Introduction to Protein Structure Prediction

Björn Wallner

Feb. 25, 2013
Today's topics

- Protein Folding
- Intro Protein structure prediction
- How can we predict the structure of a protein
- How can we tell which part of structure that needs to be improved.
- What can bioinformatics do?
- Finally, perhaps the most significant finding in protein structure prediction research the last 10 years! - The rebirth of correlated mutations.
From sequence to structure
Why do we need structure prediction?
DNA sequencing is getting cheap

Cost per Megabase of DNA Sequence

Moore's Law

National Human Genome Research Institute

genome.gov/sequencingcosts

Monday, February 25, 13
From sequence to structure
Two main routes

EQLTKCEVFQKLKDLKD....

folding

Physics
De Novo prediction

Evolution
Homology Modeling

Monday, February 25, 13
De Novo Prediction

• Requires only the amino acid sequence
• Assumption: The native fold is the structure with the lowest free energy
• Aim: Find the structure with the lowest free energy
De Novo Prediction

- Energy function - measure how “good” a certain protein structure is.
- A method to sample different structures.
De Novo Prediction

• Energy function - measure how “good” a certain protein structure is.

• A method to sample different structures.
De Novo Prediction

• Energy function - measure how “good” a certain protein structure is.

• A method to sample different structures.
De Novo Prediction

- Energy function - measure how “good” a certain protein structure is.
- A method to sample different structures.
De Novo Prediction

• Energy function - measure how “good” a certain protein structure is.

• A method to sample different structures.
De Novo Prediction

- Energy function - measure how “good” a certain protein structure is.
- A method to sample different structures.
De Novo Prediction

- Energy function - measure how “good” a certain protein structure is.
- A method to sample different structures.
De Novo Prediction

• Energy function - measure how “good” a certain protein structure is.

• A method to sample different structures.
De Novo Prediction

- Energy function - measure how “good” a certain protein structure is.
- A method to sample different structures.
De Novo Prediction

- Energy function - measure how “good” a certain protein structure is.
- A method to sample different structures.
De Novo Prediction

• Energy function - measure how “good” a certain protein structure is.
• A method to sample different structures.
De Novo Prediction

• Energy function - measure how “good” a certain protein structure is.

• A method to sample different structures.
De Novo Prediction

- Energy function - measure how “good” a certain protein structure is.
- A method to sample different structures.
De Novo Prediction

• Energy function - measure how “good” a certain protein structure is.

• A method to sample different structures.
What determines how proteins fold?

- Are protein folding pathways designed into sequences by natural selection?
- Make large random libraries, select out those sequences which fold based on function
- Compare their folding and stability to the starting structure
A

Stabilities of random variants almost always less than WT

B

But folding rates are sometimes faster!

Thus, folding rates probably not optimized by evolutionary selection

David Kim ~1998
What determines the rate of folding?

• If not dependent on details of sequence, perhaps topology of native structure?
• Contact Order: average sequence separation between residues in contact in native structure

LOW CONTACT ORDER

HIGH CONTACT ORDER
Folding rates correlates with contact order

- Graph showing the correlation between contact order and log of folding rate.

The x-axis represents contact order (%) and the y-axis represents log of folding rate. The data points are color-coded and differentiated by symbols.
What have we learnt from folding studies?

1. Local interactions bias but do not uniquely determine conformations sampled by short segments of the chain.

2. Folding occurs when local structure segments oriented so as to bury hydrophobic residues, pair beta strands, etc.


4. Folding rates are largely determined by contact order of native structure. Short folding times -> low contact order structures.

5. Native interactions on average stronger/ more consistent than non native interactions -> native minima broader than non native minima.
Assume distribution of conformations sampled by sequence segment during folding similar to distribution of conformations adopted by segment in protein 3D structures.
• 1000-100,000 short simulations to generate a population of 'decoys'
• Remove very low contact order structures
• Select lowest energy structures
• Select broadest minima using cluster analysis

Kim Simons, David Shortle ~1998
Lowest energy structures sampled on independent trajectories
Lowest energy structures sampled on independent trajectories
Lowest energy structures sampled on independent trajectories.
Lowest energy structures sampled on independent trajectories
• Soluble proteins, multimeric proteins, heterodimers, RNAs, membrane proteins, etc.

