Protein Folding: Predicting Structure from Sequence

Pralay Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur
• **Protein folding** is the process by which a protein structure assumes its functional shape or conformation from random coil.

• Each protein exists as an unfolded polypeptide or random coil when translated from a sequence of mRNA to a linear chain of amino acids.

Short term course on Bioinformatics, Oct 31, 2013, IIT KGP
Hemoglobin structure

High-resolution crystal structure of human hemoglobin with mutations at tryptophan 37beta: structural basis for a high-affinity T-state. (PDB ID: 1A00)
Energy landscape

In physics and biochemistry, an energy landscape is a mapping of all possible conformations of a molecular entity, or the spatial positions of interacting molecules in a system, and their corresponding energy levels, typically Gibbs free energy, on a two- or three-dimensional Cartesian coordinate system.
Energy landscape
End of the story ??

This is just the beginning.
Nobel 2013 in Chemistry

The computer – your Virgil in the world of atoms

Chemists used to create models of molecules using plastic balls and sticks. Today, the modelling is carried out in computers. In the 1970s, Martin Karplus, Michael Levitt and Arieh Warshel laid the foundation for the powerful programs that are used to understand and predict chemical processes. Computer models mirroring real life have become crucial for most advances made in chemistry today.

Protein folding

Input Sequence:

NSTNLPRNPSMADYEARIIFTFGTWIYSVNKEQLARAGFYALGEGDKVKC…..

Output Structure:
Protein folding

- Approaches
  - Ab initio
  - Template Based
- Simulation
  - Simulated Annealing
  - Monte Carlo
  - Genetic Algorithms
- Score/energy function
  - Physics based
  - Evolution information based
  - Hybrid
QUARK: Ab initio Prediction Method

Xu and Zhang (2012)
Proteins 1715:1735
QUARK:
The Flow

Xu and Zhang (2012)
Proteins 1715:1735
QUARK: The Flow

Xu and Zhang (2012)
Proteins 1715:1735

Short term course on Bioinformatics, Oct 31, 2013, IIT KGP
Identification of protein fold

>1A00:A | PDBID | CHAIN | SEQUENCE
VLSPADKTNVKAAGKVGAGHAGEYGAENAME
FLSFPTTKTYFPHFDLSHGSAQVKGHGKVAD
ALTNAVHVDMPNAL
SALSDLHALHKLRTVPNFKLLSHCNYLVTLLAH
LPAEFTPAVHASLDKFLASVSTVLTSKYR
Side Chain Fitting
Side Chain Fitting
Side Chain Fitting
QUARK: Result

Xu and Zhang (2012)
Proteins 1715:1735
QUARK: Result

Xu and Zhang (2012)
Proteins 1715:1735
QUARK: Result

A

B

C

D

T0547-D3  RMSD  5.88Å  TM-score  0.653  GDT-TS  68.99
T0618-D1  RMSD  11.46Å  TM-score  0.478  GDT-TS  41.77

T0624-D1  RMSD  8.22Å  TM-score  0.378  GDT-TS  44.56
T0529-D1  RMSD  23.37Å  TM-score  0.170  GDT-TS  10.87

Xu and Zhang (2012)
Proteins 1715:1735
Use Evolutionary Information whenever is available.
Protein folding – template based

Input Sequence:
NSTNLPRNPSMADYEARIIFTFGTWIYSVNKEQLARAGFYALGEGDKVKC.....

Output Structure:

2OPZ 1SE0 1TFT 2VM5 1QBH 3M0A 1E31

2VM5 1SE0 1TFT 1E31 3M0A 1QBH
TM-score=0.66 TM-score=0.65 TM-score=0.65 TM-score=0.61 TM-score=0.6 TM-score=0.57

Short term course on Bioinformatics, Oct 31, 2013, IIT KGP
I-TASSER Method

Figure 1 | A schematic representation of the I-TASSER protocol for protein structure and function predictions. The protein chains are colored from blue at the N-terminus to red at the C-terminus.

CASP

Critical Assessment of protein Structure Prediction

(http://predictioncenter.org/)
CASP
Critical Assessment of protein Structure Prediction
(http://predictioncenter.org/)

Best public CASP-certified protein structure prediction servers

- I-TASSER
- ROBETTA
- HHpred
- METATASSER
- MULTICOM
- Pcons
- SAM-T08
- 3D-Jury
- THREADER
Protein-Protein Docking: Prediction of Protein Association

Pralay Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur
Background

Proteins are the building blocks of the cells; perform bulk of the functions of the cell. The functionality of a protein is determined by its interaction with other proteinous or non-proteinous molecules.

