Comparative Protein Structure Modeling of Genes And Genomes

Xuesong Hu 胡雪松

School of life sciences, Peking University

Catalogue

- What is comparative protein structure modeling?
- Why could we do comparative modeling?
- Why is comparative modeling important?
- How to do comparative modeling?

  Fold assignment and template selection

  Target – template alignment

  Model building

  Model evaluation

- The application of comparative modeling
- Comparative modeling in structural genomics

1. What Is Comparative Protein Structure Modeling?

- ***Comparative protein structure modeling*** predicts the three-dimensional structure for a given protein sequence of unknown structure (target) on the basis of sequence similarity to proteins of known structure (the templates).

2. Why Could We Do Comparative Modeling?

- Small changes in the protein sequence usually result in small changes in its 3D structure. If similarity between two proteins is detectable at the sequence level, structural similarity can usually be assumed.

- The number of unique structural folds that proteins adopt is limited and because the number of experimentally determined new structures is increasing exponentially.

3. Why Comparative Modeling Is Important?

- ***It is an efficient way to obtain useful information about the proteins of interest.***

- Designing mutants to test hypotheses about a protein’s function
- Identifying active and binding
- Identifying, designing and improving ligands for a given binding site
- Modeling substrate specificity
- Predicting antigenic epitopes
- Simulating protein–protein docking
- Inferring function from a calculated electrostatic potential around the protein
- Facilitating molecular replacement in x-ray structure determination
- Refining models based on NMR constraints
- Testing and improving a sequence-structure alignment
- Confirming a remote structural relationship
- Rationalizing known experimental observations.

4. How To Do Comparative Modeling?

- Fold assignment and template selection
- Target – template alignment
- Model Building
- Model evaluation

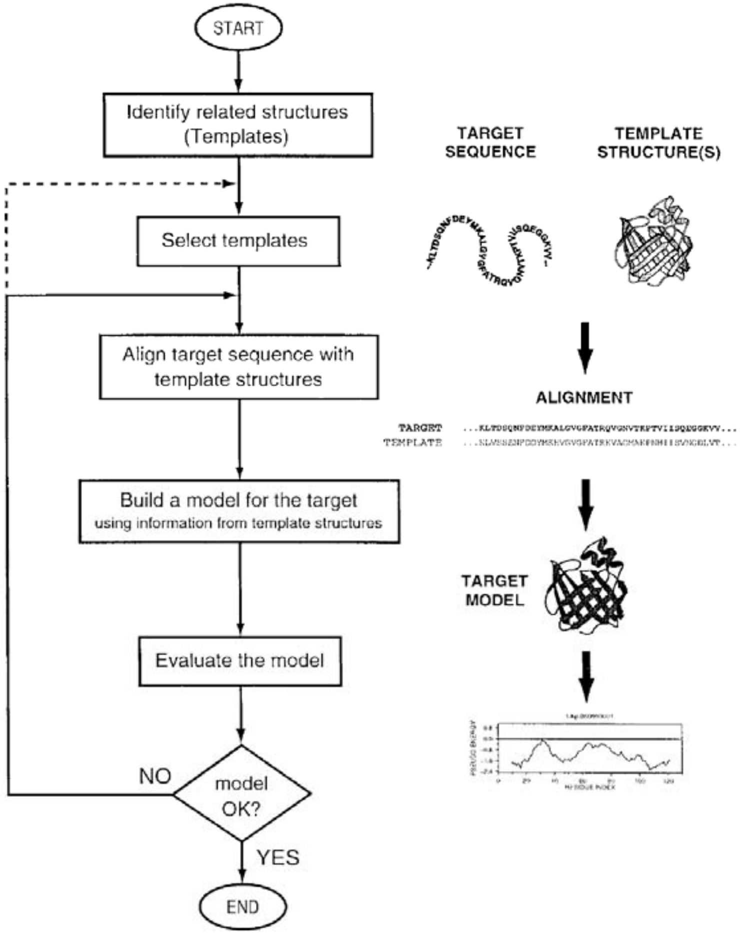

START
Identify related structures (Templates)
Select templates
Align target sequence with template structures
Build a model for the target using information from template structures
Evaluate the model
NO
model OK?
YES
END
TARGET SEQUENCE
TEMPLATE STRUCTURE(S)
ALIGNMENT
TARGET ...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...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

