Computational Biochemistry

Lecture 5
Homology Modeling

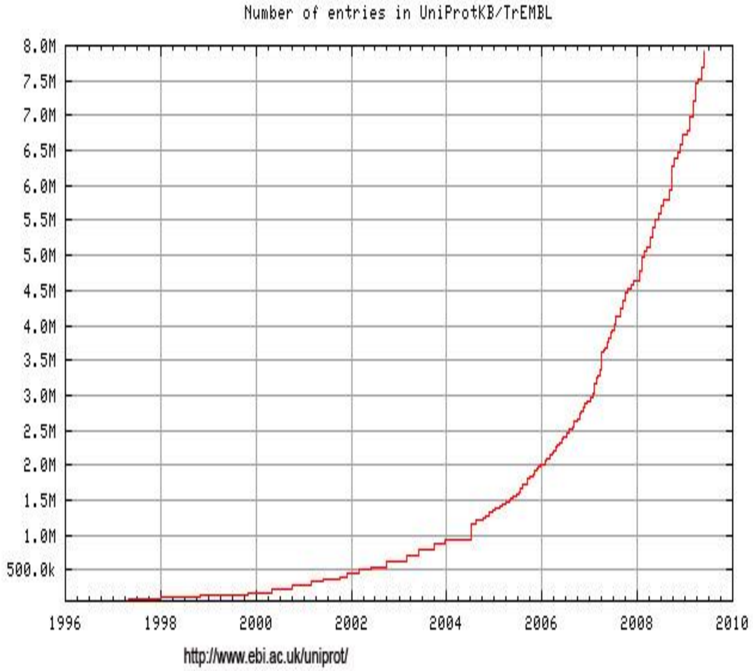
Methods for determining protein 3D structure

Method	Applicability	Strengths	Limitations
X-ray Crystallography	Most proteins as long as they are stable in solution – generally not suitable for membrane bound proteins	Gives the 3D atomic positions from fitting directly to electron density maps Almost no size limitation	• Gives only a static view of the protein • Presents a picture of the protein in a solid state – not in solution • Requires a very good quality crystal • Requires a large amount (~10mg) of pure protein • Hydrogen positions are not detected
NMR-Spectroscopy	Most classes of proteins as long as they are stable in solution. May be applied to membrane bound proteins under certain conditions	Presents a picture of the protein in solution May be used to follow reactions May be used to look at protein dynamics May be used to determine K_D, IC50	• Severe size limitations – protein must be <~50 kDa • 3D structure is inferred from inter-proton "contacts" – that is, the atomic positions are not directly detected • The "structure" represents the average of all conformers present in solution • Requires an isotopically-labelled protein • Still requires significant amounts of protein (~5mg)
Homology Modelling	May be applied to almost any protein as long as there is already something similar deposited in the protein database (PDB)	Requires only the protein sequence	• Uncertain accuracy, depends on homology with known protein 3D structures • Side chain conformation is harder to predict accurately than the overall protein fold • Not well suited to subsequent ligand docking

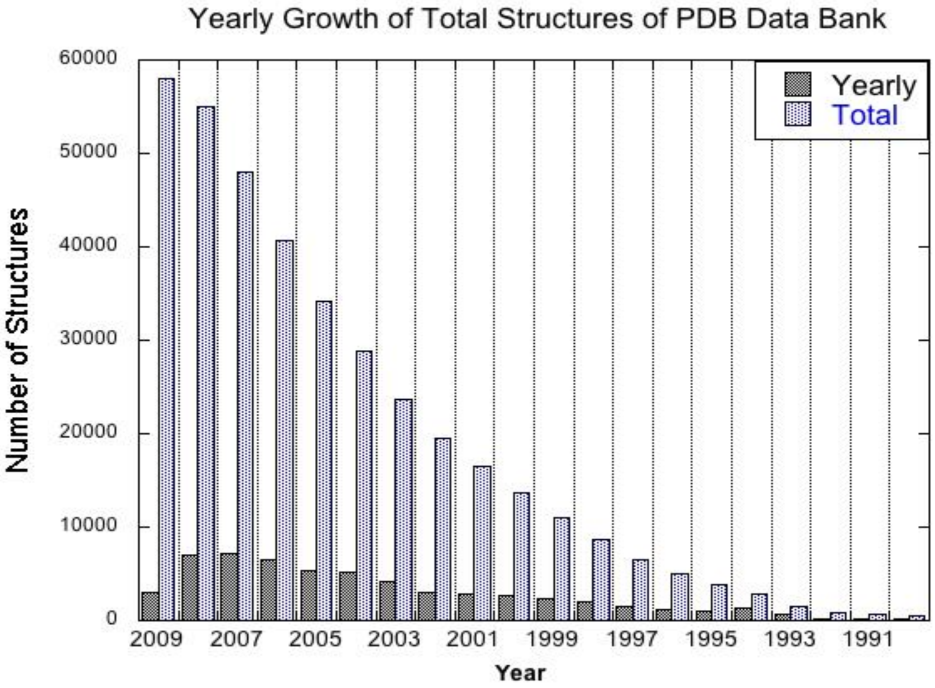
Protein structure prediction

- Protein structure prediction is the inference of the three-dimensional structure of a protein from its amino acid sequence — that is, the prediction of its folding and its secondary and tertiary structure from its primary structure.

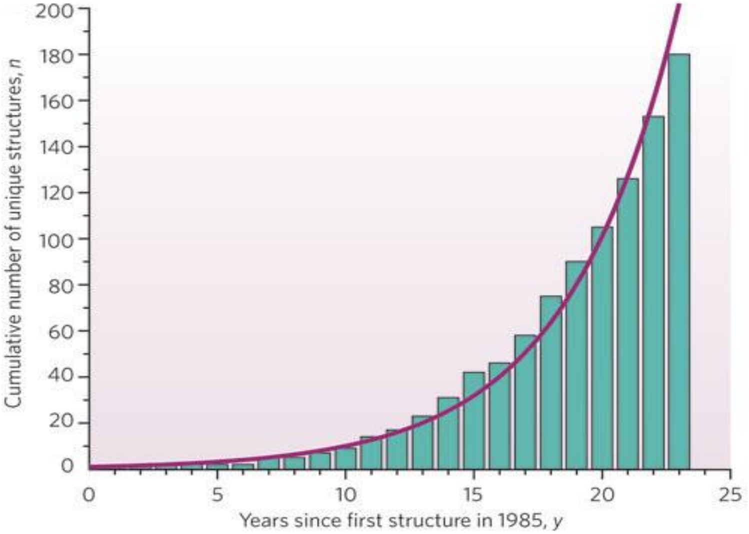
Why Do We Need Homology Modeling?

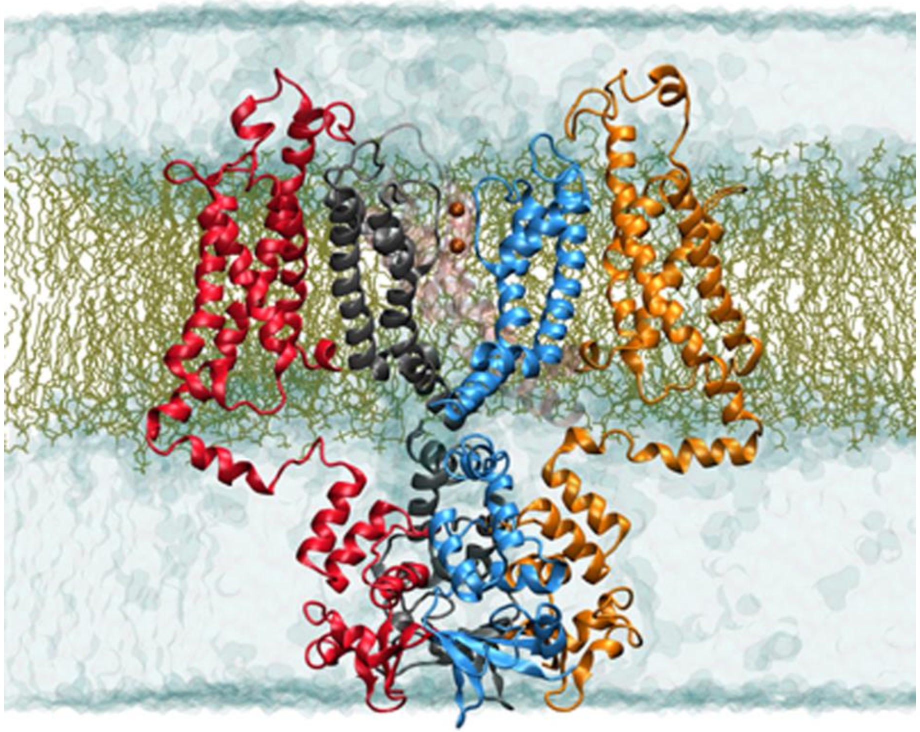


Yearly growth of number of protein sequence entries in UniPort Data Bank.



Progress in Determining Membrane Protein Structures





Stephen H white (2009) Nature, 459,344

Methods of Protein Structure Prediction

Threading and comparative methods

- Sequence similarity
- At least one known structure

De novo or *ab initio* methods

- Predict the structure from sequence alone
- Native state of a protein is at the global energy minimum

Arthur M Lesk (2007) Introduction to genomics, Oxford University Press,USA

Comparative or Homology modeling

The aim is to build a 3-D model for a protein of unknown 3D structure (*target*) on the basis of sequence similarity to proteins of known 3D structure (*templates*).

Accuracy varies from simply identifying the correct fold to generating a high resolution model

Homology modeling is the most accurate protein structure prediction method – but that doesn’ t mean it works perfectly!

