TELKOMNIKA, Vol.15, No.4, December 2017, pp. 1785~1793
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI:10.12928/TELKOMNIKA.v15i4.5813

Pairwise Sequence Alignment between HBV and HCC Using Modified Needleman-wunsch Algorithm

Lailil Muflikhah*, Edy Santoso
Faculty of Computer Science, Brawijaya University, 8 Veteran Road, Malang Indonesia
*Corresponding author, e-mail: laililmf@gmail.com

Abstract

Ths paper aims to find similarity of Hepatitis B virus (HBV) and Hepatocelluler Carcinoma (HCC) DNA sequences. It is very important in bioformatics task. The similarity of sequence allignments indicates that they have similarity of chemical and physical properties. Mutation of the virus DNA in X region has potential role in HCC. It is observed using pairwise sequence alignment of genotype-A in HBV. The complexity of DNA sequence using dynamic programming, Needleman-Wunsch algorithm, is very high. Therefore, it is purpose to modifiy the method of Needleman Wunsch algorithm for optimum global DNA sequence alignment. The main idea is to optimize filling matrix and backtracking proccess of DNA components.This method can also solve various length of the both sequence alignment. This research is applied to DNA sequence of 858 hepatitis B virus and 12 carcinoma patient, so that there are 10,296 pairwis of sequences. They are aligned globally using the purposed method and as a result, it is achieved high similarity of 96.547% and validity of 99.854%. Furhthermore, this method has reduced the complexity of original Needleman-Wunsch algorithm The reduction of computational time is as 34.6% and space complexity is as 42.52%.

Keywords: *hepatitis, sequence alignment, DNA, Needleman-Wunsch, optimum global*

1. Introduction

Hepatitis B virus (HBV) is a DNA virus which can causes acute and chronic hepatitis in humans. The chronic hepatitis can further progress to liver cirrhosis and carcinoma of the liver (hepatocellular carcinoma, HCC). There is prediction of more than 500,000 incidents in 2000, HCC is the most common malignant tumor in the world and the attack rate is rising in many countries, including Indonesia. It is the third rank in cause of death after lung cancer and stomach [1].

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. The correction for DNA sequence can be observed before analyze the sequence. This reseearch has been conducted using Fuzzy-based spectral alignment [2]. Another related research of DNA sequence is applied to identify Breast Cancer disease [3].

However, the growth of DNA sequence databases makes the sequence alignment computation, as one of bioinformatics tasks, increase significantly [4]. The complexity of sequence alignment is O (LN), where L is the length of each sequence and N is the number of sequences [5]. The longer and the more sequences to be aligned the longer it's computational time. The optimal time of sequence alignment is increased exponentially with increaseing of the number and length of sequences [6].

Basically, there are two methods of sequence alignment including dynamic programming method and heuristic method approach. The Needleman-Wunsch algorithm is a method of optimal global sequence alignment, but Smith-Waterman is a method of optimal local alignment which are based on dinamic programming approach algorithm. Meanwhile, several tools that use heuristic approach are FASTA and Basic Local Alignment Search Tool (BLAST). The produced sequent alignment is not optimal eventhough their computational speed is fast [7]. Several methods or tools are needed to obtain optimal computational process. The previous research to optimize the computational protein sequence alignment used a dynamic programming algorithm but the computational complexity is high of *O (mn)* by Zhou and

Received August 1, 2017; Revised November 8, 2017; Accepted November 23, 2017

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]. Another research is also parallel computing utilized on clusters of computers known as Distributed Memory. This research used star algorithms in parallel environment using MPI to distribute computing of DNA Multiple sequence Alignment [10].

This research is purposed 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 difference of sequence length cannot more than of 10 [11]. Therefore, we develop this prevous reseach by modifying Needleman-Wunsch method as a dynamic programming approarch in single memory for pairwise sequence alignment in order to get the similarity rate between HBV and HCC. This method is address to get high performance in computational process with limited reseource of memory.

2. Research Method

Implementation of pairwise sequence alignment is designed within three parts as shown in Figure 1. First, the input sequences is data set of HBV and HCC and parameter for evaluation. The second is applied the purposed method to align the sequences. The pusposed method is address to improve the performance, including computaional time and space complexity.

