FASTSP: linear time calculation of alignment accuracy

Siavash Mir arabbaygi
Research Preparation Exam
FastSP

- Objective: Comparing very large Multiple Sequence Alignments efficiently (in linear time)


- Software: http://www.cs.utexas.edu/~phylo/software/fastsp/
DNA Sequence Evolution

-3 mil yrs
-2 mil yrs
-1 mil yrs
today

AGGGCAT
TAGGCCCC
TAGACTT
AGCACAA
AGCGGCTT

...AGGGCAT...
...TAGGCCCC...
...TAGACTT...
...AGCACAA...
...AGCGGCTT...
Insertions and Deletions (indels)

Evolutionary Truth:

- ...ACGGTG CAGTT TCCAA...
- ...AC - - - - CAGT C CCAA...

Estimated Alignment:

- ...ACGGTG CAGTT TCC - AA - ...
- ...AC - - - - CAGT - C CCAA A...
Multiple Sequence Alignment (MSA)
MSA Estimation Methods

Basis: score alignments based on a similarity matrix and gap penalties

Most formulations of the problem are NP-complete.
   Polynomial for two sequences (dynamic programming)

There are plenty of methods to estimate alignments:

- **Progressive methods**: use a guide tree to align sequences two at a time, from most similar to more distantly related.
- **Iterative methods**: similar to progressive, but allow updating pairwise alignments if scores are improved
- **Hidden Markov models**: model “current” alignment as a Markov model, and use Viterbi algorithm to successively add new sequences to the current alignment
Alignment Comparison

- Many ways to estimate alignments
- Alignments need to be compared
Alignment Comparison: performance Study

- Assessing accuracy in performance studies

Example:

Alignment Comparison: Phylogenetic Uncertainty

- Different MSA methods produce alignments that differ enough to introduce phylogenetic uncertainty (Wong et al., 2008)

- Alignment error increases with the size of the dataset (Liu et al., 2009, 2010)

- Using several alignments, and comparisons of these alignments
Alignments Comparison Metrics

- The Developer score = SP-score (sum-of-pairs):
  Percentage of **Homologies** in Reference Alignment that are found in the estimated alignment (shared homologies).
Homologies

• The Developer score = SP-score (sum-of-pairs):
  Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

  Homology: Any pair of characters in the same column of a MSA

<table>
<thead>
<tr>
<th></th>
<th>012345678</th>
<th></th>
<th>0123456789</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>AGTGCTTC-</td>
<td>0</td>
<td>AGTGCTTC-</td>
</tr>
<tr>
<td>1</td>
<td>A---CTCCCA</td>
<td>1</td>
<td>A---CT-CCA</td>
</tr>
<tr>
<td>2</td>
<td>AC-CGTCCA</td>
<td>2</td>
<td>ACC-GT-CCA</td>
</tr>
</tbody>
</table>
Homologies

- The Developer score = SP-score (sum-of-pairs):
  Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

Homology: Any pair of characters in the same column of a MSA

```
012345678         0123456789
  0 AGTGTTC-       0 AGTGTTC- -
  1 A- - CTCCA     1 A- - CT-CCA
  2 AC-CGTCCA      2 ACC-GT-CCA
```
Homologies

- The Developer score = SP-score (sum-of-pairs):

  Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

  Homology: Any pair of characters in the same column of a MSA

<table>
<thead>
<tr>
<th></th>
<th>012345678</th>
<th>0123456789</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>AGTGCTTC-</td>
<td>AGTGCTTC-</td>
</tr>
<tr>
<td>1</td>
<td>A- - - CTCCA</td>
<td>A - - CT - CCA</td>
</tr>
<tr>
<td>2</td>
<td>AC - CGTCCCA</td>
<td>ACC - GT - CCA</td>
</tr>
</tbody>
</table>
Homologies

- The Developer score = SP-score (sum-of-pairs):
  Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

