Contemporary Mathematics and Applications
Vol. 7, No. 1, 2025, pp. 26-35.

Determining Optimal Hierarchical Clustering by Combining Needleman Wunsch and Jukes Cantor Algorithms in Tuberculosis (TB) Disease Clustering

Hildatul Anizah[1], Tony Yulianto[2], Kuzairi[3], Ira Yudistira[4], Rica Amalia[5]

[1,2,3,4,5]Universitas Islam Madura

[5]Corresponding author: ricaamalia5@gmail.com

Abstract. Tuberculosis (TBC) is an infectious disease affecting the respiratory system, caused by the bacterium Mycobacterium tuberculosis. Tuberculosis (TBC) remains a global concern, and to date, no country is completely free from TB. This disease continues to be one of the leading causes of mortality. Therefore, it is essential to categorize the spread of TBC. The percentage of identity in genetic codes will reveal the proportion of mutations. The percentage of identity in genetic codes will demonstrate that, although the symptoms caused by a disease may be quite similar, the protein sequences are not necessarily the same. In this study, the researchers employed the Hierarchical Clustering method, integrating the Needleman-Wunsch and Jukes-Cantor algorithms, resulting in two groups. The first group consists of 9 interconnected rows, while the second group consists of 7 interconnected rows.

Keywords: *jukes cantor; needleman wunsch; optimal hierarchical clustering; tuberculosis;*

1 Introduction

The development and progress of today's world cannot be separated from mathematics. Almost all human activities are related to mathematics. Mathematics is used as an important tool in various fields, including natural sciences, medical technology, and social sciences such as economics and psychology. The utilization of mathematics in everyday life is clearly seen in its application to all aspects of human life [1].

One of the events that occur in human life that can be implemented with mathematical models is an epidemic. An epidemic is an event where a disease spreads or occurs in an area. This article specifically discusses the mathematical model of the tuberculosis epidemic.

Tuberculosis (TB) is an infectious disease that mainly affects the respiratory system caused by the bacterium Mycobacterium tuberculosis. Although the disease usually affects the lungs, it can also be implicated in other organs of the body. The bacteria that causes tuberculosis was first identified by Robert Koch on March 24, 1882. Symptoms commonly experienced by individuals with TB include prolonged cough, chest pain, shortness of breath, loss of appetite, weight loss, fever, chills, and fatigue [2].

Received: October 23, 2024
Accepted: February 27, 2025

Tuberculosis (TB) remains the focus of global attention, and to date, no country is completely free of TB [2]. Tuberculosis is still one of the leading causes of death, with an estimated 1.3 million victims overall [3].

Facing these problems, effective handling efforts are needed. One of the steps that can be taken is to categorize tuberculosis transmission areas and various types of virus strains [3]. This grouping utilizes DNA data from the pathogen that causes tuberculosis.

DNA or deoxyribonucleic acid is a biomolecule in the form of nucleic acid found in the cell nucleus (nucleus), which functions to store the genetic information of an organism. DNA structure is in the form of a double strand connected by hydrogen bonds between the bases in both strands. These bases include Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). DNA can also be used as a method to prove kinship is through sequence alignment, which is the process of arranging or adjusting a sequence with one or more other sequences, so that the similarity of these sequences can be clearly identified [4].

As for the research, the needleman wunschn model is used to classify DNA data, the jukes cantor model to calculate the distance difference in TB DNA data and the hierarchical clustering model to refine the phylogenetic image. Researchers implemented tuberculosis disease clustering by using the hierarchical clustering model by combining Needleman Wunsch and Jukes Cantor Algorithm.

2 Literature Review

2.1 Hierarchical Clustering

The hierarchical clustering method aims to group objects in such a way that every object that has high similarity with other objects is in the same cluster. The main characteristic of cluster analysis is high homogeneity (similarity) between members in one cluster and high heterogeneity (difference) between different clusters [5]. In agglomerative clustering, the process starts by placing objects in different clusters and then gradually grouping objects into larger clusters, while in devisive clustering it starts by placing all objects as one cluster. Then the objects are gradually separated into different clusters, two clusters, three clusters, and so on [6].

