Distance based phylogeny reconstruction

Usman Roshan
Distance-based methods

TRUE TREE

DNA SEQUENCES
S_1 ACAAATTAGAAC
S_2 ACCCTTGAAC
S_3 ACCATTCGAAC
S_4 ACCAGACCAAC

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES

METHODOLOGY SUCH AS NEIGHBOR JOINING

INFERRED TREE

DISTANCE MATRIX

<table>
<thead>
<tr>
<th></th>
<th>S_1</th>
<th>S_2</th>
<th>S_3</th>
<th>S_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_1</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>S_2</td>
<td></td>
<td>0</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>S_3</td>
<td></td>
<td></td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>S_4</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Evolutionary distances

- We will use the Jukes Cantor model

\[ d(x, y) = -\frac{3}{4} \ln(1 - \frac{4}{3} p(x, y)) \]

- \( x \) and \( y \) are DNA/protein/RNA sequences
- \( p(x, y) \) is the number of mismatches in the alignment divided by the length of the alignment. This is also called the normalized Hamming distance.