**Homology**: Any pair of characters in the same column of a MSA

<table>
<thead>
<tr>
<th></th>
<th>012345678</th>
<th>0123456789</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>AGTGCTTC-</td>
<td>AGTGCTTC-</td>
</tr>
<tr>
<td>1</td>
<td>A- - - CTCCA</td>
<td>A - - - CT - CCA</td>
</tr>
<tr>
<td>2</td>
<td>AC - CGTCCA</td>
<td>ACC - GT - CCA</td>
</tr>
</tbody>
</table>
Homologies (count)

• The Developer score = SP-score (sum-of-pairs):
  Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

Number of Homologies: two chose number of characters per column

<table>
<thead>
<tr>
<th></th>
<th>AGTGCTTC-</th>
<th>AGTGCTTC-</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>A---------</td>
<td>A---------</td>
</tr>
<tr>
<td>1</td>
<td>A---------</td>
<td>A---------</td>
</tr>
<tr>
<td>2</td>
<td>AC---------</td>
<td>ACC---------</td>
</tr>
</tbody>
</table>

310133331 total=18
3110330311 total=16
Representing Characters

- The Developer score = SP-score (sum-of-pairs):
  Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

  **Character Representation**: a pair \((a,b)\) where
  - \(a\) indicates the row in the alignment matrix
  - \(b\) indicates the position of the character in *unaligned* sequence

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1234567</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>---12345</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>01-234567</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1234567</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>---12345</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>01-234567</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Representing Homologies (homology)

- The Developer score = SP-score (sum-of-pairs):
  Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

  Homology Representation: a pair <(a,b),(c,d)> where (a,b) each represent a character in the alignment, and (a,b) and (c,d) are in the same column of the alignment.

  \[
  \begin{align*}
  0 & \quad 1 & \quad 2 & \quad 3 & \quad 4 & \quad 5 & \quad 6 & \quad 7 & \quad 8 \\
  0 & | & \quad 0 & \quad 1 & \quad 2 & \quad 3 & \quad 4 & \quad 5 & \quad 6 & \quad 7 \\
  1 & | & \quad 0 & \quad - & \quad - & \quad 1 & \quad 2 & \quad 3 & \quad 4 & \quad 5 \\
  2 & | & \quad 0 & \quad 1 & \quad - & \quad 2 & \quad 3 & \quad 4 & \quad 5 & \quad 6 & \quad 7 \\
  \end{align*}
  \]

  \[
  \begin{align*}
  0 & \quad 1 & \quad 2 & \quad 3 & \quad 4 & \quad 5 & \quad 6 & \quad 7 & \quad 8 & \quad 9 \\
  0 & | & \quad 0 & \quad 1 & \quad 2 & \quad 3 & \quad 4 & \quad 5 & \quad - & \quad - \\
  1 & | & \quad 0 & \quad - & \quad - & \quad 1 & \quad 2 & \quad - & \quad 3 & \quad 4 \\
  2 & | & \quad 0 & \quad 1 & \quad - & \quad 2 & \quad - & \quad 3 & \quad 4 & \quad 5 & \quad 6 & \quad 7 \\
  \end{align*}
  \]

  Note: the order doesn't matter: <(a,b),(c,d)> = <(c,d),(a,b)>
Shared Homology

- The Developer score = SP-score (sum-of-pairs):
  Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

**Shared Homologies**: two homologies are shared between the two alignments if they have the exact same representation.
**Shared Homology**

- The Developer score = SP-score (sum-of-pairs):
  
  Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

**Shared Homologies**: two homologies are shared between the two alignments if they have the exact same representation.

\[
\begin{array}{c}
\text{012345678} \\
0 \ 01234567- \\
1 \ 0-\cdot\cdot\cdot12345 \\
2 \ 01-234567 \\
\end{array}
\quad \quad \quad \quad \quad
\begin{array}{c}
\text{0123456789} \\
0 \ 01234567-- \\
1 \ 0-\cdot\cdot\cdot12-345 \\
2 \ 012-34-567 \\
\end{array}
\]
SP-Score

• The Developer score = SP-score (sum-of-pairs):

  Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

  SP-Score: find all homologies in both alignments, find those that are shared, and divide by the number of homologies in the reference alignment.

