Modeling for 3D structure prediction
What is a predicted structure?

- A structure that is constructed using as the sole source of information data obtained from computer based data-mining.

- However, mixing of preexisting information (data mining/prior knowledge) and newly obtained experimental data is also common in obtaining experimental structures.

- It is mostly the relative weight of experimental versus computer generated information that determines whether the structure is considered « experimental » or « model »
Example: protonation in X-RAY

**FIGURE 2.** Comparison of the theoretical protonated and deprotonated carboxyl groups to experimental electron density in the crystal structure at 0.93 Å resolution. (a) electron density of Asp-102 compared with theoretical protonated propionic acid. In the DFT calculation, the H-O-C-C torsional angle was constrained to 109° to match the putative hydrogen position observed in the mF_{obs} − DF_{calc} electron density. (b) electron density of Asp-189 compared with theoretical deprotonated propionic acid. Theoretical electron densities of propionic acid are contoured at 2.39 e⁻/Å³, whereas the experimental 2mF_{obs} − DF_{calc} electron density map (blue) is contoured at 3.01 e⁻/Å³ (4.5 σ), and the positive mF_{obs} − DF_{calc} density map (green) is contoured at 0.22 e⁻/Å³ (2.5 σ).

**FIGURE 3.** Alternative protonation states of the imidazole ring as calculated by density functional theory and experimentally observed at residue His-57 in the electron density maps at 0.93 Å resolution. The DFT electron density of 4-ethyl-imidazole is contoured at 2.01 e⁻/Å³, whereas the 2mF_{obs} − DF_{calc} (blue) and mF_{obs} − DF_{calc} (green) electron density maps at 2.34 e⁻/Å³ (3.5 σ) and 0.26 e⁻/Å³ (3.0 σ), respectively. The experimentally observed electron density is more similar to an unprotonated imidazole ring.
Predicted structures

• Input needed: amino acid sequence only

• Which data are going to enter the modelling?
• What is the expected quality of the structure?
• How can I identify problems in the structure?
• How can I improve the structure?

• What do I need the structure for?
100% Predicted structures

Structural prediction methods

- Close homologous modeling
- Threading or distant homologous modeling
- Free modeling

Examples of biological usefulness

- Drug design
- Drug screening
- Ligand docking
- Molecular replacement
- Mutagenesis design
- Detection of enzyme active sites
- Disease substitutions
- Alternative splicing
- Domain boundary
- Topology recognition
- Family assignment
Data entering the modelling

- Experimental structures of proteins that differ in sequence from the protein of interest
- Close sequence homology: homology modelling
- More distant sequence relationship/shared fold with no obvious sequence relationship: threading. Aim at recognizing a related fold.
- No fold assigned: Folding from sequence alone.
modelling by homology
or knowledge-based modelling
Why *in silico* modelling of protein structure?

- it is the only way to obtain structural information if experimental techniques fail.
  - insoluble proteins like membrane receptors
  - proteins too large for NMR analysis
  - proteins that could not be crystallized for X-ray diffraction
- slow increase of number of solved 3D-structures, yet ca. 28000 structures available in the PDB (templates)
- fast increase of number of protein sequences with no structural informations available
Why it is useful to know 3D structure of a protein, not only its sequence?

**biochemical function** = interaction with other molecules

**biological function** = consequence of these interactions

3D structure is more informative than the sequence because interactions are determined by amino acids that are **close in space** yet frequently **distant in sequence**.

Function depends more directly on structure than sequence.
Examples of 3D structure uses

• identification of active and binding sites (substrates, co-factors, inhibitors, modulators, proteins, nucleic acids..);

• designing and improving ligands of a given binding site (e.g., structure-based drug design);

• understanding functional /physical aspects of multimeric assemblies;

• testing function hypothesis (e.g. site-directed mutagenesis);

• rational design of proteins with increased stability or novel functions;

• characterizing new sequences (e.g., detecting remote relationships between proteins);

• antigenic behavior (prediction of epitopes);

• characterization of biological networks
Why not all structures experimentally?