There are six agglomerative hierarchical methods in cluster formation, namely:

a. Single Linkage
b. Complete Linkage
c. Average Linkage
d. Ward's Method
e. Centroid Method
f. Median Linkage Method

2.2 Needleman Wunsch Algorithm

The Needleman-Wunsch algorithm is an implementation of dynamic programming, which is an algorithm for solving problems by decomposing the solution into a set of stages. The Needleman-Wunsch algorithm is used to determine the level of similarity or compatibility of two texts. This algorithm is also used to find an alignment that has an optimal value in global alignment in two sequences. This algorithm was created by Saul Needleman and Christian Wunsch in 1970 [4].

In bioinformatics, the degree of difference is usually determined by a score matrix. For example, in dynamic programming algorithms), $A' = (a'_1, a'_2, \dots, a'_n)$, $B' = (b'_1, b'_2, \dots, b'_n)$ is the optimal alignment for paired sequences $A = (a_1, a_2, a_3, \dots, a_n a)$, $B = (b_1, b_2, b_3, \dots, b_n b)$.

The steps for working on the Needleman Wunsch algorithm are:

1. Initialize the matrix
2. Matrix filling
3. Traceback step
4. Determining the alignment result

2.3 Jukes Cantor

This technique is often known as the distance correction method, which aims to 'correct' observed distances by estimating the amount of evolutionary change that has occurred. The aim of the distance correction method is to recover the amount of evolutionary change that has occurred and to 'correct' unobserved distances. As such, it seeks to 'straighten' the lines that represent the observed differences. In this study, the Jukes Cantor evolutionary model was applied as a correction to the distance method. The Jukes Cantor model is one of the simplest sequence evolution models, which assumes that all four nucleotides have the same frequency and that all substitutions occur with equal probability.

3 Research Methods

This research will be divided into several stages as follows:

1. Literature Study
   At this stage of the literature study, a study of the theory related to determining the optimal Hierarchical clustering through the combination of the Needleman Wunsch Algorithm and Jukes antor in clustering tuberculosis diseases is carried out. This process includes finding relevant references to support research such as articles, journals, theses and books.

2. Types and sources of data
   The type of data used in this research is GenBank data obtained from the National

Biotechnology Information Center. GenBank, a database managed by the National Center for Biotechnology Information (NCBI), stores DNA and RNA sequence data from various organisms and provides open access to a vast amount of genetic information. This data includes details about genetic structure and variation, as well as information related to biotechnology and biomedical studies. The data used in this research was accessed through the following link: https://www.ncbi.nlm.nih.gov/nuccore/CP013839.1?report=fasta.

3. Data Processing Technique

In this study, the type of data used was GenBank taken from the national biotechnology information center. With the application of 3 methods. The first method used is the Needleman Wunsch method. According to [4] in this model, alignment is carried out on tuberculosis DNA data using Equation (1).

$$S(i,j) = max\begin{Bmatrix} S(i-1,j-1) + s(a_i, b_j) \\ S(i-1,f) - d \\ S(i,j-1) - d \end{Bmatrix} \quad (1)$$

where:

$S(i-1,j-1)$ = upper left diagonal element of matrix S
$S(i,j-1)$ = matrix element S on the left $S(i,j)$
$S(i-1,j)$ = elements of the matrix S above $S(i,j)$
$s(a_i, b_j)$ = matrix element substitution of residue *i* in sequence *a* and residue *j* in sequence *b*
d = gap penalty / researcher score on virtual symbol

Next, calculate the difference distance between two lines using the jukes cantor method with the Equation (2):

$$d = -\frac{3}{4}\ln\left(1 - \frac{3}{4}p\right) \quad (2)$$

The last step is to form a phylogenetic tree using the hierarchical clustering method. This method uses 6 kinds of models, namely: Average, centroid, median, ward, weighted and complete. Where each model is made into 3 clusters.

