Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
17/2 (2024), 167-174. DOI: http://dx.doi.org/10.21609/jiki.v17i2.1273

Detecting Type and Index Mutation in Cancer DNA Sequence Based on Needleman–Wunsch Algorithm

Untari Novia Wisesty[1,*], Tati Rajab Mengko[2], Ayu Purwarianti[2,4], and Adi Pancoro[3]

[1]School of Computing, Telkom University, Bandung, Indonesia
[2]School of Electrical and Information Engineering, Bandung Institute of Technology, Bandung, Indonesia
[3]School of Life Sciences and Technology, Bandung Institute of Technology, Bandung, Indonesia
[4]U-CoE AI-VLB

E-mail: untarinw@telkomuniversity.ac.id

Abstract

Detecting DNA sequence mutations in cancer patients contributes to early identification and treatment of the disease, which ultimately enhances the effectiveness of treatment. Bioinformatics utilizes sequence alignment as a powerful tool for identifying mutations in DNA sequences. We used the Needleman-Wunsch algorithm to identify mutations in DNA sequence data from cancer patients. The cancer sequence dataset used includes breast, cervix uteri, lung, colon, liver and prostate cancer. Various types of mutations were identified, such as Single Nucleotide Variant (SNV)/substitution, insertion, and deletion, locate by the nucleotide index. The Needleman Wunch algorithm can detect type and index mutation with the average F1-scores 0.9507 for all types of mutations, 0.9919 for SNV, 0.7554 for insertion, and 0.8658 for deletion with a tolerance of 5 bp. The F1 scores obtained are not correlated with gene length. The time required ranges from 1.03 seconds for a 290 base pair gene to 3211.45 seconds for a gene with 16613 base pairs.

Keywords: *Cancer early detection, DNA sequence, Mutation detection, Needleman-Wunsch, Sequence alignment.*

1. Introduction

The rising number of cancer-related deaths and newly diagnosed patients emphasizes the critical importance of early cancer detection. Delayed examination leads to a late diagnosis, which is one of the factors that causes increased mortality. In patients with late diagnosis, cancer develops to an advanced stage, and treatment becomes less effective. Performing a biopsy on solid tumors can be particularly challenging in cases where a tumor is not formed or when the cancer is located in a hard-to-reach organ. So, DNA tests using patients' blood samples can be used for early detection of cancer. An abnormality (mutation) in the DNA can cause cancer. Different types of mutations can occur, including single nucleotide variant (SNV)/substitution, insertion, and deletion. Different types of cancer cause different mutations in the chromosomes. For example, in breast cancer, gene mutations often occur in the BRCA1, BRCA2 [1], [2], [3], [4], [5], and PALB2 genes [6].

Bioinformatics can be used as an alternative to detect these gene mutations. An effective method to identify mutations in cancer involves the alignment approach, which entails comparing the patient's DNA sample to a comprehensive database of reference DNA sequences. The effectiveness of the alignment process can be evaluated by measuring the match/mismatch value of the matrix created for each nucleotide or amino acid residue that is part of the alignment. Several studies were conducted to detect mutations using an alignment approach. Dicks built a system called AutoCSA to detect somatic variants using the ABI Genescan software [7]. In 2008, Huang and his team introduced an innovative mapping technique designed to significantly accelerate the alignment process. This technique involves converting each nucleotide into unique integer values, ranging from h1 to h4. By ensuring that the values are distinct, Huang's method revolutionizes the alignment process [8]. Teer conducted research on mutation detection using the GATK Unified Genotyper

software, without the need to match the test sequences to normal sample sequence data. Not matching the test sample with a normal sample leads to less precise results compared to using a normal sample [9].

IsoMut, a tool developed by Pipek and colleagues in 2017, is the solution for identifying mutations in multiple isogenic samples in experimental scenarios [10]. IsoMut is capable of decreasing the occurrence of false positives when compared to traditional tools. It surpasses them by a wide margin in its ability to accurately identify not only single nucleotide variations, but also detect insertions and deletions. This remarkable tool exhibits its superiority in as many as 30 isogenic samples. Schmidt's two-line sweep-based technique is capable of managing high throughput alignment, offering superior accuracy when compared to previous aligners, which struggled with multiple insertions and deletions [11]. When Bivartect converts the entire sequence to a bit string, it only needs to store a small part of the suffix reads in memory – specifically, the position at which the sequence alignment process deviates. Bivartect can identify multiple single nucleotide variations by comparing short reads with normal samples. [12].