Reference: 012345678  Estimated: 0123456789

0 01234567-  0 01234567- -
1 0- - -12345  1 0- - -12-345
2 01-234567  2 012-34-567

ALL: 310133331 =18  SP-Score=13/18=72%

SHARED: 310033111 =13
Modeler Score

The Modeler score:

Percentage of Homologies in the estimated Alignment that are found in the reference alignment (shared homologies).

**SP-Score:** find all homologies in both alignments, find those that are shared, and divide by the number of homologies in the reference alignment.

Reference: 012345678  Estimated: 0123456789

<table>
<thead>
<tr>
<th>Reference</th>
<th>Estimated</th>
<th>Modeler Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 01234567-</td>
<td>0 01234567- -</td>
<td>16</td>
</tr>
<tr>
<td>1 0 - - - 12345</td>
<td>1 0 - - - 12 - 345</td>
<td>81%</td>
</tr>
<tr>
<td>2 01 - 234567</td>
<td>2 012 - 34 - 567</td>
<td>13</td>
</tr>
<tr>
<td>ALL: 310133331</td>
<td>SHARED: 3100330111</td>
<td>16/16 = 81%</td>
</tr>
</tbody>
</table>
Total Column Score

- Total Column (TC) score:
  
  Percentage of *aligned* columns in the reference alignment that are found in the estimated alignment.

Reference: 012345678  Estimated: 0123456789

<table>
<thead>
<tr>
<th></th>
<th>Reference</th>
<th>Estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>01234567-</td>
<td>01234567-</td>
</tr>
<tr>
<td>1</td>
<td>0- - -12345</td>
<td>0- - -12-345</td>
</tr>
<tr>
<td>2</td>
<td>01-234567</td>
<td>012-34-567</td>
</tr>
</tbody>
</table>

**ALIGNED:** YYNYYYYYYY = 8

**SHARED:** YYNYYYYYY = 6

**TC Score = 6/8 = 75%**
Definitions

\( k = \) number of characters in the longest sequence
\( k_1 = \) number of sites in the reference alignment
\( k_2 = \) number of sites in the estimated alignment
\( n = \) number of sequences

Reference: 012345678
0 01234567-
1 0- - -12345
2 01-234567

Estimated: 0123456789
0 01234567--
1 0- - -12-345
2 012-34-567

\( k=7 \quad k_1=9 \)
\( n=3 \quad k_2=10 \)
Brute Force Calculation

- Homologies in each alignment can be represented as a presence/absence matrix with $n.k$ rows and columns.

- $O(n^2k^2)$ time and memory.
FastSP: Objectives

Show that all three scores can be calculated in linear time (with respect to k.n)

Implement an efficient algorithm to calculate alignment scores
FastSP: Idea

- Characters in each column (x) of the reference alignment are dispersed in one or more columns in the estimated alignment.

- Divide characters in column x into equivalence classes, such that all characters in the same equivalence class are in the same column in the estimated alignment.

- Number of shared homologies contributed by column x is
  - sum (for all equivalence classes S of x) \(|S| \choose 2\)

Reference: 012345678

0 01234567-
1 0- - -12345
2 01- 234567

Estimated: 0123456789

0 01234567--
1 0- - -12-345
2 012- 34-567

\( \frac{1}{2} \) + \( \frac{2}{2} \) = 1
FastSP: Algorithm

Reference:

```
0  AGTGCTTC-
1  A---CTCCA
2  AC-CGTCCA
0  01234567-
1  0- - -12345
2  01- 234567
```

1- Read reference alignment and save it with this character representation

- (also find k and n).
2- Read estimated alignment and create a n by k matrix S such that

- $S[i,j] = x$ iff $\text{Estimated}_\text{Alignment}[i,x] = j$. 