It will take few decades to experimentally determine all the protein complex structures at atomic level resolution.

An alternative: computational modeling of protein-protein interactions; commonly known as protein-protein docking.

Aloy et al. (2004). Nat. Biotechnology

Epstein–Barr virus (EBV) – human protein interaction network

Protein-Protein Docking

*Ab initio Protein-protein docking* is the determination of the molecular structure of *complexes* formed by two or more proteins without the need for *experimental* measurement.
Docking Types

- Based on crystallization information
  - Bound docking
  - Unbound docking

- Based on protein flexibility
  - Rigid Body
  - Flexible Body
Docking Strategy

Generating

- Generate the orientations so that a number of decoys (complexes) is formed.

Pruning

- Reduce the number of decoys (i.e., search space) by some coarse grain method.

Scoring

- Assign score to each decoys

Ranking

- Rank the decoys based on score
Docking Search Strategies

- **Pseudo Random**
  - Simulated Annealing / Monte Carlo
  - Genetic Algorithms

- **Directed Search**
  - Geometric Hashing
  - Spherical Harmonic Surface Triangles

- **Brute-Force Search**
  - Explicit Grid Correlations
  - Fast Fourier Transform (FFT) Correlations
  - Spherical Polar Fourier Correlations
Geometric Hashing

- Models are represented in a redundant affine invariant way and stored in a table (off-line).
- Hashing is used for organizing and searching the table.
Geometric Hashing

- **Pro:**
  - Faster

- **Con:**
  - Storage requirement is very high and increases with the increase in object points.
  - Proper identification of object points are crucial for the success.
Generation methods

- Tagline – “Higher the decoys; better the possibility of having a hit”
- How many is good?
- Move to discrete space
Generation methods

On an average some brute force method can generate $\sim 10^7$ decoys.
Fast Fourier Technique

\[ \bar{a}_{l,m,n} = \begin{cases} 1 & \text{on the surface of the molecule} \\ \rho & \text{inside the molecule} \\ 0 & \text{outside the molecule,} \end{cases} \]

and

\[ \bar{b}_{l,m,n} = \begin{cases} 1 & \text{on the surface of the molecule} \\ \delta & \text{inside the molecule} \\ 0 & \text{outside the molecule,} \end{cases} \]
Fast Fourier Technique*

\[
\overline{c}_{\alpha,\beta,\gamma} = \frac{1}{N^3} \sum_{o=1}^{N} \sum_{p=1}^{N} \sum_{q=1}^{N} \exp[2\pi i (o\alpha + p\beta + q\gamma)/N] \cdot C_{o,p,q}
\]

*Katchalski-Katzir et al, (1992) PNAS
Docking Strategy

**Generating**
- Generate the orientations so that a number of decoys (complexes) is formed.

**Pruning**
- Reduce the number of decoys (i.e., search space) by some coarse grain method.

**Scoring**
- Assign score to each decoys

**Ranking**
- Rank the decoys based on score
Generation methods

On an average some brute force method can generate $\sim 10^7$ decoys.

Assuming processing of each decoy takes 1 sec; total processing time $\sim 115$ days.
Fast Fourier Technique*

\[ \overline{c}_{\alpha,\beta,\gamma} = \frac{1}{N^3} \sum_{o=1}^{N} \sum_{p=1}^{N} \sum_{q=1}^{N} \exp\left[2\pi i \left(o \alpha + p \beta + q \gamma \right)/N\right] \cdot C_{o,p,q} \]

*Katchalski-Katzir et al, (1992) PNAS
Pruning methods

Integrated Scoring

FTDock (Katchalaski-Katzir (1992), PNAS)

Edge Scoring

Docking Strategy

Generating

Protein A

Protein B

Generate the orientations so that a number of decoys (complexes) is formed.

Pruning

Reduce the number of decoys (i.e., search space) by some coarse grain method.

Scoring

Assign score to each decoy.

Ranking

Rank the decoys based on score.

Short term course on Bioinformatics, Oct 31, 2013, IIT KGP
Scoring methods

Ab initio scoring (Physics based)

- Contact Area
- Contact Packing
- Non-bonded interactions
- Solvation Energy
- Etc.

