Modeling Protein Structure From Its Sequence

Marc A. Marti-Renom, Bozidar Yerkovich and Andrej Sali

Laboratories of Molecular Biophysics

Pels Family Center for Biochemistry and Structural Biology

The Rockefeller University, 1230 York Ave, New York, NY 10021, USA

KEYWORDS:

Comparative Modeling, MODELLER, ModWeb, ModBase, ModView

Corresponding authors:

Andrej Sali
The Rockefeller University,
1230 York Ave, New York, NY 10021, USA.
tel: (212) 327 7550; fax: (212) 327 7540
e-mail: sali@rockefeller.edu

Marc A. Marti-Renom
The Rockefeller University,
1230 York Ave, New York, NY 10021, USA.
tel: (212) 327 7206; fax: (212) 327 7540
e-mail: martim@rockefeller.edu

Version: 25 July 2002

Functional characterization of a protein sequence is one of the most frequent problems in biology. This task is usually facilitated by accurate three-dimensional (3D) structure of the studied protein. In the absence of an experimentally determined structure, comparative or homology modeling can sometimes provide a useful 3D model for a protein that is related to at least one known protein structure. Comparative modeling predicts the 3D structure of a given protein sequence (target) based primarily on its alignment to one or more proteins of known structure (templates). The prediction process consists of fold assignment, target-template alignment, model building, and model evaluation. The number of protein sequences that can be modeled and the accuracy of the predictions are increasing steadily because of the growth in the number of known protein structures and because of the improvements in the modeling software. 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, Baker, 2001), comparative modeling remains the only method that can reliably predict the 3D structure of a protein with an accuracy comparable to a low-resolution experimentally determined structure (Marti-Renom et al., 2000). Even models with errors may be useful, because some aspects of function can be predicted from just coarse structural features of a model (Marti-Renom et al., 2000;Baker, Sali, 2001).

There are several computer programs and web servers that automate the comparative modeling process (see Table 5.1.1 and unit 5.2). 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., 2001a).

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

In this unit, we first describe generic considerations in all four steps of comparative modeling (Figure 5.1.1). We then illustrate these considerations by a detailed discussion of modeling of lactate dehydrogenase from *Trichomonas vaginalis* using our program MODELLER (Sali, Blundell, 1993;Sali, Overington, 1994;Fiser et al., 2000;Sali et al., 2002). Finally, we outline types of applications of comparative modeling (Figure 5.1.2) and typical errors (Figure 5.1.3) in comparative protein structure models.

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 database scanning software available on the web (Altschul et al., 1994;Barton, 1998;Holm, Sander, 1996) (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, Sander, 1999), 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, Eisenberg, 1997;Huynen et al., 1998;Jones, 1999;Rychlewski et al., 1998;Sanchez, Sali, 1998;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, Giancarlo, 1998). The performance of these methods in searching for related protein sequences and structures has been evaluated exhaustively (Sauder et al., 2000;Thompson et al., 1999). Frequently used programs in this class include FASTA (Pearson, Lipman, 1988;Pearson, 1995) and BLAST (Altschul et al., 1990).

The second set of methods relies on multiple sequence comparisons to improve the sensitivity of the search (Altschul et al., 1997;Henikoff, Henikoff, 1994;Gribskov, 1994;Krogh et al., 1994;Rychlewski et al., 1998). A widely used program in this class is PSI-BLAST (Altschul et al., 1997), which iteratively expands the set of homologs of the target sequence. For a given sequence, an initial set of homologs from a sequence database is collected, a weighted multiple alignment is made from the query sequence and its homologs, a position specific scoring matrix is constructed from the alignment, and the matrix is used to search the database for additional homologs. These steps are repeated until no additional homologs are found. In comparison to

BLAST, PSI-BLAST finds homologs of known structure for approximately twice as many sequences (Park et al., 1998;Sternberg et al., 1999).

The third class of methods is the so-called threading or 3D template matching methods (Bowie et al., 1991;Jones et al., 1992;Godzik et al., 1992), reviewed in (Jones, 1997;Smith et al., 1997;Torda, 1997;Levitt, 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 3D folds is predicted by an optimization of the alignment with respect to a structure dependent scoring function, independently for each sequence-structure pair. That is, the target sequence is threaded through a library of 3D folds. These methods are especially useful when there are no sequences clearly related to the modeling target, and thus the search cannot benefit from the increased sensitivity of the sequence profile methods.

A useful fold assignment approach is to accept an uncertain assignment provided by any of the methods, build an all-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 (Sanchez, Sali, 1997;Guenther et al., 1997;Miwa et al., 1999).

Once a list of all related protein structures was 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. Several other factors should be taken into account when selecting the templates:

- The family of proteins, which includes the target and the templates, can frequently be organized in sub-families. The construction of a multiple alignment and a phylogenetic tree

(Felsenstein, 1985) can help in selecting the template from the sub-family that is closest to the target sequence.