- 3D structures of proteins in a given family are more conserved than their sequences
- Approximately 1/3 of all sequences are recognizably related to at least one known structure
- The number of unique protein folds is limited

Key concepts

Homology Modeling Stages:

1) Protein Sequence Alignment
2) Model Building
   - a. Fold selection/generation
   - b. Side chain positioning
   - c. Loop generation
   - d. Energy optimization
3) Evaluating the Model

Steps in homology modeling

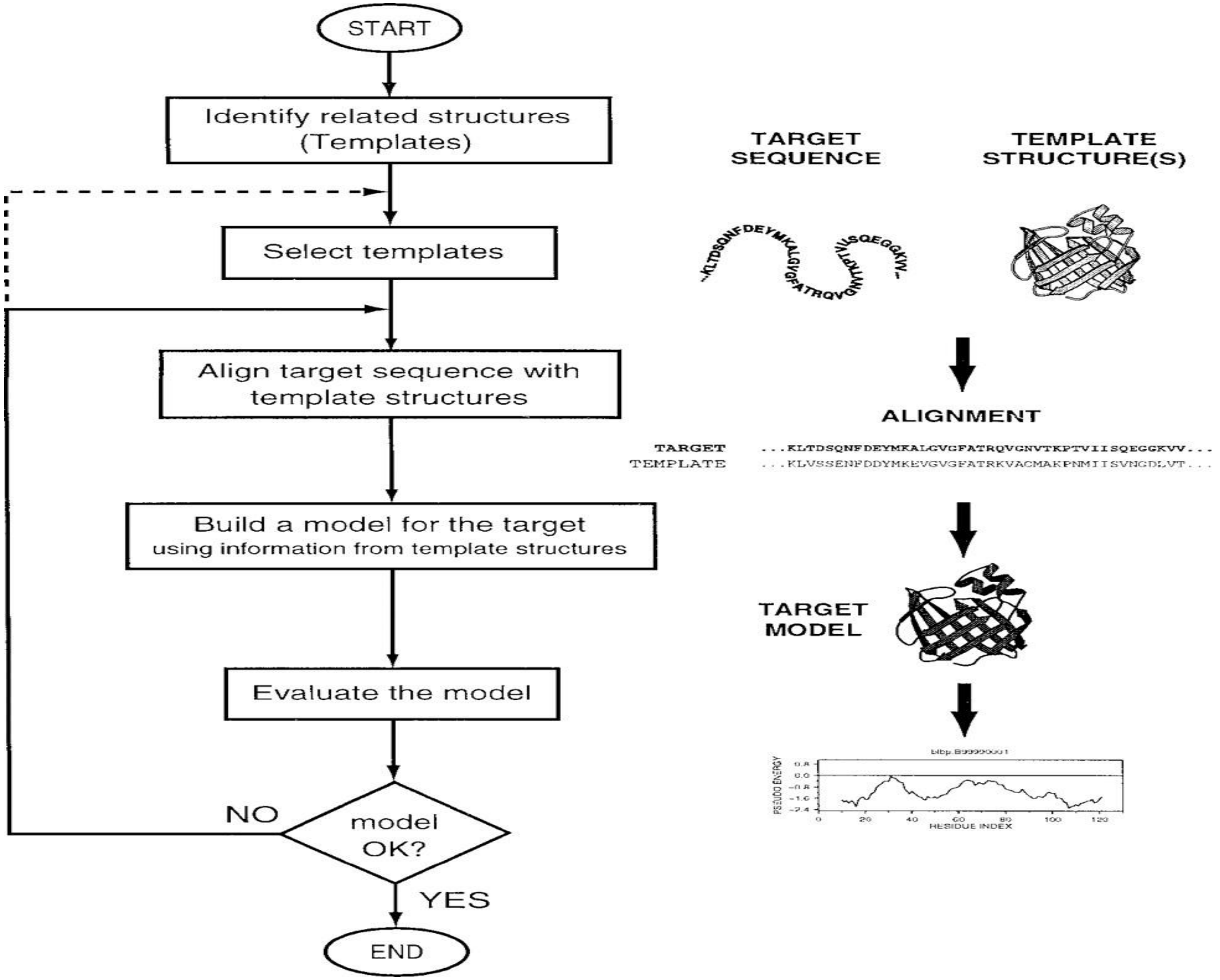

START
Identify related structures (Templates)
Select templates
Align target sequence with template structures
Build a model for the target using information from template structures
Evaluate the model
NO
model OK?
YES
END
TARGET SEQUENCE
TEMPLATE STRUCTURE(S)
...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
ALIGNMENT
TARGET ...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
TEMPLATE ...KLVSSENFDDYMKEVGVGFATRKVACMAKPNMIISVNGDLVT...
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 from MA Marti-Renom and A. Sali "Modeling Protein Structure from Its Sequence" *Current Prototocols in Bioinformatics (2003).* 5.1.1-5.1.32

Steps in homology modeling

Crucial

Steps

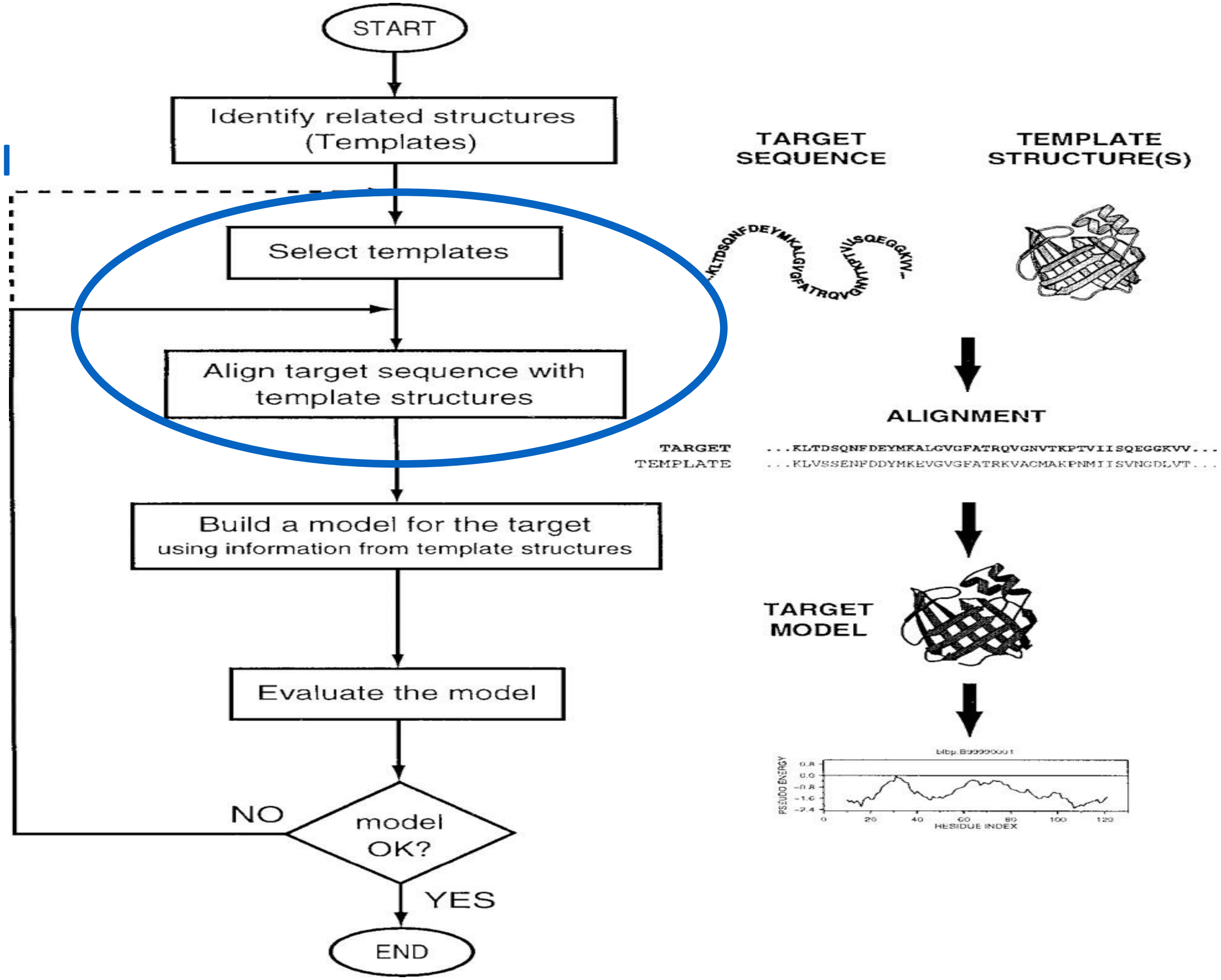
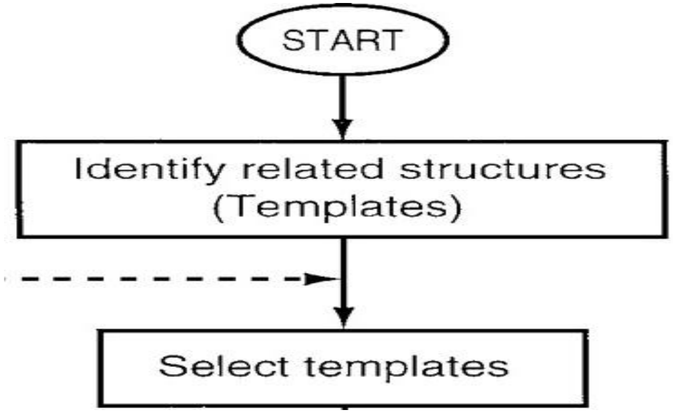