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

4.1 Fold Assignment And Template Selection

- ***Three main classes of protein comparison methods :***

1. Comparing the target sequence with each of the database sequences independently. *Program : BLAST, FASTA etc.*

2. Using multiple sequence comparisons to improve the sensitivity of the search. *Program : PSI-BLAST etc.*

   **especially useful when the sequencing identity below 25%*

3. Threading or 3D template matching methods.

   **especially useful when there are no sequences clearly related to the modeling target.*

4.1 Fold Assignment And Template Selection

- **Template selection :**

  A higher sequence similarity,

  The family of proteins,

  The quality of template structure,

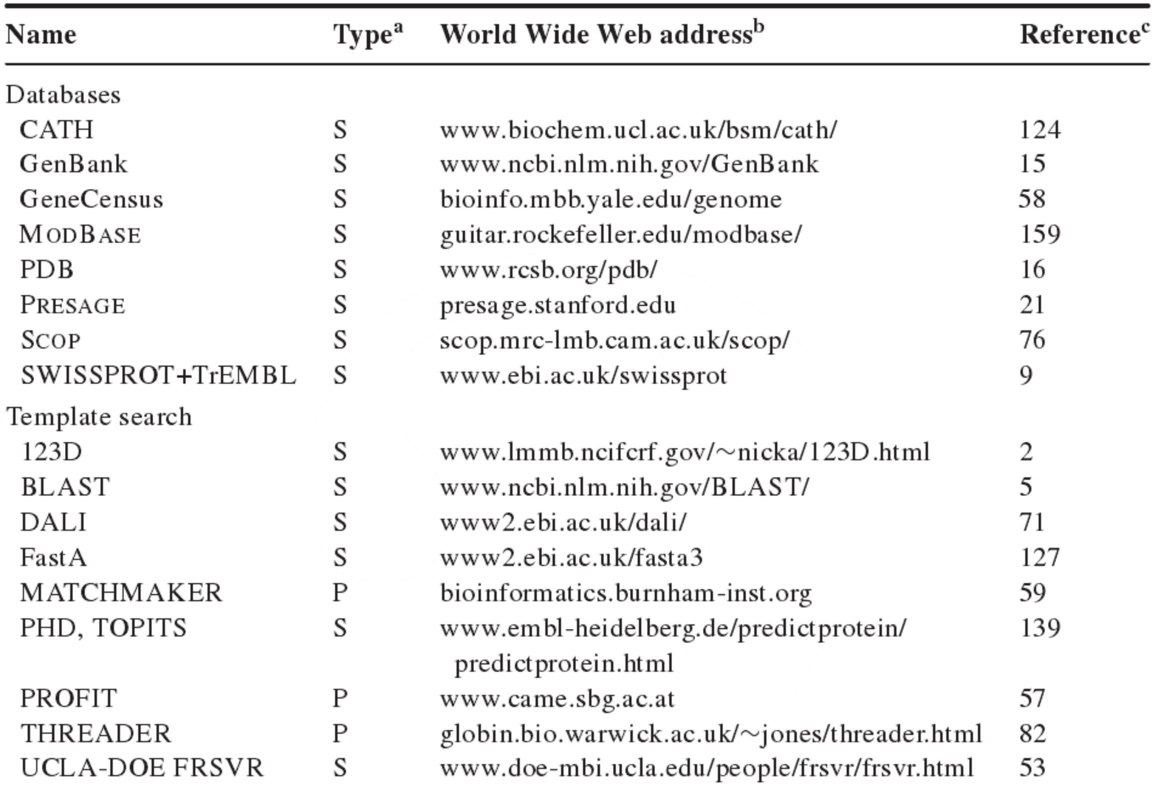
  Solvent, pH, ligands…

- **Potential problems:**

  Distantly related proteins used as templates (i.e., less than 25% sequence identity) may produce an unreliable model.

4.1 Fold Assignment And Template Selection

- *The databases and Programs you may use in this step:*



Name	Type[a]	World Wide Web address[b]	Reference[c]
Databases			
CATH	S	www.biochem.ucl.ac.uk/bsm/cath/	124
GenBank	S	www.ncbi.nlm.nih.gov/GenBank	15
GeneCensus	S	bioinfo.mbb.yale.edu/genome	58
ModBase	S	guitar.rockefeller.edu/modbase/	159
PDB	S	www.rcsb.org/pdb/	16
Presage	S	presage.stanford.edu	21
Scop	S	scop.mrc-lmb.cam.ac.uk/scop/	76
SWISSPROT+TrEMBL	S	www.ebi.ac.uk/swissprot	9
Template search			
123D	S	www.lmmb.ncifcrf.gov/~nicka/123D.html	2
BLAST	S	www.ncbi.nlm.nih.gov/BLAST/	5
DALI	S	www2.ebi.ac.uk/dali/	71
FastA	S	www2.ebi.ac.uk/fasta3	127
MATCHMAKER	P	bioinformatics.burnham-inst.org	59
PHD, TOPITS	S	www.embl-heidelberg.de/predictprotein/ predictprotein.html	139
PROFIT	P	www.came.sbg.ac.at	57
THREADER	P	globin.bio.warwick.ac.uk/~jones/threader.html	82
UCLA-DOE FRSVR	S	www.doe-mbi.ucla.edu/people/frsvr/frsvr.html	53

[a] S, server , P, program

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

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

4.2 Target – Template Alignment

- Once templates have been selected, a specialized method should be used to align the target sequence with the template structures. *Program : CLUSTAL etc.*

- The alignment becomes difficult in the "twilight zone" of less than 30% sequence identity. (Only 20% of the residues are likely to be correctly aligned when two proteins share 30% sequence.)

Similarity of BLOSUM62 is 62%, also ~45 & ~80.

4.2 Target – Template Alignment

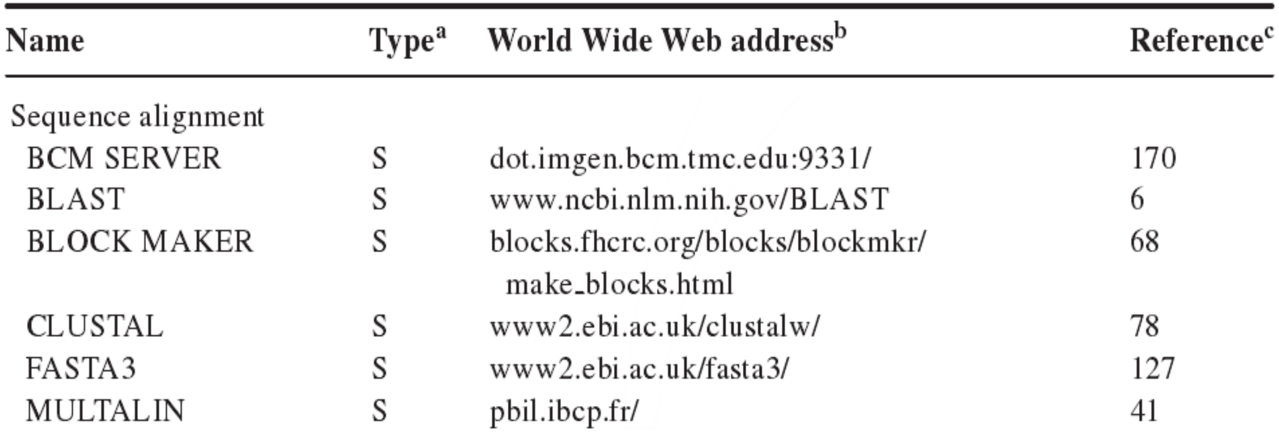
- In difficult cases, it is frequently beneficial to rely on multiple structure and sequence information. The information from structures helps to avoid gaps in secondary structure elements, in buried regions, or between two residues that are far in space.

