Structural Bioinformatics
(C3210)
Conformational Analysis
Protein Folding
Protein Structure Prediction
Conformational Analysis
Conformational Analysis

- Properties of molecules depend on their three-dimensional structures (i.e. conformations)
- **Conformational analysis** is the study of the conformations of a molecule and their influence on its properties
- Conformational analysis is used in drug design to search conformations of small molecules (putative drugs)
- In protein folding this is used to find protein 3D structure with minimal energy that usually corresponds to biologically active structure
- A key component of conformational analysis is the conformational search, the objective of which is to identify 'preferred' conformations, i.e. conformations with low energies
The conformations of a molecule are those arrangements of its atoms that can be interconverted purely by rotation about single bonds.

Different conformations have different energy because of electronic changes, steric clashes, non-bonding interactions, etc.

For molecules with one rotatable bond the “conformational potential surface” consists of the curve representing the molecular energy as a function of the dihedral (torsion) angle. The minima of the curve correspond to low energy conformations.

For molecules with two rotatable bonds the total energy is represented as a function of two variable torsion angles.
Conformational Potential Surface

- Most molecules have more than two variable torsion angles. Their conformational potential surface is defined in more than 3 dimensions, which is called potential energy hypersurface.
- Nevertheless, their surface can be represented schematically as a curve displaying the energy as a function of a parameter (p) that has no real geometric meaning.
**Energetic Minima and Maxima**

- The conformational potential surface contains **local minima** (low energy conformers), **barriers** (high energy conformers) and a **global minimum** (the conformation of lowest energy).
- One of the most common goals of conformational analysis is to find a global minimum, i.e. the **conformer with lowest energy**.
- Conformers with the lowest energy are **most highly populated** and they are usually responsible for chemical and biological properties of the molecule.
Searching Minima

• The conformer corresponding to **local energy minimum** can be easily found using **optimization methods** if other conformer with higher energy is known that is not separated by energy barrier.

• The strategy in **finding global minima** is to generate suitable set of conformers that are used as a starting point for local minimization methods which finds the nearest local minima.
Conformation Searching Methods

- The following algorithms are used to find global minima on conformational potential surface:
  - **Systematic search** (also called 'grid search', 'exhaustive search' or 'brute force search')
  - **Random search** (Monte Carlo search)
  - **Simulated annealing**
  - **Genetic algorithms**
  - **Distance-geometry algorithms**
  - **The fragment approach**
  - **Chain growth**
  - **Rule-based systems**
Systematic Search

- A simple method for exploring the conformational potential surface is systematic scanning of all geometries of the molecule.
- Within this systematic searches, dihedral angles are changed systematically by specified increment value.
- Too big increment values can result in missing the global minima while small increment values significantly increase algorithm complexity.
- Complexity of systematic search grows rapidly with number of rotatable bonds \((\sim(360/m)^n)\) where \(n\) is number of rotatable bonds and \(m\) is increment value.
Random Search

- Random search methods (also called Monte Carlo) combine local minimization with random change of dihedral angle value.
- At each iteration, a random change is made to the 'current' conformation. The new structure is locally optimized. If the minimised conformation has not been found previously, it is stored. Then the process is repeated with new conformation which is selected based on specific rules (different for various implementations of the method).
- This algorithm is faster than systematic search but finding global minima is not guaranteed (unlike a grid search with a small step size).

