Introduction to Bioinformatics

Iosif Vaisman

Email: ivaisman@gmu.edu

Multiple alignment

**Computational complexity**

Alignment of protein sequences with 200 amino acid residues:

<table>
<thead>
<tr>
<th># of sequences</th>
<th>CPU time</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1 sec</td>
</tr>
<tr>
<td>3</td>
<td>200 sec</td>
</tr>
<tr>
<td>10</td>
<td>$200^8$ sec</td>
</tr>
</tbody>
</table>

Multiple sequence alignment

Given $k$ ($k > 2$) sequences, $s_1, \ldots, s_k$, each sequence consisting of characters from an alphabet $\mathcal{A}$, a **multiple alignment** is a rectangular array, consisting of characters from the alphabet $\mathcal{A}' (\mathcal{A} + ".")$, that satisfies the following 3 conditions:

1. There are exactly $k$ rows.
2. Ignoring the gap character, row number $i$ is exactly the sequence $s_i$.
3. Each column contains at least one character different from ".".

Consensus

**Consensus sequence** - idealized sequence in which each position represents the amino acid most often found when many sequences are compared.

- **Plurality** - minimum number of votes for a consensus
- **Threshold** - scoring matrix value below which a symbol may not vote for a coalition.
- **Sensitivity** - minimum score to select consensus
- **Profiles** - blocks of prealigned sequences

Column cost: the sum of costs for all possible pairs
Multiple alignment algorithm

1. Pairwise alignments (progressive pairwise alignments)
2. Distance matrix calculation
3. Guide tree creation (hierarchical clustering)
4. New sequence addition

Scoring system (distances)

\[ D(ij) = -\ln \frac{S_{\text{real}}(ij) - S_{\text{rand}}(ij)}{S_{\text{real}}(i) - S_{\text{rand}}(i)} \times 100 \]

- \( S_{\text{real}}(ij) \) - observed similarity score for two aligned sequences \( i \) and \( j \)
- \( S_{\text{ideal}}(ij) \) - average of the two scores for each sequence aligned with itself
- \( S_{\text{rand}}(ij) \) - average score determined from 100 global randomizations of the two sequences

The distances \( D(ij) \) are used to generate the distance matrix from which the approximate guide tree is generated.

Segment - line joining two vertices
Each unit m-dimensional cube in the lattice contains \( 2^m - 1 \) segments

Alignment Path for 3 Sequences
(0,0,0), (1,0,0), (2,1,0), (3,2,0), (3,3,1), (4,3,2)

Pairwise Projections of the Alignment
Alignment statistics

<table>
<thead>
<tr>
<th></th>
<th>Rabctp</th>
<th>Ratlpb</th>
<th>Humcetp</th>
<th>Macctep</th>
<th>Humlpb</th>
<th>Bovbpi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>478</td>
<td>67%</td>
<td>65%</td>
<td>19%</td>
<td>19%</td>
<td>18%</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>82%</td>
<td>80%</td>
<td>39%</td>
<td>39%</td>
<td>36%</td>
</tr>
<tr>
<td>2</td>
<td>327</td>
<td>58%</td>
<td>16%</td>
<td>16%</td>
<td>16%</td>
<td>39%</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>75%</td>
<td>38%</td>
<td>38%</td>
<td>35%</td>
<td>62%</td>
</tr>
<tr>
<td>3</td>
<td>318</td>
<td>284</td>
<td>18%</td>
<td>18%</td>
<td>17%</td>
<td>40%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>367</td>
<td>0</td>
<td>38%</td>
<td>38%</td>
<td>35%</td>
</tr>
</tbody>
</table>

1: The number of residues that match exactly (identical residues) between the two sequences.
2: The number of residues whose juxtaposition yields a greater than zero score in the current scoring table (certain residues are considered identical).
3: The number of residues aligned up with a gap character.

Alignment score

<table>
<thead>
<tr>
<th></th>
<th>Rabctp</th>
<th>Ratlpb</th>
<th>Humcetp</th>
<th>Macctep</th>
<th>Humlpb</th>
<th>Bovbpi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4777</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5358</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5323</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8103</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>8109</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>8547</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>6474</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6392</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Alignment visualization (tree)

1: Rabctp
2: Humcetp
3: Macctep
4: Ratlpb
5: Humlpb
6: Rabctpb
7: Humbpi
8: Bovbpi

Sequence Logos: a quantitative graphical display for binding sites and proteins

Sequence Logos

Multiple Alignment Programs

- **Pileup (GCG):** Needleman and Wunsch algorithm for pairwise alignment and UPGMA method for tree construction.
- **CLUSTAL:** Wilbur and Lipman algorithm for pairwise alignment (CABIOS 8:189, 1992).
- **PIMA:** Pattern-matching based algorithm (PNAS 87:118, 1990).
- **TreeAlign:** Phylogenetic algorithm (Meth. Enzymol. 18:626, 1990).