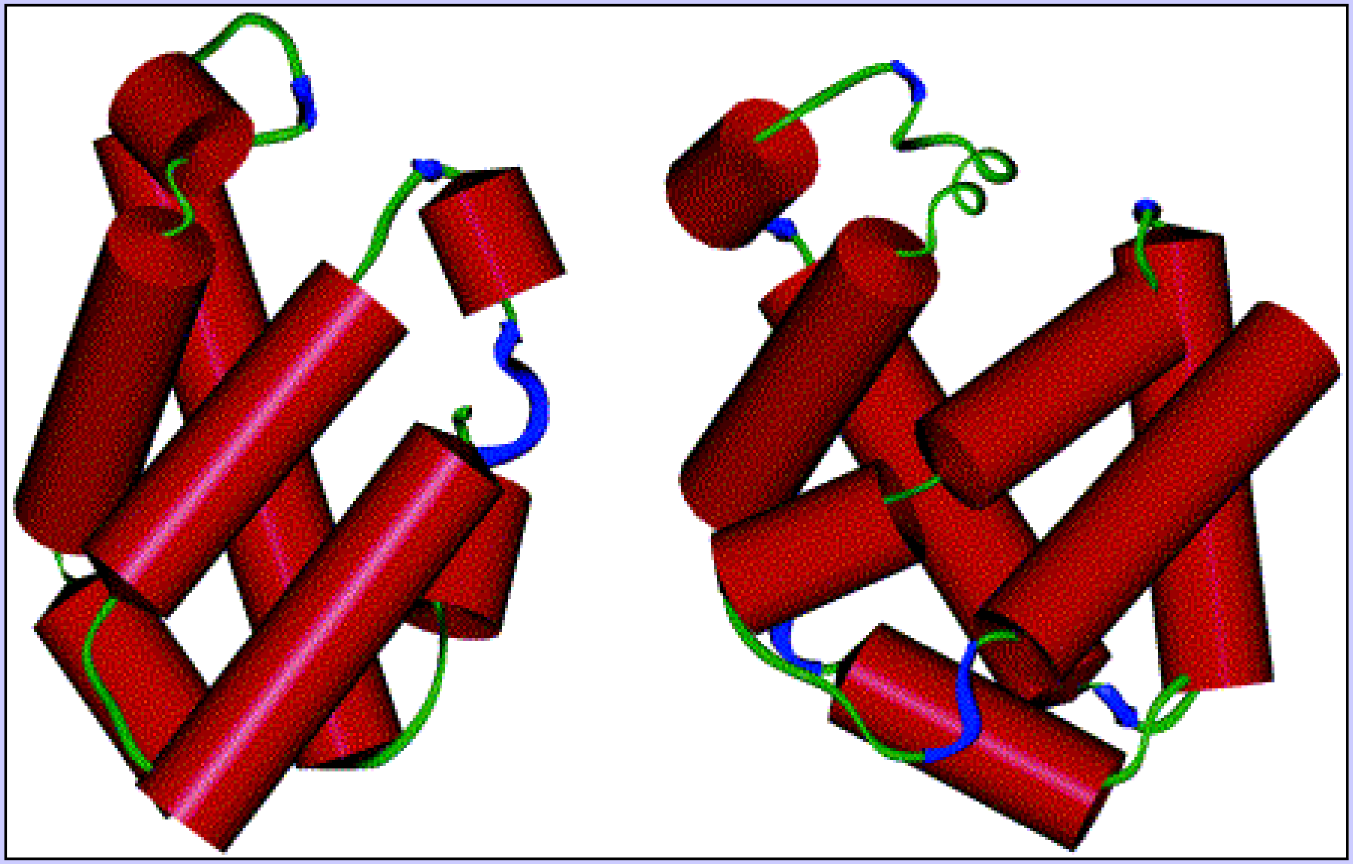
Building 3D models of proteins



Why make a structural model for your protein ?

The structure can provide clues to the function through
structural similarity with other proteins

With a structure it is easier to guess the location of active sites

With a structure we can plan more precise experiments in the lab

We can apply docking algorithms to the structures
(both with other proteins and with small molecules)

Basic princples for structural modeling

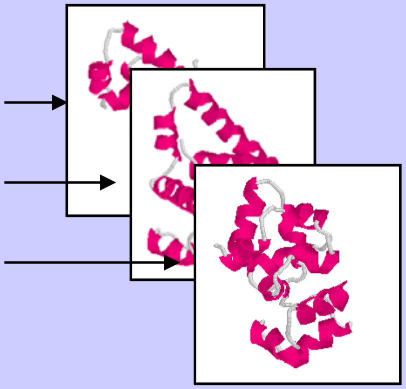
Use any piece of information the you have from the existing data bases regarding the protein you wish to model and its family

Choose the most suitable algorithm according to the amount if information that you have

Building by homology (Homology modelling)

Alignment with proteins of known structure

M	**A**	**A**	**G**	**Y**	**A**	**Y**	**G**	**V**	**L**	**S**
-	A	T	G	F	D	-	-	V	I	D
-	A	S	G	F	E	-	-	V	V	E
-	A	K	A	Y	L	-	-	V	L	S



structural model

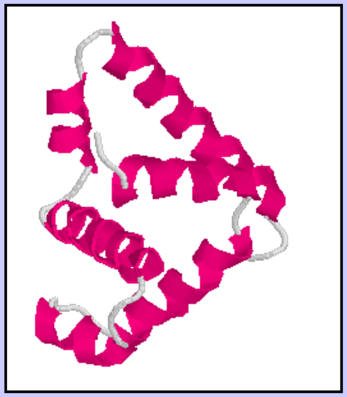
Fold recognition (Threading)

The sequence:

M	A	A	G	Y	A	V	L	S

+

Known protein folds



→

structural model

Ab initio

The sequence

M	A	A	G	Y	A	V	L	S

→

structural model

Building by homology

There are hundreds of thousands of protein sequences but only several thousand protein folds

For every second protein that we randomly pick from the structural data base there is “close” homolog (identity > 30%). This homolog almost always has the same fold.

In the current projects for experimental determination of protein structures, priority is given to determine structures of protein without homologs in the structural databases (proteomics)

We believe that in several years we will have almost all the basic folds

Steps of buiding by homology

Look for homology with the given sequence against the structural databases. Known algorithms such as Blast and FastA can be used.

If no hits were obtained, it is possible to use multiple alignment of the family for the search. This might be more sensitive.

