Bioinformatics Master Course

DNA/Protein Structure-function Analysis and Prediction

Lecture 10: Protein structure prediction (iii): rotamers and molecular modeling
Synopsis

• Given a backbone structure of a protein structure (i.e. given the main-chain atoms or C-alpha atoms only), for example resulting from homology modelling or fold recognition, how can we build in the side-chains?

• This problem has been referred to as the ‘Jigsaw Puzzle problem’ or ‘Jigsaw Problem’
  – The idea is that each side-chain has an influence on the positioning of every other side-chain in the structure
  – This leads to a combinatoric problem.
  – But is this the true scale of the problem?
Ramachandran plot

- Only certain combinations of values of phi (\(\phi\)) and psi (\(\psi\)) angles are observed.

This is the situation with main-chain atoms. The Ramachandran plot attempts to bring some order in conformational space.

Can we do something similar with side-chain atoms?
Rotamers: highly populated combinations of side-chain dihedral angles ($\chi_1, \chi_2, \ldots$ angles)
Example: **Lys has four $\chi$ angles**

Torsion Axes and Dihedral Angles of the side chain of Lysine

The sample amino acid Lysine has four torsion axes within its side chain. The torsion axes are symbolized as arrows, the dihedral angles are labeled $\chi_1$ to $\chi_4$. 
Side-chains have positional preferences for types of interaction

The pi-system of a tyrosine residue. The out-of-plane region prefers hydrophobic (green) contacts, whereas the in-plane region prefers hydrogen-bonding (red) contacts.

The beta carbon of alanine (non-pi-system atom). The green region indicates the fairly symmetric preference for hydrophobes around the atom.

From http://www.chemcomp.com/journal/rotexpl.htm
Side-chains turn out to have preferences for discrete parts in space..

Fig. 1. Example of two overlapping rotamer distributions.

Fig. 2. Examples of rotamer distributions with different rotamer freedom. Rotamers are shown, superposed on the corresponding local backbone structure. The rotamer freedom is indicated (1.0 means no freedom; 0.33 means maximal freedom).
Rotamers

- are usually defined as low energy side-chain conformations.
- the use of a library of rotamers allows the modeling a structure while trying the most likely side-chain conformations, saving time and producing a structure that is more likely to be correct.
- This only happens if the rotamers used really are the correct low energy conformations.

To make a rotamer library:

- use only very high resolution structures (1.7 Å or better),
- remove side chains whose position may be in doubt using a number of filters,
- we use the mode rather than the mean of observed conformations (which has a number of advantages), and
- make efforts to remove systematically misfit conformations.

This is done by:

Example: rotamer libraries for Arg and Val

| Res | Rotamer | n(r1) | n(r1234) | p(r1234) | sig | p(r234|r1) | sig | chi1  | sig1  | chi2  | sig2  | chi3  | sig3  | chi4  | sig4  |
|-----|---------|------|---------|---------|-----|-----------|-----|-------|-------|-------|-------|-------|-------|-------|-------|
| ARG | 1 1 1 1 | 600  | 3       | 0.05    | 0.55| 0.24      |     | 63.1  | 6.8   | 84.3  | 11.9  | 64.4  | 9.3   | 81.1  | 7.5   |
| ARG | 2 1 1 1 | 2115 | 43      | 0.66    | 0.08 | 2.02     | 0.25| -179.2| 10.7  | 65.3  | 8.3   | 59.6  | 8.3   | 84.7  | 10.5  |
| ARG | 3 1 1 1 | 3738 | 10      | 0.17    | 0.04 | 0.30     | 0.07| -78.9 | 13.7  | 88.3  | 16.3  | 69.0  | 28.2  | 8 8.7 | 13.6  |
| VAL | 1 0 0 0 | 891  | 891     | 7.71    | 0.20 | 100.00   | 0.00| 64.7  | 12.6  |       |       |       |       |       |       |
| VAL | 2 0 0 0 | 8469 | 8469    | 73.25   | 0.34 | 100.00   | 0.00| 175.6 | 7.4   |       |       |       |       |       |       |
| VAL | 3 0 0 0 | 2201 | 2201    | 19.04   | 0.30 | 100.00   | 0.00| -61.2 | 9.3   |       |       |       |       |       |       |

Here, chi1- chi4 and sig1- sig4 denote the side chain dihedral angles and standard deviations (red box). If you want details about the statistics outside the red box, consult R. L. Dunbrack, Jr. and F. E. Cohen. "Bayesian statistical analysis of protein sidechain rotamer preferences." Protein Science, 6, 1661-1681 (1997).

