Optimization of Metagenome Sequence Identification with Naive Bayes and Certainty Factor

IOP Conf. Series: Materials Science and Engineering **621** (2019) 012002 doi:10.1088/1757-899X/621/1/012002

Optimization of Metagenome Sequence Identification with Naive Bayes and Certainty Factor

Dian Kartika Utami[1], Herfina[2] and Iyan Mulyana[3]

[1] Diploma, Universitas Pakuan, Bogor, Indonesia
[2,3] Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Pakuan, Bogor, Indonesia

E-mail: diankartikautami@unpak.ac.id

Abstract. Metagenome studies are an important step in taxonomic grouping. Taxonomic grouping can be done using the binning method. Binning is a process to determine the contigs of each group of phylogenetic species. In this study, Binning was carried out using the Supervise Learning approach. We use the Naïve Bayes Classifier method and Certainty Factor. The classification process is carried out on phylum taxon levels. many of the organisms used were 50 organisms and the length of the fragments used was 500 bp and many readings were 1000 readings. The accuracy results obtained by the Naive Bayes method are 62.5%. While the accuracy obtained in the Certainty Factor method is 54.45%. From the results of the two methods of testing, it can be concluded that Naive Bayes is the best method of classification compared to Certainty Factor.

Keywords: Metagenome Sequence, Binning, Naïve Bayes, Certainty Factor, Metagenome Sequence

1. Introduction

Metagenom is a microorganism which is based on the sequencing and analysis of the DNA of a community of microorganisms in nature, such as land, air, sea, or the entrails of living beings. The scope for the genome is based on the extent and depth of the environment of the genome [1]. The DNA sequencing process of the microorganism community directly produces fragments from various organisms that mix, so it is necessary to do a binning process to reduce errors in DNA assembly. Based on composition and homology the binning method consists of two approaches [2].

Based on research that has been done [3] to classify metagenom fragments using training data from 381 organisms. The method used is a Multiclass Support Vector Machine (SVM) with Spaced K-mers frequency as its features. Accuracy results at 400bp fragment length is quite good, namely 65.3% in genus taxon, 72.0% in taxon order, 78.2% in class taxon, 82.1% in phylum taxon. At 10Kbp fragment length the accuracy obtained is 95.4% - 97.6%. Studied the classification of metagenom fragments using the Naive Bayes Classifier (NBC) method and used 381 organisms and used variations in fragment lengths of 400 bp, 800 bp, 1 kbp, 3 kbp, 5 kbp and 10 kbp and used feature extraction mers. The fragment with a length of 400 bp resulted in an accuracy of 49.34% for extraction of 3-mer features and for extraction of 4-mer features to 53,95%, while for fragment lengths of 10 kbp the accuracy increased to 82.23% for 3-mer features extraction and for extraction of 4-mer feature to 85,89% [4]. The research conducted by Harun was a classification of metagenom fragments using the method Oblique Decision Tree method with genetic algorithm optimization (ODT-GA) [5]. The data used consists of only 10 organisms belonging to 3 genera. The features used are limited to k = 2, k = 3, and k = 4. For the length of the fragments used are 200 bp, 500 bp, 1 Kbp, 5 Kbp, and 10 Kbp.

Based on previous metagenom fragment classification studies, a study was conducted to optimize the identification of metagenomic sequences consisting of fragments of 50 organisms. Optimization using Naive Bayes and Certainty Factor.

2. Literature Study

2.1. Metagenome Sequence

Metagenomes are sequences and DNA analysis of microorganisms taken directly from the environment without a culture process. The fragments obtained from this metagenome contain various organisms so that the grouping process needs to be done to avoid errors in assembly. Taxonomic grouping is an important step in the metagenom studies [6].

2.2. Binning

Binning is one of the important steps for the assembly sequence in reaching contigs (adjacent fragments) which consist of a single grouping genome [7]. In binning there are 2 approaches, namely homology and composition. The homology approach is an approach that compares DNA fragments to databases which are references in aligning data sequences. Whereas composition-based approach has several advantages, namely as a shortcut to the need for sequence alignment, as input vectors resulting from extraction of base pair characteristics can be calculated as a composition feature and as input for observation by observation.

2.3. Naïve Bayes

Naïve Bayes is based on the Bayes Theorem, assuming that each feature in the classification does not depend on each other [8]. Naïve Bayes is used to calculate the probability of a class of each group of attributes that exist and determine where the most optimal class. The process of data classification with NBC illustrated in Figure 1.