Reference:

```
 012345678
0 01234567-
1 0- -12345
2 01- 234567
```

Estimated:

```
012345679
0 01234567--
1 0- -12-345
2 012-34-567
```

Matrix S:

```
01234567
0 01234567-
1 0- -12-345
2 012-34-567
```
3- For each column of reference alignment (x)

- Mu= An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
  - Increment Mu[ S[r][M] ]
- Shared [x] = \(\sum_{j=1}^{\frac{k_2}{2}} \binom{\text{Mu}_j}{2}\)
3- For each column of reference alignment

- **Mu** = An array of length \( k_2 \) initialized by 0 (or a dictionary)
- For character \( M \) in row \( r \)
  - Increment \( \text{Mu}[S[r][M]] \)
- **Shared** \( [x] = \sum_{j=1}^{k_2} \binom{\text{Mu}_j}{2} \)
3- For each column of reference alignment

- $\text{Mu} = \text{An array of length } k_2 \text{ initialized by 0 (or a dictionary)}$
- For character $M$ in row $r$
  - Increment $\text{Mu}[ S[r][M] ]$
- Shared $[x] = \sum_{j=2}^{M} \left( \frac{\text{Mu}_j}{2} \right)$
3- For each column of reference alignment

- \( \text{Mu} = \) An array of length \( k_2 \) initialized by 0 (or a dictionary)
- For character M in row r
  - Increment \( \text{Mu}[ S[r][M] ] \)
- \( \text{Shared } [x] = \sum \binom{\text{Mu}_j}{2} \)
### FastSP

#### Reference:

<table>
<thead>
<tr>
<th></th>
<th>012345678</th>
<th>01234567-</th>
<th>0- - - -12345</th>
<th>01- -234567</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>01234567</td>
<td>01234567</td>
<td>0---12345</td>
<td>012-34567</td>
</tr>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>0---12-345</td>
<td>012-34-567</td>
</tr>
<tr>
<td>2</td>
<td>01-234567</td>
<td>01234567-</td>
<td>045789-</td>
<td>01245789</td>
</tr>
</tbody>
</table>

#### Estimated:

<table>
<thead>
<tr>
<th></th>
<th>0123456789</th>
<th>01234567-</th>
<th>0- - - -12345</th>
<th>01- -234567</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>01234567</td>
<td>01234567</td>
<td>0---12345</td>
<td>012-34567</td>
</tr>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>0---12-345</td>
<td>012-34-567</td>
</tr>
<tr>
<td>2</td>
<td>01-234567</td>
<td>01234567-</td>
<td>045789-</td>
<td>01245789</td>
</tr>
</tbody>
</table>

#### Matrix S:

$$
\sum \left( \begin{array}{c}
\text{Mu}_j \\
\end{array} \right)
$$

$$
\text{Mu} = [3 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]
$$

#### 3- For each column of reference alignment

- **Mu** = An array of length $k_2$ initialized by 0 (or a dictionary)
- For character $M$ in row $r$
  - Increment $\text{Mu}[ S[r][M] ]$
- Shared $[x] = \sum \binom{\text{Mu}_j}{2}$
3- For each column of reference alignment

- Mu = An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
  - Increment Mu[S[r][M]]
- Shared [x] = ∑(Mu_j)

Reference:

0 12345678
1 0-----12345
2 01-234567
3

Estimated:

0 123456789
1 01234567--
2 0412-34-567
3

Matrix S:

0 1234567
1 0 0 0 0 0 0 0 0 0
2 0 0 0 0 0 0 0 0 0
3

Shared = \( \sum \binom{3}{2} + \binom{0}{2} + \binom{0}{2} + \binom{0}{2} + \binom{0}{2} + \cdots = 3 \)