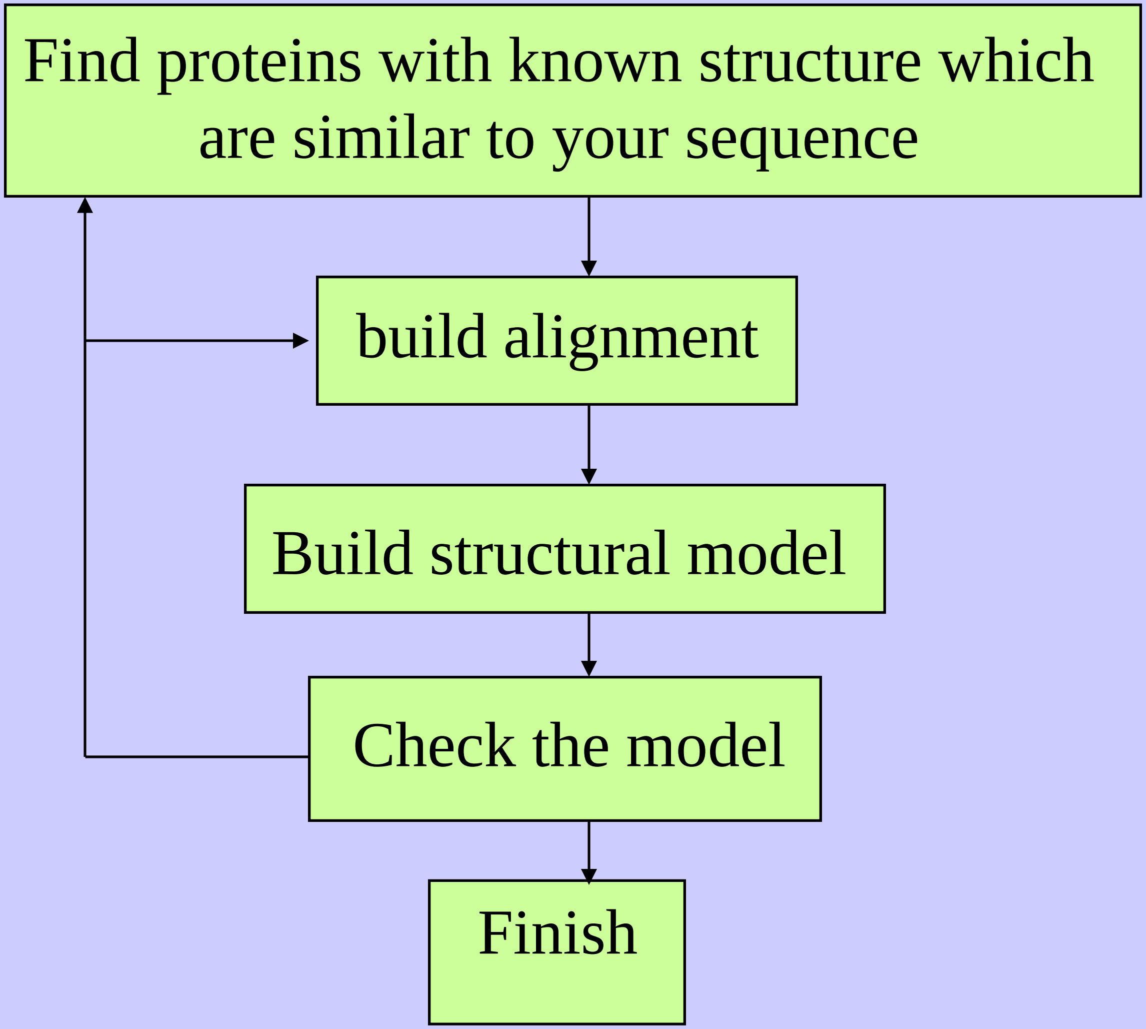
Construct accurate alignment between the query seqence and all the hits that will be serve as the template during the building

Correct alignment is crucial for this step. Any mistake can lead to significant errors in the final model

A possible algorithm is for example ClustalW. Manual intervention is often required, especially for weak homology.

More sequences within the protein family increase the chance for correct alignment.

Therefore, it is sometimes recommended to also search sequence databases (such as swissprot) and not only structural databases.



Find proteins with known structure which are similar to your sequence

build alignment

Build structural model

Check the model

Finish

Building the model itself

Determine the secondary structures according to the alignment

Determine structural reference according to the coordinates of the known structures. For every conserved region we can take the coordinates of the most similar homolog in that particular region.

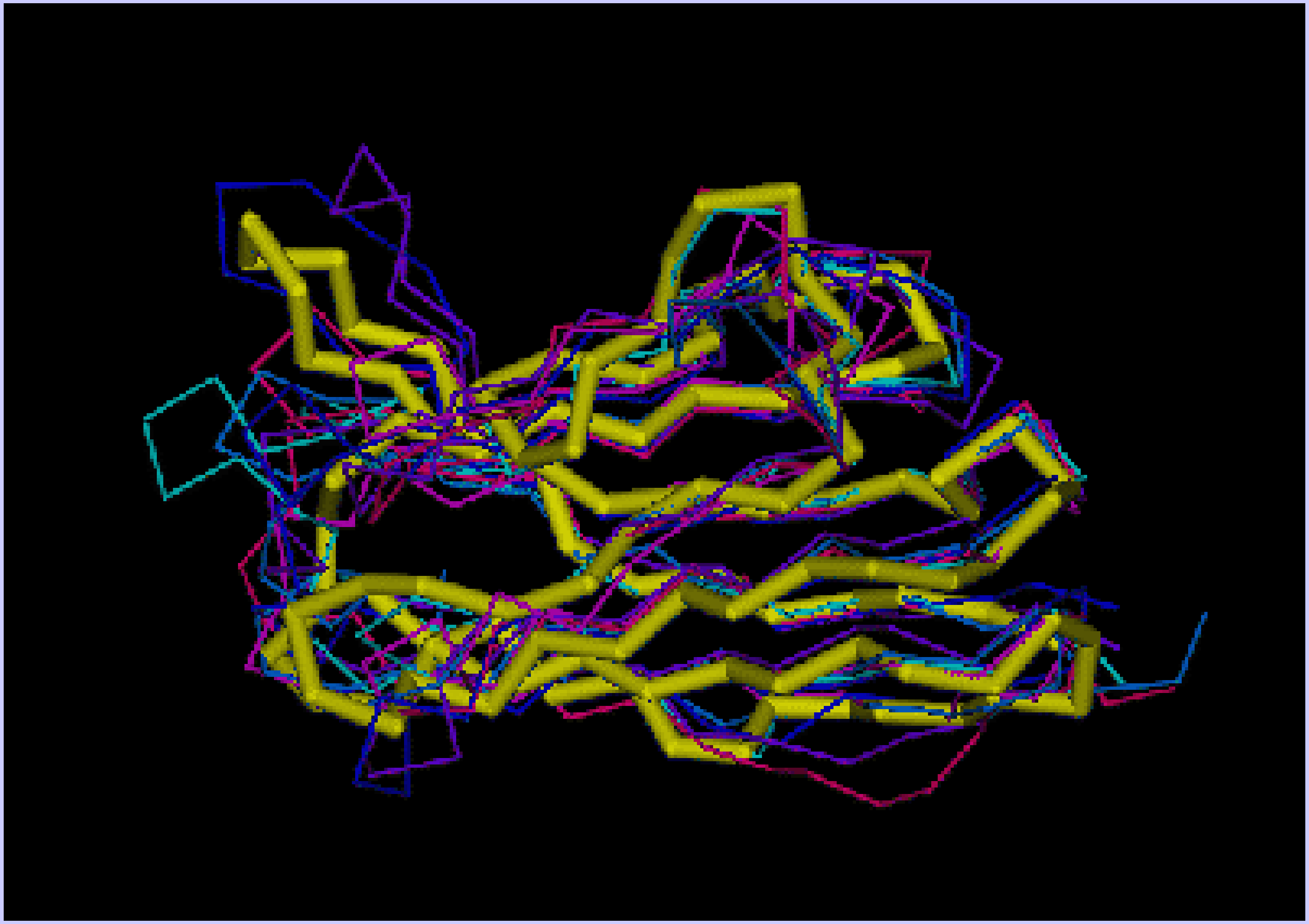
The main methods to determine structural reference during homology modeling are:

Fragment based homology

Build the conserved regions (usually secondary structures) and then build the loops