- **Potential problems:**

  Although you can use the methods aforementioned, misalignment may occur especially when the target-template sequence identity decreases below 30%.

4.2 Target – Template Alignment

- *Programs and World Wide Web servers you may use in this step:*



Name	Type[a]	World Wide Web address[b]	Reference[c]
Sequence alignment			
BCM SERVER	S	dot.imgen.bcm.tmc.edu:9331/	170
BLAST	S	www.ncbi.nlm.nih.gov/BLAST	6
BLOCK MAKER	S	blocks.fhcrc.org/blocks/blockmkr/make_blocks.html	68
CLUSTAL	S	www2.ebi.ac.uk/clustalw/	78
FASTA3	S	www2.ebi.ac.uk/fasta3/	127
MULTALIN	S	pbil.ibcp.fr/	41

[a] S, server , P, program

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

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

4.3 Model Building

- ***Three classes of methods can be used to construct a 3D model:***

1. Modeling by Assembly of Rigid(刚性的) Bodies

   Assemble a model from a small number of rigid bodies obtained from aligned protein structures.

2. Modeling by Segment Matching or Coordinate Reconstruction

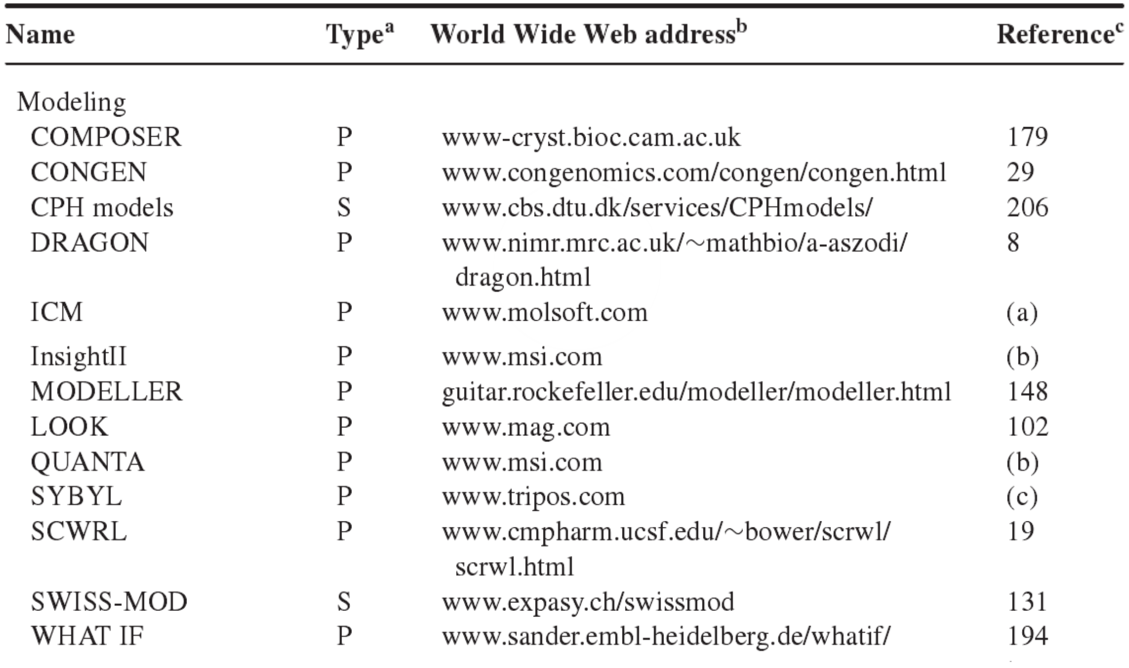
   Use a subset of atomic positions from template structures as “guiding” positions, and by identifying and assembling short, all-atom segments that fit these guiding positions.

3. Modeling by Satisfaction of Spatial(空间的) Restraints(约束)

   Generate many constraints or restraints on the structure of the target sequence, using its alignment to related protein structures as a guide.

4.3 Model Building

- *Programs and World Wide Web servers you may use in this step:*



Name	Type[a]	World Wide Web address[b]	Reference[c]
Modeling			
COMPOSER	P	www-cryst.bioc.cam.ac.uk	179
CONGEN	P	www.congenomics.com/congen/congen.html	29
CPH models	S	www.cbs.dtu.dk/services/CPHmodels/	206
DRAGON	P	www.nimr.mrc.ac.uk/~mathbio/a-aszodi/dragon.html	8
ICM	P	www.molsoft.com	(a)
InsightII	P	www.msi.com	(b)
MODELLER	P	guitar.rockefeller.edu/modeller/modeller.html	148
LOOK	P	www.mag.com	102
QUANTA	P	www.msi.com	(b)
SYBYL	P	www.tripos.com	(c)
SCWRL	P	www.cmpharm.ucsf.edu/~bower/scrwl/scrwl.html	19
SWISS-MOD	S	www.expasy.ch/swissmod	131
WHAT IF	P	www.sander.embl-heidelberg.de/whatif/	194

