Methods of ab initio prediction of protein structure

Computational aspects of molecular structure

Lecture 7
Teresa Przytycka, Ph.D
Ab initio prediction of protein structures

Ab initio = prediction of protein structure without using homolog (or analog) structure

Why ab-initio?
• To predict fold of proteins that has no homolog of known structure
• To understand the folding process
Models that have attempted to describe protein folding

• Assumption: Protein in its native conformation assumes lowest free energy minimum

• Models
  – hierarchical process where local secondary structures form first, super-secondary structures later and so on until the entire polypeptide folds.
  – spontaneous collapse of the polypeptide into a compact state. This collapsed state is known as a molten globule.
  – Combination of the two views
Ab initio protein fold prediction methods are usually formulated as optimization processes

Distinguished by

• by the representation of a protein and its degrees of freedom
  – Restrict or not the possible position of the Ca atoms points on some grid.
  – Level of details in representing side chains (from none to full)
• the function that defines the energy for each of the allowed conformations
  – atom-based potentials from molecular mechanics packages such as CHARMM, AMBER
  – statistical potentials of mean force derived from many known protein structures (see one of the coming classes)
• the optimization method that attempts to find the global minimum on a given energy surface
  – Molecular dynamic simulation
  – Statistical Monte Carlo approach
  – Genetic algorithms
  – Statistical assembly of predefined protein fragments
Molecular Dynamic

• Model folding process in small time steps using Newton's laws

• **Problems:** Since integration of Newton's laws only gives reliable results at time steps of 1 femtosecond, the simulation of protein folding is not within reach of present day computers.

• Can be used for structure refinement

• Folding at home project –distributed simulation protein folding
  
  http://folding.stanford.edu/
Monte Carlo Methods

- Randomly sample the space of all possible folds
- Problem: Levinthal paradox: There is astronomically large number of possible conformations for each protein and the fact that proteins fold rapidly suggests that protein does not explore whole availed space (thus the idea of folding pathways)
- In Monte Carlo search can be seen as a random walk in the protein fold space.
Directed graphs and random walks

\[ G = (V, E) \]

\( V \)- set of vertices

\( E \subseteq V \times V \)

\[ E = \{(v,w) | v, w \in V\} \]

(edges are ordered pairs of vertices)

Weighted graph – each edge has associated with it an edge
Random walks on graphs

Given is a weighted directed graph such that for each vertex the sum of the weights over all outgoing edges is equal to one.

A random walk is a stochastic process which progresses in steps. Starting from some vertex, each step consists in moving to another vertex with probability equal to the weight of the edge leading to this vertex.

Markov process – a random walk in such graph
Markov property: the probability of next step depends on current stage and not on the history.

Absorbing state
Random walks and matrix multiplication

- $P(i,j)$ = probability of moving from $i$ to $j$ in one step.

- This is a stochastic matrix which means that for each row the sum of the entries equals 1.

$$P = \begin{bmatrix}
0 & .8 & 0 & 0 & .2 \\
0 & .4 & 0 & .3 & .3 \\
.5 & 0 & 0 & .5 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & .2 & 0 & .8 \\
\end{bmatrix}$$
Probability distribution after one random step

- Let \( v \) be the vector such that \( v[i] \) equals probability of being in state \( i \) before taking the step.
- Let \( w \) be the vector such that \( w[i] \) is the probability of being in state \( i \) after one step.
- Observation:

\[
    w = v P
\]

e.g. the probability \( w[1] \) of ending in state assuming initial probability distribution \( v \) is \( v[1]*P[1,1] + v[2]*P[2,1] + v[3]*P[3,1] + ... v[5]*P[5,1] \)
\( P^2 (i,j) \) is the probability of getting from it to \( j \) in two steps:

\[
P^2 (i,j) = \sum_{k=1..n} P(i,k)P(k,j)
\]

\( P(i,k)P(k,j) \) = probability of going from \( i \) to \( k \) in the first step and from \( k \) to \( j \) in the second

Example: Probability of getting from 1 to 5 in 2 steps:

\[
\begin{array}{cccc}
0 & .8 & 0 & .2 & 0 \\
0 & .4 & 0 & .3 & .3 \\
.5 & 0 & 0 & .5 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & .2 & 0 & .8 \\
\end{array}
\]

\( 0 + .8 \times .3 + 0 + .2 \times 0 + 0 \times 0.8 = .24 \)
Stationary (equilibrium) distribution

$P - n \times n$ stochastic matrix

**Stationary distribution** is a probability distribution vector $w^*$ such that

$$w^* = w^* P$$

We assume that folded state is an equilibrium state where the native conformations has overwhelming probability.
Stationary Matrix

\[ w = v \, P^n \]

- \( v \) – initial probability distribution
- \( w \) – probability distribution after \( n \) steps

as \( n \) goes to infinity (assuming that \( P \) has properties given below) \( w \) approaches stationary distribution and \( P^n \) approaches stochastic matrix called stationary matrix.

