Mining Residue Contacts in Proteins using Local Structure Predictions

Mohammed J. Zaki
Computer Science Dept
http://www.cs.rpi.edu/~zaki

with Chris Bystroff (Biology, RPI)
Talk Outline

- DNA and Protein Structure
- Hybrid Mining Approach
- Experiments
Building Blocks of Biological Systems:

*nucleotides and amino acids*

**DNA** (nucleotides, 4 types): information carrier/encoder.

**RNA**: bridge from DNA to protein.

**Protein** (amino acids, 20 types): action molecules.
Four types of nucleic acids of DNA

Deoxyribonucleic Acid (DNA) contains four nucleotide bases.

The Nucleotides of DNA

![Chemical structures of DNA nucleotides]

Adenine | Guanosine | Thymine | Cytosine

Purines | Pyrimidines

Note that A pairs with T; and G pairs with C.
Primary Structure of DNA

- Unbranched polymer
- Sequence of nucleotide bases
- Double stranded

atgaatcgta ggggttttgaacgctgggcaaat
acgatgacattctcaagcggacaacattggacgcac
ggcagctggaaggccgtctcgcaggggctgga……
Structure of DNA
Processes

• Replication of DNA
• Transcription of gene (DNA) to messenger RNA (mRNA)
• Translation of mRNA into proteins
• Folding of proteins into 3D from
• Biochemical or structural functions of proteins
Transcription

Strand being transcribed

mRNA

5' 3'

5' 3'
Translation: Universal Genetic Code

- Translation form nucleotide code to amino acid code.

```
atgaatcgta ggggttttgaa cgctggcaaat
acgatgacctt ctcaagcgaa cattgacgac
ggcagctgga aggccgtctc cgagggcgga ..... 
```

`MNRRGLNAGNTMTSQANIDDGSWKAVSEGG ...`
Sequence of Amino Acids: Protein

- Unbranched polymer
- Peptide backbone
- Twenty side chain types
- 3D structure the key
Amino Acid
Amino acids with hydrophobic side groups:
- Phenylalanine (Phe)
- Methionine (Met)
- Isoleucine (Ile)
- Leucine (Leu)
- Valine (Val)

Amino acids with hydrophilic side groups:
- Glutamic acid (Glu)
- Aspartic acid (Asp)
- Arginine (Arg)
- Lysine (Lys)
- Histidine (His)

Amino acids that are in between:
- Alanine (Ala)
- Glycine (Gly)
- Cysteine (Cys)
- Serine (Ser)
- Threonine (Thr)
- Proline (Pro)
Polypeptide Chain
Torsion Angles

Peptide torsion angles.
Protein Primary Structure

- Unbranched polymer
- 20 side chains

MNRRGLNAGNTMTSQANIDDGSWKAVSEGG ...
3D: Higher Order Protein Structures

• Higher order structures
  – Secondary: local in sequence
  – Tertiary: 3D fold of one polypeptide chain
  – Quaternary: Chains packing together
Protein Folding

• Proteins the action molecules of life
• 3D Structure critical to function
• Many protein sequences fold spontaneously to native 3D fold
Right-handed alpha-helix.

White dots show the hydrogen bonds.
Antiparallel Beta-Sheet
The Protein Folding Problem
An HMM for Local Predictions

- HMMSTR (Chris Bystroff, Biology, RPI)
- Build a library of short sequences that tens to fold uniquely across protein families: the I-Sites Library
- Treat each motif as a Markov chain
- Merge the motifs into a global HMM for local structure prediction
I-Sites Motifs (Initiation Sites)

Helix C-Cap

Beta to Alpha

Beta Hairpin
What is association mining?

- Given a set of items/attributes, and a set of objects containing a subset of the items
- Find rules: if I1 then I2 (sup, conf)
- I1, I2 are sets of items
- I1, I2 have sufficient support: P(I1∪I2)
- Rule has sufficient confidence: P(I2|I1)
### Example

<table>
<thead>
<tr>
<th>OID</th>
<th>Items</th>
<th>Support (%)</th>
<th>Itemsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A C T W</td>
<td>100% (6)</td>
<td>C</td>
</tr>
<tr>
<td>2</td>
<td>CD W</td>
<td>83% (5)</td>
<td>CW</td>
</tr>
<tr>
<td>3</td>
<td>A C T W</td>
<td>67% (4)</td>
<td>A, D, T, AC, AW, CD, CT, ACW</td>
</tr>
<tr>
<td>4</td>
<td>A C D W</td>
<td>50% (3)</td>
<td>AT, DW, TW, ACT, ATW, CDW, CTW, ACTW</td>
</tr>
<tr>
<td>5</td>
<td>A C D T W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>C D T</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Maximal Frequent Itemsets
ACTW, CDW
Contact Map

• Amino acids Ai and Aj are in contact if their 3D distance is less than threshold (7å)
• Sequence separation is given as |i-j|
• Contact map C is an N x N matrix, where
  – C(i,j) = 1 if Ai and Aj are in contact
  – C(i,j) = 0 otherwise
• Consider all pairs with |i-j| >= 4
Contact Map (2igd PDB)
Training the HMM

• Build I-sites Library
  – Short sequence motifs (3 to 19)
  – Exhaustive clustering of sequences
  – Non-redundant PDB dataset (< 25% similarity)