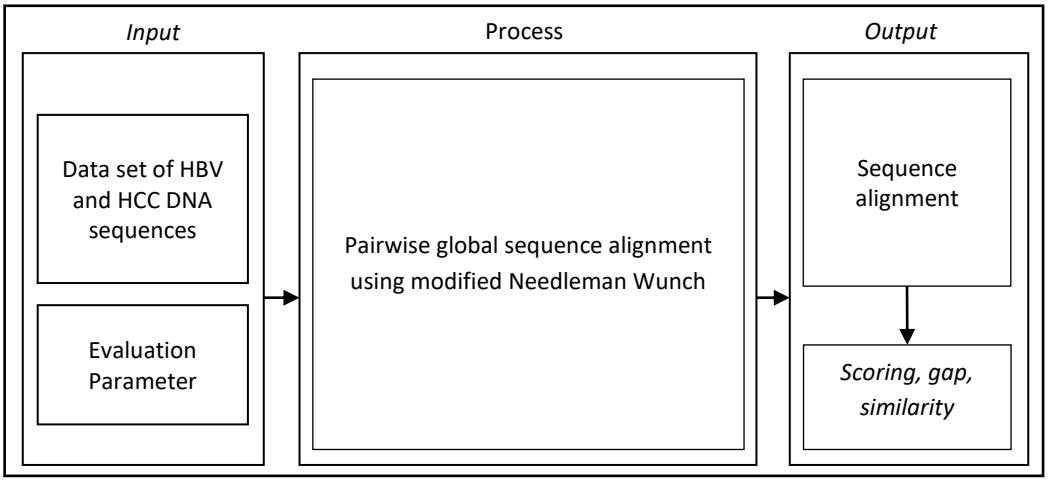

Figure 1. Block diagram of pairwise sequence alignment

2.1. Data Sets

There are various ways to identify the disease, either isolating the virus in the laboratory or analysing DNA sequences of the viruses. Hepatitis B virus is a DNA virus to mutate and it influences to be chronic disease. It is known as hepatitis with liver cancer (HCC). One way to know that the HBV DNA is mutated, can be applied sequence alignment based on the similarity percentage.

This research is examined to data set of HBV DNA sequences, with spesific genotype-A in X region. It can be downloaded from website of HBV database at url: https://hbvdb.ibcp.fr/HBVdb/HBVdbDataset?view=/data/nucleic/fasta/A_X.fas&seqtype=0gnl|hbvcds|AB014370_X_P-A. There 858 HBV DNA sequences and 12 HCC DNA sequences. They are applied to sequence in pairwise, so that there are totally 10296 in pairs.

2.2. Sequence Alignment

Sequence Alignment is the process whereby a sequence compared by looking for patterns of of the most common characters and inter-related sequences. Pairwise sequence alignment is a alignment process between two sequences based on similarity in order to search on the database and multiple sequence alignment. This similarity can be expressed in percentages. It means that the sequences have similarities to physical-chemical properties. Another term, the degree of similarity in amino acid sequence is similar between the two protein sequences [12]. The similarity may indicate a functional relationship, or the structural and evolutionary relationships [13].

To perform sequence alignment, it is necessary to shift sequence at certain positions in order to get the maximum similarity. There are two method of sequence alignment, ie global and local alignment. The global alignment method is to compare the whole sequences. This method is suitable to perform sequence alignment that has a high degree of similarity of the whole part of the sequence. Another hand, the local alignment method is to compare a local area or part of the sequence.

2.1.1. Global Alignmnt

Global alignment creates end-to-end alignment even though there is a difference in some region. This approach is suitable to align similar sequence [8]. As illustration, it can be showed in Figure 2.

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

Figure 2. Global alignment

2.1.2. Local Alignment

Local alignment only aligns the most similar region within a sequence. There is no need to align all part of sequence, only the region with big similarity according to some criteria. By using the same sequence is as Figure 3, local alignment becomes as follow:

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

Figure 3. Local alignment

2.2. Needleman Wunch Algorithm

Needleman-Wunsch algorithm is a dynamic programming method to find global optimal sequence alignment and it allows a gap. This algorithm is widely used in bioinformatics to align nucleotide sequences. It works by creating an optimal sequence alignment. As illustration, there are two DNA sequences of length m and n where $a=a_1a_2 \ldots a_m$ and $b=b_1b_2 \ldots b_n$. To find a parallel sequence with the highest score, it must be constructed a matrix F where F (i, j) is the value of the best sequence alignment between $a_{1:i}$ and $b_{1:j}$. The first, it is constructed matrix F and is filled in F (0,0)=0, F (i, 0)=-i and F (0, j)=j. Furthermore, it is applied as Equation 1 [14]:

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

It is.defined $s(a_i, b_j, p)$ which is the returned function to give 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 build by doing backtrack from F(M,N) to F(0,0). The procedure is compare to the 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 backtrack is recorded an insertion in b_j [2]. To get better performance, the data of an insertion, deletion or gap can be recorded during filling the matrix. Sometimes, several path can be generated from this process. This indicates that there are more than one alignment that can be generated with same score. The steps of the Needleman Wunsch is illustrated in Figure 4.

