Biomodeling *Biotech*

by Stefano Moro

Molecular Modeling Section (MMS)

Department of Pharmaceutical and Pharmacological Sciences

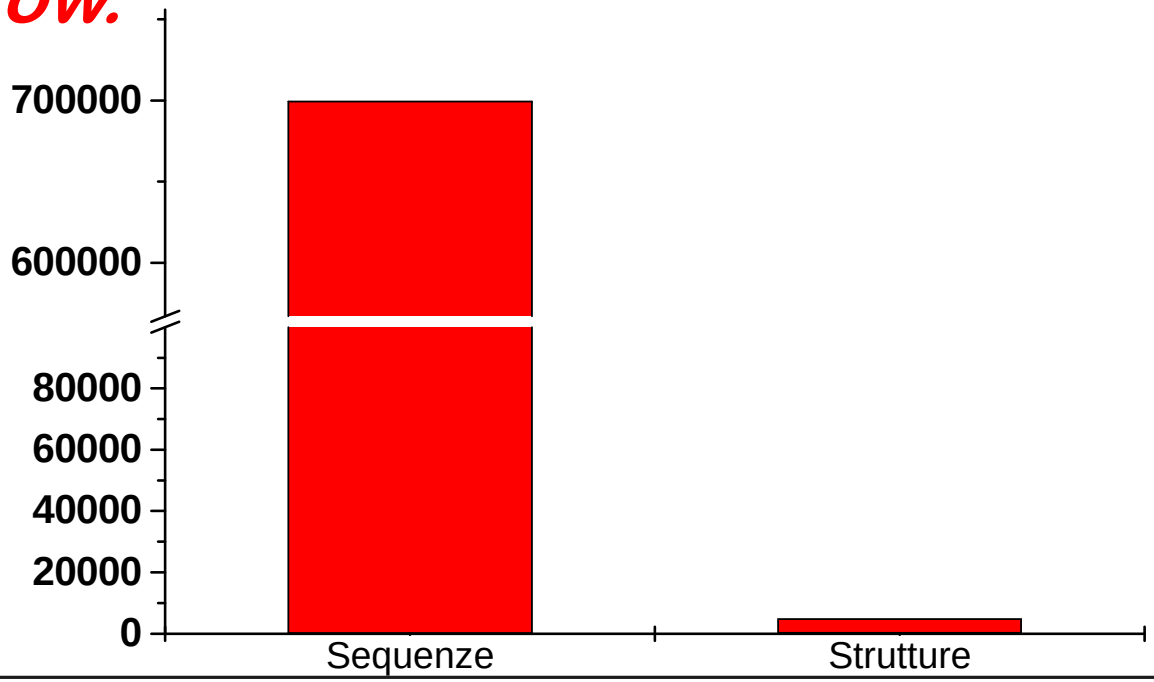
University of Padova

Homology modeling

… the best example of *copy&paste* in bioinformatics!

Why do we need homology modeling?

- ⇨ 700,000+ protein sequences
- ⇨ ~ 20,000 structures, ~ 5,000 unique
- ⇨ *The gap between sequences and structures continues to grow.*



Why do we need homology modeling?

Each polypeptide chain can potentially adopt an astronomical number of conformations:

the Levinthal paradox

J. Chim. Phys. PCB 65, 44-45 (1968).

Why do we need homology modeling?

- Many proteins fold in seconds or less: how is this possible?
- Cyrus Levinthal tried to estimate how long it would take a protein to do a random search of conformational space for the native fold.
- Imagine a 100-residue protein with three possible conformations per residue. Thus, the number of possible folds = $3^{100} = 5 \times 10^{47}$.
- Let us assume that protein can explore new conformations at the same rate that bonds can reorient (10^{13} structures/second).
- Thus, the time to explore all of conformational space = $5 \times 10^{47}/10^{13} = 5 \times 10^{34}$ seconds = 1.6×10^{27} years >> age of universe

Why do we need homology modeling?

- The structure of a protein is “uniquely” determined by its amino acid sequence
  (but sequence is sometimes not enough):
  - prions
  - pH, ions, cofactors, chaperones

- Structure is conserved much longer than sequence in evolution.
  - Structure > Function >> Sequence

Why do we need homology modeling?