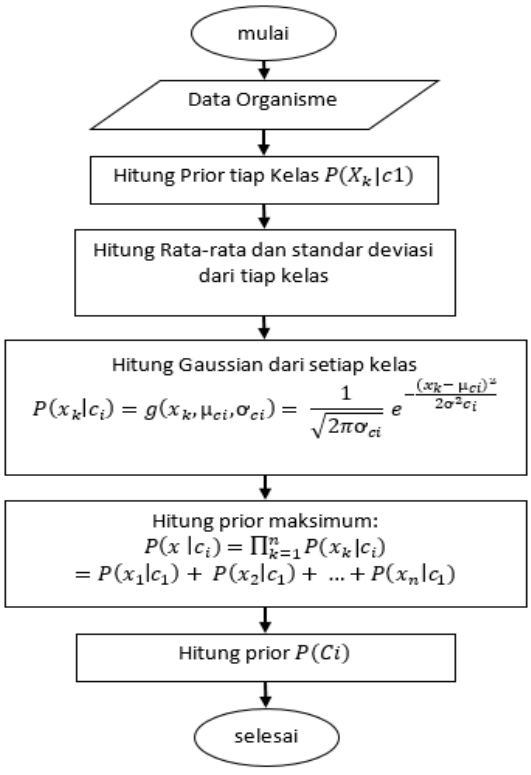

Figure 1. Classification stages with the Naive Bayes

2.4. Certainty Factor

Certainty Factor expresses confidence in an event (fact or hypothesis) based on evidence or expert judgment [9]. Certainty factor uses a value to assume an expert's degree of confidence in a data.

$$CF[H,E] = MB[H, E] - MD[H, E] \quad (1)$$

Information :

CF [H, E] = certainty of the hypothesis factor which is influenced by evidence e is known with certainty.

MB [H, E] = measure of believe on hypothesis H, if given evidence E (between 0 and 1).

MD [H, E] = measure of disbelieve against evidence H, if given evidence E (between 0 and 1).

The data classification process with Certainty Factor is illustrated in Figure 2.

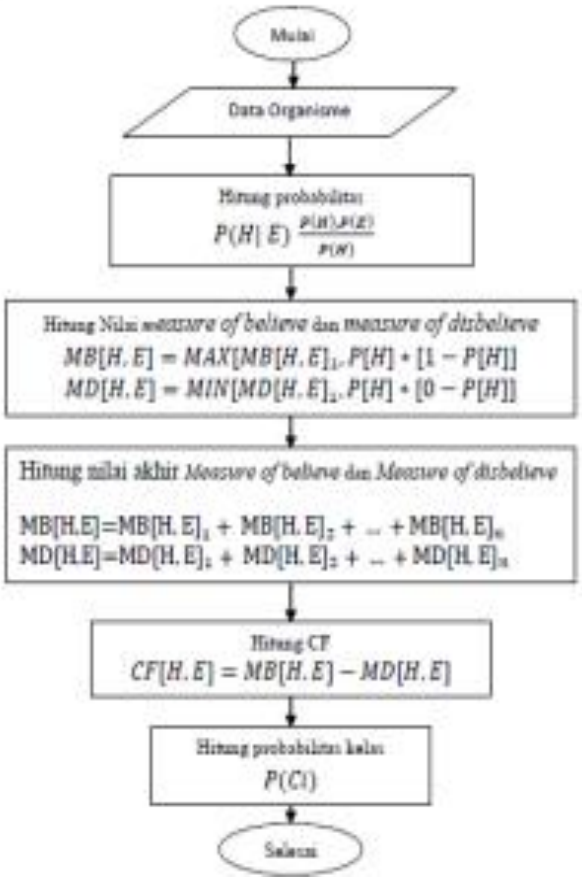

Figure 2. Classification Stages with Certainty Factor

3. Result

In this study metagenom classification was carried out by the Naive Bayes method and Certainty Factor, and the data pre-processing method used K-mers as feature extraction.