START
Identify related structures (Templates)
Select templates
Align target sequence with template structures
Build a model for the target using information from template structures
Evaluate the model
NO
model OK?
YES
END
TARGET SEQUENCE
TEMPLATE STRUCTURE(S)
...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
ALIGNMENT
TARGET ...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
TEMPLATE ...KLVSSENFDDYMKEVGVGFATRKVACMAKPNMIISVNGDLVT...
TARGET MODEL
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 from MA Marti-Renom and A. Sali “Modeling Protein Structure from Its Sequence” *Current Prototocols in Bioinformatics (2003).* 5.1.1-5.1.32

Step 1: Template Selection



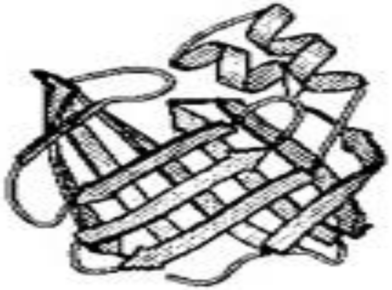
START

Identify related structures (Templates)

Select templates

TARGET SEQUENCE

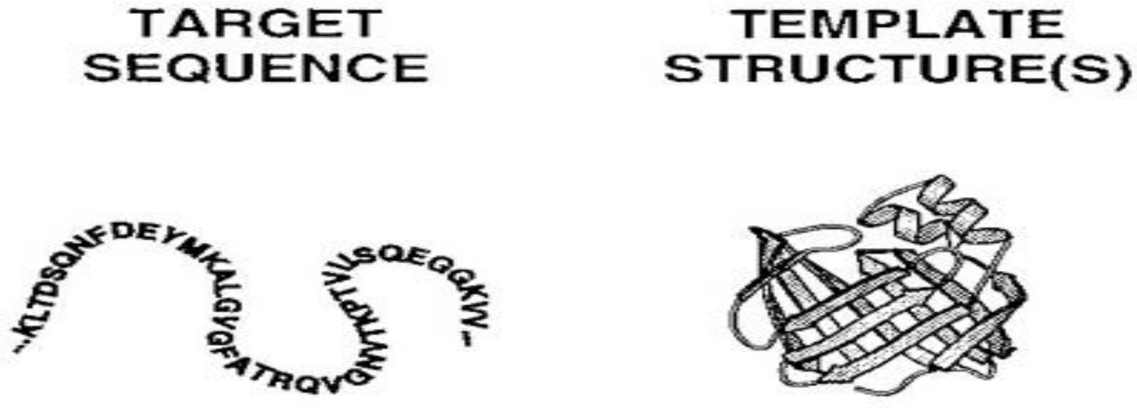
TEMPLATE STRUCTURE(S)



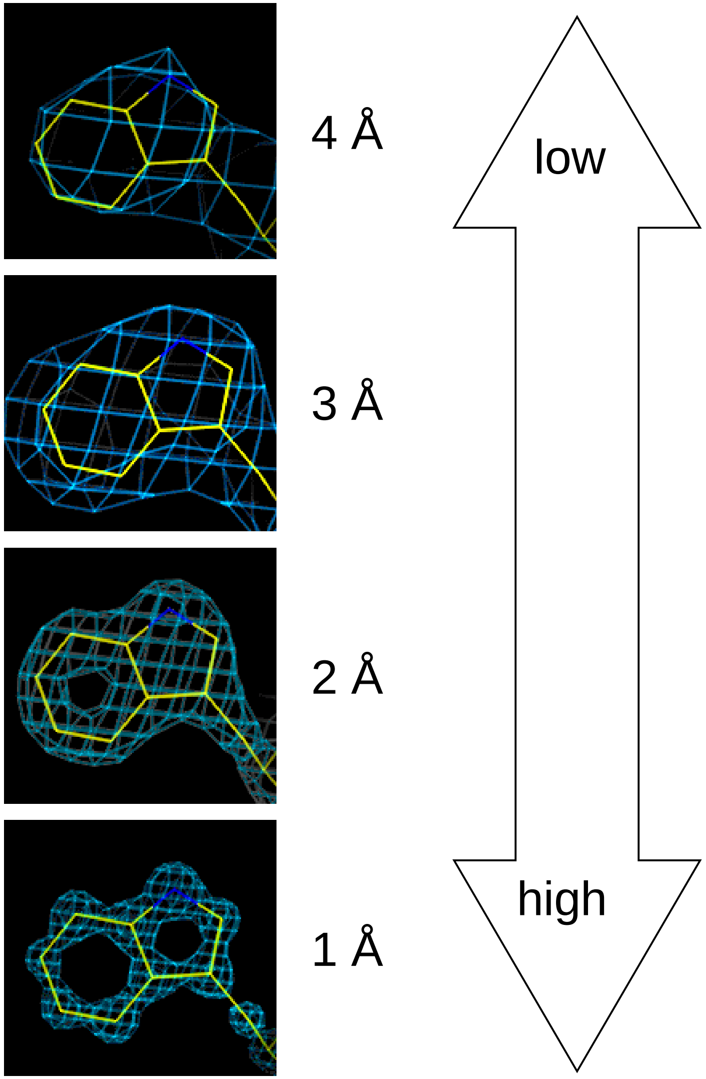
PDB www.rcsb.org/pdb/

Template Search

BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
FastA	http://www.ebi.ac.uk/fasta33/
SSM	http://www.ebi.ac.uk/msd-srv/ssm/
PredictProtein	http://www.predictprotein.org/
123D; SARF2; PDP	http://123d.ncifcrf.gov/
GenTHREADER	http://bioinf.cs.ucl.ac.uk/psipred/
UCLA-DOE	http://fold.doe-mbi.ucla.edu/



The Importance of Resolution



- In X-ray crystallography it is not always possible to flawlessly resolve the crystal density of the protein of interest.
- This results in a lower resolution structure.
- The lower the resolution the more likely the structure is wrong.
- The resolution of the template structure also reflects in the quality of the homology model.

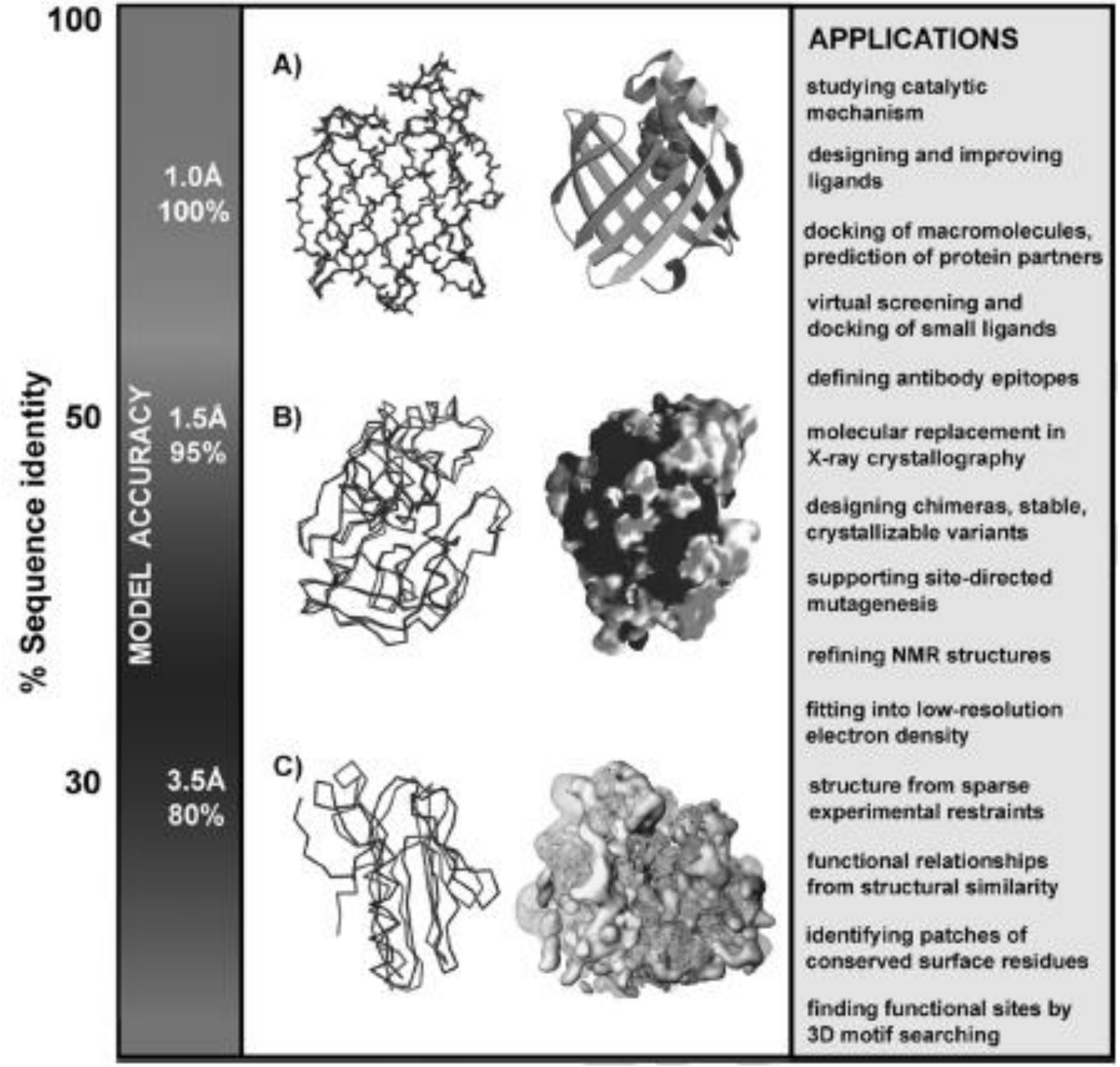
The Importance of Resolution

Quantitative comparison between model and experimental 3D structure using RMSD

- 0.0-0.5 Å → Essentially Identical
- <1.5 Å → Very good fit
- < 5.0 Å → Moderately good fit
- 5.0-7.0 Å → Structurally related
- > 7.0 Å → Dubious relationship
- > 12.0 Å → Completely unrelated

