr. et al., 1998) are combined into an objective function. The general form of the objective function is similar to that in molecular dynamics programs, such as CHARMM22 (Brooks et al., 1983). The objective function depends on the Cartesian coordinates of ~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 law, Lennard-Jones potential, and cubic splines. The geometric features presently include a distance, an angle, a dihedral angle, a pair of dihedral angles between two, three, four, and eight atoms, respectively, the shortest distance in the set of distances, solvent accessibility in Å^2, and atom density that is expressed as the number of atoms around the central atom. Some restraints can be used to restrain pseudo-atoms such as the gravity center of several atoms.

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

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

Accuracies of the various model building methods are relatively similar when used optimally (Marti-Renom et al., 2002a). 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 enough to calculate models based on several templates, and 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 of 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.

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 (Kabsch and Sander, 1984; Mezei and Guarnieri, 1998). Some additional restraints are provided by the core anchor regions that span the loop and by the structure of the rest of a protein that cradles the loop. Although many loop-modeling methods have been described, it is still not possible to model, correctly and confidently, loops longer than ~8 residues (Fiser et al., 2000).

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 (Jones and Thirup, 1986; Chothia and Lesk, 1987); and (2) the conformational search approaches that rely on an optimization of a scoring function (Moult and James, 1986; Bruccoleri and Karplus, 1987; Shenkin et al., 1987). There are also methods that combine these two approaches (van Vlijmen and Karplus, 1997; Deane and Blundell, 2001).

The database-search approach to loop modeling is accurate and efficient when a database of specific loops is created to address the modeling of the same class of loops, such as α-hairpins (Sibanda et al., 1989), or loops on a specific fold, such as the hypervariable regions in the immunoglobulin fold (Chothia and Lesk, 1987; Chothia et al., 1989). There are attempts to classify loop conformations into more general categories, thus extending the applicability of the database-search approach to more cases (Ring et al., 1992; Rufino et al., 1997; Oliva et al., 1997). 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 4 to 7 residues long have most of their conceivable conformations present in the database of known protein structures (Fidelis et al., 1994; Lessel and Schomburg, 1994). Even according to the more optimistic estimate, ~30% and 60% of all the possible 8 and 9 residue loop conformations, respectively, are missing from the database (Fidelis et al., 1994). 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 (Claessens et al., 1989). 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 to 12 residues (Fiser et al., 2000).

To overcome the limitations of the database-search methods, conformational search methods were developed (Moult and James, 1986; Bruccoleri and Karplus, 1987). There are many such methods, exploiting different protein representations, objective function terms, and optimization or enumeration algorithms. The search algorithms include the minimum perturbation method (Fine et al., 1986), molecular dynamics simulations (Bruccoleri and Karplus, 1990; van Vlijmen and Karplus, 1997), genetic algorithms (Ring et al., 1993), Monte Carlo and simulated annealing (Higo et al., 1992; Collura et al., 1993; Abagyan and Totrov, 1994), multiple copy simultaneous search (Zheng et al., 1993), self-consistent field optimization (Koehl and Delarue, 1995), and an enumeration based on the graph theory (Samudrala and Moult, 1998). 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 (Xiang et al., 2002). 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, 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 (Fiser et al., 2001; Felts et al., 2002; de Bakker et al., 2003; DePristo et al., 2003).

The loop-modeling module in MODELLER implements the optimization-based approach (Fiser et al., 2000; Fiser and Sali, 2003). The main reasons are the generality and conceptual simplicity of scoring function minimization, as well as the limitations on the database approach that are imposed by a relatively small number of known protein structures (Fidelis et al., 1994). Loop prediction by optimization is applicable to simultaneous modeling of several loops and loops interacting with ligands, which is not straightforward for the database-search approaches. Loop optimization in MODELLER relies on conjugate gradients and molecular dynamics with simulated annealing. The pseudo energy function is a sum of many terms, including some terms from the CHARMM22 molecular mechanics force field (MacKerell, Jr. et al., 1998), and spatial restraints based on distributions of distances (Sippl, 1990; Melo et al., 2002) 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 (Fiser et al., 2000).

Errors in Comparative Models

The evaluation of three dimensional comparative models is a difficult task (Bujnicki et al., 2001; Eyrich et al., 2001; Marti-Renom et al., 2002a). 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 with the experiments by CASP (Critical Assessment of Techniques for Proteins Structure Prediction; Zemla et al., 2001) and CAFASP (Critical Assessment of Fully Automated Structure Prediction; Fischer et al., 2001). However, both CASP and CAFASP assess methods only over a limited number of target protein sequences (Bujnicki et al., 2001; Marti-Renom et al., 2002a). To overcome this limitation, two additional evaluation experiments have been described, LiveBench (Bujnicki et al., 2001; Table 5.1.1) and EVA (Eyrich et al., 2001; Koh et al., 2003; Table 5.1.1). 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, nonexpert 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 (Sanchez and Sali, 1997; also see Fig. 5.1.3):

1. Errors in side-chain packing (Fig. 5.1.13A). As the sequences diverge, the packing of side chains in the protein core changes. Sometimes even the conformation of identical side chains is not conserved, a pitfall for many comparative modeling methods. Side-chain errors are critical if they occur in regions that are involved in protein function, such as active sites and ligand-binding sites.

2. Distortions and shifts in correctly aligned regions (Fig. 5.1.13B). 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 (Srinivasan and Blundell, 1993; Sanchez and Sali, 1997).

3. Errors in regions without a template (Fig. 5.1.13C). Segments of the target sequence that have no equivalent region in the template structure (i.e., insertions or loops) are the most difficult regions to model. If the insertion is relatively short, less than 9 residues long, some methods can correctly predict the conformation of the backbone (van Vlijmen and Karplus, 1997; Fiser et al., 2000). Conditions for successful prediction are the correct alignment and an accurately modeled environment surrounding the insertion.

4. Errors due to misalignments (Fig. 5.1.13D). The largest single source of errors in comparative modeling are misalignments, especially when the target-template sequence identity decreases below 30%. However, alignment errors can be minimized in two ways. First, it is usually possible to use a large number of sequences to construct

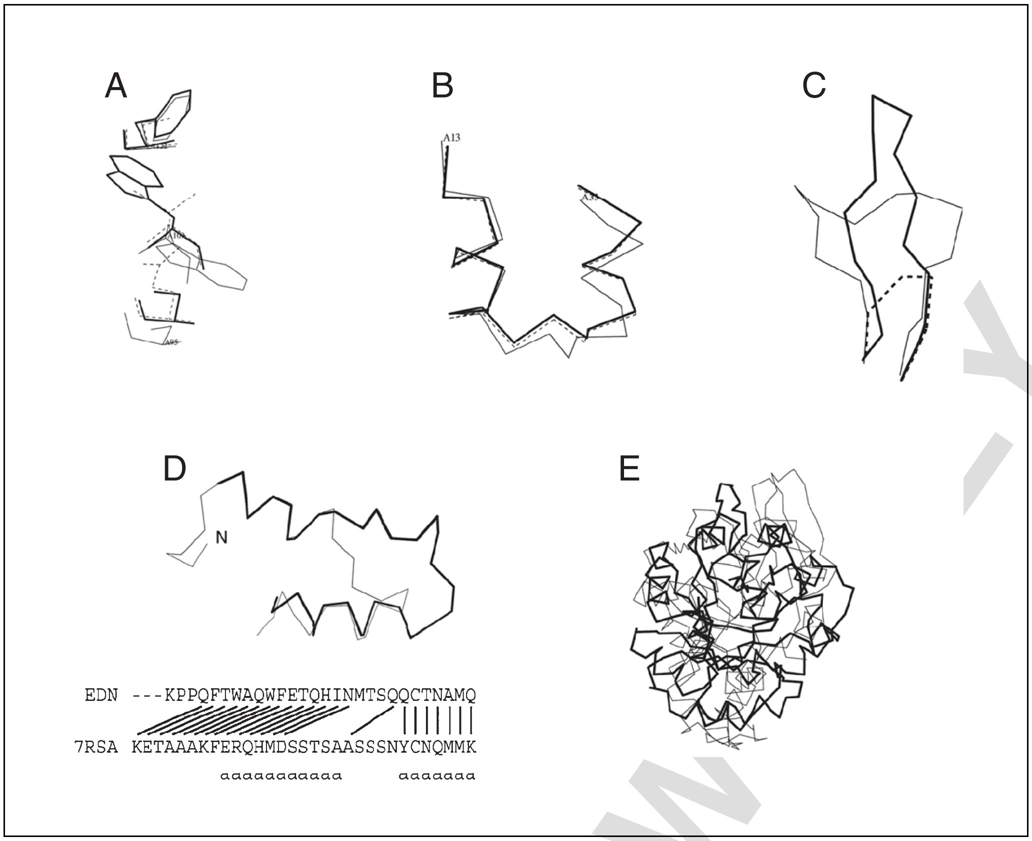
A
B
A13
C
D
N
E
EDN ---KPPQFTWAQWFETQHINMTSQQCTNAMQ
7RSA KETAAAKFERQHMDSSTSAASSSNYCNQMMK
aaaaaaaaaaa aaaaaaa