Figure: Starting conformer 1 is provided by user. Local minimum 2 is found by local optimization techniques. Random change leads to 3. Its local optimization gives 2 which is abandoned. New random change of 2 provides 4 which is locally optimized to get 5. This is accepted as new conformer. New starting point is selected (2 or 5) and process continues.
Molecular Dynamics

- Each atom of a protein has a potential energy and therefore feels a force exerted on it
- This force can be used to simulate motion of protein atoms using equations of motion from classical mechanics
- Position of each atom can be calculated along a series of extremely small time steps and the resulting series of snapshots of structures over time is called a trajectory
- Molecular dynamics simulation should mimic behaviour of real molecule
- Within molecular dynamics simulation the molecule adopts different conformations thus it searches conformational space

- Video - molecular dynamics of peptide molecule: https://www.youtube.com/watch?v=HtRbtQ12eOI
Simulated Annealing

- A simulated annealing algorithm is a molecular dynamics simulation, in which the amount of kinetic energy in the molecule (the simulation temperature) is high at the beginning and it slowly decreases over the course of the simulation.
- At the beginning of the simulation, many high-energy structures are being examined and high-energy barriers can be crossed. At the end of calculation, only structures that are close to the best-known low-energy structures are examined.
- This algorithm is most effective for finding low-energy conformers that are similar in shape to the starting geometry.
Genetic Algorithms

- Genetic algorithms are inspired by biological evolution, they stem from the observation that the evolution process tends to produce increasingly well-adapted populations.
- At the start, the algorithm creates random population of individuals (individual = conformer). Each individual is encoded by genes (gen = value of one dihedral angle).
- Fitness of individuals is evaluated (i.e. energy is calculated) and individuals with high fitness (low energy) are reproduced to make new generation of individuals.
- Some genes are mutated (some their dihedral angle values are randomly changed) and crossover is performed (values of dihedral angle are switched between some pairs of individuals).
- The process is repeated until it converges (i.e. individuals of new generation have almost the same energy as several previous generations).
- A genetic algorithm is an effective way of generating large number of low-energy conformers. However, there is no guarantee that a global minimum will be found.
- Practical test have shown genetic algorithms to be superior to simulated annealing and random search.
Other Algorithms

- **Distance-geometry algorithms** are used if some atom distances in the structure are known (typically from NMR experiment). These distances are used as constraints within a conformation search.

- **The fragment approach** is based on optimization of one part of a molecule at a time. For instance, protein side chains are individually optimized and subsequently the backbone is optimized while keeping the side chains fixed.

- Within the **chain growth algorithm**, the full molecule is built up one unit at a time. As each unit is added, its conformation is searched without changing the rest of the chain.

- **Rule-based systems** try to identify certain subsequences of amino acids that tend to have a particular secondary structure (α-helices, β-threads, etc.). These sections can be held rigid while the conformations of the connecting fragments are searched.
Protein Folding
Importance of Protein 3D Structures

- The formation of a protein in its biologically active form requires the folding of the protein into a precise three-dimensional structure.
- Changes in this structure (for example by denaturation) are associated with loss of function of that protein.
Protein Folding Problem

- The process by which a linear amino acid sequence of a protein is converted into a precise three-dimensional architecture is a key issue in biochemistry, and is known as the protein folding problem.
Anfinsen's Experiment

- Christian Anfinsen experimentally demonstrated (in 1961) that denaturated ribonuclease enzyme spontaneously refolds into its active conformation \textit{in vitro}.
- This proved that there is sufficient information in the protein sequence to determine its correct folding.
Anfinsen's Dogma and Levinthal's Paradox

- The underlying principle that can be deduced from Anfisen's experiment is that the energetically favored state of a protein coincides with its correctly folded one. However, Levinthal demonstrated that a systematic random search of all possible conformations of the protein would take almost infinite amount of time until the most stable form is found, whereas in real cells this occurs within seconds.

The Levinthal Paradox

If a chain of 100 amino acids is considered and each amino acid assumed to have 3 possible conformations, there are $3^{100}$ possible ways to arrange this chain. Bond rotation can be estimated to occur in a rate of roughly $10^{14}$ sec$^{-1}$. This means that random searching of the right conformation would take the order of $10^{34}$ seconds or $10^{26}$ years.