- Centroid Linkage Method (Centroid)

The linkage centroid is the average of all objects in the cluster. The centroid algorithm starts by defining a matrix $D = \{d_{ij}\}$ to obtain the distance of the most similar objects, for example U and V. The centroid of the newly formed cluster is based on Equation (3).

$$d_{(uv)w} = \frac{n_u d_{(uw)} + n_v d_{(vw)}}{n_{(uv)}} - \frac{n_u n_v d_{(uv)}}{n_{(uv)^2}} \quad (3)$$

- Average Linkage Method

Average Linkage is a clustering process based on the average distance between objects. The average linkage algorithm begins by defining a matrix $\boldsymbol{D} = \{dij\}$ to obtain the closest object, for example U and V, then this object is combined into a cluster (UV) and then the distance between (UV) and other clusters W, therefore it can be written in Equation (4):

$$d_{(uv)w} = \frac{d_{(uw)} + d_{(vw)}}{n_{(uv)}} \quad (4)$$

- Ward Linkage Method

The ward linkage method is a clustering method to minimize the variation between objects that are in one cluster. The ward algorithm starts by defining a matrix $\boldsymbol{D} = \{dij\}$ to obtain the most similar objects, for example U and V, then these objects are combined into a cluster (UV) and then the distance between (UV) and other clusters W, therefore it can be written as follows in Equation (5).

$$d_{(uv)w} = \frac{[(n_w + n_u)d_{(uw)} + (n_w + n_v)] - n_w d_{(uv)}}{n_w + n_{(uv)}} \quad (5)$$

- Complete Linkage Method

Complete Linkage is a clustering process based on the furthest distance between objects defined in Equation (6).

$$d_{(uv)w} = \max(d_{uw}, d_{vw}) \quad (6)$$

Where d_{uw} is the distance of the furthest neighbour of clusters U and W and d_{vw} is the distance of the furthest neighbour of clusters V and W.

- Median Linkage Method

Median linkage is a grouping method based on distance based on the minimum distance between object defined in Equation (7).

$$d_{(uv)w} = \frac{n_u d_{(uw)} + n_v d_{(vw)}}{n_{(uv)}} - \frac{n_u n_v d_{(uv)}}{n_{(uv)^2}} \quad (7)$$

- Weighted Linkage Method

The weighted method is one of the algorithms that can be used in hierarchy clustering process of agglomerative or centralize methods [7]. The equation is defined in Equation (8).

$$d_{(uv)w} = \frac{d_{(uw)} + d_{(vw)}}{2} \quad (8)$$

4. Simulation
   In this step, the data to be used is applied to the Needleman Wunsch, Jukes Cantor, and Hierarchical Clustering models in Matlab software.

5. Analysis and Discussion
   The alignment results will be analysed to determine the identity percentage of each alignment series, as well as the pattern of tuberculosis spread to various countries.

6. Conclusions
   At this stage is the last stage in completing the research. After obtaining results from the application of the needleman wunsch, jukes cantor and hierarchical clustering methods, conclusions and suggestions can be drawn from this research.

4 Results and Discussion

This study uses 1000 pairs of DNA sequences taken from Genbank, a large database containing DNA and RNA sequence information, managed by the National Centre for Biotechnology Information (NCBI). The data comes from various countries such as Ethiopia, Vietnam, Japan, Ireland, Brazil, and others, representing different organisms or global samples. The sequence data is encoded in FASTA format, which is a standard format for storing biological sequences. This format allows efficient processing and analysis of data using bioinformatics software. The study involves analyzing genetic data from various countries in a format that is easy to access and analyze [8].

4.1 Needleman Wunsch

At this initial stage, DNA data was aligned using the Needleman Wunsch model, and the results are shown below.

GCC – CTTC – T - - T - G-GC – GATGA…….. - - T - ACA - - - -

: || | : : | : | | | | | | | : |……. | | ||

CCCTCGGCGAGGTCGCGCGGATCA……. GATGACACCGG

Sequence same R = 1973, sequence different type P = 514.

4.2 Jukes Cantor

From the results of the alignment, the distance of the data differences can be calculated based on the obtained sequence. Below is displayed the matrix of the distance calculations using Jukes-Cantor.

```
Columns 1 through 13

         0         0         0         0         0         0
         0         0    0.1177    0.3222    0.0390    0.1028
         0    0.1177         0    0.1466    0.2537    0.4544
         0    0.3222    0.1461         0    0.1026    0.2290
         0    0.0390    0.2547    0.1026         0    0.2873
         0    0.1028    0.4535    0.2290    0.2873         0
         0         0         0         0         0         0
         0    0.1241    0.2808    0.1645    0.2418    0.4234
         0    0.1265    0.0022    0.1515    0.2438    0.4333
         0    0.1659         0    0.0599    0.0002    0.0104
         0    0.1208    0.4656    0.2746    0.2354    0.3946
         0    0.1293    0.0018    0.1526    0.2397    0.4251
         0    0.1309    0.0024    0.1570    0.2339    0.4130
         0         0    0.1177    0.3222    0.0390    0.1028
         0    0.1291    0.0016    0.1517    0.2426    0.4261
         0    0.1242    0.2773    0.1649    0.2474    0.4274
```