- The template "environment" should be compared to the required environment for the model. The term environment is used in a broad sense and includes all factors that determine protein structure, except its sequence (*e.g.*, solvent, *p*H, ligands, and quaternary interactions).
- The quality of the experimental template structure is another important factor in template selection. The resolution and the R-factor of a crystallographic structure and the number of restraints per residue for an NMR structure are indicative of its accuracy.

The priorities of the criteria for template selection depend on the purpose of the comparative model. For instance, if a protein-ligand model is to be constructed, the choice of the template that contains a similar ligand is probably more important than the resolution of the template. On the other hand, if the model is to be used to analyze the geometry of the active site of an enzyme, it is preferable to use a high-resolution template. It is not necessary to select only one template. In fact, the use of several templates approximately equidistant from the target sequence generally increases the model accuracy (Srinivasan, Blundell, 1993;Sanchez, Sali, 1997).

Target-template alignment

Most fold assignment methods produce an alignment between the target sequence and template structures. However, this alignment 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 (Taylor, 1996;Holm, Sander, 1996;Briffeuil et al., 1998;Baxevanis, 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, 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.

There is a great variety of protein sequence alignment methods, many of which are based on dynamic programming techniques (Needleman, Wunsch, 1970;Smith, Waterman, 1981). A frequently used program for multiple sequence alignment is CLUSTAL (Thompson et al., 1994;Higgins et al., 1996), which is also available as a web server (Table 5.1.1).

In the more difficult alignment cases, it is frequently beneficial to rely on multiple structure and sequence information (Barton, Sternberg, 1987;Taylor et al., 1994). First, the alignment of the potential templates is prepared by superposing their structures. Next, the sequences that are clearly related to the templates and are easily aligned with them are added to the alignment. The same is done for the target sequence. Finally, the two profiles are aligned with each other, taking structural information into account as much as possible (Sali et al., 2002;Koretke et al., 1998;Thompson et al., 1994;Yang, Honig, 2000;Al Lazikani et al., 2001b).

Model building

Once an initial target-template alignment was built, a variety of methods can be used to construct a 3D model for the target protein. The original and still widely used method is modeling by rigid-body assembly (Browne et al., 1969;Greer, 1990;Blundell et al., 1987). Another family of methods, modeling by segment matching, relies on the approximate positions of conserved atoms in the templates (Jones, Thirup, 1986;Unger et al., 1989;Claessens et al., 1989;Levitt, 1992). 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 (Havel, Snow, 1991;Srinivasan et al., 1993;Sali, Blundell, 1993;Brocklehurst, Perham, 1993;Aszodi, Taylor, 1996;Kolinski et al., 2001). 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. There are many reviews of comparative model building methods (Blundell et al., 1987;Sanchez, Sali, 1997;Sali, 1995;Johnson et al., 1994;Bajorath et al., 1993;Marti-Renom et al., 2000;Al Lazikani et al., 2001a). A number of programs and web servers for comparative modeling are listed in Table 5.1.1.

Model evaluation

The accuracy of a model determines its usefulness. 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, 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 model fold is generally increased by a high sequence similarity to the closest template, a pseudo-energy Z-score (Sippl, 1993;Sanchez, 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, 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, Lesk, 1986), the "geometrical" nature of modeling that forces the model to be as close to the template as possible (Sali, Blundell, 1993), and the inability of any current modeling procedure to recover from an incorrect alignment (Sanchez, 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 expected accuracy of a single model. Models that deviate significantly from the average accuracy are frequent. It is in such cases that model evaluation methods are particularly useful.

In addition to the target-template sequence similarity, the environment can strongly influence the accuracy of a model. For instance, some calcium-binding proteins undergo large conformational changes when bound to calcium. If a calcium-free template is used to model the calcium-bound state of the target, it is likely that the model will be incorrect irrespective of the target-template similarity or accuracy of the template structure (Pawlowski et al., 1996). This also applies to the

experimental determination of protein structure; a structure must be determined in the functionally meaningful environment.

A basic requirement for a model is to have good stereochemistry. Some useful programs for evaluating stereochemistry are PROCHECK (Laskowski et al., 1998), PROCHECK-NMR (Laskowski et al., 1996), AQUA (Laskowski et al., 1996), SQUID (Oldfield, 1992), and WHATCHECK (Hooft et al., 1996a). The features of a model that are checked by these programs include bond lengths, bond angles, peptide bond and sidechain ring planarities, chirality, mainchain and sidechain torsion angles, and clashes between non-bonded pairs of atoms.

There are also methods for testing 3D models that implicitly take into account many spatial features compiled from high resolution protein structures. These methods are based on 3D profiles and statistical potentials of mean force (Sippl, 1990;Luthy et al., 1992). Programs implementing this approach include VERIFY3D (Luthy et al., 1992), PROSAII (Sippl, 1993), HARMONY (Topham et al., 1994), and ANOLEA (Melo, 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 prediction errors and energy peaks is greatly weakened, resulting in the loss of predictive power of the energy profile. Despite these concerns, error profiles have been useful in some applications (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 program MODELLER 6 (Sali et al., 2002). All files described in this section, including the MODELLER program, are available at `http://guitar.rockefeller.edu/modeller/tutorials.shtml`.