• Build an HMM
  – Each of 262 motifs is a chain of Markov states
  – Each state has seq and structure for one position
  – Merge I-sites motifs hierarchically to get one global HMM for all the motifs
HMM Output

- Total of 282 States in the HMM
- Each state produces or “emits”:
  - Amino acid profile (20 probability values)
  - Secondary structure (D) (helix, strand or loop)
  - Backbone angles (R) (11 dihedral angle symbols)
  - Finer structural context (C) (10 context symbols)
Data Format and Preparation

• Take the 794 PDB proteins

• Compute optimal alignment to HMM
  – Find best state sequence for the observed acids
  – Output probability distribution of a residue over all the 282 HMM states

• Integrate the 3 datasets
  – Alignment probability distribution (Nx282)
  – Amino acid and context information (D, R, C)
  – Contact map (NxN)
Classification Problem

- A pair of amino acids \((A_i, A_j)\) is an instance
- The class: C or NC (contact or non-contact)
- Highly skewed class distribution
  - \((1.7\% \text{ C and } 98.3\% \text{ NC}; 300K \text{ C vs } 17,3M \text{ NC})\)
- Features for each instance
  - \(A_i, A_j, D_i, D_j, R_i, R_j, C_i, C_j\)
  - Profile: \(p_i1, p_i2, \ldots, p_{120}, p_j1, p_j2, \ldots, p_{j20}\)
  - HMM States: \(q_i1, q_i2, \ldots, q_{j282}, q_j1, q_j2, \ldots, Q_j282\)
- Class: C or NC
Predicting Protein Contacts

- Predict contacts for new sequence

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>D</th>
<th>T</th>
<th>S</th>
<th>K</th>
<th>C</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ai</th>
<th>Aj</th>
<th>F1</th>
<th>Fn</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>T</td>
<td>..</td>
<td>..</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>..</td>
<td>..</td>
</tr>
<tr>
<td>D</td>
<td>S</td>
<td>..</td>
<td>..</td>
</tr>
<tr>
<td>D</td>
<td>P</td>
<td>..</td>
<td>..</td>
</tr>
<tr>
<td>T</td>
<td>S</td>
<td>..</td>
<td>..</td>
</tr>
<tr>
<td>S</td>
<td>C</td>
<td>..</td>
<td>..</td>
</tr>
<tr>
<td>K</td>
<td>P</td>
<td>..</td>
<td>..</td>
</tr>
</tbody>
</table>
Association Mining

• Mine (closed) Frequent Itemsets in C data
• Compute Frequency in NC dataset
• Prune mined results
  – compute probability of occurrence in C and NC
  – prune pattern if ratio of contact to non-contact probability is less than some threshold
  – i.e., keep only the patterns highly predictive of contacts
Testing Phase

- 90-10 split into training and testing
- 2.4 million pairs, with 36K contacts (1.5%)

Evidence calculation:
- Compute matching patterns for each instance
- Compute cumulative frequency in C and NC
- Compute evidence: ratio of freq in C over NC

Prediction:
- Sort all test examples based on evidence
- Predict the top PR fraction of examples as contacts
Experiments

• 794 Proteins from Protein Data Bank
• Distinct structures (< 25% similarity)
• Longest: 907, Smallest: 35 amino acids
• 90-10 split for training-testing
• Total pairs: 20 million (> 2.5 GB)
• Contacts: 330 thousand (1.6%)
• Highly uneven class distribution
Evaluation Metrics

- \( Na \): set of all pairs
- \( Na^* \): all pairs with positive evidence
- \( Ntc \): true contacts in test data
- \( Ntc^* \): true contacts with positive evidence
- \( Npc \): predicted contacts
- \( Ntpc \): correctly predicted contacts
- Accuracy = \( \frac{Ntpc}{Npc} \)
- Coverage = \( \frac{Ntpc}{Ntc} \)
- Prediction Ratio (PR) = \( \frac{Ntc^*}{Na^*} \)
- Random Predictor Accuracy = \( \frac{Ntc}{Na} \)
Results (Amino Acids Only)

Crossover: 7% accuracy and 7% coverage; 2 times over Random
1-100: 12% accuracy (A) and coverage (C); 100-170: 6% A and C
170-300: 4.5% A and C; 300+: 2% A and C
HMM + Amino Acids

Crossover: 19% accuracy and 19% coverage; 5 times over Random
1-100: 27% accuracy (A) and coverage (C); 100-170: 22% A and 10% C
170-300: 13% A and 8% C; 300+: 10% A and 8% C
HMM + AA + R,D,C

Crossover: 17% accuracy and 17% coverage; 5 times over Random
1-100: 30% accuracy(A) and coverage (C); 100-170: 17% A and C
170-300: 10% A and C; 300+: 6% A and C
Predicted Contact Map (2igd)
Summary

• Challenging prediction problem
• In essence, we have to predict a contact matrix for a new protein
• Hybrid HMM/Associations approach
• Best results to-date: 19% A and C
  – 14.4% Accuracy (Fariselli, Casadio ‘99; NN)
  – 13% Accuracy (Thomas et al ‘96)
  – Short proteins: 26% (vs. our 30%) (Olmea, Valencia, ‘97)