Sequences analysis and phylogenetic of leukemia DNA

D Lusiyanit, N Nacong1 and M I Irawan2

1 Mathematics Study Program of Science Faculty, Tadulako University, Indonesia
2 Dapertemen Mathematics, Faculty of Mathematics, Computing and Data Science, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

*Email: Desy.lusiyanti@untad.ac.id

Abstract. DNA is one part of a biological sequence other than RNA and protein. DNA is a nucleic acid that stores all the unique biological information of every living thing and several viruses. Someone who has an illness can be seen in the structure of the DNA which can then be compared with other DNA structures. This process is part of sequence analysis which is the core of bioinformatics. This study aims to analyze the results of sequences alignment of cancer. The sequence was obtained from DNA leukemia taken from GenBank (www.ncbi.nlm.nih.gov). In the sequence analysis, the most basic thing is the sequence alignment. Alignment results can provide information about the level of similarity in the structure of the two sequences. This study will apply the Super Pairwise Alignment method that is implemented in JAVA-based programming languages. Alignment results are carried out by giving different values of gap penalty and gap extend. The results of the study obtained different alignment values, this proves that there is a significant effect of the selection of core parameters used in alignment.

1. Introduction
The main requirement in bioinformatics analysis is the presence of biological data in the database. One is a database of DNA and protein that is Genbank. The availability of this basic database is open, this is a great opportunity to explore valuable information such as sequence alignment and sequence mutation.

Mutations are changes in the structure of the genetic sequences in living organisms. Mutations are caused by impaired encoding in the deoxyribonucleic acid chain or so-called DNA. DNA (Deoxyribonucleic acid) is biomolecules that serves to store genetic information in living organisms [1]. Mutations in DNA can occur due to various reasons such as environmental factors, unhealthy lifestyles or viral or bacterial infections. This can lead to the occurrence of cancer one of which is cancer of the blood, or commonly called with Leukemia.

Leukemia is a type of blood cancer that originates in the spinal cord, where blood cells are made. Leukemia may be acute (deteriorate rapidly) or chronically (worsening slowly). In this study, the sequences of acute and chronic leukemia DNA sequences were performed using software MEGA. Alignment results obtained maximum similarity level of two sequences. The DNA leukemia data used were obtained from GenBank NCBI which can be accessed for free.

2. Materials and methods
2.1. Mutation
Mutations are changes in the genetic structure of living organisms. Mutations can be classified into 4 types, namely Type 1, Type 2, Type 3 and type 4.
Type I is a type of mutation that occurs due to a change in nucleotides, for example "g" changes to "a". Type II is a mutation that occurs because a part of the nucleotide changes the position sequence, for example, the "guace" part changes the sequence to "accgu". These type I and type II mutations are substitution mutations, because the position of all nucleotides does not change.

Type III is a type of mutation that is caused by the insertion of a new segment into the sequence, for example insertion of a "cc" in the middle of the "gguugg" segment will change the segment to "gguccugg" whereas Type IV is a type of mutation that occurs due to the deletion of a nucleotide segment, for example the deletion of the "ac" nucleotide from the "agacuua" segment so that the segment turns into 'aguua'. Type III and IV mutations are called displacement mutations, because there is a shift in the position of all nucleotides. [2]

2.2. DNA structure
DNA (deoxyribonucleic acid) or deoxyribose nucleic acid (ADN) is a storage place for an organism's information. DNA is a polymer consisting of three main components, a phosphate group, a deoxyribose sugar and a nitrogenous base. The three components in the DNA monomer unit are called nucleotides, so that DNA is included as a polynucleotide. DNA can have millions of nucleotides arranged in chains.

DNA sequences are a series of letters representing the primary structure of DNA molecules [4]. The letters used are a, c, g, and t which represent the four nucleotides, namely adenine, cytosine, guanine, thymine. The DNA sequence is described as follows:

\[ A = \{x_i\}_{i=1}^{n} \quad \text{and} \quad B = \{y_j\}_{j=1}^{n} \]

A and B are sequences, \(x_i\) and \(y_j\) are the elements in each sequence, \(m\) and \(n\) are the lengths of the X and Y sequences [5].

2.3. Sequence alignment
Sequence alignment is the process of arranging or setting two or more sequences so that the similarities between the sequences appear real. Information on the level of similarity or dissimilarity between these sequences is used to study the evolutionary process of a biological sequence. The purpose of sequence analysis is looking for a relationship between two or more sequences.

Basically, based on the alignment, sequence alignment is classified as global alignment and local alignment. Global alignment is a way of aligning the sequence in a whole length, with a sequence as a reference. Local alignment identifies isolated areas that have high similarity [6].

2.4. Global alignment of two DNA sequences
There are two terms: mutation and indel. Mutation is seen as a mismatch on sequences alignment and indels which are insertions and deletions in alignment. Mutations and indels are represented by gaps.

Two sequences are given \(X = \{x_i\}_{i=1}^{n}\) dan \(Y = \{y_j\}_{j=1}^{m}\) with each length being \(n\) and \(m\), let \(c\) be the total length of the alignment then obtained: \(\max(n, c) \leq c \leq n + m\).

