6

Comparative Protein Structure Modeling

Introduction and Practical Examples with Modeller

Roberto Sánchez and Andrej Šali

1. Introduction

1.1. What is Comparative Protein Structure Modeling?

A useful three-dimensional (3D) model for a protein of unknown structure (the target) can frequently be built based on one or more related proteins of known structure (the templates). This is the aim of comparative or homology protein structure modeling. The necessary conditions are that the similarity between the target sequence and the template structures is detectable and that the correct alignment between them can be constructed. For reviews of comparative modeling, *see* **refs. *1–5***. This approach to structure prediction is possible because a small change in the protein sequence usually results in a small change in its 3D structure ***(6,7)***.

1.2. Why is Comparative Modeling Useful?

The biochemical function of a protein is defined by its interactions with other molecules and the biological function is a consequence of these interactions. Although protein function is best determined experimentally ***(8)***, it can sometimes be predicted by matching the sequence of a protein with proteins of known function ***(8–10)***. One way to improve sequence-based predictions of function is to rely on the known native 3D structure of proteins. The 3D structure of a protein generally provides more information about its function than sequence because interactions of a protein with other molecules are determined by amino acid residues that are close in space but are frequently distant in sequence. For example, several mouse mast cell proteases have a conserved surface region of positively charged residues that binds proteoglycans ***(11)***.

From: *Methods in Molecular Biology*, vol. 143: *Protein Structure Prediction: Methods and Protocols*
Edited by: D. Webster

Table 1
Common Uses of Comparative Protein Structure Models

Designing (site-directed) mutants to test hypotheses about function
Identifying active and binding sites
Searching for ligands of a given binding site
Designing and improving ligands of a given binding site
Modeling substrate specificity
Predicting antigenic epitopes
Protein–protein docking simulations
Inferring function from calculated electrostatic potential around the protein
Molecular replacement in X-ray structure refinement
Testing a given sequence–structure alignment
Rationalizing known experimental observations
Planning new experiments

This region is not easily recognizable in sequence because the constituting residues occur at variable and sequentially nonlocal positions that form a binding site only when the protease is fully folded.

Comparative modeling remains the only method that can reliably predict the 3D structure of a protein with an accuracy comparable to that of low-resolution experimental structures *(1)*. Even such low resolution models are useful to address biological questions, because function can sometimes be predicted from only coarse structural features of a model. Typical uses of comparative models are listed in **Table 1**. For a review of comparative modeling applications *see* **refs.** *2* and *3*.

Three-dimensional structure of proteins from the same family is more conserved than their sequences *(12)*. Therefore, if similarity between two proteins is detectable at the sequence level, structural similarity can usually be assumed. Moreover, proteins that share low or even nondetectable sequence similarity many times also have similar structures. It has been estimated that approximately one third of all sequences are related to at least one protein of known structure *(13)*. Because there are approx 450,000 known protein sequences *(14)*, comparative modeling could, in principle, be applied to approx 150,000 proteins. This is an order of magnitude more proteins than the number of experimentally determined protein structures (approx 10,000) *(15)*. Furthermore, the usefulness of comparative modeling is steadily increasing because the number of different structural folds that proteins adopt is limited *(16)*, and because the number of experimentally determined new structures is increasing exponentially *(17)*. It is predicted that, in less than 10 yr, at least one example

of most structural folds will be known, making comparative modeling applicable to most globular domains in most protein sequences *(1,17)*.

2. Steps in Comparative Modeling

Comparative modeling usually consists of the following five steps: search for templates, selection of one or more templates, target–template alignment, model building, and model evaluation (*see* **Fig. 1**). If the model is not satisfactory, some or all of the steps can be repeated. Each of these steps is described as follows.

2.1. Search for Templates

Comparative modeling usually starts by searching the database of known protein structures (Protein Data bank, PDB) *(15)* using the target sequence as the query. This is generally done by comparing the target sequence with the sequence of each of the structures in the database. A variety of sequence-sequence comparison methods can be used *(18–20)*. Sometimes, the availability of many sequences related to the target makes it possible to do more sensitive searching with profile methods and Hidden Markov Models (HMM) *(21–24)*. It is also possible to search for templates by evaluating directly the compatibility between the target sequence and each of the structures in the database. This is achieved by fold-recognition methods also known as "threading" *(25–29)*. Threading uses sequence–structure fitness functions, such as low-resolution, knowledge-based force-fields, to evaluate potential target-template matches. In doing so, threading methods generally do not rely on sequence similarity. This sometimes allows recognition of structural similarity between proteins with no detectable sequence similarity *(30)*.

A good starting point for template searches are the many database search servers on the World Wide Web (WWW) (*see* **Table 2**). The most useful ones are those that search directly against the PDB. If nothing is found with sequence similarity searches, threading programs and fold-recognition WWW servers can be used (**Table 2**). In general, it is useful to try many different methods to find as many templates as possible. This is especially important when the target sequence is only remotely related to known structures.

2.2. Template Selection

Once a list of potential templates has been obtained using one or more template searching methods, it is necessary to select the templates that are appropriate for the particular modeling problem. Usually, the higher the overall sequence similarity (i.e., higher percentage of identical residues, and lower number and shorter length of gaps in the alignment) between the target and the template sequences, the better the template is likely to be. Other factors should also be taken into account when selecting a template:

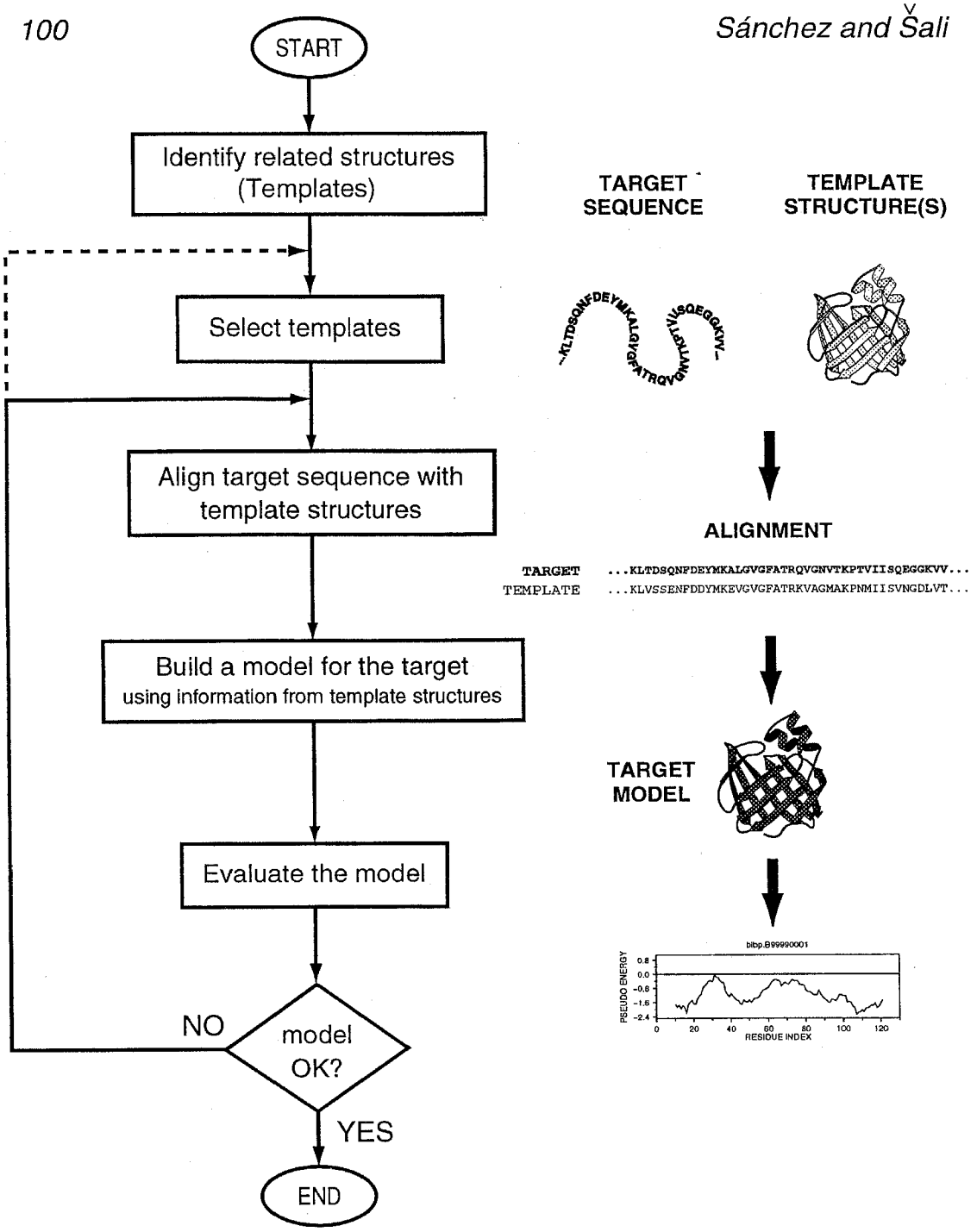

Fig. 1. Steps in comparative protein structure modeling. See text for description of each step.

