NEEDLEMAN-WUNSCH AND SMITH-WATERMAN COMBINATIONS IN PAIRWISE ALIGNMENT

Asril Adi Sunarto[1], Prajoko[2]
[1,2] Teknik Informatika, Universitas Muhammadiyah Sukabumi
Jl. Syamsudin SH. No. 50 Kota Sukabumi Jawa Barat, Indonesia
asriladi@ummi.ac.id

ABSTRAK

Identification of Deoxyribonucleic acid (DNA), Ribonucleic acid (RNA), or protein needs to be done to find functional, structural, or evolutionary relationships between two sequences. There are various applications that already exist such as one of them EMBOSS either web or desktop versions. There are drawbacks to this application, such as repeatedly processing each user who needs sequence alignment results locally and globally at the same time. Therefore, we designed an application that can generate two sequence alignment outputs both locally and globally at the same time with pairwise alignment Needleman-Wunsch and Smith-Waterman. The result show that methods can be produces two outputs of sequence alignment in the same process. The impact is it can reduce the waiting time for users.

Keyword : *DNA, RNA, EMBOSS, Smith-Waterman, Needleman-Wunsch, Sequence Alignment.*

1. INTRODUCTION

Sequence alignment of DNA is the first step in understanding basic concepts of bioinformatics, which aims to find functional, structural or evolutionary similarities between these sequences. It is usefuly in revealing the relationship between species or groups of organisms, so that it can predict the function of unknown biological molecules based on their similarities to molecules previously.

In the medicine, sequence alignment is a short reads to get precision medicine. Researchers can design drugs that intervene with these targets by finding sequences associated with certain diseases or conditions, including understanding genetic variations between individuals in a population. Such as whole genome sequencing analysis for cancer genomics and precision medicine. Among the methods for sequence alignment are the Needleman–Wunsch (NW) and Smith-Waterman (SW) methods. Both of them compare one by one and in pairs during the alignment process.

2. LITERATURE REVIEW

Sequence Alignment is a method of aligning Deoxyribonucleic acid (DNA), Ribonucleic acid (RNA), or protein sequences to identify areas of similarity that may be a consequence of functional, structural, or evolutionary relationships between two sequences [1]. Sequence alignment also provides a lot of information about the composition of a sample from the taxonomic classification and functional classification of an organism [2] to find similarities and homology between organisms compared [3]. Its homologous sequences can contain information on their evolutionary histories [4] and including to know a vaccine from a virus or bacteria [5] that utilize cellulase for the bioconversion of lignocellulosic materials into an energy source [6]. Therefore, it is very important that this Sequence Alignment is done.

There is a crucial issue in the sequence alignment process, namely fast processing - less accurate and vice versa, slow processing - high accuracy [7]. To determine the level of similarity and determine the phylogenetic form of an organism, a method that has high accuracy is needed. Broadly speaking, there are two types of determining similarity with high accuracy, namely globally and locally. The globally alignment is carried out from beginning till end of the sequence to find out the best possible alignment. Whereas, the locally alignment focus on sequences which are suspected to have similarity or even dissimilar sequences on local regions with high level of similarity. One method for sequence alignment is the NW method which defines how to find the best global results from two sequences [8]. The local similarity algorithm only looks for subsequences, and a single comparison can produce several different subsequences [9].

Even so, both of these methods use dynamic programming, which is a method that has a strategy of breaking down problems into smaller ones and using the smaller solutions to build larger solutions with a complexity of up to O(n^2) [12]. An application that has used the NW and SW methods is EMBOSS [11] which can be accessed via the site “http://www.ebi.ac.uk/Tools/psa/. The Basic Local Alignment Search Tool (BLAST) [10] can also use SW to process sequence alignment which can be accessed through the website “http://www.genome.jp/tmp/blast/”. Both applications are web-based. The classic problem with web-based applications is that there is a network connection that can work if an online network is available. Other applications such as