[a] S, server , P, program

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

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

4.3.1 Loop Modeling

- Loops often determine the functional specificity of a given protein framework. They contribute to active and binding sites.
- Loop modeling can be seen as a mini–protein folding problem, but they are generally too short to provide sufficient information about their local fold.
- ***Three methods:***

  1) *Ab initio* methods

  2) Database search techniques

  3) Both

4.3.2 Sidechain Modeling

- Side chain conformations are predicted from similar structures and from steric(立体的) or energetic considerations.
- They are modeled using structural information from proteins in general and from equivalent disulfide(二硫) bridges in related structures.
- ***Two effects on sidechain conformation:***

  1) The coupling between the main chain and side chains

  2) The continuous nature of the distributions of side-chain dihedral angles(二面角)

- ***Three different side-chain prediction methods*** :

  1)The packing of backbone-dependent rotamers(旋转异构体)

  2)The self-consistent mean-field approach to positioning rotamers based on their **van der Waals** interactions

  3)The segment-matching method of Levitt

4.3.3 Potential Problems

- According to a recent survey analyzed the accuracy of 3 modeling methods, they can only correctly predict approximately 50% of χ_1 angles and 35% of both χ_1 and χ_2 angles.
- 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, especially when the insertion is more than 9 residues long.
- Some correctly aligned segments of a model, the template is locally different (<3 A˚) from the target, resulting in errors in that region.
- As the sequences diverge, the packing of side chains in the protein core may changes.

4.4 Model Evaluation

- ***Typical errors in comparative models :***

  1. Errors in side-chain packing
  2. Distortions and shift in correctly aligned regions.
  3. Errors in regions without a template
  4. Errors due to misalignments
  5. Incorrect template.

4.4 Model Evaluation

- **Typical errors in comparative models :**

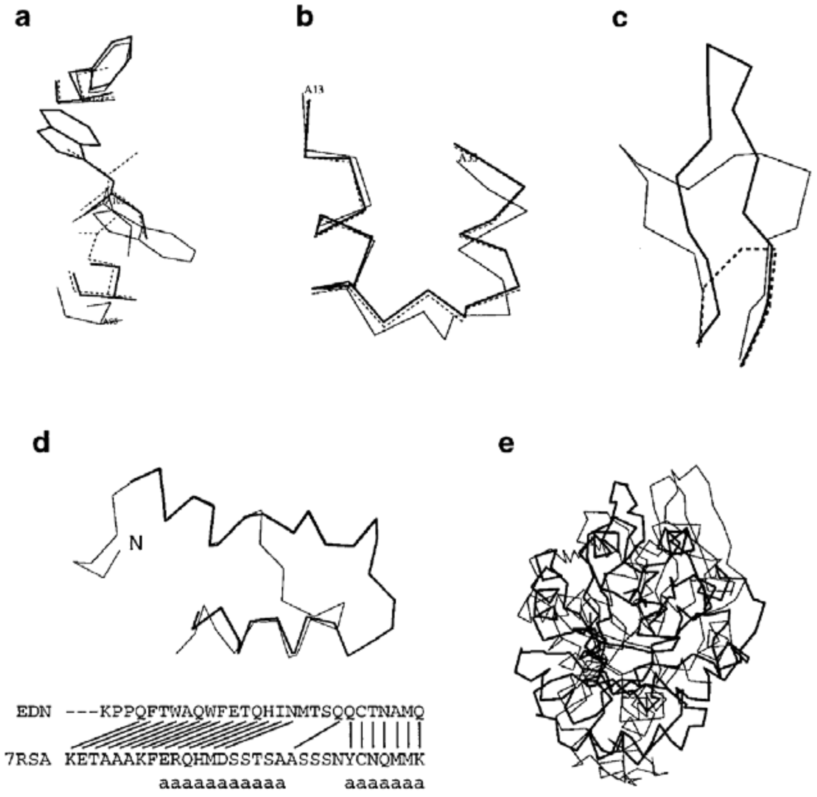
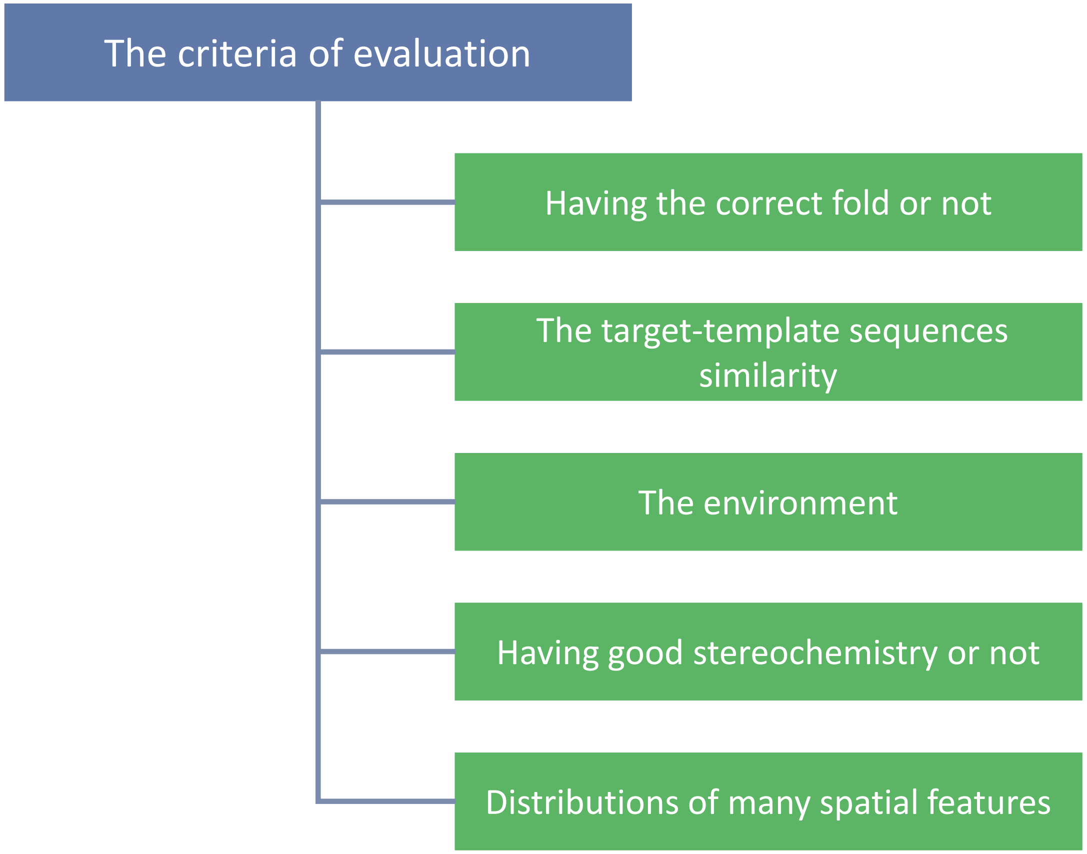