Evolutionary scoring (Template based)
Ab initio method

- Interface area (IA)
- Normalized interface packing (NIP)
- Normalized surface complementarity (NSc)
- Non-bonded energy (NE):

\[ NE = \sum_{i<j}^{\text{atoms}} \left( \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{4\Pi \varepsilon R_{ij}} \right) \]

- Solvation energy (SE):*

\[ SE = \sum_{\text{interface atoms}} \Delta \sigma(\text{Atom Type}) \times \Delta \text{ASA} \]

*Eisenberg and McLachlan (1986) *Nature*
NSc and NIP at protein interface*

Correlation coefficient of NIP and NSc is $+0.95$

Scoring methods

Correlation among the four physico-chemical properties at the protein interfaces

\[ SC = 0.6547 (\pm 0.0010) \times IP + 0.1495 (\pm 0.0002) \]

Scoring and Ranking*

Compute $IP$, $SC$, $NE$ and $SE$ at the decoy interface

Group the decoys such that all decoys with RMSD<1.0Å and difference in SP<0.04 is in a group $G$, where

$$SP = |SC-IP \times 0.6547 - 0.1495|$$

Nonbonded energy for a group $G$: $NE^G = \overline{NE} - \sigma(NE)$
Solvation energy for a group $G$: $SE^G = \overline{SE} + \sigma(SE)$

$NE^G_i$: $NE^G$ bin number in all groups’ $NE$ histogram
$SE^G_i$: $SE^G$ bin number in all groups’ $SE$ histogram

$$Score = \sqrt{((NE^G_i \times NE^G_i) + (SE^G_i \times SE^G_i)) + SP^G \times 10.0}$$

where, $SP^G$ is minimum SP of the group $G$.

Rank of a decoy is its position in the sorted list

Sort (in ascending order) the group of decoys based upon their scores.

Docking types

• Bound docking
  • The crystal structure of complex is available. Interacting/docking partners are taken from that complex structure.
  • Easy to model since the side chain orientation is proper.

• Predictive/Unbound docking
  • The docking partners and complex structure is separately crystallized.
  • Side chain refinement is required
The Dataset

Bound

Download data from PDB with Resolution ≤2.5 Å and R-factor ≤0.2

Remove proteins which are NOT dimers (consult PQS, PISA or literature wherever is required)

Reject PDB if it has ligand mediated interaction

Reject PDB if both the subunits are not of size >25

Make it non-redundant at 90%

828 homodimers + 119 heterodimers = 947 protein dimers

Unbound heteromers

Compile data from Benchmark 3.0, Gottschalk et al. 2004 and from Bernauer et al. 2007

26 unbound-unbound + 6 bound-unbound = 32 protein hetero dimers
Evaluating bound dataset

(A) Variation of accuracy with rank. The darker curve shows the accuracy where the dimers could be successfully screened by IA filter. The lighter curve shows the accuracy over the whole dataset.

(B) Variation of accuracy with rank when the cases screened by IA filter was divided into various interface area categories.

Example prediction (PDB: 1EX2)

- Red: Charged
- Green: Aromatic
- Yellow: Hydrophobic
- Brown: Polar

PDB and PQS structure

Residue property at the interface of the protein - a conserved *Bacillus subtilis* protein Maf

ZRANK

\[
Score = w_{vdW-a}E_{vdW-a} + w_{vdW-r}E_{vdW-r} + w_{elec-sra}E_{elec-sra} \\
+ w_{elec-srr}E_{elec-srr} + w_{elec-lra}E_{elec-lra} \\
+ w_{elec-lrr}E_{elec-lrr} + w_{ds}E_{ds}
\]

\[
E_{vdW}(i,j) = \varepsilon_{ij} \left( \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right)
\]

Van der Wall interaction

\[
E_{elec}(i,j) = 332 \frac{q_i q_j}{r_{ij}^2}
\]

Electrostatic Interaction

\[
E_{ds}(i,j) = \alpha_{ij}
\]

Desolvation energy
ZRANK

Pierce and Weng (2007)
Proteins, 67:1078–1086
PatchDock and FireDock

• **PatchDock**: Molecular Docking Algorithm Based On Shape Complementarity Principles

• **FireDock**: Includes three main steps:

  (1) Side-chain optimization: The side-chain flexibility of the receptor and the ligand is modeled by a rotamer library. The optimal combination of rotamers for the interface residues is found by solving an integer LP problem.