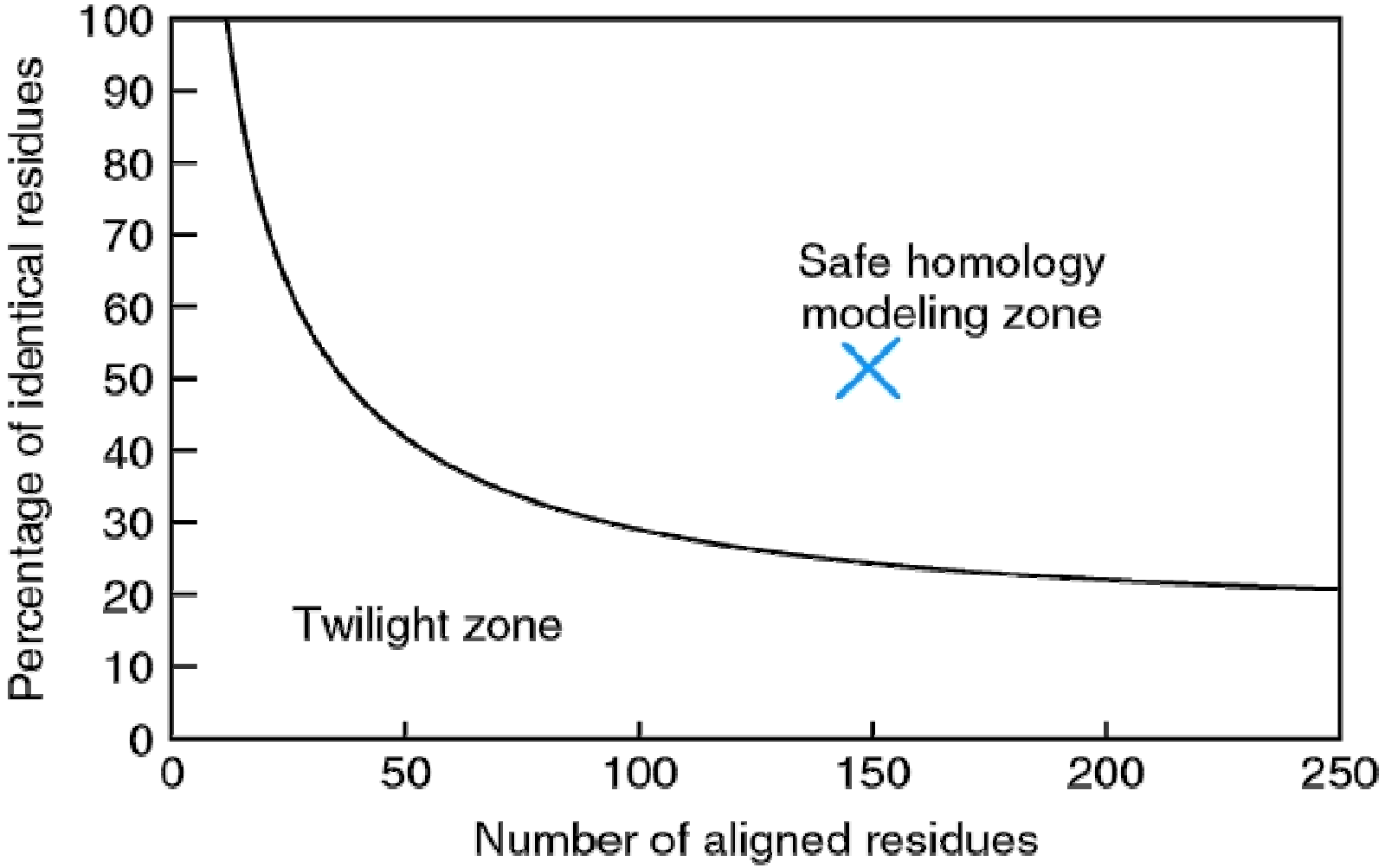
In the context of evolutionary biology, ***homology*** is the existence of shared ancestry between a pair of proteins or genes.

Remember: homology is a qualitative property. Beware of those who say: " % of homology" !!!

What is homology modeling?

