COMP598: Advanced Computational Biology Methods & Research

Protein Fold Recognition

Jérôme Waldispühl
School of Computer Science, McGill
Comparative Modeling

• Homology modeling
  – identification of homologous proteins through sequence alignment
  – structure prediction through placing residues into “corresponding” positions of homologous structure models

• Protein threading
  – make structure prediction through identification of “good” sequence-structure fit
Only a few folds are found in nature

PROTEINS

One thousand families for the molecular biologist

*Cyrus Chothia

How many families are there? By putting together the information to be found in papers published over the past few months we can make an initial estimate, and my calculation suggests that the large majority of proteins come from no more than one thousand families.

Proteins are clustered into families. crystallography, NMR and molecular modelling will produce, at least in outline, structures for most proteins in time for the completion of the genome projects.

<table>
<thead>
<tr>
<th>Genome projects</th>
<th>Total number of genes</th>
<th>Genes related to those previously determined</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caenorhabditis elegans chromosome III (part)</td>
<td>32</td>
<td>14 (44%)</td>
<td>1</td>
</tr>
<tr>
<td>Yeast chromosome III chromosome IX (part)</td>
<td>182</td>
<td>52–66 (29–36%)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>46</td>
<td>15 (33%)</td>
<td>*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Large libraries of expressed genes</th>
<th>Total number of clones</th>
<th>Clones related to previously determined protein sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human brain</td>
<td>~1,400†</td>
<td>406 (~30%)</td>
</tr>
<tr>
<td>Caenorhabditis elegans St Louis–Cambridge NIH</td>
<td>1,517</td>
<td>512 (34%)</td>
</tr>
<tr>
<td>NIH</td>
<td>585</td>
<td>210 (36%)</td>
</tr>
</tbody>
</table>
• Only a few thousand unique folds in nature
• 90% of new structures deposited to PDB in the past three years have similar structural folds
Comparative Modeling

1. Find homologous proteins

1. Identify cores and loops: conserved segments are cores, otherwise loops

1. Core modeling: copy backbone coordinates from the homologous one with known structure

1. Loop modeling

1. Side chain modeling

1. Refinement
Homology Modeling

Query Sequence:

DRVYIHPFADRVYIHPFA

Protein sequence classification database

The Best Match

- PSI-BLAST
- HMM
- Smith-Waterman algorithm
Protein Threading

Query Sequence: DRVYIHPFADRVYIHPF
Threading Example

Protein Structure

Protein Sequence

Positions or residues in red are gaps
Protein Threading (2)

- Make a structure prediction through finding an optimal alignment (placement) of a protein sequence onto each known structure (structural template)
  - “alignment” quality is measured by some statistics-based scoring function
  - best overall “alignment” among all templates may give a structure prediction

Target sequence

MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

Template library
Fold recognition or threading

Target sequence = SHPALTQLRALRYCKEIPALDPQLLDWLLLLEDS...

Library of possible folds
(thesè have known sequences AND structures):

[Diagrams of possible folds]
Sequence-structure alignment

Target = SHPALTQLRALRYCKEIPALDPQLLDDLWLLLLDSMTKRFEEQQ...
= t_1 t_2 t_3 t_4 t_5 ... t_n

C

Sequence for known fold = s_1 s_2 s_3 s_4 s_5 ... s_n

Positions for known fold = p_1 p_2 p_3 p_4 p_5 ... p_n

Fold recognition is used when sequence identity is low.
How do you align the sequence to the structure?
Plan A

Sequence alignment

(Bowie et al., 1991)
Linking the sequence to structural properties
3D-1D comparison

- Describe the structure by a sequence of terms representing the structural environment of each residue

- How buried it is
- Polar/non-polar nature of the environment
- Local secondary structure

6 x 3 environment classes

Bowie & Eisenberg, Science (1991) 253, 164-170
Different amino acids prefer different environments

- Quantify preference of each amino-acid type for each environment using statistical preferences (log odds score)

\[ score_{ij} = \ln \left( \frac{P(j \_in \_environment \_i)}{P(j \_in \_any \_environment)} \right) \]

<table>
<thead>
<tr>
<th>environment class</th>
<th>Trp</th>
<th>Phe</th>
<th>Tyr</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1α</td>
<td>1.00</td>
<td>1.32</td>
<td>0.18</td>
</tr>
<tr>
<td>B1β</td>
<td>1.17</td>
<td>0.85</td>
<td>0.07</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Make a scoring matrix = 3D profile

