Introduction to Bioinformatics for Medical Research

Gideon Greenspan
gdg@cs.technion.ac.il

Lecture 8
Multiple Alignment and Phylogeny
Multiple Alignment & Phylogeny

- Multiple Alignment
  - Scoring
  - Complexity
- Feasible multiple alignment
  - ClustalW
- Phylogenetic Trees
  - Tree reconstruction
  - Phylodendron
Multiple Alignment

• Like pairwise alignment
  – $n$ input sequences instead of 2
  – Add indels to make same length
  – Local and global alignments

• Score columns in alignment
  – Positive score for agreement
  – Negative for disagreement or indels

• Seek an alignment to maximize score
Alignment Example

Same allele in 3 or 4

Lone allele

1 | 23
GTCGTA\text{GTCGAC} | 1 | 15
GTCN\text{TAGCGAGC} | 2 | 6
GCGA\text{AGAGCGAGC} | 3 | 39
GCCGTCG\text{CGTCGTAAC} | 4 | 16

2 | 30
GTC\text{TAGCGAGCGTGAT} | 2 | 6
GCGA\text{AGAGCGAGC} | 3 | 6
GCCGTCG\text{CGTCGTAAC} | 4 | 4

3 | 6
GCGA\text{AGAGCGAGC} | 3 | 6
GCGA\text{AGAGCGAGC} | 3 | 6
GCCGTCG\text{CGTCGTAAC} | 4 | 4

4 | 4
GCGA\text{AGAGCGAGC} | 4 | 4
GCGA\text{AGAGCGAGC} | 4 | 4
GCCGTCG\text{CGTCGTAAC} | 4 | 4
Motivation

- **Determine consensus sequence**
  - Ignore variations, e.g. for genome sequencing
  - Identify variations, e.g. for SNP mapping

- **Build sequence families**
  - Orthologous genes, protein groups

- **Build phylogenetic trees**
  - Infer evolutionary relationship
Scoring (1)

- Sum of pairwise alignment scores
  - For $n$ sequences, there are $\frac{n(n-1)}{2}$ pairs

$$\text{GTCTAGTCG} - \text{GC} - \text{TCGAC}$$

$$\text{GTC} - \text{TAG} - \text{CGAGCGT} - \text{GAT}$$

$$\text{GC} - \text{GAAG} - \text{AG} - \text{GCG} - \text{AG} - \text{C}$$

$$\text{GCCGTCG} - \text{CG} - \text{TCGTA} - \text{AC}$$
Scoring (2)

• Sum of column scores
  – Minus number of different values
    • Example: −1 , −2 , −2 , −3
  – Minus number not in consensus
    • Example: 0 , −1 , −2 , −2
  – Maximum profile probability
    • Example: 1.0 , 0.1 , 0.06 , 0.02
Complexity (1)

- **Pairwise A–B alignment table**
  - Cell \((i,j)\) = score of best alignment between first \(i\) elements of A and first \(j\) elements of B
  - Complexity: length of A \(\times\) length of B

- **3-way A–B–C alignment table**
  - Cell \((i,j,k)\) = score of best alignment between first \(i\) elements of A, first \(j\) of B, first \(k\) of C
  - Complexity: length A \(\times\) length B \(\times\) length C
Complexity (2)

- \( n \)-way \( S_1 \rightarrow S_2 \rightarrow \ldots \rightarrow S_{n-1} \rightarrow S_n \) alignment table
  - Cell \((x_1, \ldots, x_n)\) = best alignment score between first \( x_1 \) elements of \( S_1 \), \( \ldots \), \( x_n \) elements of \( S_n \)
  - Complexity: length \( S_1 \) \( \cdots \) \( \cdots \) length \( S_n \)

- Example: protein family alignment
  - 100 proteins, 1000 amino acids each
  - Complexity: \( 10^{300} \) table cells
  - Calculation time: beyond the big crunch!
Feasible Approach

• Based on pairwise alignment scores
  – Build $n$ by $n$ table of pairwise scores
• Align similar sequences first
  – After alignment, consider as single sequence
  – Continue aligning with further sequences
• Complexity is $n^2 \cdot l^2$
  – 100 proteins of 1000 amino acids: $\sim 5,000$ alignments of $10^6$ cells $\sim 5 \cdot 10^9$ operations
Example (1)

1. GTCGTAGTGGCTCGAC
2. GTCTAGCGAGCGTGAT
3. GCGAAGAGGGCGAGC
4. GCCGTCGCGGTCGTAAC

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>
Example (2)