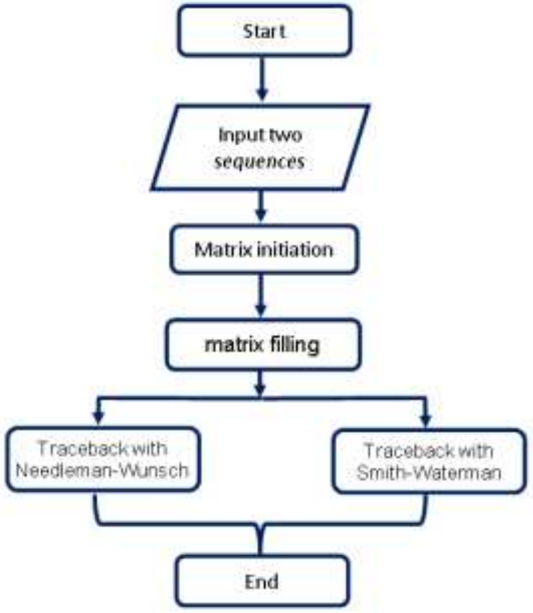
EMBOSS are designed to process almost all requirements related to processing of molecular biology. The three applications when the sequence alignment process with global and local methods are carried out separately. This requires extra time for the user to wait for the sequence alignment results. Based on these two issues, we developed a similar desktop-based application. Here is summarize advantages and disadvantages of EMBOSS and BLAST that can be seen ini Table 1 below.

Table 1. Advantages and disadvantages of EMBOSS and BLAST

	EMBOSS	BLAST
Avaibility	Web	Web, Desktop
Complexity	O(n^2)	O(n^2)
Times processes NW and SW	2	2

3. RESEACH METHODS

The method we use is combining the NW [8] and SW [9] methods in one process. The following is the flow of the research process which can be seen in Figure 1 below:



Start
Input two sequences
Matrix initiation
matrix filling
Traceback with Needleman-Wunsch
Traceback with Smith-Waterman
End

Figure 1. Research Method

To support this research, we need a set of computers and materials taken from the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov) with the names of the organisms Chlamydophila caviae (ref. NC_004720.1) and Clostridium difficile (ref. .NC_008226.1). Matrix initiation is performed using parameters *match* = 7, *mismatch* = -3, and *gap* = -2. The first organism becomes a row (A) and another organism becomes a column (B) that are in the matrix are denoted by Equation (1) and Equation (2) for filling out the matrix (M) using the NW method below.

$$A_i = A_{i-1} + gap,$$
$$B_j = B_{j-1} + gap \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \ (1)$$

$$M[i,j] = max \begin{cases} M[i-1,j-1] + sub(A[i],B[j]); \\ M[i-1,j] + del(A[i]); \\ M[i,j-1] + ins(B[j]) \end{cases} \quad ..(2)$$

$$i = position\ of\ A;\ j = position\ of\ B$$

In the next step, the backward tracing process, there are differences in the location of the prefix position used. It causes NW starts from position M[i,j] as the last position goes to M[0,0]. Meanwhile, SW begins from the highest value of the M matrix to M[0,0]. To find out the highest value of Matrix, we can use Equation (3) below. After getting the highest score, do the backtracking process by including the alignment of two organisms with use the Equation (4) below.

$$Mi,j = max \begin{cases} M[0,0]; \\ \ldots. \\ \ldots. \\ M[i,j] \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots.(3)$$

$$max \begin{cases} diag = M[i-1,j-1] + sub(A[i-1],B[j-1])\ ?match : mismath; \\ left = M[i-1,j] + gap; \\ up = M[i,j-1] + gap \end{cases} \quad (4)$$

4. RESULT AND DISCUSSION

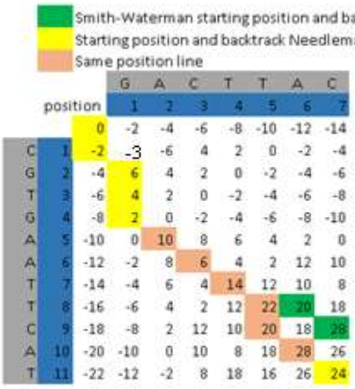
Based on the method above, we compared two existing applications. As an example input of sequence alignment, we process Sequence1 = “CGTGAATTCAT, and Sequence2 = "GACTTAC". The matrix (M) initiation set parameters *match* = 7, *mismatch* = -3, and *gap* = -2 to complete filling matrix using Equation 1 that A_0 to A_n and B_0 to B_n can be seen in Figure 2 below. Furthermore, Equation 2 is used to fill in the remaining unfilled matrix for example M[1,1]. Parameter match and mismatch are comparing between A[i] and B[i] which is notated in as Sub(A[i], B[i]). If A[i]=B[i] then parameter *match* is taken, else mismatch. Here calculation to fill M[1,1] due Equation 2 below.