4.3 Hierarchical Clustering

In this analysis, 6 methods were used, namely Centroid Linkage Method, Complete Linkage, Average Linkage, Ward Linkage Method, Median Linkage Method, and Weighted Linkage Method. The clustering process in this study was conducted using MATLAB R2019a. Below are the results of the clustering that has been performed.

Table 1 Value of each cluster method

	Average2	**Average 3**	**Centroid 2**	**Centroid 3**	**Median 2**	**Median 3**
Sequence 1	2	2	2	2	2	2
Sequence 2	2	2	2	2	2	2
Sequence 3	2	2	2	2	2	2
Sequence 4	2	2	2	2	2	2
Sequence 5	2	2	2	2	2	2
Sequence 6	2	1	2	1	2	1
Sequence 7	2	2	2	2	2	2
Sequence 8	2	2	2	2	2	2
Sequence 9	2	2	2	2	2	2
Sequence 10	2	2	2	2	2	2
Sequence 11	1	3	1	3	1	3
Sequence 12	2	2	2	2	2	2
Sequence 13	2	2	2	2	2	2
Sequence 14	2	2	2	2	2	2
Sequence 15	2	2	2	2	2	2
Sequence 16	2	2	2	2	2	2
	Ward 2	**Ward 3**	**Weighted2**	**Weighted3**	**Complete2**	**Complete3**
Sequence 1	2	2	2	2	1	3

Sequence 2	1	3	2	2	2	2
Sequence 3	2	1	2	2	1	3
Sequence 4	2	2	2	2	1	3
Sequence 5	1	3	2	2	2	2
Sequence 6	1	3	2	1	2	1
Sequence 7	2	2	2	2	1	3
Sequence 8	2	2	2	2	1	3
Sequence 9	2	1	2	2	1	3
Sequence 10	2	2	2	2	1	3
Sequence 11	1	3	1	3	2	2
Sequence 12	2	1	2	2	1	3
Sequence 13	2	1	2	2	1	3
Sequence 14	1	3	2	2	2	2
Sequence 15	2	1	2	2	1	3
Sequence 16	2	2	2	2	1	3

4.4 Selection of the Best Method

According to Mingoti and Lima (2006), the quality of clustering results can be seen from the internal distribution within the group. ICD-rate is a metric used to evaluate clustering quality, where the smaller the value, the better the clustering results, as it indicates that the data within the cluster is more homogeneous [9]. The results of the best method selection can be seen in Table 2.

Table 2 Icdarate values for each model

	SST	SSW	R_Squared	ICD-rate
Average Linkage Cluster 2	4,4195	3,8694	0,2569	0,7431
Average Linkage Cluster 3	4.4195	3,8806	0,0864	0,9136
Centroid Linkage Cluster 2	4,4195	4,2976	0,0284	0,9716
Centroid Linkage Cluster 3	4,4195	4,0679	0,0864	0,09136
Median Linkage Cluster 2	4,4195	4,2976	0,0284	0,9716
Median Linkage Cluster 3	4,4195	3,8694	0,0864	0,9136
Ward Linkage Cluster 2	4,4195	3,8806	0,1422	0,8578
Ward Linkage Cluster 3	4,4195	4,1980	0,1389	0,8611
Weighted Linkage 2	4,4195	4,1980	0,0528	0,9472
Weighted Linkage 3	4,4195	4,0679	0,0864	0,9136
Complete Linkage 2	4,4195	3,7839	0,1680	0,8320

Complete Linkage 3	4,4195	3,8489	0,1483	0,8517

From the results of the analysis, it can be concluded that the best method to use is Average Linkage in the second grouping, where the Icdarate value is smaller than the other methods, namely 0.7431.