3.1. Pre-Processing

In the preprocessing phase of the data, the DNA sequence of metagenome that has been selected from the NCBI site will be broken down using MetaSim software. The steps taken using MetaSim are:

1. Fail containing the DNA sequences of microorganisms that have been downloaded from NCBI incorporated into the software MetaSim. The stages of inserting files into MetaSim software can be seen in Figure 3.
2. After entering the data from NCBI MetaSim into the software, the next process is to choose some of the DNA sequences of microorganisms that have been available in the database according to the needs of research. Database of DNA sequences of microorganisms can be seen in Figure 4.
3. The next step is to set the preset. The setting at preset done to set the length of fragments and the number of readings to be used. Preset settings can be seen in Figure 5.
4. The last step is to run the simulator according to the presets that have been set. The results of the DNA sequences of microorganisms that have been successfully described can be seen in Figure 6.

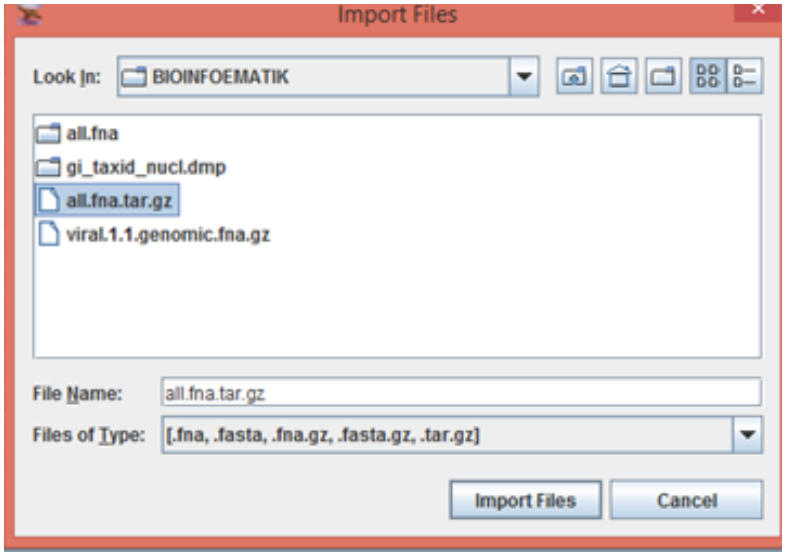

Figure 3. The process of entering metagenomic data on MetaSim

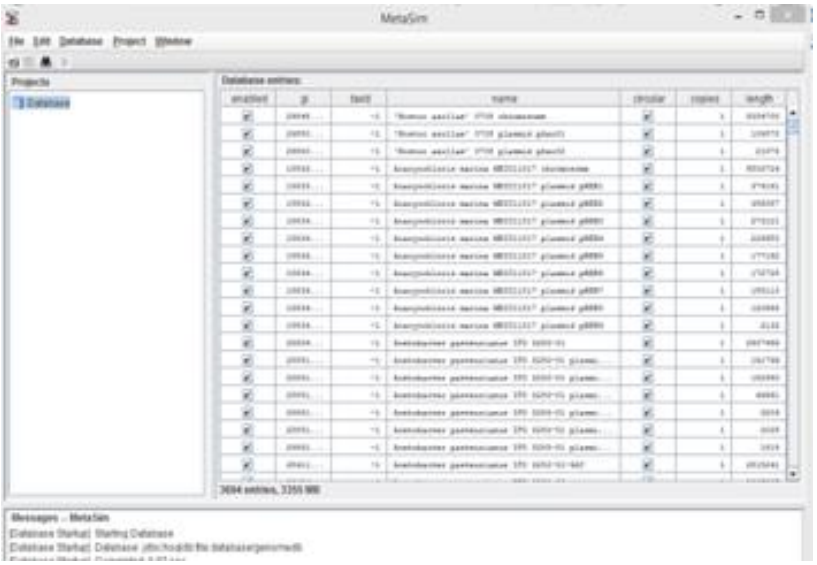

Figure 4. Database of DNA sequences of microorganisms on MetaSim

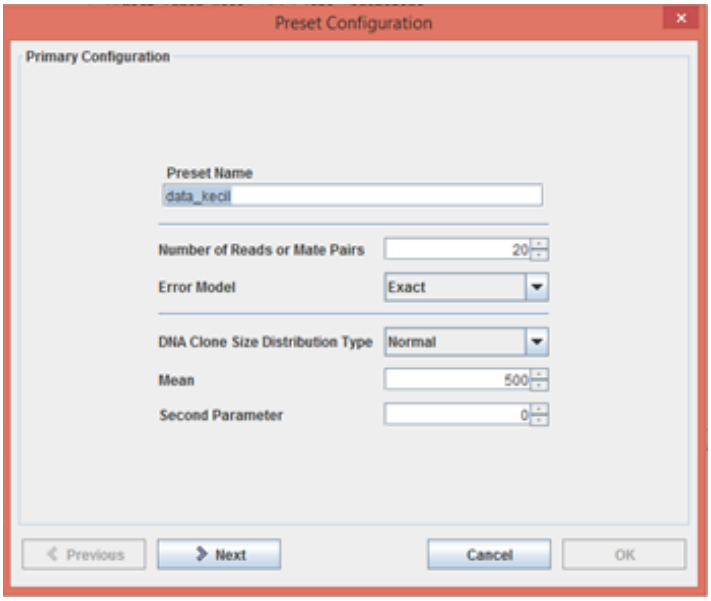

Figure 5. Preset

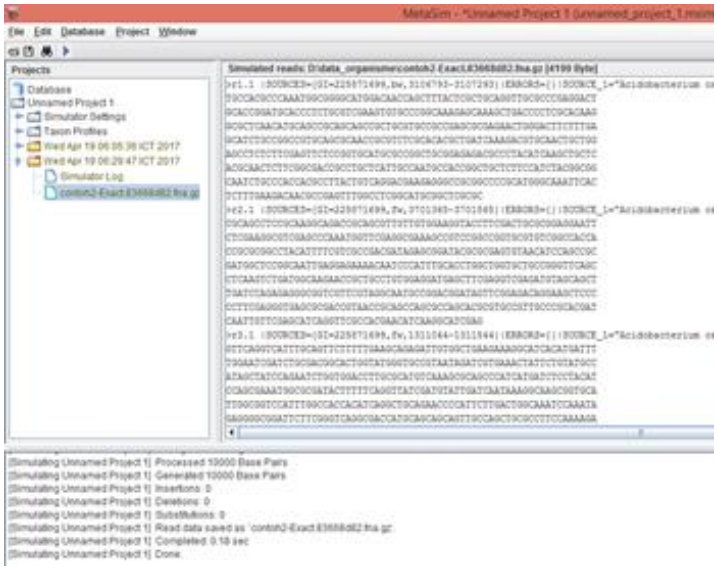

