Molecular Modeling -- Spr ’14
lecture 17

• Protein folding theory
  • pathways
• Protein structure prediction
  • CASP
  • Rosetta
• Scoring predictions

Please read Anfinsen article
17.1 Protein folding theory
Anfinson’s Thermodynamic Hypothesis of protein folding

“Scrambled” ribonuclease -- reduced and denatured (using urea) ribonuclease is oxidized first, then dialyzed into water. ~1% of the activity remained. There are 105 different ways to make 4 disulfides (=7x5x3x1) so ~1% activity means the disulfides are random.

But after adding a disulfide exchange enzyme (thioredoxin), all of the activity returns. **Thermodynamics alone selects the native conformation.**
Levinthal’s Paradox

Let’s say a polypeptide chain has 100 amino acids and each can have 3 states: [i.e. alpha, beta, or loop]. How many conformational states are possible?

\[ 3 \times 3 \times 3 \times 3 \times 3 \ldots = 3^{100} \]

If a protein can search \(10^8\) conformations per second. How many years will it take to sample all possible states once?**

\[ 10^{50} \text{ states} / 10^8 \text{ states/s} = 10^{42} \text{ s} \approx 10^{37} \text{ d} \approx 10^{34} \text{ y} \]

**Levinthal’s estimate of the folding half-life of a protein is off by about 40 orders of magnitude!
The existence of folding pathways resolves Levinthal’s paradox.

An analogy:
It takes 100 steps in the right direction to get you from here to the front door of the Folsom Library. You can step N, S, E or W. How long will it take you to get to the library? (How long will it take you to get out of the room?)

Now imagine you have a guide (called “Energy”). Each time you take a step the guide says “on path” or “off path”. If you are off path, you go back one step. Now how long will it take to get to the library?
Anfinson reasoned that small chunks of the protein could have a preference toward the native state (lower part of figure), and that assembling these chunks would be possible in the sub-second timescale.

A pathway is a sequence of folding events, such as these, ending in the “native state.”
Short bits of chain can fold by random conformational search. Long stretches cannot.

Whole chain. $3^{100}$ states. Takes $10^{34}$ y to fold by random search.

10-residue segment. $3^{10}$ states. Takes 0.000001 s to fold by random search.

If short segments of the protein fold independently, then they also fold fast. Folded segments reduce the total number of states of the polypeptide chain.
A *folding pathway* is a defined sequence of folding events, each stepping downhill on the energy landscape.

- Early in folding, local segments fold independently.

- Local segments condense or extend.

- Each condensation or extension adds another short piece of folded chain.

- Equilibrium drives toward the fully folded state, since it is lowest in energy.

**Sequence position** (folded=red, non-folded=black)
Small natively-folded intermediates *should* exist. But do they?

Anfinsen proved they do, in 1972.
Anfinsen (1969) shows that antibodies that bind the native state bind to a fragment.
Small natively-folded segments exist in "Flickering equilibria"

Since 99-149 binds to Ab that also inhibit the native state, 99-149 must be in equilibrium with the native state.
What other evidence is there for early folding units?

1) Secondary structure prediction uses only local sequence, and it works.

2) Some peptides fold in isolation, as shown by presence of NOEs in NMR.

3) Mutations have effects on the folding rate that are different from the effect on the folding stability.

![Energy landscape diagram showing mutation effects on folding rate and folding stability.](image)
17.2 Protein structure prediction
without using a template!
Protein Structure Prediction

...two very different UNDERLYING PRINCIPLES

**Darwin:**
Proteins with a common ancestor have the same fold.

**Boltzmann:**
Proteins adopt a minimum the free energy conformation.

---

query sequence ➔ best alignment

millions of years

query sequence ➔ lowest energy

microseconds to seconds
FOLD RECOGNITION

Sequence-based methods say -- if the sequence, sequence pattern, features, environments, contacts or any contrived 1D variable is similar, then the structure is similar.

environment, backbone angles, surface area burial, contacts.
Mixed messages: HOMOLOG detection using energy calculations

Target is aligned to each known protein template, all possible ways ("threading"). For each alignment, calculate the energy. The best energy wins.