Taken from: http://redpoll.pharmacy.ualberta.ca/bioinfo301/Bioinf301-3Dvisualize.ppt

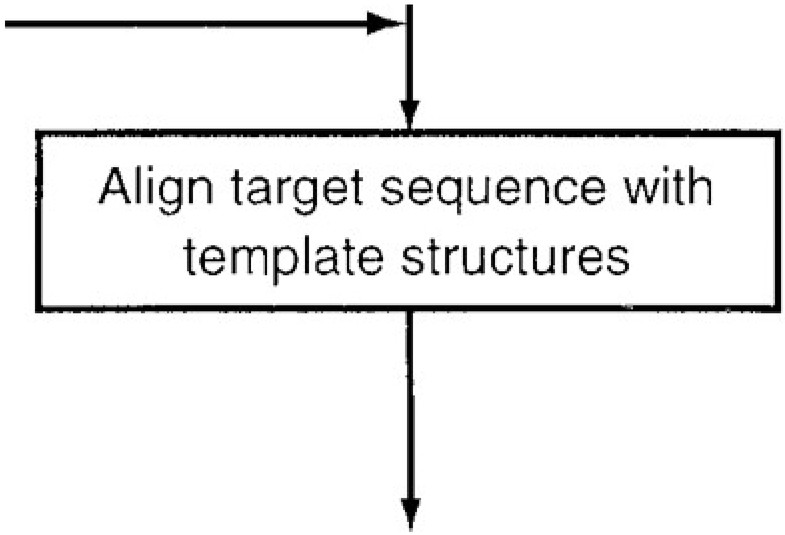
Required accuracy for intended application



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

Step 2: Sequence Alignment



Align target sequence with template structures

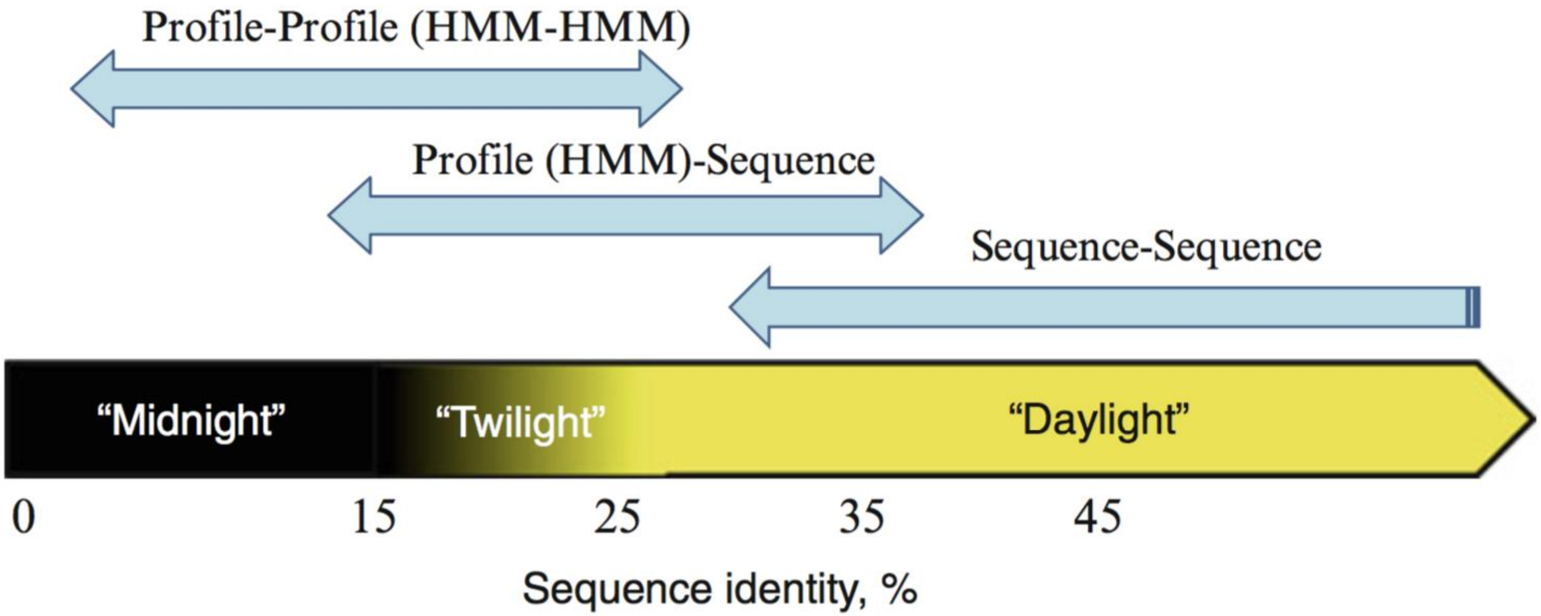
ALIGNMENT

```
TARGET     ...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
TEMPLATE   ...KLVSSENFDDYMKEVGVGFATRKVACMAKPNMIISVNGDLVT...
```

This is the most crucial step in the process

Homology modeling cannot recover from a bad initial alignment

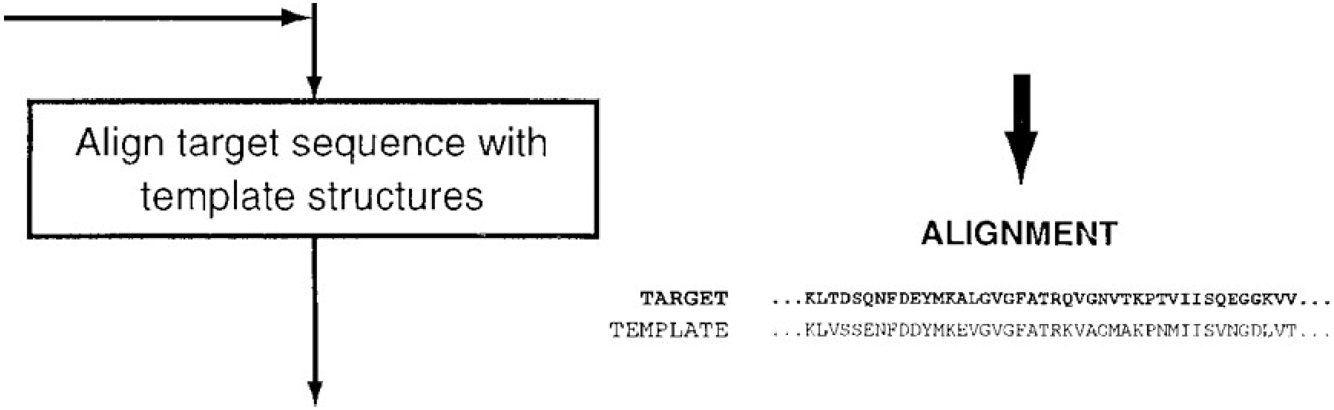
Homology Detection and Alignment Methods



Sequence similarity is partitioned into three approximate intervals corresponding to the decreasing difficulty of identifying homology from sequence: the "midnight" zone (<15% sequence identity), the "twilight" zone (~15–25%), and the "daylight" zone (>25%).

Step 2: Sequence Alignment

EMBOSS	http://www.ebi.ac.uk/emboss/align/
Tcoffee	http://www.igs.cnrs-mrs.fr/Tcoffee
ClustalW	http://www.ebi.ac.uk/clustalw/
SwissModel	http://www.expasy.org/spdbv/
BCM	http://searchlauncher.bcm.tmc.edu/multi-align/
POA	http://www.bioinformatics.ucla.edu/poa/
STAMP	http://www.ks.uiuc.edu/Research/vmd/



ALIGNMENT

```
TARGET     ...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
TEMPLATE   ...KLVSSENFDDYMKEVGVGFATRKVACMAKPNMIISVNGDLVT...
```

Sequence Alignment
Example: α-crystallin in various species

α-crystallin is a water-soluble structural protein found in the lens and the cornea of the eye accounting for the transparency of the structure.