Figure 6. The results of the DNA sequences of microorganisms are successfully parsed

3.2 Data Fragment Metagenom

The processed data is obtained from the preprocessing data, the DNA sequences of metagenomes that have been selected from the NCBI site, and the fragments have been described using the MetaSim simulator. Data to be processed will be read many times according to research needs. In the example of this study, the prepared data was read 1000 times. This means that there are 1000 lines of microorganism fragments. The length of the fragment used is 500 base pair (bp). Fragment data that has been described using metasim can be seen in Figure 7.

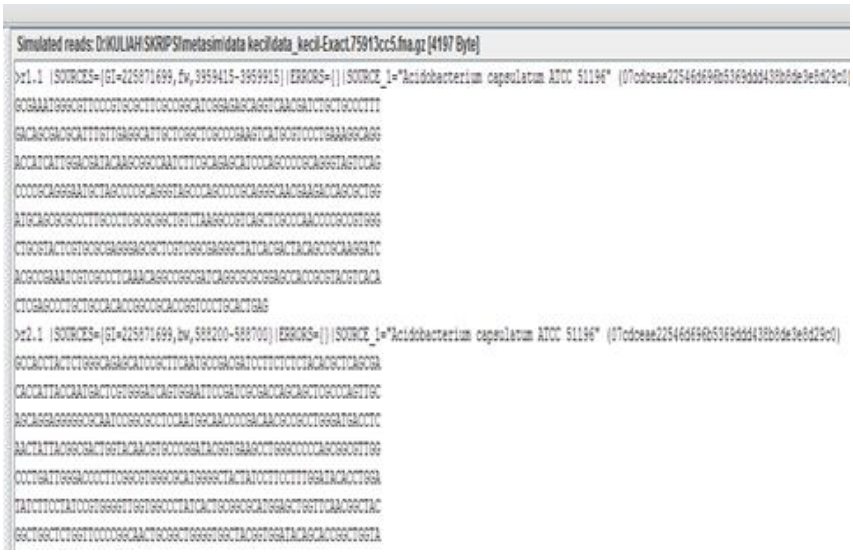

Figure 7. Data Fragment yang sudah di uraikan menggunakan metasim

3.3 Feature Extraction

Feature extraction is done by reading the frequency of nucleotide combinations that may be formed by using k-mers for k = 4. The pattern of occurrence k is a pattern that displays k at a time in a sequence. The pattern of occurrence of k in the sequences is calculated using four main bases (A, C, T, and G) raised by the base pair sequence that you want to use (pattern of occurrence: 4 ^ k, with k ≥ 1). In this study k = 4 means that there will be 4 ^ 4 = 256 patterns of occurrence that are formed. An example for feature extraction using the 4-mers feature can be seen in Figure 8. And it illustrates the calculation results can be seen in Table 1.