1. The family of proteins that includes the target and the templates frequently can be organized in subfamilies. The construction of a multiple alignment and a phy-

Table 2
Programs and World Wide Web Servers Useful in Comparative Modeling

Name	Type[a]	World Wide Web or e-mail address	Reference
Template search			
BLASTP	S	www.ncbi.nlm.nih.gov/BLAST/	*(64)*
FASTA	S	www.pdb.bnl.gov/pdb-bin/pdbmain	*(65)*
123D	S	www-lmmb.ncifcrf.gov/~nicka/123D.html	*(66)*
PHDTHREADER	S	www.embl-heidelberg.de/predictprotein/predictprotein.html	*(67)*
UCLA-DOE FRSVR	S	www.doe-mbi.ucla.edu/people/frsvr/frsvr.html	*(68)*
PROFIT	P	www.came.sbg.ac.at	*(69)*
THREADER	P	globin.bio.warwick.ac.uk/~jones/threader.html	*(26)*
MATCHMAKER	P	www.scripps.edu/adam/home.html	*(27)*
Modeling			
COMPOSER	P	felix.bioc.cam.ac.uk/soft-base.html	*(70)*
CONGEN	P	bruc@dino.squibb.com	*(71)*
DRAGON	P	www.nimr.mrc.ac.uk/~mathbio/a-aszodi/dragon.html	*(42)*
MODELLER	P	guitar.rockefeller.edu/modeller/modeller.html	*(40)*
NAOMI	P		*(41)*
WHAT IF	P	www.sander.embl-heidelberg.de/whatif/	*(72)*
INSIGHTII	P	www.msi.com	(a)
LOOK	P	www.mag.com	*(37)*
QUANTA	P	www.msi.com	(a)
SYBYL	P	www.tripos.com	(b)
SWISS-MOD	S	www.expasy.ch/SWISS-MODEL.html	*(73)*
Model evaluation			
PROCHECK	P	www.biochem.ucl.ac.uk/~roman/procheck/procheck.html	*(48)*
WHATCHECK[c]	P	www.sander.embl-heidelberg.de/whatcheck/	*(49)*
PROSAII[c]	P	www.came.sbg.ac.at	*(47)*
PROCYON[d]	P	www.horus.com/sippl/	*(47,69)*
BIOTECH	S	biotech.embl-ebi.ac.uk:8400/	*(48,49)*
VERIFY3D	S	www.doe-mbi.ucla.edu/verify3d.html	*(46)*
ERRAT	S	www.doe-mbi.ucla.edu/errat_server.html	*(74)*

[a]S = server, P = program.
[b](a) Molecular Simulations Inc., San Diego (b) Tripos, St Louis.
[c]PROCYON is a new software package that includes PROSAII, PROFIT, and other programs.
[d]The BIOTECH server uses PROCHECK and WHATCHECK for structure evaluation.

logenetic tree *(31)* can help in selecting the template from the subfamily that is closest to the target sequence.

2. The similarity between the "environment" of the template and the environment in which the target needs to be modeled should also be considered. The word "environment" is used here in a broad sense, including everything that is not the protein itself: solvent, pH, ligands, quaternary interactions, and the like (*see* **Subheadings 3.1.2.** and **4.2.**). In particular, the template(s) bound to the same or similar ligand(s) as the model should be used whenever possible.
3. The quality of the experimental template structure is another important factor in template selection. The resolution and R-factor of a crystallographic structure and the number of restraints per residue for a nuclear magnetic resonance (NMR) structure are indicative of the accuracy of the structure. This information can generally be obtained from the template PDB files or from the articles describing structure determination. If two templates have comparable sequence similarity to the target, the one determined at the highest resolution should be used.

The criteria for selecting templates also depend on the purpose of a comparative model. For instance, if a protein–ligand model is to be constructed, the choice of the template that contains a similar ligand is probably more important than the resolution of the template. On the other hand, if the model is to be used to analyze the geometry of the active site of an enzyme, it is preferable to use a high-resolution template. It is not necessary to select only one template. In fact, the use of several templates generally increases the model accuracy (*see* **Subheading 3.2.** and **Notes**).

2.3. Target-Template Alignment

To build a model, all comparative modeling programs depend on a list that establishes structural equivalences between the target and template residues. This is defined by the alignment of the target and template sequences. Although many template search methods will produce such an alignment, it is usually not the optimal target–template alignment. Search methods tend to be tuned for detection of remote relationships, not for optimal alignments. Therefore, once templates have been selected, a specialized method should be used to align them with the target sequence. The alignment is relatively simple to obtain when the target–template sequence identity is above 40%. In most such cases, an accurate alignment can be obtained automatically using standard sequence–sequence alignment methods. If the target–template sequence identity is lower than 40%, the alignment generally has gaps and needs manual intervention to minimize the number of misaligned residues. In these low-sequence identity cases, the alignment accuracy is the most important factor affecting the quality of the resulting model. Alignments can be improved by including structural information from the template. For example, gaps should be avoided in secondary-structure elements, in buried regions, or between two residues that are

far apart in space. Some alignment methods take such criteria into account (*see* **Subheading 3.1.3.**). However, it is always important to check and edit the alignment by inspecting the template structure visually, especially if the target–template sequence identity is low. A misalignment by only one residue position will result in an error of approximately 4 Å in the model because the current modeling methods cannot recover from errors in the alignment.

2.4. Model Building

Once an initial target–template alignment has been built, a variety of methods can be used to construct a 3D model for the target protein. The original and still most widely used method is modeling by rigid-body assembly *(5,32,33)*. This method constructs the model from a few core regions and from loops and sidechains, that are obtained from dissecting related structures. Another family of methods, modeling by segment matching, relies on the approximate positions of conserved atoms from the templates to calculate the coordinates of other atoms *(34–37)*. The third group of methods, modeling by satisfaction of spatial restraints, uses either distance geometry or optimization techniques to satisfy spatial restraints obtained from the alignment of the target sequence with the template structures *(38–42)*. Accuracies of the various model-building methods are relatively similar when used optimally. Other factors such as template selection and alignment accuracy usually have a larger impact on the model accuracy, especially for models based on less than 40% sequence identity to the templates. However, it is important that a modeling method allows a degree of flexibility and automation, which will make it easier and faster to obtain better models. For example, a method should allow for an easy recalculation of a model when a change is made in the alignment; it should be straightforward to calculate models based on several templates; and the method should provide the tools to incorporate prior knowledge about the target (e.g., experimental data, or predicted features such as secondary-structure). Here we will describe automated comparative model building by satisfaction of spatial restraints as implemented in program MODELLER *(40)*. Reviews of comparative model building methods have been published elsewhere *(1–4)*. Several programs for comparative modeling are listed in **Table 2**.

2.4.1. Comparative Modeling with Program MODELLER

MODELLER is a computer program that models protein structure by satisfaction of spatial restraints (*see* the Appendix at the end of the chapter for information on how to obtain MODELLER). It can be used in all stages of comparative modeling described so far, including template search, target–template alignment and model building. Once a target–template alignment is obtained, the calculation of the 3D model of the target by MODELLER is com-

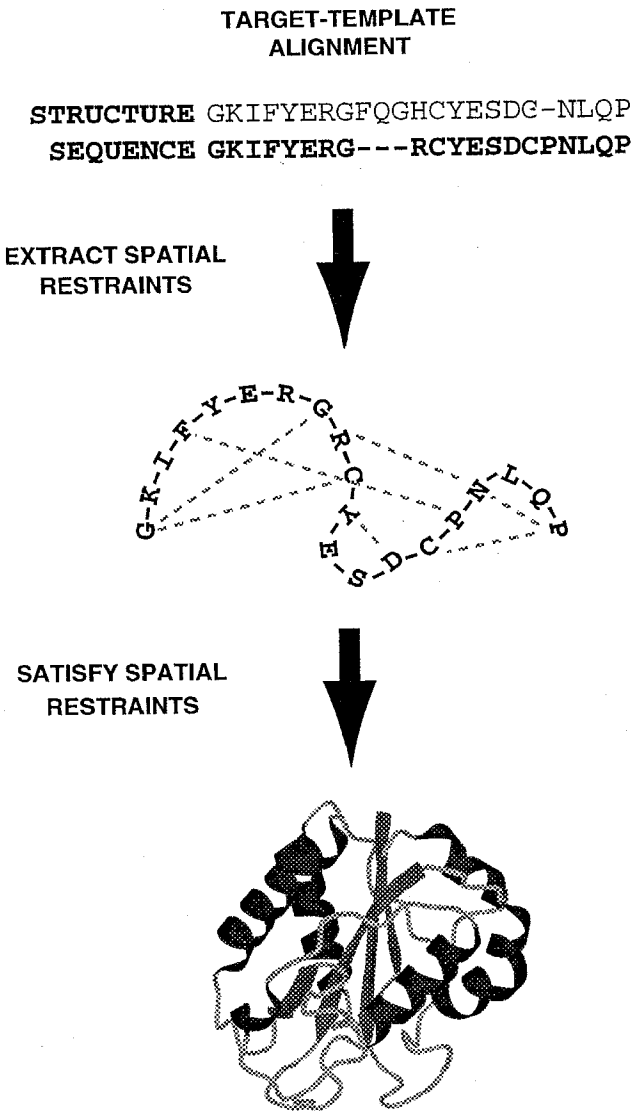

Fig. 2. Comparative modeling by program MODELLER. First, spatial restraints in the form of atom–atom distances and dihedral angles are extracted from the template structure(s). The alignment is used to determine equivalent residues between the target and the template. The restraints are combined into an objective function. Finally, the model for the target is optimized until a model that best satisfies the spatial restraints is obtained. This procedure is similar to the one used in structure determination by NMR.

pletely automated. The program extracts atom–atom distance and dihedral angle restraints on the target from the template structure(s) and combines them with general rules of protein structure such as bond length and angle preferences. The model is then calculated by an optimization procedure that minimizes violations of the spatial restraints (*see* **Fig. 2**). The procedure is

conceptually similar to the one used in the determination of protein structures from NMR data. More detailed descriptions of MODELLER can be found elsewhere *(40,43–45)*.

2.5. Model Evaluation

After a model has been built, it is important to check it for possible errors. Two types of evaluation should be carried out: (1) "internal" evaluation of self-consistency that checks whether or not the model satisfies the restraints used to calculate it and (2) "external" evaluation that relies on information that was not used in calculating the model *(46,47)*.

When the model is based on less than approx 30% sequence identity to the template, the first purpose of the external evaluation is to test whether or not a correct template was used. This is especially important when the alignment is only marginally significant or several alternative templates with different structures are to be evaluated. A complication is that at low similarities the alignment generally contains many errors, making it difficult to distinguish between an incorrect template on one hand and an incorrect alignment with a correct template on the other hand. It is only possible to recognize a correct template if the alignment is also approximately correct. This complication can sometimes be overcome by trying several alternative alignments for each template. One way to predict whether or not a template is correct is to compare the PROSAII Z-score *(47)* for the model and the template structure(s). The Z-score of a model is a measure of compatibility between its sequence and structure. The model Z-score should be comparable to the Z-score obtained for the template. However, this evaluation does not always work. It is sometimes possible that good models have bad Z-scores because the potential function used in PROSAII is not suitable for certain fold types.