**PROPERTIES:**
- **ergodicity** – there is a path from any state to any other state
- **acycliclity** – for any state it does not hold that it can be visited in steps \( p, 2p, 3p, 4p, \ldots \) (never a problem if a graph has a self-loop for each vertex)
Relevance of the convergence to equilibrium

\[ w = v \, P^n \]

\( V \) – any initial probability distribution

As \( n \) goes to infinity, the probability of ending at any given state does not depend on starting point.

Consequently: If we take a random walk from any point and iterate it long enough then the probability of being in any conformation after sufficiently many steps does not depend on the starting point but only on the vector \( w \).

Here: We started with \( P \) and if we iterate computation of \( P^n \) long enough we get equilibrium distribution vector \( w \).

In protein folding problem, we know theoretically (see next slide) the target equilibrium distribution and we attempt to correctly design matrix \( P \) so that sufficiently long random walk would leave us in the native (free energy minimum) state.
Equilibrium distribution in protein folding problem

S – protein conformation; P(S) – prob. of the conformation
E – energy

\[ P(S) = \frac{e^{-E(S)/kT}}{Z} \]

Z – partition function = sum over all states S of terms \( e^{-E(S)/kT} \)

Problem: Z cannot be computed.
Theorem: \( w(i) \) is the target equilibrium distribution if matrix \( P \) is such that Markov chain that satisfies the detailed balance principle: \( P(i,j)w(i) = P(j,i)w(j) \)

**Typical Metropolis setting:**
States have equal number of neighbor states (the degrees of nodes in the graph are equal)

In each random move from node \( i \):
1. Neighbor \( j \) is selected uniformly at random and \( w(j) \) is computed.
Let \( b \) be the probability of selection (it is equal for any node)
2. Move to selected neighbor with probability

\[
\min\left(1, e^{-\Delta w/kT}\right)
\]
P(i,j) defined in the Metropolis algorithm satisfies detailed balance principle

Metropolis setting for transition probability

\[ P(i,j) = b \min \left( 1, e^{-\Delta E/kT} \right) \]

Our target equilibrium distribution:

\[ w(i) = e^{-E(i)/kT} \]

We should confirm detailed balance:

\[ P(i,j)w(i) = P(j,i)w(j) \]

Case 1: E(i) < E(j) then E(j) – E(i) is positive thus -\(\Delta E/kT\) is negative and \(e^{-\Delta E/kT} < 1\)

Thus we have

\[ P(i,j) = b e^{-\Delta E/kT}. \]

However in this case \(P(i,i) = b\)

\[ \frac{P(i,j)}{P(j,i)} = b e^{-(E(j)-E(i))/kT} \quad / \quad b = e^{-E(j)/kT} / e^{-E(i)/kT} = w(i)/w(j) \]
Case $E(i) > E(j)$ is symmetric and in case of equality $\Delta E = 0$ and both probabilities are 1.

Otherwise the move that lowers the energy is accepted with probability one an “uphill” move is accepted with probability depending on how big the energy difference is - that is how non-favorable the move is.
Random edge selection:
Select randomly an number 0,1,2 and assume that lowest number corresponds to the lowest edge label highest to highest.

1. Check detailed balance principle between yellow nodes.
2. Perform several moves in this network
   a) Your random selection of integers in 0-2:
   b) Your random selection of real numbers in 0-1

Select. prob $b = 1/3$
Transition prob =
$$\min\left(1, e^{-\Delta S}\right)$$
we dropped $kT$ term for simplicity. That why $E$ (energy) is changed to $S$ (score)
Monte Carlo, Metropolis Algorithm

• At each state $i$ choose uniformly at random one of neighboring conformations $j$.

• Compute $p(i,j) = \min(1, e^{-\Delta E/kT})$

• With probability $p(i,j)$ move to state $j$.

THE FACT THAT EACH STATE HAS THE SAME NUMBER OF NEIGHBOURS IS IMPORTANT (otherwise one has to use a variant of the algorithm called Metropolis-Hastings which deals with non-uniform selection)
In theory

Equilibrium (i.e. at stationary distribution) is equal to the Boltzman distribution as required.