- **Given the sequence of an unknown protein, make an informed guess on its 3D structure based on the structure of an homologous sequence:**
  - Search structure databases for *homologous* sequences
  - Transfer coordinates of known protein onto unknown

How well we can do it?



How is it done?

- Identify template(s) – initial alignment
- Improve alignment
- Backbone generation
- Loop modelling
- Side chains
- Refinement
- Validation ←

Template identification

- Search with sequence
  - Blast
  - Psi-Blast
  - Fold recognition methods
- Use biological information
- Functional annotation in databases
- Active site/motifs

Improving the alignment

1	2	3	4	5	6	7	8	9	10	11	12	13	14
PHE	ASP	ILE	CYS	ARG	LEU	PRO	GLY	SER	ALA	GLU	ALA	VAL	CYS
PHE	ASN	VAL	CYS	ARG	THR	PRO	---	---	---	GLU	ALA	ILE	CYS
PHE	ASN	VAL	CYS	ARG	---	---	---	THR	PRO	GLU	ALA	ILE	CYS

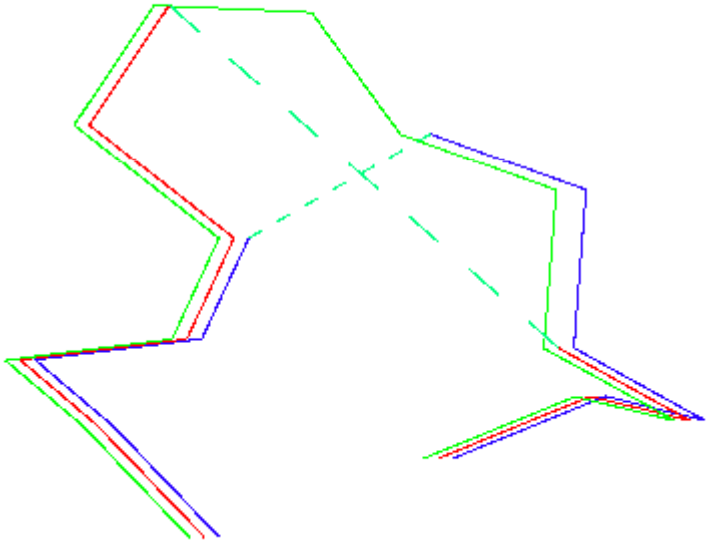
	F	D	I	C	R	L	P	G	S	A	E	A	V	C
F	6	-2	0	-3	-2	2	-2	-3	-1	-2	-3	-2	0	-3
N	-3	2	-2	-2	0	-2	-2	0	2	0	1	0	-2	-2
V	0	-2	2	-2	-1	2	-1	-1	-1	0	-1	0	5	-2
C														
R	-2	-2	-2	-2	5	-1	0	0	1	-1	0	-1	-1	-2
T														
P														
E	-3	2	-2	-3	0	-2	1	0	1	1	5	1	-1	-3
A	-2	0	-1	-2	-1	-1	1	0	1	5	1	5	0	-2
I	0	-3	5	-2	-2	2	-2	-2	-1	-1	-2	-1	2	-2
C	-3	-2	-2	8	-2	-3	-3	-2	-1	-2	-3	-2	-2	8

1	2	3	4	5	6	7	8	9	10	11	12	13	14
PHE	ASP	ILE	CYS	ARG	LEU	PRO	GLY	SER	ALA	GLU	ALA	VAL	CYS
PHE	ASN	VAL	CYS	ARG	THR	PRO	---	---	---	GLU	ALA	ILE	CYS
PHE	ASN	VAL	CYS	ARG	---	---	---	THR	PRO	GLU	ALA	ILE	CYS

	F	D	I	C	R	L	P	G	S	A	E	A	V	C
F	6	-2	0	-3	-2	2	-2	-3	-1	-2	-3	-2	0	-3
N	-3	2	-2	-2	0	-2	-2	0	2	0	1	0	-2	-2
V	0	-2	2	-2	-1	2	-1	-1	-1	0	-1	0	5	-2
C	-3	-2	-2	8	-2	-3	-3	-2	-1	-2	-3	-2	-2	8
R	-2	-2	-2	-2	5	-1	0	0	1	-1	0	-1	-1	-2
T	-2	0	0	-1	0	0	0	-1	2	0	1	0	0	-1
P	-2	0	-2	-3	0	-2	8	0	0	1	1	1	-1	-3
E	-3	2	-2	-3	0	-2	1	0	1	1	5	1	-1	-3
A	-2	0	-1	-2	-1	-1	1	0	1	5	1	5	0	-2
I	0	-3	5	-2	-2	2	-2	-2	-1	-1	-2	-1	2	-2
C	-3	-2	-2	8	-2	-3	-3	-2	-1	-2	-3	-2	-2	8

