Protein Structure Prediction - cont.

11/14/05

Protein Structure Prediction & Modeling

Protein-nucleic acid interactions; protein-ligand docking (no time, sorry!)

Bioinformatics Seminars

Baker Center/BCB Seminars:

Nov 14 Mon 1:10 PM Doug Brutlag, Stanford
Discovering transcription factor binding sites

Nov 15 Tues 1:10 PM Ilya Vakser, Univ Kansas
Modeling protein-protein interactions

both seminars will be in Howe Hall Auditorium

Bioinformatics Seminars

Nov 14 Mon 12:10 IG Seminar in 101 Ind Ed II
Building Using Eyes to Study Developmental Change During Evolution
Jeanne Serb, EEOB

Nov 15 Tues 2:10 PM An Sci Seminar in 1204 Kildee
Lab of Milk & Honey: Bioinformatics for Bovine & Bee
Chris Elsik, Texas A&M

Protein Structure Prediction

Genome Analysis

Mon Protein 3’ structure prediction
Wed Genome analysis & genome projects
Comparative genomics; ENCODE, SNPs, HapMaps, medical genomics
Thur Lab Protein structure prediction
Fri Experimental approaches: microarrays, proteomics, metabolomics, chemical genomics

Reading Assignment (for Mon–Fri)

Mount Bioinformatics

• Chp 11 Genome Analysis
• pp. 495 - 547
• Ck Errata: http://www.bioinformaticsonline.org/help/errata2.html

BCB 544 Additional Readings

Required:

• Gene Prediction
  Burge & Karlin 1997 JMB 268:78
  Prediction of complete gene structures in human genomic DNA
• Human HapMap (Nature 437, Oct 27, 2005)
  Commentary (437:1233)
  http://www.nature.com/nature/journal/v437/n7063/full/4371233a.html
  News & Views (437: 1241)
  http://www.nature.com/nature/journal/v437/n7063/full/4371241a.html

Optional:

• Article (437:1299) A haplotype map of the human genome
  The International HapMap Consortium

D Dobbs ISU - BCB 444/544X
Review last lecture:

Protein Structure Prediction  
**focus on:**  
Tertiary Structure

### Structural Genomics
- 2 X 10^6 proteins sequences in UniProt
- 3 X 10^4 structures in the PDB

- Experimental determination of protein structure lags far behind sequence determination
- **Goal:** Determine structures of "all" protein folds in nature, using combination of experimental structure determination methods (X-ray crystallography, NMR, mass spectrometry) & computational structure prediction

- ~ 30,000 "traditional" genes in human genome (not counting alternative splicing, miRNAs)
- ~ 3,000 proteins expressed in a typical cell

### Structural Genomics Projects
- **TargetDB:** database of structural genomics targets  
  [http://targetdb.pdb.org](http://targetdb.pdb.org)

### Protein Structure Prediction

- "Major unsolved problem in molecular biology"

- **In cells:** spontaneous  
  assisted by enzymes  
  assisted by chaperones

- **In vitro:** many proteins fold spontaneously  
  & many do not!

### Deciphering the Protein Folding Code
- Protein Structure Prediction  
  or "Protein Folding" Problem  
  given the amino acid sequence of a protein, predict its 3-dimensional structure (fold)

- "Inverse Folding" Problem  
  given a protein fold, identify every amino acid sequence that can adopt its 3-dimensional structure

### Protein Structure Determination?

- **High-resolution structure determination**  
  - X-ray crystallography (<1Å)
  - Nuclear magnetic resonance (NMR) (~1-2.5Å)

- **Lower-resolution structure determination**  
  - Cryo-EM (electron-microscopy) ~10-15Å

- **Theoretical Models?**  
  - Highly variable - now, some equiv to X-ray!
Tertiary Structure Prediction
Fold or tertiary structure prediction problem can be formulated as a search for minimum energy conformation
- Search space is defined by psi/phi angles of backbone and side-chain rotamers
- Search space is enormous even for small proteins
- Number of local minima increases exponentially of the number of residues

Computationally it is an exceedingly difficult problem!

