Protein Folding
Protein Structure Prediction
Protein Design

Brian Kuhlman
Department of Biochemistry and Biophysics
Protein Folding

- The process by which a protein goes from being an unfolded polymer with no activity to a uniquely structured and active protein.

Why do we care about protein folding?

- If we understand how proteins fold, maybe it will help us predict their three-dimensional structure from sequence information alone.

- Protein misfolding has been implicated in many human diseases (Alzheimer's, Parkinson's, …)
Protein folding *in vitro* is often reversible (indicating that the final folded structure is determined by its amino acid sequence)
How Do Proteins Fold?
Do proteins fold by performing an exhaustive search of conformational space?

- Cyrus Levinthal tried to estimate how long it would take a protein to do a random search of conformational space for the native fold.

- Imagine a 100-residue protein with three possible conformations per residue. Thus, the number of possible folds = $3^{100} = 5 \times 10^{47}$.

- Let us assume that protein can explore new conformations at the same rate that bonds can reorient ($10^{13}$ structures/second).

- Thus, the time to explore all of conformational space = $5 \times 10^{47}/10^{13} = 5 \times 10^{34}$ seconds = $1.6 \times 10^{27}$ years >> age of universe

- This is known as the Levinthal paradox.
How do proteins fold?
Do proteins fold by a very discrete pathway?

Flat landscape (Levinthal paradox)  Tunnel landscape (discrete pathways)  Realistic landscape ("folding funnel")
How do proteins fold?

Typically, proteins fold by progressive formation of native-like structures.

Folding energy surface is highly connected with many different routes to final folded state.
How do proteins fold?

Interactions between residues close to each other along the polypeptide chain are more likely to form early in folding.
Protein Folding Rates Correlate with Contact Order

$$Abs\_CO = \frac{1}{N} \sum_{i=1}^{N} \Delta L_{ij}$$

$N$ = number of contacts in the protein

$\Delta L_{ij}$ = sequence separation between contacting residues
Protein misfolding: the various states a protein can adopt.
Molecular Chaperones

• Nature has developed a diverse set of proteins (chaperones) to help other proteins fold.

• Over 20 different types of chaperones have been identified. Many of these are produced in greater numbers during times of cellular stress.
Example: The GroEL(Hsp60) family

- GroEL proteins provide a protected environment for other proteins to fold.

Binding of U occurs by interaction with hydrophobic residues in the core of GroEL. Subsequent binding of GroES and ATP releases the protein into an enclosed cage for folding.
Hsp60 Proteins

The Chaperonin - GroEL
Protein misfolding: the various states a protein can adopt.
Amyloid fibrils

- rich in $\beta$ strands (even if wild type protein was helical)
- forms by a nucleation process, fibrils can be used to seed other fibrils
- generally composed of a single protein (sometimes a mutant protein and sometimes the wildtype sequence)
Amyloid fibrils implicated in several diseases

- Amyloid fibrils have been observed in patients with Alzheimer's disease, type II diabetes, Creutzfeldt-Jakob disease (human form of Mad Cow's disease), and many more.

- In some cases it is not clear if the fibrils are the result of the disease or the cause.

- Fibrils can form dense plaques which physically disrupt tissue.

- The formation of fibrils depletes the soluble concentration of the protein.
### Folding Diseases: Amyloid Formation