• Many more sequences than structures
• New structures often « look like » old ones
• Redundancy of structural information in the PDB:
  
  less than 3000 unique folds
• 2 peptidic chains share the same fold whether
  
  – RMSD < 3.0 Å
  
  – number of aligned positions ≥ 70% total chain length
The purpose of the Protein Data Bank is to collect and organize 3D structures of proteins, nucleic acids, protein-nucleic acid complexes and complexes with drug molecules and inhibitors.

The PDB is the recognized worldwide repository for 3D structures.

After validation of deposited structures, the coordinates are made available to all researchers worldwide and free-of-charge.

http://www.pdb.org
PDB: fold number

new (blue) & old (red)
Homology modelling is based on 2 observations:

1. The structure of a protein is determined by its amino acid sequence. Knowing the sequence should, at least in theory, suffice to obtain the structure.

2. During evolution, the structure is more stable and changes much slower than the associated sequence:
   - similar sequences adopt practically identical structures
   - distantly related sequences may still fold into similar structures
SSAP score: measure of structural similarity

• When the sequence identity is $\geq 30\%$ between 2 sequences, and experimental 3D information is available for one of the sequences, it is possible to build in silico a 3D model for the sequence of unknown structure

• accuracy of a structure predicted from its sequence is much lower than the results achieved experimentally: no homology modelling structures in the PDB!

• theoretical protein 3D structures represent "low-resolution" models, and hence should be used ACCORDINGLY
high sequence similarity results in structure similarity:

- rmsd of the C$\alpha$ co-ordinates for protein cores sharing 50% residue identity is expected to be around 1Å
- two sequences of identical length which share more than 30% sequence identity virtually adopt a similar structure.

Comparative model building consists of the extrapolation of the structure for a new (target) sequence from the known 3D-structure of related family members (templates).
homology modelling workflow

1. Template recognition and initial alignment
2. Alignment correction
3. Conserved regions modelling
4. Loops modelling
5. Side-chain modelling
6. Model optimization
7. Model validation
Identification of modelling templates

query using FastA and BLAST to retrieve from the PDB sequence of target homologous proteins

A template must share at least 30% residue identity with the target.

whether many template are available

• best template structure (highest sequence similarity to the target) = reference
• all structures used to generate structurally correct multiple alignment of the sequences
• Choose the best structure: the quality of the model could not be better than that of the template
Check the template structure

- The **PDBREPORT** database (C. Sander, EMBL) indicates problems in structures for every PDB entry. http://www.cmbi.kun.nl/gv/pdbreport/

- Generally **X-ray structures** better than NMR structures

- X-ray structure quality attested by the **resolution factor**

- Yet also pay attention to **temperature** (conformation freezeed at low temperature, see local flexibility indicated by B-factor)

- **External conditions**: pH, ions, ligand, ....
Aligning the target sequence with the template sequence

Some residues should not be used for model building, for example those located in non-conserved loops

Analyzing the alignment

- Identification of conserved regions
- No insertion / deletion in conserved regions

The quality of alignment determine the quality of the modelled structure!
Steps in constructing a Multiple Alignment

Muliple sequence alignment
BAliBASE reference 3: aldehyde dehydrogenase-like
Lessons from evolution...

Upon evolution

• mutation are observed rather than insertion or deletion (indel)
• shorter indels are easier to make than longer ones
• active site residues are conserved
• Cysteines involved in disulfide bridges are conserved
• Core residues are better conserved than those exposed at the protein surface
Lessons from evolution...

Upon evolution

• residues tend to mutate into similar residues
  (e.g. V <-> I; S <-> T; see scoring matrix like PAM and BLOSUM )
• residues mutate more easily to residues encoded by similar codons
• Some residues cannot be easily mutated because of their unique structural features (e.g., accessible phi-psi values for Glycine, Proline)
Additional informations that can be used to correct sequence alignment...