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. We 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

First, it is necessary to put the target TvLDH sequence into the PIR format readable by MODELLER (file "TvLDH.ali").

```
>P1;TvLDH
sequence:TvLDH:::::::0.00: 0.00
```

```
MSEAAHVLITGAAGQIGYILSHWIASGELYG-DRQVYLHLLDIPPAMNRLTALTMELEDCAFPHLAGFVATTDPK
AAFKDIDCAFLVASMPLKPGQVRADLISSNSVIFKNTGEYLSKWAKPSVKVLVIGNPDNTNCEIAMLHAKNLKPE
NFSSLSMLDQNRAYYEVASKLGVDVKDVHDIIVWGNHGESMVADLTQATFTKEGKTQKVVDVLDHDYVFDTFFKK
IGHRAWDILEHRGFTSAASPTKAAIQHMKAWLFGTAPGEVLSMGIPVPEGNPYGIKPGVVFSFPCNVDKEGKIHV
VEGFKVNDWLREKLDFTEKDLFHEKEIALNHLAQGG*
```

The first line of the file 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 "*" 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 "seq_search.top").

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

The SEQUENCE_SEARCH command has many options (Sali et al., 2002), 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 “seq_search.top” script is written to the “seq_search.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 very long, a separate data file (“seqsearch.dat”) is created that contains the summary of the search. The example below shows only the top 10 hits from such file.

#	CODE_1	CODE_2	LEN1	LEN2	NID	%ID1	%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	9ldtA	335	331	95	28.4	28.7	181720.	4.7
10	TvLDH	1cdb	335	105	69	29.6	65.7	80141.	3.8

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 proteins 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 approximately 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 (Sali et al., 2002). 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% 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 bacteria). 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 and 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 following script performs all pairwise comparisons among the selected proteins (file "compare.top").

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

The READ_ALIGNMENT command 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 3D structures. 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 mainchain and sidechain 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)]. 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 below (file “compare.log”).

```
>> 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
```

The comparison above shows that *5mdhA* and *4mdhA* are almost identical, both sequentially and structurally. They were solved at similar resolutions, 2.4 and 2.5Å, 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), 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, Wunsch, 1970), it is different from standard sequence-sequence alignment methods because it takes into account structural information 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 Cα 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 following MODELLER script aligns the TvLDH sequence in file "TvLDH.seq" with the *4mdhA* structure in the PDB file "4mdh.pdb" (file "align2d.top").

```
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'
```

In the first line, MODELLER reads the *4mdhA* structure file. The SEQUENCE_TO_ALI command transfers the sequence to the alignment array and assigns it the name of "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 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").

```
_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 *  *   *     **** *  ** ***    * ****   **   * ****      *   *


 _aln.pos    210       220       230       240       250       260       270
4mdhA     KEVGVYEAVKDDSWLKGEFITTVQQRGAAVIKARKLSSAMSAAKAICDHVRDIWFGTPEGEFVSMGII
TvLDH     KTQKVVDVLDHDYVFDTFFKKIGHRAWDILEHRGFTSAASPTKAAIQHMKAWLFGTAPGEVLSMGIPV
 _consrvd *   *      *      *                 * *     **           *


 _aln.pos      280       290       300       310       320       330
4mdhA     SDGNSYGVPDDLLYSFPVTIK-DKTWKIVEGLPINDFSREKMDLTAKELAEEKETAFEFLSSA-
TvLDH     PEGNPYGIKPGVVFSFPCNVDKEGKIHVVEGFKVNDWLREKLDFTEKDLFHEKEIALNHLAQGG
 _consrvd   ** **       ***           ***   **  *** * * * *  *** *   *
```

Model building

Once a target-template alignment is constructed, MODELLER calculates a 3D model of the target completely automatically. The following script will generate five models of TvLDH based on the *4mdhA* template structure and the alignment in file "TvLDH-4mdh.ali" (file "model-single.top").

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

The first line 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 “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 “TvLDH.B99990001”, which contains the model coordinates in the PDB format. The model can be viewed by any program that reads the PDB format, such as ModView (`http://guitar.rockefeller.edu/modview/`) (Ilyin et al., 2002) or RasMol (`http://www.rasmol.org`) (Bernstein, 2000;Sayle, Milner-White, 1995)

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 in the sense that it can only be used to rank models calculated from the same alignment.

Once a final model is selected, there are many ways to assess it. In this example, PROSAII (Sippl, 1993) is used to evaluate the model fold and PROCHECK (Laskowski et al., 1998) is used to check the model's stereochemistry. 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 (Sali et al., 2002)). Both PROSAII and PROCHECK confirm that a reasonable model was obtained, with a 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, Sali, 2002).