Distance constraints

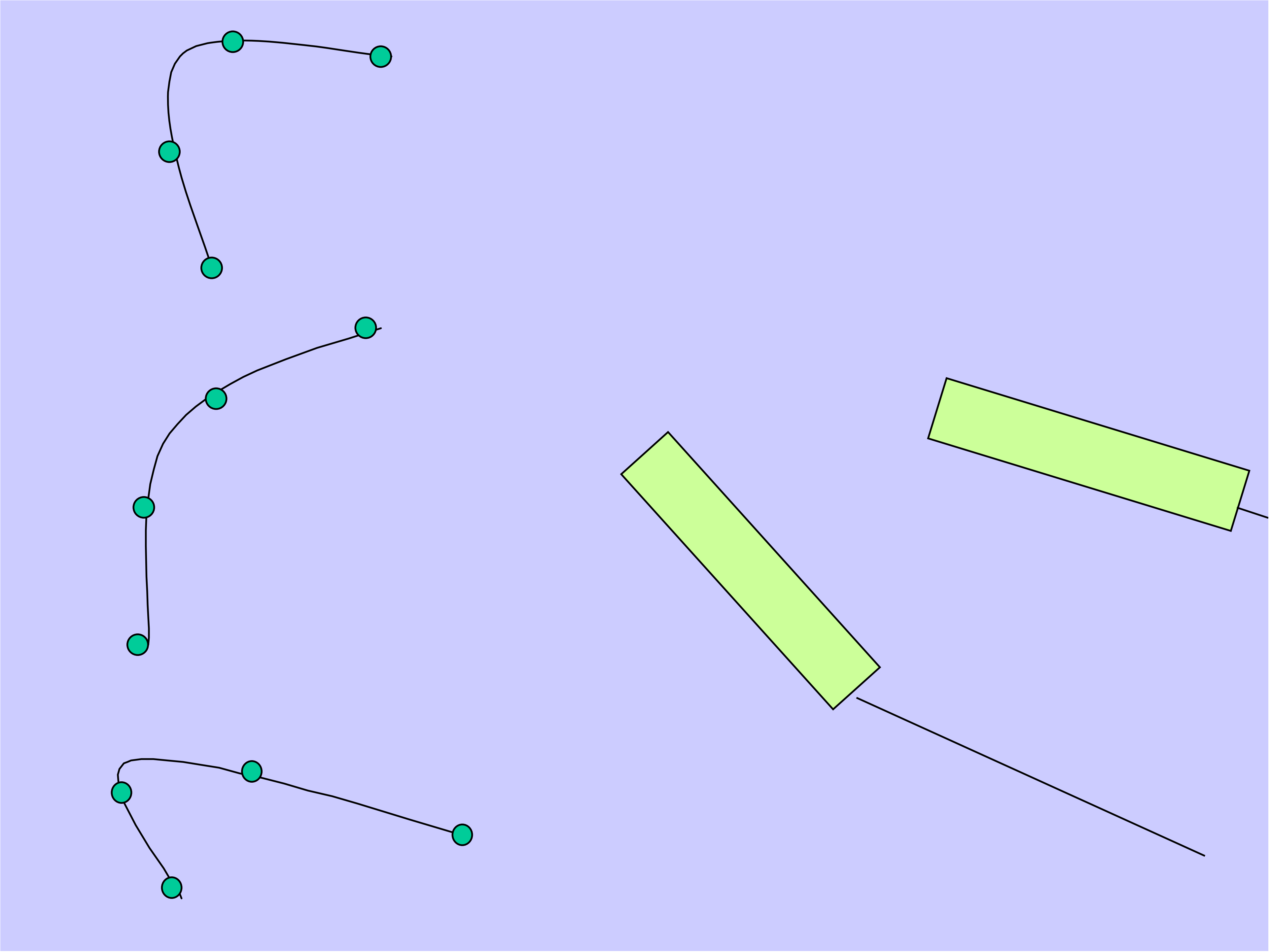
Building in one step the entire model based on distance constraints derived from the known structures and from the chemical nature of the molecule.

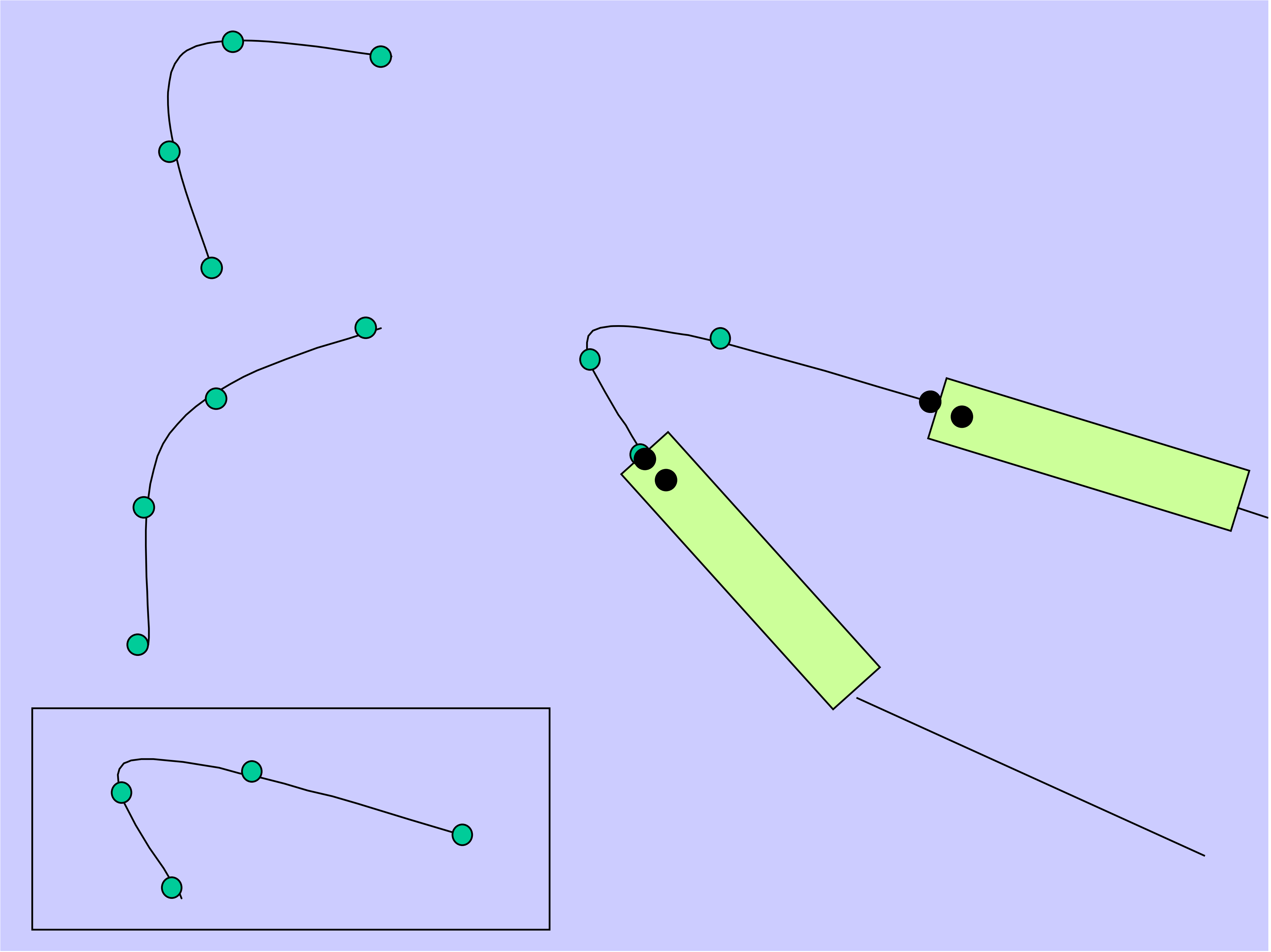


Construction of loops might be done by:

Using database of loops which appear in known structures. The loops could be catagorised by their length or sequence

***Ab initio* methods -** without any prior knowledge. This is done by empirical scoring functions that check large number of conformations and evaluates each of them.