3.4 Data Sharing

The data used in this study were divided into two parts, namely the training data and test data. The data used consisted of 50 organisms including 12 different phylum groups. Training data and test data are divided using k-fold cross validation with k = 4. The k-fold cross validation method repeats many times to divide a sample set randomly into mutually independent k-subset. Each replication has one subset used for testing while the rest is used for training.

From one thousand fragments, the division is divided into training data and the test data uses 4-fold cross validation. So there are 4 repetitions for each group of data and in each iteration 250 fragments become test data while the other 750 become training data.

Figure 8. K-mers feature extraction

Tabel 1. Illustration of the results of k-mers frequency calculations on the data of 50 organisms that belong to the phylum 12 to 1000 readings.

No	ID	X1	X2	X3	X4	…	X256	Class
		AAAA	AAAT	AAAC	AAAG	….	GGGG	
1	CCGCCTCGCGGCCACCCGGTGTCAAGGGCGTGAGCGGCAGCGGCAATCTCTGCG……	2	0	0	2	….	0	C1
2	ACGGATGGGAGAAGCGGCCAACGCAGTGGGGCGTGTTCATGGCGCACTTCC……	1	0	2	0	….	3	C1
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
1000	AATTCCAAAATTTAATTGGACATCATTGAGATCTCTTAAAAAATTAAATA……	17	17	2	3	….	0	C12

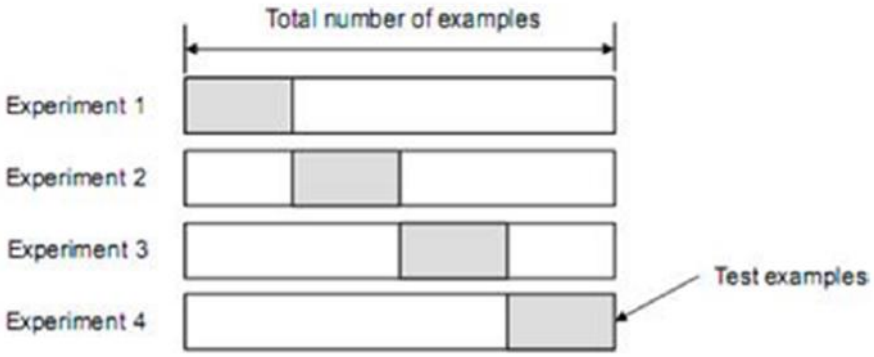

Figure 9. Illustration of k-fold validation [5]

3.5. Calculation of Naïve Bayes Classification

In the process of classification using Naive Bayes, Sequencing DNA data that has been extracted enters the first process. The first process in the classification Naive Bayes is looking for a probability value of each class by using the following formula:

$$P(x_k|Ci) = P(AAAA|C1) = \frac{\frac{2}{496582}}{\frac{1}{256}\cdot\frac{2}{3609} + \frac{1}{256}\cdot\frac{0}{2748} + \ldots\ldots + \frac{1}{256}\cdot\frac{0}{1685}} = 0.007532$$

$$P(x_k|Ci) = P(AAAT|C1) = \frac{\frac{0}{496582}}{\frac{1}{256}\cdot\frac{2}{3609} + \frac{1}{256}\cdot\frac{0}{2748} + \ldots\ldots + \frac{1}{256}\cdot\frac{0}{1685}} = 0$$

After the above process then looks up the value of the average and standard deviation. The following are the average values and standard deviations.