1	2	3	4	5	6	7	8	9	10	11	12	13	14
PHE	ASP	ILE	CYS	ARG	LEU	PRO	GLY	SER	ALA	GLU	ALA	VAL	CYS
PHE	ASN	VAL	CYS	ARG	THR	PRO	---	---	---	GLU	ALA	ILE	CYS
PHE	ASN	VAL	CYS	ARG	---	---	---	THR	PRO	GLU	ALA	ILE	CYS

The second one is better because it leads a small gap compare to the huge hole of the first alignment.

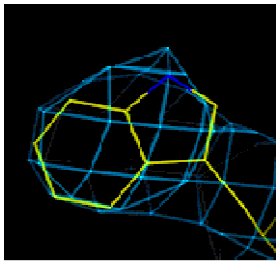


From ”Professional Gambling” by Gert Vriend
http://www.cmbi.kun.nl/gv/articles/text/gambling.html

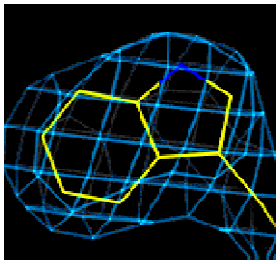
Template quality

- Selecting the best template is crucial!
- The best template may not be the one with the highest % id (best p-value…)
  - Template 1: 93% id, 3.5 Å resolution ☹
  - Template 2: 90% id, 1.5 Å resolution ☺