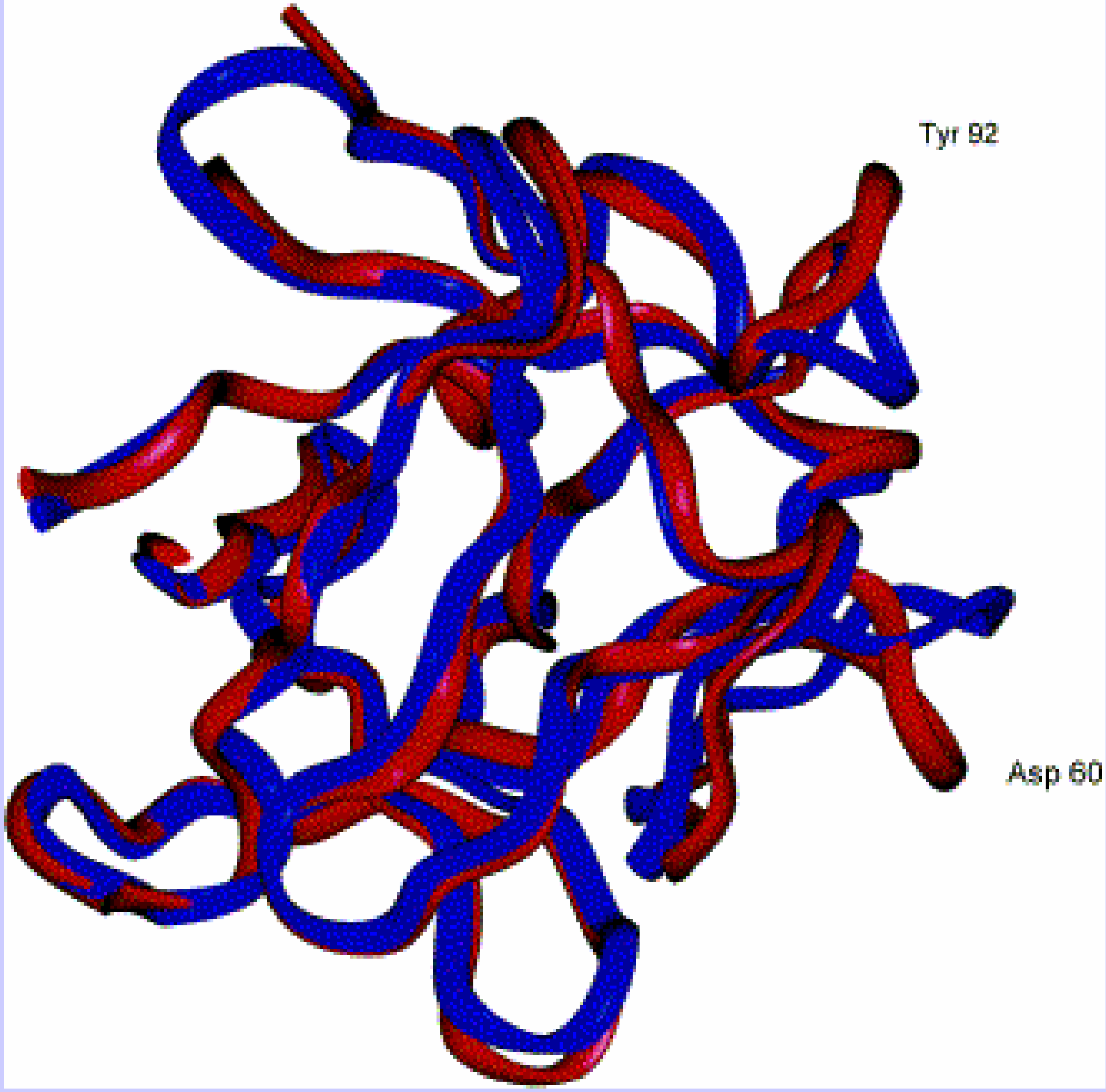
Several web pages for homology modeling

COMPOSER – felix.bioccam.ac.uksoft-base.html

MODELLER – guitar.rockefeller.edu/modeller/modeller.html

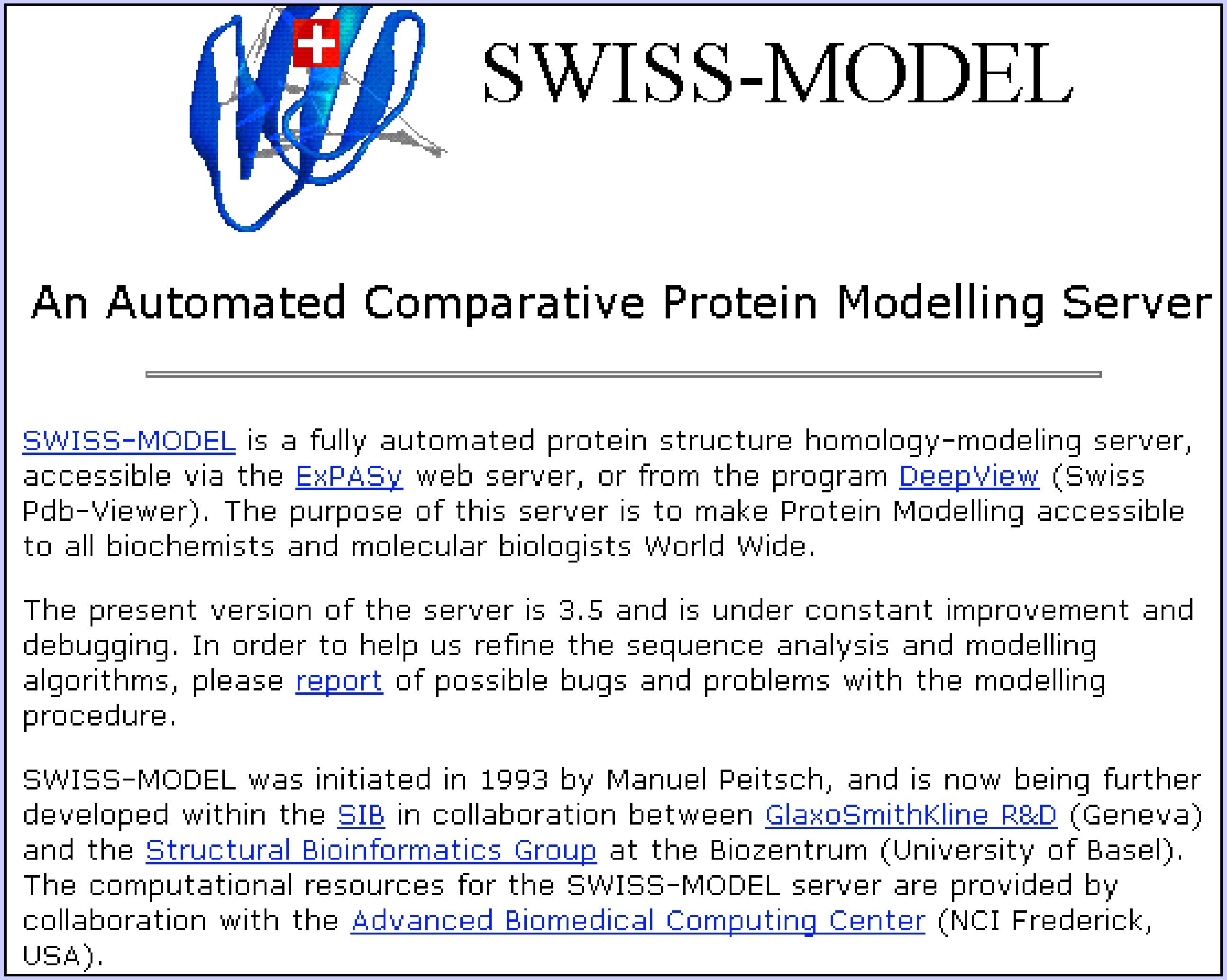
WHAT IF – www.sander.embl-heidelberg.de/whatif/

