Protein folding

Native
Predicted

Chignolin (crn025) 1.0 Å
Trp-cage (2qof) 1.4 Å
BBA (1fme) 1.6 Å
Villin (24k) 1.3 Å

WW domain (2121) 1.2 Å
NTL9 (2hba) 0.5 Å
BBL (2wxc) 4.8 Å
Protein B (1prb) 3.3 Å

Homeodomain (2p6) 3.6 Å
Protein G (1mio) 1.2 Å
α3D (2a3d) 3.1 Å
λ-repressor (1mb) 1.8 Å
Basic protein structure

Amino acid (Phenylalanine)

Polypeptide

‘Surface’ representation

‘Cartoon’ representation

beta-sheet

alpha-helix
What happens if proteins misfold?

Mad Cow disease

– In a bovine epidemic that struck the UK starting in 1986, cows appeared to be mad: they drooled and staggered, were extremely nervous, or bizarrely aggressive. They all died. As the brains of the dead “mad” cows resembled a sponge, the disease was called bovine spongiform encephalopathy, or BSE.

Prions are proteins that are found in the nerve cells of all mammals. Many abnormally-shaped prions are found in the brains of BSE-infected cows.

Called (variant) Creutzfeldt–Jakob disease in humans

http://www.uvm.edu/~wschaeff/101PrionsMadCow.html
What happens if proteins misfold?

- Alzheimer's disease (AD)
  - Worsens as it progresses, eventually leading to death
  - The Aβ peptides are believed to be involved in AD: they can oligomerize and be released into the interstitial fluid of brain, where soluble oligomers may diffuse into synaptic clefts and interfere with synaptic function by unknown mechanisms.
  - Multiple drugs failing in Phase III trials (even as of two weeks ago!).

http://www.nature.com/ncb/journal/v6/n11/full/ncb1104-1054.html
What happens if proteins misfold?

- Alzheimer's disease (AD)
  - Worsens as it progresses,
  - The \( \text{A}^{\beta} \) peptides are believed to be involved in AD: they can oligomerize and be released into the interstitial fluid of brain, where soluble oligomers may diffuse into synaptic clefts and interfere with synaptic function by unknown mechanisms.
  - multiple drugs failing in Phase III trials (even as of two weeks ago!).

http://www.nature.com/ncb/journal/v6/n11/full/ncb1104-1054.html

https://www.the-scientist.com/news-opinion/biogen--eisai-end-two-late-stage-trials-for-alzheimers-treatment-66431
Protein Folding: Three Questions

“The protein-folding problem came to be three main questions:

1. The physical folding code: How is the 3D native structure of a protein determined by the physicochemical properties that are encoded in its 1D amino-acid sequence? (Anfisen’s dogma - 1972 Nobel Prize)

2. The folding mechanism: A polypeptide chain has an almost unfathomable number of possible conformations. How can proteins fold so fast?

3. Predicting protein structures using computers: Can we devise a computer algorithm to predict a protein’s native structure from its amino acid sequence?

• The Protein-Folding Problem, 50 Years On. Ken A. Dill and Justin L. MacCallum Science 338, 1042 (2012);
Q1 The Physical Folding Code

Forces Governing Protein Folding:
1. Hydrogen bonds
2. van der Waals interactions
3. Backbone angle preferences
4. Electrostatic interactions.
5. **Hydrophobic interactions.**
6. Chain entropy.
Q2 The Folding Mechanism

• Levinthal’s paradox
  – The conformation of a protein is largely determined by its backbone angles: $\phi$ and $\psi$.
  – If we only focus on $\psi$, and assume that it can adopt one of three possible values. How long does it take for a protein with 101 amino acids to fold?
    • Possible conformations:
      $3^{100} \approx 5 \times 10^{47}$.
Dawkins’ weasel

*Original used to explain evolution by random mutation*

Given enough time, a monkey bashing away at random on a typewriter could produce all the works of Shakespeare.

“Let us limit the task facing our monkey somewhat. Suppose that he has to produce, not the complete works of Shakespeare but just the short sentence 'Methinks it is like a weasel', and we shall make it relatively easy by giving him a typewriter with a restricted keyboard, one with just the 26 (capital) letters, and a space bar. How long will he take to write this one little sentence?”

Richard Dawkins

\[27^{28} = 10^{40}; \text{ 1 letter/second} \rightarrow 10^{32} \text{ years!}\]

Dawkins’ resolved this by allowing for **cumulative selection**.
Solution

• Dawkins’ weasel
  – Restrictions for the monkey:
    • not allowed to change those letters that are already correctly in place. *How does it know what the target is?*

• Levinthal’s paradox
  – Consider amino acid interactions: native-fold interactions maintained as they form  

This solution only offers a mathematical explanation of why protein folding is possible; it doesn’t actually solve the problem.
Protein Folding Funnel

- Protein folding landscapes are narrower at the bottom; there are few low-energy, native-like conformations and many more open unfolded structures.

- A protein folds by taking random steps that are mostly incrementally downhill in energy.

- Different molecules of the same protein sequence may each follow microscopically different routes to the same native structure.