According to the research mentioned earlier, the alignment method is capable of accurately identifying mutations in DNA sequence data. In this paper, we employed the Needleman-Wunsch algorithm, an alignment approach, to identify type and index mutation in DNA sequences of cancer patients. The Needleman-Wunsch algorithm is a powerful global pairwise alignment algorithm that efficiently aligns two sequences starting from the first nucleotide and ending at the last nucleotide of the tested sequence [13]. The data used in this study was obtained from the the Catalog of Somatic Mutation in Cancer public database [14]. This comprehensive database covers various types of cancer, including breast, cervical, lung, colon, liver, and prostate cancers. The data is then preprocessed, and the type and location of mutations are detected using the Needleman–Wunsch algorithm. Various types of mutations such as SNV/substitution, insertion, and deletion were identified, pinpointing the nucleotide index with mutations in specific gene sequences.

2. Materials and Methods

Mutations are alterations that occur in one or more nucleotides within the DNA sequence. Mutations can be caused by various factors such as errors during cell replication, inheritance, or other factors. This research revealed DNA sequence mutations, such as SNV/substitution and base-pair mutations (insertion and deletion). During SNV/substitution, one or more nucleotide is altered without affecting the overall length of the DNA. When one or more genes are added or removed, it is called a base-pair mutation. This mutation impacts the length of the DNA sequence [15]. Insertion and deletion mutations can occur simultaneously and are often referred to as Delins mutations. Duplication mutations can be classified as insertion mutations, where these mutations can occur by duplicating one or more certain nucleotides. Table 1 shows an example of a DNA sequence mutation in the original sequence "CGACCAACGGCG". Red nucleotides represent the mutated nucleotides.

Table 1. Example of DNA sequence mutation.

Mutation Type	Sequence Example
SNV/substitution	CGTCCAACGGCG
Insertion	CGACCAGAACGGCG
Deletion	CGAC—ACGGCG

The Needleman–Wunsch algorithm proposed is specifically tailored for identifying mutations in cancer DNA sequences, determining both their type and index. The research involves preprocessing data and designing a mutation detection system using the Needleman-Wunsch algorithm. The detected mutations encompass SNV, insertion, and deletion types, with the mutation location representing the indices of mutated nucleotides within the DNA sequence.

2.1. Data Preprocessing

The DNA sequence data used in this study originate from various types of cancer and were obtained from COSMIC (Catalogue of Somatic Mutation in Cancer) – a public database source [14]. We analyzed patients' gene mutation data (mutation call data) across various types of cancer, including breast, cervical, lung, colon, liver, and prostate cancers, along with reference gene data. Gene data consists of genes with the highest number of mutations in specific types of cancer. Each gene is characterized by its length and reference gene. Gene length refers to the count of nucleotides present in a single gene sequence.

The mutation call data includes gene name, gene transcript, sample name, sample ID, protein mutation (AA mutation), DNA mutation (CDS mutation), primary tissue, and other information (Table 2). Mutations encompass a range of genetic changes, such as substitution (SNV), insertion, deletion, delins (deletion/insertion), and duplicates. Extracted from the CDS mutation, the type, location, and length of the mutation are derived by parsing, abiding by the following rules:

Table 2. Mutation call data from COSMIC database.

Gene Name	Transcript	Sample ID	AA Mutation	CDS Mutation	Primary Tissue	Tissue Subtype 1
TP53	ENST00000269305.8	1361573	p.E221D	c.663G>C	Large intestine	Colon
TP53	ENST00000269305.8	1394231	p.E221G	c.662A>G	Large intestine	Colon
TP53_ENST00000359597	ENST00000359597.8	1677081	p.Q331H	c.993G>C	Large intestine	Colon
TP53_ENST00000359597	ENST00000359597.8	2433480	p.Q331R	c.992A>G	Large intestine	Colon
TP53_ENST00000413465	ENST00000413465.6	1361590	p.R248Q	c.743G>A	Large intestine	Colon
TP53_ENST00000413465	ENST00000413465.6	1399686	p.R248Q	c.743G>A	Large intestine	Colon

- Deletion-Insertion/Delins (example: c.778_782delinsTAGAT) → 778_782 (position of deletion mutation), TAGAT (inserted nucleotides).
- Deletion (example: c.348_367del) → 348_367 (position of deletion mutation)
- Insertion (example: c.357_358insTCCTG) → 357_358 (insert mutation position), TCCTG (inserted nucleotides)
- Duplicate (example: c.77dup) → 77 (duplicate mutation position)
- SNV (example c.280A>G) → 280 (SNV position), A (initial nucleotide), G (mutated nucleotide)
- Data cleaning (example: c.?): delete data whose location and type of mutation are unknown.

