PROTEIN FOLDING: A PARADIGM FOR SOLVING HARD PROBLEMS IN BIOLOGY

Michael Levitt
Structural Biology & Computer Science
Stanford

http://csb.stanford.edu/levitt
OUTLINE

- Simulation
  - Basic methods
  - Hydrophobic effect
  - Unfolding, folding

- Prediction
  - Special potentials
  - Minimization
  - Monte Carlo

- Hard Problems
INTRODUCTION
DNA of bacteria with 2,000 genes ≈ 2,000 proteins

- From DNA sequence, predict all protein structures
- From protein structures, predict all function
PROTEIN FOLDING IS CENTRAL

Sequence ➔ Structure ➔ Function

- Unfolded protein is a chain of amino acids
- Folded protein
- Function depends on protein shape

- Highly mobile
  - Inactive
- Almost unique shape
  - Precisely ordered
  - Stable
  - Active
- Specific associations
  - Precise reactions

©Michael Levitt 01
SIMULATION
TOTAL POTENTIAL ENERGY

\[ U = \sum \frac{1}{2} K_b (b - b_0)^2 + \sum \frac{1}{2} K_\theta (\theta - \theta_0)^2 \]

All Bonds

All Angles

\[ + \sum K_\phi \left[ 1 - \cos(n\phi + \delta) \right] \]

All Torsion Angles

\[ + \sum \varepsilon \left[ \left( \frac{r}{r_0} \right)^{12} - 2 \left( \frac{r}{r_0} \right)^6 \right] \]

All nonbonded pairs

\[ + \sum \frac{332 q_i q_j}{r} \]

All partial charges

ENCAD Parameters from 1979 (Lifson)
TOTAL POTENTIAL ENERGY. 2

- The total potential energy or enthalpy fully defines the system, \( U \).
- The forces are the gradients of the energy.
- The energy is a sum of independent terms for: Bonds, Bond angles, Torsion angles and non-bonded atom pairs.

\[ F(x) = -\frac{dU}{dx} \]
**MOLECULAR DYNAMICS THEORY**

- Force = -dU/dx (slope of potential, U); acceleration, \( m \cdot a(t) = \text{Force} \).
- All atoms move together so force between atoms change with time.
- Analytical solution for \( x(t) \) and \( v(t) \) is impossible; numerical solution is trivial.

\[
\begin{align*}
  x(t + \Delta t) &= x(t) + v(t) \Delta t + \left[ \frac{4a(t) - a(t - \Delta t)}{6} \right] \Delta t^2 / 6 \\
  v(t + \Delta t) &= v(t) + \left[ \frac{2a(t + \Delta t) + 5a(t) - a(t - \Delta t)}{6} \right] \Delta t / 6
  
end{align*}
\]

- **Kinetic energy**
  \[ U_{\text{kinetic}} = \frac{1}{2} \sum m_i \cdot v_i(t)^2 = \frac{1}{2} n \cdot k_B \cdot T \]

- Total energy \((U_{\text{potential}} + U_{\text{kinetic}})\) must not change with time.

©Michael Levitt 01
HYDROPHOBIC EFFECT
SIMULATING THE HYDROPHOBIC EFFECT

Tanya Raschke

- 1 nanosecond MD simulations in periodic water boxes with from 30mM to 3 Molar hydrocarbon solution. Encad with F3C water (1996).

- Measure cluster formation by Voronoi. \( d(AB) = d(BC), \) but only A, B touch.

Box with periodic boundaries.
MOVIE OF BENZENE MOLECULAR DYNAMICS IN WATER AT ROOM TEMPERATURE
HYDROPHOBIC ENERGY IS COOPERATIVE

\[ \Delta G_N = -kT \log \left[ \frac{C_N}{(C_{N-1} C_1)} \right] \]

- Assume clusters are close-packed spheres:
  \[ V_N = NV_1 \]
  \[ A_N = \alpha (V_N)^{2/3} = \beta (N)^{2/3} \]
  \[ \Delta A_N = \beta [(N)^{2/3} - (N-1)^{2/3}] \]

- If \( \Delta G_N = \gamma \Delta A_N \), then
  \[ \Delta G_N = \gamma \beta [(N)^{2/3} - (N-1)^{2/3}] \]

- Determine \( \gamma \) by fitting with \([ (N)^{2/3} - (N-1)^{2/3} ] \).
HYDROPHOBIC ENERGY DEPENDS ON BURIED SURFACE

Constant of proportionality matches experiment.
UNFOLD THE $\alpha$-HELIX

13 Alanine residues

- Start as an ideal $\alpha$-helix in a box of water.
- Run 200 ps (100,000 time steps) of molecular dynamics at six different temperatures.
- Record percentage $\alpha$-helix formed for last 50 ps.
- See temperature-induced melting on picosecond time-scale.