Calculate the average value of each class (μci):

$$\mu ci = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{P(ci)}$$

$$\mu c1(AAAA) = \frac{0.007532 + 0.003766 + \ldots + 0}{27} = 0{,}006117$$

$$\mu c1(AAAT) = \frac{0 + 0 + \ldots + 0.002844}{27} = 0{,}002753$$

Calculate the standard deviation of each class (*σci*):

$$\sigma ci = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

$$\sigma c1(AAAA) = \sqrt{\frac{(0.001415)^2 + (-0.002351)^2 + \cdots\ldots + \cdots..}{27}} = 0{,}010687$$

$$\sigma c1(AAAT) = \sqrt{\frac{(-0{,}002753)^2 + (-0{,}002573)^2 + \cdots + \cdots}{11}} = 0{,}005285$$

Looking Gaussian value using the following formula:

$$P(x_k|c_i) = g(x_k, \mu_{ci}, \sigma_{ci}) = \frac{1}{\sqrt{2\pi\sigma_{ci}}}\, e^{-\frac{(x_k - \mu_{ci})^2}{2\sigma^2 c_i}}$$

Where:

$$\pi = 3{,}14$$

$$e = 2{,}71828$$

$$P(AAAA|C1) = g(x_k, \mu_{ci}, \sigma_{ci}) = \frac{1}{\sqrt{2(3{,}14 \times 0{,}010687)}} 2{,}71828^{-\frac{(2-0{,}006117)^2}{2\times 0{,}010687^2}} = 3{,}826334$$

After the input into the probability density function gauss then find the maximum value for each classification. Looking for maximum prior values using the following formula:

$$P(x\ |c_i) = \prod_{k=1}^{n} P(x_k|c_i)$$

$$\text{P(x | C1)} \qquad = P(x_1|c_1) + P(x_2|c_1) + \ \ldots\ + P(x_n|c_1)$$

$$= P(AAAA|C1) + P(AAAT|C1) + \ldots\ldots + P(GGGG|C1)$$

$$= 3{,}826334 + 3{,}276811 + \ \ldots\ldots + \ \ldots\ldots = 35771{,}589737099$$

After determining the maximum value we determine the probability by dividing the maximum value Phylum the overall maximum amount. Next the results of the probability can be seen in Figure 10.

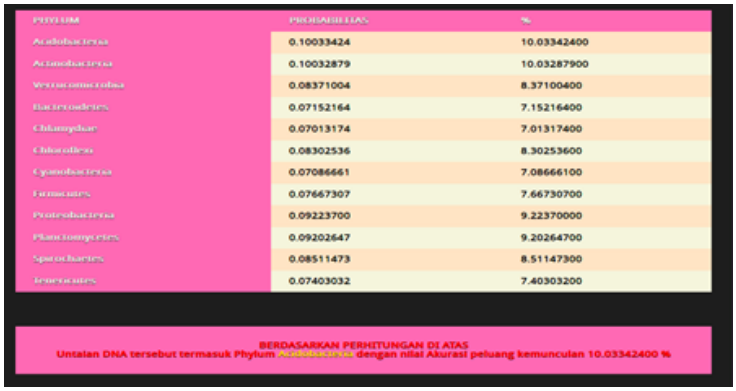

Figure 10. The results of the probability value

3.6. Calculation of the Certainty Factor Classification

In the process of classification using certainty factor, Sequencing DNA data that has been extracted enters the first process. The first process in the classification with the certainty factor is to find the probability values by using the following formula:

$$P(H|\ E)\ \frac{P(H),P(E)}{P(H)}$$

$$P(AAAA|C1) = \frac{\frac{2}{3609}}{\frac{1}{256}} = 0{,}141868$$

Value 2 (two) is an emergence pattern AAAA feature 2 (two) times, 3609 is the total value of the pattern emergence AAAA feature divided by 1/256 emergence pattern features base pairs. After the above process then looks up the value Measure of believe and disbelieve Measure of value. The following are the values of MB and MD.

Calculate the value of MB and MD by using the formula:

MB[H.E]=MAX[MB[H.E]_1.P[H]*[1-P[H]]
=[Max[0.141868 ,0.027]-0.027]*[1-0.027] = 0.111766

MD[H.E]=MIN[MB[H.E]_1.P[H]*[0-P[H]]
=[Min[0.141868,0.027]-0.027]*[0-0.027] = 0

The value of 0.027 is the average value of class C1. After the above process, then add the final value of Measure of believe and the value of Measure of disbelieve each data.