• Reflection of very large free energy gaps required for existence of single unique native state - think about statistical mechanics! \( \exp(-[\text{energy gap}]/(kT)) \)

• Prediction possible because (magnitude of actual free energy gap) >> (error in free energy calculation)

• Challenge: how to sample close to native state?
Rosetta folding video

Searching...  Accepted

Low Energy  Native  RMSD

Accepted Energy

mem widd run03 centroid A 1waz SAVE ALL OUT IGNORE THE REST 22158_317131

Stage: MembraneNormalPerturbationMover
CPU time: 0 hr 25 min 27 sec
The-Real-Link - Total credit: 73356.5 - RAC: 2902.51
DeathCom Multimedia

25.45% Complete
Model: 6 Step: 7163
Accepted Energy: 0.2536324
Accepted RMSD: 14.03
Low Energy: -10.77007
Low RMSD: 12.53

Rosetta@home v2.17 http://boinc.bakerlab.org/rosetta/
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**
  - Hard: Transcription Polymerase
  - Hard: Translation Ribosome
  - Easy: Folding is free by laws of physics
Two main routes

EQLTKCEVFQKLKDLKD....

folding

Physics

*De Novo* prediction

Evolution

Homology Modeling
Protein Evolution - In the Lab

(a few) random mutations → select/screen

Parent gene (= parent protein)

Evolved gene (= evolved protein)

repeat

NO

YES

Monday, February 25, 13
Protein Evolution - In the vivo

(a few) random mutations

Survival?

NO

Parent gene
(= parent protein)

repeat

YES

Evolved gene
(= evolved protein)
Two related proteins (homologs)
Two related proteins (homologs)
Homology modeling

- **Assumption:** Proteins with high sequence similarity have the same overall structure

- **Idea:** Look in the PDB _if_ there is a structure with a similar sequence that can be used as starting point (template) for modeling.

- **Example:** If the human protein (red) is not known and we want to create a model, using the fly protein (yellow) as a *template* is really good starting point.
Homology modeling

- **Assumption:** Proteins with high sequence similarity have the same overall structure

- **Idea:** Look in the PDB _if_ there is a structure with a similar sequence that can be used as starting point (template) for modeling.

- **Example:** If the human protein (red) is not known and we want to create a model, using the fly protein (yellow) as a template is really good starting point
Does high sequence similarity \(<\rightarrow\) similar structure?
“Look” in the PDB

• Perform sequence search against all proteins with known 3D-structure (e.g. BLAST search against PDB, or more sensitive methods like PSI-blast, HHblits or jackhmmer)

• Compare sequences using Sequence Alignment
• Maximize the number of positions that are in agreement between two sequences
Global Alignments:
LGPSTKDFGKISESREFDN
 |     |       |   |
LNQLERSFGKINM−RLEDA

Local Alignments:
-------------------FGKI-------------------
 |     |     |
-------------------FGKI-------------------
Sequence similarity – dot plot

How to do this? Simplest approach: Dot plot.
Sequence similarity – dot plot

Filter short matches (<3 positions)
Similar amino acids score higher

BLOSUM62
[amino acid substitution matrix]
Model Building Steps in Practice

- Find template
- Align target sequence with template
- Generate model:
  - add loops
  - add side chains
- Refine model
Modeling Overview

- Identify related structures (templates)
- Align target sequence to template structures
- Build a model for the target sequence using information from template structures
- Evaluate the model
- NO model OK?
- YES

START

TARGET SEQUENCE

TEMPLATE STRUCTURE

ALIGNMENT

TARGET ...SCDKLLDDELDDIACAKILAIKGID...
TEMPLATE ...SCDKFLDDITDDDIMCAKKILDIGID...