The sequence data from patients is grouped by the mutation calls, using a unique combination of sample ID and gene name (transcript gene). The appropriate reference gene is then mutated based on the results of this grouping of the mutation call data. After completing the preprocessing, we have extracted the patient sequences and identified the type and location of mutations for each patient. Table 3 provides detailed specifications of the preprocessed data.

Table 3. Specification of the preprocessed data.

Cancer Type	Gene Name	Number of Initial Mutations	Preprocessing Results	
			Number of Sequence	Number of Mutation
Breast cancer	BRCA1, BRCA 2	1,599	1,495	1,597
Cervix uteri	FBXW7, KMT2C, KMT2D, PIK3CA	1,004	889	1,002
Lung cancer	EGFR	30,057	18,201	19,901
Colon cancer	APC, TP53	41,649	38,508	41,502
Liver cancer	TP53	27,616	25,878	27,615
Prostate cancer	ERBB4, LRP1B, PTPRT	719	583	718

2.2. Needleman Wunsch Algorithm

This research introduces a pairwise alignment approach, where one sequence is compared to another using alignment techniques. The Needleman–Wunsch algorithm, which is a global alignment algorithm, is applied in pairwise alignment. Global alignment technique aligns the entire sequence from beginning to end [16]. The Needleman-Wunsch algorithm leverages dynamic programming, an optimization technique that divides the problem into subproblems and stores the results in a matrix [17], [18]. Next, the backtrace method is employed on the constructed matrix to combine the results of solving these problems. To construct the matrix, it is essential to have a score function or recurrence relation that can be used to determine the value for each element within the matrix. Dynamic programming can determine the optimal solution to a problem by using the appropriate scoring function.

The score function utilized in the Needleman-Wunsch algorithm to populate each value in the dynamic programming matrix is demonstrated by Equation 1 [19]. The size of the constructed matrix is determined by adding one to the number of columns in the first sequence (m) and one to the number of rows in the second sequence (n). The matrix is filled with values in an iterative process that begins at column 0, row 0 and continues until column m + 1, row n + 1. The substitution score, denoted as s(i,j), represents the score assigned when comparing nucleotides at a specific index in two sequences. The score is assigned when the nucleotides match or when a substitution takes place. The study displays the substitution score used in Figure 1, assigning a value of "+1" for a match and "−1" for a mismatch. The gap penalty is imposed whenever an insertion or deletion occurs. The corresponding gap values we have observed are -1, -2, and -3.

$$C(i,j) = max\begin{cases} C(i-1,j-1) + s(i,j) \\ C(i-1,j) - gap \\ C(i,j-1) - gap \end{cases} \quad (1)$$

with $\boldsymbol{C(i,j)}$ is dynamic programming table, $\boldsymbol{s(i,j)}$ is substitution score, $\boldsymbol{gap}$ is penalty of insertion and deletion, and $\boldsymbol{i,j}$ is index of dynamic programming table.

	A	C	T	G	N
A	1	-1	-1	-1	-1
C	-1	1	-1	-1	-1
T	-1	-1	1	-1	-1
G	-1	-1	-1	1	-1
N	-1	-1	-1	-1	1

Figure 1. Substitution score.

For a more streamlined traceback process, a matrix of the same size as the previous score matrix is essential. The traceback matrix B(i,j) contains "up," "left," and "diag" (diagonal) values that specify the direction for the traceback operation. Equation 2 is used to determine the direction during the process. The path for backtrace is determined from the matrix index (m+1, n+1) to the matrix index 0, 0 using the following directions contained in matrix B:

- "diag" indicates match or SNV mutation/substitution,
- "up" indicates insertion mutation,
- "left" denotes the deletion mutation.

$$B(i,j) = \begin{cases} diag,\ if\ C(i,j) = C(i-1,j-1) + s(i,j) \\ up,\ if\ C(i,j) = C(i-1,j) - gap \\ left,\ if\ C(i,j) = C(i,j-1) - gap \end{cases} \quad (2)$$

```
['A', 'T', 'A', 'G', 'T', 'T', 'T']
['A', 'T', '-', '-', 'T', 'G', 'T']
['Del', 'Del', 'SNV']
[3, 4, 6]
```

Figure 2. An example of alignment and mutation detection using the Needleman–Wunsch algorithm.