SWISS-MODEL – www.expasy.ch/SWISS-MODEL.html


Tyr 92
Asp 60

Swiss-Model

http://www.expasy.ch/swissmod/SWISS-MODEL.html



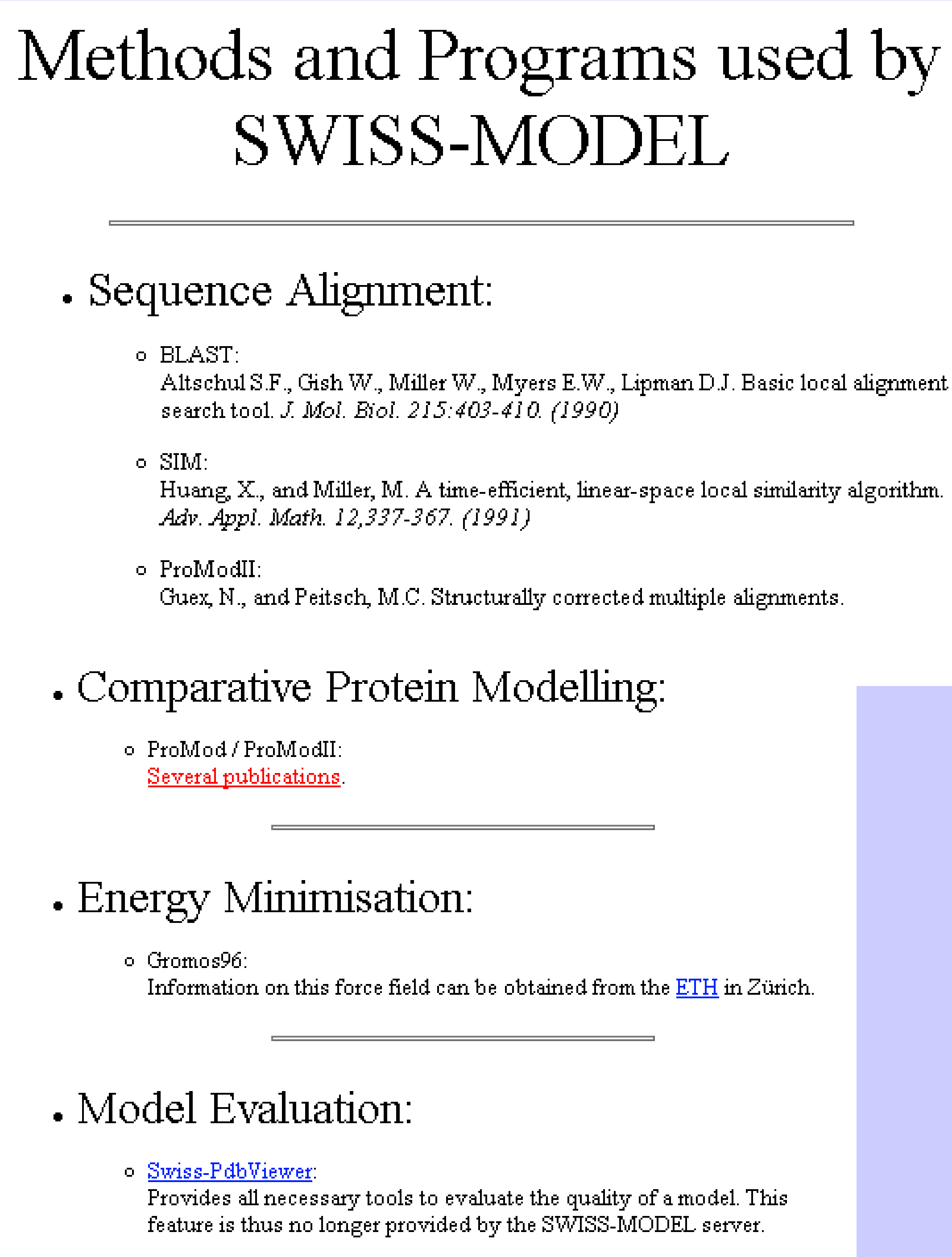
SWISS-MODEL

An Automated Comparative Protein Modelling Server

SWISS-MODEL is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make Protein Modelling accessible to all biochemists and molecular biologists World Wide.

The present version of the server is 3.5 and is under constant improvement and debugging. In order to help us refine the sequence analysis and modelling algorithms, please report of possible bugs and problems with the modelling procedure.

SWISS-MODEL was initiated in 1993 by Manuel Peitsch, and is now being further developed within the SIB in collaboration between GlaxoSmithKline R&D (Geneva) and the Structural Bioinformatics Group at the Biozentrum (University of Basel). The computational resources for the SWISS-MODEL server are provided by collaboration with the Advanced Biomedical Computing Center (NCI Frederick, USA).



Methods and Programs used by SWISS-MODEL

- ## Sequence Alignment:
  - BLAST:
    Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J. Mol. Biol. 215:403-410. (1990)*
  - SIM:
    Huang, X., and Miller, M. A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math. 12,337-367. (1991)*
  - ProModII:
    Guex, N., and Peitsch, M.C. Structurally corrected multiple alignments.

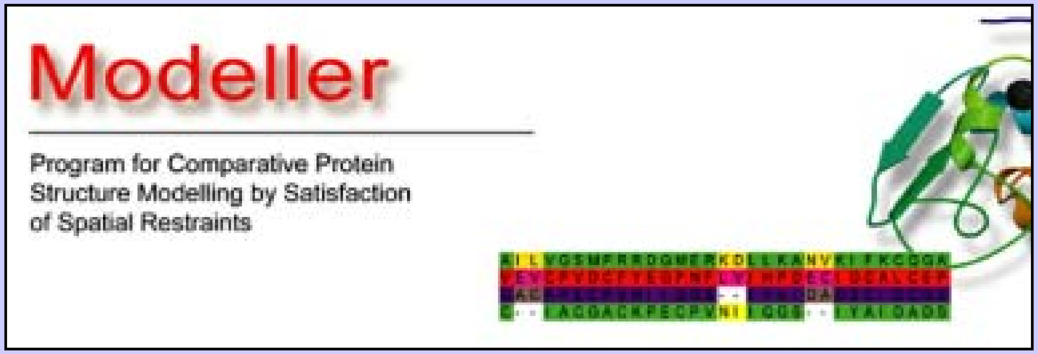
- ## Comparative Protein Modelling:
  - ProMod / ProModII:
    Several publications.