Mu = [3 0 0 0 0 0 0 0 0]
3- For each column of reference alignment

- \( \text{Mu} = \) An array of length \( k_2 \) initialized by 0 (or a dictionary)
- For character \( M \) in row \( r \)
  - Increment \( \text{Mu}[ S[r][M] ] \)
- \( \text{Shared} [x] = \sum \binom{\text{Mu}_j}{2} \)
3- For each column of reference alignment

- Mu= An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
  - Increment Mu[ S[r][M] ]
- Shared [x]= \( \sum_{j}^{ \binom{2}{2} } (Mu_j) \)
3- For each column of reference alignment

- \( \text{Mu} = \) An array of length \( k_2 \) initialized by 0 (or a dictionary)
- For character M in row r
  - Increment \( \text{Mu}[ S[r][M] ] \)
- Shared \( [x] = \sum_{j=2}^{M_u} \binom{M_u}{2} \)
### FastSP

<table>
<thead>
<tr>
<th>Reference:</th>
<th>Estimated:</th>
<th>Matrix S:</th>
</tr>
</thead>
<tbody>
<tr>
<td>012345678</td>
<td>0123456789</td>
<td>012345677</td>
</tr>
<tr>
<td>0</td>
<td>01234567--</td>
<td>012345677</td>
</tr>
<tr>
<td>1</td>
<td>0---12345</td>
<td>045789--</td>
</tr>
<tr>
<td>2</td>
<td>01-234567</td>
<td>01245789</td>
</tr>
<tr>
<td>310</td>
<td>012-34-567</td>
<td></td>
</tr>
<tr>
<td>310</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3- For each column of reference alignment

- **Mu** = An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
  - Increment **Mu**[ S[r][M] ]
- **Shared** [x] = \( \sum_{j=2}^{n} \binom{\text{Mu}_j}{2} \)
3- For each column of reference alignment

- \( \text{Mu} = \) An array of length \( k_2 \) initialized by 0 (or a dictionary)
- For character \( M \) in row \( r \)
  - Increment \( \text{Mu}[S[r][M]] \)
- \( \text{Shared} [x] = \sum \binom{\text{Mu}_j}{2} \)
3- For each column of reference alignment

- Mu= An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
  - Increment Mu[ S[r][M] ]
- Shared [x]= \( \sum_{j=1}^{k2} Mu_j \)
**FastSP**

<table>
<thead>
<tr>
<th>Reference:</th>
<th>Estimated:</th>
<th>Matrix S:</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 12345678</td>
<td>0123456789</td>
<td>01234567</td>
</tr>
<tr>
<td>0 1234567-</td>
<td>0 1234567-</td>
<td>0 1234567</td>
</tr>
<tr>
<td>1 0----12345</td>
<td>1 0----12-345</td>
<td>1 045789-</td>
</tr>
<tr>
<td>2 01-234567</td>
<td>2 012-34-567</td>
<td>2 01245789</td>
</tr>
</tbody>
</table>

**3- For each column of reference alignment**

- Mu= An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
  - Increment Mu[ S[r][M] ]
- Shared \[x]= \sum_{j=1}^{2} \left( \frac{\text{Mu}_j}{2} \right)
3- For each column of reference alignment

- \( \text{Mu} = \) An array of length \( k_2 \) initialized by 0 (or a dictionary)
- For character \( M \) in row \( r \)
  - Increment \( \text{Mu}[ S[r][M] ] \)
- \( \text{Shared} [x] = \sum \binom{\text{Mu}_j}{2} \)
3- For each column of reference alignment

- \( \text{Mu}= \) An array of length \( k_2 \) initialized by 0 (or a dictionary)
- For character M in row \( r \)
  - Increment \( \text{Mu}[ S[r][M] ] \)
- \( \text{Shared} [x] = \sum \binom{\text{Mu}_j}{2} \)
FastSP

3- For each column of reference alignment

- \( \text{Mu} = \) An array of length \( k_2 \) initialized by 0 (or a dictionary)
- For character \( M \) in row \( r \)
  - Increment \( \text{Mu}[S[r][M]] \)
- \( \text{Shared} \ [x] = \sum \binom{\text{Mu}_j}{2} \)
FastSP

