A comparison of four pair-wise sequence alignment methods

Nadia Essoussi¹ and Sondes Fayech¹, *

¹Department of Computer Science, Higher Institute of Management, Tunis, Tunisia; Sondes Fayech* - E-mail: sondes.fayech@isg.rnu.tn;
Phone: 216 71 58 85 14; Fax: 216 71 58 84 87; * Corresponding author

received August 24, 2007; revised November 23, 2007; accepted December 09, 2007; published online December 28, 2007

Abstract:
Protein sequence alignment has become an essential task in modern molecular biology research. A number of alignment techniques have been documented in literature and their corresponding tools are made available as freeware and commercial software. The choice and use of these tools for sequence alignment through the complete interpretation of alignment results is often considered non-trivial by end-users with limited skill in Bioinformatics algorithm development. Here, we discuss the comparison of sequence alignment techniques based on dynamic programming (N-W, S-W) and heuristics (LFASTA, BL2SEQ) for four sets of sequence data towards an educational purpose. The analysis suggests that heuristics based methods are faster than dynamic programming methods in alignment speed.

Keywords: sequence alignment techniques; Needleman & Wunsch; Smith & Waterman; LFASTA; BL2SEQ

Methodology
Datasets
Dataset #1: DS-R
It contains 200 protein sequences selected randomly from Universal Protein Resource (UNIPROT, www.uniprot.org). This dataset is thereafter designated as DS-R.

Dataset #2: DS-20
The PISCES server is used to create this dataset [7]. PISCES is a protein sequence culling server (http://dunbrack.fccc.edu/PISCES.php) with sequences culled from the Protein Data Bank [8] (PDB, http://www.rcsb.org/pdb) based on maximum sequence similarity. We downloaded S-20 (containing non-redundant sequences at less than 20% sequence similarity) dataset from PISCES. We extracted 200 sequences from S-20 in a random manner and created a dataset designated as DS-20. It contains non-redundant sequences at 20% sequence similarity cut-off.

Dataset #3: DS-40
We downloaded S-40 (containing non-redundant sequences at less than 40% sequence similarity) dataset from PISCES. We extracted 200 sequences from S-40 in a random manner and created a dataset designated as DS-40. It contains non-redundant sequences at 40% sequence similarity cut-off.

Dataset #4: DS-90
We downloaded S-90 (containing non-redundant sequences at less than 90% sequence similarity) dataset from PISCES. We extracted 200 sequences from S-90 in a random manner and created a dataset designated as DS-90. It contains non-redundant sequences at 90% sequence similarity cut-off.

Data statistics
The distribution of sequences with varying lengths for datasets #1 to #4 is summarized in Table 1 (supplementary material).
**Hypothesis**

**Sequence comparison**

We performed pair-wise alignment for randomly selected sequences from one dataset to sequences in other datasets such as DS-R, DS-20, DS-40 and DS-90 using N-W, S-W, LFASTA and BL2SEQ in a one-to-many alignment manner.

**Discussion:**

Sequence alignment is an important task in sequence based molecular biology experiments in modern research. A number of sequence alignment tools are available in the internet for varying purposes (see EMBOSS). However, selection of specific tools for a Biologist who is not an expert in the field of Bioinformatics is non-trivial. Here, we describe the comparison of pair-wise sequence alignment using methods N-W, S-W, LFASTA and BL2SEQ described elsewhere [1-4]. These techniques and their corresponding tools are developed by authors with strong mathematical knowledge. This is not the case with end-users who often have difficulties in selecting tools and interpreting alignment results. The performance of these methods has been discussed extensively in graduate level TEXT books for Bioinformatics. However, a comparative study on the performance of these techniques is not explicitly available. In this study, we use execution time (alignment speed) as a parameter to compare four alignment methods. For the purpose of simplicity, the experiment is conducted in a 2.4 GHZ Pentium-IV processor with 512 MB of RAM under windows platform.

Figure 1 gives the profile for execution time (alignment speed) versus sequence length for all the four methods used in the analysis using four different datasets (DS-R, DS-20; DS-40; DS-90). The analysis shows that alignment speed for heuristic methods such as LFASTA and BL2SEQ are faster than dynamic programming methods such as N-W, S-W. This provides insight to the selection of several programs that are 

**Alignment execution time**

The execution time is the time needed to perform an alignment between two protein sequences for a given method in a 2.4 GHZ Pentium-IV processor with 512 MB of RAM.

**Sequence alignment tools**


**Figure 1:** Performance of N-W, S-W, LFASTA and BL2SEQ for data sets DS-R, DS-20, DS-40 and DS-90 is given. The alignment speed for BL2SEQ is high with low execution time for all the four dataset.
available for sequence alignment in the internet for end-users who often use them for biological investigations. The time taken by N-W is the largest for all the four datasets. This is followed by S-W (S-W is faster than N-W). The least time is taken by the heuristic method BL2SEQ. LFASTA is slower than BL2SEQ and faster than S-W. Thus, BL2SEQ is the preferred method of choice in terms of alignment speed. The performance of the methods is not affected by dataset type and length of sequences. Although, this comparison experiment is simple, the profiles explicitly show the method that is quick to perform pair-wise sequence alignment given the choices.

Conclusion:
The comparison of sequence alignment techniques such as N-W, S-W, LFASTA and BL2SEQ for four sets of sequence data is discussed. The analysis suggests that heuristic methods such as LFASTA and BL2SEQ are faster than dynamic programming methods such as N-W, S-W. This comparison is useful for educational purpose to non-experts in Bioinformatics algorithm development.

References:

Reference:
Edited by P. Kangeane

Citation: Essoussi & Fayech, Bioinformation 2(4): 166-168 (2007)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

Supplementary material

<table>
<thead>
<tr>
<th>Sequence length (residues)</th>
<th>Dataset size (number)</th>
<th>DS-R</th>
<th>DS-20</th>
<th>DS-40</th>
<th>DS-90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein &lt; 100</td>
<td></td>
<td>50</td>
<td>57</td>
<td>14</td>
<td>27</td>
</tr>
<tr>
<td>100 ≤ Protein &lt; 200</td>
<td></td>
<td>45</td>
<td>59</td>
<td>66</td>
<td>53</td>
</tr>
<tr>
<td>200 ≤ Protein &lt; 300</td>
<td></td>
<td>60</td>
<td>39</td>
<td>51</td>
<td>72</td>
</tr>
<tr>
<td>300 ≤ Protein &lt; 400</td>
<td></td>
<td>25</td>
<td>23</td>
<td>34</td>
<td>24</td>
</tr>
<tr>
<td>400 ≤ Protein &lt; 500</td>
<td></td>
<td>12</td>
<td>13</td>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>500 ≤ Protein</td>
<td></td>
<td>8</td>
<td>9</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

Table1: Distribution of sequences in different datasets based on protein sizes. Description on datasets DS-R, DS-20, DS-40 and DS-90 is given in methodology. The total number of randomly chosen sequences in each dataset is 200.