Ab Initio Prediction
1. Develop energy function
   - bond energy
   - bond angle energy
   - dihedral angle energy
   - van der Waals energy
   - electrostatic energy
2. Calculate structure by minimizing energy function
   (usually Molecular Dynamics or Monte Carlo methods)

- Ab initio prediction - not practical in general
  - Computationally? very expensive
  - Accuracy? Usually poor for all but short peptides
  (but see Baker review!)

Ab initio prediction - provides both folding pathway & folded structure

Comparative Modeling
Two primary methods
1) Homology modeling
2) Threading (fold recognition)

- Note: both rely on availability of experimentally determined structures that are "homologous" or at least structurally very similar to target

Homology Modeling
1. Identify homologous protein sequences (PSI-BLAST)
2. Among available structures, choose the one with closest sequence match to target as template
   (combine steps 1 & 2 by using PDB-BLAST)
3. Build model by placing residues in corresponding positions of homologous structure & refine by "tweaking"

- Homology modeling - works "well"
  - Computationally? not very expensive
  - Accuracy? higher sequence identity \(\Rightarrow\) better model

Threading - Fold Recognition
Identify "best" fit between target sequence & template structure
1. Develop energy function
2. Develop template library
3. Align target sequence with each template & score
4. Identify best scoring template (1D to 3D alignment)
5. Refine structure as in homology modeling

- Threading - works "sometimes"
  - Computationally? Can be expensive or cheap, depends on energy function & whether "all atom" or "backbone only" threading
  - Accuracy? in theory, should not depend on sequence identity (should depend on quality of template library & "luck")
  - But, usually higher sequence identity \(\Rightarrow\) better model

Threading: more details
1. Align target sequence with template structures (fold library) from the Protein Data Bank (PDB)
2. Calculate energy score to evaluate goodness of fit between target sequence & template structure
3. Rank models based on energy scores
Threading Goals & Issues

- Find “correct” sequence-structure alignment of a target sequence with its native-like fold in PDB
- Structure database - must be complete: no decent model if no good template in library!
- Sequence-structure alignment algorithm:
  - Bad alignment → Bad score!
- Energy function (scoring scheme):
  - must distinguish correct sequence-fold alignment from incorrect sequence-fold alignments
  - must distinguish “correct” fold from close decoys
- Prediction reliability assessment - how determine whether predicted structure is correct (or even close?)

Threading: Structure database

- Build a template database
  (e.g., ASTRAL domain library derived from PDB)
- Supplement with additional decoys, e.g., generated using ab initio approach such as Rosetta (Baker)

Threading: Energy function

- Two main methods (and combinations of these)
  - Structural profile (environmental)
    physico-chemical properties of aa’s
  - Contact potential (statistical)
    based on contact statistics from PDB
    (Miyazawa & Jernigan - Jernigan now at ISU)

Protein Threading: typical energy function

- MTKLRLNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE
- What is "probability" that two specific residues are in contact?
- How well does a specific residue fit structural environment?
- Alignment gap penalty?
- Total energy: $E_s + E_c + E_g$
- Find a sequence-structure alignment that minimizes the energy function

New today:
Protein Structure Prediction

A Rapid Threading Approach for Protein Structure Prediction

- Kai-Ming Ho, Physics
- Haibo Cao
- Yungok Ihm
- Zhong Gao
- James Morris
- Cai-zhuang Wang
- Drena Dobbs, GDCB
- Jae-Hyung Lee
- Michael Terribilini
- Jeff Sander
Protein Structure Prediction - cont.

