An Improved Needleman-Wunsch Algorithm for Pairwise Sequence Alignment of Protein-Albumin

Lailil Muflikhah, Dian Eka R.
Faculty of Computer Science, Brawijaya University
lailil@ub.ac.id

***Abstract*—This paper aims to improve the method of optimal global sequence alignment in order to increase the computational performance. The huge number of genome sequences is main problem of alignment. One of global sequence alignment methods is Needleman-Wunsch algorithm. This algorithm is implemented by constructing a MxN matrix, which M is the length of first sequence and N is the length of second sequence. All cells of the matrix are filled to compute the score for constructing global pairwise sequence alignment, so that the time and space complexity are very high. Therefore, the improved Needleman-Wunsch algorithm (INWA) is addressed to compute partially for the score in the cells. The test set consisted of 1250 pairwise sequence alignments of human protein-albumin and it is compared to the original method. As a result shows that the space and time compexity of INWA is O(N) instead of O(MN).**

***Index Terms*— Genome; Needleman-Wunsch; Protein-Albumin; Sequence Alignment;**

I. INTRODUCTION

Information technology in the field of molecular biology which is known as bioinformatics is growing very rapidly. The activities include mapping and DNA analysis, protein sequencing, aligning different DNA, constructing and performing models of protein structures in three dimensions. Several methods are developed in order to increase the performance result of DNA computing in biomolecular analysis. One of them is to compare two methods for encoding process using Restriction Enzym for DNA computation in cutting process [1]. Another activity is protein sequence alignment. It is a technique in bioinformatics to discover regions of similarity that may indicate functional, structural, and/or evolutionary relationship between two biological sequences (protein or nucleic acid)[2]. Basically, there are two methods of sequence alignment including optimal global sequence alignment and optimal local sequence alignment. Several algorithms of dynamic programming approach give global optimal result (e.g. Needleman-Wunsch [3], Smith-Waterman[4]). However, the computational time of these algorittms is high.The other algorithms are using heuristic approach (e.g. FASTA[5], BLAST[6]). The last one gives fast computational time but it only achieves local optimal sequence aligment.

In October 8th 2015, protein sequence data which is stored at National Center for Biotechnology Information (NCBI) is 51.933.925 [7]. The amount of genome sequences become main problem in bioinformatics due to time requirement for analysis. The related research has been conducted to optimize the computational protein sequence alignment using a dynamic programming algorithm but the computational complexity is still high of O (mn) by Zhou and Chen [8]. Furthermore, the developing performance of dynamic programming has been applied through sharing memory to speed up the alignment process. The dynamic programming method for sequence alignment has developed with share memory system using four different data partitioning schemas: blocked columnwise, rowwise, antidiagonal, and revised blocked columnwise [9]. Furthermore, the other study for improving the accuracy of alignment result has been applied by optimizing scoring function [12].[10]

However, this research is conducted to improve the previous work conducted by Shehab et,al. using Fast Dynamic Algorithm without sharing memory. The previous research had limitation in scalability, which the input sequences cannot have different length more than of ten elements[11]. Therefore, we develop this prevous research by modifying Needleman-Wunsch method as a dynamic programming approarch in single memory for pairwise sequence alignment for protein-albumin. This method is address to get high performance in computational process with limited reseource of memory.

II. BACKGROUND

In sequence alignment, basically there are two approaches to create sequence alignments, i.e. global alignment and local alignment.[2]

A. *Local Alignment*

Local alignment only aligns the most similar region in a sequence. There is no need to align the whole sequence, but it is only the region whose highest similarity based on certain criteria. The local alignment is presented as in Figure 1.

```
EARDFNQYYSSIKRSGSI
       :::::::
EPKLFIQYYSSIKRTMGH
```

Figure 1. Local alignment

Short sequence can be aligned simply by performing its structure. However, it becomes complicated for the long sequence. Pairwise alignment is now widely used in bioinformatics to align two sequences.

B. *Global Alignment*

Global alignment creates end-to-end alignment even though there is a difference in some regions. This approach is suitable for aligning similar sequence [2].

```
EARDFNQYYSSIKRSGSI
:...:..:::::::....
EPKLFIQYYSSIKRTMGH
```

Figure 2. Global alignment

Furthermore, the objective of sequence alignment for protein-albumin data is to match as much as possible for identical amino acid. It is the same as global sequence alignment. As illustration, given an alignment between the sequences $A = GTASCDG$ and $B = GTASNND$.

```
A = GTASC-DG
    :::::. :
B = GTASNND-
```