The second kind of external evaluation is to recognize unreliable regions in the model. One way to approach this problem is to calculate an energy profile of the model by a program such as PROSAII. The profile reports the energy for each position in the model. It is sometimes possible to detect errors in the model because they appear as peaks of positive energy in the profile. Such regions of the model should be inspected carefully. Another way of finding unreliable regions of a model is to evaluate the stereochemistry (bond length and angles, dihedral angles, atom-atom overlaps, etc.) of the model with programs such as PROCHECK *(48)* and WHATCHECK *(49)*. Although errors in stereochemistry are rare and less informative than errors detected by profiles, a cluster of stereochemical errors in the same segment of the model could indicate that the corresponding region also contains other errors (*see* **Table 2** for a list of evaluation programs and servers). Finally, an important evaluation tool is the experimental knowledge about the protein structure and its function. A model should

be consistent with experimental observations such as site-directed mutagenesis, crosslinking data, ligand binding, and so on.

2.6. The Cycle of Alignment–Modeling–Evaluation

In cases where the best template selection and alignment are not clear, one powerful way of improving a comparative model is to change the alignment and/or the template selection and recalculate the model iteratively until no improvement in the model is detected *(50,51)*. The more exhaustive is the exploration of the templates and alignments, the more likely it is that the accuracy of the final model will improve.

3. Examples

This section contains examples of typical comparative modeling cases. All the examples use program MODELLER and other freely available software. The first example shows each of the five steps of comparative modeling. The other three examples concentrate on specific variations of the basic modeling procedure. The examples are necessarily concise. For more information, the MODELLER manual *(52)* and the literature *(40,43–45,50,53–55)* should be consulted. All the example files can be obtained as explained in the Appendix at the end of the chapter.

3.1. Example 1: Modeling with a Single Template

THE CASE OF HUMAN BRAIN LIPID-BINDING PROTEIN

Brain lipid-binding protein (BLBP) is a brain-specific member of the fatty acid-binding protein (FABP) family. When the sequence of this protein was determined, its function was not known. Thus, a model of the structure of BLBP was built by comparative modeling, and combined with site-directed mutagenesis and binding experiments to understand its ligand specificity *(56)*. The individual modeling steps are described in **Subheading 3.1.1**.

3.1.1. Search for Templates

First, it is necessary to put the target sequence (BLBP sequence) into a format that is readable by MODELLER.MODELLER reads files is the format similar to the widely used FASTA format *(65)*.

File: `blbp.seq`

```
>P1;blbp
sequence:blbp:::::::::
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEE
TSIDDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

The first line contains `'>P1;` followed by the sequence name, `'blbp'` in this case. The second line has 10 fields (separated by colons ":") of which only

two are used in this case: `'sequence'` (indicating that the file contains a sequence without known structure) and `'blbp'`, the sequence name again. The rest of the file contains the sequence of BLBP, with `'*'` marking the end of the sequence. A search for structures that have similar sequence can be performed by the `SEQUENCE_SEARCH` command of MODELLER. The following command file (`TOP` file) will use the query sequence with the name `'blbp'` (`ALIGN_CODES`) from the file `blbp.seq`.

File: `search.top`

```
SET SEARCH_RANDOMIZATIONS = 100
SEQUENCE_SEARCH FILE = 'blbp.seq', ALIGN_CODES = 'blbp'
```

The `SEQUENCE_SEARCH` command has many options *(52)*, but in this example only `SEARCH_RANDOMIZATIONS` is set to a nondefault value. `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 scores will be. To execute the `TOP` command file, type `'mod search.top'`.

3.1.2. Template Selection

The output of the search.top command file is written to the search.log file. If there is any problem with the command file, it will be reported in the log file.* At the end of this long file, MODELLER lists the hits sorted by alignment significance. The example shows only the top 10 hits.

File: `search.log`

#	CODE_1	CODE_2	LEN1	LEN2	NID	%ID	%ID	SCORE	SIGNI	SIGNI2	SIGNI3
1	blbp	1hmt	131	131	81	61.8	61.8	96904.	29.9	-999.0	-999.0
2	blbp	1cbs	131	137	55	40.1	42.0	83725.	19.9	-999.0	-999.0
3	blbp	1ifc	131	131	37	28.2	28.2	76909.	15.1	-999.0	-999.0
4	blbp	1mdc	131	130	37	28.2	28.5	72299.	9.7	-999.0	-999.0
5	blbp	1eal	131	127	34	26.0	26.8	69104.	9.1	-999.0	-999.0
6	blbp	1iltA	131	143	25	17.5	19.1	64604.	3.8	-999.0	-999.0
7	blbp	1bgk	131	37	18	13.7	48.6	7774.	3.5	-999.0	-999.0
8	blbp	1tdx	131	133	25	18.8	19.1	64750.	3.3	-999.0	-999.0
9	blbp	1thjA	131	213	43	20.2	32.8	59771.	3.3	-999.0	-999.0
10	blbp	1amy	131	403	55	13.6	42.0	35790.	3.3	-999.0	-999.0

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

*MODELLER always produces a log file. Errors and warnings in log files can be found by searching for the `'_E>'` and `'_W>'` strings (e.g., with the UNIX grep utility).

PDB code in each line is the representative of a group of PDB sequences that share 30% 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 MODELLER's `CHAINS_3.0_30_XN.grp` file. The `'%ID'` column reports the percentage sequence identity between the two sequences (BLBP and each PDB sequence in this case). In general, a `'%ID'` value above 25–30% indicates a suitable template unless the alignment is short (less than 100 residues). A better measure of the significance of the alignment is given by the SIGNI column *(52)*. A value above 6.0 is generally significant regardless of the sequence identity. In the foregoing example, five PDB structures have significant alignments with the BLBP sequence: 1HMT, 1CBS, 1IFC, 1MDC, 1EAL. All five proteins belong to the family of fatty acid binding proteins. The most similar to BLBP is 1HMT (human muscle fatty acid binding protein) with 61.8% sequence identity and a significance score of 29.9. By inspecting the PDB database (`http://www.pdb.bnl.gov`) or the `CHAINS_3.0_30_XN.grp` file, we find additional structures for the same sequence: 1HMS, 1HMR, 2HMB, and 1HMT all have identical sequences. The main difference between these four structures is the ligand to which the protein is bound. The ligands are stearic acid, oleic acid, elaidic acid, and 1-hexyldecanoic acid for 1HMT, 1HMS, 1HMR, and 2HMB, respectively. Thus, the four proteins are in different "environments." Assuming the interest is in studying the BLBP/oleic acid interaction, the template of choice is 1HMS. 1HMS is also a good template because it is a high resolution structure (1.4 Å). The coordinate file for 1HMS can be retrieved from the PDB database.

3.1.3. Target–Template Alignment

A good way of aligning a sequence (BLBP) and a structure (1HMS) is the `ALIGN2D` command in MODELLER. Although this command is based on the dynamic programming algorithm *(57)*, it is different from standard sequence–sequence alignment methods because it takes into account structural information from the template when constructing an alignment. This 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 *(58)*. As a result, the alignment errors are reduced to approximately one-half of those that occur with standard sequence alignment techniques. This becomes more important as the similarity (sequence identity) between the sequences decreases and the number of gaps increases. In this example, the similarity between template and target is so high that almost any alignment method with reasonable parameters will result in the same alignment. The following MODELLER TOP file will align

the BLBP sequence in file `blbp.seq` with the 1HMS structure in file 1hms.pdb, which is the coordinate file retrieved from the PDB database.

File: `align2d-1.top`

```
READ_MODEL FILE = '1hms.pdb'
SEQUENCE_TO_ALI ALIGN_CODES = '1hms'
READ_ALIGNMENT FILE = 'blbp.seq', ALIGN_CODES = ALLIGN_CODES 'blbp',
ADD_SEQUENCE = on
ALIGN2D
WRITE_ALIGNMENT FILE = 'blbp-1hms.ali', ALIGNMENT_FORMAT = 'PIR'
WRITE_ALIGNMENT FILE = 'blbp-1hms.pap', ALIGNMENT_FORMAT = 'PAP'
```

In the first line, MODELLER reads the 1HMS structure. The `SEQUENCE_TO_ALI` command transfers the sequence from the structure to the alignment in memory and assigns it the name `'1hms'` (`ALIGN_CODES`). The third line reads the BLBP sequence from file blbp.seq, assigns it the name `'blbp'` (`ALIGN_CODES`) and adds it to the alignment in memory (`'ADD_SEQUENCE = on'`). The fourth line calls the `ALIGN2D` command to perform the alignment. Finally, the alignment is written out in two formats, `'PIR'` and `'PAP'`. The PIR format is used by MODELLER in the subsequent model building stage. The PAP alignment is easier to inspect visually. The TOP file is executed by typing `'mod align2d-1.top'`. The output goes to files `blbp-1hms.ali` and `blbp-1hms.pap`:

File: `blbp-1hms.ali`

```
>P1;1hms
structureX:1hms:   1  :  :  131  :  :undefined:undefined:-1.00:-1.00
VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGVEFDETTA
DDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE*
>P1;blbp
sequence:blbp:     :  :     :  :  :  :  0.00:  0.00
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSI
DDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

File: `blbp-1hms.pap`

```
 _aln.pos        10        20        30        40        50        60
1hms      VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNT
blbp      VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNT
 _consrvd **** **** ** *** *** **********   **** **   *      *  *****

 _aln.pos        70        80        90       100       110       120
1hms      EISFKLGVEFDETTADDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHG
blbp      EINFQLGEEFEETSIDDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFG
 _consrvd ** * ** ** **  ***  ** * *** ** * ***** **   **  ***   *** *

 _aln.pos       130
1hms      TAVCTRTYEKE
blbp      DIVAVRCYEKA
 _consrvd   *  * ***
```

Due to the high similarity and equal lengths of BLBP and 1HMS, there are no gaps in the alignment. In the PAP format, all identical positions are marked with a '*'. The PIR format contains the starting and ending residue numbers from the 1HMS PDB file (1 and 131, in this case).