<table>
<thead>
<tr>
<th>fold position</th>
<th>environ. class</th>
<th>Trp</th>
<th>Phe</th>
<th>Tyr</th>
<th>...</th>
<th>gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B1β</td>
<td>1.17</td>
<td>0.85</td>
<td>0.07</td>
<td>...</td>
<td>200</td>
</tr>
<tr>
<td>2</td>
<td>E loop</td>
<td>-2.14</td>
<td>-1.90</td>
<td>-0.94</td>
<td>...</td>
<td>2</td>
</tr>
</tbody>
</table>

and use it to align the sequence to the environment string using dynamic programming

\[
\begin{align*}
S_1 & e_1 e_2 e_3 e_4 e_5 e_6 e_7 \ldots \\
S_2 & \\
S_3 & \\
S_4 & \\
\vdots & \\
\end{align*}
\]

environment class
Fold recognition by 3D-1D

- Compare the target sequence alignment to the template against a large number of other possible sequences

\[
Z_{score} = \frac{score - <score>}{\sigma}
\]

- Z-scores > 7 represent a good match
Result from original paper: structure profile for sperm whale myoglobin

Bowie, Luty, Eisenberg
Science (1991) 253, 164-170

Fig. 6. Results of a compatibility search for the structure of sperm whale myoglobin. Myoglobin sequences are represented by black bars, other globin sequences are represented by white bars, and all other sequences are shown in gray bars. Sperm whale myoglobin is the eighth highest scoring protein ($Z$ score = 23.7). Gaps were not allowed in helical regions (as defined in the protein data bank file). In nonhelical regions, a gap-opening penalty of 2.0 and a gap-extension penalty of 0.02 was used.
Improvements to 3D-1D scoring

- Better to use more classes - this is possible now that we have a lot more structural data
- Incorporate predicted properties of the target (e.g. $2^\circ$ structure)
- H3P2 uses 5 scoring dimensions
  - 3 for the fold
    - 7 residue classes (e.g. \{C\}, \{P\}, \{G\}, \{W,F,Y\}...)
    - 3 secondary structures
    - 2 burial groups
  - 2 for the sequence
    - 7 residues classes
    - predicted secondary structure
- $7 \times 3 \times 2 \times 7 \times 3 = 882$ different elements in the scoring matrix
- Derive values for the matrix from 119 structurally similar pairs with < 30% sequence identity

Fold recognition by 3D-1D alignment

**Advantages**
- fast $O(mn)$
- incorporates structural information
- reasonable performance

**Disadvantage**
- assumes independence of positions
- assumes conservation of environment
Incorporating dependencies between positions

- Score based on a pair-wise contact potential

\[ S = \sum_{i} \sum_{j>i} \text{score}(i, j) \]

\[ \text{score}(i, j) = f(p_i, p_j, t_{r_i}, t_{r_j}) \]

- \( t_{r_j} \) is the amino acid from the target sequence that is mapped to structure position \( i \)
Knowledge-based contact potentials

- Again, use observed frequencies in the pdb to compute scores

Example
Define a contact as occurring if 2 residues are < 6 Å apart (Cα-Cα distance)

\[
\text{score}(i,j) = -\ln \left( \frac{P(i,j \mid \text{contact})}{\text{normalization}} \right)
\]

Normalization based on the expected rate of seeing i and j in contact, given no interaction energy between the two.
Knowledge-based threading potentials

- Some statistical potentials include a distance dependence

\[
\text{score}(aa_i, aa_j, r_{ij}, d_{ij}) = -\ln \left( \frac{f(aa_i, aa_j, r_{ij}, d_{ij})}{f(r_{ij}, d_{ij})} \right)
\]

\(d_{ij}\) is separation in sequence
\(r_{ij}\) is separation in space
at \(d_{ij} = 4\) compare potentials for

Benefits of contact potentials

- Fast to evaluate
- Insensitive to small differences in atomic positions
- Show good performance for a variety of applications - particularly fold recognition

Problem with contact potentials

- Don’t represent physical potentials well (compare last slide to VdW potential!)
- Hard to interpret
- Can lead to some strange artifacts. E.g. +/-, +/ and -/ are similarly good at distances > 4 Å since they are often all found on the surface; hydrophobic groups repel each other at long distances
Problem with using contact potentials for threading

- The contacts depend on the alignment
- The alignment depends on the contacts

To calculate the score for putting a residue in a certain position, you need to know what residues are in other positions. These aren’t yet determined!

Performing an alignment using a pairwise scoring function while allowing variable-length gaps is an **NP-complete** problem - it can’t be solved in polynomial time. Note similarity to problem of finding common 3D sub-structures.
What to do?