Example:
Given X = {cttaacgtta} and Y = {cgateca}, then \(n = 10\) and \(m = 7\). then a possible alignment of X and Y is [5]

\[ A(X, Y) = \begin{array}{cccccc}
C & T & C & A & A & T
\end{array} \quad \begin{array}{cccc}
G & A & G & T & T & A
\end{array} \]

*Mutation *Deletion *Insertion

2}
3. Results and discussion
The first thing to do is to align two sequences. Sequence pairs to be aligned are as follows:

DNA sequence data 1: Homo sapiens t (9; 22) (q34, q11) reciprocal chromosomal translocation breakpoint, patient 5349 with chronic myeloid leukemia (FN869142.1) with a sequence length of 775 bp.

DNA sequence data 2: Homo sapiens t (9; 22) (q34, q11) chromosomal translocation breakpoint, patient 6359 with acute lymphoblastic leukemia (FN820268.2) with a sequence length of 604 bp.

Alignment results using software MEGA as follows:

![Figure 1. Pairs alignment](image)

Homology obtained at 99%. Where as mutation changes occur in nucleotides as follows:

| From/To | A      | T      | C      | G      |
|---------|--------|--------|--------|--------|
| A       | 3,3247 | 2,9124 | 18,3035|
| T       | 3,2340 | 18,1020| 3,4738 |
| C       | 3,2340 | 20,6646| 3,4738 |
| G       | 17,0401| 3,3247 | 2,9124 |

From the table above, obtained nucleotide frequencies are A = 24.98%, T = 25.68%, C = 22.50%, and G = 26.84%. To estimate the ML value, the tree topology is automatically calculated. Possible maximum log for this calculation is -3349,672. This analysis involves 8 nucleotide sequences. Codon positions included are 1st + 2nd + 3rd + Noncoding. There are a total of 1149 positions in the last data set.

Furthermore, the same is done for acute and chronic DNA leukemia sequence pair, and the results obtained as in the following table:

| Acute Leukemia | Chronic Leukemia | Homology |
|----------------|------------------|----------|
| FN820209.1     | FN869167.1       | 99%      |
| FN820259.1     | FN820245.1       | 99%      |
| FN869137.1     | FN820258.1       | 99%      |
Based on experiments conducted on sequences of acute and chronic DNA leukemia showed homology sequences 99%. This suggests that no significant mutations between acute and chronic leukemia. To see the relationship between acute and chronic leukemia DNA sequence with phylogenetic tree approach as follows:

**Figure 2** Phylogenetic tree acute and chronic leukemia DNA sequences

In Figure 2 there are two large groups of leukemia types, the first group consists of 4 types of Leukemia are Acute Myelostic Leukemia, Acute Lymphocytic Leukemia, Chronic Myelocytic Leukemia and Chronic Lymphostic Leukemia while the second group consists of 3 types of Leukemia are Acute Myelostic Leukemia, Chronic Myelocytic Leukemia and Chronic Lymphostic Leukemia. The length of each pair of branches shows the distance between each of the fourth sequence sequences of the type of leukemia, the distance between each pair of sequences indicating the bootstrap value of each sequence pair. The bootstrap value is a value to test how well the model data set is used. The grouping of the four types of leukemia obtained is stable with a bootstrap value of 100, the optimal tree with the number of branch lengths = 2.15499245. The closest genetic distance is 0.00000000 and the furthest is 1.6190937063.

4. Conclusion
Based on these results, it can be concluded that the DNA results for sequence alignment of acute and chronic leukemia indicate that DNA mutations originate from the same host and by using MEGA software the average mutation rate is 10% with a frequency of 24.98 per nucleotide, %A = 25.68%, %C = 22.50%, and %G = 26.84%. The DNA sequence similarity for acute and chronic leukemia is 90%.

Acknowledgement
The research was funded by a research grant funding contract of Pekerti 872 / UN6.3 / LT / 2017. The authors would like to thank the Rector of Tadulako University and the Ministry of Research, Technology and Higher Education of the Republic of Indonesia. Also thanks to the anonymous reviewer for valuable checking this paper.

References
[1] Utomo W A 2013 *Penjajaran Global Sekuen DNA Menggunakan Algoritme Needleman-Wunsch*
Departemen Ilmu Komputer Fakultas Matematika Dan Ilmu Pengetahuan Alam Institut Pertanian Bogor
[2] Shen S N 2007 *Theory and Mathematical Methods for Bioinformatics, Biological and Medical Physics, Biomedical Engineering*, Springer
[3] Christianini N, Hahn M W 2006 *Introduction to Computational Genomics studies A Case Approach*, Cambridge University Press, New York
[4] Isaev A 2006 *Introduction to Mathematical Methods in Bioinformatics, Springer*

[5] Escarino C R 2009 *A two-base encoded DNA sequence alignment problem in computational biology* MathInIndustry Project, National Institute Of Genomic Medicine, Mexico

[6] Nacong N, Lusiyanti D, Irawan M I 2018 *AIP Conference Proceedings* 020010-1 – 020010-4