APPLICATIONS OF COMPARATIVE MODELING

Comparative modeling is an increasingly efficient way to obtain useful information about the proteins of interest. For example, comparative models can be helpful in designing mutants to test hypotheses about a protein's function (Boissel et al., 1993;Wu et al., 1999), identifying active and binding sites (Sheng et al., 1996), identifying, designing and improving ligands for a given binding site (Ring et al., 1993), modeling substrate specificity (Xu et al., 1996), predicting antigenic epitopes (Sali et al., 1993), simulating protein-protein docking (Vakser, 1997), inferring function from a 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), testing and improving a sequence-structure alignment (Wolf et al., 1998), confirming a remote structural relationship (Guenther et al., 1997;Miwa et al., 1999), and rationalizing known experimental observations. For exhaustive reviews of comparative modeling applications see (Johnson et al., 1994;Fiser, Sali, 2002;Baker, Sali, 2001).

Fortunately, a 3D model does not have to be absolutely perfect to be helpful in biology, as demonstrated by the applications listed above. However, the type of a question that can be addressed with a particular model does depend on its accuracy (Figure 5.1.2). Three levels of model accuracy and some of the corresponding applications are as follows.

- At the low end of the spectrum, there are models that are based on less than 30% sequence identity and have sometimes less than 50% of their atoms within 3.5Å of their correct positions. Such models still have the correct fold, which is frequently sufficient to predict its approximate biochemical function. More specifically, only nine out of 80 fold families known in 1994 contained proteins (domains) that were not in the same functional class, although 32% of all protein structures belonged to one of the nine superfolds (Orengo et al., 1994). Models in this low range of accuracy combined with model evaluation can be used for confirming or rejecting a match between remotely related proteins (Sanchez, Sali, 1998;Sanchez, Sali, 1997).
- In the middle of the accuracy spectrum are the models based on approximately 30%-50% sequence identity, corresponding to 85% of the atoms modeled within 3.5Å of their correct positions. Fortunately, the active and binding sites are frequently more conserved than the rest of the fold, and are thus modeled more accurately (Sanchez, Sali, 1998). In general, medium resolution models frequently allow refinement of the functional prediction based on sequence alone because ligand binding is most directly determined by the structure of the binding site rather than by its sequence. It is frequently possible to correctly predict important features of the target protein that do not occur in the template structure. For example, the location of a binding site can be predicted from clusters of charged residues (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 more than 50% sequence identity. The average accuracy of these models approaches that of low resolution X-ray structures (3Å resolution) or medium resolution NMR structures (10 distance restraints per residue) (Sanchez, Sali, 1997). The alignments on which these models are based generally contain almost no errors. In addition to the already listed applications, high quality models can be used for docking of small ligands (Ring et al., 1993) or whole proteins onto a given protein (Totrov, Abagyan, 1994;Vakser, 1997).

ERRORS IN COMPARATIVE MODELS

As the similarity between the target and the templates decreases, the errors in the model increase. Errors in comparative models can be divided into five categories (Sanchez, Sali, 1997) (Figure 5.1.3):

- 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. Sidechain errors are critical if they occur in regions that are involved in protein function, such as active sites and ligand-binding sites.
- Distortions and shifts in correctly aligned regions. As a consequence of sequence divergence, the mainchain 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 (< 3Å) 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, Blundell, 1993;Sanchez, Sali, 1997).

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

- Errors due to misalignments. The largest source of errors in comparative modeling is misalignments, especially when the target-template sequence identity decreases below 30%. However, alignment errors can be minimized in two ways. First, it is usually possible to use a large number of sequences to construct a multiple alignment, even if most of these sequences do not have known structures. Multiple alignments are generally more reliable than pairwise alignments (Barton, 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, Sali, 1997).

- Incorrect templates. This is a potential problem when distantly related proteins are used as templates (*i.e.*, less than 25% sequence identity). Distinguishing between a model based on an incorrect template and a model based on an incorrect alignment with a correct template is difficult. In both cases, the evaluation methods will predict an unreliable model. The

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

An informative way to test automated protein structure modeling methods is provided by EVA-CM (Eyrich et al., 2001) and LiveBench (Bujnicki et al., 2001b) (Table 5.1.1)

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, Sali, 2001;Pieper et al., 2002). The magnitude of errors in fold assignment, alignment, and the modeling of sidechains and loops, has decreased measurably. These improvements are a consequence of both better techniques and a larger number of known protein sequences and structures. Nevertheless, all the errors remain significant and demand future methodological improvements. In addition, there is a great need for more accurate detection of errors in a given protein structure model. Error detection is useful both for refinement and interpretation of the models.

To maximize the number of proteins that can be modeled reliably, a concerted effort towards structure determination of new folds by X-ray crystallography and nuclear magnetic resonance spectroscopy is in order, as envisioned by structural genomics (Terwilliger et al., 1998;Sali, 1998;Montelione, Anderson, 1999;Zarembinski et al., 1998;Burley et al., 1999;Vitkup et al., 2001;Sanchez et al., 2000). 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). 4000 of these families 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.