- ## Energy Minimisation:
  - Gromos96:
    Information on this force field can be obtained from the ETH in Zürich.

- ## Model Evaluation:
  - Swiss-PdbViewer:
    Provides all necessary tools to evaluate the quality of a model. This feature is thus no longer provided by the SWISS-MODEL server.

Modeller

http://guitar.rockefeller.edu/modeller/about_modeller.shtml



Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints

Advanced program for homology modeling

Based on distance constraints

Implemented in several popular modelling packages
such as InsightII

The source is available for unix platforms at the above URL

Threading (fold recognition)

The input sequence is threaded on different folds from library of known folds

Using scoring functions we get a score for the compatability between the sequence and the structure

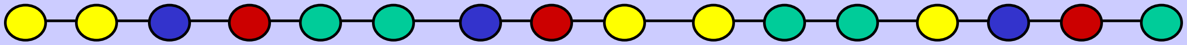
Statisticaly significant score tells that the input protein adopts similar 3D structure to that of the examined fold

This method is less accurate but could be applied for more cases

When the “real” fold of the input sequence is not represented in the structural database we can not get correct solution by this method

The most important part is the accuracy of the scoring function. The scoring function is the major difference between different programs for fold recognition

Input:



sequence

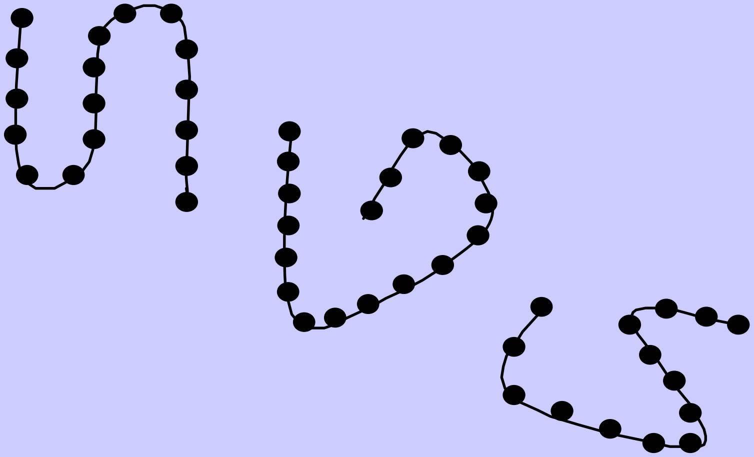
🔵 H bond donor

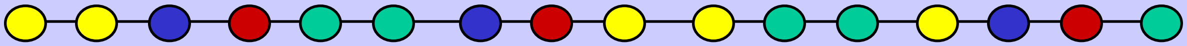
🔴 H bond acceptor

🟢 Glycin

🟡 Hydrophobic

Library of folds of known proteins



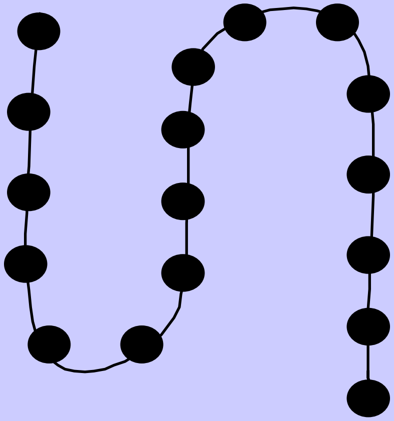


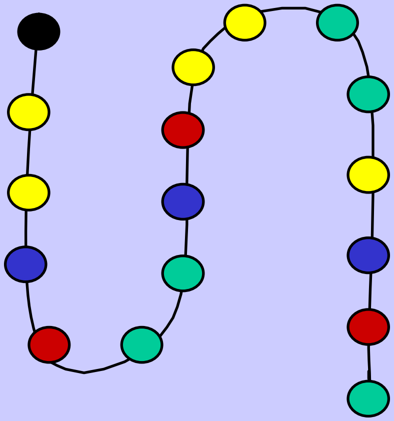
🔵 H bond donor

🔴 H bond acceptor

🟢 Glycin

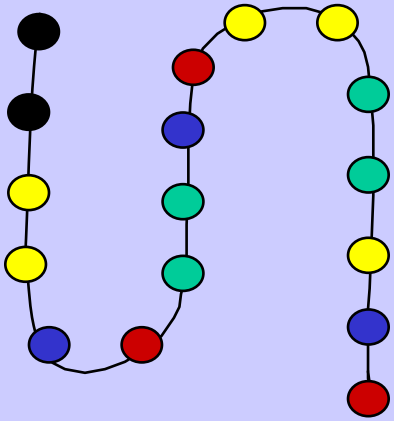
🟡 Hydrophobic





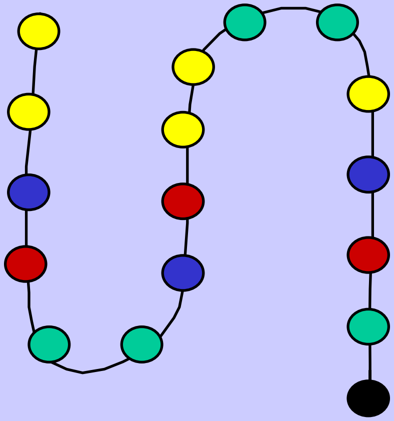
S=-2

Z= -1



S=5

Z=1.5



S=20
Z=5

Scoring functions for fold recognition

There are 2 basic methods to evaluate sequence-structure (1D-3D) compatibility

In methods based on structural profile, for every fold a profile is built based on structural features of the fold and compatibility of every amino acid to the features.

The structural features of each position are determined based on the combination of secondary structure, solvent accessibility and the property of the local environment (hydrophobic/hydrophlic)

The profile is a defined mathematical structure, adjusted for pair-wise comparisons and dynamic programming

Amino acid type

Position on sequence

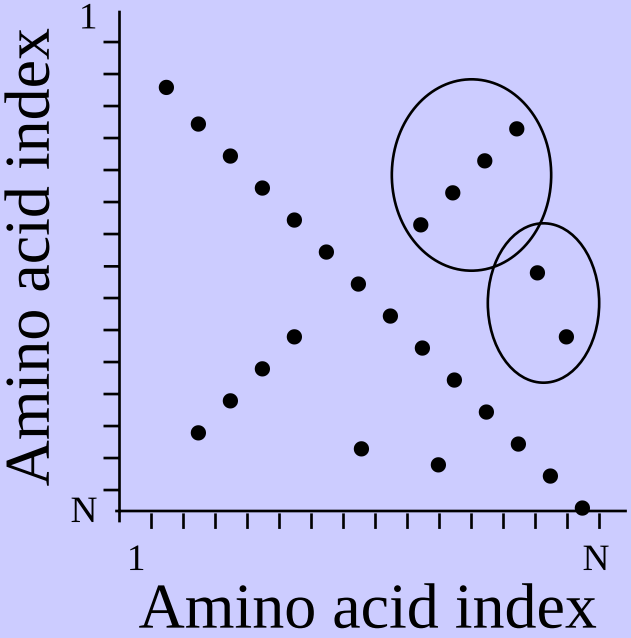
	A	C	D	…	Y	G_{op}	G_{ext}
1	10	-50	101		-80	100	10
2	-24	87	-99		167	100	10
:	:	:	:	:	:	:	:
N						100	10

Contact potentials

This method is based on predefined tables which include pseudo-energetic scores to each pair-wise interaction of two amino acids.