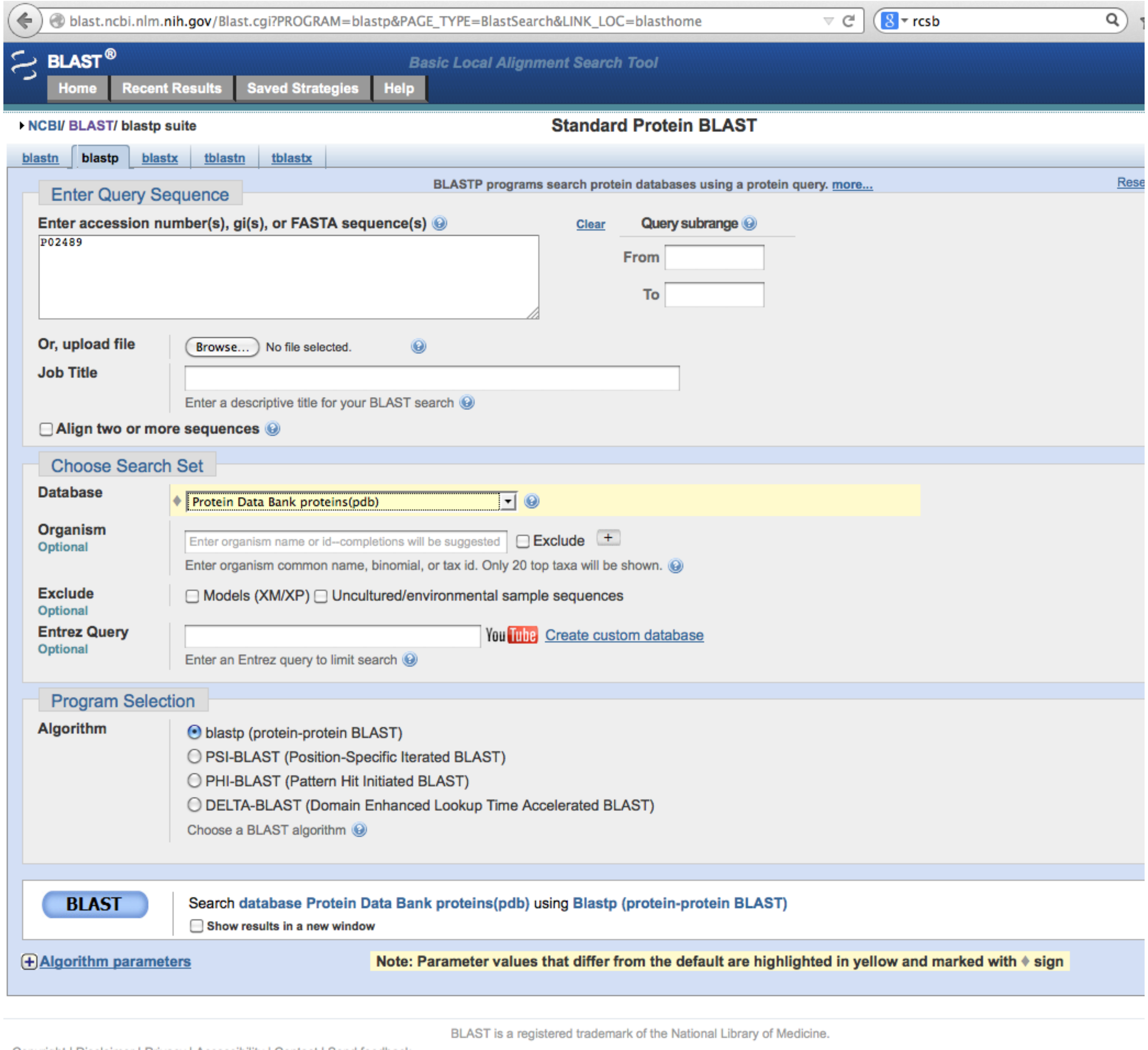
Every reported protein sequence has a unique identifier (!)
α-crystallin: UniProt ID: P02489

What are the differences between human, rhesus monkey and mouse sequences?

Go to BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) and enter the protein ID or sequence

Why do you care? Suppose you have isolated a new α-crystallin and want to know what it looks like. Which of the reported structures is most similar?

Sequence Alignment
Example: α-crystallin in various species



blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

rcsb

BLAST®

Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ blastp suite

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Rese

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

P02489

Clear

Query subrange

From

To

Or, upload file: Browse... No file selected.

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Protein Data Bank proteins(pdb)

Organism Optional: Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query Optional: YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Algorithm

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST

Search database Protein Data Bank proteins(pdb) using Blastp (protein-protein BLAST)

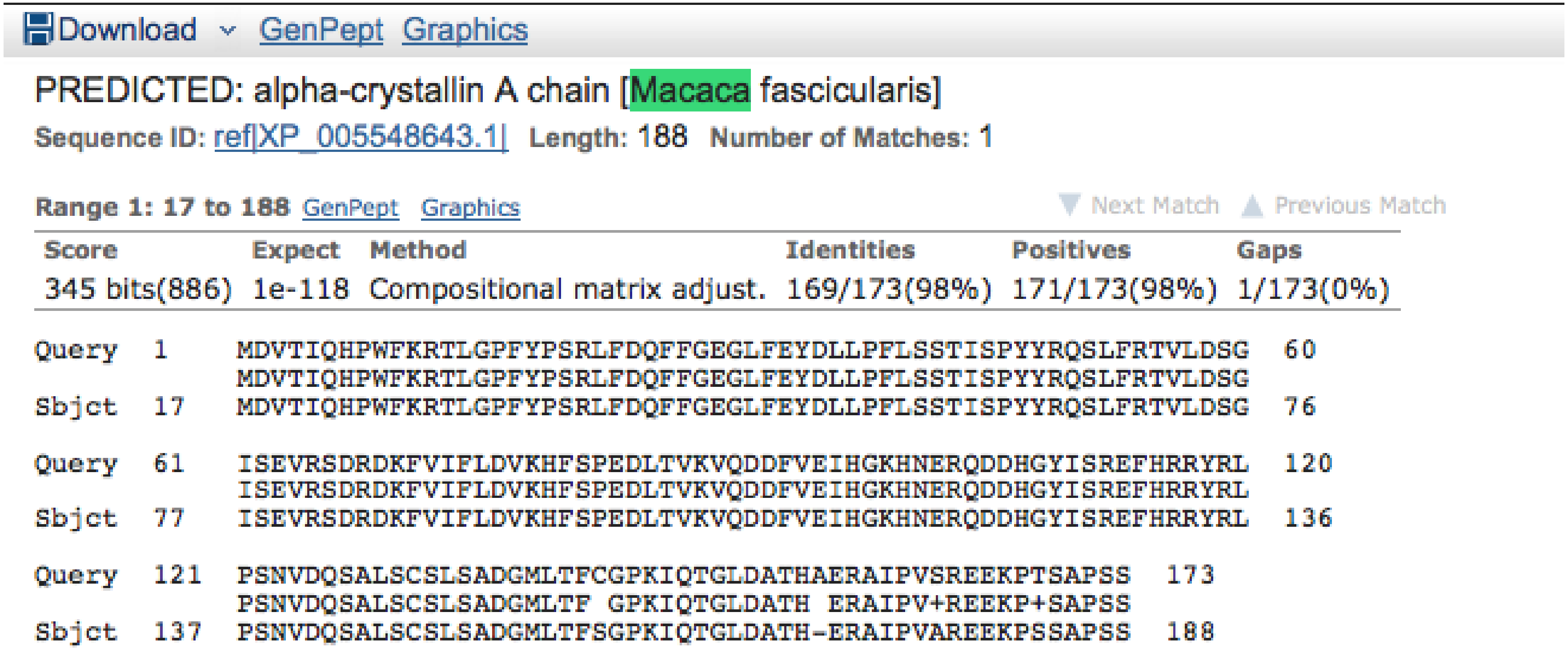
□ Show results in a new window

+ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

BLAST is a registered trademark of the National Library of Medicine.

Sequence Alignment
Example: human α-crystallin versus rhesus monkey



PREDICTED: alpha-crystallin A chain [Macaca fascicularis]

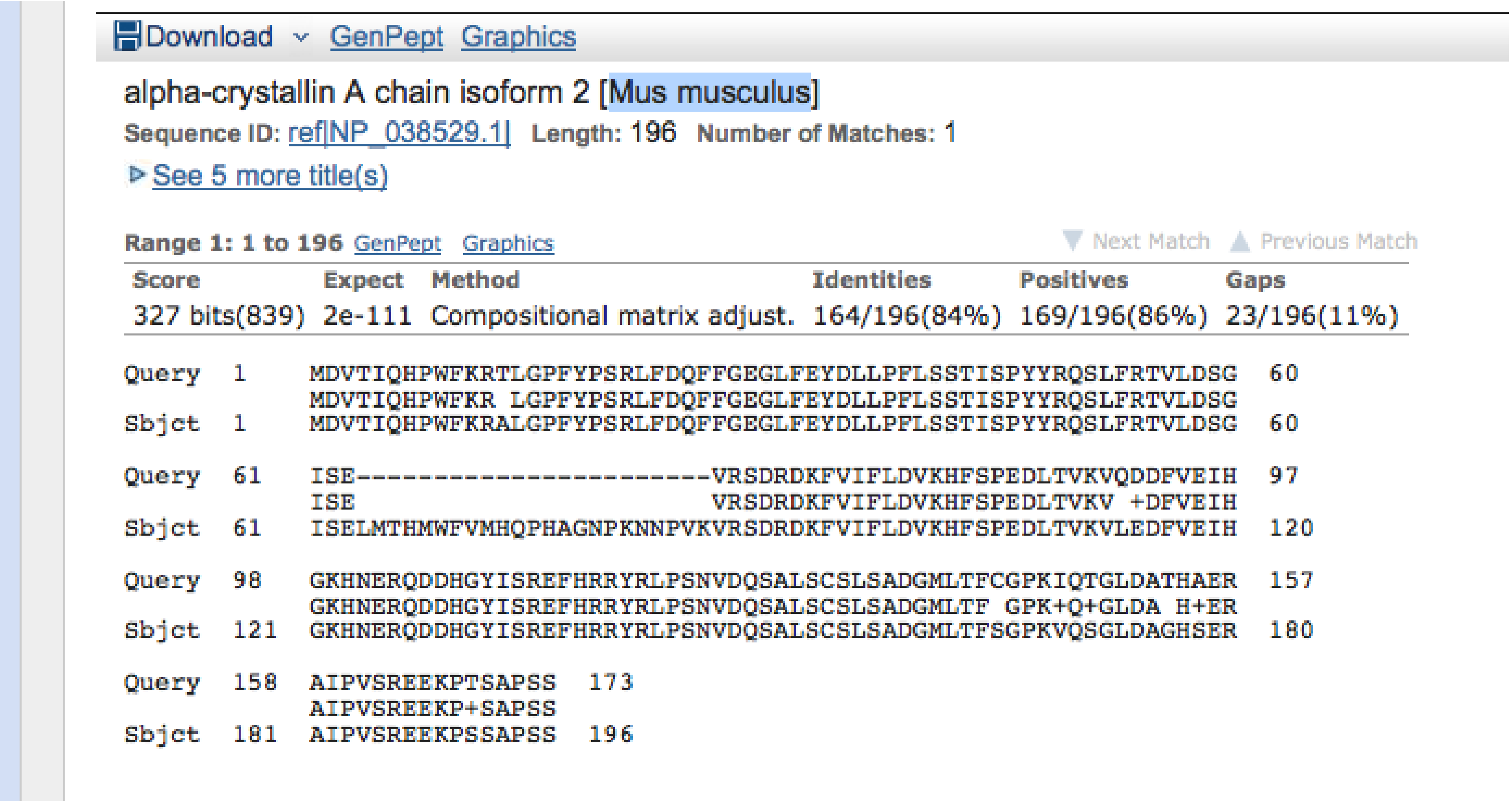
Sequence ID: ref|XP_005548643.1| Length: 188 Number of Matches: 1