Acknowledgments

We are grateful to all members of our research group, especially to Dr. András Fiser, for many discussions about comparative protein structure modeling. MAM-R was Burroughs Wellcome Fund Postdoctoral Fellow and is currently Rockefeller University Presidential Postdoctoral Fellow. AS 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 based on (Marti-Renom et al., 2002b)

Legend to figures:

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 (violet) 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 (*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 15Å 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 (*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 Cα trace of the 112–117 loop is shown for the X-ray structure of human eosinophil neurotoxin (thin line), its model (thick line), and the template ribonuclease A structure (residues 111–117; broken line). (d) Errors due to misalignments. The N-terminal region in the crystal structure of human eosinophil neurotoxin (thin line) is compared with its model (thick line). The corresponding region of the alignment with the template ribonuclease A is shown. The black lines show correct equivalences, that is residues whose Cα atoms are within 5Å 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.

Legend to tables:

Table 5.1.1. Programs and web servers useful in comparative modeling. [a]S, server; P, program. [b]Some of the sites are mirrored on additional computers. [c](a) MolSoft Inc., San Diego. (b) Accelrys Inc., San Diego. (c) Tripos Inc., St Louis. [d]The BIOTECH server uses PROCHECK and WHATCHECK for structure evaluation.

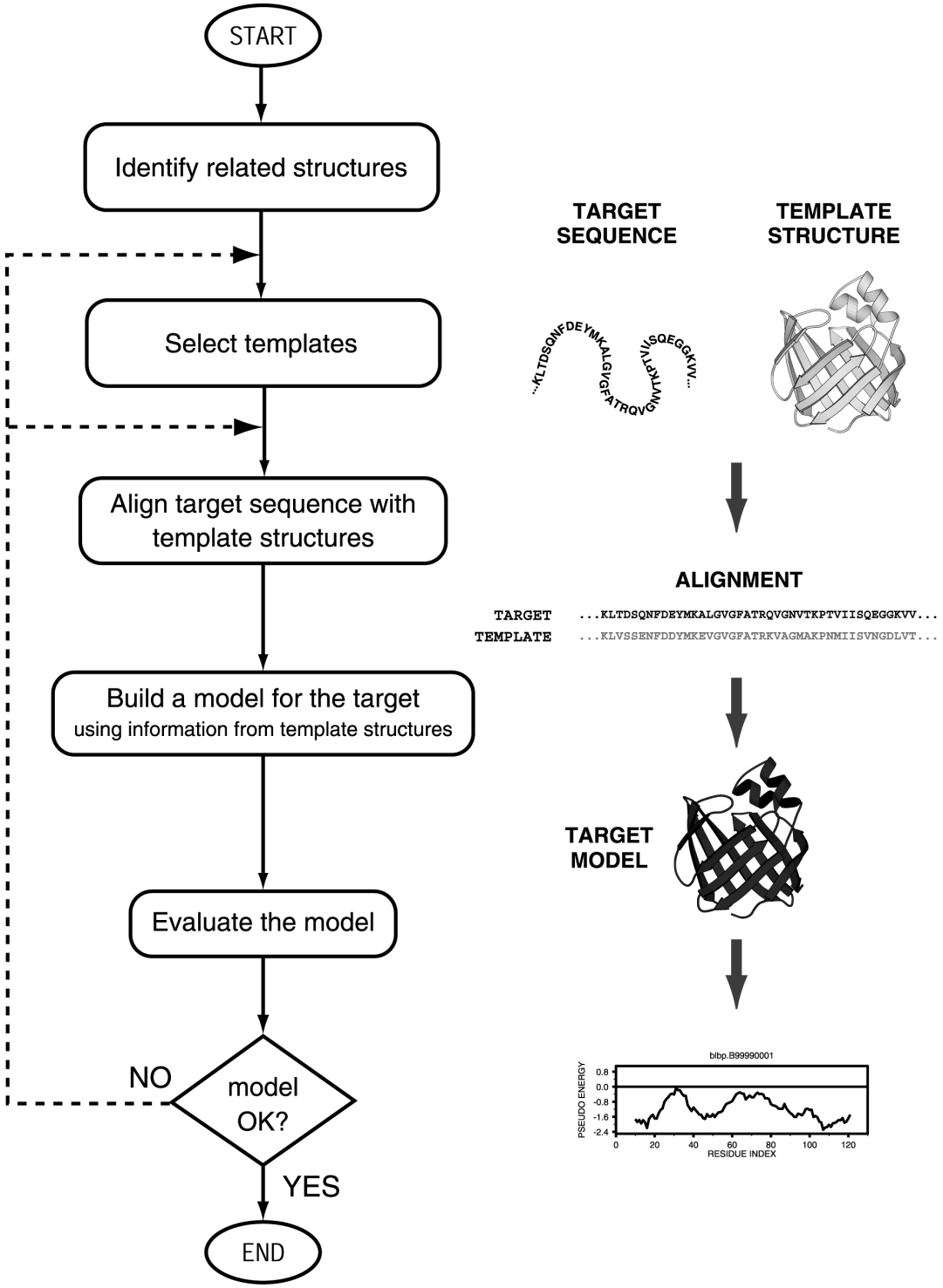
START
Identify related structures
Select templates
Align target sequence with template structures
Build a model for the target
using information from template structures
Evaluate the model
model OK?
NO
YES
END
TARGET SEQUENCE
TEMPLATE STRUCTURE
...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
ALIGNMENT
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