Figure 3. Sequence alignment

In Figure 3, there are six match elements and they are marked by colon (:), one mismatch element is marked by dot (.) and two gaps are marked by dash (-) in the sequence. The alignment can be achieved by insertion, deletion or substitution. The alignment result reflects the changes that have occurred during evolution [12]. The figure can be interpreted that sequence A is transformed into sequence B by following process:

1. To substitute the first C for a N
2. To insert N after substitution
3. To delete the last G

After alignment is created, a score will be assigned to each pair of the aligned letter according to the selected score. The scored scheme will determine the grade for match, mismatch and gap. The final score is computed by summing up all pairwise sequence of letters. For example, the scored scheme of match= +2,mismatch=-1 and gap = -1, so that score of the alignment as in Fig 3 can be computed as (2×5)+(1×(-1))+(2×(-1))=7. The best alignment is chosen by the highest score among all possible alignment. The best alignment is also known as optimal alignment or global optimal alignment. The score value of pairwise alignment depends on the scoring scheme.

There are several previous works of sequence alignment based on dynamic programming. Dynamic programming is a method to build solution by resolving the previous sub problem and then to combine all sub solution as final solution [9]. The first bioinformatics algorithms using dynamic programming are Needleman-Wunsch (for global alignment) and Smith-Waterman (for local alignment). .

A Needleman-Wunsch algorithm computes similarity between two sequences by making M×N matrix. M is the length of the first sequence, N is the length of the second sequence). Then it is filled the entire matrix of the score to perform alignment. This algorithm always produces optimal solution for the alignment. After filling the entire matrix, this algorithm will do a backtracking form end to end point to build the aligned sequence [3].

For example, there are two sequence proteins, $A = A_1, A_2, A_3, \dots A_m$ and $B = B_1, B_2, B_3, \dots B_n$. To perform the alignment, matrix F is constructed which $F(i,j)$ has the highest score of the alignment between $A_{1:i}$ and $B_{1:j}$ By initialize $F(0,0) = 0, F(i,0) = -i$ and $F(0,j) = -j$, the matrix F is constructed by the following recursion:

$$F(i,j) = max \begin{cases} F(i-1,j-1) + s(A_i, B_j, p) \\ F(i-1,j) \\ F(i,j-1) \end{cases} \quad (1)$$

We define $s(A_i, B_j, p)$ as the function by returning the scoring scheme for a match if $A_i = B_j$ and a mismatch if $A_i \neq B_j$. The final score can be obtained by looking the value of $F(M,N)$. The alignment is constructed by backtracking from $F(M,N)$ to $F(0,0)$. The procedure is compare with value on F(i,j) to its top left and diagonal using Equation 1. For example if $F(i,j) = F(i,j-1)$, then the backtrack record of an insertion in B_j [3]. To get the best performance, the data of an insertion, deletion or gap can be recorded during filling the matrix. Sometimes, it can be generated from this process more than one path. This indicates that there are more than one generated alignments with the same score.

The related research of DNA sequence is developed a parallel solution to achieve optimal solution [13]. The main work is shared memory parallel architecture and multicore CPUs as well as advanced shared memory platform, which is hardware improvement. Hardware improvement is a limited factor thus we have to develop software improvement.

Recently, the study on sequence alignment is developed a solution to reduce computational time by scoring only three main diagonal of matrix [14]. This work produces high performance and low complexity. If a pair sequences has the same length then the space complexity is $3M + 1$. Otherwise, if a pair sequence has different length then the space complexity is $3M + 2$. Unfortunately, this work cannot produce global optimal solution if the difference in length of two sequences is quite long. For example, the first sequence has length of 35 and the second sequence has length of 20. The three main diagonal cannot cover optimal alignment as shown in Figure 4.

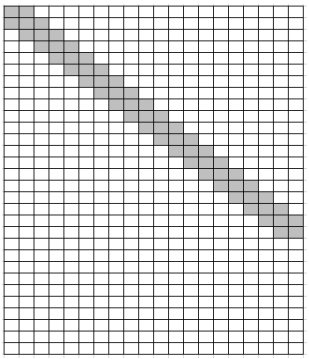

Figure 4. Three main diagonal of 35x20 matrix

The grey block in Figure 4 is the main diagonal which contains the alignment result and the remain is empty data. The alignment result is not optimal due to there is no backtracking from end to end point (bottom-right-corner to top-left-corner).