Knowledge about

• proteolytic cleavage sites
• metal binding residues
• ligand binding residues
• Structure:
  - structure of homologous protein
  - predicted secondary structure

In any cases, use **multiple alignment** rather than pairwise alignment
1. **Conserved regions**

   Conserved regions are built using template structural informations,
   All methods keep the structure of the template in conserved regions with few adjustments (adjustements are necessary when there are deletions in target with respect to template)

2. **Non conserved regions**

   2a. Backbone of inserted sequences (mostly loops), need to be constructed or refined

   2b. Side chain of mutated residues need to be reconstructed/refined.
   This may involve local changes around the mutation in conserved regions (repacking)
Example: modelling by spatial restraints

Spatial restraints

molecular mechanics using distance restraints extracted from the structures of one or several targets
Modeller
(Andrej Sali)
The 3D structure of ECD1–CRF-R2β.
First extracellular domain of a type B1 G protein-coupled receptor

(A) A ribbon diagram of the lowest energy conformer highlighting the β-sheets in cyan and the disulfide bonds in yellow.

(B) Superposition of 20 conformers representing the 3D NMR structure. Only amino acid residues 44–119 are shown. The bundle is obtained by superimposing the backbone Cα carbons of residues 58–83 and 99–113.

Grace C R R et al. PNAS 2004;101:12836-12841
Molecular dynamics as a tools for structure determination

Molecular dynamics as a means for finding structures that satisfy restraints. MD explores many structures, and finds stable ones.
• Energy surface: classical MM force field

\[ E_{\text{Total}}(X^N) = E_{\text{bonded}}(X^N) + E_{\text{nonbonded}}(X^N) \]

\( X^N \) : atomic positions

• Energy surface with restraints

\[ E_{\text{Total}}(X^N) = E_{\text{bonded}}(X^N) + E_{\text{nonbonded}}(X^N) + E_{\text{RESTRAINTS}}(X^N) \]

\( X^N \) : atomic positions

• Example of restraint: spring between two atoms that are known to be close in space, from some independant observation (artificial bond)

i and j are NOT chemically linked, but the artificial bond will pull them close together:

advantage of computer simulation: the potential surface can be « manipulated ». 
Example: modelling by spatial restraints

« Artificial bonds » will pull atoms close in space, even if there is no real chemical bond between them.
Incorporate restraints into potential energy function

\[ E_{Total}(X^N) = E_{bonded}(X^N) + E_{nonbonded}(X^N) + \sum_i E_{RESTRAINTS}^i(X^N) \]

Then return to the formalism developed for unbiased molecular dynamics simulations …

\[ F_i = m_i \cdot a_i = m_i \cdot \frac{dv_i}{dt} = m \cdot \frac{d^2r_i}{dt^2} \]

\[ F_i = -\nabla_i E \]
MD with restraints

• To satisfy experimental observations

• Most noted example is NMR structure determination:
  Many experimentally derived restraints: the structure is MOSTLY determined by this data

• Few experimental restraints: the structure is a « mix » of computational and experimental information

• Restraints from other source than experiment, expl, from a sequence alignment : HOMOLOGY MODELLING
  Protein A should « look » like protein B, because the sequence of A and B are similar. If 3D structure of B is known, then A can be modelled « by homology » to B.
Restraints obtained from an NMR experiment

Nuclear Overhauser effect (NOE)

Chemical shifts

Hydrogen bonds

$^{3}J$ scalar couplings
The 3D structure of ECD1–CRF-R2β. First extracellular domain of a type B1 G protein-coupled receptor

(A) A ribbon diagram of the lowest energy conformer highlighting the β-sheets in cyan and the disulfide bonds in yellow.

(B) Superposition of 20 conformers representing the 3D NMR structure. Only amino acid residues 44–119 are shown. The bundle is obtained by superimposing the backbone Cα carbons of residues 58–83 and 99–113.

Grace C R R et al. PNAS 2004;101:12836-12841
spatial restraints from sequence alignment

template: PDKNI SRA

| | | | | | |

target: PDKNALTRA

restraint = harmonic energy

\[
E_{Cα-Cα}^{P→A}(target) = \frac{1}{2} k \left[ r_{Cα-Cα}^{P→A}(target) - r_{Cα-Cα}^{P→A}(template) \right]^2
\]

PDB: template

MODEL: target
Non conserved parts

• In modelling by spatial restraints, the missing loops and mutated side chain are incorporated automatically in the structure.