$$MB[H.E] = MB[H,E]_1 + MB[H,E]_2 + \dots + MB[H,E]_n$$

$$= MB(AAAA|C1) + MB(AAAT|C1) + \dots\dots + MB(GGGG|C1)$$

$$= 0.111766 + 0 + \dots\dots + \dots\dots = 54{,}506$$

$$MD[H.E] = MD[H,E]_1 + MD[H,E]_2 + \dots + MD[H,E]_n$$

$$= MD(AAAA|C1) + MD(AAAT|C1) + \dots\dots + MD(GGGG|C1)$$

$$= 0 + 0.000729 + \dots\dots + \dots\dots = 0.051179$$

Calculate CF by using the formula:

$$CF[H,E] = MB[H,E] - MD[H,E]$$

$$CF[H,E] = 54{,}50661 - 0.051179$$

$$= 54.454851$$

The value of CF 54.4548 is obtained from the reduction in the values of mb and md obtained from the previous stage. After obtaining the certainty factor value, the results are obtained along with the type of phylum.

3.7. Sensitivity and Specificity

In this study, in addition to the application of methods Naive Bayes and Certainty Factor in classifying sequences metagenom to seek maximum accuracy, it has also been conducted in the process of testing the accuracy of classification using Sensitivity and Specificity approach. Sensitivity is used to measure the proportion of true positives correctly identified, while specificity to measure the proportion of negatives which are correctly identified. Sensitivity and specificity calculations performed on each phylum to see the accuracy of the model was generated against each phylum. The results of calculation of sensitivity can be seen in Figure 11 and Figure 12, the calculation of specificity can be seen in Figure 13 dan 14.