In practice:

• The process may be unable to reach the equilibrium state in reasonable time. It is likely to be trapped in a local minimum.
• Various techniques are used to negotiate the barriers (e.g. replica exchange).
Simulated annealing-idea

- In 1983 Kirkpatrick et al. introduced a temperature cooling schedule:
- Recall the probability of accepting a move:
  \[
  P(i,j) = \min\left(1, e^{-\Delta E/kT}\right)
  \]
- Note that the probability grows with temperature: In higher temperature a move “uphill” is more likely thus we can explore the energy landscape easier. Once we get to what we hope is a “valley” that contains a global minimum we start to reduce temperature.
Exercise

\[ P(i,j) = \min \left( 1, e^{\frac{-\Delta E}{kT}} \right) \]

• Assume \( \exp \left\{ (E_j - E_i)/kT \right\} = 16 \)
• What is \( p(i,j) \)?
• Assume that \( T \) has be doubled. What is new \( p(i,j) \)?
Simulated annealing - practical algorithm

T = T_high
i = initial state
while (T > T_min) repeat M times:
    choose j as a random neighbor of i
    {if E(j) <= E(i) then i = j
     else x = random(0,1)
     if x < e^{-[E(j)-E(i)]/T} then i = j (accept) }
    T = 0.8 * T
}
return i and E(i)
Replica exchange

• Perform several independent simulations in so called replicas
• Replicas have same starting point, different temperatures and are simulated independently
• At certain points of simulations replicas are exchanged
Replica exchange cont.

• To keep the equilibrium of the system it is necessary that detained balance condition is satisfied
• Let $\beta_A = (k T_A)^{-1}$; $\Delta = (\beta_A - \beta_B)(E_B - E_A)$
• The exchange probability is
  1 if $\Delta < 0$
  $\exp(\Delta)$ otherwise
Replica exchange Exercise.

Let $\beta_A = (k T_A)^{-1}$; $\Delta = (\beta_A - \beta_B)(E_B - E_A)$

The exchange probability is

$1$ if $\Delta < 0$

$\exp(-\Delta)$ otherwise

Let $\beta_A = (k T_A)^{-1} = 1/100$

$\beta_B = (k T_A)^{-1} = 1/200$

$E_B = 100$

$E_A = 50$

Compute exchange probability .

Repeat computation with values of $E_A$ and $E_B$ switched. When exchange is easy?
Key points in setting a Metropolis algorithm

• Decide what is the Energy function
• Decide what is your move set (that is which nodes in your graph are connected)
• Decide when to stop (or max iterations in one run of simulated annealing)
Lattice model – simplification of search space

- Residues are represented using uniform sizes
- Bond lengths are uniform
- Every conformation is a self-avoiding walk in $\mathbb{Z}^3$.
- Energy function is associated with each contact:
HP-model (Lau & Dill, 1989)

• Each amino acid is classified as one of two types:
  – H-hydrophilic
  – P-hydrophobic

• Contact energy: -1 for H-H contact, 0 otherwise

• Lowest energy conformation = self avoiding walk maximizing the number of H-H contacts.

\[
\text{Energy} = -2
\]
HP-model properties

• Captures hydrophobic “core” concept
• Yields compact structure
• Some degenerate properties – the distance between nodes have to be even to make a contact possible
• Protein folding on 3D lattice in HP-model was formally proven to be NP complete (Berger, Leighton, RECOMB 98; following a number of partial results by others.)
Local move set

- **end**: move end segment in any of 5 possible directions
- **corner** (Figure a)
- **crankshaft** (figure b)
Failures of Inverse Folding and Threading with gapped alignment

Gordon M. Crippen, Proteins, 1996, 168-171

**Question:** Given a conformation, can one determine which sequences fold into this conformation (the reverse folding problem)
Model

- Model 2D grid,
- Two types of residues A,B
- Contact = two residues non-adjacent in the sequence are adjacent in the grid.
- Contact potentials: $e_{AA}, e_{BB}, e_{AB}$ depending on contact type
Method

• Enumerate all sequence of certain length (8, 10, …)
• Define a “native” sequence as a sequence that has a unique minimum conformation and consider only these sequence
• Examine how such native sequences score on non-native structures (structures native for other sequences but not for the sequence in question)
Results

• For sequence length 11 there are
  – 5 513 conformations,
  – 44 native sequences
  – 17 different native conformations (some of 44 sequences share conformations)
  – They found 1 case where a score of a sequence on non native structure was better then the score of the sequence for which the structure was native.

• If gaps are allowed several cases when native alignment of the native sequence on its own native structure scores no better than some alignment of a differently folding sequence.

• Suggests possible problem with reverse protein design.
Towards more realistic lattices

(Kolinski, Skolnik, and coworkers)

- The protein is described as a lattice chain connecting side chain centers of mass (SICHO model) or triple: C_alpha, C_beta, SG) where SG is center of mass of the side group (CABS model).

- These interaction centers (beads) are projected onto an underlying cubic lattice with a lattice spacing of 1.45 Å (alternatively 0.87).