Range 1: 17 to 188 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
345 bits(886)	1e-118	Compositional matrix adjust.	169/173(98%)	171/173(98%)	1/173(0%)

```
Query  1    MDVTIQHPWFKRTLGPFYPSRLFDQFFGEGLFEYDLLPFLSSTISPYYRQSLFRTVLDSG  60
            MDVTIQHPWFKRTLGPFYPSRLFDQFFGEGLFEYDLLPFLSSTISPYYRQSLFRTVLDSG
Sbjct  17   MDVTIQHPWFKRTLGPFYPSRLFDQFFGEGLFEYDLLPFLSSTISPYYRQSLFRTVLDSG  76

Query  61   ISEVRSDRDKFVIFLDVKHFSPEDLTVKVQDDFVEIHGKHNERQDDHGYISREFHRRYRL  120
            ISEVRSDRDKFVIFLDVKHFSPEDLTVKVQDDFVEIHGKHNERQDDHGYISREFHRRYRL
Sbjct  77   ISEVRSDRDKFVIFLDVKHFSPEDLTVKVQDDFVEIHGKHNERQDDHGYISREFHRRYRL  136

Query  121  PSNVDQSALSCSLSADGMLTFCGPKIQTGLDATHAERAIPVSREEKPTSAPSS  173
            PSNVDQSALSCSLSADGMLTF GPKIQTGLDATH ERAIPV+REEKP+SAPSS
Sbjct  137  PSNVDQSALSCSLSADGMLTFSGPKIQTGLDATH-ERAIPVAREEKPSSAPSS  188
```

Sequence Alignment
Example: human α-crystallin versus rhesus monkey



alpha-crystallin A chain isoform 2 [Mus musculus]

Sequence ID: ref|NP_038529.1| Length: 196 Number of Matches: 1

▶ See 5 more title(s)

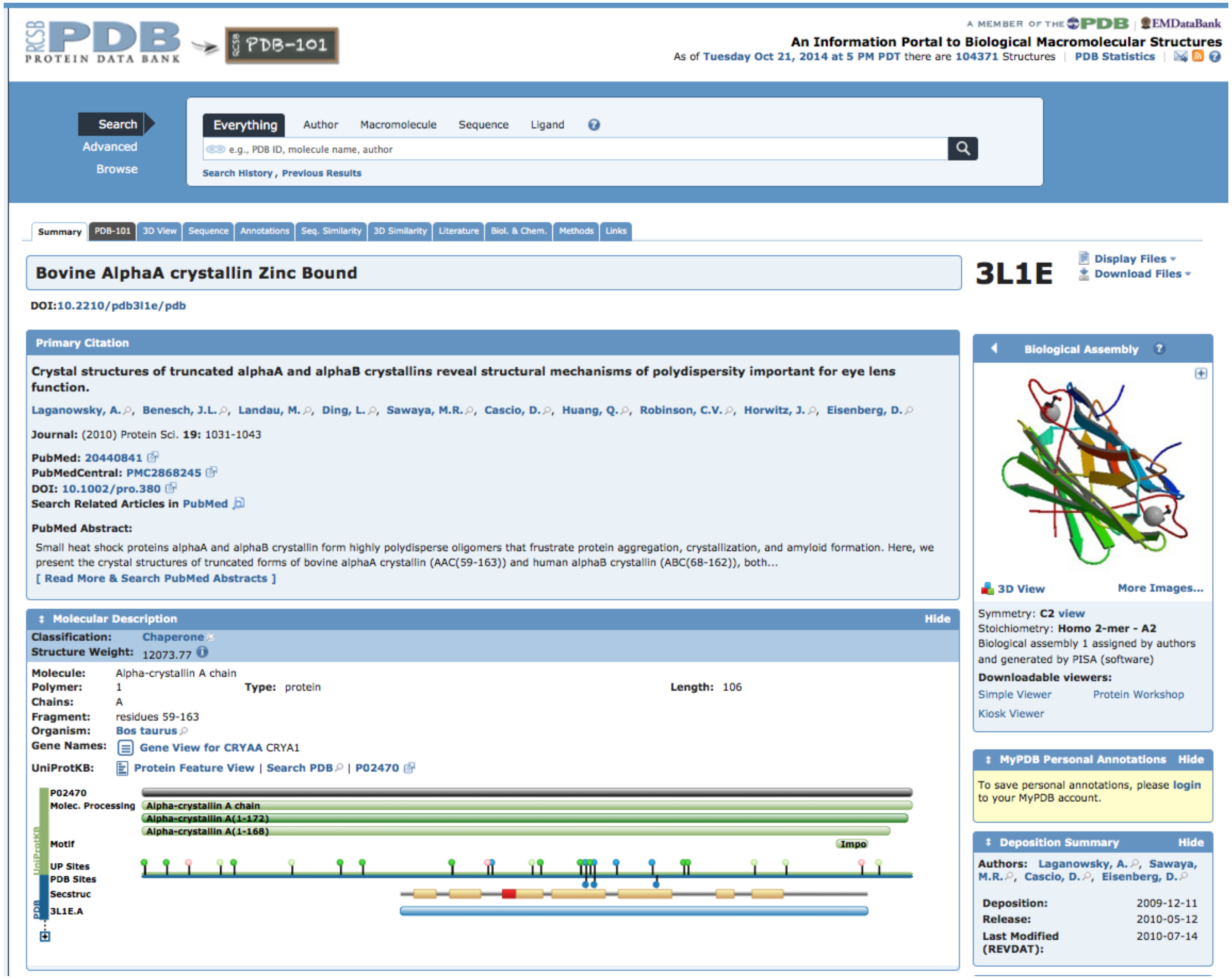
Range 1: 1 to 196 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
327 bits(839)	2e-111	Compositional matrix adjust.	164/196(84%)	169/196(86%)	23/196(11%)

```
Query  1    MDVTIQHPWFKRTLGPFYPSRLFDQFFGEGLFEYDLLPFLSSTISPYYRQSLFRTVLDSG  60
            MDVTIQHPWFKR LGPFYPSRLFDQFFGEGLFEYDLLPFLSSTISPYYRQSLFRTVLDSG
Sbjct  1    MDVTIQHPWFKRALGPFYPSRLFDQFFGEGLFEYDLLPFLSSTISPYYRQSLFRTVLDSG  60

Query  61   ISE-----------------------VRSDRDKFVIFLDVKHFSPEDLTVKVQDDFVEIH  97
            ISE                       VRSDRDKFVIFLDVKHFSPEDLTVKV +DFVEIH
Sbjct  61   ISELMTHMWFVMHQPHAGNPKNNPVKVRSDRDKFVIFLDVKHFSPEDLTVKVLEDFVEIH  120

Query  98   GKHNERQDDHGYISREFHRRYRLPSNVDQSALSCSLSADGMLTFCGPKIQTGLDATHAER  157
            GKHNERQDDHGYISREFHRRYRLPSNVDQSALSCSLSADGMLTF GPK+Q+GLDA H+ER
Sbjct  121  GKHNERQDDHGYISREFHRRYRLPSNVDQSALSCSLSADGMLTFSGPKVQSGLDAGHSER  180

Query  158  AIPVSREEKPTSAPSS  173
            AIPVSREEKP+SAPSS
Sbjct  181  AIPVSREEKPSSAPSS  196
```

Retrieve 3D Structure for Template (α-crystallin in Bos taurus)

Once a suitable Template (known 3D structure with high homology) is found, retreive the 3D structure from the protein structure database (pdb):
www.rcsb.org/pdb



An Information Portal to Biological Macromolecular Structures

As of Tuesday Oct 21, 2014 at 5 PM PDT there are 104371 Structures | PDB Statistics

Bovine AlphaA crystallin Zinc Bound **3L1E**

DOI:10.2210/pdb3l1e/pdb

Primary Citation

Crystal structures of truncated alphaA and alphaB crystallins reveal structural mechanisms of polydispersity important for eye lens function.