Figure 5.1.3 Typical errors in comparative modeling. *(***A***)*. Errors in side-chain packing. The Trp 109 residue in the crystal structure of mouse cellular retinoic acid binding protein I (thin line) is compared with its model (thick line), and with the template mouse adipocyte lipid-binding protein (broken line). (**B**) Distortions and shifts in correctly aligned regions. A region in the crystal structure of mouse cellular retinoic acid binding protein I is compared with its model and with the template fatty acid binding protein using the same representation as in panel A. (**C**) Errors in regions without a template. The C_α trace of the 112–117 loop is shown for the X-ray structure of human eosinophil neurotoxin (thin line), its model (thick line), and the template ribonuclease A structure (residues 111–117; broken line). (**D**) Errors due to misalignments. The N-terminal region in the crystal structure of human eosinophil neurotoxin (thin line) is compared with its model (thick line). The corresponding region of the alignment with the template ribonuclease A is shown. The black lines show correct equivalences, that is, residues whose C_α atoms are within 5 Å of each other in the optimal least-squares superposition of the two X-ray structures. The "a" characters in the bottom line indicate helical residues. (**E**) Errors due to an incorrect template. The X-ray structure of α-trichosanthin (thin line) is compared with its model (thick line) which was calculated using indole-3-glycerophosphate synthase as the template.

a multiple alignment, even if most of these sequences do not have known structures. Multiple alignments are generally more reliable than pairwise alignments (Barton and Sternberg, 1987; Taylor et al., 1994). The second way of improving the alignment is to iteratively modify those regions in the alignment that correspond to predicted errors in the model (Sanchez and Sali, 1997).

5. Incorrect templates (Fig. 5.1.13E). This is a potential problem when distantly related proteins are used as templates (i.e., <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.

Predicting the Model Accuracy

The quality of the predicted model 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 (Laskowski et al., 1998; Wilson et al., 1993; Table 5.1.1).

The first step in model evaluation is to determine if the model has the correct fold (Sanchez and Sali, 1998). 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, a significant energy-based Z-score (Sippl, 1993; Sanchez and Sali, 1998), and 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 (Sanchez and Sali, 1998). Sequence identity above 30% is a relatively good predictor of the expected accuracy. The reasons are the well-known relationship between structural and sequence similarities of two proteins (Chothia and Lesk, 1986), the “geometrical” nature of modeling that forces the model to be as close to the template as possible (Sali and Blundell, 1993), and the inability of any current modeling procedure to recover from an incorrect alignment (Sanchez and Sali, 1997). The dispersion of the model-target structural overlap increases with the decrease in sequence identity. If the target-template sequence identity falls below 30%, the sequence identity becomes unreliable as a measure of accuracy of a single model. Models that deviate significantly from the average accuracy are frequent. It is in such cases that model-evaluation methods are particularly useful.

In addition to the target-template sequence similarity, the environment can strongly influence the accuracy of a model. For instance, some calcium-binding proteins undergo large conformational changes when bound to calcium. If a calcium-free template is used to model the calcium-bound state of the target, it is likely that the model will be incorrect, irrespective of the target-template similarity or accuracy of the template structure (Pawlowski et al., 1996). This qualification 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 (see Table 5.1.1 for URLs) are PROCHECK (Laskowski et al., 1998), PROCHECK-NMR (Laskowski et al., 1996), AQUA (Laskowski et al., 1996), SQUID (Oldfield, 1992), and WHATCHECK (Hooft et al., 1996b). The features of a model that these programs check include bond lengths, bond angles, peptide bond and side-chain ring planarities, chirality, main-chain and side-chain torsion angles, and clashes between nonbonded pairs of atoms.

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 (Sippl, 1990; Luthy et al., 1992). Programs implementing this approach (see Table 5.1.1 for URLs) include VERIFY3D (Luthy et al., 1992), PROSAII (Sippl, 1993), HARMONY (Topham et al., 1994), and ANOLEA (Melo and Feytmans, 1998). The programs evaluate the environment of each residue in a model with respect to the expected environment as found in the high-resolution X-ray structures. There is a concern about the theoretical validity of the energy profiles for detecting regional errors in models (Fiser et al., 2000). 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 model 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 (Miwa et al., 1999).

EXAMPLE OF COMPARATIVE MODELING: MODELING LACTATE DEHYDROGENASE FROM *TRICHOMONAS VAGINALIS* BASED ON A SINGLE TEMPLATE

This section contains an example of a typical comparative modeling application. It demonstrates each of the five steps of comparative modeling, using the program MODELLER6 (see Internet Resources). All files described in this section, including the MODELLER program, are available at *http://salilab.org/modeller/user_manual.shtml* (also see Internet Resources).

A novel gene for lactate dehydrogenase was identified from the genomic sequence of *Trichomonas vaginalis* (*TvLDH*). The corresponding protein had a higher similarity to the malate dehydrogenase of the same species (TvMDH) than to any other LDH. The authors hypothesized that TvLDH arose from TvMDH by convergent evolution relatively recently (Wu et al., 1999). Comparative models were constructed for TvLDH and TvMDH to study the sequences in the structural context and to suggest site-directed mutagenesis experiments for elucidating specificity changes in this apparent case of convergent evolution of enzymatic specificity. The native and mutated enzymes were expressed and their activities were compared (Wu et al., 1999).

Searching for Structures Related to TvLDH

It is necessary to put the target TvLDH sequence into the PIR format readable by MODELLER (see Internet Resources for MODELLER Web site; file `TvLDH.ali`). The first line of the file (see Fig. 5.1.4) contains the sequence code, in the format `>P1;code`. The second line, with ten fields separated by colons, generally contains information about the structure file, if applicable. Only two of these fields are used for sequences: `sequence` (indicating that the file contains a sequence without known structure) and `TvLDH` (the model file name). The rest of the file contains the sequence of TvLDH, with a “*” marking its end. A search for potentially related sequences of known structure can be performed by the `SEQUENCE_SEARCH` command of MODELLER. The following script uses the query sequence TvLDH assigned to the variable `ALIGN_CODES` from the file `TvLDH.ali` assigned to the variable `FILE` (file `seqsearch.top`; Fig. 5.1.5).

The `SEQUENCE_SEARCH` command has many options (see Internet Resources for MODELLER Web site), but in this example only `SEARCH_RANDOMIZATIONS` and `DATA_FILE` are set to non-default values. `SEARCH_RANDOMIZATIONS` specifies the number of times the query sequence is randomized during the calculation of the significance score for each sequence-sequence comparison. The higher the number of randomizations, the more accurate the significance score. `DATA_FILE = ON` triggers creation of an additional summary output file (`seqsearch.dat`).

Selecting a Template

The output of the `seqsearch.top` script is written to the `seqsearch.log` file. MODELLER always produces a log file. Errors and warnings in log files can be found by searching for the `E>` and `W>` strings, respectively. At the end of the log file, MODELLER lists the hits, sorted by alignment significance. Because the log file is sometimes

```
>P1;TvLDH
sequence:TvLDH::::::0.00: 0.00
MSEAAHVLITGAAGQIGYILSHWIASGELYG-DRQVYLHLLDIPPAMNRLTALTMELEDCAFPHLAGFVATTDPK
AAFKDIDCAFLVASMPLKPGQVRADLISSNSVIFKNTGEYLSKWAKPSVKVLVIGNPDNTNCEIAMLHAKNLKPE
NFSSLSMLDQNRAYYEVASKLGVDVKDVHDIIVWGNHGESMVADLTQATFTKEGKTQKVVDVLDHDYVFDTFFKK
IGHRAWDILEHRGFTSAASPTKAAIQHMKAWLFGTAPGEVLSMGIPVPEGNPYGIKPGVVFSFPCNVDKEGKIHV
VEGFKVNDWLREKLDFTEKDLFHEKEIALNHLAQGG*
```