III. Proposed Algorithm

The proposed algorithm of this research is a modified Needleman-Wunsch algorithm to fill and to compute the score scheme into matrix partially. It is called as INWA. There are two possibilities for aligning sequences. First, the pairwise sequences have the same length and second, the pairwise sequences have the different length. The INWA can handle these two possibilities by different method in selecting

area to fill the matrix. There are four main steps of the INWA as below:

Step 1: Marking area to fill the matrix

The INWA will detect whether input sequences have the same length or not. If the both sequences have the same length, then three main diagonal will be marked. In this case, INWA method is similar to the Fast Dynamic Algorithm [11]. The three diagonals consist of one main diagonal(D), one diagonal above the main diagonal ($D+1$) and one diagonal below the main diagonal ($D-1$). The marked area can be seen at Figure 5.

D	D+1			
D-1	D	D+1		
	D-1	D	D+1	
		D-1	D	D+1
			D-1	D

Figure 5. Marked area for the same length of input sequences

If the input sequences have the different length, then the matrix will have two main diagonals. In this case, the marked areas are the first main diagonal ($D1$), the second main diagonal ($D2$), area between $D1$ and $D2(x)$, one diagonal above $D2$ ($D2+1$) and one diagonal below $D1$ ($D1-1$). The other hand, the marked areas of the sequences input with different length can be seen at Figure 6.

D1	x	x	x	D2	D2+1		
D1-1	D1	x	x	x	D2	D2+1	
	D1-1	D1	x	x	x	D2	D2+1
		D1-1	D1	x	x	x	D2

Figure 6. Marked area for the different length of input sequences

The space and time complexity which is required to fill the marked area of INWA is $O(N)$, where N is length of the longest input sequence. This performance is applied to the both cases, either the same length of input sequences or different length of input sequences. In the original Needleman-Wunsch algorithm, time and space complexity which is required to fill the entire matrix is $O(MN)$. Therefore, the performance of INWA is better than the performance of the original Needleman-Wunsch algorithm.

Step 2: Initialization

The initialization process of INWA is the same as initialization on the original Needleman-Wunsch algorithm as explained in related works. This process aims to fill the first column and first row in marked area.

Step 3: Filling the marked area in matrix

The process of filling the matrix in INWA has the same procedure as the original Needleman-Wunsch as explained in previous works. How ever, in INWA case, it is not applied to get maximum value as stated on Equation 1. However, it must obtain at least one value. The missing value is not to be considered.

Step 4: Backtracking and build the alignment

The backtracking phase in INWA and original Needleman-Wunsch algorithm is exactly same. It is started from $F(M,N)$ until $F(0,0)$ using procedure that is explained in related works.

A. INWA Case Study 1

This case assumes that the input sequences has same length. Given the first sequence, $A = GESKC$ and the second sequence, $B = GTASC$. Scoring scheme for the alignment is $match = 2, mismatch = -1$ *and* $gap = -2$. The alignment process using INWA is illustrated in Figure 7. The marked area in this case study is three main diagonals.

A.1. Marking area to fill in the matrix

	1	2	3	4	5	6	7
1		_	G	T	A	S	C
2	_						
3	G						
4	E						
5	S						
6	K						
7	C						

Figure 7. Marked area from case study 1

A.2. Initialization

Initially, filling $F(0,0)$ is 0 ($zero$). Then, the first row and first column of marked area is iterated gap value as in Figure 8. In this case, the form of flow data value is represented by arrow (↑, ← and ↖). This arrow is useful for backtrack process.

	1	2	3	4	5	6	7
1		_	G	T	A	S	C
2	_	0	-2←				
3	G	-2↑					
4	E						
5	S						
6	K						
7	C						

Figure 8. Initialization from case study 1

A1.3. Filling the marked area in matrix

After all marked areas are filled, the matrix is shown as in Figure 9.

	1	2	3	4	5	6	7
1		_	G	T	A	S	C
2	_	0	-2←				
3	G	-2↑	2↖	0←			
4	E		0↑	1↖	-1←		
5	S			-1↑	0↖	1↖	
6	K				-2↑	-1↑	0↖
7	C					-3↑	1↖

Figure 9. All marked area has been filled from case study 1

A1.4. Backtracking and build the alignment

The backtracking process can be seen at Figure 10.

	1	2	3	4	5	6	7
1		␣	G	T	A	S	C
2	␣	0	-2				
3	G	-2	2↖	0			
4	E		0	1↖	-1←		
5	S			-1	0	1↖	
6	K				-2	-1↑	0
7	C					-3	1↖

Figure 10. Backtracking path from case study 1

The diagonal arrow (↖) means substitution, vertical arrow (↑) means insertion of a gap in B and horizontal arrow (←) means insertion of a gap in A. It is constructed the alignment steps according to Table in Figure 10 and the result is as follow:

Sequence A : G E - S K C
Sequence B : G T A S - C
Score : 2 -1 -2 2 -2 2

The final score can be computed by summing up $2 + (-1) + (-2) + 2 + (-2) + 2 = 1$ or it is shown at $F(7,7) = 1$.