Laganowsky, A., Benesch, J.L., Landau, M., Ding, L., Sawaya, M.R., Cascio, D., Huang, Q., Robinson, C.V., Horwitz, J., Eisenberg, D.

Journal: (2010) Protein Sci. **19:** 1031-1043

PubMed: 20440841
PubMedCentral: PMC2868245
DOI: 10.1002/pro.380
Search Related Articles in PubMed

PubMed Abstract:

Small heat shock proteins alphaA and alphaB crystallin form highly polydisperse oligomers that frustrate protein aggregation, crystallization, and amyloid formation. Here, we present the crystal structures of truncated forms of bovine alphaA crystallin (AAC(59-163)) and human alphaB crystallin (ABC(68-162)), both...
[Read More & Search PubMed Abstracts]

Molecular Description Hide

Classification: Chaperone
Structure Weight: 12073.77

Molecule: Alpha-crystallin A chain
Polymer: 1 **Type:** protein **Length:** 106
Chains: A
Fragment: residues 59-163
Organism: Bos taurus
Gene Names: Gene View for CRYAA CRYA1
UniProtKB: Protein Feature View | Search PDB | P02470

Biological Assembly

Symmetry: **C2** view
Stoichiometry: **Homo 2-mer - A2**
Biological assembly 1 assigned by authors and generated by PISA (software)
Downloadable viewers:
Simple Viewer Protein Workshop
Kiosk Viewer

MyPDB Personal Annotations Hide

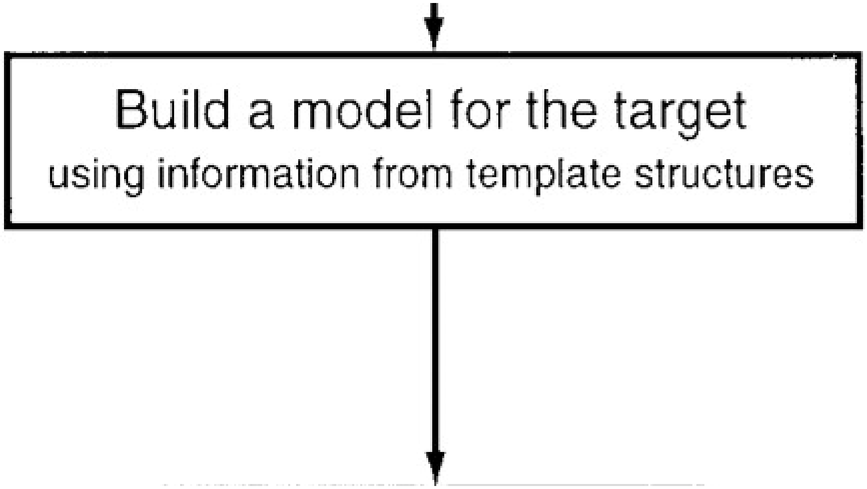
To save personal annotations, please login to your MyPDB account.

Deposition Summary Hide

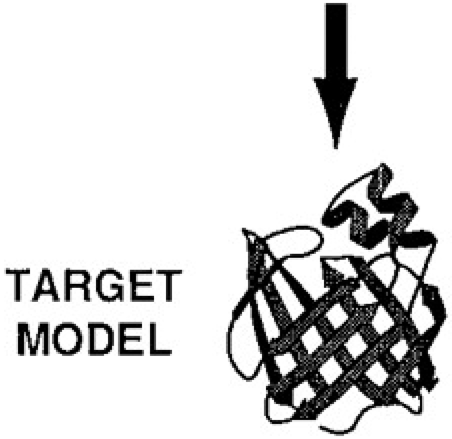
Authors: Laganowsky, A., Sawaya, M.R., Cascio, D., Eisenberg, D.

Deposition:	2009-12-11
Release:	2010-05-12
Last Modified (REVDAT):	2010-07-14

Step 3: Model Building



Build a model for the target
using information from template structures



Overlap template structures and generate backbone

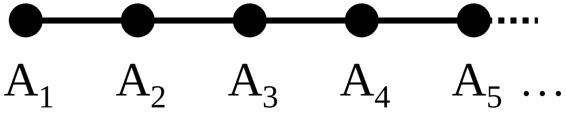
Generation of loops (data based or energy based)

Side chain generation based on known preferences

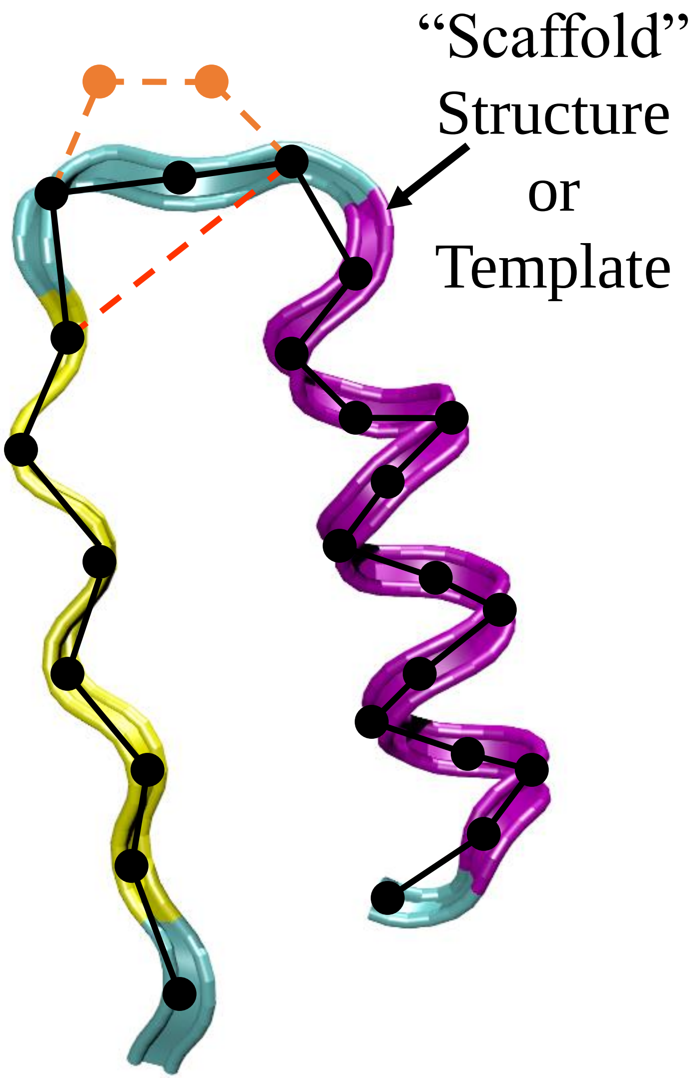
Overall model optimization (energy minimization)

Homology Modeling: Scoring

Target sequence



Alignment between target(s) and scaffold(s)



Quality of Prediction can be Ranked, based on "Energy"

Energy includes contributions from matches (favorable) , gaps (unfavorable), and hydrogen bonds.

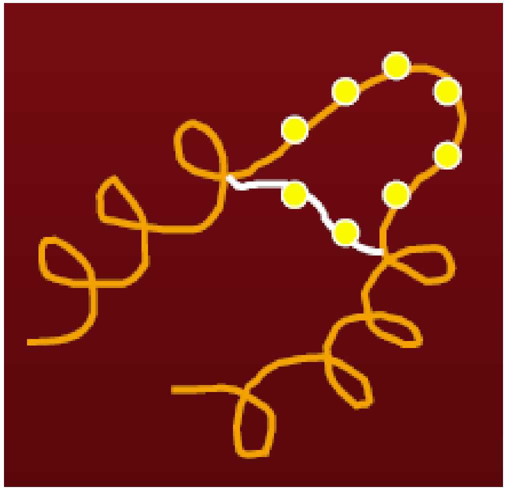
*R. Goldstein, Z. Luthey-Schulten, P. Wolynes (1992, PNAS), K. Koretke et.al. (1996, Proteins)

Why Modeling Loops is Difficult

Difference in the symmetry contacts in the crystals of the template and the real structure to be modeled.

Loops are flexible and can be distorted by neighboring residues

The mutation of a residue to proline within the loop



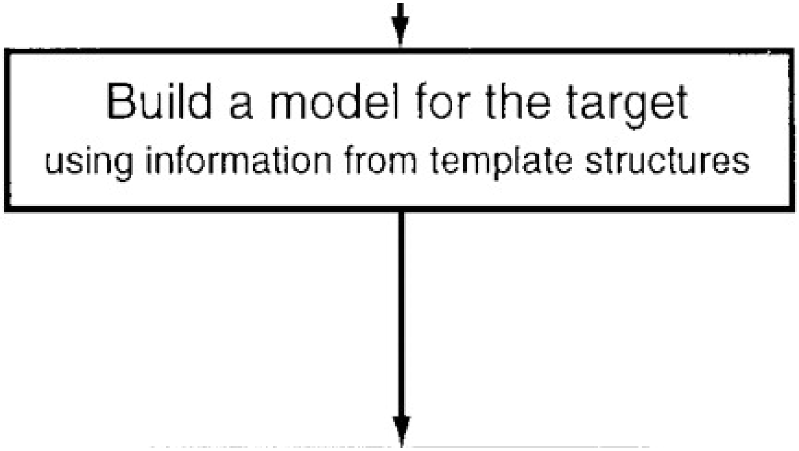
It is currently not possible to confidently model loops > 8 aa. There are two approaches

1) Data-base searches
2) Conformational searches using energy scoring functions (SwissModel)