1. GTCGTA\textcolor{red}{G}T\textcolor{green}{C}G\textcolor{red}{G}C\textcolor{green}{C}TCGAC
2. GTC\textcolor{green}{G}T\textcolor{red}{A}G\textcolor{green}{C}G\textcolor{red}{C}G\textcolor{green}{C}G\textcolor{red}{T}\textcolor{green}{G}T\textcolor{red}{G}A\textcolor{green}{T}
3. GC\textcolor{red}{G}AAG\textcolor{red}{A}G\textcolor{red}{G}C\textcolor{red}{G}A\textcolor{red}{G}C
4. GCCGTCGCGGTCGTA\textcolor{red}{A}AC

\begin{array}{ccc}
3 & 4 \\
12 & -3 & -1 \\
3 & 2 \\
\end{array}
Example (3)

1. GTCGTAGTCG–GC–TCGAC
2. GTC–TAG–CGAGCGT–GAT
3. GC–GAAGAGGGCG–AGC
4. GCCGTCGCCTCGTAAAC

1. GTCGTA–GTCG–GC–TCGAC
2. GTC–TA–G–CGAGCGT–GAT
ClustalW Input

- Fast alignment?
- Alignment format
- Alignment options
- Gap scoring
- Phylogenetic trees
- Scoring matrix
- Input sequences
ClustalW Output (1)

Sequence format is Pearson
Sequence 1: CFTR_BOVIN    1481 aa
Sequence 2: CFTR_HUMAN    1480 aa
Sequence 3: CFTR_MOUSE    1476 aa
Sequence 4: CFTR_RABIT    1450 aa
Sequence 5: CFTR_SHEEP    1481 aa
Sequence 6: CFTR_SQUAC    1492 aa
Sequence 7: CFTR_XENLA    1485 aa

Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score: 90
Sequences (1:3) Aligned. Score: 77
Sequences (1:4) Aligned. Score: 88
Sequences (1:5) Aligned. Score: 97
Sequences (1:6) Aligned. Score: 69
Sequences (1:7) Aligned. Score: 75
Sequences (2:3) Aligned. Score: 78
Sequences (2:4) Aligned. Score: 92
Sequences (2:5) Aligned. Score: 90
Sequences (2:6) Aligned. Score: 71
Sequences (2:7) Aligned. Score: 77
Sequences (3:4) Aligned. Score: 78
Sequences (3:5) Aligned. Score: 77
Sequences (3:6) Aligned. Score: 64
Sequences (3:7) Aligned. Score: 70
Sequences (4:5) Aligned. Score: 89
Sequences (4:6) Aligned. Score: 70
Sequences (4:7) Aligned. Score: 76
Sequences (5:6) Aligned. Score: 69
Sequences (5:7) Aligned. Score: 76
Sequences (6:7) Aligned. Score: 68

Guide tree file created: [/ebi/extserv/old-work/clustalw-20030614-08620820.dnd]

Start of Multiple Alignment
There are 6 groups
Aligning...
Group 1: Sequences:  2  Score: 31747
Group 2: Sequences:  2  Score: 30344
Group 3: Sequences:  4  Score: 30056
Group 4: Sequences:  5  Score: 28656
Group 5: Sequences:  6  Score: 28090
Group 6: Sequences:  7  Score: 27936
Alignment Score 152319
CLUSTAL-Alignment file created  [/ebi/extserv/old-work/clustalw-20030614-08620820.aln]
ClustalW Output (2)

**Sequence names**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Sequence</th>
<th>Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFTR_BOVIN</td>
<td>MQRSPLEKASVSVSKVFFSWTRPLKKGYRQRLELSIYHSSDSADNLSEKLREWDR</td>
<td>60</td>
</tr>
<tr>
<td>CFTR_SHEEP</td>
<td>MQRSPLEKASVSVSKLFFSWTRPLKKGYRQRLELSIYHSSDSADNLSEKLREWDR</td>
<td>60</td>
</tr>
<tr>
<td>CFTR_HUMAN</td>
<td>MQRSPLEKASVSVSKLFFSWTRPLKKGYRQRLELSIYHSSDSADNLSEKLREWDR</td>
<td>60</td>
</tr>
<tr>
<td>CFTR_RABIT</td>
<td>MQRSPLEKAGVLSKLFSSWTRPLKKGYRQRLELSIYQIPSADSDNLSEKLREWDR</td>
<td>60</td>
</tr>
<tr>
<td>CFTR_MOUSE</td>
<td>MQKSKLEFKASIFSKLFFSWTPILKKGYRHHLELSIYQAPSADSDHLSKLREWDR</td>
<td>60</td>
</tr>
<tr>
<td>CFTR_XENLA</td>
<td>MQKTPLEKASIFQSIFSSFSTKPPILKGYRQRELSDIYQIPDSDADNLSEKLREWDR</td>
<td>60</td>
</tr>
<tr>
<td>CFTR_SQUAC</td>
<td>MQRSPEKANAFSLLFFSWRPPILKGYRQKLLELSIYQIPDSDADELSEKLERWDR</td>
<td>60</td>
</tr>
</tbody>
</table>