  (2) Rigid-body minimization: This minimization stage is performed by a MC technique that attempts to optimize an approximate binding energy by refining the orientation of the ligand structure.

  (3) Scoring and ranking: This final ranking stage attempts to identify the near-native refined solutions. The ranking is performed according to a binding energy function that includes a variety of energy terms: desolvation energy, van der Waals interactions, partial electrostatics, hydrogen and disulfide bonds, $p$-stacking and aliphatic interactions, rotamer’s probabilities and more.

Predictive Docking - The Unbound Dataset*

Docking from sequence
Application to Genome-wide scale
COTH – docking from sequence

Mukherjee and Zhang (2011)
Structure 19, 955–966
COTH – docking from sequence

The native complex (Ran-Importin β complex) is represented in cyan.
Critical Assessment of PRediction of Interactions (CAPRI)
## Critical Assessment of PRediction of Interactions (CAPRI)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Affiliation</th>
<th>Software</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abagyan</td>
<td>Scripps</td>
<td>ICM</td>
<td>Force Field</td>
</tr>
<tr>
<td>Camacho/Vajda</td>
<td>Boston</td>
<td>CHARMM</td>
<td>Force Field Refinement</td>
</tr>
<tr>
<td>Gardiner</td>
<td>Sheffield</td>
<td>GAPDOCK</td>
<td>Shape+Area GA</td>
</tr>
<tr>
<td>Sternberg/Smith</td>
<td>Imperial</td>
<td>FTDOCK</td>
<td>FFT</td>
</tr>
<tr>
<td>Bates/Fitzjohn</td>
<td>ICRF</td>
<td>Guided Docking</td>
<td>Force Field</td>
</tr>
<tr>
<td>Ten Eyck/Mitchell</td>
<td>SDSC</td>
<td>DOT</td>
<td>FFT</td>
</tr>
<tr>
<td>Vakser/Tovchigrechko</td>
<td>SUNY/MUSC</td>
<td>GRAMM</td>
<td>FFT</td>
</tr>
<tr>
<td>Olson</td>
<td>Scripps</td>
<td>Harmony</td>
<td>Spherical Harmonics</td>
</tr>
<tr>
<td>Weng/Chen</td>
<td>Boston</td>
<td>ZDOCK</td>
<td>FFT</td>
</tr>
<tr>
<td>Eisenstein</td>
<td>Weizmann</td>
<td>MolFit</td>
<td>FFT</td>
</tr>
<tr>
<td>Wolfson/Nussinov</td>
<td>Tel Aviv</td>
<td>BUDDA/PPD/FireDock</td>
<td>Geometric Hashing</td>
</tr>
<tr>
<td>Iwadate</td>
<td>Kitasato</td>
<td>TSCF</td>
<td>Force Field+Solvent</td>
</tr>
<tr>
<td>Ritchie/Mustard</td>
<td>Aberdeen</td>
<td>Hex</td>
<td>Spherical Polar Fourier</td>
</tr>
<tr>
<td>Palma</td>
<td>Lisbon</td>
<td>BIGGER</td>
<td>Geometric+Electrostatic</td>
</tr>
<tr>
<td>Gray/Baker</td>
<td>Washington/JHU</td>
<td>RosettaDock</td>
<td>Monte Carlo+Flexibility</td>
</tr>
<tr>
<td><strong>Mitra and Pal</strong></td>
<td>IISc</td>
<td>PROBE/PRUNE</td>
<td>FFT</td>
</tr>
</tbody>
</table>

---

T50, T53
Parallel Implementation

• At the generation phase:
  • The protein can be divided into different parts that are mutually exclusive.

• At the scoring phase:
  • All the decoys are mutually independent; thus they can be processed separately on different processors.
Summary

✓ The bound test set is easy to predict, but the real benchmark set is unbound data set.

✓ Refining the side chain of the unbound docked complexes are still an active area of research.

✓ Computationally flexible docking is more challenging than rigid body docking.

✓ Evolutionary information can be integrated to improve the performance of the method.
Acknowledgement

- Prof. Yang Zhang
  - University of Michigan Medical School, Ann Arbor, USA
- Prof. Debnath Pal
  - Indian Institute of Science, Bangalore, India

Funding
National Science Foundation (NSF), USA
AICTE, India
Thank you for your attention

http://cse.iitkgp.ac.in/~pralay/