Solvation can have a large effect on loops

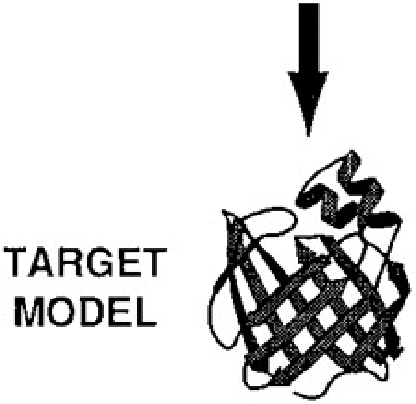
Modeling Servers

SwissModel	http://swissmodel.expasy.org/SWISS-MODEL.html
Modeller	http://salilab.org
Geno3D	http://geno3d-pbil.ibcp.fr
ESyPred	http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/
3D-jigsaw	http://www.bmm.icnet.uk/servers/3djigsaw/
CPHmodels	http://www.cbs.dtu.dk/services/CPHmodels/

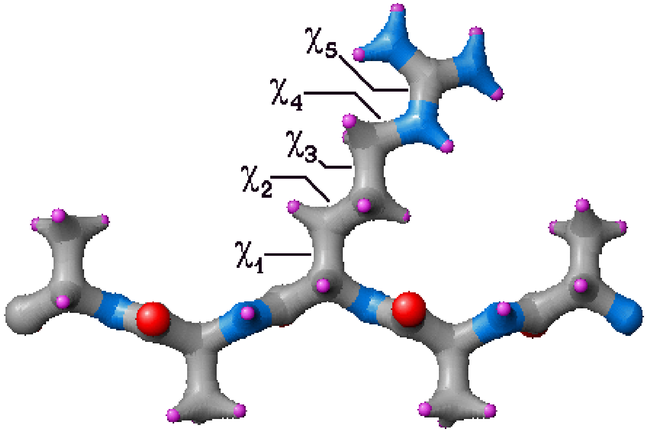


Build a model for the target

using information from template structures



Side Chain Modeling: Rotamer Libraries



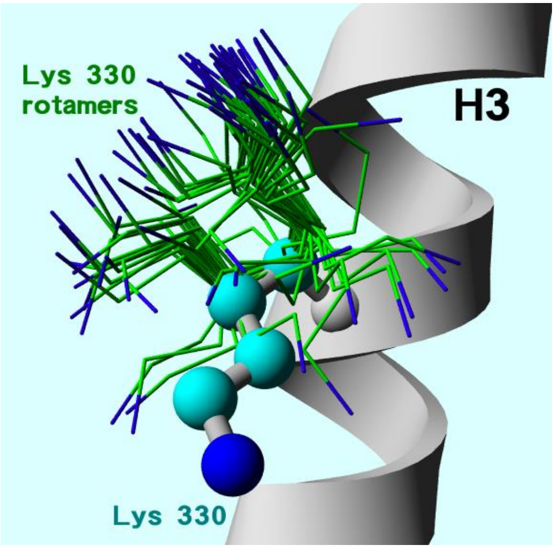
When we study the rotamers of residues that are conserved in different proteins with known 3D structure we observe in more than 90% of all cases similar side chain orientations.

The problem of placing side chains is thus reduced to concentrating on those residues that are not conserved in the sequence.

Two sub-problems:

1) finding potentially good rotamers,

2) determining the best one among the candidates.



SC Lovell et. al. “The Penultimate Rotamer Library”
Proteins: Structure Function and Genetics 40, 389-408 (2000).

Evaluating the Model: Looking for Unlikely Structures

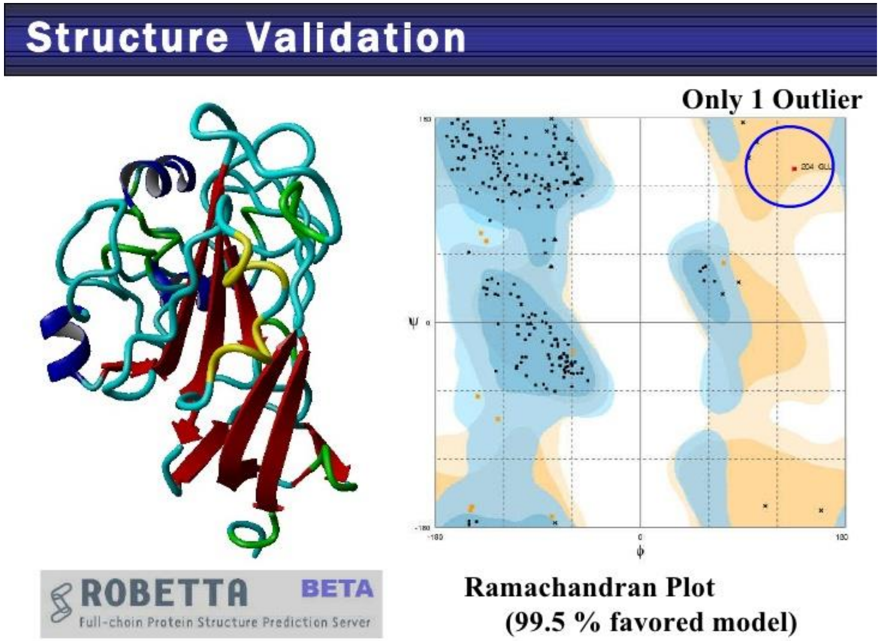
Errors in side chain packing

Template distortions because of crystal packing forces

Loop generation

Misalignments

Incorrect templates

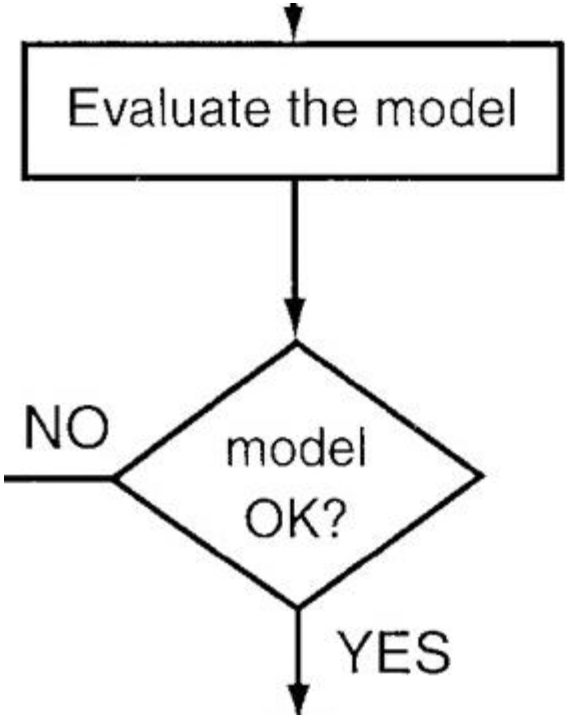


Structure Validation

Only 1 Outlier

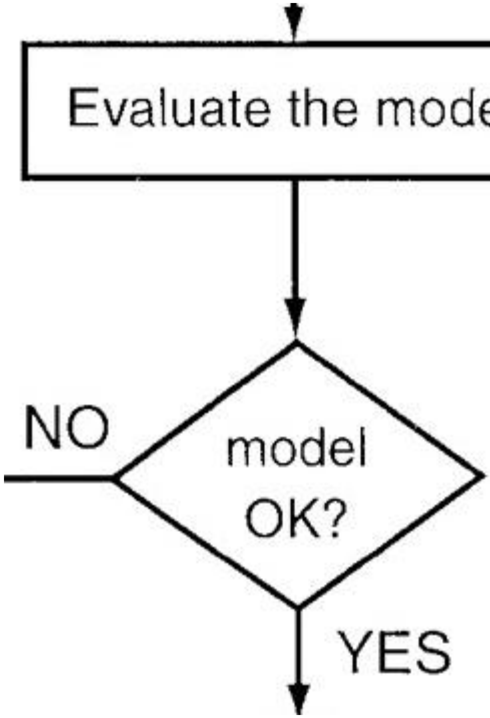
Ramachandran Plot

(99.5 % favored model)



Evaluation Servers

COLORADO3D	http://genesilico.pl/
PROCHECK	http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html
VERIFY3D	http://fold.doe-mbi.ucla.edu/
PROSAII	http://www.came.sbg.ac.at/
WHATCHECK	http://swift.cmbi.kun.nl/WIWWWI/modcheck.html



Homology Modeling Conclusions

Percent sequence identity	Total number of models	Percent models with rmsd lower than 1 Å	Percent models with rmsd lower than 2 Å	Percent models with rmsd lower than 3 Å	Percent models with rmsd lower than 4 Å	Percent models with rmsd lower than 5 Å	Percent models with rmsd higher than 5 Å
25-29	125	0	10	30	46	67	33
30-39	222	0	18	45	66	77	23
40-49	156	9	44	63	78	91	9
50-59	155	18	55	79	86	91	9
60-69	145	38	72	85	91	92	8
70-79	137	42	71	82	85	88	12
80-89	173	45	79	86	94	95	5
90-95	88	59	78	83	86	91	9

Homology modeling is better at predicting protein folds, worse at predicting side chain positions

When it comes to template selection: garbage in → garbage out

Protein–ligand interactions depend heavily on side chain positions, therefore use caution when proposing to understand such details based on homology models

Key Points

- Homology modeling is heavily dependent on the quality and percent identity of the template structure
- Insertions and deletions in the sequence degrade accuracy of the model
- Small errors in the backbone conformation can have large impacts on the accuracy of side chain placement and shape of the binding site – thus homology models are rarely suitable for subsequent use in ligand docking
- Homology models can be useful for generating hypotheses, for independent testing, such as by point mutagenesis