Protein Domains: Prediction from Amino Acid Sequence

I690/B680: Structural Bioinformatics

February 15, 2006
Introduction

• protein domain identification, given only a sequence, is an important problem
  – gap between sequence and structure data is increasing

• Usual steps given a new sequence
  – if homology can be identified by sequence alignment (to domain databases) the problem is typically easy
  – if homology cannot be reliably identified, the problem is much more difficult

• Databases to look at: Pfam, SMART etc.
Practical Importance

- cost-reduction of experimental methods
  - X-ray crystallography
  - NMR spectroscopy

- to help local multiple sequence alignment
  - difficulties arise when whole sequences with “shuffled” domains need to be aligned

- to help fold recognition methods
  - it is easier to loot at a sequence in terms of its constituent domains then as a whole chain
DomSSEA: idea

- alignment of (predicted) secondary structures can be used for sequence alignment when identity is low
- prediction of secondary structure is sufficiently accurate to be useful (77% or higher; 3-state)

- Idea:
  - use Secondary Structure Element Alignment (SSEA)
  - it has been used for fold recognition when no detectable homology existed (using standard sequence alignments)

- Goal:
  - to have a fast method capable of predicting on whole genomes
  - result: DomSSEA
Analysis: Domain Lengths
Analysis: Domain Lengths

=> Domains can be predicted using the size of proteins
DomSSEA: algorithm

- predict secondary structure on a new (query) protein
  - PSIPRED (Jones, 1999)

- align query sequence of predicted secondary structures to all sequences (of domains) in the database
  - dynamic programming (McGuffin et al, 2001)
  - adjusted scoring scheme (Przytycka et al, 1999)

- collect top hit from the pair with the highest score
  - use PSI-BLAST to find homologs of the best hit

- make a cut within a coil region
Accuracy

Standard measures

• accuracy in domain number prediction

• accuracy of domain boundary prediction
  – ±20 residues from the CATH assignment ⇒ correct

Table 1. Prediction of one, two, or three or more domain chains

<table>
<thead>
<tr>
<th>Prediction of number of domains</th>
<th>% Correctly assigned</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
</tr>
<tr>
<td>PUU</td>
<td>79.0</td>
</tr>
<tr>
<td>DomSSEA observed secondary structure</td>
<td>75.4</td>
</tr>
<tr>
<td>DomSSEA predicted secondary structure</td>
<td>73.3</td>
</tr>
<tr>
<td>DGS-M</td>
<td>76.7</td>
</tr>
<tr>
<td>DGS-W</td>
<td>76.7</td>
</tr>
<tr>
<td>Absolute difference in length</td>
<td>66.2</td>
</tr>
<tr>
<td>Fasta</td>
<td>60.9</td>
</tr>
<tr>
<td>Random (weighted)</td>
<td>61.4</td>
</tr>
<tr>
<td>Random (basic)</td>
<td>37.9</td>
</tr>
<tr>
<td>Sum of squares</td>
<td>62.0</td>
</tr>
</tbody>
</table>

The percentage of chains given a correctly assigned domain number (top prediction), for single, two, and three or more domain chains, as well as for all chains in the representative set.
Accuracy

<table>
<thead>
<tr>
<th>Methods</th>
<th>% Correctly assigned boundaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUU</td>
<td>81.8</td>
</tr>
<tr>
<td>Consensus</td>
<td>52.5</td>
</tr>
<tr>
<td>L/(N-1)</td>
<td>49.5</td>
</tr>
<tr>
<td>DomSSEA observed secondary structure</td>
<td>49.5</td>
</tr>
<tr>
<td>DomSSEA predicted secondary structure</td>
<td>49.0</td>
</tr>
<tr>
<td>DGS-M</td>
<td>46.0</td>
</tr>
<tr>
<td>Absolute difference in length</td>
<td>44.6</td>
</tr>
<tr>
<td>DGS-W</td>
<td>37.1</td>
</tr>
<tr>
<td>FASTA</td>
<td>30.0</td>
</tr>
<tr>
<td>Random (weighted)</td>
<td>26.8</td>
</tr>
</tbody>
</table>
Some Other Methods

- **SnapDRAGON**
  - George and Heringa, 2002
  - *ab initio* protein folding method: DRAGON
    - predicts secondary structure, makes sequence alignment, then folds each chain based on conserved hydrophobicity
    - each folded chain is evaluated for domain boundaries
    - large number of structures are evaluated for domain consistency
  - 72% accuracy for # of domains; 51% overall cut prediction

- **CHOPnet**
  - Liu and Rost, 2004
  - uses aa composition, many predicted features, sequence alignments and combines them using neural networks
  - 69% accuracy for # of domains; 50% cut prediction (for 2 domain) chains