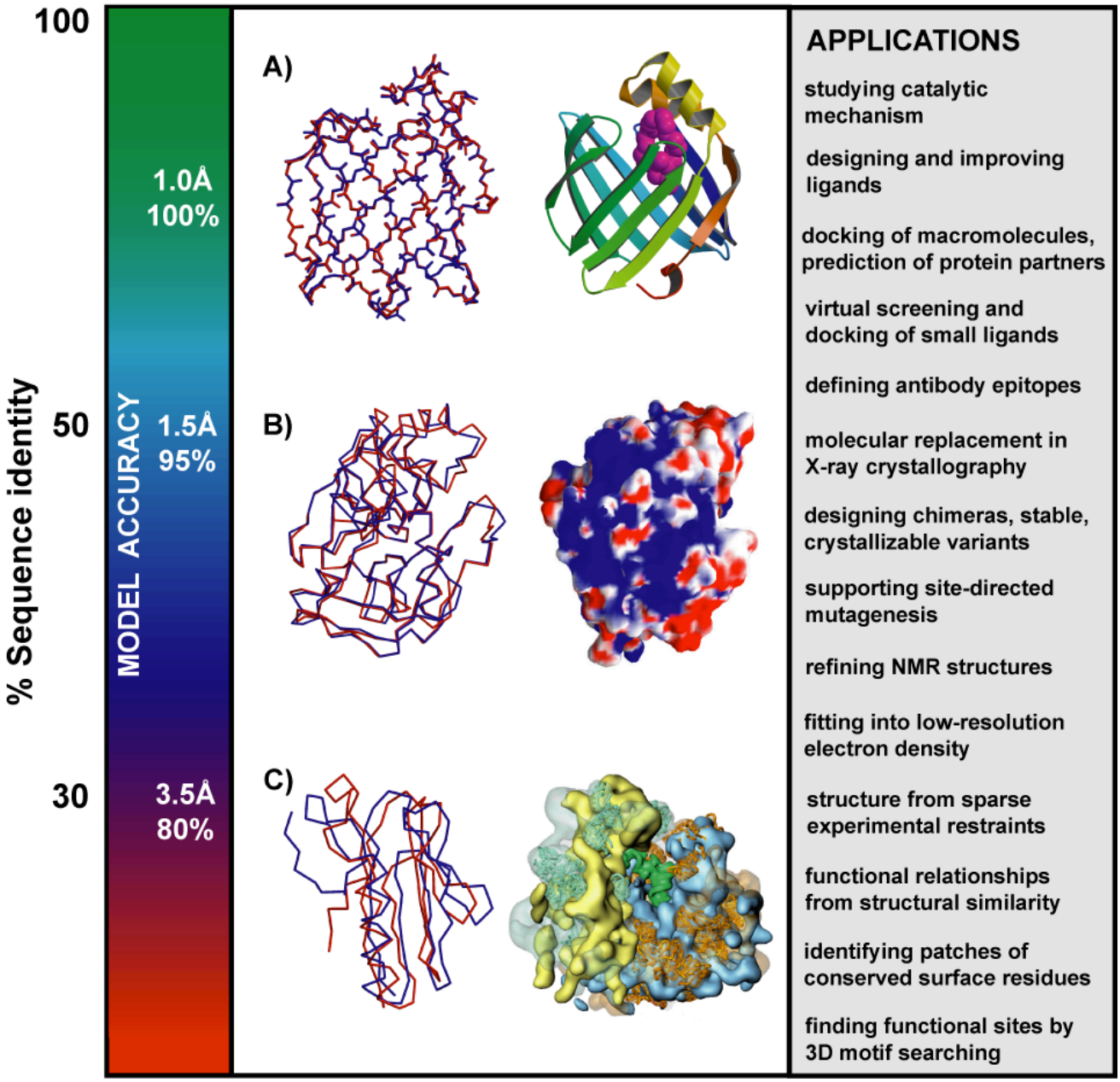
100
50
30
% Sequence identity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
A)
B)
C)
APPLICATIONS
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screening and docking of small ligands
defining antibody epitopes
molecular replacement 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