**Table 1** Standardized nomenclature for amyloid and amyloidosis

<table>
<thead>
<tr>
<th>Amyloid protein&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Protein precursor</th>
<th>Protein type of variant</th>
<th>Clinical</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA&lt;sup&gt;b&lt;/sup&gt;</td>
<td>SAA</td>
<td>Aκ, Aλ, (e.g. AκIII)</td>
<td>Reactive (secondary), familial Mediterranean fever, familial amyloid nephropathy with urticaria and deafness (Muckle–Wells syndrome)</td>
</tr>
<tr>
<td>AL</td>
<td>Kappa, lambda (e.g. κIII)</td>
<td>Aγ1</td>
<td>Idiopathic (primary), myeloma-associated or macroglobulinaemia-associated</td>
</tr>
<tr>
<td>AH</td>
<td>IgG 1 (γ1)</td>
<td>Aγ1</td>
<td>Familial amyloid polyneuropathy (Portuguese)</td>
</tr>
<tr>
<td>ATTR</td>
<td>Transthyretin</td>
<td>e.g. Met30&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Familial amyloid cardiomyopathy (Danish), systemic senile amyloidosis</td>
</tr>
<tr>
<td>AapoAl</td>
<td>apoAl</td>
<td>Arg26</td>
<td>Familial amyloid polyneuropathy (Iowa)</td>
</tr>
<tr>
<td>AGel</td>
<td>Gelsolin</td>
<td>Asn187&lt;sup&gt;d&lt;/sup&gt;(15)</td>
<td>Familial amyloidosis (Finnish)</td>
</tr>
<tr>
<td>ACys</td>
<td>Cystatin C</td>
<td>Gln68</td>
<td>Hereditary cerebral haemorrhage with amyloidosis (Icelandic)</td>
</tr>
<tr>
<td>AFib</td>
<td>Fibrinogen Aα chain</td>
<td>e.g. Leu554</td>
<td>Hereditary renal amyloidosis</td>
</tr>
<tr>
<td>ALys</td>
<td>Lysozyme</td>
<td>e.g. His</td>
<td>Nonneuropathic hereditary amyloidosis</td>
</tr>
<tr>
<td>Aβ</td>
<td>β protein precursor (e.g. βPP&lt;sub&gt;60&lt;/sub&gt;) Gln693(22)</td>
<td></td>
<td>Alzheimer disease, Down syndrome, hereditary cerebral haemorrhage amyloidosis (Dutch)</td>
</tr>
<tr>
<td>Aβ&lt;sub&gt;2&lt;/sub&gt;M</td>
<td>β&lt;sub&gt;2&lt;/sub&gt;-microglobulin</td>
<td>Pr&lt;sub&gt;p&lt;/sub&gt;&lt;sup&gt;Sc&lt;/sup&gt;, Pr&lt;sub&gt;p&lt;/sub&gt;&lt;sup&gt;CJD&lt;/sup&gt;</td>
<td>Associated with chronic dialysis</td>
</tr>
<tr>
<td>AprP</td>
<td>Pr&lt;sub&gt;p&lt;/sub&gt;&lt;sup&gt;c&lt;/sup&gt;-cellular prion protein</td>
<td>e.g. P102L, A117V, F198S, Q217R</td>
<td>Scrapie, Creutzfeldt–Jakob disease, kuru</td>
</tr>
<tr>
<td>ACa1</td>
<td>(Pro)calcitonin</td>
<td>(Pro)calcitonin</td>
<td>Gerstmann–Sträussler–Scheinker syndrome</td>
</tr>
<tr>
<td>AANF (atrial natriuretic factor)</td>
<td></td>
<td></td>
<td>Medullary carcinoma of the thyroid</td>
</tr>
<tr>
<td>AIAPP (islet amyloid polypeptide)</td>
<td></td>
<td></td>
<td>Isolated atrial amyloid</td>
</tr>
<tr>
<td></td>
<td>(Pro)calcitonin</td>
<td></td>
<td>Islets of Langerhans, diabetes type II, insulinoma</td>
</tr>
</tbody>
</table>
Misfolded proteins can be infectious (Mad Cow’s Disease, Prion proteins)

Stanely Prusiner: 1997 Nobel Prize in Medicine
Structure Prediction

DEIVKMSPIIRFYSSGNAGLRTYIGDHKSCVMCTYWQNNLTYESGILLPQRSRTSR
Prediction Strategies

Homology Modeling
- Proteins that share similar sequences share similar folds.
- Use known structures as the starting point for model building.
- Can not be used to predict structure of new folds.