Arg has four chi (χ) angles, Val has only one.
Lovell et al., 2000

• All-atom contact analysis shows that all published rotamer libraries to date contain serious van der Waals overlaps (side-chain clashes)
• This should not occur as rotamers, being the more common conformations, should have the lower energy states.
• Using a select database of 240 high resolution, low-clash score, low $R_{\text{cryst}}$ structures and then filtering it by B-factor and clash score, Lovell et al. composed a rotamer library, consisting of 153 conformers, which they think is more faithful to the rotamer concept and will improve accuracy of new structures.
• The library is available as an O database.
\( \chi_1 \) angle

\[ X = \text{CHL-1; total residues 67608} \]

\[ Y = \text{Nr of observations in 1-degree bin} \]

Noise 187 Statt peaks 3 Merged 3 SUM% 94.8
\( \chi_2 \) angle

\( X = \text{CHI-2}; \text{total residues} \ 4726 \ l \\
Y = \text{Nr of observations in 1-degree bin} \\
\text{Noise 130 Stat peaks 10 Merged 5 SUM\% 96.7} \)
χ₃ angle

X = CHL-3; total residues  13416
Y = Nr of observations in 1-degree bin
Noise 37 Start peaks 19 Merged + SUM% 96.5
$\chi_4$ angle

$X = \text{CHI-4;}$ total residues 5781
$Y = \text{Nr of observations in 1-degree bin}$
Noise 16 Start peaks 13 Metged 4 SUM\% 87.2
# Backbone-dependent rotamer libraries

Based on the backbone-dependent rotamer library of Dunbrack and Karplus (1993), Bower et al. (1997) present a method for rapidly predicting the conformations of protein side-chains, starting from main-chain coordinates alone.

The method involves using fewer than ten rotamers per residue from a backbone-dependent rotamer library and a search to remove steric conflicts. The method is initially tested on 299 high resolution crystal structures by rebuilding side-chains onto the experimentally determined backbone structures.

A total of 77% of chi1 and 66% of chi(1 + 2) dihedral angles were predicted within 40 degrees of their crystal structure values.


*Modeling by homology is about placing the polypeptide backbone and adding side-chains.*
Rotamers: to be or not to be?


- Please read these papers. Have you got criticisms? (don’t worry, your teacher can handle it).

**Strengths/weaknesses?**
Non-rotamericity

Many side-chains are outside 20° (or even 40°) of the nearest rotamer (defined by the $\chi_1$ and $\chi_2$ angle) -- potentially leading to unfavourable and high-energy sites.
A cluster of five non-rotameric side-chains (further than 20° away from nearest rotamer ($\chi_1, \chi_2$)) in the oligopeptide binding protein from Salmonella typhimurium(2olb chain A). Cluster constituent side chains are Leu297A, Arg299A, Ile302A, Trp382A and Val388A.
Self-Consistent Mean Field (SCMF) modeling

Mean field Energy

\[ E(i,k) = U(i,k) + U(i,k,\text{Backbone}) + \sum_{j=1}^{N} \sum_{I=1}^{N_{\text{rot}(i)}} P(j,I)U(i,k,j,I) \]

Updating the probabilities

\[ P_{\text{new}}(i,j) = \frac{\exp\left(-\frac{E(i,j)}{kT}\right)}{\sum_{l=1}^{N_{\text{rot}(i)}} \exp\left(-\frac{E(i,l)}{kT}\right)} \]

Modeling side-chain conformation using SCMF. SCMF is based on a multi-copy sampling in conformation space: the protein is replaced by an effective system containing one copy of the backbone, and multicopies of all sidechains. Each copy \( k \) of a residue \( i \) is given a weight \( P(i,k) \). The weights are initialized to follow a uniform distribution: \( P(i,k) = 1/K_i \), where \( K_i \) is the number of copies of residue \( i \). The meanfield energy "felt" by copy \( k \) of residue \( i \) is the sum of three terms:

1. the internal energy of rotamer \( k \) (green term),
2. the interaction of rotamer \( k \) with the backbone (red term),
3. sum of the interactions of rotamer \( k \) with all copies of all other residues in the effective system (blue term).