Real protein fold over a time scale of 0.1 to 1000 seconds!
The Pathway Theory and Energy Funnels

- The most plausible hypothesis was that folding follows a limited number of pathways, until the protein reaches the final fold. Thus, the search for a free energy minima is more like balls rolling down funnels rather than balls rolling aimlessly on a flat surface. The funnel energy landscape thus accommodates folding kinetics and thermodynamic stability.
Mechanisms of Protein Folding

- Many intricate mechanisms were put forward to describe simple pathways of protein folding. Some of these models include the initial formation of secondary structure elements, the initial formation of compact shapes by hydrophobic collapses, or even more complicated combinations of these two processes which are beyond the scope of this chapter.
Protein Structure Prediction
Prediction of Protein 3D Structure

- Number of known protein sequences is much higher than number of known protein 3D structures
- Knowledge of 3D structure is necessary for understanding chemical and biological function of the protein
- The prediction of the 3D structure of a protein from sequence data is a challenge for current bioinformatics research
- Although reliable method for 3D protein structure prediction still has not been developed, few approaches are used with some success
- Currently used methods can be divided into the following groups:
  - Homology modeling
  - Fold recognition methods
    - Profile-based methods
    - Threading Methods
  - Ab initio protein modeling methods
Homology Modeling

- Homology modeling methods (also called comparative modeling) predict 3D structure of a protein based on knowledge of 3D structure of other homologous proteins.
- Homologous proteins are proteins that are derived from a common "ancestor". These proteins have similar 3D structure and more or less similar sequence. Homologous proteins can usually be detected based on similarity of their sequences.
- 3D structure of a protein is much more conserved than sequence during evolution.
- If sequence similarity between proteins is detected, structural similarity can be assumed.
- Homology modeling uses 3D structure of other homologous protein as a template for building 3D model of target protein for which only sequence is known.
Homology Modeling

1. Identify Related Structures (Templates)
2. Select Templates
3. Align Target Sequence with Template Structures
4. Build a Model for the Target (Using information from template structures)
5. Evaluate the Model
   - Model OK? (YES) → END
   - Model OK? (NO) → Return to Select Templates
Homology Modeling

- Homology modeling procedure consists of the following steps:
  1. **Template recognition** – template protein is searched among the same proteins from different species or proteins with a similar function or from the same metabolic pathway or fast approximate search for many sequences is performed.
  2. **Alignment** – the alignment algorithms scores how well pair match: high scores for identical pairs, medium score for similar pairs (i.e. both hydrophobic), low score for dissimilar pairs.
  3. **Backbone generation** – structurally conserved regions of a protein (α-helices, β-sheets, active sites of enzymes, etc.) are placed in space with an orientation and conformation best matching those of a template. This often results in having sections of the model compound that are unconnected.
  4. **Modeling loops** – the generation of loops is necessary to connect the sections. Loops can be determined by searching database of similar loops or conformation search methods are used.
  5. **Side-chain modeling** – conformation search techniques are usually used.
  6. **Model optimization** – molecular mechanics energy minimization is done or molecular dynamics or simulated annealing are used.
  7. **Model verification** – e.g. hydrophobic residues should point inward.
Accuracy of Homology Modeling

- Accuracy of the result depends on sequence similarity:
  - >90% similarity, homology model can match crystallography within the experimental error, with the biggest difference in side chain rotations
  - 75 - 90% similarity can still result in very good result. The model optimization is the most critical point.
  - 50 – 75% similarity can be expected to result in a RMS error of about 1.5 Å, with large errors in some sections of the molecule
  - 25 - 50% similarity can give adequate results with manual intervention. These cases are limited by the quality of alignment.
  - <25% similarity – homology modeling is unreliable
Advantages and Limitations of Homology Modeling

- Homology modeling is the most used method for predicting the structure of proteins because quality of models is more accurate that those produced with different techniques
- Homology techniques are computationally less demanding than alternative methods
- Limitation of the method is that it requires a known structure with a similar sequence
- With growing number of known protein 3D structures, the increased use of homology modeling methods can be expected in the future
**Fold-recognition methods**