$$M[1,1] = max \begin{cases} M[0,0] + -3 = 0 + -3 = -3; \\ M[0,1] + -2 = -2 + -2 = -4; \\ M[1,0] + -2 = -2 + -2 = -4; \end{cases}$$

	B➔		G	A	C	T	T	A	C
	Position		1	2	3	4	5	6	7
A🡻		0	-2	-4	-6	-8	-10	-12	-14
C	1	-2	-3						
G	2	-4							
T	3	-6							
G	4	-8							
A	5	-10							
A	6	-12							
T	7	-14							
T	8	-16							
C	9	-18							
A	10	-20							
T	11	-22							

Figure 2. Filling Matrix M[1,1] using NW / SW

After filling in the matrix M is complete, the next step is backtracking from position M[i,j] for NW and the position with the highest value for SW. Overall stage with NW and SW that using Equation 1 to Equation 4 become such as in Figure 3 below.

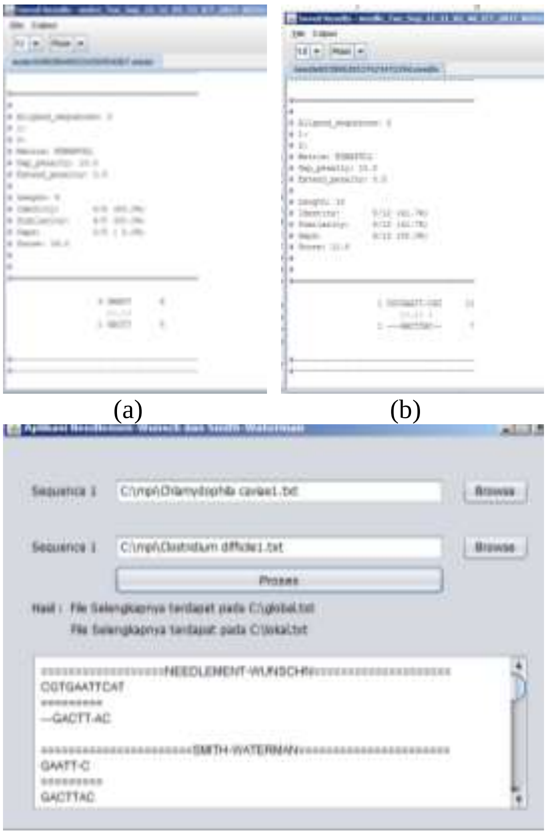


🟩 Smith-Waterman starting position and backtrack
🟨 Starting position and backtrack Needleman-Wunsch
🟧 Same position line

			G	A	C	T	T	A	C
	position		1	2	3	4	5	6	7
		0	-2	-4	-6	-8	-10	-12	-14
C	1	-2	-3	-6	4	2	0	-2	-4
G	2	-4	6	4	2	0	-2	-4	-6
T	3	-6	4	2	0	-2	-4	-6	-8
G	4	-8	2	0	-2	-4	-6	-8	-10
A	5	-10	0	10	8	6	4	2	0
A	6	-12	-2	8	6	4	2	12	10
T	7	-14	-4	6	4	14	12	10	8
T	8	-16	-6	4	2	12	22	20	18
C	9	-18	-8	2	12	10	20	18	28
A	10	-20	-10	0	10	8	18	28	26
T	11	-22	-12	-2	8	18	16	26	24

Figure 3. NW and SW Matrix and Backtracking

As a result, each result has its stages. Figure 4 shows sequence alignment using two methods at once in one process. The analysis and evaluation of the data follow the appropriate theoretical study formula. The appearance of successive applications can be seen in Figure 1 as follows:



(a) (b)

Sequence 1 C:\mp\Chlamydophila caviae1.txt
Sequence 1 C:\mp\Clostridium difficle1.txt
Hasil : File Selengkapnya terdapat pada C:\global.txt
File Selengkapnya terdapat pada C:\lokal.txt

```
==================NEEDLEMENT-WUNSCH====================
CGTGAATTCAT
=========
---GACTT-AC

=====================SMITH-WATERMAN=====================
GAATT-C
=========
GACTTAC
```

(c)

Figure 4. User Interface and Results of the Desktop-based EMBOSS Sequence Alignment. (a) using SW method. (b) Using NW method. (c) Using combine NW and SW methods

The final results of the Needlemen-Wunsch and SW alignment sequences can see in Table 1 and Table 2 below. The sequence alignment results between EMBOS and the method designed show a slight difference in the SW backtracking process. The difference is because the EMBOS program calculates the highest total score among the matched characters. The program that we make on the initial position starts from the highest value in the backtracking process.

Table 2. Sequence alignment results NW

C	G	T	G	A	A	T	T	C	A	T
			|	|		|	|		|	
-	-	-	G	A	C	T	T	-	A	C

Table 3. SW sequence alignment results

G	A	A	T	T	-	C
|	|	|	|	|		|
G	A	C	T	T	A	C

5. CONCLUSION

Based on a comparison between the existing program and the program we designed, it produces a sequence alignment with two outputs in the same process, which is different from existing programs. The impact is it can reduce the waiting time for users. This research needs to be considered as a form of attention from researchers in bioinformatics and has better features than existing programs.

REFERENCES

[1] F.I. Sholeh, “Constraint Programming-based Refinement of Multiple Sequence Alignments”, Master Thesis, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa. 2016.

[2] Ainsworth David, Michael J.E. Sternberg, Come Raczy, dan Sarah A. Butcher, “k-SLAM: accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets”. *Nucleic Acids Research*, Volume 45, Issue 4, 28 February 2017, Pages 1649–1656.

[3] Baxevanis AD, Ouellette BFF. 2001. Bioinformatics A Practical Guide to the Analyses of gene and Proteins. A John Wiley & Sons, Inc., Publication.

[4] S Petti, N Bhattacharya, R Rao, J Dauparas, N Thomas, J Zhou, A.M. Rush, P Koo, andS Ovchinnikov,” End-to-end learning of multiple sequence alignments with differentiable Smith–Waterman” Bioinformatics, Volume 39, Issue 1, p 1-7 https://doi.org/10.1093/bioinformatics/btac724. 2022

[5] 4 A. Baby Jerald and T. R. Gopalakrishnan Nair, "Influenza virus vaccine efficacy based on conserved sequence alignment," *2012 International Conference on Biomedical*

Engineering (ICoBE), Penang, 2012, pp. 327-329. doi: 10.1109/ICoBE.2012.6179031.

[6] MB. Ulum, “Perancangan Alternatif Potensial Primer Selulase Dengan Teknik *Multiple Sequence Alignment”. Master Tesis. Dept. Ilmu Komputer IPB. Bogor. 2013.*

[7] AA Sunarto, WAKusuma, H Sukoco. 2013. “Parallelization of star alignment”. Conference: IEEE Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME), DOI: 10.1109/ICICI-BME.2013.6698486

[8] Needleman S. and Wunsch, C. D. 1987. A general method applicable to the search for similarities in the amino acid *sequence* of two proteins. J. Mol. Biol. 48, 443-453.

[9] Smith and Waterman. 1981. Identification of Common Molecular Subsequences. J. Mol. *Bwl.* (1981), 147, 195-197.

[10] Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421. doi:10.1186/1471-2105-10-421.

[11] Rice Peter, Ian Longden, dan Alan Bleasby. 2000. Emboss: The European Molecular Biology Open Software Suite. Rice, P., I. Longden, and A. Bleasby. "EMBOSS: the European Molecular Biology Open Software Suite." *Trends in genetics: TIG* 16.6 (2000): 276-277.

[12] NC. Jones, PA. Pevzner. An Introduction to Bioinformatics Algorithms. 2004