The effective energies are used to update the probabilities, following a Boltzmann-like law [18]. The whole procedure is iterated until the total energy of the system does not change, i.e., when it has reached self-consistency. The sidechain conformation of residue \( i \) is then chosen to be the copy with the highest converged probability.

Molecular modelling helped by Experimental Data

Many experimental data can aid the structure prediction process. Some of these are:
• Disulphide bonds, which provide tight restraints on the location of cysteines in space
• Spectroscopic data and secondary structure prediction, which can give you and idea as to the secondary structure content of your protein
• Site directed mutagenesis studies, which can give insights as to residues involved in active or binding sites
• Knowledge of proteolytic cleavage sites, post-translational modifications, such as phosphorylation (at Tyr sites) or glycosylation (e.g. N-glycosylation sites are specific to the consensus sequence Asn-Xaa-Ser/Thr) can suggest residues that must be accessible

Remember to keep all of the available data in mind when doing predictive work. Always ask yourself whether a prediction agrees with the results of experiments. If not, then it may be necessary to modify what you've done.
Importance of Molecular Modelling

• The 1998 Nobel Chemistry Prize was awarded to Pople and Kohn for their work in Computational Chemistry and Molecular Modelling.

• The 1999 Nobel Chemistry Prize was awarded to Ahmed Zewail for his work in developing spectroscopic methods for studying reactions and in particular transition states, an essential aspect of molecular modelling.
Simple Definition of Molecular Modelling:

Molecular modelling is a collection of (computer based) techniques for deriving, representing and manipulating the structures and reactions of molecules, and those properties that are dependent on these three dimensional structures.

- Molecular visualisation
- Molecular mechanics
- Geometry minimisation and transition state location
- Semi-empirical and \textit{ab initio} molecular orbital theories
- Modern computer programs for performing molecular modelling
Search algorithms in sidechain conformation space

- Two classes of search algorithms in scientific computing: stochastic and deterministics.
- Stochastic algorithms such as Monte Carlo [15] and genetic algorithms [16] follow probabilistic trajectories and converge, but are not guaranteed to reach the global minimum of the system. Their outcome is also dependent on their initial conditions and on the random number generator seed.
- Deterministic methods such as the Dead End Elimination [17] and SCMF [18] will find the same results for a given set of parameters. They do not always converge, most of the time because of the computational time they require.
- Both classes of algorithms have been applied to the problem of modeling sidechain conformation. The same methods can be used for protein design.
Modelling new IgM structure using two templates and an antigen (peptide ligand)

The aim of this study was the construction of a model of the immunocomplex between MIR analogue and anti-AChR autoantibodies

The MIR decapeptide (the peptide ligand) is a Torpedo MIR analogue which has about twofold enhancement of binding capacity to mAb198 in comparison with the human MIR analogue.

The Antibody structure AChR was modelled by homology

The Antibody-peptide complex (AChR - MIR) contacts were determined by NMR

“Directed modelling”: the AChR antibody loops that need to be modelled are influenced by the peptide ligand binding in the binding groove and vise versa.
Immunoglobulin basic structure

This is a schematic cartoon of an IgG molecule showing some of the features of the molecule including the flexibility of the Fab and Fc regions. This schematic can be compared with the other images shown here which have been rendered from crystal structures of the fragments of Ig molecules.
Immunoglobulin basic structure

This is a ray-traced image of the model of human IgG1 showing the two heavy chains in red, the two light chains in yellow and the carbohydrate attached to the heavy chains in purple. The rotational symmetry about a vertical axis can be clearly seen in this picture.
Some loops shift more than others...
Compared “top” stereo view (MOLSCRIPT56) representation of all CDRs (thick ribbons) in the context of the light and heavy chain variable domains for the scFv198 (up) and Pot IgM (down) antibodies. The light chain is situated on the left side.

CDRs: Complementarity-Defining Regions, i.e. hypervariable parts of the variable domains that interact with an antigen.