TARGET MODEL

QUALITY PROFILE
• How can we tell if a model is good or bad?
Sometimes it is easy
Sometimes it is _not_ so easy
Sometimes it is _not_ so easy
Sometimes it is _not_ so easy

Native
Sometimes it is _not_ so easy

Native

85% correct
70% correct
Assessing models in 3D

• Scoring Function
• Sequence-Structure alignment must make sense in 3D.
• Few gaps in secondary structure
• Good strand pairing
Scoring Function

how preferable to put two particular residues nearby: $E_p$
(Pairwise potential)

alignment gap penalty: $E_g$
(gap score)

how well a residue fits a structural environment: $E_s$
(Fitness score)

sequence similarity between query and template proteins: $E_m$
(Mutation score)

How consistent of the secondary structures: $E_{ss}$

$E = E_p + E_s + E_m + E_g + E_{ss}$

Minimize $E$ to find a sequence-template alignment
Statistical Potentials

• Construct a database of all 20x20 or 21*20/2 amino acid pairs

• Derive a potential using:
  \[ P(a,b) = \exp(-E(a,b)/kT) \]

• \[ E(a,b) = -kT \ln P(a,b) \]

• Predict a given sequence using the pairwise potentials
Problem: Many different parameters

- Pair Energy
- Solv. Energy
- Alignment score
- Alignment Length
- Len1 (Struct)
- Len2 (Seq)
Neural Network Could Make the Decision

Pair Energy → Input Layer
Solv. Energy → Hidden Layer
Alignment score → Hidden Layer
Alignment Length → Hidden Layer
Len1 (Struct) → Hidden Layer
Len2 (Seq) → Hidden Layer

Output Layer → Proteins related
Output Layer → Proteins unrelated
Lower alignment score can be compensated by better pairwise energy
Neural Network provides Confidence Scores

<table>
<thead>
<tr>
<th>Low</th>
<th>Medium (80%)</th>
<th>High (99%)</th>
<th>Certain (100%)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Monday, February 25, 13
The structure of a protein is *uniquely* determined by its amino acid sequence except for:

- prions
- pH, ions, cofactors, chaperones

Structure is more conserved than sequence

- Sequence -> Structure -> Function
- Possible to infer: Sequence -> Function
• Chothia & Lesk (1986): Correlation between structural divergence and sequence similarity
50% of all protein families can be modeled

<table>
<thead>
<tr>
<th>Coverage type</th>
<th>Percentage coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Per sequence</td>
</tr>
<tr>
<td>Integr8_263 SP-trEMBL Integr8_263 SP-trEMBL Integr8_263 SP-trEMBL</td>
<td>52.4</td>
</tr>
<tr>
<td>All Sequences</td>
<td>54.4</td>
</tr>
<tr>
<td>- excluding transmembrane &amp; problematic sequences</td>
<td>/</td>
</tr>
<tr>
<td>- excluding transmembrane problematic &amp; singleton sequences</td>
<td>/</td>
</tr>
<tr>
<td>Marsden et al. BMC Bioinformatics 2007 8:86 doi:10.1186/1471-2105-8-86</td>
<td>71.1</td>
</tr>
</tbody>
</table>

Families ordered by size

or.... 50% can not be modeled...

<table>
<thead>
<tr>
<th>Coverage type</th>
<th>Percentage coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Per sequence</td>
</tr>
<tr>
<td>Integ8_263 SP-trEMBL Integ8_263 SP-trEMBL Integ8_263 SP-trEMBL</td>
<td></td>
</tr>
<tr>
<td>All Sequences</td>
<td>52.4</td>
</tr>
<tr>
<td>- excluding transmembrane &amp; problematic sequences</td>
<td>/</td>
</tr>
<tr>
<td>- excluding transmembrane problematic &amp; singleton sequences</td>
<td>/</td>
</tr>
</tbody>
</table>

Finally some cutting edge research
Correlated mutations carry information about distance relationships in protein structure.
Problem indirect interactions

- *Indirect* correlations between A-C and A-D because they interact with B that *interacts* directly with A

- Number of indirect correlations $\gg$ direct correlation
Prediction of domain-domain interactions and structural modeling
Prediction of domain-domain interactions and structural modeling
From Evolutionary Constraints to 3D Structures
Summary

• Protein Folding
• Intro Protein structure prediction
• How can we predict the structure of a protein
• How can we tell which part of structure that needs to be improved.
• What can bioinformatics do?
• Finally, perhaps the most significant finding in protein structure prediction research the last 10 years! - The rebirth of correlated mutations.