- Put limits on gap lengths and positions (e.g. don’t allow gaps in core secondary structure elements)
- Use heuristics
  
  Example: in the “frozen” approximation you first use the target sequence to compute the scores at each position.
  
  In subsequent iterative rounds you use the residue that was there in the last round of alignment.

Plan B

Using linear programming techniques

(Raptor: Xu et al., 2003)
Threading Model

- Each template is parsed as a chain of cores. Two adjacent cores are connected by a loop. Cores are the most conserved segments in a protein.
- No gap allowed within a core.
- Only the pairwise contact between two core residues are considered because contacts involved with loop residues are not conserved well.
- Global alignment employed
Scoring Function

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

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

alignment gap penalty: $E_g$
(gap 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
Scoring: Fitness Score

\[ \text{FitnessScore}(a, s) = -\log \frac{P(a,s)}{P(a)P(s)} \]

- occurring probability of amino acid a with s
- occurring probability of amino acid a
- occurring probability of solvent accessibility s
Scoring: Pairwise Potential

\[ \text{PairwisePotential}(a, b) = - \log \frac{P(a,b)}{P(a)P(b)} \]

occurring probability of a and b with distance < cutoff

occurring probability of amino acid a

occurring probability of amino acid b
Contact Graph

1. Each residue as a vertex
2. One edge between two residues if their spatial distance is within given cutoff.
3. Cores are the most conserved segments in the template
Simplified Contact Graph

Original Contact Graph

No gap allowed within cores

Simplified Contact Graph
Alignment Example

Original Contact Graph

No gap allowed within cores

Simplified Contact Graph

Sequence
Alignment Example

Original Contact Graph

No gap allowed within cores

Simplified Contact Graph

Sequence
Calculation of Alignment Score

Alignment Score = Singleton Score + Pairwise Score + Gap Penalty

Singleton Score \((S1,T1)\) = Mutation Score \((S1,T1)\) + Fitness Score \((S1,T1)\) + SS \((S1,T1)\)

Pairwise Score = Pairwise Score \((S1, S2, \text{dist}(T1,T2))\) + ...

Filled small circles are unaligned template positions or sequence residues
Linear & Integer Program

maximize

\[ z = 6x + 5y \]

Subject to

\[ 3x + y \leq 11 \]
\[ -x + 2y \leq 5 \]
\[ x, y \geq 0 \]

\[ x, y \text{ integer} \]
• $x(i, l)$ denotes core $i$ is aligned to sequence position $l$
• $y(i, l, j, k)$ denotes that core $i$ is aligned to position $l$ and core $j$ is aligned to position $k$ at the same time.
Minimize

\[ E = \sum a_{i,l} x_{i,l} + \sum b_{(i,l)(j,k)} y_{(i,l)(j,k)} \]

s.t.

\[ x_{i,l} = \sum_{k \in R[i,j,l]} y_{(i,l)(j,k)}, \forall l \in D[i] \]

\[ x_{j,k} = \sum_{l \in R[j,k,i]} y_{(i,l)(j,k)}, \forall k \in D[j] \]

\[ \sum_{l \in D[i]} x_{i,l} = 1 \]

\[ x_{i,l}, y_{(i,l)(j,k)} \in \{0,1\} \]

a: singleton score parameter
b: pairwise score parameter

Each y variable is 1 if and only if its two x variable are 1

Each core has only one alignment position
Template Ranking

• Alignment score
  – residue composition bias
  – template length bias

• Z-score (S.H. Bryant et al., 1995)
  – statistical test, cancel out bias
  – time-consuming to calculate, sequences must be shuffled and threaded many times (~100 times)

• Classification-based methods
  – A threading pair is positive if they have similar structures
  – Noise caused by bad alignments

• Regression-based method
  – Predict alignment accuracy
  – Rank templates based on predicted accuracy
Feature Extraction

- topology of a predicted structure
  - sizes
  - # pairwise contacts
  - independent domain?
- sequence-template alignment
  - alignment scores
  - sequence identify
  - # gaps

Features
CASP6 Example

T0268_1 (PDB id: 1wg8)
MaxSub score: 0.9
Sequence identity: 50%

T0267 (PDB id: 1wk4)
MaxSub score: 0.6
Sequence identity: 19%
CASP6 Example

T0224 (PDB id: 1rHX or 1x9a)
  MaxSub score: 0.5
  Sequence identity: 15%

T0228_1 (PDB id: 1vlp)
  MaxSub score: 0.3
  Sequence identity: 8%
CASP6 Example

T0238 (PDB id: 1w33)
MaxSub score: 0.2
Sequence identity: 9%

T0242 (PDB id: 2blk)
MaxSub score: 0.17
Sequence identity: 10%