Figure 2 displays how the Needleman-Wunsch algorithm is used to align and detect mutation types and locations, where deletion mutations are located at indexes 3 and 4, while SNV/substitution mutations occur at index 6. The accuracy of the system is determined by the presence or absence of each type of mutation at the corresponding nucleotide point in the sequence, using metric precision, recall, and F1-score [20]. Equation 3 - 5 shows precision, recall, and F1-score metrics, where True Positive (TP) is a mutated nucleotide that is correctly predicted as a mutated nucleotide, False Positive (FP) is a normal nucleotide that is incorrectly detected as a mutated nucleotide, and False Negative (FN) is a mutated nucleotide that is incorrectly detected as a normal nucleotide.

$$\boldsymbol{precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (3)$$

$$\boldsymbol{recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (4)$$

$$\boldsymbol{F1-score} = \frac{\boldsymbol{2*precision*recall}}{\boldsymbol{precision+recall}} \quad (5)$$

3. Results and Discussion

In order to assess the effectiveness of detecting different types and locations of mutations in various cancers, we conducted a thorough evaluation schemes, including:

- Comparison of exact match and 5 bp tolerance scheme for lung, colon, liver, breast, prostate, and cervical cancer mutation detection.
- Observation of gap parameters in the Needleman Wunsch algorithm, where the observed gap values are -1, -2, and -3.
- Comparison of the best F1-score values achieved for lung, breast, liver, colon, prostate, and cervical cancer datasets.

Our evaluation included measuring precision, recall, F1 score, and running time. Furthermore, we incorporated penalty gap parameters into the alignment process to improve mutation detection. This enabled us to not only compare the performance across all cancer data, but also analyze the processing time. The results provide valuable insights into the accuracy and efficiency of our mutation detection methods.

3.1. Mutation Detection Performance for Each Cancer Type

We conduct tests on each cancer dataset to accurately identify the specific mutations and their respective index. Detection of the mutation index involves two schemes, as follows:

- Exact match scheme, a predicted mutation point is considered correct only when it precisely matches the actual point mutation data.
- The 5 bp tolerance scheme, allows for a range of ±5 nucleotides from the actual point mutation data, making it acceptable for the predicted mutation point to fall within this range.
- Comparison of the best F1-score values achieved for lung, breast, liver, colon, prostate, and cervical cancer datasets.

The detected mutations included SNV/substitution, insertion, and deletion. The precision, recall, and F1 score were used to quantify the accuracy of mutation detection performance. The performance measures (precision, recall, and F1 score) for detecting mutations in DNA sequence data are showed in tables 4 to 9. These tables specifically present the findings for different types and locations of mutations in lung, colon, liver, breast, prostate, and cervical cancers. SNV demonstrated exceptional performance, exceeding thresholds of 0.9 for

precision, recall, and F1 score in both the exact match and 5 bp tolerance detection schemes. Insertion and deletion mutations have low performance, especially for insertion, when it comes to exact match detection. The sequence data is affected by a nucleotide shift, which can be addressed by implementing a 5 bp tolerance to enhance performance. The term "All" encompasses all forms of mutations, including SNV, insertions, and deletions.

Table 4. Mutation detection performance for lung cancer.

Detection Scheme	Mutation Type	Precision	Recall	F1-score
Exact Match	All	0.7795	0.7834	0.7814
	SNV	0.9879	0.9855	0.9867
	Insertion	0.1123	0.1281	0.1197
	Deletion	0.7869	0.7870	0.7869
5bp Tolerance	All	0.9698	0.9747	0.9722
	SNV	0.9879	0.9855	0.9867
	Insertion	0.8435	0.9629	0.8992
	Deletion	0.9738	0.9739	0.9738

Table 5. Mutation detection performance for colon cancer.

Detection Scheme	Mutation Type	Precision	Recall	F1-score
Exact Match	All	0.8161	0.8170	0.8166
	SNV	0.9967	0.9990	0.9979
	Insertion	0.2021	0.1967	0.1994
	Deletion	0.5457	0.5482	0.5470
5bp Tolerance	All	0.9533	0.9543	0.9538
	SNV	0.9973	0.9996	0.9984
	Insertion	0.7731	0.7528	0.7628
	Deletion	0.8945	0.8987	0.8966

Table 6. Mutation detection performance for liver cancer.