WPSGTECIAKYNHGTAEQDLPFCKGDVLTVAVTKDPNWYKAKNKVREGIIP

(lowest wins)

What is the unspoken assumption in this method?
... That a protein folds by sampling every possible structure???
Homolog detection using a library of profile HMMs

1. MYSEQUENCE

2. Pick the model with the max P

Constructing a protein from fragments

For representative proteins, backbones can be assembled from a library of 1000 different 5-residue fragments. All proteins can be constructed to within RMSD $\leq 2.7\text{Å}$. 

Folding initiation site motifs

Find all short, similar sequences that also fold to similar structures independent of context. Do they fold early? (Bystroff & Baker, JMB, 1998)

Non-homologous sequences

<table>
<thead>
<tr>
<th>Recurrent sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDFPIEGGDSPMQTIFFFWSNANAKLISHGY</td>
</tr>
<tr>
<td>CPYDNIWMQTIFFFNQSAAVYSVLHLIFT</td>
</tr>
<tr>
<td>IDMNPQGSIEMQTIFFGYAESA</td>
</tr>
<tr>
<td>ELSPVVFLEEMQTIFFFISGFTQTANSID</td>
</tr>
<tr>
<td>INWGSMQTIFFEWQLMNVMKIPS</td>
</tr>
<tr>
<td>IFNESKKKGIAMQTIFFILSGR</td>
</tr>
<tr>
<td>PPPMQTIFFFVIVYNESKHALWCSVD</td>
</tr>
<tr>
<td>PWMWNLMQTIFFISQQVIEIPS</td>
</tr>
<tr>
<td>MQTIFFFVFSHDEQMKLKLKGLKGA</td>
</tr>
</tbody>
</table>

Is there a recurrent sub-structure here?
I-sites motifs: folding initiation sites

I-sites Library = a catalog of fragment motifs.

**Verified by MD and NMR**

(Yi et al, J Mol Biol. 1998)
Fragment insertion Monte Carlo

Randomly choose a fragment

...change backbone angles

Accept or reject changes

Evaluate energy

Convert to 3D

Reference:
Simons et al, PNAS. 1999
Knowledge-based potentials

The energy of an interaction can be calculated from the database if enough examples of that interaction exist in the database.

\[ \text{Energy} = -RT \ln \left( \frac{p}{1 - p} \right), \]

where \( p \) is the probability of the interaction

- For example, \( E \) and \( K \) interact. If the interaction is strong, then the number of \( E \) bound to \( K \) relative to the number of \( E \) is greater. If the interaction is weak, the number of \( E \) bound to \( K \) relative to the number of \( E \) is lower.

\[ E + K \rightleftharpoons E \cdot K \]

- Knowledge-based potentials assume that the data are representative of the interaction frequency. Meaning, increasing the database base size will have no effect.
Energy = -RT ln( p/(1 - p)), where p is the probability of the structure

model structure vector representation as (phi, theta, sigma, Sep, r)
Topologically correct (rmsd=5.9Å) but loop is mis-predicted as helix.
CASP prediction

T0122 57-153 (97 residues)

...contains a 53 residue stretch with max deviation = 96°
Successful blind predictions using YASARA knowledge-based potential:

http://www.yasara.org/kbpotentials.htm
Ways of expressing prediction accuracy in CASP

Each line is one model.
x-axis = % of model
y-axis = best RMSD for that %

Quality of alignment
green=%correct
yellow=% off by one
some protein folding talking points

- Proteins can’t possible sample all conformational states (**Levinthal’s paradox**)
- Proteins fold along a **pathway**, from **local to non-local**.
- Anfinsen proved **natively-folded substructures** exist along the pathway.
- Template-based prediction depends on the (. .very good!..) assumption that **similar sequences have similar structures**.
- Non-template based prediction depends on the assumption that the **true structure is the one with the lowest energy**.
- Homology based (**Darwin**) predictions tend to be more successful than energy based (**Boltzmann**) ones. (why?)
- Knowledge-based free energy = -RT log( Probability of interaction )
- Knowledge-based energy functions assume that the database is a **Boltzmann distribution**. And that it is **representative**.
- All proteins are made up of **recurrent short motifs**.
http://fold.it/portal/
Excellent blind prediction in CASP9 by Fold-It gamers!
Homework 9 - an essay

• Develop a winning strategy for Fold-IT.
• The puzzle: Fold a protein from scratch.
• Save as PDF, Upload by midnight April 4
• Grade = (content + clarity)/length