- Fold-recognition methods can be used to determine 3D structure of proteins for which no homologous protein of known structure is available.
- Analyses of large datasets of protein structures showed that proteins that do not share evolutionary relationship can have similar structure – they share a similar fold.
- Fold-recognition methods try to find folds that are compatible with the target sequence irrespective of the existence or detectability of an evolutionary relationship.
- The problem is how to evaluate the fitness function of a sequence and a structure. Usually, this can be done using two alternative approaches: profile based method and sequence threading.
Profile-based methods

- Profile-based methods for fold-recognition are based on idea that the physico-chemical properties of the amino acids of the target protein must 'fit' with the environment in which they are placed in the modelled structure.
- Each amino acid of the target sequence is coded according to its properties, for example secondary structure propensity, hydrophobicity, average solvent accessibility in protein structures.
- The next step is to analyse selected set of known protein structure and assign to each position in the structure a symbol coding its environment (secondary structure, number of hydrophobic contacts with other residues, exposure to solvent).
- Subsequently, the target 'property coded' sequence is compared with 'property coded' sequences of these known proteins.
- The comparison algorithm are based on similar principles as amino acid sequence alignment algorithms (but property matching criteria is used instead of amino acid matching).
- Structure of known protein which has the best similarity score is used as a template to model target structure in a similar way as in homology modeling methods.
Threading Methods

- **Threading methods** try to 'thread' (i.e. place, align) each amino acid of the target sequence to a position in the template structure, and evaluate its fitness with the proposed template structure.

- The **fitness** is usually evaluated based on a **solvation potential** and **residue pair potentials** (atomic interactions between residues are attributed to a single point within each residue), i.e. it considers the full 3-D structure of the protein template.

- The procedure must be repeated for each of the template candidates.

- After the best-fit template is selected, the structural model of the sequence is built based on the alignment with the chosen template (similarly as in homology modeling methods).
Ab Initio Methods

- **Ab initio methods** (also called *de novo* methods) seek to build 3D protein models from sequence, based on physical principles without using any knowledge of previously solved structures
- The methods are based on computing free energy of different conformation with the goal to find the conformation with the lowest energy
- The problem can be divided into two sub-problems - evaluation of the free energy of a given conformation and the search strategy for finding all possible conformations
- The energy is usually evaluated using methods developed for molecular mechanics. Solvent effects can be included by direct inclusion of water molecules (explicit solvent) or using approximations (implicit solvent)
- Because atom based energy evaluation is often too slow for large structures such as proteins, **residue-based strategy** is often used where all atomic interactions between residues are attributed to a single point within each residue
Ab Initio Methods

- Conformational space of a protein is searched using molecular dynamics, Monte Carlo or genetic algorithms. Also specific modifications of these algorithms has been developed.
- Unfortunately, sufficient sampling of large conformational space of proteins is often not achievable by these searching methods.
- Among many programs for *ab initio* protein structure prediction, the most known is Rosetta@home that uses idea of distributed computing to get maximal computational resources for better sampling of conformational space.
CASP Experiment

- **CASP** (Critical Assessment of Techniques for Protein Structure Prediction) is a community-wide experiment (or competition) for testing techniques of protein structure prediction from its sequence.
- It takes place every two years since 1994.
- The experiment is conducted in a double-blind fashion: Neither predictors nor the organizers know the structures of the target proteins at the time when predictions are made. Targets are chosen from among those proteins whose structures are expected to be solved shortly.
- Evaluation of the results is carried out in several categories from which the tertiary structure prediction is the most important and it is subdivided into:
  - homology modeling
  - fold recognition
  - *Ab initio (de novo)* structure prediction
CASP Experiment

There are two types of participants in the CASP experiment

- **Expert teams** using methods that require user intervention – they have a period of several weeks to complete their model
- **Automatic servers** are given 48 hour

Evaluation of results is conducted by expert assessors in the field and conclusions are published in the journal *Proteins*

Web site of CASP experiment: http://predictioncenter.org/