**Sequence positions**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Sequence</th>
<th>Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFTR_BOVIN</td>
<td>LA-SKKNPKILALRRCFWRMF YGIIYLYGEVTAVQPLLGGRIIASYDPDKVERSI</td>
<td>119</td>
</tr>
<tr>
<td>CFTR_SHEEP</td>
<td>LA-SKKNPKILALRRCFWRMF YGIIYLYGEVTAVQPLLGGRIIASYDPDKVERSI</td>
<td>119</td>
</tr>
<tr>
<td>CFTR_HUMAN</td>
<td>LA-SKKNPKILALRRCFWRMF YGIIYLYGEVTAVQPLLGGRIIASYDPDKVERSI</td>
<td>119</td>
</tr>
<tr>
<td>CFTR_RABIT</td>
<td>LA-SKKNPQILALRRCFWRFLGYIILYGEVTAVQPLLGGRIIASYDPDKVERSI</td>
<td>119</td>
</tr>
<tr>
<td>CFTR_MOUSE</td>
<td>QA-SKKNPQILALRRCFWRFLGYIILYGEVTAVQPLLGGRIIASYDPDKVERSI</td>
<td>119</td>
</tr>
<tr>
<td>CFTR_XENLA</td>
<td>VATSKKPNKILALRRCFWKFLFYGIIYLYGEVTAVQPLLGGRIIASYDPDKVERSI</td>
<td>119</td>
</tr>
<tr>
<td>CFTR_SQUAC</td>
<td>LATSKKPNKVLNALRRCFWRFLGYIILYFVEFATAVQPLLGGRIIASYDKKTYEREI</td>
<td>120</td>
</tr>
</tbody>
</table>

**Match strength in decreasing order:** * : .
Phylogenetic Trees

• Represent closeness between many entities
  – In our case, genomic or protein sequences

- Observed entity
- Distance representation
- Unobserved commonality
  - chimp
  - monkey
  - human
Rooting Trees

- A tree can be hung from a root
  - Adds directional information
  - Requires addition of ‘outgroup’

So we hang the tree from where it joins

We know this is furthest

pig  monkey  human  chimp
Phylogeny and Evolution

Evolutionary Time

Common Ancestor

Speciation

Number of mutations

Evolutionary Time
Tree Reconstruction

- Build tree based on organism sequences
- Distance-based methods
  - Use pairwise alignment scores to build tree
  - Ignores sequences after initial alignments
- Character-based methods
  - Learn a tree with intermediate sequences that minimizes total number of mutations
  - Slower but generally better results
Distance-based Example (1)

```
<table>
<thead>
<tr>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>
```

```
1 2

3 4
```
Distance-based Example (2)
Distance-based Example (3)
Newick Tree Format

(CFTR_SHEEP:0.01457,
(CFTR_HUMAN:0.16153,
(CFTR_MOUSE:0.70599,
(CFTR_RABIT:2.76042,
(CFTR_SQUAC:1.27192,
CFTR_XENLA:0.28818)
):3.42183)
:0.77076)
:0.65873)
:0.73937,
CFTR_BOVIN:0.00953);}
Phylodendron Input

Tree styles
- tree diag
- cladogram
- phenogram
- eurogram
- curvogram
- smoopogram

Graphical style

Newick tree description

Tree data (newick / nh format)
Upload tree file: [Browse...]
or paste data or URL in box below

Title:
See sample data 1 and data 2

Submit  Clear

Extra options
- Format: PDF (with hyperlinks)
- width: 500  height: 500 (pixels)
- For image maps, make hyperlinks to labels
- Base URL for labels (URL's in node comments will be hyperlinked)
  http://iubio.bio.indiana.edu/.bin/genbankq.html
- Font Helvetica  style plain  size 10

Tree size

Orientation
- horizontal
- vertical
- regular

Tree growth
- use node lengths

Node position
- intermediate
- centered
- V shaped
- weighted
- inner