3.1.4. Model Building

Once a target–template alignment has been constructed, MODELLER calculates a 3D model of the target in a completely automated way. The following TOP file will generate one model for BLBP based on the 1HMS template structure and the alignment in file blbp-1hms.ali.

File: model1.top

```
INCLUDE
SET ALNFILE = 'blbp-1hms.ali'
SET KNOWNS = '1hms'
SET SEQUENCE = 'blbp'
SET STARTING_MODEL = 1
SET ENDING_MODEL = 1
CALL ROUTINE = 'model'
```

The first line includes many standard variable and routine definitions. The following five lines set parameter values for the 'model' routine. ALNFILE is the name of the file that contains the target–template alignment in the PIR format. KNOWNS is the name that corresponds to the template(s) (the known structure(s)) in ALNFILE (blbp-1hms.ali). SEQUENCE corresponds to the name of the target sequence in ALNFILE. STARTING_MODEL and ENDING_MODEL define the number of models that will be calculated for this alignment. Since STARTING_MODEL and ENDING_MODEL are the same in this case, only one model will be calculated. The last line in the file calls the 'model' routine that actually calculates the model. Typing 'mod model1.top' will execute the command file. The most important output files are:

1. model1.log: This file reports warnings, errors, and other useful information including restraints that remain violated in the final model.
2. blbp.B99990001: The actual model coordinates in the PDB format. This file can be viewed by any program that reads the PDB format (e.g., RASMOL *[59]* http://www.umass.edu/microbio/rasmol/).

3.1.5. Model Evaluation

As discussed before, there are many alternatives for model evaluation. In this example, PROSAII *(47)* is used to evaluate the model fold and PROCHECK *(48)* is used to check the model's stereochemistry. Before doing any external evaluation of the model, one should check the log file from the

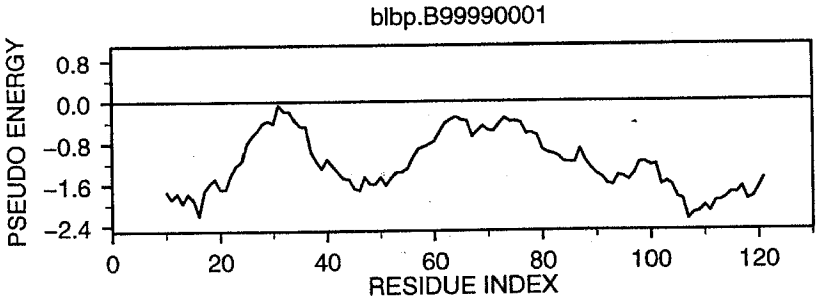

Fig. 3. PROSAII *(47)* energy profile for the BLBP model (*see* Example 1).

modeling run for errors (`model1.log` in this example) and restraint violations (*see* the MODELLER manual for more details on this *(52)*).

First, an energy profile of the model is obtained using the PROSAII program. It is sometimes possible to identify errors in the model because they appear as regions of positive energy in the PROSAII profile. In the case of the BLBP model, no errors were found (*see* **Fig. 3**). This is not surprising given the high similarity between the template and the target. PROSAII is not able to detect all errors, but if a region of the model has a positive profile, one should try alternative alignments in that region.* The stereochemistry of the model can be checked by program PROCHECK. The output of PROCHECK is a series of POSTSCRIPT files with evaluations of different aspects of the model's stereochemistry. One of the most important charts is the Ramachandran plot (*see* **Fig. 4**) which points out those residues that have anomalous combinations of ϕ and ψ angles. As mentioned before, a few deviations of this type are usual even in experimentally determined structures. For example, in **Fig. 4**, alanine 6 and aspartate 98 are in disallowed regions of the plot. However, if several errors cluster in the same region of the model, it is likely that other errors, such as misalignments, have occurred. In this example, both PROSAII and PROCHECK confirm that a good quality model was obtained.

3.2. Example 2: Modeling of a Protein/Ligand Complex

ADDING OLEIC ACID TO BLBP

A better way of analyzing the interaction between BLBP and oleic acid is to add the ligand molecule to the model. To add the ligand that is present in the

*When using profiles, one should always calculate the profile for the template as well. Sometimes a positive peak appears in the model's profile as a consequence of a similar peak in the template's profile. This does not necessarily mean that there is an error in the template structure but more likely the evaluation method is reporting a false error for that particular structure. In such a case, the positive peak in the model probably does not correspond to an error.

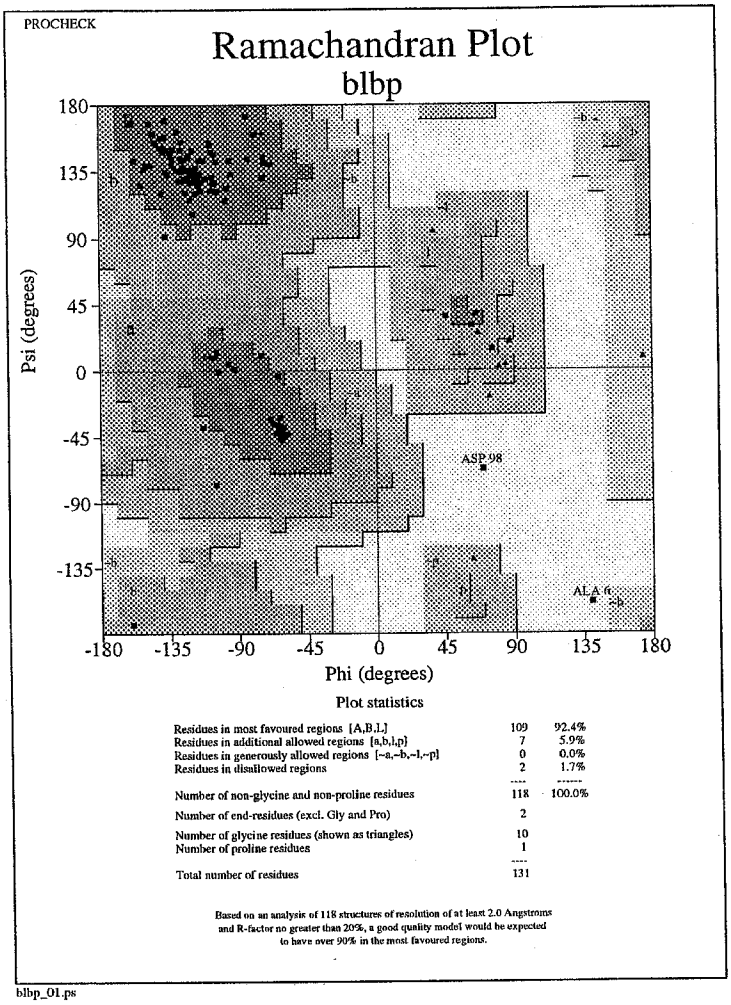

Fig. 4. Evaluation of model stereochemistry. The Ramachandran plot was created for the BLBP model by the PROCHECK program *(48)* (*see* Example 1).

1HMS template (oleic acid) to the BLBP model, all we need to modify is the alignment file `blbp-1hms.ali` and the modeling TOP file `model1.top`. The new files are shown next:

File: `blbp-1hms-ola.ali`

```
>P1;1hms
structureX:1hms:   1  :   :  133 : :undefined:undefined:-1.00:-1.00
VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGVEFDETTA
DDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE.*
>P1;blbp
sequence:blbp:     : :     : : : : 0.00: 0.00
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSI
DDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA.*
```

The second line in the alignment file now specifies that the template is to be used from residue 1 to residue 133 (the oleic acid molecule is residue 133 in 1HMS). The second change in this file is the appearance of the '.' character at the end of each sequence. This character represents the oleic acid molecule in the alignment.*

The modeling command file `model2.top` has two changes with respect to `model1.top`. First, the name of the alignment file assigned to `ALNFILE` was updated. The second change is the addition of `'SET HETATM_IO = on'`. `HETATM_IO` is a flag that indicates to MODELLER whether or not heteroatoms (e.g., nonstandard residues, such as oleic acid) should be read in from the PDB files.

File: `model2.top`

```
INCLUDE
SET ALNFILE = 'blbp-1hms-ola.ali'
SET KNOWNS = '1hms'
SET SEQUENCE = 'blbp'
SET STARTING_MODEL = 1
SET ENDING_MODEL = 1
SET HETATM_IO = on
CALL ROUTINE = 'model'
```

MODELLER can be started with this `TOP` file by typing `'mod model2.top'`. The BLBP model containing the oleic acid residue docked into the binding pocket will be written to `blbp.B99990001`.

It is possible to add ligands which are not present in the template by using predefined ligands in the MODELLER residue topology libraries. These ligands include water molecules, metal ions, heme groups, and others. To place such ligands in the model, additional protein–ligand distance restraints have to be supplied to MODELLER *(52)*.

3.3. Example 3: Modeling Based on More Than One Template

IMPROVING THE BLBP MODEL

Using more than one template usually improves the quality of the model because MODELLER is generally able to combine the best regions from each template when constructing the model *(50)*. Another good template for modeling of BLBP is adipocyte lipid binding protein (ALBP), which is 56% identical to BLBP. Furthermore, a structure of ALBP in complex with oleic acid is available (PDB code 1LID). To calculate a model for BLBP using both templates, an alignment of all three sequences was constructed.

*The dot ('.') character in MODELLER represents a generic residue called a "block" residue. It can be used to represent any nonstandard residue. For more details, *see* the MODELLER manual *(52)*.

