Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

A comparison of four pair-wise sequence alignment methods

Nadia Essoussi[1] and Sondes Fayech[1,*]

[1]Department 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 Novemeber 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

Background:
Protein sequence alignment is an important step in understanding molecular functions from sequences. Sequence alignments help to infer functions for new sequences by detecting similarity with sequences of known function. Sequence comparison using pair-wise alignment techniques such as Needleman & Wunsch (N-W) **[1]**, Smith & Waterman (S-W) **[2]**, LFASTA **[3]**, BL2SEQ **[4]** and several others are available. The use of these techniques has been elaborately described in graduate level TEXT books for Bioinformatics.

Sequence alignment techniques such as N-W, S-W, LFASTA and BL2SEQ are routinely used in molecular biology laboratory (research) and drug discovery (development) environment. The N-W algorithm performs global alignment (comparison of entire sequences) between sequences and the S-W algorithm performs local alignment (comparison of local stretches of sequences for the identification of motifs). The LFASTA and BL2SEQ methods use heuristic (rule of thumb) to compare protein sequences. The measure of similarity in these methods is scored using similarity matrices **[5, 6]**.

The availability of several protein sequence comparison tools provide a wide range of choice for selecting appropriate tools for specific purposes. Generally these tools show varying degree of difference between them. These differences at a fine level are seldom used correctly by end-users who are non-experts in Bioinformatics developments. Here, use we execution time as a parameter to compare sequence alignment tools using scoring matrices such as BLOSUM 45, BLOSUM 62 and BLOSUM 80 **[5, 6].** This comparison is of help to biologist who are non-expert in Bioinformatics to select appropriate sequence tools for specific tasks based on the available in-house infra-structural facilities.

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 Databank **[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).

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **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.