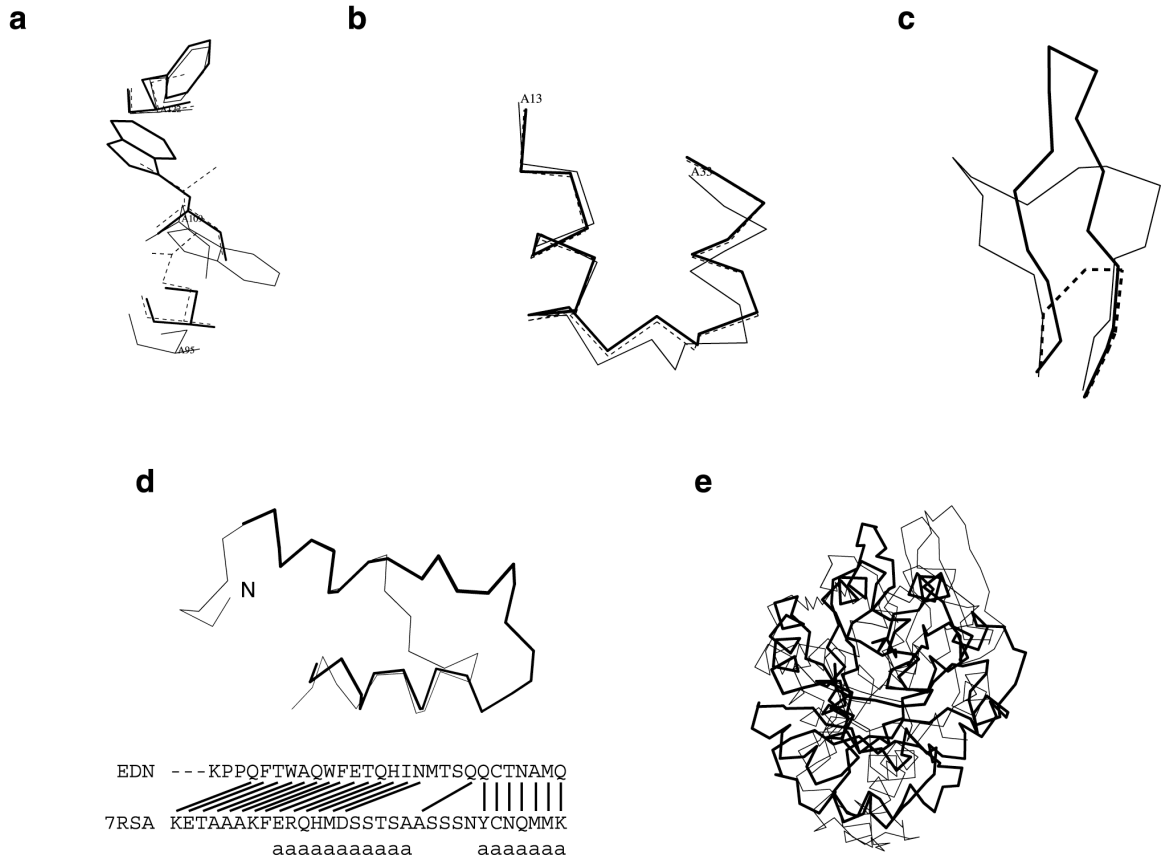
a
b
c
A13
A35
d
e
N
EDN ---KPPQFTWAQWFETQHINMTSQQCTNAMQ
7RSA KETAAAKFERQHMDSSTSAASSSNYCNQMMK
aaaaaaaaaaa aaaaaaa

Figure 5.1.3

Table 5.1.1.

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

Table 5.1.1. Continuation…

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

References

Al Lazikani B, Jung J, Xiang Z, Honig B. 2001a. Protein structure prediction. *Curr Opin Chem Biol 5*:51-56.

Al Lazikani B, Sheinerman FB, Honig B. 2001b. Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. *Proc Natl Acad Sci U S A 98*:14796-14801.

Alexandrov NN, Nussinov R, Zimmer RM. 1996. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. *Pac Symp Biocomput*53-72.

Altschul SF, Boguski MS, Gish W, Wootton JC. 1994. Issues in searching molecular sequence databases. *Nat Genet 6*:119-129.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403-410.

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

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

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

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

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

Barton GJ, Sternberg MJ. 1987. A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J Mol Biol 198*:327-337.

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

Beckmann R, Spahn CMT, Eswar N, Helmers J, Penczek PA, Sali A, Frank J, Bloebel G. 2001. Analysis of the protein-conducting channel associated with the translating 80S ribosome. *submitted*.

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

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

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

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

Bower MJ, Cohen FE, Dunbrack RL, Jr. 1997. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J Mol Biol 267*:1268-1282.

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

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

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

Brocklehurst SM, Perham RN. 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.

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

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

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. 2001a. LiveBench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci 10*:352-361.

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. 2001b. LiveBench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci 10*:352-361.

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

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

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

Colovos C, Yeates TO. 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 MJ, Hunter IW. 2000. 3D-1D threading methods for protein fold recognition. *Pharmacogenomics 1*:445-455.

Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, 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.

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

Fischer D, 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 AR, Dunbrack RL, Jr. 2001. CAFASP2: The second critical assessment of fully automated structure prediction methods. *Proteins 45 Suppl 5*:171-183.

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

Fiser A, Sali A. 2002. Comparative protein structure modeling with Modeller: A practical approach. *Methods Enzymol in press*.

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

Gerstein M, 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, Skolnick J. 1992. Topology fingerprint approach to the inverse protein folding problem. *J Mol Biol 227*:227-238.