Figure 5.1.4 File TvLDH.ali. Sequence file in the PIR MODELLER format.

```
SET SEARCH_RANDOMIZATIONS = 100
SEQUENCE_SEARCH FILE = 'TvLDH.ali', ALIGN_CODES = 'TvLDH', DATA_FILE = ON
```

Figure 5.1.5 File seqsearch.top. The TOP script file for template search using MODELLER.

```
 # CODE_1       CODE_2   LEN1 LEN2 NID  %IDI  %ID2    SCORE   SIGNI
 -------------------------------------------------------------------
  1 TvLDH       1bdmA     335  318 153  45.7  48.1  212557.   28.9
  2 TvLDH       1lldA     335  313 103  30.7  32.9  183190.   10.1
  3 TvLDH       1ceqA     335  304  95  28.4  31.3  179636.    9.2
  4 TvLDH       2hlpA     335  303  86  25.7  28.4  177791.    8.9
  5 TvLDH       1ldnA     335  316  91  27.2  28.8  180669.    7.4
  6 TvLDH       1hyhA     335  297  88  26.3  29.6  175969.    6.9
  7 TvLDH       2cmd      335  312 108  32.2  34.6  182079.    6.6
  8 TvLDH       1db3A     335  335  91  27.2  27.2  181928.    4.9
  9 TvLDH       91dtA     335  331  95  28.4  28.7  181720.    4.7
 10 TvLDH       1cdb      335  105  69  29.6  65.7   80141.    3.8
```

Figure 5.1.6 Top 10 hits from file seqsearch.dat. The summary output file for the SEQUENCE_SEARCH command in MODELLER.

very long, a separate data file (seqsearch.dat) is created that contains the summary of the search (Fig. 5.1.6). The example in Figure 5.1.6 shows only the top 10 hits from such a file.

The most important columns in the SEQUENCE_SEARCH output are the CODE_2, %ID, and SIGNI columns. The CODE_2 column reports the code of the PDB sequence that was compared with the target sequence. The PDB code in each line is the representative of a group of PDB sequences that share 40% or more sequence identity to each other and have less than 30 residues or 30% sequence length difference. All the members of the group can be found in the MODELLER CHAINS_3.0_40_XN.grp file. The LEN1 and LEN2 are lengths of the protein sequences in the CODE_1 and CODE_2 columns, respectively. NID represents the number of aligned residues. The %ID1 and %ID2 columns report the percentage sequence identities between TvLDH and a PDB sequence, normalized by their lengths, respectively. In general, a %ID value above ~25% indicates a potential template unless the alignment is short (i.e., less than 100 residues). A better measure of the significance of the alignment is given by the SIGNI column (see Internet Resources for MODELLER Web site). A value above 6.0 is generally significant irrespective of the sequence identity and length. In this example, one protein family represented by 1bdmA shows significant similarity with the target sequence, at more than 40%

```
SET ALIGN_CODES = '1bmdA' '4mdhA' '5mdhA' '7mdhA'
READ_ALIGNMENT FILE = '$(LIB)/CHAINS_all.seq'
MALIGN
MALIGN3D
COMPARE
ID_TABLE
DENDROGRAM
```

Figure 5.1.7 TOP script file `compare.top`.

sequence identity. While some other hits are also significant, the differences between 1bdmA and other top scoring hits are so pronounced that we use only the first hit as the template. As expected, 1bdmA is a malate dehydrogenase (from a thermophilic bacterium). Other structures closely related to 1bdmA (and thus not scanned against by `SEQUENCE_SEARCH`) can be extracted from the `CHAINS_3.0_40_XN.grp` file: 1b8vA, 1bmdA, 1b8uA, 1b8pA, 1bdmA, 1bdmB, 4mdhA, 5mdhA, 7mdhA, 7mdhB, and 7mdhC. All these proteins are malate dehydrogenases. During the project, all of them, as well as other malate and lactate dehydrogenase structures, were compared and considered as templates (there were 19 structures in total). However, for the sake of illustration, we will investigate only four of the proteins that are sequentially most similar to the target: 1bmdA, 4mdhA, 5mdhA, and 7mdhA. The script in Figure 5.1.7 performs all pairwise comparisons among the selected proteins (file `compare.top`).

The `READ_ALIGNMENT` command shown in Figure 5.1.7 reads the protein sequences and information about their PDB files. MALIGN calculates their multiple sequence alignment, used as the starting point for the multiple structure alignment. The `MALIGN3D` command performs an iterative least-squares superposition of the four 3-D structures. The `COMPARE` command compares the structures according to the alignment constructed by `MALIGN3D`. It does not make an alignment, but it calculates the RMS and DRMS deviations between atomic positions and distances, differences between the main-chain and side-chain dihedral angles, percentage sequence identities, and several other measures. Finally, the `ID_TABLE` command writes a file with pairwise sequence distances that can be used directly as the input to the `DENDROGRAM` command—or the clustering programs in the PHYLIP package (Felsenstein, 1985; *UNIT 6.3*). `DENDROGRAM` calculates a clustering tree from the input matrix of pairwise distances, which helps visualizing differences among the template candidates. Excerpts from the log file are shown in Figure 5.1.8 (file `compare.log`).

The comparison in Figure 5.1.8 shows that 5mdhA and 4mdhA are almost identical, both sequentially and structurally. They were solved at similar resolutions, 2.4 and 2.5 Å, respectively. However, 4mdhA has a better crystallographic R-factor (16.7 versus 20%), eliminating 5mdhA. Inspection of the PDB file for 7mdhA reveals that its crystallographic refinement was based on 1bmdA. In addition, 7mdhA was refined at a lower resolution than 1bmdA (2.4 versus 1.9 Å), eliminating 7mdhA. These observations leave only 1bmdA and 4mdhA as potential templates. Finally, 4mdhA is selected because of the higher overall sequence similarity to the target sequence.

Aligning TvLDF with the Template

A good way of aligning the sequence of TvLDH with the structure of 4mdhA is the `ALIGN2D` command in MODELLER. Although `ALIGN2D` is based on a dynamic-programming algorithm (Needleman and Wunsch, 1970; *UNIT 3.1*), it is different from standard sequence-sequence alignment methods because it takes into account structural informa-

```
>> Least-squares superposition (FIT)          :       T

   Atom types for superposition/RMS (FIT_ATOMS) : CA
   Atom type for position average/variability (DISTANCE_ATOMS[1]): CA

   Position comparison (FIT_ATOMS):

      Cutoff for RMS calculation:   3.5000

      Upper = RMS, Lower = numb equiv positions

           1bmdA    4mdhA    5mdhA    7mdhA
1bmdA      0.000    1.038    0.979    0.992
4mdhA        310    0.000    0.504    1.210
5mdhA        308      329    0.000    1.173
7mdhA        320      306      307    0.000

>> Sequence comparison:

      Diag=numb res, Upper=numb equiv res, Lower = % seq ID
            1bmdA    4mdhA    5mdhA    7mdhA
1bmdA         327      168      168      158
4mdhA          51      333      328      137
5mdhA          51       98      333      138
7mdhA          48       41       41      351
          .------------------------------------  1bmdA @1.9  49.0000
          |
          |                                  .--  4mdhA @2.5   2.0000
          |                                  |
      .--------------------------------------  5mdhA @2.4  55.5000
      |
    .----------------------------------------  7mdhA @2.4

    +--+--+--+--+---+--+--+--+--+--+--+--+
 57.6400   48.0100   38.3800   28.7500   19.1200   9.4900  -0.1400
   52.8250 43.1950 33.5650 23.9350 14.3050 4.6750
```

Figure 5.1.8 Excerpts from the log file `compare.log`.

tion from the template when constructing an alignment. This task is achieved through a variable gap penalty function that tends to place gaps in solvent-exposed and curved regions, outside secondary structure segments, and between two C_α positions that are close in space. As a result, the alignment errors are reduced by approximately one third relative to those that occur with standard sequence alignment techniques. This improvement becomes more important as the similarity between the sequences decreases and the number of gaps increases. In the current example, the template-target similarity is so high that almost any alignment method with reasonable parameters will result in the same alignment. The MODELLER script shown in Figure 5.1.9 aligns the TvLDH sequence in file `TvLDH.seq` with the 4mdhA structure in the PDB file `4mdh.pdb` (file `align2d.top`).