Template structure (reduced) representation

Template structure \( C \) (N \times N contact matrix)
- \( C_{ij} = 1 \), if \( r_{ij} \leq 6.5 \text{ Å} \) (contact)
- \( C_{ij} = 0 \), otherwise (non-contact)

Energy function
Assumption: At residue level, pair-wise hydrophobic interaction is dominant:
\[
E = \sum_{i,j} C_{ij} U_{ij}
\]
- \( C_{ij} \): contact matrix
- \( U_{ij} = U(\text{residue } I, \text{residue } J) \)
  - MJ matrix: \( U = U_{ij} \)
  - LTW: \( U = Q_i Q_j \)
  - HP model: \( U = (1,0) \)

Contact energy: pairwise interactions
Miyazawa-Jernigan (MJ) matrix:
Statistical potential:
210 parameters
\[
M = \begin{pmatrix}
0.46 & 0.54 & -0.20 & 0.49 & -0.01 & 0.06 & 0.57 & 0.01 & 0.03 & -0.08 & 0.52 & 0.18 & 0.10 & -0.01 & -0.04
\end{pmatrix}
\]
Li-Tang-Wingreen (LTW):
Factorize the MJ interaction matrix to reduce the number of parameters from 210 to 20 \( q \) values associated with 20 amino acids

Residue interaction scheme (Ho)
Interaction "counts" only if two hydrophobic amino acid residues are in contact
Miyazawa-Jernigan (MJ) model: inter-residue contact energy \( M(i,j) \) is a quasi-chemical approximation; based on contact statistics extracted from known protein structures in PDB

Li-Tang-Wingreen (LTW): Factorize the MJ interaction matrix to reduce the number of parameters from 210 to 20 \( q \) values associated with 20 amino acids

Residue interaction scheme
- \( q_i^2 \) - solubility
- \( Q_i \) - hydrophobicity
- \( C \) - contact matrix

Contact Energy:
\[
E = \sum (Q_i C_i + \beta Q_i)
\]

Trick for Fast Threading?
- ALKKGF_HFDTS
- Sequence - Structure (1D - 3D problem)
- Sequence - Contact Matrix (1D - 2D problem)
- >Sequence - 1D Profile (1D - 1D problem)
**1D profile? first eigenvector of contact matrix**

Hydrophobic Contacts

1. $\mathbf{CT} = \mathbf{C} \mathbf{T}$
2. $\mathbf{C} = \sum \lambda_i \mathbf{v}_i \mathbf{v}_i^T$

$C$: contact matrix

$\lambda_i$: $i$-th eigenvalue of $C$

$v_i$: $i$-th eigenvector

$T$: protein sequence of the template structure

$r_i$: fraction of hydrophobic contacts from $i$-th eigenvector

**Weights of eigenvectors for real proteins**

1. First eigenvector of contact matrix dominates the overlap between sequence and structure
2. Higher ranking (rank > 4) eigenvectors are "sequence blind"

**Fast threading alignment algorithm**

1. **1D Profile** $\mathbf{P} = \mathbf{V}_1$
2. Maximize the overlap between the Sequence ($\mathbf{S}$) and the profile ($\mathbf{P}$) allowing gaps

New profile $\mathbf{P} = \mathbf{C} \mathbf{P}$

Calculate contact energy using the alignment $\mathbf{E}$

**Parameters for alignment?**

**Gap penalty:**
- Insertion/deletion in helices or strands strongly penalized; small penalties for in/dels in loops
- but, gap penalties do not count in energy calculation

**Size penalty:**
- If a target residue $\mathbf{A}$ and aligned template residue differ in radius by $> 0.5 \text{Å}$ and if the residue is involved in $> 2$ contacts, alignment contribution is penalized
- but, size penalties do not count in energy calculation

**How incorporate secondary structure?**

Predict secondary structure of target sequence (PSIPRED, PROF, JPRED, SAM, GOR V)

- $N_+ = $ total number of matches between the predicted secondary structure and the template structure
- $N_- = $ total number of mismatches
- $N_s = $ total number of residues selected in alignment

"Global fitness" $f = 1 - (N_+ - N_-) / N_s$

$E_{modify} = f \cdot E_{threading}$
Finally, calculate "relative" score:

How much better is this "fit" than random?

\[ E_{\text{modify}} : \text{Sequence vs Structure} \]

(adjusted for 2' structure match)

\[ E_{\text{shuffled}} : \text{Shuffled Sequence vs Structure} \]

(randomize amino acid order in target sequence 50-200 times, calc. score for each shuffled sequence, then take average)

\[ E_{\text{relative}} = E_{\text{modify}} - E_{\text{shuffled}} \]

Performance Evaluation?

in a "Blind Test"

CASP5 Competition

(Critical Assessment of Protein Structure Prediction)

\[ \text{Given: Amino acid sequence} \]

\[ \text{Goal: Predict 3-D structure} \]

(before experimental results published)

Typical Results: (well, actually, our BEST Results):

HO = top-ranked CASP5 prediction for this target!

Target 174  PDB ID = 1MG7

Predicted Structure  Actual Structure

Overall Performance in CASP5 Contest

Ho = 8th out of ~180 (by M. Levitt, Stanford)

FR Fold Recognition

(targens manually assessed by Nick Grishin)

<table>
<thead>
<tr>
<th>Rank</th>
<th>E-Score</th>
<th>NgNW</th>
<th>NpNW</th>
<th>NgNW</th>
<th>NpNW</th>
<th>Group-name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24.26</td>
<td>9.00</td>
<td>12.00</td>
<td>9</td>
<td>12</td>
<td>Ginalski</td>
</tr>
<tr>
<td>2</td>
<td>21.64</td>
<td>7.00</td>
<td>12.00</td>
<td>7</td>
<td>12</td>
<td>Skolnick</td>
</tr>
<tr>
<td>3</td>
<td>19.55</td>
<td>8.00</td>
<td>12.50</td>
<td>9</td>
<td>14</td>
<td>Baker</td>
</tr>
<tr>
<td>4</td>
<td>16.88</td>
<td>6.00</td>
<td>10.00</td>
<td>6</td>
<td>10</td>
<td>BIOINFO.PL</td>
</tr>
<tr>
<td>5</td>
<td>15.25</td>
<td>7.00</td>
<td>7.00</td>
<td>7</td>
<td>7</td>
<td>Shortle</td>
</tr>
<tr>
<td>6</td>
<td>14.56</td>
<td>5.00</td>
<td>11.50</td>
<td>7</td>
<td>13</td>
<td>BAKER-Rosetta</td>
</tr>
<tr>
<td>7</td>
<td>13.49</td>
<td>4.00</td>
<td>11.00</td>
<td>4</td>
<td>11</td>
<td>Brooks</td>
</tr>
<tr>
<td>8</td>
<td>11.34</td>
<td>3.00</td>
<td>6.00</td>
<td>3</td>
<td>6</td>
<td>Ho-Kai-Ming</td>
</tr>
<tr>
<td>9</td>
<td>10.45</td>
<td>3.00</td>
<td>5.50</td>
<td>3</td>
<td>6</td>
<td>Jones-NewFold</td>
</tr>
</tbody>
</table>

FR NgNW = number of good predictions without weighting for multiple models

FR NpNW = number of total predictions without weighting for multiple models

Protein Structure Prediction

Servers & Software

Three basic approaches:

1) Homology modeling (need >30% sequence identity)

   PredictProtein META, SWISS-MODEL, Cn3D

2) Threading (if <30% sequence identity)

   Best? Hmm - see CASP & EVA

3) Ab initio (if no template available & many CPUs)

   Best? Rosetta (Baker) - see CASP & EVA

Protein Structure Prediction

Servers & Software

Three basic approaches:

1) Homology modeling (need >30% sequence identity)

   PredictProtein META, SWISS-MODEL, Cn3D

2) Threading (if <30% sequence identity)

   Best? Hmm - see CASP & EVA

3) Ab initio (if no template available & many CPUs)

   Best? Rosetta (Baker) - see CASP & EVA