File: align2d-3.top

```
SET ALIGN_CODES = '1hms' '1lid'
SET ATOM_FILES = '1hms.pdb' '1lid.pdb'
MALIGN3D
SET ADD_SEQUENCE = on, ALIGN_BLOCK = NUMB_OF_SEQUENCES
READ_ALIGNMENT FILE = 'blbp.seq', ALIGN_CODES = ALIGN_CODES 'blbp'
ALIGN2D
WRITE_ALIGNMENT FILE = 'blbp-1hms-1lid.ali'
WRITE_ALIGNMENT FILE = 'blbp-1hms-1lid.pap', ALIGNMENT_FORMAT = 'PAP'
```

The first three lines in the Top file produce a structural alignment of 1HMS and 1LID using the `MALIGN3D` command. The BLBP sequence in file `blbp.seq` is then added to the structural alignment using the `ALIGN2D` command (lines 4–6). The resulting alignment file in the PIR format, `blbp-1hms-1lid.ali`, has to be edited manually to include the oleic acid residues as block residues (*see* previous example). The edited file is shown here.

File: blbp-1hms-1lid-2.ali

```
>P1;1hms
structureX:1hms:1    : :133    : :undefined:undefined:-1.00:-1.00
VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGVEFDETTA
DDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE.*
>P1;1lid
structureX:1lid:1    : :131    : :undefined:undefined:-1.00:-1.00
CDAFVGTWKLVSSENFDDYMKEVGVGFATRKVAGMAKPNMIISVNGDLVTIRSESTFKNTEISFKLGVEFDEITA
DDRKVKSIITLDGGALVQVQKWDGKSTTIKRKRDGDKLVVECVMKGVTSTRVYERA-*
>P1;blbp
sequence:blbp:     : :     : : : : 0.00: 0.00
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSI
DDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA.*
```

Because the conformations of the oleic acid molecules in 1HMS and 1LID are different, only the 1HMS oleic acid is used as a template. This is done by replacing the 1LID oleic acid residue in the alignment by a gap character (`'-'`). It would be straightforward to produce a BLBP model with the 1LID oleic acid molecule by changing the `blbp-1hms-1lid.ali` alignment. Models for both complexes could be used to design mutants that discriminate between the two binding modes.

Using the TOP file shown below, MODELLER will generate an "ensemble" of five models. Because MODELLER uses different starting coordinates for each model, it is possible that the final models have different conformation in some regions, especially for sidechains. Those regions of the structure that are more variable among the models are likely to be modeled less reliably than the structurally more conserved regions.

File: model3.top

```
INCLUDE
SET ALNFILE = 'blbp-1hms-1lid-2.ali'
SET KNOWNS = '1hms' '1lid'
SET SEQUENCE = 'blbp'
SET STARTING_MODEL = 1
SET ENDING_MODEL = 5
SET HETATM_IO = on
CALL ROUTINE = 'model'
```

After execution of the Top file, the models will be contained in five files `blbp.B99990001` through `blbp.B99990005`. A quick way of evaluating the variability of the models is to superpose their structures. This can be done with the `MALIGN3D` command of MODELLER.

File: malign3d.top

```
SET ATOM_FILES = 'blbp.B99990001' 'blbp.B99990002' 'blbp.B99990003' ;
 'blbp.B99990004' 'blbp.B99990005'
SET WRITE_FIT = on
MALIGN3D
```

The first line specifies the five coordinate files containing the models. The second line directs MODELLER to write the superposed structures to new files. The `MALIGN3D` command finally superposes the five models and actually writes the superposed structures in the new orientations to five files `blbp.B99990001.fit` through `blbp.B99990005.fit`. An easy way to view the superposed models is to concatenate the files with the UNIX `'cat'` command, `'cat blbp.B9999*.fit > sup.pdb'` and display the `sup.pdb` file with RASMOL. The superposed models are shown in **Fig. 5**.

The "best" model can be selected by looking at the value of the MODELLER objective function in the second line of the model PDB files and choosing the one with the lowest value. The value of the objective function in MODELLER is not an absolute measure. It can only be used to compare models calculated from the same templates and alignments, and rank them accordingly.

File: blbp.B99990001

```
REMARK Produced by MODELLER: 19-Dec-97 00:49:51                          1
REMARK MODELLER OBJECTIVE FUNCTION:      623.0785
ATOM      1  N    VAL  1      27.443  41.227  41.628  1.00  0.15      1SG  2
ATOM      2  CA   VAL  1      26.733  41.202  42.923  1.00  0.15      1SG  3
ATOM      3  CB   VAL  1      27.576  41.899  43.956  1.00  0.15      1SG  4
.
.
.
```

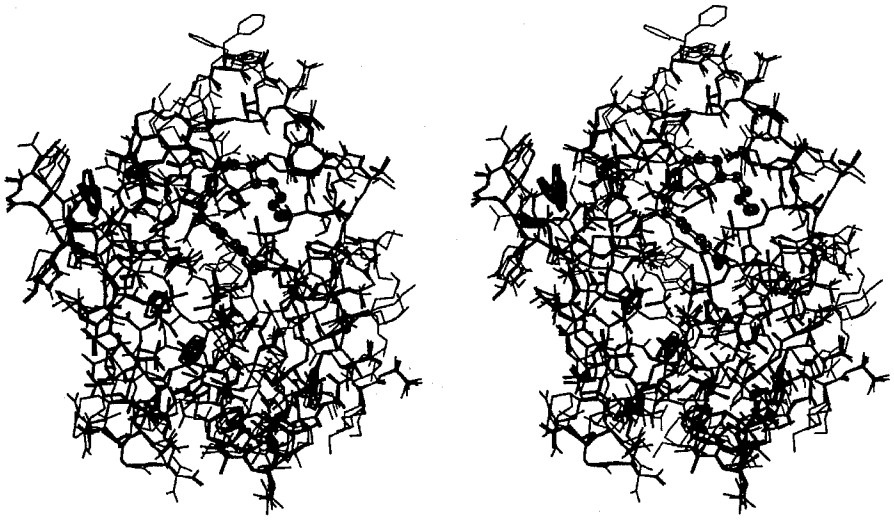

Fig. 5. Stereo plot of the superposition of five BLBP models from Example 3. The oleic acid molecule is shown in ball-and-stick representation *(75)*.

3.4. Example 4: The Alignment–Modeling–Evaluation Cycle

THE CASE OF *Haloferax Volcanii* DIHYDROFOLATE REDUCTASE

Several structures of dihydrofolate reductase (DHFR) are known. However, the structure of DHFR from *Haloferax volcanii* was not known and its sequence identity with DHFRs of known structure is rather low (approx 30%). A model of *H. volcanii* DHFR (HVDFR) was constructed before the experimental structure was solved. Once the crystallographic structure was available, it was possible to compare it with the model ***(50)***. This example illustrates the power of the iterative alignment–modeling–evaluation approach to comparative modeling.

Of all the available DHFR structures, HVDHFR has the sequence most similar to DHFR from *Escherichia coli*. The PDB entry 4DFR corresponds to a high resolution (1.7 Å) *E. coli* DHFR structure. It contains two copies of the molecule — named chain A and chain B. According to the authors, the structure for chain B is of better quality than that of chain A ***(60)***. The following TOP file aligns HVDFR and chain B of 4DFR.

File: `align2d-4.top`

```
READ_MODEL FILE = '4dfr.pdb', MODEL_SEGMENT '@:B' 'X:B'
SEQUENCE_TO_ALI ALIGN_CODES = '4dfr'
READ_ALIGNMENT FILE = 'hvdfr.seq', ALIGN_CODES = ALIGN_CODES 'hvdfr', ADD_SEQUENCE
= on
ALIGN2D
WRITE_ALIGNMENT FILE = 'hvdfr-4dfr.ali'
WRITE_ALIGNMENT FILE = 'hvdfr-4dfr.pap', ALIGNMENT_FORMAT = 'PAP', ;
ALIGNMENT_FEATURES = 'indices helix beta'
```

The new options used in this example include `MODEL_SEGMENT`, which is used to indicate chain B of 4DFR; and `ALIGNMENT_FEATURES`, which is used to output information such as secondary-structure, to the alignment file in the PAP format.

File: `hvdfr-4dfr.pap`

```
 _aln.pos        10        20        30        40        50        60
4dfr      M-ISLIAALAVDRVIGMENAMPW-NLPADLAWFKRNTLDKPVIMGRHTWESIGRPLPGRK
hvdfr     MELVSVAALAENRVIGRDGELPWPSIPADKKQYRSRIADDPVVLGRTTFESMRDDLPGSA
 _helix                          999999999999      999999999
 _beta    9 999999999                            999999           999

 _aln.pos        70        80        90       100       110       120
4dfr      NIILSSQPGT--DDRVTWVKSVDEA--IAACGDVPEIMVIGGGRVYEQFLPKAQKLYLTH
hvdfr     QIVMSRSERSFSVDTAHRAASVEEAVDIAASLDAETAYVIGGAAIYALFQPHLDRMVLSR
 _helix                    99999  99999         99999999
 _beta    99999          99999              9999999            9999999

 _aln.pos       130       140       150       160
4dfr      IDAEVEGDTHFPDYEPDDWESVFSEFHDADAQNSHSYCFKILERR
hvdfr     VPGEYEGDTYYPEWDAAEWELDAETDHEGF--TLQEWVRSASSR
 _helix
_beta     99                   999999999999      999999999999
```

Using the alignment file `hvdfr-4dfr.ali`, an initial model is calculated.