In the first line of the script shown in Figure 5.1.9, MODELLER reads the 4mdhA structure file. The `SEQUENCE_TO_ALI` command transfers the sequence to the alignment array and assigns it the name 4mdhA (`ALIGN_CODES`). The third line reads the TvLDH sequence from file `TvLDH.ali`, assigns it the name TvLDH (`ALIGN_CODES`), and adds it to the alignment array (`ADD_SEQUENCE = ON`). The fourth line executes the `ALIGN2D` command to perform the alignment. Finally, the alignment is written out in two formats, PIR (`TvLDH-4mdhA.ali`) and PAP (`TvLDH-4mdhA.pap`). The PIR

```
READ_MODEL FILE = '4mdh.pdb'
SEQUENCE_TO_ALI ALIGN_CODES = '4mdhA', ATOM_FILES = '4mdhA'
READ_ALIGNMENT FILE = 'TvLDH.ali', ALIGN_CODES = ALIGN_CODES 'TvLDH', ADD_SEQUENCE = ON
ALIGN2D
WRITE_ALIGNMENT FILE='TvLDH-4mdhA.ali', ALIGNMENT_FORMAT = 'PIR'
WRITE_ALIGNMENT FILE='TvLDH-4mdhA.pap', ALIGNMENT_FORMAT = 'PAP'
```

Figure 5.1.9 The `align2d.top` file that contains the MODELLER script executing the `ALIGN2D` command.

```
_aln.pos          10        20        30        40        50        60
4mdhA     GSEPIRVLVTGAAGQIAYSLLYSIGNGSVFGKDQPIILVLLDITPMMGVLDGVLMELQDCALPLLKDV
TvLDH     MSEAAHVLITGAAGQIGYILSHWIASGELYG-DRQVYLHLLDIPPAMNRLTALTMELEDCAFPHLAGF
_consrvd   **   ** ******* * *   *  *   * *    * **** * *  *    *** *** * *

_aln.p    70        80        90       100       110       120       130
4mdhA     IATDKEEIAFKDLDVAILVGSMPRRDGMERKDLLKANVKIFKCQGAALDKYAKKSVKVIVVGNPANTN
TvLDH     VATTDPKAAFKDIDCAFLVASMPLKPGQVRADLISSNSVIFKNTGEYLSKWAKPSVKVLVIGNPDNTN
_consrvd   **     **** * * ** ***   *  * **   *  ***  *  * * ** **** * *** ***

_aln.pos   140       150       160       170       180       190       200
4mdhA     CLTASKSAPSIPKENFSCLTRLDHNRAKAQIALKLGVTSDDVKNVIIWGNHSSTQYPDVNHAKVKLQA
TvLDH     CEIAMLHAKNLKPENFSSLSMLDQNRAYYEVASKLGVDVKDVHDIIVWGNHGESMVADLTQATFTKEG
_consrvd  *  *   *     **** *  ** ***    * ****   **   * ****      *   *
```

Figure 5.1.10 The alignment file between sequences TvLDH and 4mdhA in the PAP MODELLER format.

format is used by MODELLER in the subsequent model-building stage. The PAP alignment format is easier to inspect visually. Due to the high target-template similarity, there are only a few gaps in the alignment. In the PAP format, all identical positions are marked with a "`*`" (file `TvLDH-4mdhA.pap`; Fig. 5.1.10).

Model Building

Once a target-template alignment is constructed, MODELLER calculates a 3-D model of the target completely automatically. The script in Figure 5.1.11 will generate five models of TvLDH based on the 4mdhA template structure and the alignment in file `TvLDH-4mdh.ali` (file `model-single.top`). The first line (see Fig. 5.1.11) includes MODELLER variable and routine definitions. The following five lines set parameter values for the "model" routine. `ALNFILE` names the file that contains the target-template alignment in the PIR format. `KNOWNS` defines the known template structure(s) in `ALNFILE` (`TvLDH-4mdh.ali`). `SEQUENCE` defines the name of the target sequence in `ALNFILE`. `STARTING_MODEL` and `ENDING_MODEL` define the number of models that are calculated (their indices will run from 1 to 5). The last line in the file calls the "model" routine that actually calculates the models. The most important output files are (1) `model-single.log`, which reports warnings, errors, and other useful information including the input restraints used for modeling that remain violated in the final model, and (2) `TvLDH.B99990001`, `TvLDH.B9990002`, etc., which contain the model coordinates in the PDB format. The models can be viewed by any program that reads the PDB format, such as ModView (*http://guitar.rockefeller.edu/modview/*; Ilyin et al., 2003) or RasMol (*http://www.rasmol.org*; Sayle and Milner-White, 1995; Bernstein, 2000).

```
INCLUDE
SET ALNFILE = 'TvLDH-4mdhA.ali'
SET KNOWNS = '4mdhA'
SET SEQUENCE = 'TvLDH'
SET STARTING_MODEL = 1
SET ENDING_MODEL = 5
CALL ROUTINE = 'model'
```

Figure 5.1.11 TOP script file `model-single top`.

Evaluating a Model

If several models are calculated for the same target, the "best" model can be selected by picking the model with the lowest value of the MODELLER objective function, which is reported in the second line of the model PDB file. The value of the objective function in MODELLER is not an absolute measure; it can only be used to rank models calculated from the same alignment.

Once the final model is selected, there are many ways to assess it. Before any external evaluation of the model, one should check the log file from the modeling run for runtime errors (`model-single.log`) and restraint violations; see the MODELLER manual for details (at the MODELLER Web site listed in Internet Resources). Next, PROSAII (Sippl, 1993) is used to evaluate the model fold and PROCHECK (Laskowski et al., 1998) is used to check the model's stereochemistry. Both PROSAII and PROCHECK confirm that a reasonable model was obtained, with a PROSAII Z-score comparable to that of the template (-10.53 and -12.69 for the model and the template, respectively).

Additional detailed examples of MODELLER applications can be found in Fiser and Sali (2003).

APPLICATIONS

Comparative modeling is often an efficient way to obtain useful information about the protein of interest. For example, comparative models can be helpful in designing mutants to test hypotheses about the protein's function (Vernal et al., 2002; Wu et al., 1999); identifying active and binding sites (Sheng et al., 1996); searching for, designing, and improving ligand binding strength for a given binding site (Ring et al., 1993); modeling substrate specificity (Xu et al., 1996); predicting antigenic epitopes (Sali and Blundell, 1993); simulating protein-protein docking (Vakser, 1995); inferring function from calculated electrostatic potential around the protein (Matsumoto et al., 1995); facilitating molecular replacement in X-ray structure determination (Howell et al., 1992); refining models based on NMR constraints (Modi et al., 1996; Barrientos et al., 2001); testing and improving a sequence-structure alignment (Wolf et al., 1998); confirming a remote structural relationship (Guenther et al., 1997; Wu et al., 2000); and rationalizing known experimental observations. For a review of comparative modeling applications see Baker and Sali (2001) and Johnson et al. (1994).

Fortunately, a 3-D model does not have to be absolutely perfect to be helpful in biology, as demonstrated by the applications listed above. The type of a question that can be addressed with a particular model does depend on its accuracy (Fig. 5.1.3).

At the low end of the accuracy spectrum, there are models that are based on less than 25% sequence identity and have sometimes less than 50% of their C_α atoms within 3.5 Å of their correct positions. However, such models still have the correct fold, and even knowing only the fold of a protein may sometimes be sufficient to predict its approximate biochemical function. Models in this low range of accuracy ,combined with model evaluation, can be used for confirming or rejecting a match between remotely related proteins (Sanchez and Sali, 1997b, 1998).

In the middle of the accuracy spectrum are the models based on approximately 35% sequence identity, corresponding to 85% of the C_α atoms modeled within 3.5 Å of their correct positions. Fortunately, the active and binding sites are frequently more conserved than the rest of the fold, and are thus modeled more accurately (Sanchez and Sali, 1998). In general, medium-resolution models frequently allow a refinement of the functional prediction based on sequence alone, because ligand binding is most directly determined by the structure of the binding site rather than its sequence. It is frequently possible to correctly predict important features of the target protein that do not occur in the template structure. For example, the location of a binding site can be predicted from clusters of charged residues (Matsumoto et al., 1995), and the size of a ligand may be predicted from the volume of the binding site cleft (Xu et al., 1996). Medium-resolution models can also be used to construct site-directed mutants with altered or destroyed binding capacity, which in turn could test hypotheses about the sequence-structure-function relationships. Other problems that can be addressed with medium-resolution comparative models include designing proteins that have compact structures, without long tails, loops, and exposed hydrophobic residues, for better crystallization, 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 (3-Å resolution) or medium-resolution NMR structures (10 distance restraints per residue; (Sanchez and Sali, 1997b). The alignments on which these models are based generally contain almost no errors. In addition to the already listed applications, high-quality models can be used for docking of small ligands (Ring et al., 1993) or whole proteins onto the given protein (Totrov and Abagyan, 1994; Vakser, 1995).