a
b
c
A13
d
N
e
EDN ---KPPQFTWAQWFETQHINMTSQQCTNAMQ
7RSA KETAAAKFERQHMDSSTSAASSSNYCNQMMK
aaaaaaaaaaa aaaaaaa

Figure 2 Typical errors in comparative modeling (151, 156). (*a*). Errors in side chain packing. The Trp 109 residue in the crystal structure of mouse cellular retinoic acid binding protein I (thin line) is compared with its model (thick line), and with the template mouse adipocyte lipid-binding protein (broken line). (*b*) Distortions and shifts in correctly aligned regions. A region in the crystal structure of mouse cellular retinoic acid binding protein I is compared with its model and with the template fatty acid binding protein using the same representation as in panel a. (*c*) Errors in regions without a template. The C_α trace of the 112–117 loop is shown for the X-ray structure of human eosinophil neurotoxin (thin line), its model (thick line), and the template ribonuclease A structure (residues 111–117; broken line). (*d*) Errors due to misalignments. The N-terminal region in the crystal structure of human eosinophil neurotoxin (thin line) is compared with its model (thick line). The corresponding region of the alignment with the template ribonuclease A is shown. The black lines show correct equivalences, that is residues whose C_α atoms are within 5 Å of each other in the optimal least-squares superposition of the two X-ray structures. The "a" characters in the bottom line indicate helical residues. (*e*) Errors due to an incorrect template. The X-ray structure of α-trichosanthin (thin line) is compared with its model (thick line) which was calculated using indole-3-glycerophosphate synthase as the template.

4.4 Model Evaluation



The criteria of evaluation

Having the correct fold or not

The target-template sequences similarity

The environment

Having good stereochemistry or not

Distributions of many spatial features

4.4 Model Evaluation

1) Having the correct fold or not

A model will have the correct fold if the correct template is picked and if that template is aligned at least approximately correctly with the target sequence.

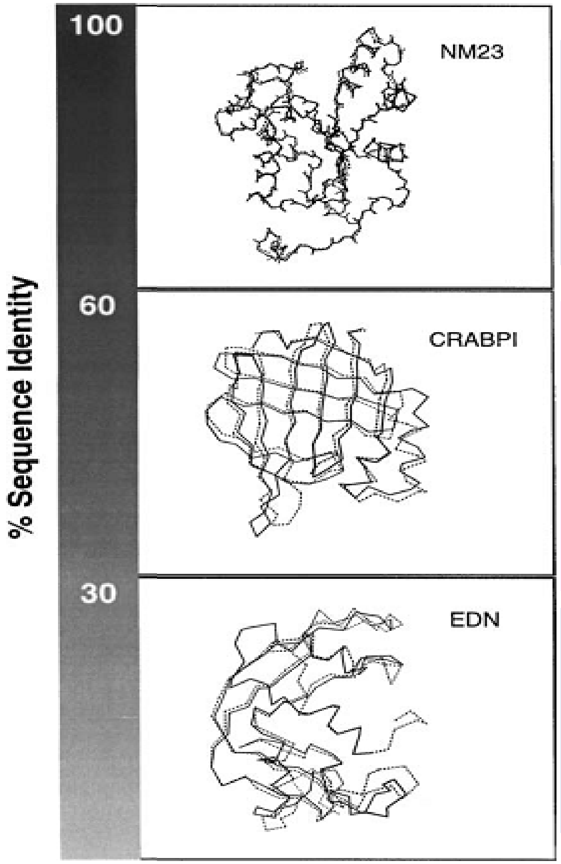
The fold of a model can be assessed by a high sequence similarity with the closest template, an energy based Z-score, or by conservation of the key functional or structural residues in the target sequence.

2) The target-template sequences similarity

Sequence identity above 30% is a relatively good predictor of the expected accuracy.