<table>
<thead>
<tr>
<th>Reference:</th>
<th>Estimated:</th>
<th>Matrix S:</th>
</tr>
</thead>
<tbody>
<tr>
<td>012345678</td>
<td>0123456789</td>
<td>01234567</td>
</tr>
<tr>
<td>001234567-</td>
<td>001234567--</td>
<td>001234567</td>
</tr>
<tr>
<td>000-12345</td>
<td>000---12-345</td>
<td>00045789--</td>
</tr>
<tr>
<td>0201-234567</td>
<td>02012-34-567</td>
<td>0201245789</td>
</tr>
</tbody>
</table>

\[ \text{SP-Score} = \frac{13}{18} \]

4- Report \((\text{sum of shared})/(\text{sum of reference})\) as SP-Score
Running Time Analysis

1- read reference alignment and save it with our character representation
   \( O(n.k1) \)

2- read estimated alignment and create a \( n \) by \( k \) matrix \( S \) such that
   - \( S[i,j]=x \) iff \( \text{EstimatedAlignment}[i,x]=j \).
   \( O(n.k2) \)

3- For each column of reference alignment (\( k1 \))
   - \( \text{Mu} = \) An array of length \( k2 \) initialized by 0 (or a dictionary)
   - For character \( M \) in row \( r \) (\( n \))
     - Increment \( \text{Mu}[ S[r][M] ] \)
   \( O(n.k1) \)

4- Report (sum of shared)/(sum of reference)
   \( O(k1) \)

\[ \text{Overall}=O(\max(k1,k2).n) \]
Memory Analysis

1- read reference alignment and save it with our character representation
   \[ O(n.k1) \]

2- read estimated alignment and create a n by k matrix S such that
   - \[ S[i,j]=x \iff \text{Estimated\_Alignment}[i,x]=j. \]
   \[ O(n.k) \]

3- For each column of reference alignment
   - \[ \text{Mu}= \text{An array of length k2 initialized by 0 (or a dictionary) } O(k2) \text{ or } O(n) \]
   - For character M in row r
     - Increment \[ \text{Mu}[ S[r][M] ] \]
     - Shared \[ [x]= \sum \left( \frac{\text{Mu}_j}{2} \right) \]

4- Report \( \frac{\text{sum of shared}}{\text{sum of reference}} \)

Overall=\( O((k1+k).n) \)
Memory Analysis

• Trick: choose smallest of k1 and k2 as reference alignment and the other as estimated alignment. Number of shared homologies will be the same either way.

Overall=O((min(k1,k2)+k).n)
Modeler and TC scores

• Both Modeler and TC scores can be calculated with FastSP algorithm without any sacrifice to running time and memory

• Modeler: we just need to calculate total number of homologies in the estimated alignment

• TC: As we go through column of reference alignment, if we get only one equivalence class, we have a correct column.
Implementation

• Implemented in Java
  • Computes SP, Modeler, and TC in one run

• 420 LOC

• Available publicly at
  http://www.cs.utexas.edu/~phylo/software/fastsp/
# Performance Study: datasets

