Homology Modeling

**AIM:** to get a model equivalent to what would be determined by experimental techniques (in terms of accuracy)

1) structure based drug design
2) analysis of protein function
3) interactions
4) antigenic behavior
5) rational design of proteins (stability, function)
Structure from Sequence

• Structure determined by sequence, so sequence theoretically should have all the information necessary

• Structure changes slower than sequence during evolution.

de novo Vs. Homology based methods
Structural similarity in Sequence identity vs. sequence length

75-100% identity: same accuracy as NMR

50-75% identity: fine tune details of model, like side chains

25-50% identity: best possible alignment

0-25% identity:
7 Steps of Homology Modeling:

a) Template recognition and initial alignment
b) Alignment correction
c) Backbone generation
d) Loop modeling
e) Side chain modeling
f) Model optimization
g) Model validation
h) Iteration
a) Template recognition and initial alignment

BLAST against PDB database
Best/First match usually the good template but
a) cofactors
b) active/inactive states
c) interacting partners
d) partial matches
b) Alignment Corrections:

ex: impossible aligning of YAYAYAYA and LTLTLTLT until TYTYTYTY comes along

```
Y A Y A Y A Y A
T Y T Y T Y T Y
L T L T L T L L
```

MSA could add valuable information. MUSCLE and T-COFFEE refine alignments based on profiles. Other programs may add “structure’ information to alignment.
c) Backbone Generation

Actual model building starts here
Simple transfer of co-ordinates of C-alpha, N, C, O atoms
Side chains of identical residues can also be put
In practice, C-beta atoms also can be used

Problems with model template
Multiple templates
**d) Loop modeling**
Alignment leads to gaps in model & template sequences
Gap in model sequence: omit residues from template, a hole in model that must be closed.
Gap in template sequence: inserting missing residues in the backbone

Both lead to conformational changes
Conformational changes do not occur in middle of Sec. Structures
So all changes can be shifted to loops and turns.

Loop conformation prediction is difficult b’cos
a) Surface loops may be involved in crystal contacts
b) Small to bulky aa below the loop lead to conformational changes
c) Mutation of proline/glycine to other reduces conformational flexibility
d ) Loop modeling

Three main approaches

a) Knowledge based: search PDB for loops with similar endpoints.

b) Hybrid: loop divided into small fragments and searched in PDB, then stitched together.

c) Energy based: *ab-initio*, energy function used to select the conformation of the loop.

Only loops of 5-7 aa can be predicted good. Loops may actually have multiple conformations or different conformation in crystal structure vs in solution.
e) Side Chain modeling

a) Certain favorable rotamers exist for each amino acid.
b) In homologous proteins (>40%) similarity, even C-gamma atoms will be conformationally similar.
c) As similarity among two proteins lowers, side chain conformations will also begin to change.

Rotamer prediction will be computationally demanding

b’cos each aminoacid will have multiple probabilities & b’cos conformation of one side chain affects the conformation of others.

SO conserved residue side chains can be extracted from template.

then combinations of rotamers predicted from using backbone information.
e) Side Chain modeling
Backbone aminoacid in sec structure opt for side-chain rotamers different from the ones in strands/loops. “position specific” rotamer libraries.

Rotamer prediction accuracy is higher for hydrophobic cores than for rotamers of surface residues.

a) surface side chains are flexible and have multiple conformations
b) the energy function used is better for hydrophobic packing (Van-der Waals). Electrostatic interactions are more difficult to predict.

c) crystal packing may induce changes not seen in solution.

Rotamer prediction on “correct” backbone will be more accurate than on “incorrect” backbone.

“Chicken and Egg problem”
**f ) Model optimization**

iterative modeling of sidechain rotamers and backbone.  

“Energy minimization” = global or local minima of  

“potential energy surface”.

- accuracy of energy function becomes an issue...  
- extensive energy minimization could solve big mistakes but introduce small ones at each step.
Threading (Fold recognition)

- is useful when a target sequence of interest lacks identifiable sequence matches BUT may have folds similar to other proteins of known structure.

- may assume a fold that is already characterized due to
  a) convergent evolution
  b) homologous but extremely distantly related

- Input sequence is parsed into fragments and ‘threaded” onto a library of known structures, combined to give models of protein

- Scored against known folds to see how compatible the sequence is to these folds.
Threading (Fold recognition)

3D-PSSM
PHYRE 2
FUGUE
LIBRA I
UCLA-DOE
123D