B. INWA Case Study 2

This case assumes that the length of input sequences have different length. If it is given the sequence $A = GPTGTGESKC$ and sequence $B = GTASC$, then scoring scheme for the alignment are $match = 2, mismatch = -1$ and $gap = -2$. The alignment steps using INWA are as follow:

B.1. Marking area to fill in the matrix

The marked area in this case can be seen at Figure 11.

	1	2	3	4	5	6	7
1		␣	G	T	A	S	C
2	␣						
3	G						
4	P						
5	T						
6	G						
7	T						
8	G						
9	E						
10	S						
11	K						
12	C						

Figure 11. Marked area from case study 2

B.2. Initialization

The matrix $F(0,0)$ is input by 0. Then, filling at the first row and column of marked area are iterated by gap value as in Figure 12.

	1	2	3	4	5	6	7
1		␣	G	T	A	S	C
2	␣	0	-2←				
3	G	-2↑					
4	P	-4↑					
5	T	-6↑					
6	G	-8↑					
7	T	-10↑					
8	G	-12↑					
9	E						
10	S						
11	K						
12	C						

Figure 12. Initialization from case study 2

B.3. Filling the marked area in matrix

All marked cells of matrix are filled as in Figure 13.

	1	2	3	4	5	6	7
1		␣	G	T	A	S	C
2	␣	0	-2←				
3	G	-2↑	2↖	0←			
4	P	-4↑	0↑	1↖	-1←		
5	T	-6↑	-2↑	2↖	0←	-2←	
6	G	-8↑	-4↑	0↑	1↖	-1←	-3←
7	T	-10↑	-6↑	-2↑	-1↑	0↖	-2←
8	G	-12↑	-8↑	-4↑	-3↑	-2↑	-1↖
9	E		-10↑	-6↑	-5↑	-4↑	-3↑
10	S			-8↑	-7↑	-3↖	-5↑
11	K				-9↑	-5↑	-4↖
12	C					-7↑	-3↖

Figure 13. All marked area has been filled from case study 2

B.4. Backtracking and build the alignment

The backtracking process can be seen in Figure 14. It is constructed by the score alignment as follow:

Sequence A : G P T G T G E S K C
Sequence B : G - T A - - - S - C
Score : 2 -2 2 -1 -2 -2 -2 2 -2 2+(-2)

The final score can be computed by sum up $(2 + (-2) + 2 + (-1) + (-2) + (-2) + (-2) + 2 + (-2) + 2 = -3$ or we can see at $F(12,7) = -3$.

	1	2	3	4	5	6	7
1		␣	G	T	A	S	C
2	␣	0	-2				
3	G	-2	2↖	0			
4	P	-4	0↑	1	-1		
5	T	-6	-2	2↖	0	-2	
6	G	-8	-4	0	1↖	-1	-3
7	T	-10	-6	-2	-1↑	0	-2
8	G	-12	-8	-4	-3↑	-2	-1
9	E		-10	-6	-5↑	-4	-3
10	S			-8	-7	-3↖	-5
11	K				-9	-5↑	-4
12	C					-7	-3↖

Figure 14. Backtracking path from case study 2

IV. EXPERIMENTAL RESULT

This research is examined to data set of 215 protein-albumin sequences (1250 pairwise sequence}. It can be downloaded from National Center of Biotechnology Information (NCBI) at url: http://www.ncbi.nlm.nih.gov. As illustration, there are two protein sequences, i.e. isoform CRA_q and isoform CRA_p as follows:

isoform CRA_q :

```
MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEE
NFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADES
AENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERN
ECFLQHKDDNPNLPRLVRPEVDVMCTAFHDNEETFLKKYLY
EIARRHPYFYAPELLFFAACCQSSMNFGMKGRLRLPNRDSS
VPVSKNLEKELSKHGQ
```

isoform CRA_p :

```
MSQLKICELFEQLGEYKFQNALLVRYTKKVPQVSTPTLVEV
SRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKT
PVSDRVTKCCTESLVNRRPCFSALEVDETYVPKEFNAETFT
FHADICTLSEKERQIKKQTALVELVKHKPKATKEQLKAVMD
DFAAFVEKCCKADDKETCFAEEGKKLVAASQAALGL
```