Detection Scheme	Mutation Type	Precision	Recall	F1-score
Exact Match	All	0.8450	0.8535	0.8492
	SNV	0.9966	0.9800	0.9883
	Insertion	0.0697	0.0925	0.0795
	Deletion	0.6702	0.6719	0.6711
5bp Tolerance	All	0.9386	0.9481	0.9433
	SNV	0.9984	0.9817	0.9900
	Insertion	0.5282	0.7014	0.6026
	Deletion	0.9103	0.9126	0.9115

Table 7. Mutation detection performance for breast cancer.

Detection Scheme	Mutation Type	Precision	Recall	F1-score
Exact Match	All	0.7423	0.7419	0.7421
	SNV	0.9688	1.0000	0.9842
	Insertion	0.3977	0.3309	0.3612
	Deletion	0.6832	0.6907	0.6869
5bp Tolerance	All	0.8523	0.8519	0.8521
	SNV	0.9688	1.0000	0.9842
	Insertion	0.8333	0.6934	0.7570
	Deletion	0.8016	0.8104	0.8060

Table 8. Mutation detection performance for prostate cancer.

Detection Scheme	Mutation Type	Precision	Recall	F1-score
Exact Match	All	0.9764	0.9764	0.9764
	SNV	0.9986	0.9874	0.9929
	Insertion	-	-	-
	Deletion	0	0	0
5bp Tolerance	All	0.9847	0.9847	0.9847
	SNV	0.9986	0.9874	0.9929
	Insertion	-	-	-
	Deletion	0.6000	1.0000	0.7500

Table 9. Mutation detection performance for cervical cancer.

Detection Scheme	Mutation Type	Precision	Recall	F1-score
Exact Match	All	0.9980	0.9980	0.9980
	SNV	1.0000	0.9980	0.9990
	Insertion	-	-	-
	Deletion	0.7500	1.0000	0.8571
5bp Tolerance	All	0.9980	0.9980	0.9980
	SNV	1.0000	0.9980	0.9990
	Insertion	-	-	-
	Deletion	0.7500	1.0000	0.8571

3.2. Gap Parameter Observation

In this section, we will examine the significance of the parameter "gap" in assessing the effectiveness of mutation detection. In the Needleman-Wunsch algorithm, the parameter gap serves as a penalty for any insertions or deletions that may occur during the alignment process. The values of the observed gap for liver and cervical cancer are -1, -2, and -3. The precision, recall, and F1 score results for different gap values in the mutation detection in liver cancer (Tables 10 and 11) show consistent values for all types of mutations, both in exact match and 5 bp tolerance. In terms of cervical cancer (Tables 12 and 13), we observed a modest improvement in the accuracy of detecting single nucleotide variations, deletions, and their combination. In this scenario, the values of -2 and -3 for the gap are superior to -1. Minimizing the occurrence of insertions and deletions is crucial, as the penalty for these operations increases with the value of the gap.

Table 10. Observation of gap value on mutation detection performance in exact match detection on liver cancer.

	Gap	-1	-2	-3
Precision	All	0.8450	0.8450	0.8450
	SNV	0.9966	0.9966	0.9966
	Insertion	0.0697	0.0697	0.0697
	Deletion	0.6702	0.6702	0.6702
Recall	All	0.8535	0.8535	0.8535
	SNV	0.9800	0.9800	0.9800
	Insertion	0.0925	0.0925	0.0925
	Deletion	0.6719	0.6719	0.6719
F1-Score	All	0.8492	0.8492	0.8492
	SNV	0.9883	0.9883	0.9883
	Insertion	0.0795	0.0795	0.0795
	Deletion	0.6711	0.6711	0.6711

Table 11. Observation of gap value on mutation detection performance with 5bp tolerance on liver cancer.

	Gap	-1	-2	-3
Precision	All	0.9386	0.9386	0.9386
	SNV	0.9984	0.9984	0.9984
	Insertion	0.5282	0.5282	0.5282
	Deletion	0.9103	0.9103	0.9103
Recall	All	0.9481	0.9481	0.9481
	SNV	0.9817	0.9817	0.9817
	Insertion	0.7014	0.7014	0.7014
	Deletion	0.9126	0.9126	0.9126
F1-Score	All	0.9433	0.9433	0.9433
	SNV	0.9900	0.9900	0.9900
	Insertion	0.6026	0.6026	0.6026
	Deletion	0.9115	0.9115	0.9115

Table 12. Observation of gap value on mutation detection performance in exact match detection on cervical cancer.