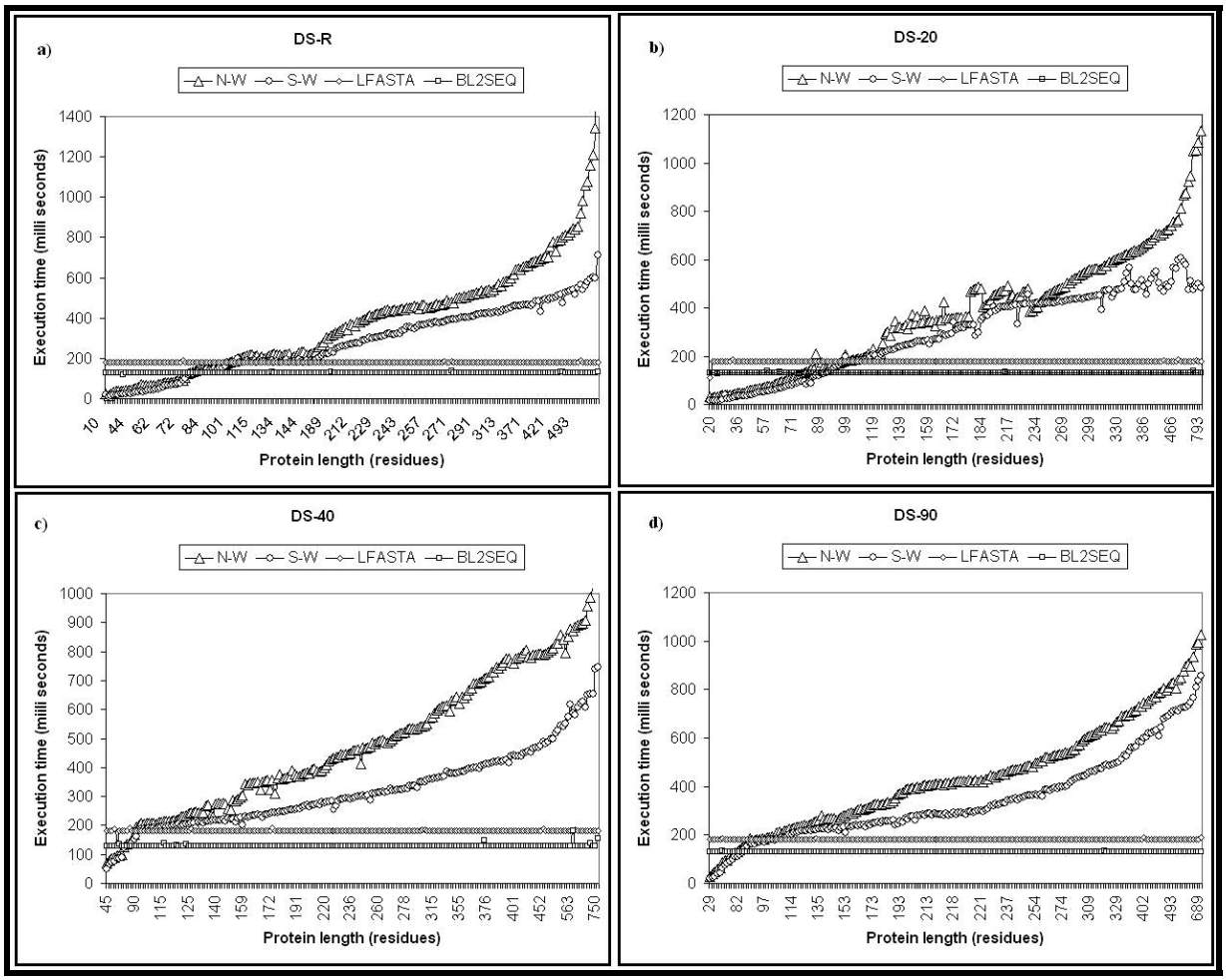

Figure 1: Performance of N-W, S-W, LFASTA and BL2SEQ for datasets DS-R, DS-20, DS-40 and DS-90 is given. The alignment speed for BL2SEQ is high with low execution time for all the four dataset

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

The alignment tools N-W and S-W are downloaded from EMBOSS (ftp://emboss.open-bio.org/pub/EMBOSS/). LFASTA is downloaded from FASTA website (http://faculty.virginia.edu/wrpearson/fasta/win32_fasta/) and BL2SEQ is downloaded from BLAST website (http://www.ncbi.nlm.nih.gov/BLAST/download.shtml).

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

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:

[1] S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.*, 48: 443 (1970) [PMID: 5420325]

[2] T. F. Smith and M. S. Waterman, *J. Mol. Biol.*, 147: 195 (1981) [PMID: 7265238]

[3] W. R. Pearson and D. J. Lipman, *Proc. Natl. Acad. Sci.*, 85: 2444 (1988) [PMID: 3162770]

[4] T. A. Tatusova and T. L. Madden, *FEMS Microbiol. Lett.*, 177: 247 (1999) [PMID: 10339815]

[5] S. Henikoff and J. G. Henikoff, *Proc. Natl. Acad. Sci.*, 89: 10915 (1992) [PMID: 1438297]

[6] S. Henikoff and J. G. Henikoff, *Proteins*, 17: 49 (1993) [PMID: 8234244]

[7] G. Wang and R. L. Dunbrack, *Bioinformatics*, 19: 1589 (2003) [PMID: 12912846]

[8] H. M. Berman, *et al*., *Nucleic Acids Res.*, 28: 235 (2000) [PMID: 10592235]

Edited by P. Kangueane

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

Supplementary material

Sequence length (residues)	Dataset size (number)			
	DS-R	**DS-20**	**DS-40**	**DS-90**
Protein < 100	50	57	14	27
100 ≤ Protein < 200	45	59	66	53
200 ≤ Protein < 300	60	39	51	72
300 ≤ Protein < 400	25	23	34	24
400 ≤ Protein < 500	12	13	18	13
500 ≤ Protein	8	9	17	11
Total	**200**	**200**	**200**	**200**

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