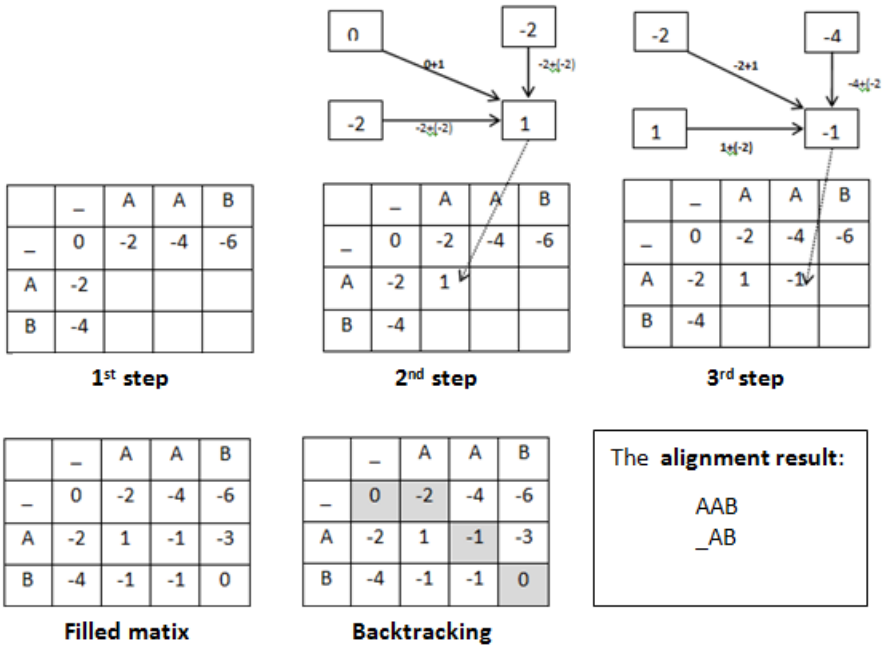

Figure 4. The steps of Needleman Wunch Algorithm

3. The purposed Method

The Needleman Wunsch algrithm is a dynamic programming approach to align global optimally. How ever, this algoritm needs high computational time and space complexity. Therefore, it is need to optimize the filling and backtracking process. The main deferrence of the purposed method is filling process of matrix alignment with spesific area as in Figure 5. If the both sequences have the same length, then filling mattrix is on the main diagonal (D), (D-1) and (D+1). The (D-1) is a diagonal below the main diagonal and (D+1) is a diagonal above the main diagonal. In this case, we use the same with FAST Needleman Wunsch algorithm [5].

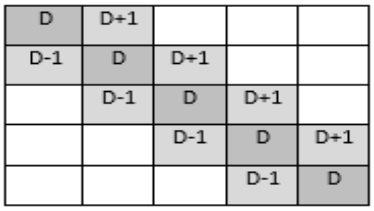

Figure 5. Ilustration of filling matrix with the same sequence

The other hand, if the input sequence has different length, the matrix will have two main diagonals. In this case, the marked area are including 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) as shown in 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 if the sequences input have different length

Furthermore, the stages of the modified Needleman Wunsch are as follows:

a. 1st step: Matrix Initialization

In this step, it is to construct matrix NW and to put a certain value for marking into specific area as in Figure 5 or Figure 6. It will be given parameter value of gap openning and gap extension.

b. 2nd step: Filling to the matrix

There are two operation in this step as follows:

1. To count Linear evaluation

Each nucleotide is given the predetermined parameter value, including match, mismatch, gap opening and gap extention. As illustration, it is given the sequence A=GESKC and sequence B=GTASC. The scoring scheme for the alignment is match=2, mismatch=-1 and gap=-2. Filling F(0,0) by 0 and also filling the first row and first column of marked area is iterated gap value as in Figure 7. In this case study, the data where is the value come form is represent 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. 7. Initialization stage

2. To fill NW matrix and tracer

After all the rows and columns of the marked matrix NW are counted, then the next stage is to fill the matrix NW and tracer. To perform this stage, it takes the output from the previous process in the form of a direction to tracer process later (D=diagonal), (H=horizontal) and (V=vertical). After all marked area has been filled, the matrix will be look like Figure 8.