- Number of neighbors 646
TOUCHSTONE II: A New Approach to Ab Initio Protein Structure Prediction

Yang Zhang,* Andrzej Kolinski,*† and Jeffrey Skolnick*
*Center of Excellence in Bioinformatics, University at Buffalo, Buffalo, New York 14203; and †Faculty of Chemistry, Warsaw University, 02-093 Warsaw, Poland
Towards more realistic lattices

(Kolinski, Skolnik)

• Energy terms (Touchstone program):
  – E-generic – bias towards protein like conformation
  – E-specific – sequence specific bias to predicted sec. structure and other predicted fragments, pairwise contact potentials.
  – E-rest

• Replica exchange Monte Carlo (100-150 days of computer time for 50 runs for a protein)
Move set

a

b

c

d

e
LINUS an of-lattice hierarchical approach.

- Srinivasan Rajgopal and George D. Rose. (1995) LINUS - A hierarchic procedure to predict the fold of a protein
- Full atom model
- Monte Carlo simulation
- The move set: select at random tree consecutive amino-acid
- Choose a conformation from one of selected Ramachandran region: helix, strand, other.
- Perform Monte Carlo simulation, during which collect statistics about frequency with which each amino acid is in each of the tree states.
- Energy in this phase - only local interactions
LINUS continued

Second phase:

• Use the statistics from the previous phase to bias your move set: the probability of selecting beta/alpha/other conformation for a given.

• Energy in this phase - local and middle range interaction (middle range = interactions between residues at most constant d (about 12 res) apart.

Subsequent phase: as second phase but all interaction are counted in the energy function.

Energy function: hydrogen bounds, salt bridges, etc (controlled by user).
The meaning of the phases:

- The governing principle is the assumption that protein folding is hierarchic with local conformational preferences emerging first, the secondary structures, subsequently supersecondary structures (hairpins etc) and finally the complete fold.
GO model-extreme simplification of energy function

- Model designed to study the dynamics of folding: only native contact have a non-zero energy contribution.
- This potentially can show which contacts emerge first.
• Construct a data base fragment library:
  – Cluster short PDB fragments according sequence similarity (let’s call these cluster super-clusters). Within each super-cluster identify sub-clusters by their structural similarity.

• Monte Carlo simulation
  – Move set: select a location in the current conformation at random
  – Replace the fragment at this location with a fragment from the same super-cluster. Chose a sub-cluster conformation with probability proportional to the size of the sub-cluster.

• Energy function has terms for compactness, burial hydrophobic residues, beta-sheet formation etc,

• Simulated Annealing
Rosetta, continued

• Cluster all good structures from the simulation by structural similarity
• Choose the representative of the most largest cluster as the predicted structure – the idea is that the structure that emerges most frequently is the correct one.
Rosetta CASP 4 prediction

T087 - PPase (Domain 2: 202-307)

Native

Model 3

RMSD = 6.2 Å (85 Cα)
T106 - secreted frizzled protein 3

RMSD = 6.4 Å (106 Cα)
RMSD = 4.5 Å (75 Cα)
LINUS, Rosetta and detailed balance principle

• Theoretically neither LINUS nor Rosetta satisfies detailed balance principle: the move set is biased either towards secondary structures (LINUS) or frequent special fragments (Rosetta)

• Is this a problem?

• Not really but one need to be aware that bias = additional energy term favoring the biased choice.
Genetic algorithm for Protein Folding

- Unger and Moult 1993 (Journal of Molecular Biology)
- Here we will describe basic idea using 2D lattice HP model
- Subsequently genetic algorithms have been used in more realistic models
Basic Idea of a genetic algorithm

- **Given** is some optimization function (in our case energy minimum)
- **Start** with an initial population
- **Iterate:**
  - Produce of individuals for next generation using
    - Crossover operation
    - Mutation operation
  - Evaluate of fitness of each individual
  - Select survivors for next generation
Mutation Operation

• Mutation occurs with some probability.
• Mutation is a local random change (usually change of a letter in a sequence). Choose a location.
• Here mutation is a single Metropolis step of the corresponding lattice model Monte Carlo algorithm.
Crossover Operation

• General scheme:
  – Choose parents to combine (presumably with some beneficial traits)
  – Pick crossover point
  – Combine parents
  – Produce offspring

• In this application:
  – Choose parent conformation with probability proportional to the absolute value of energy
  – Choose a crossover point
  – Take first part of the confirmations from one parent and the second form the other parent - find a legal (no-overlapping) embedding of the two parts
Genetic Algorithm for folding in HP Model. (2D lattice)

Initialize population of $P(t)$ random coils
best = best feet element form P
repeat  do a local conformation change
     n=0
     while (n < P)
       select two elements m,f
       produce child, c, by crossover;
       if the child's score is better than parents average
         place it in the new generation; n++
       else z = random(0,1)
         if z < (some function of c and parents average)
           place c in the new generation; n++
       update best
     until convergence