Put it in a box of water.
\(\alpha\)-HELIX LESS STABLE IN WATER

- In vacuo the helix is very stable even at high temperature.
- In water the helix is unstable at high temperature.
- The rate of melting depends on temperature.
- Happens because water molecules stabilize the transition state.
WATER ALLOWS HYDROGEN BONDS TO BREAK

- Water catalyzes the breakage of hydrogen bonds by stabilizing the transition state.
SIMULATE FOLDING
Simulating folding is difficult?

- Simulation of 1 millisecond requires 10,000 CPU years!
- Must get over high barriers & many degrees of freedom.
MASSIVE COMPUTATIONAL RESOURCES

• Empty Supercomputers.
• Blue Gene (IBM).
• Folding@home (Vijay Pande).
FOLDING@HOME

Using Folding@home

• Project Goals: solving the protein folding problem
• How you can help
• Download (New! Version 1.33)
• How to install our software
• Frequently asked questions (FAQ)
• Contact Folding@home (Help Center)
• Folding@home discussion board

Fold proteins on 20,000 computers using the program as a Screen Saver!

http://www.stanford.edu/group/pandegroup/Cosm/

Join Folding@home by running our screen saver or client software

Like SETI@home

©Michael Levitt 01
PANDE MOVIE

α-HELIX FOLDING
50 YEARS OF SIMULATION

- We have 10,000,000 times more resources.
- Systems have become larger (100 times).
- Runs have become longer (100,000 times).
- Energy functions have become simpler.
- Fit reality well. Nothing bad has happened!

1955 Argon  1970 Water  1988 Protein in Water

©Michael Levitt 01
PREDICTION
WHAT DRIVES FOLDING?

- Protein is a chain.
- Self-avoiding and close packed.
- Residue preferences:
  - Inside/Outside
  - Specific Neighbors

Red are hydrophobic, like to be away from water
Green are hydrophilic, like contact with water

Hydrophobic  Hydrophilic  All Residues
Discrimination Paradigm
A PARADIGM FOR PREDICTING STRUCTURE

DECOYS
- Construct a large number of possible folded shapes (Decoys).

DISCRIMINATION
- Select the correct, native fold.

Need a good energy function
THE CASP EFFECT

- Critical Assessment of Structure Prediction.
- Predict what no one knows.
- Predict what is about to be known.
- Carefully control evaluation and assessment. Competition?
- Meet to discuss what went wrong and what went right.
- Have had CASP1 (‘94) through CASP4 (‘00).

HIEARCHICAL REDICTION
1998
Hierarchical Structure Prediction

Use a $\sqrt{3}$-state lattice model.

Use a 4-state off-lattice model.

Yu Xia  Ram Samudrala

1 of 10,000 low-energy shapes.

David Hinds

Predicted secondary structure

Britt Park

Add all atoms in full detail.

©Michael Levitt 01
Hierarchical prediction does well


T46/adg 7.5 Å (49 residues; 66113)

* T56/dnab 6.8 Å (60 residues; 67126)

** T59/smd3 6.7 Å (46 residues; 3075)

** T61/hdea 7.4 Å (66 residues; 974)

SPECIAL POTENTIALS
2000
SAMPLING ANT LION TOWN POTENTIALS

Uniform Exhaustive Search

Random start Minimization or Monte Carlo
Energy Minimization
ALL-ATOM ENERGY MINIMIZATION

- Minimize all-atom energy with respect to all torsion angles.
- Augment the normal potential energy function with:
  - Cooperative hydrogen bonds.
  - Cooperative hydrophobic interactions.
  - Forced exposure of charges.

Chen Keasar
\[ U = \sum K_{\phi} [1 - \cos(n\phi + \delta)] \]
\[ + \sum \varepsilon [(r_0/r)^2 - 2(r_0/r)] \]
\[ + \sum 332 q_i q_j / r \]

A protein with \( N \) residues has about
\( 4N (\phi, \psi, \chi) \) single bond torsion
angles. The same protein has about
\( 50N \) Cartesian coordinates \((x, y, z)\).
COOPERATIVE HYDROPHOBIC PACKING

- Cooperative hydrophobic compaction makes a good core.