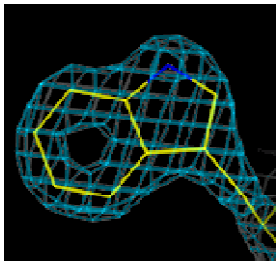
The importance of the resolution



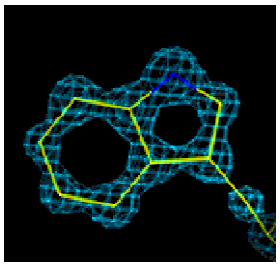
4 Å



3 Å



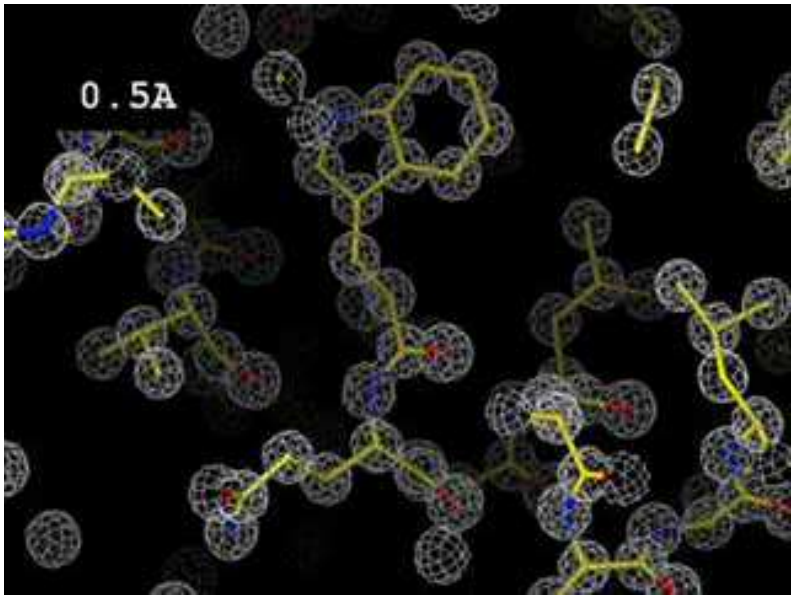
2 Å



1 Å

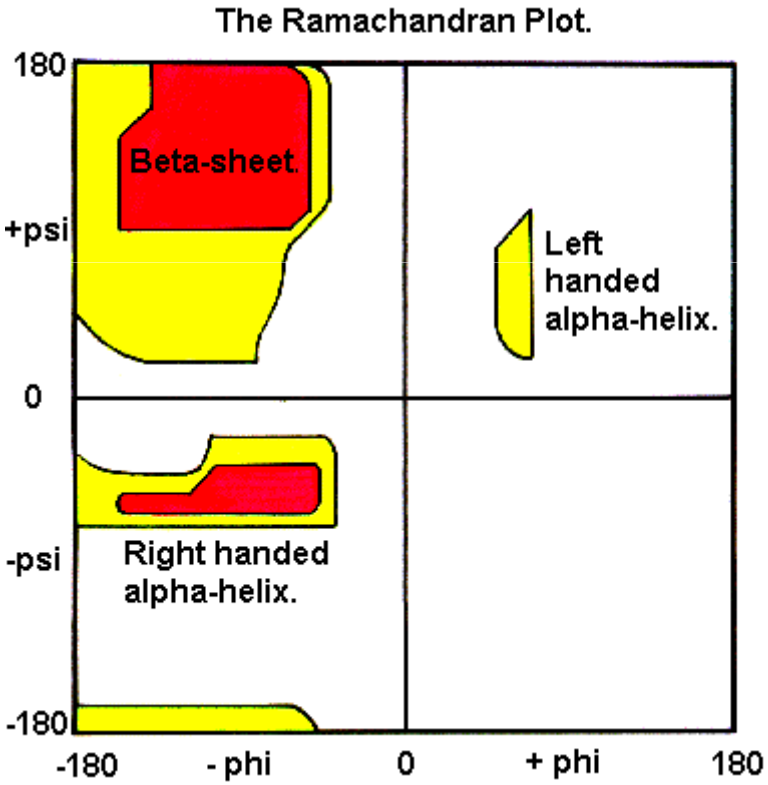
low

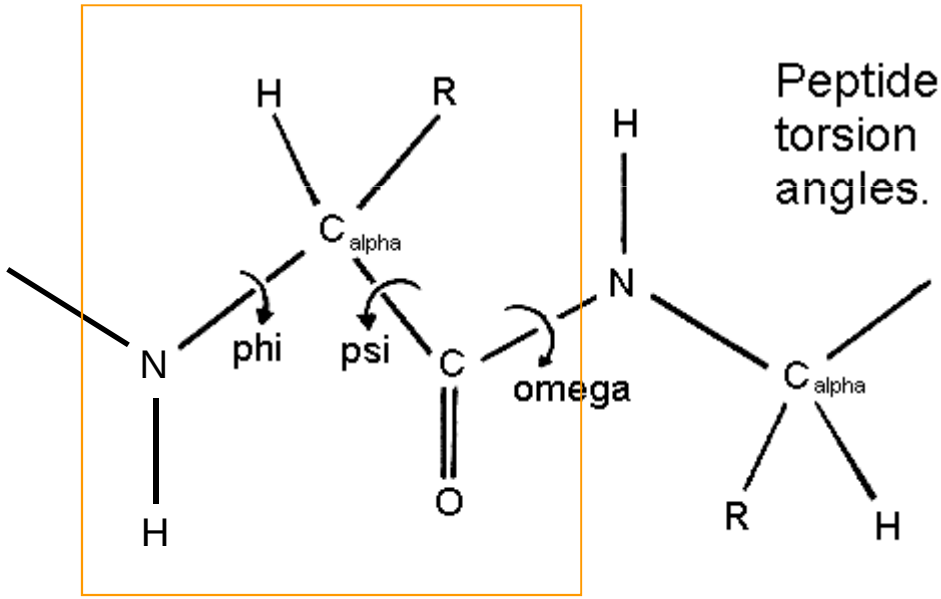