4.4 Model Evaluation

Average model accuracy as a function of the template-target sequences similarity



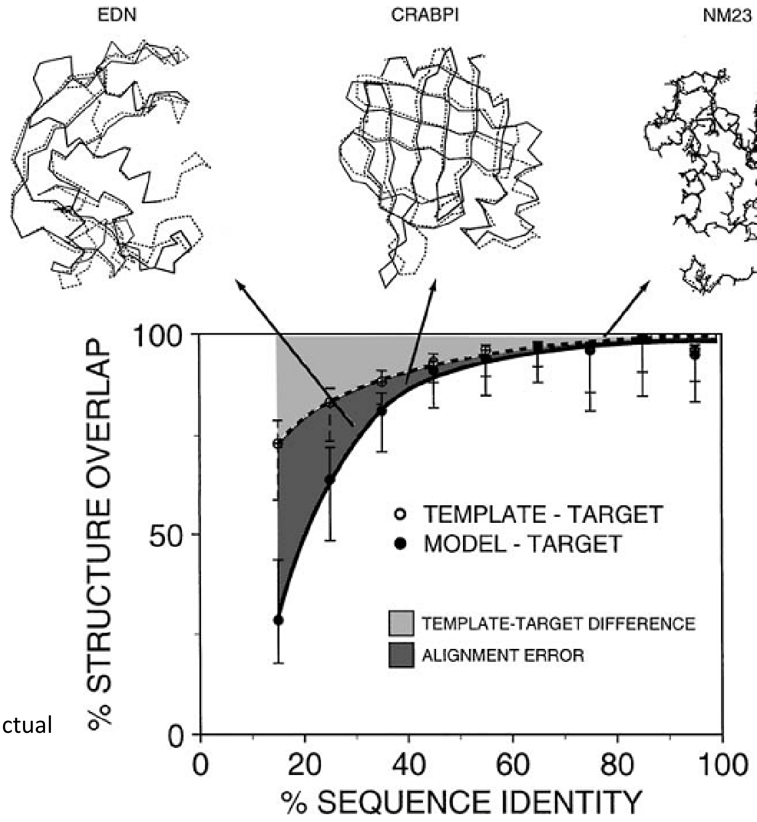
NM23H2: Human nucleoside diphosphate kinase, is a single domain protein consisting of a central 4-stranded antiparallel β-sheet surrounded by 8 α-helices. For the NM23H2 model, all but one C_α atom superpose within 3.5 Å of the X-ray structure; rms difference is 0.41 Å.

CRABPI: mouse cellular retinoic acid binding protein I, is a single domain protein composed of interacting α-helices packed at the edge of two orthogonal, 4- and 6-stranded antiparallel β-sheets. For the CRABPI model, 90% of C_α atoms superpose within 3.5 Å of their counterparts in the X-ray structure; the rms error is 1.31 Å.

EDN: human eosinophil neurotoxin, is a ribonuclease with 3 α-helices and 2 three-stranded antiparallel β-sheets arranged in a single domain.

4.4 Model Evaluation

Average model accuracy as a function of the template-target sequences similarity



Solid line:
sample models

Dotted line: corresponding actual structures

- Percentage structure overlap is defined as the fraction of equivalent residues.
- Two residues are equivalent when their C_α atoms are within 3.5 Å of each other upon rigid-body, least-squares superposition of the two structures.

4.4 Model Evaluation

3) The environment

Example: 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.

4) Having good stereochemistry or not

Including bond lengths, bond angles, peptide bond and side-chain ring planarities, chirality, main-chain and side-chain torsion angles, and clashes between nonbonded pairs of atoms.

5) Distributions of many spatial features

Such features include packing, formation of a hydrophobic core, residue and atomic solvent acces sibilities, spatial distribution of charged groups, distribution of atom-atom distance, atomic volumes, and main-chain hydrogen bondin.

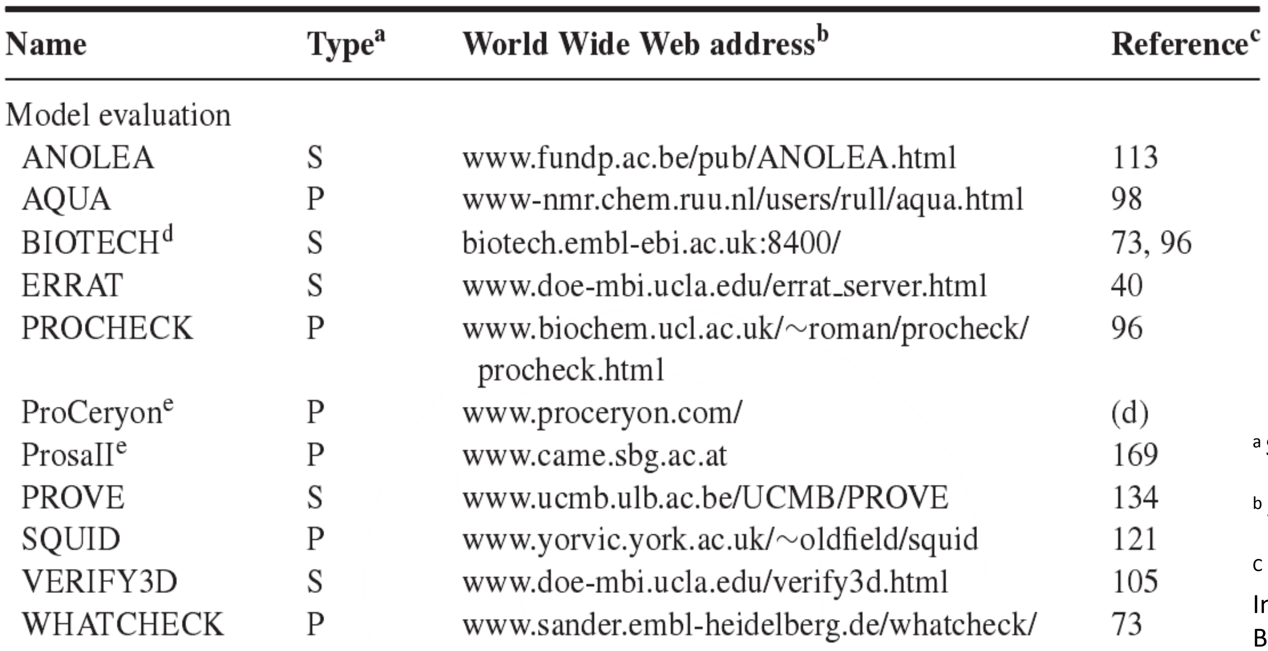
4.4 Model Evaluation

- There are also methods for testing 3D models that implicitly take into account many of the criteria listed above. These methods are based on 3D profiles and statistical potentials of mean force.