De Novo Structure Prediction
- Do not rely on global similarity with proteins of known structure
- Folds the protein from the unfolded state.
- Very difficult problem, search space is gigantic
Protein Structure Prediction: Targets and Methods

- **Target Type:**
  - Comparative Modeling
  - Fold Recognition
  - BLAST
  - PSIBLAST
  - Homologous Analogous
  - New Fold

- **Method of choice:**
  - Homology-Based Modeling
  - De novo prediction

- **Limiting problem:**
  - Side chains, loops, insertions, distortions
  - detection, alignment
  - topology, architecture

- Similar structure(s) exist
  - Sequence detectable
  - Not sequence detectable

- No significant structural matches
De Novo Structure Prediction
Fragment-based Methods (Rosetta)

- Hypothesis, the PDB database contains all the possible conformations that a short region of a protein chain might adopt.
- How do we choose fragments that are most likely to correctly represent the query sequence?
Fragment-based Methods (Rosetta)

- Hypothesis, the PDB database contains all the possible conformations that a short region of a protein chain might adopt.

- How do we choose fragments that are most likely to correctly represent the query sequence?
A unique library of fragments is generated for each 9-residue window in the query sequence.

Assume that the distributions of conformations in each window reflects conformations this segment would actually sample.

Regions with very strong local preferences will not have a lot of diversity in the library. Regions with weak local preferences will have more diversity in the library.
Monte Carlo-based Fragment Assembly

- start with an elongated chain
- make a random fragment insertion
- accept moves which pass the metropolis criterion (random number \(< \exp(-\Delta U/RT)\))
- to converge to low energy solutions decrease the temperature during the simulation (simulated annealing)
movie
Multiple Independent Simulations

- Any single search is rapidly quenched
- Carry out multiple independent simulations from multiple starting points.
Fragments are only going to optimize local interactions. How do we favor non-local protein-like structures?

- An energy function for structure prediction should favor:
Fragments are only going to optimize local interactions. How do we favor non-local protein-like structures?

• An energy function for structure prediction should favor:
  - Buried hydrophobics and solvent exposed polars
  - Compact structures, but not overlapped atoms
  - Favorable arrangement of secondary structures. Beta strand pairing, beta sheet twist, right handed beta-alpha-beta motifs, ...
  - Favorable electrostatics, hydrogen bonding

• For the early parts of the simulation we may want a smoother energy function that allows for better sampling.
Protein Design
Protein Design

• A rigorous test of our understanding of protein stability and folding

• Applications

  1. increase protein stability
  2. increase protein solubility
  3. enhance protein binding affinities
  4. alter protein-protein binding specificities (new tools to probe cell biology)
  5. build small molecule binding sites into proteins (biosensors, enzymes)
Central Problem: Identifying amino acids that are compatible with a target structure.

To solve this problem we will need:

- A protocol for searching sequence space
- An energy function for ranking the fitness of a particular sequence for the target structure
Rosetta Energy Function

1) Lennard-Jones Potential (favors atoms close, but not too close)

2) implicit solvation model (penalizes buried polar atoms)

3) hydrogen bonding (allows buried polar atoms)

4) electrostatics (derived from the probability of two charged amino acids being near each other in the PDB)

5) PDB derived torsion potentials

6) Unfolded state energy
Search Procedure – Scanning Through Sequence Space

Monte Carlo optimization

• start with a random sequence
• make a single amino acid replacement or rotamer substitution
• accept change if it lowers the energy
• if it raises the energy accept at some small probability determined by a boltzmann factor
• repeat many times (~ 2 million for a 100 residue protein)
Search Procedure

start with a random sequence
Search Procedure

try a new Trp rotamer
Search Procedure

Trp to Val
Search Procedure

Leu to Arg
Search Procedure
Search Procedure

final optimized sequence
Biosensor Design

- Specificity of ligand binding sites redesigned in periplasmic binding proteins
- Binding-linked conformational change (pre-existing) monitored by fluorescence.

Designing a Completely New Backbone

1. draw a schematic of the protein
2. Identify constraints that specify the fold (arrows)
3. Assign a secondary structure type to each residue (s = strand, t = turn)
4. Pick backbone fragments from the PDB that have the desired secondary structure
5. Assemble 3-dimensional structure by combining fragments in a way that satisfies the constraints (Rosetta).
An Example of a Starting Structure
Design Model and Crystal Structure of Top7