This method makes use of distance matrix for representation of different folds

For each pair of amino acids which are close in space the interaction energy is summed. The total sum is the indication for the fitness of the sequence into that structure


1
N
1
N
Amino acid index
Amino acid index

Web sites for fold recognition

Profiles:

3D-PSSM - http://www.bmm.icnet.uk/~3dpssm

Libra I - http://www.ddbj.nig.ac.jp/htmls/E-mail/libra/LIBRA_I.html

UCLA DOE - http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.html

Contact potentials

123D - http://www-Immb.ncifcrf.gov/~nicka/123D.html

Profit - http://lore.came.sbg.ac.at/home.html

Ab initio **methods for modelling**

This field is of great theoretical interest but, so far, of very little practical applications. Here there is no use of sequence alignments and no direct use of known structures

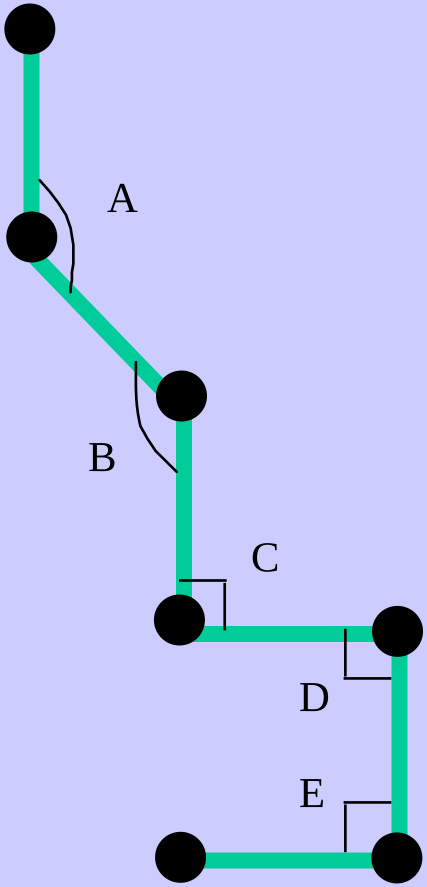
The basic idea is to build empirical function that simulates real physical forces and potentials of chemical contacts

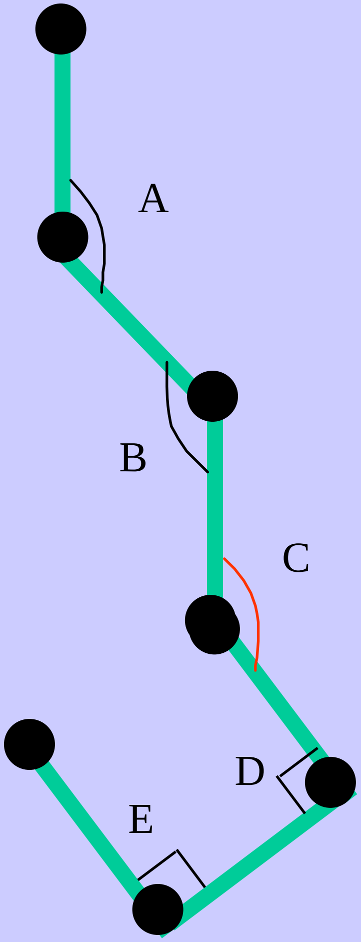
If we will have perfect function and we will be able to scan all the possible conformations, then we will be able to detect the correct fold

Algorithms for *Ab initio* prediction include:
A. Searching procedure that scans many possible structures (conformations)
B. Scoring function to evaluate and rank the structures

Due to the large search space, heuristic methods are usually applied

The parameters in the searching procedure are the dihedral angles which specify the exact fold of the polypeptide chain


A
B
C
D
E


A
B
C
D
E

Methods to evaluate structures are based on

Force fields- collection of terms that simulate the forces act between atoms

Terms based on probabilities to find pairs of amino acids or atoms within specific distances

Terms based on surface area and overlapping volume of spheres representing atoms

Side chain construction

In homology modeling, construction of the side chains is done using the template structures when there is high similarity between the built protein and the templates

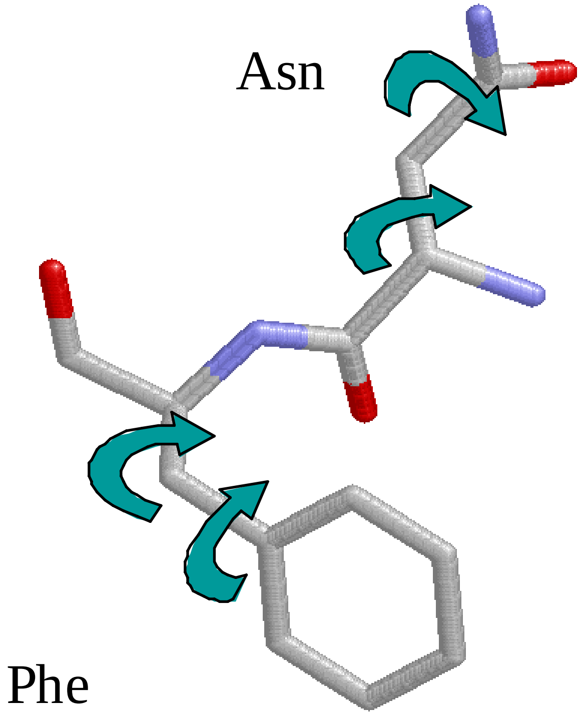
Without such similarity the construction can be done using rotamer libraries

A compromise between the probability of the rotamer and its fitness in specific position determines the score. Comparing the scores of all the rotamer for a given amino acid determines the preferred rotamer.

In spite of the huge size of the problem (because each side chain influences its neighbors) there are quite successful algorithms to this problem.

Conformation - a given set of dihedral angle which defines a structure.