- A physics-based approach to deriving energy functions has been tested for use in protein structure evaluation (1999).

4.4 Model Evaluation

- *Programs and World Wide Web servers you may use in this step:*



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

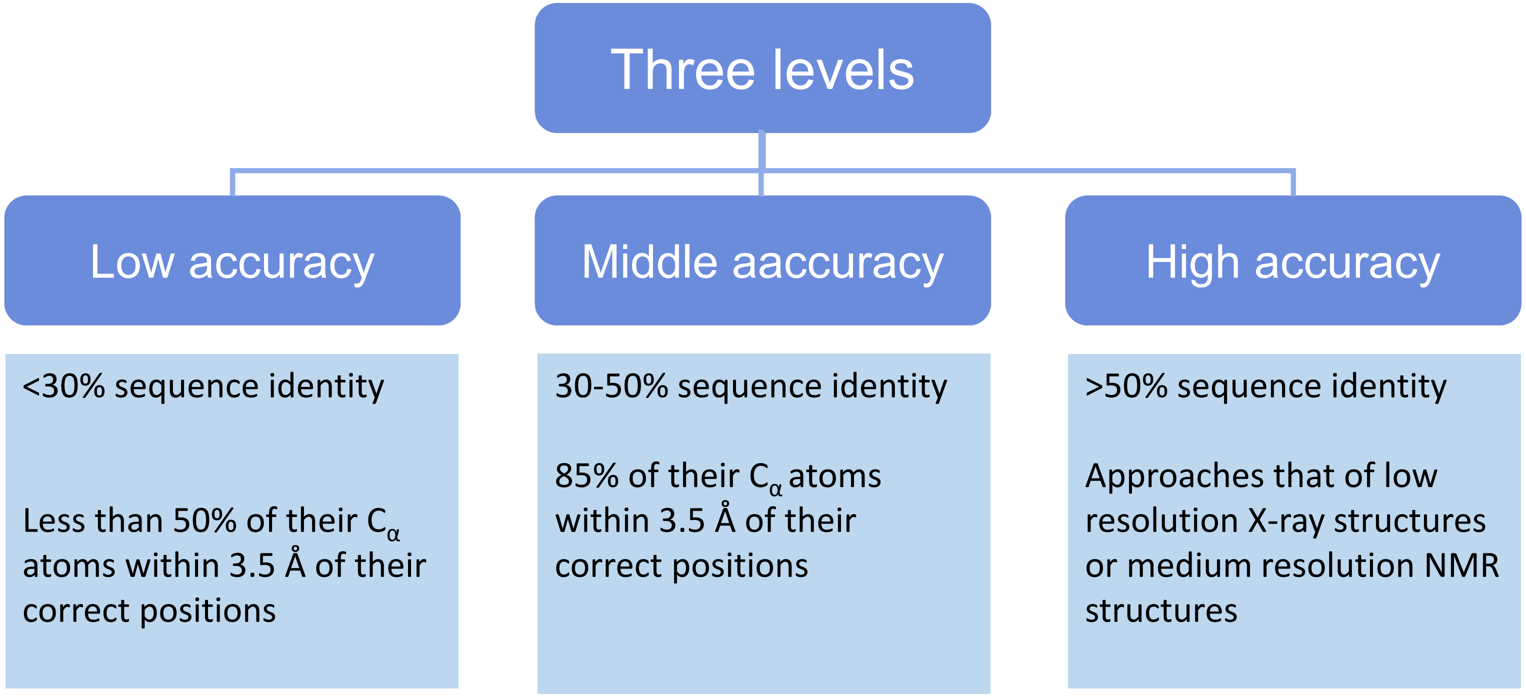
[a] S, server , P, program

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

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

5. The Application of Comparative Modeling

- ***Three levels of model accuracy and some of the corresponding applications***



Three levels

Low accuracy

<30% sequence identity

Less than 50% of their C_α atoms within 3.5 Å of their correct positions

Middle aaccuracy

30-50% sequence identity

85% of their C_α atoms within 3.5 Å of their correct positions

High accuracy

>50% sequence identity

Approaches that of low resolution X-ray structures or medium resolution NMR structures

r_w (van der Waals radius) of C atom = 1.70Å

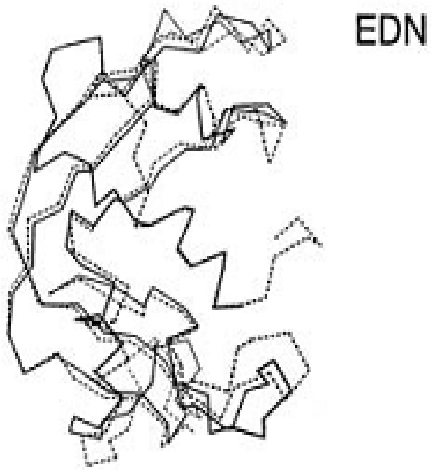
5. The Application of Comparative Modeling

- ***Applications1: low accuracy models***

- <30% sequence identity, having the correct fold
- Less than 50% of their C_α atoms within 3.5 Å of their correct positions

- ***Use:*** To confirm or reject a match between remotely related proteins

- NMR structure refinement.
- Finding binding/active sites by 3D motif searching.
- Functional annotation by fold assignment.