A sample application of comparative modeling is the SNPWEB Web server for prediction of the functional effect of a single amino acid residue substitution (Table 5.1.1; Mirkovic et al., 2003). The server takes as input the specifications of the wild-type protein structure and a single amino acid residue substitution. The output is a prediction of whether or not the function of the mutant is impaired, as well as the rationalization of the predicted impact in terms of several features of the wild-type and mutant structures. The classification of the mutation as neutral or deleterious is achieved by a decision tree. The protocol is based on the assumption that a mutation is deleterious in either of the following two ways: (1) when it is exposed to the solvent, it may substantially change the structure or chemical nature of functional sites that bind other molecules; or (2) when it is buried in the core, it may prevent folding of the domains into their native fold, or, less likely, affect only the structure of functional sites. When applied to the case of the human BRCA1 domains, the server was able to rationalize 31 of 37 point mutations with a known functional impact (Mirkovic et al., 2003).

AUTOMATED COMPARATIVE PROTEIN STRUCTURE MODELING

As described earlier, each of the steps involved in the calculation of a protein structure model, namely fold assignment, sequence-structure alignment, model building, and model evaluation, can be completely automated (Sanchez and Sali, 1998; Marti-Renom et al., 2000). Despite limits on the accuracy of each one of these steps, they can be assembled into a software pipeline and applied to automatically model many protein sequences (Sanchez et al., 2000; Pieper et al., 2002). Such completely automated solutions are needed to complement genome sequencing and structural genomics (Sanchez et al., 2000; Vitkup et al., 2001). Comparative protein structure models have been calculated for individual genomes (Dubchak et al., 1998; Sanchez and Sali, 1998; Marti-Renom et al., 2000; Gough et al., 2001) and for all the sequences in the Swiss-Prot database (Peitsch, 1997; Pieper et al., 2002). One of the aims of structural genomics, to put every protein sequence within the modeling neighborhood of a given experimental structure, relies on automated techniques to pick targets for structure determination as well as to verify the modeling coverage after the structures are determined (Marti-Renom et al., 2000). Here, the authors of this unit describe their large-scale protein structure–modeling pipeline, called MODPIPE.

MODPIPE is a completely automated software pipeline for comparative protein structure modeling, which can calculate comparative models for a large number of protein sequences, using many different template structures and sequence-structure alignments (Sanchez and Sali, 1998; Marti-Renom et al., 2000; Narayanan et al., 2003) following the steps introduced above (see Example of Comparative Modeling). Sequence-structure matches are established by aligning the PSI-BLAST (Altschul et al., 1997) sequence profile of the target sequence against each of the template sequences extracted from the PDB, as well as by scanning the target sequence against a database of template profiles using IMPALA (Schaffer et al., 1999). Alignments with significant scores, covering distinct regions of the target sequence, are chosen for modeling. Models are calculated for each of the sequence-structure matches using MODELLER (Sali and Blundell, 1993). The resulting models are then evaluated by a composite model quality criterion (Sanchez and Sali, 1998; Melo et al., 2002).

The thoroughness of a search for the best model is modulated by a number of parameters, including two E-value thresholds for identifying useful sequence-structure relationships and the degree of conformational sampling given a sequence-structure alignment. The validity of sequence-structure relationships is not prejudged at the fold-detection stage, but is assessed after the construction of the model and its evaluation. This approach enables a thorough exploration of fold assignments, sequence-structure alignments, and conformations, with the aim of finding the model with the best evaluation score.

Comparative models in MODPIPE are evaluated by a composite scoring function given by (Melo et al., 2002):

$$\mathrm{GA341} = 1 - [\cos(\mathit{sequence\_identity})]^{(\mathit{compactness}\,+\,\mathit{sequence\_identity})/\exp(\mathit{z\text{-}score})}$$

where *sequence_identity* is the fraction of positions with identical residues in the target-template alignment; *compactness* of the model is defined as the ratio between the sum of the standard volumes of the amino acid residues in the protein and the volume of the sphere with the radius equal to half of the largest dimension of the model; and, the *z-score* is calculated for the combined statistical potential energy of a model, using the mean and standard deviation of the 200 random sequences with the same composition and structure as the model (Melo et al., 2002). The combined statistical potential energy of a model is a sum of the solvent accessibility terms for all C_β atoms and distance-dependent terms for all pairs of C_α and C_β atoms. The solvent accessibility term for a given C_β atom depends on its residue type and the number of other C_β atoms within 10 Å; the nonbonded distance-dependent terms depend on the atom types spanning the distance, the distance itself, and the number of residues separating the distance-spanning atoms in sequence.

These potential terms reflect the statistical preferences observed in 760 non-redundant proteins of known structure. The GA341 scoring function was evolved by a genetic algorithm that explored many combinations of a variety of mathematical functions and model features, in order to optimize the discrimination between good and bad models in a training set of models.

The GA341 score ranges from 0 for models that tend to have an incorrect fold to 1 for models that tend to be comparable to low-resolution X-ray structures. Comparison of models with their corresponding experimental structures indicates that models with GA341 scores >0.7 generally have the correct fold with more than 35% of the backbone atoms superposable within 3.5 Å. Reliable models (GA341 score ≥ 0.7) based on alignments with more than 40% sequence identity, have a median overlap of more than 90% with the corresponding experimental structure. In the 30% to 40% sequence identity range, the overlap is usually between 75% to 90% and below 30% it drops to 50% to 75%, or even less in the worst cases.

Data storage and retrieval of the models calculated by MODPIPE is enabled by MODBASE, a comprehensive relational database of annotated comparative protein structure models (Pieper et al., 2002). MODBASE contains several model data sets, including that for all available protein sequences matched to at least one known protein structure. This data set was calculated by applying MODPIPE to all sequences in the Swiss-Prot database (March, 2002). Currently, MODBASE contains models for domains in 415,937 out of 733,239 (~57%) unique protein sequences found in Swiss-Prot.

MODBASE is queryable through its Web user interface by PDB codes, Swiss-Prot (Boeckmann et al., 2003) and GENPEPT (Benson et al., 2002) accession numbers, open reading frame names, various keywords, model reliability, model size, target-template sequence identity, alignment significance, and sequence similarity against the modeled sequences as detected by BLAST (*UNITS 3.3 & 3.4*). It is also possible to query the database directly using SQL as implemented in MySQL.

The output of a search is displayed on pages with varying amounts of information about the modeled sequences, template structures, alignments, and functional annotations. These tables also contain links to other sequence, structure and function annotation databases, such as PDB (Berman et al., 2002), GenBank (Benson et al., 2002), Swiss-Prot (Boeckmann et al., 2003), CATH (Pearl et al., 2003), Pfam (Bateman et al., 2002), and ProDom (Servant et al., 2002).

STRUCTURAL GENOMICS AND COMPARATIVE MODELING

The complete genomes of a number of organisms have been sequenced and many more genome-sequencing projects are underway. Structural biology now faces the arduous task of characterizing the shapes and dynamics of the encoded proteins to facilitate the understanding of their functions and mechanisms of action. Recent developments in the techniques of structure determination at atomic resolution, X-ray diffraction and nuclear magnetic resonance spectroscopy, have enhanced the quality and speed of structural studies (Zhang and Kim, 2003). Nevertheless, current statistics still show that the known protein sequences (~1,000,000; Boeckmann et al., 2003) vastly outnumber the available protein structures (~20,000; Berman et al., 2002). Fortunately, domains in protein sequences are gradually evolving entities that can be clustered into a relatively small number of families of domains with similar sequences and structures (i.e., folds; Vitkup et al., 2001). These evolutionary relationships make it possible to use computational methods, such as threading (Domingues et al., 2000) and comparative protein structure modeling (Blundell et al., 1987; Marti-Renom et al., 2000) to predict the structures of protein sequences based on their similarity to known protein structures.

Many structural genomics efforts, in fact, combine the experimental structure determination methods and the computational modeling techniques to determine a sufficient number of appropriately selected structures, so that most other sequences can be placed within modeling distance of at least one known structure. To maximize the number of proteins that can be modeled reliably, a concerted effort toward structure determination of new folds by X-ray crystallography and nuclear magnetic resonance spectroscopy is in order, as envisioned by structural genomics (Sali, 1998; Terwilliger et al., 1998; Zarembinski et al., 1998; Burley et al., 1999; Montelione and Anderson, 1999; Sanchez et al., 2000; Vitkup et al., 2001). It has been estimated that 90% of all globular and membrane proteins can be organized into approximately 16,000 families containing protein domains with more than 30% sequence identity to each other (Vitkup et al., 2001). Of these families, 4000 are already structurally defined; the others present suitable targets for structural genomics. The full potential of the genome-sequencing projects will only be realized once all protein functions are assigned and understood. This aim will be facilitated by integrating genomic sequence information with databases arising from functional and structural genomics. Comparative modeling will play an important bridging role in these efforts.