The both sequences are aligned globally using the proposed method with default parameter value (match=1; mismatch=-1; gap=-1) and the result is as below:

```
. Alignment score     :-9
  Length of sequence 1:221
  Length of sequence 2:200
  Sequence identity   :29/274 (0.11%)
  Positives           :49/274 (0.18%)
  Gaps                :127/274 (0.46%)
  HSSP                :-9.74 (not similar)
  Filled Matrix       :4822
  Execution Time : 49 ms
```

Table 1
The number of filled matrix areas

Pair sequence no.	Original NW	Proposed method (INWA)	The score of result
1	99645	52339	Same
2	89790	42484	Same
3	118479	71173	Same
4	94827	47521	Same
5	271998	224692	Same
6	114756	67450	Same
7	131400	84094	Sane
8	36792	9070	Same
9	76212	28906	Same
10	139941	92635	Same
…	…	…	…
…	…	…	…
1250	41925	37893	same

In the experimental result, the number of the filled matrix areas with the same score for the both methods are shown in Table 1. The evaluation is implemented by ten trials against 1250 pairwise sequence alignments of human protein-albumin. Furthermore the computational time is required using original Needleman-Wunsch of 5.382 second and the other one using INWA of 3.988 second. It means that there is time reduction as 25.9% of the original method. The alignment results for case study 1 and case study 2 by INWA are the same as Needleman-Wunsch. This means that the computational time of purposed method is reduced but it performs the optimal alignment result. Furthermore, the time and space complexity of the proposed method (INWA) are *O(N).* They are less than the original method NW of *O(MN).*

V. CONCLUSION

An improvement of Needleman-Wunsch algorithm (INWA) has been applied to align pairwise sequence for human protein-albumin. The main idea of the proposed method is to skip unused data by remaining blank area in order to obtain the least computational time and to reduce space complexity. INWA only fill cell of matrix partially. This algorithm can be applied to the both kind of sequence alignment, either the input sequences have the same length or not. Furthermore, the space and time complexity of INWA is O(N). It is better than the original Needleman-Wunch algorithm that is O(MN). Also, the running time of INWA is 25.9% faster than the original Needleman-Wunsch algorithm.

REFERENCES

[1] Rajaee, N., Ahmad S., Awang, Zulkharnain, A., "Comparison between Double Stranded DNA with Restriction Enzymes and Single Stranded DNA with Primers for Solving Boolean Matrix Multiplication", *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 8 no. 12, pp.5-8, 2016.

[2] EMBL-EBI, "Pairwise Sequence Alignment," 2015. [Online]. Available: http://www.ebi.ac.uk/Tools/psa/. [Accessed 8 10 2015].

[3] S. Needleman and C. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two sequences," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443-453, 1970.

[4] T. Smith and M. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, pp. 195-197, 1981.

[5] W. Pearson and D. Lipman, "FASTA: Improved tools for biological sequence comparison," in *Proceedings of the National Academy of Sciences*, USA, 1988.

[6] S. F. Altschul, W. Gish, W. Miller, E. Myers and D. Lipman, "Basic Local Alignment Search Tool," *Journal of Molecular Biology*, vol. 215, pp. 403-410, 1990

[7] NCBI, "NCBI," 2015. [Online]. Available: http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome. [Accessed 6 May 2015]

[8] Zhou Zm, Chen Z-w. Dynamic Programming for Protein Sequence Alignment. *International Journal of Bio-Science and Bio-Technology*. 2013; 5(2): 141-150.

[9] Rahmad A., Auriza, Sukoco H., Kusuma, A.W. Comparison of Data Partitioning Schema of Parallel Pairwise Alignment on Shared Memory System. *TELKOMNIKA, Telecomunication, Computing, Electronics and Control.* 2015; 13(2): 694-702

[10]. Yamada, Kazunori D, Optimixing Scoring Function of Dynamic Programming of Pairwise Profile Alignment using Derivative Free Network, Graduate School of Information Sciences, Tohoku University Sendai, Japan,Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan. arXiv:1708. 09097v2 [q-bio.QM]. September 2017

[11] S. A.Shehab, A. Keshk and H. Mahgoub, "Fast Dynamic Algorithm for Sequence Alignment based on Bioinformatics," International *Journal of Computer Applications*, vol. 32, no. 7, pp. 54-61, 2012

[12] J. Xiong, "Essential Bioinformatics," Cambridge University Press, p. 4. 2006.

[13] N. C. Jones and P. A. Pevzner, An Introduction to Bioinformatics Algorithms, S. Istrail, P. Pevzner and M. Waterman, Eds., Massachusetts: The MIT Press, 2004.

[14] H. A. S. A. S. M. F. T. Ahmad M. Hosny, "An Efficient Solution for Aligning Huge DNA Sequences," in ICCES, Cairo, 2011.