File: `model4.top`

```
INCLUDE
SET ALNFILE = 'hvdfr-4dfr.ali'
SET KNOWNS = '4dfr'
SET SEQUENCE = 'hvdfr'
SET STARTING_MODEL = 1
SET ENDING_MODEL = 1
CALL ROUTINE = 'model'
```

Because the sequence identity between 4DFR and HVDFR is relatively low (30%), the automated alignment is likely to contain errors. The PROSAII evaluation of the model (*see* **Fig. 6**, upper panel) shows two regions with positive energy. The first region is around residue 85, the second region is at the C-terminal end of the protein. Referring to the target–template alignment shown, (`hvdfr-4dfr.pap`), it is easy to understand why the first positive peak appears. The insertion between position 85 and 88 of the alignment was placed in the middle of an a-helix in the template (the "9" characters on the first line below the sequence mark the helices). Moving the insertion to the end of the α-helix may improve the model. The second problem, which occurs in the C-terminal region of the alignment, is less clear. The deletion in that region of

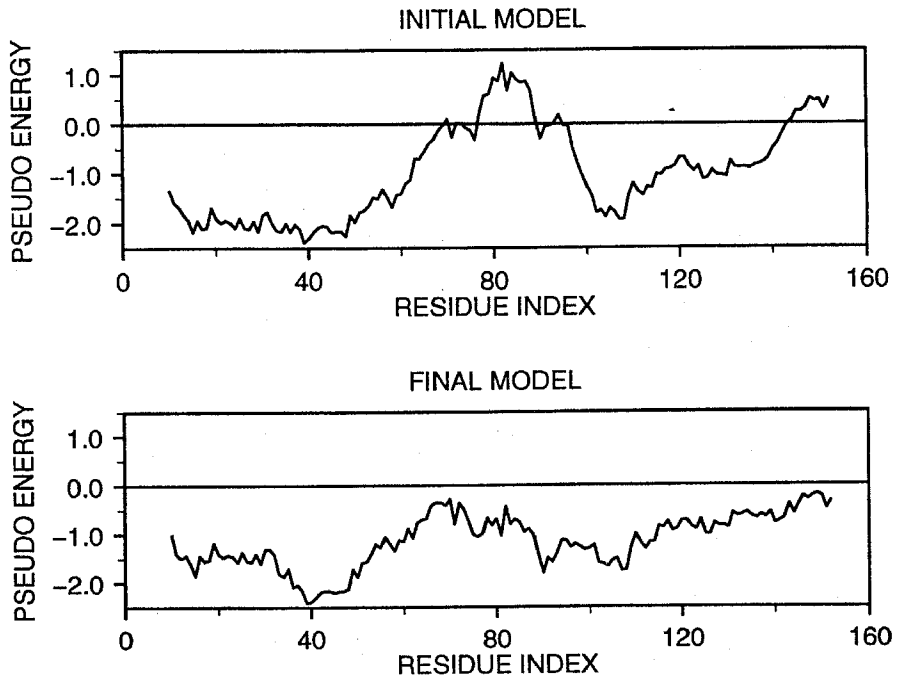

Fig. 6. PROSAII energy profiles for the initial and final HVDFR models (*see* Example 4).

the alignment corresponds to the loop between the last two β-strands of 4DFR (a β-hairpin). Since the profile suggests that this region is in error, an alternative alignment should be tried. One possibility is that the deletion is actually longer, making the C-terminal β-hairpin shorter in HVDFR. One plausible alignment based on this considerations is shown here.

File: hvdfr-4dfr-2.pap

```
 _aln.pos         10        20        30        40        50        60
4dfr      M-ISLIAALAVDRVIGMENAMPW-NLPADLAWFKRNTLDKPVIMGRHTWESIGRPLPGRK
hvdfr     MELVSVAALAENRVIGRDGELPWPSIPADKKQYRSRIADDPVVLGRTTFESMRDDLPGSA
 _helix                            999999999999       999999999
 _beta    9 999999999                              999999           999

 _aln.pos         70        80        90       100       110       120
4dfr      NIILSSQPGT--DDRVTWVKSVDEAIAACG--DVPEIMVIGGGRVYEQFLPKAQKLYLTH
hvdfr     QIVMSRSERSFSVDTAHRAASVEEAVDIAASLDAETAYVIGGAAIYALFQPHLDRMVLSR
 _helix                       99999  99999         99999999
 _beta    99999          99999              9999999           9999999

 _aln.pos        130       140       150       160
4dfr  IDAEVEGDTHFPDYEPDDWESVFSEFHDADAQNSHSYCFKILERR----
hvdfr VPGEYEGDTYYPEWDAAEWELDAETDHE------GFTLQEWVRSASSR
 _helix
_beta     99                 999999999999      999999999999
```

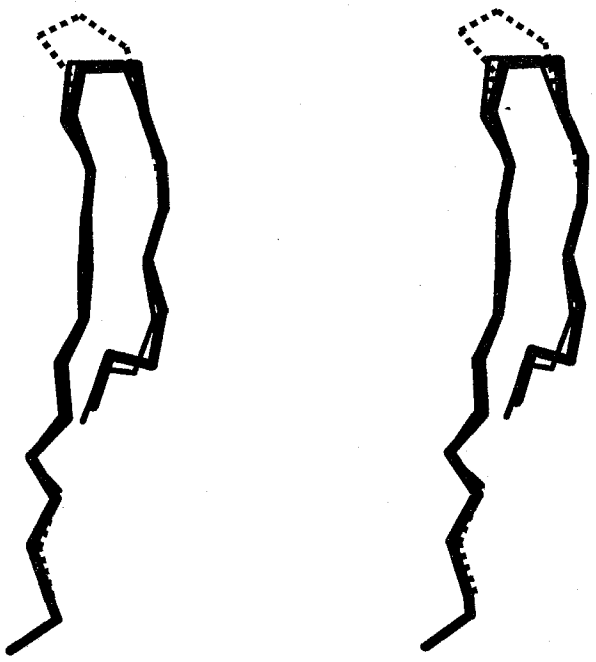

Fig. 7. Stereo plot of the superposition of the C-terminal region of the HVDFR models and the experimental structure (*see* Example 4). Initial model, dotted line; final model, thick line; experimental structure, thin line.

A new model was calculated. Its PROSAII profile is shown in **Fig. 6** (lower panel). Both positive peaks disappeared and the new profile does not contain any positive regions. **Figure 7** shows the comparison of the C-terminal β-hairpin of both models and the actual experimental structure ***(50)***. This confirms that the correct choice for the final alignment was made and that PROSAII was indeed able to detect the error in the initial alignment.

The examples shown here correspond only to the most basic comparative modeling problems. MODELLER can be used for many more complex projects, such as multiple chain models (multimers or protein–protein complexes), symmetry-constrained models, modeling of chimeric structures, and so on. It is also possible to add experimental or predicted data in the form of additional restraints (e.g., NMR or fluorescence distance measurements, disulfide bridges, secondary-structure prediction, and the like). For details and more examples, *see* the MODELLER manual ***(52)***.

4. Notes

4.1. Errors in Comparative Modeling

As the similarity between the target and the templates decreases, the errors in a model increase (*see* **Subheading 4.2.**). Errors in comparative models can be ex-

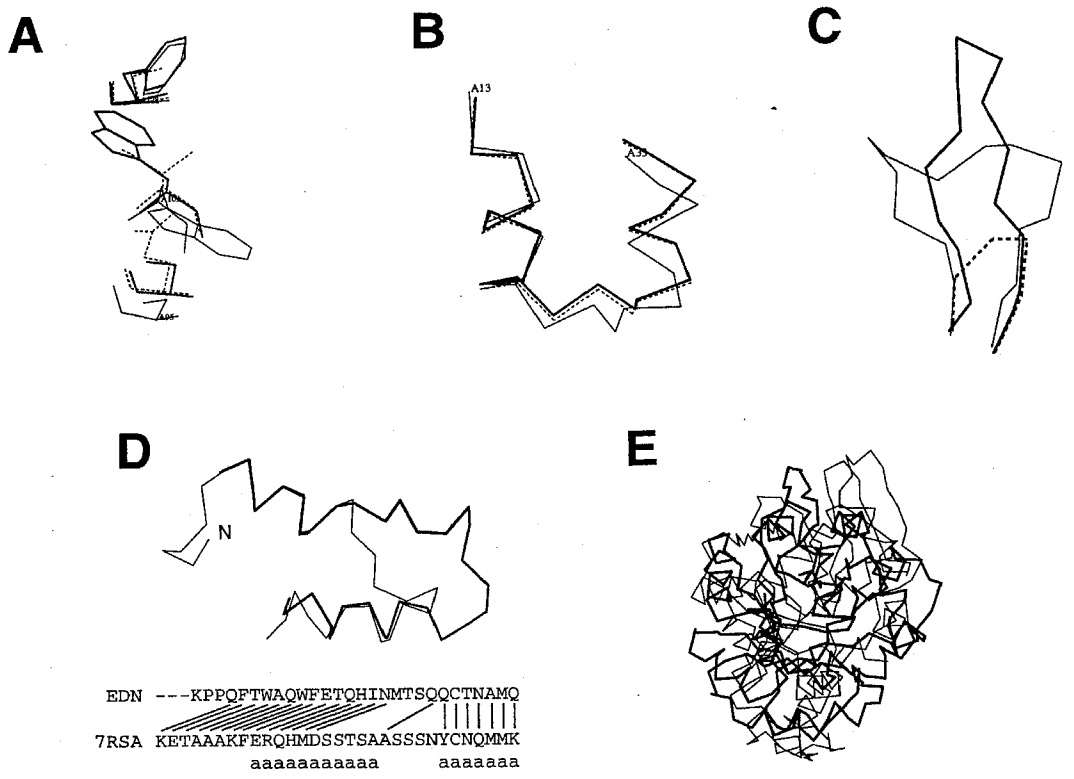

Fig. 8. Typical errors in comparative modeling *(54)* **(A)** Errors in sidechain packing. The Trp 109 residue in the crystal structure of mouse cellular retinoic acid binding protein I (thin line) is compared with its model (thick line), and with the template mouse adipocyte lipid-binding protein (broken line). **(B)** Distortions and shifts in correctly aligned regions. A region in the crystal structure of mouse cellular retinoic acid binding protein I (thin line) is compared with its model (thick line), and with the template fatty acid binding protein (broken line). **(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, i.e., 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) 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 a template.

plained based on the facts that the model resembles the templates as much as possible, and that the modeling procedure cannot recover from misalignments. The typical errors in comparative models include *(45,50,54)* (*see* **Fig. 8**):