CONCLUSION

Over the past few years, there has been a gradual increase in both the accuracy of comparative models and the fraction of protein sequences that can be modeled with useful accuracy (Marti-Renom et al., 2000; Baker and Sali, 2001; Pieper et al., 2002). The magnitude of errors in fold assignment, alignment, and the modeling of side-chains and loops has decreased measurably. These improvements are a consequence both of better techniques and a larger number of known protein sequences and structures. Nevertheless, all the errors remain significant and demand future methodological improvements. In addition, there is a great need for more accurate modeling of distortions and rigid body shifts, as well as detection of errors in a given protein structure model. Error detection is useful both for refinement and interpretation of the models.

ACKNOWLEDGEMENTS

We are grateful to all members of our research group. Andrej Sali is an Irma T. Hirschl Trust Career Scientist. Support by The Merck Genome Research Institute, Mathers Foundation, and NIH (GM 54762 and GM 62529) is also acknowledged. This review is partially based on our previous reviews (Marti-Renom et al., 2002b; Marti-Renom et al., 2003; Narayanan et al., 2003).

LITERATURE CITED

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.

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.

Altschul, S.F., Boguski, M.S., Gish, W., and Wootton, J.C. 1994. Issues in searching molecular sequence databases. *Nat. Genet.* 6:119-129.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.

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

Aszodi, A. and Taylor, W.R. 1994. Secondary structure formation in model polypeptide chains. *Protein Eng.* 7:633-644.

Aszodi, A. and Taylor, W.R. 1996. Homology modelling by distance geometry. *Fold. Des.* 1:325-334.

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

Bajorath, J., Stenkamp, R., and Aruffo, A. 1993. Knowledge-based model building of proteins: Concepts and examples. *Protein Sci.* 2:1798-1810.

Baker, D. 2000. A surprising simplicity to protein folding. *Nature* 405:39-42.

Baker, D. and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* 294:93-96.

Barrientos, L.G., Campos-Olivas, R., Louis, J.M., Fiser, A., Sali, A., and Gronenborn, A.M. 2001. ^{1}H, ^{13}C, ^{15}N resonance assignments and fold verification of a circular permuted variant of the potent HIV-inactivating protein cyanovirin-N. *J. Biomol. NMR* 19:289-290.

Barton, G.J. 1998. Protein sequence alignment techniques. *Acta Crystallogr. D. Biol. Crystallogr.* 54:1139-1146.

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.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res.* 30:276-280.

Baxevanis, A.D. 1998. Practical aspects of multiple sequence alignment. *Methods Biochem. Anal* 39:172-188.

Beckmann, R., Spahn, C.M.T., Eswar, N., Helmers, J., Penczek, P.A., Sali, A., Frank, J., and Bloebel G. 2001. Architecture of the protein-conducting channel associated with the translating 80S ribosome. *Cell* 107:361-72.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L. 2002. GenBank. *Nucleic Acids Res.* 30:17-20.

Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D., and Zardecki, C. 2002. The Protein Data Bank. *Acta. Crystallogr. D. Biol. Crystallogr.* 58:899-907.

Bernstein, H.J. 2000. Recent changes to RasMol, recombining the variants. *Trends Biochem. Sci.* 25:453-455.

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.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31:365-370.

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* 268:15983-15993.

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

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.

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.

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.

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. U.S.A.* 95:6073-6078.

Brenner, S.E., Barken, D., and Levitt, M. 1999. The PRESAGE database for structural genomics. *Nucleic Acids Res.* 27:251-253.

Briffeuil, P., Baudoux, G., Lambert, C., De, Bolle, X., Vinals, C., Feytmans, E., and Depiereux, E. 1998. Comparative analysis of seven multiple protein sequence alignment servers: Clues to enhance reliability of predictions. *Bioinformatics* 14:357-366.

Brocklehurst, S.M. and Perham, R.N. 1993. Prediction of the three-dimensional structures of the biotinylated domain from yeast pyruvate carboxylase and of the lipoylated H-protein from the pea leaf glycine cleavage system: A new automated method for the prediction of protein tertiary structure. *Protein Sci.* 2:626-639.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Waminathan, S.S., and Karplus, M. 1983. CHARMM: A program for macromolecular energy minimization and dynamics calculations. *J. Comp. Chem.* 4:187-217.

Browne, W.J., North, A.C.T., Phillips, D.C., Brew, K., Vanaman, T.C., and Hill, R.C. 1969. A possible three-dimensional structure of bovine lactalbumin based on that of hen's egg-white lysosyme. *J. Mol. Biol.* 42:65-86.

Bruccoleri, R.E. and Karplus, M. 1990. Conformational sampling using high-temperature molecular dynamics. *Biopolymers* 29:1847-1862.

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.

Burley, S.K., Almo, S.C., Bonanno, J.B., Capel, M., Chance, M.R., Gaasterland, T., Lin, D., Sali, A., Studier, F.W., and Swaminathan, S. 1999. Structural genomics: Beyond the human genome project. *Nat. Genet* 23:151-157.

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.

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.

Chothia, C. and Lesk, A.M. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5:823-826.

Chothia, C. and Lesk, A.M. 1987. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* 196:901-917.

Chothia, C., Lesk, A.M., Tramontano, A., Levitt, M., Smith-Gill, S.J., Air, G., Sheriff, S., Padlan, E.A., Davies, D., and Tulip, W.R. 1989. Conformations of immunoglobulin hypervariable regions. *Nature* 342:877-883.

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.

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.

Cohen, F.E. and Kuntz, I.D. 1989. Tertiary structure prediction. *In* Prediction of protein structure and the principles of protein conformations. (G.D. Fasman, ed.) pp. 647-705. Plenum, New York.

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

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

Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* 16:10881-10890.

David, R., Korenberg, M.J., and Hunter, I.W. 2000. 3D-1D threading methods for protein fold recognition. *Pharmacogenomics* 1:445-455.

Dainese, E., Minafra, R., Sabatucci, A., Vachette, P., Melloni, E., and Cozzani, I. 2002. Conformational changes of calpain from human erythrocytes in the presence of Ca^{2+}. *J. Biol. Chem.* 277:40296-40301.

Dayhoff, M.O. and Eck, R.V. 1968. Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Spring, Md.

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.

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.

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.

Domingues, F.S. and Koppensteiner, W.A., and Sippl, M.J. 2000. The role of protein structure in genomics. *FEBS. Lett.* 476:98-102.

Dubchak, I., Muchnik, I., and Kim, S.H. 1998. Assignment of folds for proteins of unknown function in three microbial genomes. *Microb. Comp. Genomics* 3:171-175.

Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* 14:755-763.

Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Fiser, A., Pazos, F., Valencia, A., Sali, A., and Rost, B. 2001. EVA: Continuous automatic evaluation of protein structure prediction servers. *Bioinformatics* 17:1242-1243.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783-791.

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.

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.

Fine, R.M., Wang, H., Shenkin, P.S., Yarmush, D.L., and Levinthal, C. 1986. Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. *Proteins* 1:342-362.

Fischer, D. and Eisenberg, D. 1996. Protein fold recognition using sequence-derived predictions. *Protein Sci.* 5:947-955.

Fischer, D. and Eisenberg, D. 1997. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc. Natl. Acad. Sci. U.S.A.* 94:11929-11934.

Fischer, D., Elofsson, A., Rychlewski, L., Pazos, F., Valencia, A., Rost, B., Ortiz, A.R., and Dunbrack, R.L., Jr. 2001. CAFASP2: The second critical assessment of fully automated structure prediction methods. *Proteins* 45(5):171-183.

Fiser, A. and Sali, A. 2003. Comparative protein structure modeling with Modeller: A practical approach. *Methods Enzymol.* In press.

Fiser, A., Do, R.K., and Sali, A. 2000. Modeling of loops in protein structures. *Protein Sci.* 9:1753-1773.

Flockner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M., and Sippl, M.J. 1995. Progress in fold recognition. *Proteins* 23:376-386.

Gerstein, M. 1998. Measurement of the effectiveness of transitive sequence comparison, through a third intermediate sequence. *Bioinformatics* 14:707-714.