• Yet, since there is no information on these parts in the PDB template, their position is modelled in a less reliable way.

• Specialized loop/side chain reconstruction algorithms can be applied afterwards, to improve the model.
Programmes

**segment matching / rigid body**

**Software**
- SYBYL (tripos)
- MOE (CCG)

**Freeware**
- SwissPDBViewer

**Serveurs WWW**
- Swiss-model(Expasy)

**Spatial constraints**

**Freeware**
- MODELLER (accelrys)

**WWW tools for structure analysis and validation**

- WHATIF
- PROCHECK
structure evaluation

= validation step

• **geometry analysis**: angles length, bonds & torsion values, planarity of aromatic rings, steric clashes

• pseudo energetical analysis

• **experimental validation**: choice of experiments based on structural assumption (e.g., site directed mutagenesis)
General Criteria for protein structure validation

• local Geometry :
  bond length, angle values, chirality, planarity

• general quality :
  – Ramachandran plot
  – rotamers (χ angle)
  – bumps
  – h-bonds and salt bridges network
  – buried hydrophobic
  – aromatic stacking
**planarity**

**backbone**
Torsion Angle Omega

*Trans- & Cis-conformation*

**side chains**
Ex: ARG

good  bad
Ramachandran Plot

good

bad
Reminder

Ramanchandran

Two degrees of freedom (dihedral angles phi et psi) along the protein backbone
Backbone conformations

The Ramachandran Plot

- Beta-sheet
- Left handed alpha-helix
- Right handed alpha-helix

\( \Phi \)
Backbone conformations

The Ramachandran Plot

Psi
In experimental structures, not all values of $\phi/\psi$ are observed. Some combinations are never seen because they correspond to unstable structures (steric clash - high energy)
Each point correspond to ONE AA in the structure:

**Good:** most of the AA in Allowed (low energy) regions.

**Bad:** many AA in disallowed (high energy) regions.
rotamers

Bad

Good
Steric clashes between 2 atoms of the backbone
H-bond network

Ex: Crambine
salt bridge network

bad
good
quality of the packing

bad

good
backbone Conformation

normal

strange!!
protein folding

The best model has the lowest free energy

What is difficult:

- To compute the free energy (NOT THE SAME AS ENERGY)
- Absolute free energy: very large, yet variation of free energy between 2 folded conformations very small (5-10 kcal / mol)
Empirical energy functions

E.g. Bowie, Luthy, Eisenberg
Science 253 164 (1991)
Nature 356 83 (1992)

Definition of 18 classes that describe amino acid properties:

- **accessibility** & **secondary structure**
  - E: solvent exposed
  - P1, P2: partly exposed
  - B1, B2, B3: buried
  - *α helix
  - β sheet
  - other

For the 20 amino acid types: **Score s** assigned to the 18 classes

- values from statistical analysis of known 3D structures
- High score ⇔ good environment
- E.g., W: B1. \( \alpha = 1.00, \ E \alpha = -1.35 \)
  - K: B1. \( \alpha = -1.82, \ E \alpha = 0.13 \)
Empirical energy functions

Application to 3D model evaluation

For every amino acid i:

- Side chain buried surface
- Percentage of the buried surface facing polar atoms (including water)
- Local secondary structure

score:

\[ S(\text{model}) = \sum_{i=1}^{N} s_i \]

N: number of amino acids

Best model: S1 > S2
Empirical energy functions

NOTE that the scale is arbitrary:

In the e.g. in Bowie, Luthy, Eisenberg energy function, a high score is good.

In the Modeller DOPE score, which is also an empirical function, although with different parameters, a LOW score is good.
In practice

- Databases of already built models

ModBase, A. Sali
In practice

• Databases of already built models

Swiss-Model Repository, T. Schwede