high



The Ramachandran plot

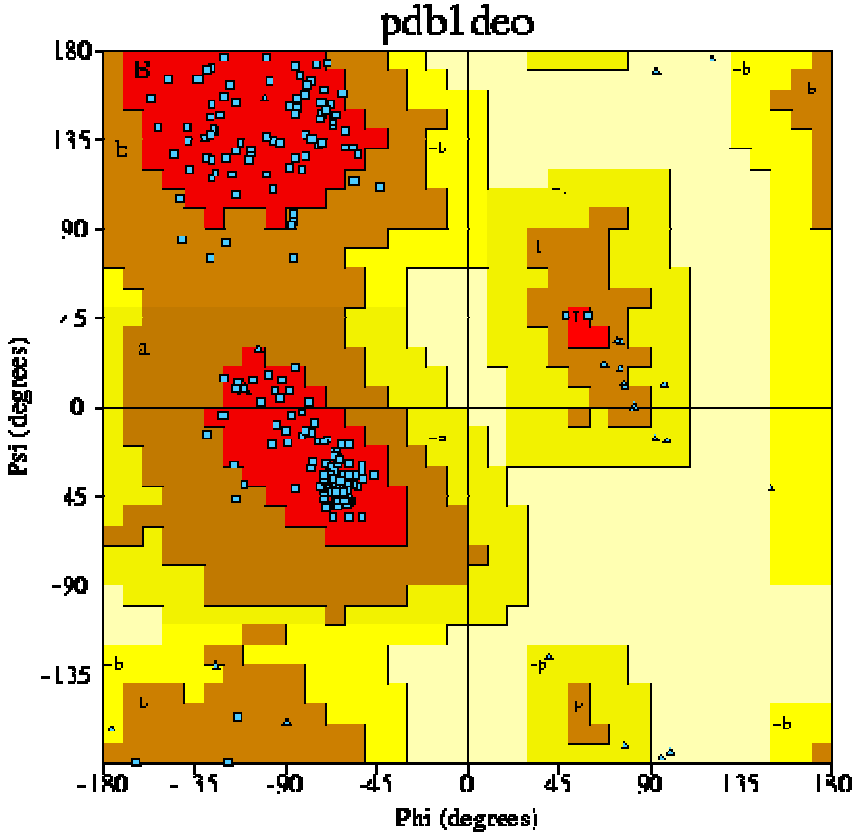
- Allowed backbone torsion angles in

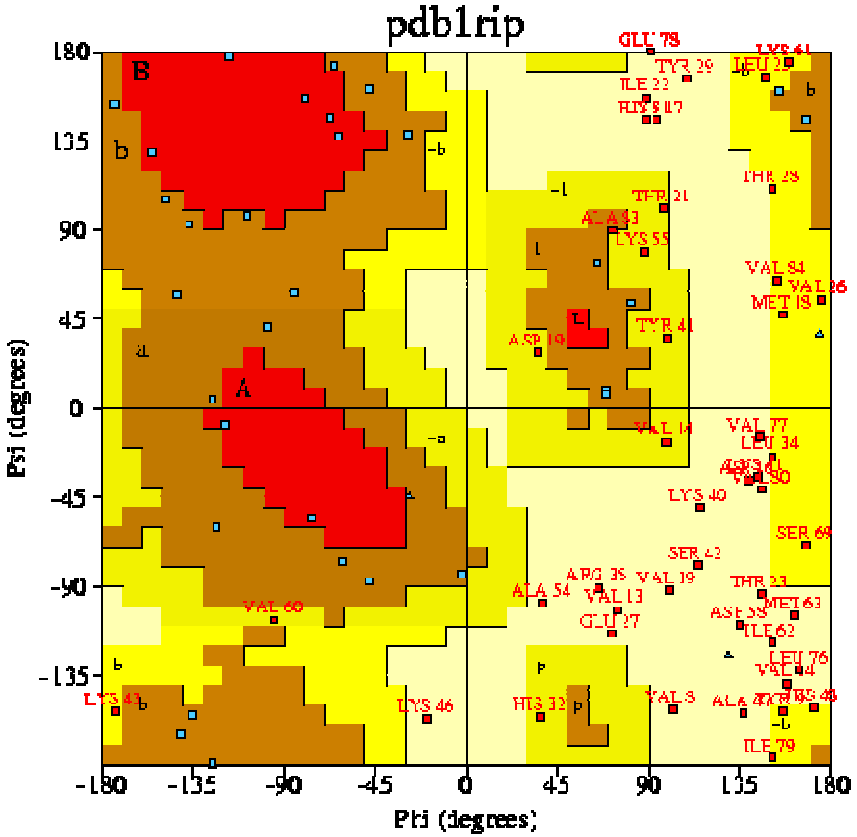




Amino acid residue

The Ramachandran plot – Template quality:





X-ray structure – good data.