Gerstein, M. and Levitt, M. 1997. A structural census of the current population of protein sequences. *Proc. Natl. Acad. Sci. U.S.A* 94:11911-11916.

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

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.

Gowri, V.S., Pandit, S.B., Karthik, P.S., Srinivasan, N., and Balaji, S. 2003. Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. *Nucleic Acids Res.* 31:486-488.

Greer, J. 1990. Comparative modeling methods: Application to the family of the mammalian serine proteases. *Proteins* 7:317-334.

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

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

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.

Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* 89:10915-10919.

Henikoff, S., Henikoff, J.G., Alford, W.J., and Pietrokovski, S. 1995. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene* 163:GC17-GC26.

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.

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.

Holm, L. and Sander, C. 1996. Mapping the protein universe. *Science* 273:595-603.

Holm, L. and Sander, C. 1999. Protein folds and families: Sequence and structure alignments. *Nucleic Acids Res.* 27:244-247.

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

Hooft, R.W., Sander, C., and Vriend, G. 1996b. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins* 26:363-376.

Howell, P.L., Almo, S.C., Parsons, M.R., Hajdu, J., and Petsko, G.A. 1992. Structure determination of turkey egg-white lysozyme using Laue diffraction data. *Acta Crystallogr. B.* 48:200-207.

Huynen, M., Doerks, T., Eisenhaber, F., Orengo, C., Sunyaev, S., Yuan, Y., and Bork, P. 1998. Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.* 280:323-326.

Ilyin, V.A., Pieper, U., Stuart, A.C., Marti-Renom, M.A., McMahan, L., and Sali, A. 2003. Visualization and analysis of multiple protein sequences and structures by ModView. *Bioinformatics* 19:165-166.

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

John, B. and Sali, A. 2003. Comparative protein structure modeling by iterative alignment, model building, and model assessment. *Nucleic Acids Res.* 31:3982-3992.

Johnson, M.S. and Overington, J.P. 1993. A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J. Mol. Biol.* 233:716-738.

Johnson, M.S., Srinivasan, N., Sowdhamini, R., and Blundell, T.L. 1994. Knowledge-based protein modeling. *Crit. Rev. Biochem. Mol. Biol.* 29:1-68.

Jones, D.T. 1997. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins* 1:185-191.

Jones, D.T. 1999. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* 287:797-815.

Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. A new approach to protein fold recognition. *Nature* 358:86-89.

Jones, T.A. and Thirup, S. 1986. Using known substructures in protein model building and crystallography. *EMBO J.* 5:819-822.

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. U.S.A.* 81:1075-1078.

Karplus, K., Barrett, C., and Hughey, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14:846-856.

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.

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.

Koh, I-Y., Y., Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Narayanan, E., Graña, O., Pazos, F., Valencia, A., Sali, A., and Rost, B. 2003. EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res.* 31:3311-3315.

Kolinski, A., Betancourt, M.R., Kihara, D., Rotkiewicz, P., and Skolnick, J. 2001. Generalized comparative modeling (GENECOMP): A combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins* 44:133-149.

Laskowski, R.A., Rullmannn, 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.

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.

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

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

Levitt, M. 1997. Competitive assessment of protein fold recognition and alignment accuracy. *Proteins* 1:92-104.

Lindahl, E. and Elofsson, A. 2000. Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* 295:613-625.

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.

Luthy, R., Bowie, J.U., and Eisenberg, D. 1992. Assessment of protein models with three-dimensional profiles. *Nature* 356:83-85.

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

Martelli, P.L., Fariselli, P., Krogh, A., and Casadio, R. 2002. A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics* 1:S46-S53.

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

Marti-Renom, M.A., Madhusudhan, M.S., Fiser, A., Rost, B., and Sali, A. 2002a. Reliability of assessment of protein structure prediction methods. *Structure* 10:435-440.

Marti-Renom, M.A., Yerkovich, B., and Sali, A. 2002b. Comparative protein structure prediction. *In* Current Protocols in Protein Science (J.E. Coligan, B.M. Dunn, D.W. Speicher, and P.T. Wingfield, eds.) pp. 2.9.1-2.9.22. John Wiley & Sons, New York.

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.* 270:19524-19531.

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

Melo, F., Sanchez, R., and Sali, A. 2002. Statistical potentials for fold assessment. *Protein Sci.* 11:430-448.

Mezei, M. and Guarnieri, F. 1998. Computer simulation studies of the fully solvated wild-type and mutated GnRH in extended and beta-turn conformations. *J. Biomol. Struct. Dyn.* 16:723-732.

Mirkovic, N., Marti-Renom, M.A., Sali, A., and Monteiro, A.N.A. 2003. Structure-based assessment of missense mutations in human BRCA1: Implications for breast and ovarian cancer predisposition. *Proc. Natl. Acad. Sci. U.S.A.* Submitted for publication.

Miwa, J.M., Ibanez-Tallon, I., Crabtree, G.W., Sanchez, R., Sali, A., Role, L.W., and Heintz, N. 1999. lynx1, an endogenous toxin-like modulator of nicotinic acetylcholine receptors in the mammalian CNS. *Neuron* 23:105-114.

Modi, S., Paine, M.J., Sutcliffe, M.J., Lian, L.Y., Primrose, W.U., Wolf, C.R., and Roberts, G.C. 1996. A model for human cytochrome P450 2D6 based on homology modeling and NMR studies of substrate binding. *Biochemistry* 35:4540-4550.

Montelione, G.T. and Anderson S. 1999. Structural genomics: Keystone for a Human Proteome Project. *Nat. Struct. Biol.* 6:11-12.

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.

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

Muller, A., MacCallum, R.M., and Sternberg, M.J. 1999. Benchmarking PSI-BLAST in genome annotation. *J. Mol. Biol.* 293:1257-1271.

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.

Narayanan, E., John, B., Mirkovic, N., Fiser, A., Ilyin, V. A., Pieper, U., Stuart, A., Marti-Renom, M.A., Madhusudhan, M.S., Yerkovich, B., and Sali, A. 2003. Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.* 31:3375-3380.

Navaratnam, N., Fujino, T., Bayliss, J., Jarmuz, A., How, A., Richardson, N., Somasekaram, A., Bhattacharya, S., Carter, C., and Scott, J. 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.

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.

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.

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.

Orengo, C.A., Bray, J.E., Buchan, D.W., Harrison, A., Lee, D., Pearl, F.M., Sillitoe, I., Todd, A.E., and Thornton, J.M. 2002. The CATH protein family database: A resource for structural and functional annotation of genomes. *Proteomics* 2:11-21.

Panchenko, A.R., Marchler-Bauer, A., and Bryant, S.H. 2000. Combination of threading potentials and sequence profiles improves fold recognition. *J. Mol. Biol.* 296:1319-1331.

Park, J., Teichmann SA, Hubbard, T, Chothia C. 1997. Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.* 273:349-354.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* 284:1201-1210.

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

Pearl, F.M., Bennett, C.F., Bray, J.E., Harrison, A.P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J., and Orengo, C.A. 2003. The CATH database: An extended protein family resource for structural and functional genomics. *Nucleic Acids Res.* 31:452-455.

Pearl, F.M., Lee, D., Bray, J.E., Buchan, D.W., Shepherd, A.J., and Orengo, C.A. 2002. The CATH extended protein-family database: Providing structural annotations for genome sequences. *Protein Sci.* 11:233-244.

Pearson, W.R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* 183:63-98.

Pearson, W.R. 1995. Comparison of methods for searching protein sequence databases. *Protein Sci.* 4:1145-1160.

Pearson, W.R. and Lipman D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* 85:2444-2448.

Peitsch, M.C. 1997. Large scale protein modelling and model repository. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 5:234-236.

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

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

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.

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.

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.

Ring, C.S., Sun, E., McKerrow, J.H., Lee, G.K., Rosenthal, P.J., Kuntz, I.D., and Cohen, F.E. 1993. Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc. Natl. Acad. Sci. U.S.A.* 90:3583-3587.

Rost, B. 1995. TOPITS: Threading one-dimensional predictions into three-dimensional structures. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3:314-321.

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

Rost, B. and Sander, C. 1995. Progress of 1D protein structure prediction at last. *Proteins* 23:295-300.

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.

Rychlewski, L., Zhang, B., and Godzik, A. 1998. Fold and function predictions for *Mycoplasma genitalium* proteins. *Fold. Des.* 3:229-238.

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.

Sali, A. 1995. Comparative protein modeling by satisfaction of spatial restraints. *Mol. Med. Today* 1:270-277.

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

Sali, A. and Blundell T.L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234:779-815.

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.

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.