	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 8. All marked area has been filled

c. 3rd step: Backtracking process

Backtracking is the last stage of sequence alignment. At this stage, the system will perform a trace back from the matrix that has been previously established. Starting from the end entry of trace process (end position of matrix), it will be executed the alignment and it results refer to the rules that have been determined. The backtracking process can be seen at 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. Backtracking path

The diagonal arrow (↖) means substitution, vertical arrow (↑) means insertion of a gap in B and horizontal arrow (←) means insertion of a gap in A. From Figure 9 we can construct the alignment is as follows:

```
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 sum up 2+(-1)+(-2)+2+(-2)+2=1 or we can see at F(7,7)=1

4. Results and Analysis

There are two scenarios in this research. First scenario is finding characteristics for DNA sequences of HBV and HCC based on similarity using either modified or original Needeman Wunsch. The second scenario is comparing performance result between the modified method and the original. This scenario is also to analysis the parameter value against the performance.

In the first scenario, it is used the default parameter (match=2; mismatch=-3; gap opening=-5; gap extension=-2). As a result, it is achieved using the modified method's performance including average similarity rate=96.549% and score=849.8 with computational time=1370 ms. How ever, the performance is achieved using original Needleman Wunsch, i.e. similarity rate= 96.548% and score = 849.8 with total computational time=64131ms. Therefore, the accuracy rate of the both method is 99.854% and there is time reduction as 78.95%. The performance of experimental result is partially showed in Table 1

Table 1. Comparisson of Performance the Both Algorithm for Pairwise Sequence Alignment

Sequence no	Modified Needleman-Wunsch			Original Needleman-Wunsch		
	score	Similarity (%)	Time (ms)	score	Similarity (%)	Time (ms)
1	900	98,71	50	900	98,71	90
2	860	96,99	38	860	96,99	28
3	875	97,63	37	875	97,63	19
4	900	98,71	5	900	98,71	10
5	900	98,71	3	900	98,71	11
6	905	98,92	36	905	98,92	8
7	895	98,49	4	895	98,49	10
8	895	98,49	7	895	98,49	11
9	895	98,49	3	895	98,49	13
10	840	96,13	6	840	96,13	13

Then, the second scenario is analisys the parameter against the computational performance. By using various parameters including match, mismatch, gap opening and gap

extention is obtained similarity rate of the original method as shown in Figure 10. The match, gap opening and gap extention show that the larger parameter value is given, the lower similarity is. Otherwise, the lower parameter value of mismatch, the higher similarity is.