	Gap	-1	-2	-3
Precision	All	0.9980	1.0000	1.0000
	SNV	1.0000	1.0000	1.0000
	Insertion	-	-	-
	Deletion	0.7500	1.0000	1.0000
Recall	All	0.9980	1.0000	1.0000
	SNV	0.9980	1.0000	1.0000
	Insertion	-	-	-
	Deletion	1.0000	1.0000	1.0000
F1-Score	All	0.9980	1.0000	1.0000
	SNV	0.9990	1.0000	1.0000
	Insertion	-	-	-
	Deletion	0.8571	1.0000	1.0000

Table 13. Observation of gap value on mutation detection performance with 5bp tolerance on cervical cancer.

	Gap	-1	-2	-3
Precision	All	0.9980	1.0000	1.0000
	SNV	1.0000	1.0000	1.0000
	Insertion	-	-	-
	Deletion	0.7500	1.0000	1.0000
Recall	All	0.9980	1.0000	1.0000
	SNV	0.9980	1.0000	1.0000
	Insertion	-	-	-
	Deletion	1.0000	1.0000	1.0000
F1-Score	All	0.9980	1.0000	1.0000
	SNV	0.9990	1.0000	1.0000
	Insertion	-	-	-
	Deletion	0.8571	1.0000	1.0000

3.3. Analysis of F1-score and Running Time of All Cancer Data

This section presents an analysis of the mutation detection system's overall performance using DNA sequence data collected from cancer patients. We have conducted a thorough analysis of the F1 scores in cancer datasets using both exact match and 5 bp tolerance detection schemes. Additionally, we have examined the correlation between F1 score and gene length, as well as the relationship between running time and gene length. Tables 14 and 15 display the F1 scores for the lung, breast, liver, colon, prostate, and cervical cancer datasets regarding exact match and 5 bp tolerance.

Table 14. Comparison of F1-Score mutation detection using exact match scheme for all types of cancer.

Cancer Type	F1-Score			
	All	SNV	Insertion	Deletion
Breast	0.7421	0.9842	0.3612	0.6869
Lung	0.7814	0.9867	0.1197	0.7869
Liver	0.8492	0.9883	0.0795	0.6711
Colon	0.8166	0.9979	0.1994	0.5470
Prostate	0.9764	0.9929	-	0.0000
Cervix	0.9980	0.9990	-	0.8571
Average score	0.8606	0.9915	0.1899	0.5915

Table 15. Comparison of F1-Score mutation detection using 5 bp tolerance scheme for all cancer types.

Cancer Type	F1-Score			
	All	SNV	Insertion	Deletion
Breast	0.8521	0.9842	0.7570	0.8060
Lung	0.9722	0.9867	0.8992	0.9738
Liver	0.9433	0.9900	0.6026	0.9115
Colon	0.9538	0.9984	0.7628	0.8966
Prostate	0.9847	0.9929	-	0.7500
Cervix	0.9980	0.9990	-	0.8571
Average score	0.9507	0.9919	0.7554	0.8658

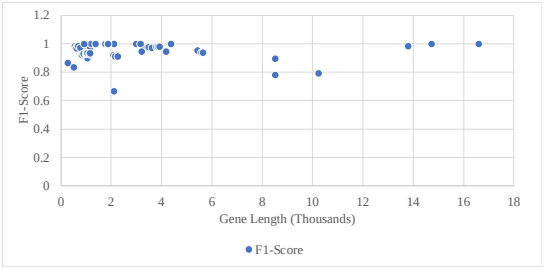

Figure 3. Relationship of F1-Score and gene length.

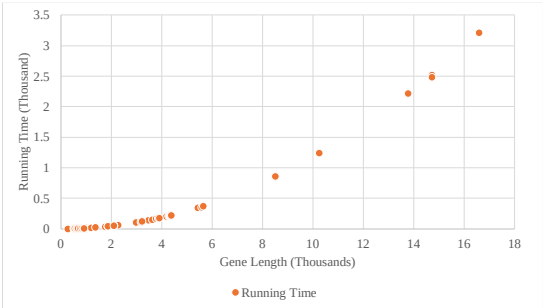

Figure 4. Relationship of running time and gene length.