1. Errors in sidechain packing: As the sequences diverge, the packing of sidechains in the protein core changes. Sometimes even the conformation of identical sidechains is not conserved, a pitfall for many comparative modeling methods. The sidechain errors are generally not important unless they occur in regions that are involved in function, such as active sites and ligand-binding sites.
2. Distortions and shifts in correctly aligned regions: As a consequence of sequence divergence, the mainchain conformation also changes even if the overall fold remains the same (*see* **Fig. 9**). 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 an incorrect model in that region. Sometimes the target–template differences are not due to differences in sequence but are a consequence of artifacts in structure determination (e.g., crystal packing) or structure determination in different environments. The simultaneous use of several templates minimizes this kind of error ***(50)***.
3. Errors in regions without a template: Segments of the target sequence that have no equivalent region in the template structure (insertions) are the most difficult regions to model. If the insertion is relatively short (usually less than eight residues), some methods are able to predict reliably the conformation of the backbone, but they usually need special attention ***(1,2)***. Conditions for the successful prediction of the conformation of an insertion are the correct alignment and an accurately modeled environment around the insertion. Insertions longer than 8 residues are generally not possible to model correctly with the current methods.
4. Errors due to misalignments: The largest source of errors in comparative modeling are misalignments, especially when the target–template similarity decreases below 40% (*see* **Fig. 9**). For example, at 30% sequence identity on the average 20% of the residues are misaligned ***(61)***. A misalignment of a residue by a single position produces a positional error of approx 4 Å in the model. The current comparative modeling methods cannot recover from alignment errors because the model building procedure is not able to modify the target–template alignment. However, alignment errors can be corrected or avoided in two ways. First, it is usually possible to use a large number of sequences, even if most of them do not have known structures, to construct a family alignment. Multiple alignments are generally more reliable than pairwise alignments ***(62)***. The second way of improving the alignment is to modify those regions of the alignment that correspond to predicted errors in the model in an iterative way, as described in **Subheading 2.6**.
5. Incorrect templates: This is a potential problem when distantly related proteins are used as templates (i.e., less than 30% sequence identity). As discussed before, models based on incorrect templates can generally be identified at the evaluation stage. The largest practical problem is to distinguish between a model based on an incorrect template and a model based on a mostly incorrect alignment with a correct template. In both cases, the evaluation methods will predict an unreliable model. A possible solution to this problem is to explore several different alignments for the target–template pair. In theory, it should be possible to find align-

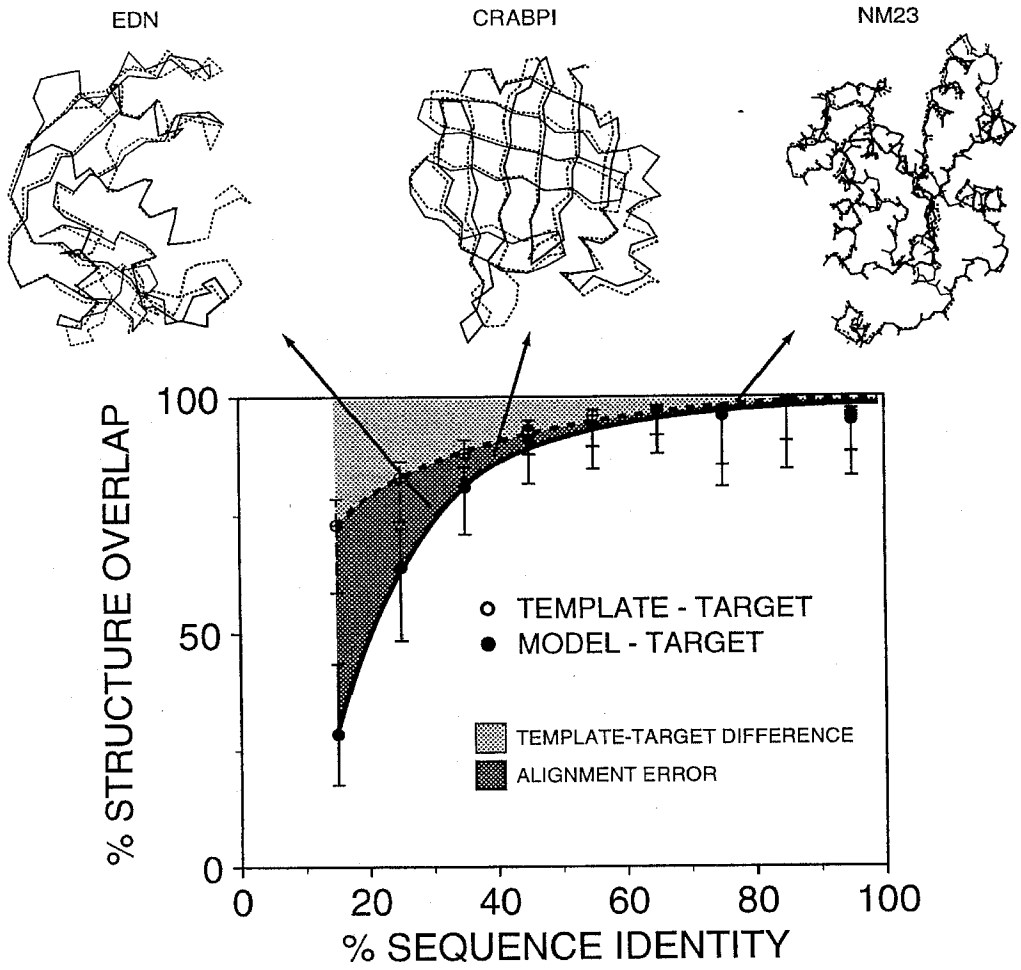

Fig. 9. Average model accuracy as a function of sequence identity. As the sequence identity between the target sequence and the template structure decreases, the average structural similarity between the template and the target also decreases (dotted line, open circles). Structural overlap is defined as the fraction of equivalent C_α atoms. For the comparison of the model with the actual structure (filled circles), two C_α atoms were considered equivalent if they were within 3.5 Å of each other and belonged to the same residue. For comparison of the template structure with the actual target structure (open circles), two C_α atoms were considered equivalent if they were within 3.5 Å of each other after alignment and rigid-body superposition by the `ALIGN3D` command in MODELLER. At high-sequence identities, the models are close to the templates, and therefore also close to the experimental target structure (solid line, filled circles). At low-sequence identities, errors in the target–template alignment become more frequent and the structural similarity of the model with the experimental target structure falls below the target–template structural similarity. The difference between the model and the actual target structure is a combination of the target–template differences (light area) and the alignment errors (dark area). The figure was constructed by calculating 3993 comparative models based on single templates of varying similarity

ments that are accurate enough to produce a good model if the template is correct. However, in practice the number of possibilities that need to be explored to find a sufficiently accurate alignment may be too large. Therefore, the only way to assure that a certain template is incorrect for a particular target is by finding another template with different structure that produces a better model for the same target.

4.2. Relationship Between Target–Template Similarity and Model Accuracy

The quality of a model can be approximately predicted from the sequence similarity between the target and the template (**Fig. 9**). Sequence identity above 30% is a relatively good predictor of the expected accuracy of a model. However, other factors, including the environment, can strongly influence the accuracy of a model. For instance, some calcium-binding proteins undergo large conformational changes when bound to calcium. If a calcium-free template is used to model the calcium-bound state of a target, it is likely that the model will be incorrect irrespective of the target–template similarity. This also applies to experimental determination of protein structure. A structure must be determined in the functionally meaningful environment. If the target–template sequence identity falls below 30%, the sequence identity becomes unreliable as a measure of expected accuracy of a single model. The reason is that the dispersion of the model–target structural overlap increases with the decrease in sequence identity. Below 30% sequence identity, it is relatively frequent to obtain models that deviate significantly, in both directions, from the average accuracy. It is in such cases, that model evaluation methods (*see* **Subheading 2.5.**) are most important to use.

4.3. Are Comparative Models Better than Their Templates?

In general, models are as close to the target structure as the templates, or slightly closer if the alignment is correct ***(50)***. This is not a trivial achievement because of the many residue substitutions, deletions, and insertions that occur when the sequence of one protein is transformed into the sequence of another. Even in a favorable modeling case with a template that is 50% identical to the target, half of the sidechains change and have to be packed in the protein core such that they avoid atom clashes and violations of stereochemical restraints.

---

to the targets. All targets had known (experimentally determined) structures, and therefore the comparison of the models and templates with the experimental structures was possible ***(63)***. The top part of the figure shows three models (solid line) compared with their corresponding experimental structures (dotted line). The models were calculated with MODELLER in a completely automated fashion before the experimental structures were available ***(54)***. The arrows indicate the target–template similarity in each case.

When more than one template is used for modeling, it is sometimes possible to obtain a model that is significantly closer to the target structure than any of the templates *(50)*. This is so because the model tends to inherit the best regions from each template, thus minimizing some of the distortions in the correctly aligned regions. Alignment errors are the main factor that may make models worse than the templates. However, to represent the target, it is always better to use a comparative model rather than the template. The reason is that the errors in the alignment affect similarly the use of the template as a representation of the target as well as the comparative model based on the same template *(50)*.

4.4. Establishing Remote Protein–Protein Relationships by Model Evaluation

Evaluation of a comparative model implied by a target–template alignment is a powerful way of confirming the significance of the alignment. It is often the case that a sequence similarity search of a database results in only a marginal or nonsignificant hit even when two proteins are homologous. A good way of confirming such a hit, when one of the proteins happens to have a known structure, is to build a comparative model for the sequence of unknown structure. If the resulting model is of good quality, according to the evaluation methods described in **Subheading 2.5.**, it is likely that the two proteins have similar structures *(50,51,63)*. This approach is also useful when structural similarity is suspected in the absence of sequence similarity.

Acknowledgments

The authors are grateful to Azat Badretdinov, Eric Feyfant, and András Fiser for discussions about comparative modeling. RS is a Howard Hughes Medical Institute predoctoral fellow. AS is an Alexandrine and Alexander Sinsheimer Medical Fund Scholar. The investigation has also been aided by grants from the National Institutes of Health (GM 54762) and the National Science Foundation (BIR-9601845).

Appendix: How to Obtain MODELLER and the Example Files

MODELLER

MODELLER is freely available to academic users. It runs on most UNIX systems, including PCs running LINUX. The program and data files can be accessed on the Web at `http://guitar.rockefeller.edu/modeler/modeller.html` or can be downloaded by FTP from `guitar.rockefeller.edu` using the anonymous account. MODELLER, with a graphical interface, is also available as part of QUANTA, INSIGNTII, and GENEEXPLORER (Molecular Simulations Inc., San Diego, CA, e-mail: `dje@msi.com`).