**S.Mirarab and T.Warnow**

---

**Table 1. Datasets and their sizes**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( n^a )</th>
<th>( L )</th>
<th>Ref.</th>
<th>MAFFT</th>
<th>OPAL</th>
<th>PART</th>
<th>PRANK</th>
<th>QUICK</th>
<th>SATé</th>
<th>SATé-II</th>
</tr>
</thead>
<tbody>
<tr>
<td>100L1-R0</td>
<td>100</td>
<td>1089</td>
<td>2287</td>
<td>1563</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1737</td>
</tr>
<tr>
<td>500L1-R0</td>
<td>500</td>
<td>1110</td>
<td>4992</td>
<td>3307</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3421</td>
</tr>
<tr>
<td>1000L1-R0</td>
<td>1000</td>
<td>1079</td>
<td>3517</td>
<td>907</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2856</td>
</tr>
<tr>
<td>Price-78K</td>
<td>78132</td>
<td>1286</td>
<td>1287</td>
<td></td>
<td>1504</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23S.E</td>
<td>117</td>
<td>5317</td>
<td>9079</td>
<td>8929</td>
<td>9860</td>
<td>11018</td>
<td>13941</td>
<td>6796</td>
<td>10352</td>
<td></td>
</tr>
<tr>
<td>23S.E.aa_ag</td>
<td>144</td>
<td>1079</td>
<td>8619</td>
<td>8123</td>
<td>9576</td>
<td>10956</td>
<td>14343</td>
<td>7017</td>
<td>9029</td>
<td></td>
</tr>
<tr>
<td>23S.M.aa_ag</td>
<td>263</td>
<td>4483</td>
<td>10305</td>
<td>7353</td>
<td>13625</td>
<td>12320</td>
<td>13471</td>
<td>5522</td>
<td>7815</td>
<td></td>
</tr>
<tr>
<td>23S.M</td>
<td>278</td>
<td>4216</td>
<td>10738</td>
<td>7478</td>
<td>10447</td>
<td>13384</td>
<td>13639</td>
<td>5311</td>
<td>8746</td>
<td></td>
</tr>
<tr>
<td>16S.M</td>
<td>901</td>
<td>2023</td>
<td>4722</td>
<td>4418</td>
<td>12812</td>
<td>9496</td>
<td>12826</td>
<td>3216</td>
<td>4776</td>
<td></td>
</tr>
<tr>
<td>16S.M.aa_ag</td>
<td>1028</td>
<td>2672</td>
<td>4907</td>
<td>4493</td>
<td>13785</td>
<td>11225</td>
<td>20856</td>
<td>3317</td>
<td>48881</td>
<td></td>
</tr>
<tr>
<td>16S.3</td>
<td>6323</td>
<td>4066</td>
<td>8716</td>
<td></td>
<td>19775</td>
<td></td>
<td>5310</td>
<td>10186</td>
<td>20414</td>
<td></td>
</tr>
<tr>
<td>16S.T</td>
<td>7350</td>
<td>4066</td>
<td>11856</td>
<td>10891</td>
<td>43797</td>
<td>25951</td>
<td></td>
<td>6109</td>
<td>12301</td>
<td></td>
</tr>
<tr>
<td>16S.B.ALL</td>
<td>27643</td>
<td>1851</td>
<td>6857</td>
<td></td>
<td>14217</td>
<td></td>
<td>3413</td>
<td>8463</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16S.GG-50K</td>
<td>50000</td>
<td>1701</td>
<td>7682</td>
<td></td>
<td>14877</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( n^a \) is the number of sequences. \( L \) is the maximum number of nucleotides in any of the sequences. Ref. is the length of the reference alignment (i.e. including gaps). The rest of the columns show the length of the alignment for each estimated alignment. An empty cell indicates that the respective alignment method was not run on that particular dataset. The first four datasets are simulated datasets, while the rest are all real biological datasets.

We have included results from two different runs of SATé-II on 16S.B.ALL dataset. The second version had a length of 8209.
Performance Study: All Techniques
Fig. 4. Comparison of FASTSP and QSCORE-default with respect to running time on machines with ‘at least’ 8 GB of main memory.
Fig. 5. Comparison of FASTSP and QSCORE-default with respect to peak memory usage on machines with ‘at least’ 8 GB of main memory.
Performance Study: Limited Memory

![Graph showing performance study results](image)

**Fig. 1.** Log-scaled running time on machines with 4 GB of main memory. QSCORE is run only in default setting, and so computes only SP- and TC-scores. Note that QSCORE fails to analyze the largest dataset.
Summary

• Two alignments can be compared in terms of SP-Score, Modeler Score, and TC in linear time (linear with respect to k.n)

• FastSP provides a memory-efficient tool for comparing alignments