Exact match mutation detection has excellent results in detecting SNV mutations, achieving an impressive average F1-score of 0.9915 (Table 14). Regarding insertion and deletion mutations, the performance is quite low. The average F1 score for insertion mutations is 0.1899, and for deletion mutations, it is 0.5915. Insertion and deletion mutations shift the nucleotide index in DNA sequences, making it difficult to predict accurate mutation points using exact match detection. In order to effectively address this issue, the mutation

detection process involves allowing for a 5 base pair tolerance from the exact point of mutation. The F1 score for SNV mutations can be significantly increased, with an impressive average of 0.9919, when allowing a 5 bp tolerance. Similarly, for insertion mutations, the F1 score can reach an average of 0.7554, and for deletion mutations, it can reach an average of 0.8658 (Table 15).

We conducted an analysis to assess the impact of gene length on F1 score and mutation detection time when using the Needleman-Wunsch alignment algorithm. The F1-score and detection time average for each gene were computed and arranged in ascending order according to the gene's length. Figure 3 demonstrates that there is no impact or correlation between gene length and the resulting F1-score value, where if the correlation calculation is carried out using the Pearson Correlation method, the Pearson score obtained is -0.0106. A Pearson score close to zero indicates that there is no correlation between the two variables being measured. Upon examining Figure 3, it becomes evident that no discernible pattern exists between the length of the gene and the resulting F1-score.

Unlike the F1 score, which is not correlated with gene length, the running time shows a strong correlation with gene length. Based on the correlation calculation between gene length and running time, the Pearson score obtained was 0.9497. The alignment process takes longer as the length of the nucleotides that form a gene increases, as shown in Figure 4. According to the graph, the gene length that requires the least amount of time is 1.03 seconds for 290 bp, whereas the gene length that demands the longest time is 16613 bp with a staggering 3211.45 seconds. The amount of time required grows exponentially as the length of the sequence increases. The Needleman-Wunsch algorithm contains various processes that contribute to its overall time complexity.

4. Conclusion

This paper presents the development of a mutation detection system for DNA sequence data, utilizing the renowned Needleman-Wunsch algorithm, an effective pairwise alignment algorithm. The Needleman-Wunsch algorithm is a powerful global alignment technique that identifies mutations in every nucleotide of a given sequence. We utilized DNA sequence data from various cancer types, such as breast, lung, cervical, colon, liver, and prostate cancers. These valuable data were obtained from the publicly available COSMIC Cancer Browser database. This research incorporates data collection, preprocessing, and the development and application of the Needleman-Wunsch algorithm to identify mutation types and their index.

According to the experiments carried out, the Needleman-Wunsch algorithm possesses the capability to identify mutations and pinpoint their index, achieving average F1 scores of 0.8606 for all types of mutations, 0.9915 for SNVs, 0.1899 for insertions, and 0.5915 for deletions. The average F1 scores will increase as follows: 0.9507 for all types of mutations, 0.9919 for SNV, 0.7554 for insertion, and 0.8658 for deletion with tolerance of 5 bp. Nevertheless, this algorithm has two vulnerabilities. Initially, the alignment process necessitates a substantial amount of time to complete. The minimum amount of time needed is only 1.03 seconds for a gene length of 290 base pairs, whereas the maximum time required is a 3211.45 seconds for a gene length of 16613 base pairs. As the length of the sequence increases, the time will increase exponentially. Another drawback is that it necessitates a reference sequence dataset containing a specific transcript for aligning and detecting mutations. Finally, the algorithm we developed cannot detect mutations in certain cases where several mutations occur at the same nucleotide index in one sequence.

Acknowledgement

We would like to express our gratitude to Telkom University and the Bandung Institute of Technology for supporting this research.

References

[1] M. Dean *et al.*, "Addressing health disparities in Hispanic breast cancer: Accurate and inexpensive sequencing of BRCA1 and BRCA2," *GigaScience*, vol. 4, no. 1, 2015, doi: 10.1186/s13742-015-0088-z.

[2] H. T. Lynch *et al.*, "Hereditary ovarian carcinoma : Heterogeneity , molecular genetics , pathology , and management," *Mol. Oncol.*, vol. 3, pp. 97–137, 2009, doi: 10.1016/j.molonc.2009.02.004.

[3] C. Alvarez *et al.*, "BRCA1 and BRCA2 founder mutations account for 78 % of germline carriers among hereditary breast cancer families in Chile," *Oncotarget*, vol. 8, no. 43, pp. 74233–74243, 2017.

[4] G. I. Lambrou *et al.*, "Computational Analysis of BRCA1 Mutations in Pediatric Patients with Malignancies and Their Mothers," *Proc. - IEEE Symp. Comput.-Based Med. Syst.*, vol. 2017-June, pp. 138–143, 2017, doi: 10.1109/CBMS.2017.111.