The performance measure including computational time and space complexity is evaluated. Using parameter randomly, there is reduction of computational time in average of 34.60% as in Table 2. In space compexity, it is evaluated by filling the matrix of sequence alignment and the partial result is showed as in Table 3.

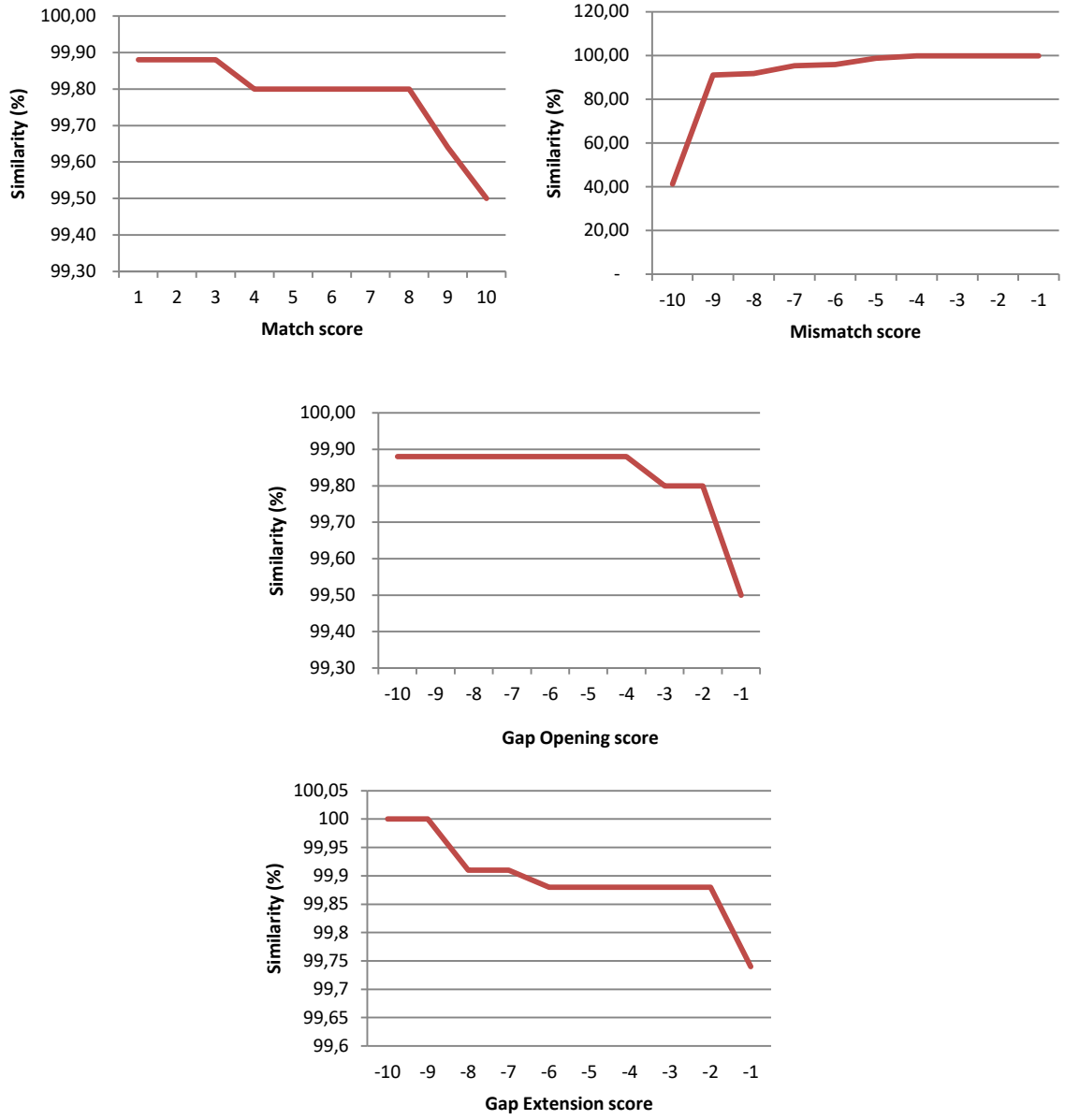

Figure 10. The effect of parameter value to similarity performance

Table 2. Computational Time for Sequence Alignment

Match	Mismatch	Gap Openning	Gap Extention	Time requirement for sequence alignment			
				Modified NW	**Original NW**	**Reduction (ms)**	**Reduction (%)**
1	-1	-5	-2	252474	322590	70116	21,74
1	-2	-5	-2	509065	618041	108976	17,63

Table 2. Computational Time for Sequence Alignment

Match	Mismatch	Gap Openning	Gap Extention	Time requirement for sequence alignment			
				Modified NW	**Original NW**	**Reduction (ms)**	**Reduction (%)**
1	-1	-2	-4	155513	306827	151314	49,32
1	-2	-2	-4	160639	313108	152469	48,70
2	-3	-2	-2	228456	320816	92360	28,79
2	-3	-2	-4	159961	311777	151816	48,69
2	-3	-4	-4	242231	373504	131273	35,15
2	-3	-5	-2	203316	387511	184195	47,53
2	-3	-6	-2	235820	329285	93465	28,38
Average				238608,33	364828,78	126220,44	34,60

Table 3. The Number of Filled Matrix Area

Sequence #	Modified NW	Original NW
1	52339	99645
2	42484	89790
3	71173	118479
4	47521	94827
5	224692	271998
6	67450	114756
7	84094	131400
8	9070	36792
9	28906	76212
10	92635	139941

Furthermore, the modified method is evaluated for the effect of total sequences alignment against the performance. How ever, the experimental result is showed that the total sequence is no effect to the performance as shown in Figure 11 and Figure 12.

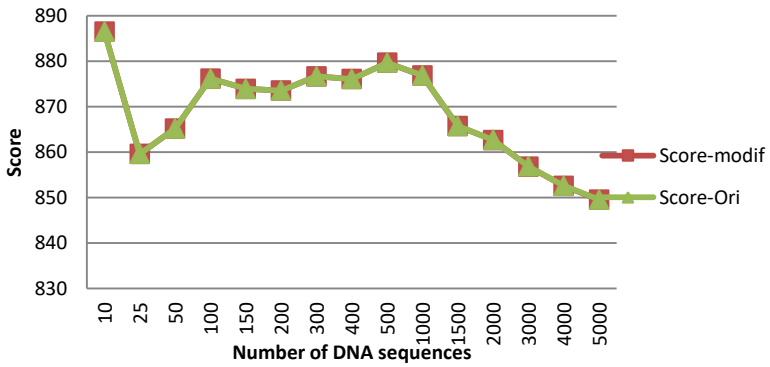

Figure 11. The performance of scoring alignment against total sequences

Figure 12. The performance of similarity against total sequences

5. Conclusion

The modified method of Needleman Wunch has been successfully implemented to perform the pairwise DNA sequence global alignment optimally between HBV and HCC:

1. The main principle in the modified method of Needleman Wunsch is the matrix region of filling DNA sequences for marking and backtracking process to get the sequence result.
2. The level of space and time complexity of the method are used by O(n) of the original method of O(mn). It is based on the experimental result that there is a reduction of filled matrix area of 42.52%.
3. Testing pairwise sequence alignment between hepatitis B virus and hepatocellular carcinoma DNA sequences is generated at high level of similarity rate in average value of 96 547%

6. Future Work