NMR structure – low quality data…

Backbone generation

- Generate the backbone coordinates from the template for the aligned regions.

- Several programs can do this, most of the groups at CASP6 use Modeller:

http://salilab.org/modeller/modeller.html

Backbone generation

If the two aligned residues differ, only the backbone coordinates (N, Cα, C and O) can be copied;

If the two aligned residues are identical, it can be included also the side chian coordinates.

Loop modeling

- **Knowledge based:**
  - Searches PDB for fragments that match the sequence to be modelled (Levitt, Holm, Baker etc.).
- **Energy based:**
  - Uses an energy function to evaluate the quality of the loop and minimizes this function by Monte Carlo (sampling) or molecular dynamics (MD) techniques.
- **Combination**

Side chain

If the seq. ID is high, the networks of side chain contacts may be conserved, and keeping the side chain rotamers from the template may be better than predicting new ones.

Side chain prediction

- Side chain rotamers are dependent on backbone conformation.

- Most successful method in CASP6 was SCWRL by Dunbrack *et al.*:
  - Graph-theory knowledge based method to solve the combinatorial problem of side chain modelling.

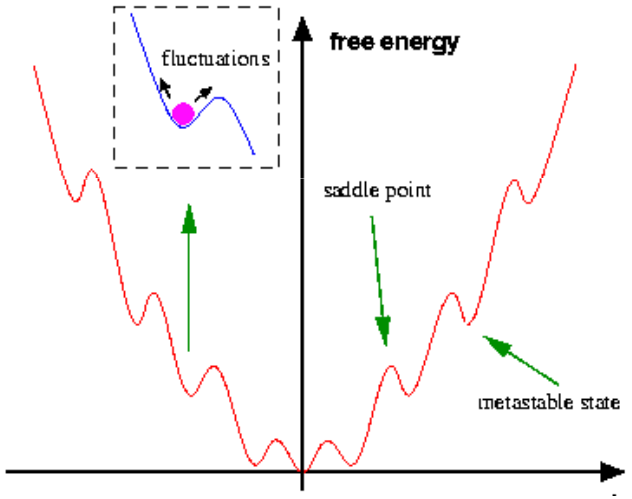
http://dunbrack.fccc.edu/SCWRL3.php

Side chain accuracy

- Prediction accuracy is high for buried residues, but much lower for surface residues
  - Experimental reasons:
    side chains at the surface are more flexible.
  - Theoretical reasons:
    much easier to handle hydrophobic packing in the core than the electrostatic interactions, including H-bonds to waters.

Model refinement

- Energy minimization
- Molecular dynamics
  - *Big errors like atom clashes can be removed, but force fields are not perfect and small errors will also be introduced – keep minimization to a minimum or matters will only get worse.*



Error recovery

- If errors are introduced in the model, they normally can NOT be recovered at a later step
  - The alignment can not make up for a bad choice of template.
  - Loop modeling can not make up for a poor alignment.
- If errors are discovered, the step where they were introduced should be redone.

Model validation

- Most programs will get the bond lengths and angles right.
- The Ramachandran plot of the model usually looks pretty much like the Ramachandran plot of the template (so select a high quality template).
- Inside/outside distributions of polar and apolar residues can be useful.
- Biological/biochemical data
  - Active site residues
  - Modification sites
  - Interaction sites

Model validation – ProQ server

- ProQ is a neural network based predictor that based on a number of structural features predicts the quality of a protein model.

- ProQ is optimized to find correct models in contrast to other methods which are optimized to find native structures.

Arne Elofssons group: http://www.sbc.su.se/~bjorn/ProQ/

Homology modeling servers

http://swissmodel.expasy.org/

SIB
Swiss Institute of Bioinformatics

http://zhanglab.ccmb.med.umich.edu/I-TASSER/

Zhang Lab

http://protein.bio.unipd.it/homer/auto.html

BioComputing