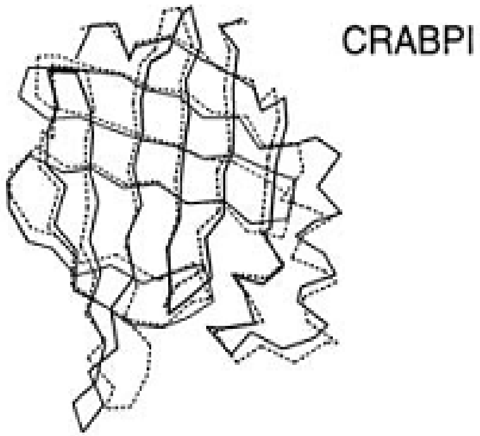
5. The Application of Comparative Modeling

- ***Applications2: middle accuracy models***

- 30-50% sequence identity
- 85% of their C_α atoms within 3.5 Å of their correct positions

- ***Use:*** Refinement of the functional prediction based on sequence to construct site-directed mutants with altered or destroyed binding capacity other problems...

- Molecular replacement in crystallography.
- Engineering of proteins.
- Support site-directed mutagenesis experiments.



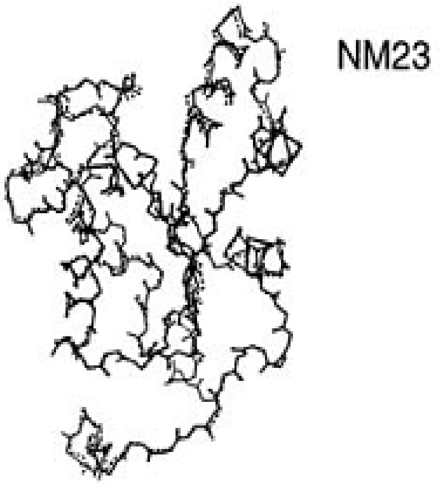
5. The Application of Comparative Modeling

- ***Applications3: high accuracy models***

- >50% sequence identity
- The average accuracy of these models approaches that of low resolution X-ray structures (3 Å resolution) or medium resolution NMR structures (10 distance restraints per residue)

- ***Use:*** For docking of small ligands or whole proteins onto a given protein.

- Comparable to medium resolution NMR.
- Fine specificity.
- Docking of small ligands, proteins.



6. Comparative modeling in structural genomics

- The aim of structural genomics is to determine or accurately predict the 3D structure of all the proteins encoded in the genomes.

- This aim will be achieved by a focused, large-scale determination of protein structures by X-ray crystallography and NMR spectroscopy, combined efficiently with accurate protein structure modeling techniques.

6. Comparative modeling in structural genomics

- For comparative modeling to contribute to structural genomics, automation of all the steps in the modeling process is essential.

- The automation of large-scale comparative modeling involves assembling a software pipeline that consists of modules for fold assignment, template selection, target–template alignment, model generation, and model evaluation.

6. Comparative modeling in structural genomics

- ***Two examples of large-scale comparative modeling for complete genomes:***

the SWISS-MODEL web server:

The sequences encoded in the *E. coli* genome have been used to build models for 10–15% of the proteins using the SWISS-MODEL web server.

MODPIPE:

MODPIPE produced models for five procaryotic and eukaryotic genomes. This calculation resulted in models for substantial segments of 17.2%, 18.1%, 19.2%, 20.4%, and 15.7% of all proteins in the genomes of *Saccharomyces cerevisiae* (6218 proteins in the genome); *Escherichia coli* (4290 proteins), *Mycoplasma genitalium* (468 proteins), *Caenorhabditis elegans* (7299 proteins, incomplete), and *Methanococcus janaschii* (1735 proteins).

6. Comparative modeling in structural genomics

- Large-scale comparative modeling will extend opportunities to tackle a myriad of problems by providing many protein models for many genomes.

Rotein evolution

Drug design

A facile comparison of ligand binding requirements and

Substitutions in and around important residues

……

A specific example:

The selection of a target protein for drug development !

7. 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.

- Further advances are necessary in recognizing weak sequence–structure similarities, aligning sequences with structures, modeling of rigid body shifts, distortions, loops and side chains, as well as detecting errors in a model.

- It is currently possible to model with useful accuracy significant parts of approximately one third of all known protein sequences.

- A major new challenge for comparative modeling is the integration of it with the torrents of data from genome sequencing projects as well as from functional and structural genomics.

Acknowledgements

- Xue-Song Hu
- Qing Shen
- Ying-La Zhang
- Xiao-Meng Liu
- Li Zhou
- Xue-Yuan Tian
- Li-wa Shao
- Rui-Dong Xue
- Meng Wang
- Yang Ding
- Yong-Xin Ye
- Xiao-Xu Yang
- Dr. Ge Gao
- Dr. Li-Ping Wei

Reference

- Martí-Renom M A, Stuart A C, Fiser A, et al. Comparative protein structure modeling of genes and genomes[J]. Annual review of biophysics and biomolecular structure, 2000, 29(1): 291-325.
- Šali A, Potterton L, Yuan F, et al. Evaluation of comparative protein modeling by MODELLER[J]. Proteins: Structure, Function, and Bioinformatics, 1995, 23(3): 318-326.
- Fiser A, Do R K G, Šali A. Modeling of loops in protein structures[J]. Protein science, 2000, 9(9): 1753-1773.

- Fiser A, Do R K G, Šali A. Modeling of loops in protein structures[J]. Protein science, 2000, 9(9): 1753-1773.

- Sánchez R, Šali A. Comparative protein structure modeling in genomics[J]. Journal of Computational Physics, 1999, 151(1): 388-401.

Thanks!