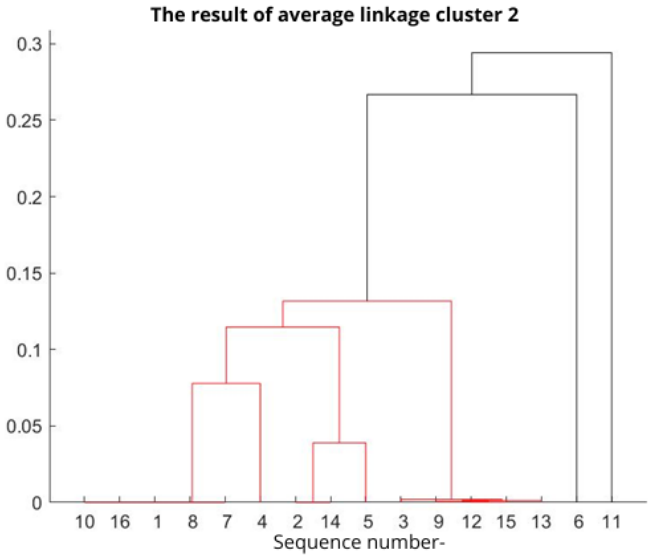

Figure 1 Group and distance

5 Conclusions

From the discussion that has been done, the following conclusions can be drawn: there are mutations between the resulting lines due to the differences detected. This mutation has a very high probability of occurring due to variations in the types of viruses that exist. Based on the simulation results that calculate the distance between lineages, a phylogenetic tree using the Jukes-Cantor method has been produced that shows 2 groups. The first group consists of 9 interconnected rows, while the second group consists of 7 interconnected rows.

6 Acknowledgments

The authors would like to thank all those who have provided support and contributions to this research. In particular, we would like to thank our supervisors for their guidance and input in the preparation of this article.

7 References

[1] Ramadhan, M. R., Waluya, S. B. & Kharis M., 2018, *Pemodelan Matematika Penyebaran Penyakit Tuberkulosis dengan Strategi DOTS*, UNNES Journal of Mathematics, 7(2), 130-141.

[2] Rafflesia, U., 2014, *Model Penyebaran Penyakit Tuberkulosis (TBC)*, Jurnal Gradien,**10**(2), 983-986.

[3] Sari, Y. P., Primajaya, A. & Irawan, A. S. Y., 2020, *Implementasi Algoritma K-Means untuk Clustering Penyebaran Tuberkulosis di Kabupaten Karawang*, Journal INOVTEK Polbeng, **5**(2), 229-239.

[4] Dewi, A. R., 2018, *Penerapan Algoritma Needleman-Wunsch untuk Mengidentifikasi Mutasi pada Sekuen DNA Virus Korona,* Undergraduate Thesis, Mathematics, Institut Teknologi Sepuluh Nopember , Surabaya.

[5] Nicolaus, Sulistianingsih, E. & Perdana, H., 2016, *Penentuan Jumlah Cluster Optimal Pada Median Linkage Dengan Indeks Validitas Silhouette*, Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster), **5**(2), 97-102.

[6] Paramadina, M., Sudarmin & Aidid, M. K., 2019, *Perbandingan Analisis Cluster Metode Average Linkage dan Metode Average Linkage dan Metode Ward (Kasus: IPM Provinsi Sulawesi Selatan)*, VARIANSI: Journal of Statistics and Its Application on Teaching and Research, **1**(2), 22-31.

[7] Dani, A. T. R., Wahyuningsih, S. & Rizki N. A., 2019, *Penerapan Hierarchical Clustering Metode Agglomerative pada Data Runtun Waktu*, Jambura Journal of Mathematics, **1**(2), 64-78.

[8] Muhamad, F., Ahmad, R., Asi, S. & Murad, M., 2018, *Performance Analysis Of Needleman-Wunsch Algorithm (Global) And Smith-Waterman Algorithm (Local) In Reducing Search Space And Time For Dna Sequence Alignment*, in IOP Conference Proceeding.

[9] Istichana Y. Y., 2015, *Pengelompokan Kabupaten/Kota Berdasarkan Pola Makan Penduduk Penderita Penyakit Stroke di Provinsi Jawa Timur Menggunakan Analisis Faktor dan Analisis Hierarchical Clustering,* Undergraduate Thesis, Mathematics, Institut Teknologi Sepuluh Nopember, Surabaya.