[5] G. J. Mann *et al.*, "Analysis of cancer risk and BRCA1 and BRCA2 mutation prevalence in the kConFab familial breast cancer resource," *Breast Cancer Res.*, vol. 8, no. 1, pp. 1–15, 2006, doi: 10.1186/bcr1377.

[6] H. R. V. Kumar, M. Elancheran, S. R. Dhamotharan, and J. C. Indrani, "Novel PALB2 deleterious mutations in breast cancer patients from South Indian population," *Gene Rep.*, vol. 17, no. July, p. 100492, 2019, doi: 10.1016/j.genrep.2019.100492.

[7] E. Dicks *et al.*, "AutoCSA, an algorithm for high throughput DNA sequence variant detection in cancer genomes," *Bioinformatics*, vol. 23, no. 13, pp. 1689–

1691, 2007, doi: 10.1093/bioinformatics/btm152.

[8] G. Huang, B. Liao, W. Zhang, and F. Gong, "A novel method for sequence alignment and mutation analysis," *Match*, vol. 59, no. 3, pp. 635–645, 2008.

[9] J. K. Teer *et al.*, "Evaluating somatic tumor mutation detection without matched normal samples," *Hum. Genomics*, vol. 11, no. 1, pp. 1–13, 2017, doi: 10.1186/s40246-017-0118-2.

[10] O. Pipek *et al.*, "Fast and accurate mutation detection in whole genome sequences of multiple isogenic samples with IsoMut," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–11, 2017, doi: 10.1186/s12859-017-1492-4.

[11] M. Schmidt, K. Heese, and A. Kutzner, "Accurate high throughput alignment via line sweep-based seed processing," *Nat. Commun.*, vol. 10, no. 1, 2019, doi: 10.1038/s41467-019-09977-2.

[12] K. Shimmura, Y. Kato, and Y. Kawahara, "Bivartect: accurate and memory-saving breakpoint detection by direct read comparison," *Bioinformatics*, no. January, pp. 1–6, 2020, doi: 10.1093/bioinformatics/btaa059.

[13] C.-T. Lee and S.-L. Peng, "A Pairwise Alignment Algorithm for Long Sequences of High Similarity," in *Information and Communication Technology*, vol. 625, D. K. Mishra, A. T. Azar, and A. Joshi, Eds., in Advances in Intelligent Systems and Computing, vol. 625. , Singapore: Springer Singapore, 2018, pp. 279–287. doi: 10.1007/978-981-10-5508-9_27.

[14] COSMIC, "Cancer Browser." Accessed: Sep. 21, 2020. [Online]. Available: https://cancer.sanger.ac.uk/cosmic/browse/tissue

[15] U. N. Wisesty, T. R. Mengko, and A. Purwarianti, "Gene mutation detection for breast cancer disease: A review," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 830, p. 032051, May 2020, doi: 10.1088/1757-899X/830/3/032051.

[16] F. Muhamad, R. Ahmad, S. Asi, and M. Murad, "Performance Analysis of Needleman-Wunsch Algorithm (Global) And Smith-Waterman Algorithm (Local) In Reducing Search Space And Time For Dna Sequence Alignment," *J. Phys. Conf. Ser.*, vol. 1019, p. 012085, Jun. 2018, doi: 10.1088/1742-6596/1019/1/012085.

[17] C. Kyal, R. Kumar, and A. Zamal, "Performance-Based Analogising of Needleman Wunsch Algorithm to Align DNA Sequences Using GPU and FPGA," in *2020 IEEE 17th India Council International Conference (INDICON)*, New Delhi, India: IEEE, Dec. 2020, pp. 1–5. doi: 10.1109/INDICON49873.2020.9342078.

[18] E. Aspland, P. R. Harper, D. Gartner, P. Webb, and P. Barrett-Lee, "Modified Needleman–Wunsch algorithm for clinical pathway clustering," *J. Biomed. Inform.*, vol. 115, p. 103668, Mar. 2021, doi: 10.1016/j.jbi.2020.103668.

[19] H. Christensen, *Introduction to Bioinformatics in Microbiology*, vol. 39, no. 3. in Learning Materials in Biosciences, vol. 39. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-99280-8.

[20] I. Anzar, A. Sverchkova, R. Stratford, and T. Clancy, "NeoMutate: An ensemble machine learning framework for the prediction of somatic mutations in cancer," *BMC Med. Genomics*, vol. 12, no. 1, pp. 1–14, 2019, doi: 10.1186/s12920-019-0508-5.