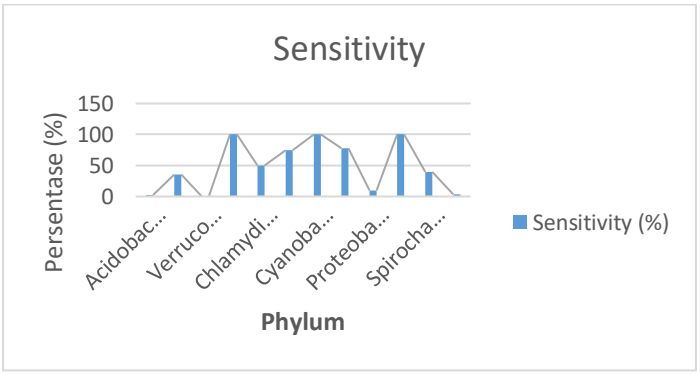

Figure 11. Sensitivity with Naïve Bayes Method

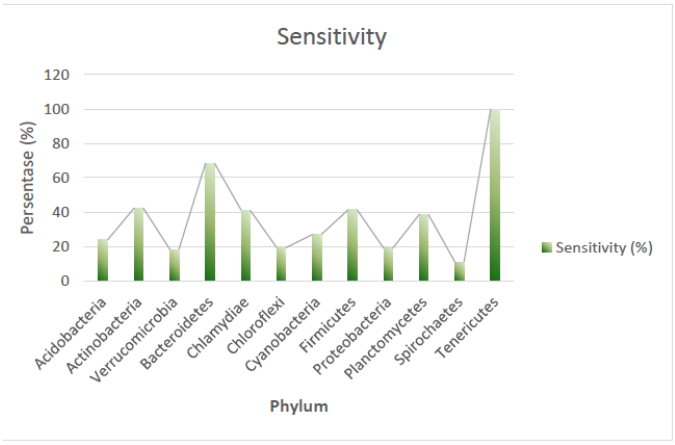

Figure 12. Sensitivity with Certainty Factor Method

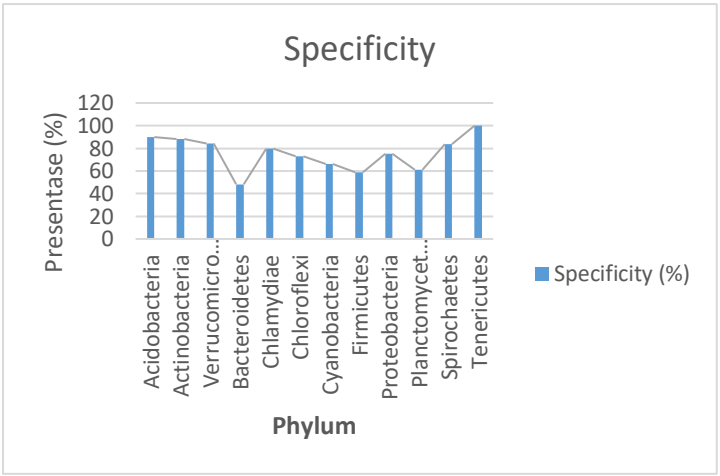

Figure 13. Specificity with Naïve Bayes Method

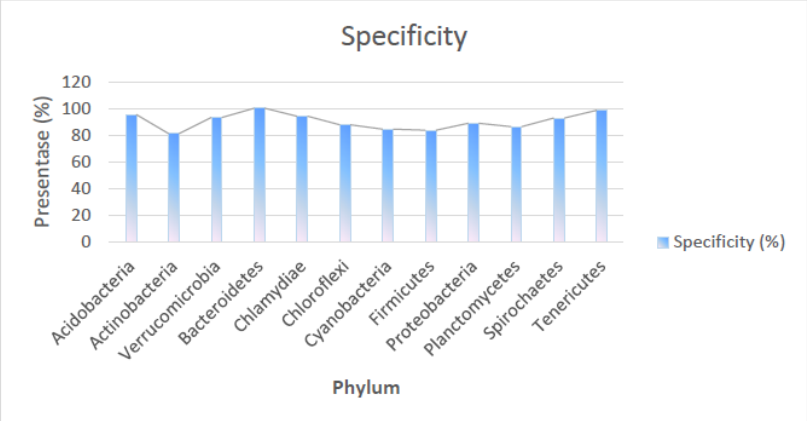

Figure 14. Specificity with Certainty Facor Method

4. Conclusion

The application of the Naïve Bayes method and Certainty Factor in the classification of metagenom uses 50 organisms belonging to the 12 phylum, namely: Acidobacteria, Actinobacteria, Verrucomicrobia, Bacteroidetes, Chlamydiae, Chlorophlexi, Cyanobacteria, Firmicutes, Proteobacteria, Planctomycetes, Spirochaetes, Tenericutes.

Based on the calculation of the values of sensitivity and specificity in the Naive Bayes method that have been carried out, the values obtained for sensitivity range from 0 to 100% while the values for specificity range from 48% to 100%. While based on the calculation of the value of sensitivity and specificity in the method of Certainty Factor, the value obtained for the sensitivity ranged from 11% to 100%, while the value for specificity ranged from 81% to 100%.

The application of the Naïve Bayes method to the classification of metagenome does not produce output that is in accordance with the class. Many sequencing DNA that is input is classified into another phylum. An example is the phylum acidobacteria the test data used by 27 sequences but only 6 sequences are read in the phylum correct. While the application of the classification method of Certainty Factor this metagenom DNA Sequencing widely read when in the classification of the phylum to which Bacteroidetes 202 sequences and a low of Verrucomicrobia read as much as 2 sequences in the phylum correct.

References

[1] Maureen A and O'Malley 2006 *Metagenomics*. Sydney-Australia. University of Sydney.

[2] Wooley J C, Godzik A and Friedberg I 2010 A primer on metagenomics. *PLos Computational Biology*. 6(2):1–13. doi: 10.1371/journal.pcbi.1000667.

[3] Ariny 2013 Klasifikasi Fragmen *Metagenome* Menggunakan Metode *Support Vector Machine* (SVM) [skripsi]. Bogor-Indonesia : Institut Pertanian Bogor.

[4] Utami D K 2014 Klasifikasi Metagenom dengan Metode *Naïve Bayes Classifier* [tesis]. Bogor-Indonesia : Institut Pertanian Bogor.

[5] Harun A S 2014 Kalsifikasi Fragmen *Metagenome* Menggunakan *Oblique Decision Tree* dengan Optimasi Algoritme Genetika [skripsi]. Bogor-Indonesia. Institut Pertanian Bogor.

[6] Higashi S, Barreto André da MS, Cantão ME and de Vasconcelos ATR 2012 *Analysis Of Composition-Based Metagenome Classification*. BMC Genomic 2012, 13(Supply 5):S1. http://www.biomedcentral.com/1471-2164/13/S5/S1

[7] Strous M, Kraft B, Bisdorf R and Tegetmeyer H 2012 *The binning of metagenomic contigs for microbial physiology of mixed cultures.* Frontiers in Microbiology 3.

[8] Han J, Kamber M 2001 *Data Mining Concepts and Techniques*. Cerra DD, Severson H, Breyer B, editor. San Diego (USA): Academic Pr. ISBN 978-0-12-381479-1.

[9] Turban E and Aronson J E 2001 Desecion Support System and Intelligent System, 6th. Edison; Prentice Hall International Edition, New Jersey.