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

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

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

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

Sanchez, R., Pieper, U., Melo, F., Eswar, N., Marti-Renom, M.A., Madhusudhan, M.S., Mirkovic, N., and Sali, A. 2000. Protein structure modeling for structural genomics. *Nat. Struct. Biol.* 7:986-990.

Saqi, M.A., Russell, R.B., and Sternberg, M.J. 1998. Misleading local sequence alignments: Implications for comparative protein modelling. *Protein Eng.* 11:627-630.

Sauder, J.M., Arthur, J.W., and Dunbrack, R.L., Jr. 2000. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins* 40:6-22.

Sayle, R..A. and Milner-White, E.J. 1995. RASMOL: Biomolecular graphics for all. *Trends Biochem. Sci.* 20:374.

Schaffer, A.A., Wolf, Y.I., Ponting, C.P., Koonin, E.V., Aravind, L., and Altschul, S.F. 1999. IMPALA: Matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* 15:1000-1011.

Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D., and Kahn, D. 2002. ProDom: Automated clustering of homologous domains. *Brief Bioinform.* 3:246-251.

Shan, Y., Wang, G., and Zhou, H.X. 2001. Fold recognition and accurate query-template alignment by a combination of PSI-BLAST and threading. *Proteins* 42:23-37.

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

Shenkin, 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.

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.

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.

Sippl, M.J. 1993. Recognition of errors in three-dimensional structures of proteins. *Proteins* 17:355-362.

Smith, R.F., Wiese, B.A., Wojzynski, M.K., Davison, D.B., and Worley, K.C. 1996. BCM Search Launcher: An integrated interface to molecular biology data base search and analysis services available on the World Wide Web. *Genome Res.* 6:454-462.

Smith, T.F. 1999. The art of matchmaking: Sequence alignment methods and their structural implications. *Structure Fold Des.* 7:R7-R12.

Smith, T.F. and Waterman, M.S. 1981. Overlapping genes and information theory. *J. Theor. Biol.* 91:379-380.

Smith, T.F., Lo, Conte, L., Bienkowska, J., Gaitatzes, C., Rogers, R.G., Jr., and Lathrop, R. 1997. Current limitations to protein threading approaches. *J. Comput. Biol.* 4:217-225.

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.

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.

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.

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.

Taylor, W.R. 1996. Multiple protein sequence alignment: Algorithms and gap insertion. *Methods Enzymol.* 266:343-367.

Taylor, W.R., Flores, T.P., and Orengo, C.A. 1994. Multiple protein structure alignment. *Protein Sci.* 3:1858-1870.

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.

Terwilliger, T.C., Waldo, G., Peat, T.S., Newman, J.M., Chu, K., and Berendzen, J. 1998. Class-directed structure determination: Foundation for a protein structure initiative. *Protein Sci.* 7:1851-1856.

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.

Thompson, J.D., Plewniak, F., and Poch, O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* 27:2682-2690.

Topham, C.M., McLeod, A., Eisenmenger, F., Overington, J.P., Johnson, M.S., and Blundell, T.L. 1993. Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. *J. Mol. Biol.* 229:194-220.

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.

Torda, A.E. 1997. Perspectives in protein-fold recognition. *Curr. Opin. Struct. Biol.* 7:200-205.

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

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.

Vakser, I.A. 1995. Protein docking for low-resolution structures. *Protein Eng.* 8:371-377.

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.

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.

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

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

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

Westbrook, J., Feng, Z., Jain, S., Bhat, T.N., Thanki, N., Ravichandran, V., Gilliland, G.L., Bluhm, W., Weissig, H., Greer, D.S., Bourne, P.E., and Berman, H.M. 2002. The protein data bank: Unifying the archive. *Nucleic Acids Res.* 30:245-248.

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.

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

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

Wu, G., McArthur, A.G., Fiser, A., Sali, A., Sogin, M.L., and Mllerm, M. 2000. Core histones of the amitochondriate protist, *Giardia lamblia*. *Mol. Biol. Evol.* 17:1156-1163.

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. U.S.A.* 99:7432-7437.

Xu, L.Z., Sanchez, R., Sali, A., and Heintz, N. 1996. Ligand specificity of brain lipid-binding protein. *J. Biol. Chem.* 271:24711-24719.

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.

Zarembinski, T.I., Hung, L.W., Mueller-Dieckmann, H.J., Kim, K.K., Yokota, H., Kim, R., and Kim, S.H. 1998. Structure-based assignment of the biochemical function of a hypothetical protein: A test case of structural genomics. *Proc. Natl. Acad. Sci. U.S.A.* 95:15189-15193.

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

Zhang, C. and Kim, S.H. 2003. Overview of structural genomics: from structure to function. *Curr. Opin. Chem. Biol.* 7:28-32.

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

INTERNET RESOURCES

http://www.salilab.org/modeller

> *MODELLER, A Protein Structure Modeling Program, Release 6v2. Sali, A., Fiser, A., Sanchez, R., Marti-Renom, M.A., Jerkovic, B., Badretdinov, A., Melo, F., Overington, J., and Feyfant, E. 2001.*

Table 5.1.1 lists a large number of additional Web sites useful in comparative modeling.

Contributed by Marc A. Marti-Renom, M.S. Madhusudhan, Narayanan Eswar, Ursula Pieper, Min-yi Shen, and Andrej Sali
Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry and The California Institute for Quantitative Biomedical Research
University of California at San Francisco
San Francisco, California

Andras Fiser
Department of Biochemistry and Seaver Foundation Center for Bioinformatics
Albert Einstein College of Medicine
Bronx, New York

Bino John and Ashley Stuart
Laboratory of Molecular Biophysics
The Rockefeller University
New York, New York

Figure 5.1.1 Steps in comparative protein structure modeling. See text for details.

Figure 5.1.2 Accuracy and application 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 (van der Waals sphere model) 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 (Xu et al., 1996), even though the ligand-specificity profile of this protein is different from that of the template structure. Typical overall accuracy of a comparative model in this range of sequence similarity is indicated by a comparison of a model for adipocyte fatty acid binding protein with its actual 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 (PDB code, 2ptn) that does not bind proteoglycans. The prediction was confirmed by site-directed mutagenesis and heparin-affinity chromatography experiments (Matsumoto et al., 1995). 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 15 Å resolution (Beckmann et al., 2001). Most of the models for 40 out of the 75 ribosomal proteins were based on approximately 30% sequence identity to their template structures. 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 *B. Stearothermophilus* with the actual structure (PDB code, 1rl2).

Figure 5.1.3 Typical errors in comparative modeling. *(*A**)**. Errors in side-chain packing. The Trp 109 residue in the crystal structure of mouse cellular retinoic acid binding protein I (thin line) is compared with its model (thick line), and with the template mouse adipocyte lipid-binding protein (broken line). (**B**) Distortions and shifts in correctly aligned regions. A region in the crystal structure of mouse cellular retinoic acid binding protein I is compared with its model and with the template fatty acid binding protein using the same representation as in panel A. (**C**) Errors in regions without a template. The C_α trace of the 112–117 loop is shown for the X-ray structure of human eosinophil neurotoxin (thin line), its model (thick line), and the template ribonuclease A structure (residues 111–117; broken line). (**D**) Errors due to misalignments. The N-terminal region in the crystal structure of human eosinophil neurotoxin (thin line) is compared with its model (thick line). The corresponding region of the alignment with the template ribonuclease A is shown. The black lines show correct equivalences, that is, residues whose C_α atoms are within 5 Å of each other in the optimal least-squares superposition of the two X-ray structures. The “a” characters in the bottom line indicate helical residues. (**E**) Errors due to an incorrect template. The X-ray structure of α-trichosanthin (thin line) is compared with its model (thick line) which was calculated using indole-3-glycerophosphate synthase as the template.

Figure 5.1.4 File `TvLDH.ali`. Sequence file in the PIR MODELLER format.

Figure 5.1.5 File `seqsearch.top`. Sequence file in the `PIR MODELLER` format. The `TOP` script file for template search using `MODELLER`.

Figure 5.1.6 op 10 hits from file `seqsearch.dat`. The summary output file for the `SEQUENCE_SEARCH` command in MODELLER.

Figure 5.1.7 TOP script file `compare.top`.

Figure 5.1.8 Excerpts from the log file `compare.log`.

Figure 5.1.9 The `ALIGN2D.top` file that contains the MODELLER script executing the `ALIGN2D` command.

Figure 5.1.10 The alignment file between sequences TvLDH and 4mdhA in the PAP MODELLER format.

Figure 5.1.11 TOP script file `model-single.top`.