Based on the experimental result, there are still many shortcomings that needed some repairs for further research as follows:

a. In this study conducted by pairwise alignment (pairs), it is advisable to do multiple sequence alignment to increase speed up the computational time.
b. This research is obtained for the scoring and similarity, therefore for further research are expected to be able to determine the position of mutation in DNA sequences. Furthermore, it can be used to early detection of carcinoma (HCC) based on the HBV DNA sequence alignment.

Acknowledgement

This research was financially supported by Faculty of Computer Science, Brawijaya University, Indonesia under program Superior Research Grant in Faculty. We would like to thank Nashi Widodo Ph.D., Brawijaya University, for his fruitful and guidance.

References

[1] Lupberger J, Hildt E. Hepatitis B Virus-induced Ocogenesis. *World J.Gastroenterol*. 2007; 13(1): 74-81
[2] Saputra S, Kana, Buono, Agus, Kusuma WA. Fuzzy-based Spectral Alighment from Generation Sequencer. *TELKOMNIKA, Telecomunication, Computing, Electronics and Control*. 2016; 14(2): 707-714.
[3] Muflikhah L, Yuliantoro I. *Identifying Cancer Disease through Deoxyribonucleic Acid (DNA) Sequential Pattern Mining*. International Journal of Intelligence Science, 2017; 7: 9-23.
[4] Liu Y, Schmidt B, Maskell DL. MSA-CUDA. Multiple Sequence Alignment on Graphics Processing Units with CUDA. *IEEE International Conference on Application-specific Systems, Architectures and Processors*. 2009: 121-128
[5] Lloyd GS. 2010. Parallel Multiple Sequence Alignment: An Overview.
[6] Edgar RC, Batzoglou S. Multiple sequence alignment. *Current opinion in structural biology*. 2006; 16(3): 368-373.
[7] Kasap S, Benkrid K, Liu Y. Design and Implementation of an FPGA-based Core for Gapped BLAST Sequence Alignment with the Two-Hit Method. *Engineering Letters.* 2008; 16(3).
[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] Satra R, Kusuma WA, Sukoco H. Accelerating Computation of DNA Multiple Sequence Alignment in Distributed Environtment. *TELKOMNIKA Indonesian Journal of Electrical Engineering.* 2014; 12 (12): 8278-8285.
[11] A Shehab S, Keshk A, Mahgoub H. Fast Dynamic Algorithm for Sequence Alignment based on Bioinformatics. *International Journal of Computer Applications.* 2012; 32(7): 54-61.
[12] Xiong J. *Essential Bioinformatics.* 2006. New York: Cambridge University Press.
[13] EMBL-EBI, 2015. Pairwise Sequence Alignment. [Online] Available at: http://www.ebi.ac.uk/Tools/psa/ [available at 8 10 2015].
[14] Nanni L, Lumini A. Generalized Needleman–Wunsch algorithm for the recognition of T-cell epitopes. *Expert Systems with Applications.* 2008; 35:1463–1467.