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

Henikoff S, Henikoff JG. 1994. Protein family classification based on searching a database of blocks. *Genomics 19*:97-107.

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

Higgins DG, Thompson JD, Gibson TJ. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol 266*:383-402.

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

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

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

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

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

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

Ilyin VA, Pieper U, Stuart AC, Marti-Renom MA, McMahan L, Sali A. 2002. Visualization and analysis of multiple protein sequences and structures by ModView. *submitted*.

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

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

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

Jones DT. 1999. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol 287*:797-815.

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

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

Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, 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.

Koretke KK, Luthey-Schulten Z, Wolynes PG. 1998. Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. *Proc Natl Acad Sci U S A 95*:2932-2937.

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

Laskowski RA, MacArthur MW, Thornton JM. 1998. Validation of protein models derived from experiment. *Curr Opin Struct Biol 8*:631-639.

Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. 1996. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR 8*:477-486.

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 Suppl 1*:92-104.

Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. 2002. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res 30*:264-267.

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

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

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

Marti-Renom MA, Yerkovich B, Sali A. 2002b. Comparative Protein Structure Prediction. *Current Protocols in Protein Science 2*.

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

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

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

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

Montelione GT, Anderson S. 1999. Structural genomics: keystone for a Human Proteome Project. *Nat Struct Biol 6*:11-12.

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

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol 48*:443-453.

Oldfield TJ. 1992. SQUID: a program for the analysis and display of data from crystallography and molecular dynamics. *J Mol Graph 10*:247-252.

Orengo CA, Bray JE, Buchan DW, Harrison A, Lee D, Pearl FM, Sillitoe I, Todd AE, Thornton JM. 2002. The CATH protein family database: A resource for structural and functional annotation of genomes. *Proteomics 2*:11-21.

Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature 372*:631-634.

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

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

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

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

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A 85*:2444-2448.

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

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

Pontius J, Richelle J, Wodak SJ. 1996. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J Mol Biol 264*:121-136.

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

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

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

Sali, A., Fiser, A., Sanchez, R., Marti-Renom, M. A., Jerkovic, B., Badretdinov, A., Melo, F., Overington, J., and Feyfant, E. MODELLER, A Protein Structure Modeling Program, Release 6v2. 2002. http://guitar.rockefeller.edu/modeller/.

Ref Type: Computer Program

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

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

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

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

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

Saqi MA, Russell RB, Sternberg MJ. 1998. Misleading local sequence alignments: implications for comparative protein modelling. *Protein Eng 11*:627-630.

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

Sayle RA, Milner-White EJ. 1995. RASMOL: biomolecular graphics for all. *Trends Biochem Sci 20*:374.

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

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

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

Smith RF, Wiese BA, Wojzynski MK, Davison DB, Worley KC. 1996. BCM Search Launcher--an integrated interface to molecular biology data base search and analysis services available on the World Wide Web. *Genome Res 6*:454-462.

Smith TF. 1999. The art of matchmaking: sequence alignment methods and their structural implications. *Structure Fold Des 7*:R7-R12.

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

Smith TF, Waterman MS. 1981. Overlapping genes and information theory. *J Theor Biol 91*:379-380.

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

Srinivasan S, March CJ, Sudarsanam S. 1993. An automated method for modeling proteins on known templates using distance geometry. *Protein Sci 2*:277-289.

Sternberg MJ, Bates PA, Kelley LA, MacCallum RM. 1999. Progress in protein structure prediction: assessment of CASP3. *Curr Opin Struct Biol 9*:368-373.

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

Taylor WR. 1996. Multiple protein sequence alignment: algorithms and gap insertion. *Methods Enzymol 266*:343-367.

Taylor WR, Flores TP, Orengo CA. 1994. Multiple protein structure alignment. *Protein Sci 3*:1858-1870.

Terwilliger TC, Waldo G, Peat TS, Newman JM, Chu K, Berendzen J. 1998. Class-directed structure determination: foundation for a protein structure initiative. *Protein Sci 7*:1851-1856.

Thompson JD, Higgins DG, Gibson TJ. 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 JD, Plewniak F, Poch O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res 27*:2682-2690.

Topham CM, Srinivasan N, Thorpe CJ, Overington JP, Kalsheker NA. 1994. Comparative modelling of major house dust mite allergen Der p I: structure validation using an extended environmental amino acid propensity table. *Protein Eng 7*:869-894.

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

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

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

Vakser IA. 1997. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins Suppl 1*:226-230.

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

Vitkup D, Melamud E, Moult J, 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 TN, Thanki N, Ravichandran V, Gilliland GL, Bluhm W, Weissig H, Greer DS, Bourne PE, Berman HM. 2002. The Protein Data Bank: unifying the archive. *Nucleic Acids Res 30*:245-248.

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

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

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

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

Yang AS, Honig B. 2000. An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. *J Mol Biol 301*:691-711.

Zarembinski TI, Hung LW, Mueller-Dieckmann HJ, Kim KK, Yokota H, Kim R, Kim SH. 1998. Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc Natl Acad Sci U S A 95*:15189-15193.