Example Files

All example files used in the text, some additional data files, as well as the links in **Table 2** can be accessed on the Web at `http://guitar.rockefeller.edu/modeller/psp/`

References

1. Sánchez, R. and Šali, A. (1997) Advances in comparative protein-structure modeling. *Curr. Opin. Struct. Biol.* **7,** 206–214.
2. Marti-Renom, M. A., Stuart, A., Fiser, A., Sá-nchez, R., and Šali, A. (????) Comparative protein structure modeling of genes and genomes. *Ann. Rev. Biophys. Biomolec. Struct.*, in press.
3. Johnson, M. S., Srinivasan, N., Sowdhamini, R., and Blundell, T. L. (1994) Knowledge-based protein modelling. *CRC Crit. Rev. Biochem. Mol. Biol.* **29,** 1–68.
4. Bajorath, J., Stenkamp, R., and Aruffo, A. (1994) Knowledge-based model building of proteins: concepts and examples. *Protein Sci.* **2,** 1798–1810.
5. Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., and Thornton, J. M. (1987) Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **326,** 347–352.
6. Chothia, C. and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5,** 823–826.
7. Hubbard, T. J. P. and Blundell, T. L. (1987) Comparison of solvent inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.* **1,** 159–171.
8. Oliver, S. G. (1996) From DNA sequence to biological function. *Nature* **379,** 597–600.
9. Koonin, E. V. and Mushegian, A. R. (1996) Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. *Curr. Biol.* **6,** 757–762.
10. Dujon, B. (1996) The yeast genome project: what did we learn? *Trends Genet.* **12,** 263–270.
11. Matsumoto, R., Šali, 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 in mouse mast cell protease-7 regulates its binding to heparin serglycin proteoglycan. *J. Biol. Chem.* **270,** 19524–19531.
12. Lesk, A. M. and Chothia, C. (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136,** 225–270.
13. Rost, B. and Sander, C. (1996) Bridging the protein sequence–structure gap by structure predictions. *Annu. Rev. Biophys. Biomol. Struct.* **25,** 113–136.
14. Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., and Ouellette, B. F. F. (1997) GenBank. *Nucleic Acids Res.* **26,** 1–7.
15. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T., and Weng, J. (1987) Protein data bank, in *Crystallographic Databases — Information, Content, Software Systems, Scientific Applications* (Allen, F. H., Bergerhoff, G., and Sievers, R., eds.), Data Commission of the International Union of Crystallography, Cambridge, pp. 107–132.

16. Chothia, C. (1992) One thousand families for the molecular biologist. *Nature* **360,** 543–544.
17. Holm, L. and Sander, C. (1996) Mapping the protein universe. *Science* **273,** 595–602.
18. Doolittle, R. F. (1990) Searching through sequence databases. *Methods Enzymol.* **183,** 99–110.
19. Altschul, S. F., Boguski, M. S., Gish, W., and Wootton, J. C. (1994) Issues in searching molecular sequence databases. *Nat. Genet.* **6,** 119–129.
20. Pearson, W. R. (1996) Effective protein sequence comparison. *Methods Enzymol.* **266,** 227–258.
21. Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* **84,** 4355–4358.
22. Gribskov, M. (1994) Profile analysis. *Methods Mol. Biol.* **25,** 247–266.
23. Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994) Hidden Markov Models in computational biology: applications to protein modeling. *J. Mol. Biol.* **235,** 1501–1531.
24. Eddy, S. R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.* **6,** 361–365.
25. Bowie, J. U., Lüthy, R., and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253,** 164–170.
26. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) A new approach to protein fold recognition. *Nature* **358,** 86–89.
27. Godzik, A., Kolinski, A., and Skolnick, J. (1992) Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.* **227,** 227–238.
28. Sippl, M. J. and Flöckner, H. (1996) Threading thrills and threats. Structure **4,** 15–19.
29. Torda, A. E. (1997) Perspectives in protein-fold recognition. *Curr. Opin. Struct. Biol.* **7,** 200–205.
30. Dunbrack Jr., R. L., Gerloff, D. L., Bower, M., Chen, X., Lichtarge, O., and Cohen, F. E. (1997) Meeting review: the second meeting on the critical assessment of techniques for protein structure prediction (CASP2), Asilomar, CA, December 13–16, 1996. *Fold. Des.* **2,** R27–R42.
31. Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39,** 783–791.
32. 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 α-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol.* **42,** 65–86.
33. Greer, J. (1981) Comparative model-building of the mammalian serine proteases. *J. Mol. Biol.* **153,** 1027–1042.
34. Jones, T. H. and Thirup, S. (1986) Using known substructures in protein model building and crystallography. *EMBO J.* **5,** 819–822.
35. Unger, R., Harel, D., Wherland, S., and Sussman, J. L. (1989) A 3-D building blocks approach to analyzing and predicting structure of proteins. *Proteins* **5,** 355–373.
36. Claessens, M., Cutsem, E. V., Lasters, I., and Wodak, S. (1989) Modelling the polypeptide backbone with "spare parts" from known protein structures. *Protein Eng.* **4,** 335–345.

37. Levitt, M. (1992) Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226,** 507–533.
38. 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.
39. Srinivasan, S., March, C. J., and Sudarsanam, S. (1993) An automated method for modeling proteins on known templates using distance geometry. *Protein Sci.* **2,** 227–289.
40. Šali, A. and Blundell, T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234,** 779–815.
41. 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 lipolyated H-protein from the pea leaf glycine cleavage system: a new automated methods for the prediction of protein tertiary structure. *Protein Sci.* **2,** 626–639.
42. Aszodi, A. and Taylor, W. R. (1996) Homology modelling by distance geometry. *Fold. Des.* **1,** 325–334.
43. Šali, A. and Overington, J. (1994) Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.* **3,** 1582–1596.
44. Šali, A. (1995) Protein modeling by satisfaction of spatial restraints. *Mol. Med. Today* **1,** 270–277.
45. Sánchez, R. and Šali, A. (1997) Comparative protein modeling as an optimization problem. *J. Mol. Struct. (Theochem.)* **398,** 489–496.
46. Lüthy, R., Bowie, J. U., and Eisenberg, D. (1992) Assessment of protein models with three-dimensional profiles. *Nature* **356,** 83–85.
47. Sippl, M. J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins* **17,** 355–362.
48. Laskowski, R. A., McArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26,** 283–291.
49. Hooft, R., Vriend, G., Sander, C., and Abola, E. (1996) Errors in protein structures. *Nature* **381,** 272.
50. Sánchez, R. and Šali, A. (1997) Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins* **(Suppl.),** 50–58.
51. Guenther, B., Onrust, R., Šali, A., O'Donnell, M., and Kuriyan, J. (1997) Crystal structure of the δ' subunit of the clamp-loader complex of *E. coli* DNA polymerase III. *Cell* **91,** 335–345.
52. Šali, A., Sánchez, R., and Badretdinov, A. (1997) MODELLER, *A Protein Structure Modeling Program,* URL `http://guitar.rockefeller.edu./Modeller/Modeller.html`
53. Šali, A. and Blundell, T. L. (1994) Comparative protein modelling by satisfaction of spatial restraints, in *Protein Structure by Distance Analysis* (Bohr, H., and Brunak, S., eds.), IOS Press, Amsterdam, The Netherlands, pp. 64–86.
54. Šali, A., Potterton, L., Yuan, F., van Vlijmen, H., and Karplus, M. (1995) Evaluation of comparative protein modeling by MODELLER. *Proteins* **23,** 318–326.

55. Sánchez, R., Badretdinov, A. Y., Feyfant, E., and Šali, A. (1997) Homology protein structure modeling. *Trans. Am. Cryst. Assoc.* **32,** 81–91.
56. Xu, L. Z., Sánchez, R., Šali, A., and Heintz, N. (1996) Ligand specificity of brain lipid binding protein. *J. Biol. Chem.* **271,** 24711–24719.
57. 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.
58. Sánchez, R. and Šali, A. Variable gap penalty function for protein sequence–structure alignments. In preparation.
59. Sayle, R. and Milner-White, E. J. (1995) RasMol: biomolecular graphics for all. *Trends Biochem. Sci.* **20,** 374.
60. Bolin, J. T., Filman, D. J., Matthews, D. A., Hamlin, R. C., and Kraut, J. (1982) Crystal structures of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase refined at 1. 7 angstroms resolution. I. General features and binding of methotrexate. *J. Biol. Chem.* **257,** 13,650.
61. 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.
62. Barton, G. J. and Sternberg, M. J. E. (1987) A strategy for the rapid multiple alignment of protein sequences; confidence levels from tertiary structure comparisons. *J. Mol. Biol.* **198,** 327–337.
63. Sánchez, R. and Šali, A. (1998) Large-scale protein structure modeling of the saccharomyces cerevisiae genome. *Proc. Natl. Acad. Sci.* **95,** 13,597–13,602
64. 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.
65. Pearson, W. R. (1990) Rapid and sensitive comparison with FASTA and FASTP. *Methods Enzymol.* **183,** 63–98.
66. Alexandrov, N. N., Nussinov, R., and Zimmer, R. M. (1995) Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. World Scientific Publishing Co., Singapore, pp. 53–72
67. Rost, B. (1995) TOPITS: threading one-dimensional predictions into three-dimensional structures. AAAI Press, Menlo Park, CA, pp. 314–321.
68. Ficher, D. and Eisenberg, D. (1996) Fold recognition using sequence-derived predictions. *Prot. Sci.* **5,** 947–955.
69. Flockner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M., and Sippl, M. J. (1995) Progress in fold recognition. *Proteins* **23,** 376–386.
70. 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.
71. Bruccoleri, R. E. (1993) Application of systematic conformational search to protein modeling. *Mol. Sim.* **10,** 151–174.
72. Vriend, G. (1990) WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8,** 52–56.
73. 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.

74. Colovos, C. and Yeates, T. O. (1993) Verification of protein structures: patterns of non-bonded atomic interactions. *Protein Sci.* **2,** 1511–1519.
75. Kraulis, P. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structure. *J. Appl. Cryst.* **24,** 946–950.