- A protein appears to first develop local structures in the chain (such as helices and turns) followed by growth into more global structures. Even though the folding process is blind, nevertheless it can be fast because native states can be reached by this divide-and-conquer, local-to-global process.

A 1D “funnel” - folding of deca-alanine helix

10-Ala helix (in vacuum)
end-to-end distance (ξ) a common RC

done in vacuum

calculated free energy shows minima in two different states
Lattice models for protein folding

Only permit folding on a lattice (unoccupied sites are solvent)

**HP model:** assigns residues into Hydrophobic (H) or Polar (P) classes,
Assumes hydrophobic collapse dominates folding free energy

Assign an energy penalty for any H-P or H-S contacts

*dark - H; light - P*

---

Figure 8.30 Physical Biology of the Cell, 2ed. (© Garland Science 2013)
Lattice models for protein folding

Only permit folding on a lattice (unoccupied sites are solvent)

**HP model**: assigns residues into Hydrophobic (H) or Polar (P) classes,

*Assumes hydrophobic collapse dominates folding free energy*

Assign an energy penalty for any H-P or H-S contacts

*dark - H; light - P*

**not a good protein, all states have the same energy!**
Lattice models for protein folding

Only permit folding on a lattice (unoccupied sites are solvent)

**HP model**: assigns residues into Hydrophobic (H) or Polar (P) classes, 
Assumes hydrophobic collapse dominates folding free energy

Assign an energy penalty for any H-P or H-S contacts

dark - H; light - P

**HPHPHP**

not a good protein, all states have the same energy!

**PHPPHP**

Here, a lowest energy minimum state exists
Lattice models for protein folding

\[
p_{\text{fold}} = \frac{e^{-2\beta \epsilon}}{e^{-2\beta \epsilon} + 2e^{-4\beta \epsilon}}
\]

Figure 8.31 Physical Biology of the Cell 2ed. © Garland Science 2013
Gō (lattice) model

In its simplest form, developed in 1975, assigns a favorable energy $\varepsilon$ for native contacts and 0 for non-native contacts.

Requires knowledge of final structure, but permits one to examine folding kinetics.

using Monte Carlo simulations, can enumerate different folding pathways.

examination of simulation results allows identification of common intermediate states.

still used in various forms today!!!

Complications to the funnel

true free-energy landscape is much more complex than a single funnel

proteins may sample a number of intermediates without native-like structure on the folding pathway


Complications to the funnel

many disease states on the right-hand side - how to avoid them?

proteins already start folding _during_ synthesis!

protein chaperones prevent aggregation


Complications to the funnel

chaperones such as Trigger Factor

d

many *in vivo* factors alter the folding process, e.g.,

macromolecular crowding


H-X Zhou
http://pubs.acs.org/cen/coverstory/88/8848cover.html?featured=1
Q3 Protein Structure Prediction

• We know up to $1000x$ more sequences than structures.
  – There is considerable value in methods that could accurately predict structures from sequences.

  – Held every second summer, CASP is a community-wide blind competition in which typically more than 100 different “target sequences” (of proteins whose structures are known but not yet publicly available) are made available to a community that numbers more than 100 research groups around the world.
The Progress of CASP

• Currently, all successful structure-prediction algorithms are based on assuming that similar sequences lead to similar structures.

• PDB: 156,000 structures but mostly redundant (4000 structural families and 1400 folds - may be all that exist?).
Remaining Challenges in CASP

• When there is no protein in the PDB with a sequence resembling the target’s, accurately predicting the structure of the target is much more difficult (free modeling, or *ab initio*, *de novo* prediction).

• Substantial improvements have been observed for free-modeling targets shorter than 100 amino acids, although no single group yet consistently produces accurate models.

• Winner in CASP12: Rosetta+MD from David Baker lab (2016)

http://robetta.bakerlab.org/
structure prediction webserver
Remaining Challenges in CASP

- AlphaFold is product of DeepMind, an AI company in the UK acquired by Google in 2014 (also made AlphaGo)

For protein sequences for which no other information was known—43 of the 90—AlphaFold made the most accurate prediction 25 times. That far outpaced the second place finisher, which won three of the 43 tests.

average margin of 15% accuracy improvement over other groups on the toughest 43 tests

AlphaFold algorithm

- deep neural networks predict: (a) the distances between pairs of amino acids and (b) the $φ/ψ$ angles between chemical bonds that connect those amino acids

- trained a generative neural network to invent new fragments, which were used to continually improve the score of the proposed protein structure

- optimized scores through gradient descent (a type of minimization)

https://deepmind.com/blog/article/alphafold
MD Simulation of Protein Folding

• Challenges
  – Timescale
    • Specialized supercomputers
  – Force field
    • Improvement made to classical MD FF:
      – AMBER; CHARMM; GROMOS, etc.
    • New FF with polarizability is increasingly used.

IBM ‘Blue Gene’

D.E. Shaw Research ‘Anton’
MD Simulations of Protein Folding

Using Anton, Shaw and co-workers observed reversible folding and unfolding in more than 400 events across 12 small proteins to structures within 4.5 Å of the experimental structure.

Beyond prediction: designer proteins

Beyond structure, engineer protein for specific functions


http://www.nature.com/nature/journal/vaop/ncurrent/full/nature12443.html