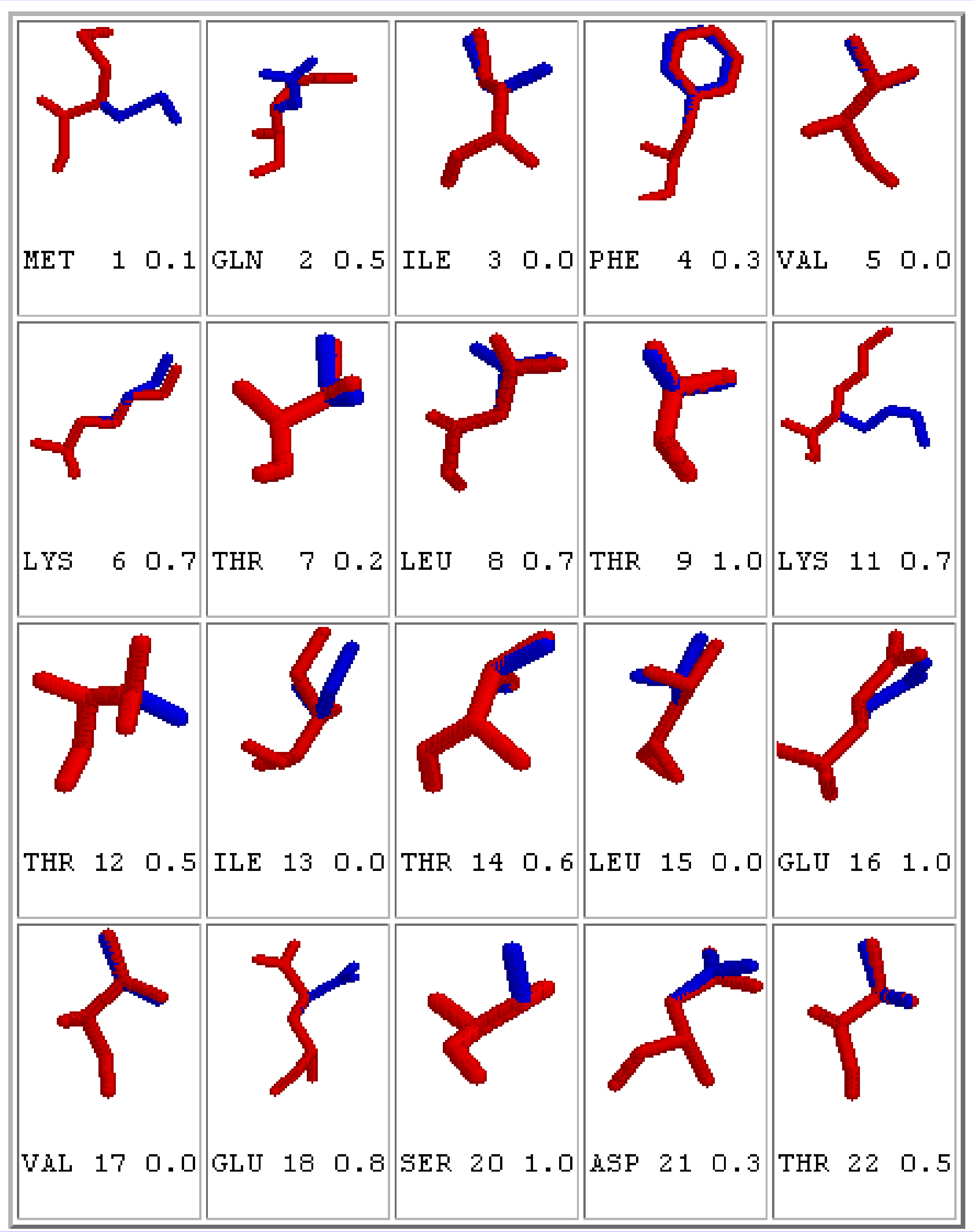
Rotamer - energetically favourable conformation.



Example to library of rotamers

SER	59.6	41.0
SER	-62.5	26.4
SER	179.6	32.6

TYR	63.6	90.5	21.0
TYR	68.5	-89.6	16.4
TYR	170.7	97.8	13.3
TYR	-175.0	-100.7	20.0
TYR	-60.1	96.6	10.0
TYR	-63.0	-101.6	19.3

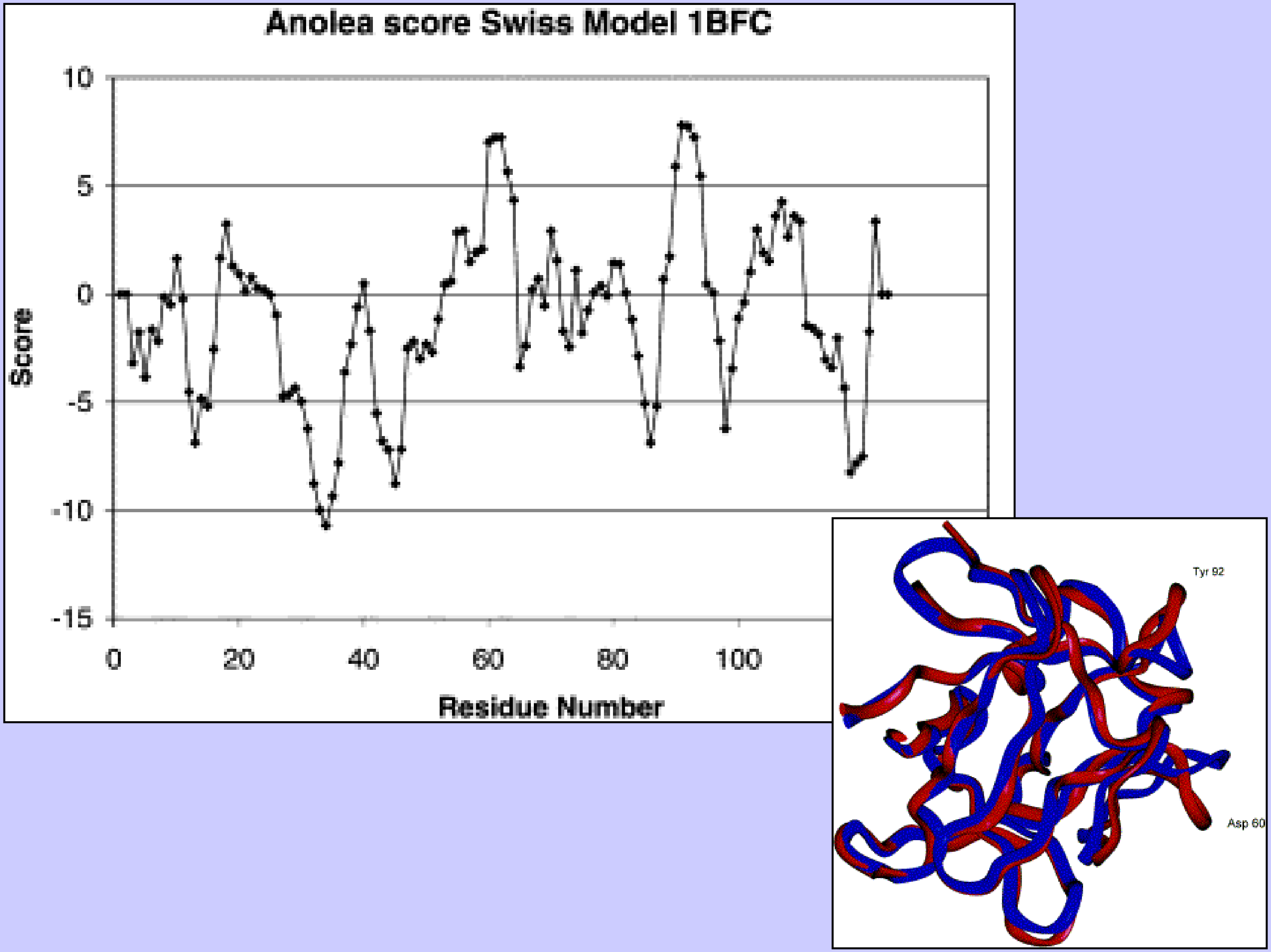

MET 1 0.1
GLN 2 0.5
ILE 3 0.0
PHE 4 0.3
VAL 5 0.0
LYS 6 0.7
THR 7 0.2
LEU 8 0.7
THR 9 1.0
LYS 11 0.7
THR 12 0.5
ILE 13 0.0
THR 14 0.6
LEU 15 0.0
GLU 16 1.0
VAL 17 0.0
GLU 18 0.8
SER 20 1.0
ASP 21 0.3
THR 22 0.5

Model evaluation

After the model is built we can check it by various methods.

If the model turns out to be bad, it is necessary to repeat several stages of the model building

The main approaches for model evaluation are:
A. Use of internal information (such as the one that used for the model construction)
B. Use of external information derived from the databases


Anolea score Swiss Model 1BFC
10
5
0
-5
-10
-15
Score
0
20
40
60
80
100
Residue Number
Tyr 92
Asp 60

Usually algorithms are checked by building models for proteins which have already solved structure and comparison between the model and the native structure

It is always possible that information from the native structure will be used in direct or indirect ways for model building

A more objective test is prediction of structures before they are publicly distributed (this is the idea of the CASP competitions)