Original Potential

Modified Potential
STRUCTURE PREDICTION BY MINIMIZATION

- Minimize special energy function with respect to torsion angles ($\phi, \psi, \chi$).
- Add energy terms for cooperative hydrogen bonds and hydrophobic compaction.

This method did well at CASP4, 2000.

10Å = 1 nm

Structure of T102

Best T102. RMS = 3.3Å
T102 Submitted. RMS = 5.0Å
COOPERATIVE HYDROGEN BONDS

- Cooperative hydrogen bonds give rise to good secondary structure.
All-β PREDICTION SUCCESS

• All-β sheet proteins are the hardest to predict.

• Torsion minimization does well on T114, an all β-protein.

Native Structure. Prediction is somewhat similar.
Segment Monte Carlo
MONTE CARLO METHODS

(a) At each step, attempt many moves.

(b) Accept the first move that obeys: Random number, \( R_n < \exp(-\Delta U/kT) \)

- Normal Monte Carlo:
  Make random moves and accept some of them (Metropolis).

- Simulated Annealing:
  Reduce \( T \), the temperature, as the run proceeds.
**FRAGMENT MONTE CARLO**

- Make a library of small fragments of similar sequence.
- Swap in a new fragment by setting six \((\phi, \psi)\) torsion angles.
- Accept move by Monte Carlo and anneal.
**KNOWLEDGE-BASED ENERGIES**

- Get distribution of distances between pairs of atom centers of a particular type, e.g. D-OD1...F-CD2.

- Normalize and take log to get Energy score:

\[ E_{ij}(r) = \log \frac{N_{ij}(r)}{M_{ij}(r)} \]
SEGMENT FOLDING PREDICTION

- Do Monte Carlo moves with respect to $(\phi, \psi)$ torsion angles. Simulated annealing.

- Use all-atom Knowledge-Based energy function.
  Add terms to enforce compaction.

- Get reasonable $(\phi, \psi)$ angles from real protein fragments.

- This method does well at CASP4, Asilomar '00.

  T0110 Fit 80 residues to 4.0Å
HARD PROBLEMS
WHY IS FOLDING SO HARD?

• Many different specific interactions.

• Cooperativity of the underlying interactions.

• Three-dimensional with very many possible spatial arrangements.

• Violates Crick’s Law of Hard Problems.
WHY ARE WE GETTING BETTER AT FOLDING?

- Peer pressure (CASP)?
- Faster computers?
- Many more sequences?
- More structures?
INFORMATION + PHYSICS = LIFE

DNA Sequence → RNA Sequence → Protein Sequence → Folded Protein

in silico

Easy: Change T to U

Easy: Triplet Code

Hard: Folding is many body simulation

in vivo

Easy: Folding is free by laws of physics

Hard: Transcription Polymerase

Hard: Translation Ribosome

©Michael Levitt 01
ACKNOWLEDGEMENTS

PEOPLE
- Tanya Raschke
- Erik Sandelin
- Boris Fain
- Patrice Koehl
- Michael Sykes
- Yu Xia
- Rachel Kolodny
- Chen Keasar
- Nizar Batada
- Chris Summa

SUPPORT
- NIH (NIGMS); DOE (SBI); NSF (ITR)

WEB
- http://csb.stanford.edu/levitt
- http://biospace.stanford.edu
- http://dd.stanford.edu
- http://astral.stanford.edu

- Papers
- Lecture
- Thesis
THE END
## Historical Record of Best Predictions at CASP

<table>
<thead>
<tr>
<th>CASP &amp; Year</th>
<th>Number of Targets</th>
<th>Best Result</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASP1 1994</td>
<td>6</td>
<td>63</td>
<td>Rost &amp; Sander</td>
</tr>
<tr>
<td>CASP2 1996</td>
<td>24</td>
<td>70</td>
<td>Rost</td>
</tr>
<tr>
<td>CASP3 1998</td>
<td>18</td>
<td>75</td>
<td>Jones</td>
</tr>
<tr>
<td>CASP4 2000</td>
<td>28</td>
<td>80</td>
<td>Jones</td>
</tr>
</tbody>
</table>

- Steady improvement of about 5% per CASP (every two years)

©Michael Levitt 01
NOTES

- PDF files on home page
- Searching www.google.com