Performance of Hidden Markov Model structure on Deoxyribo Nucleic Acid Coding Sequence of *Plasmodium falciparum*

Suhartati Agoes, Alfred Pakpahan, and Binti Solihah

Abstract—The Hidden Markov Model (HMM) structure application on coding sequence (CDS) used for exon controlling of DNA *Plasmodium falciparum*. HMM performance parameters are Correlation Coefficient (CC), Sensitivity (Sn), Specificity (Sp) and Approximate Correlation (AC). A CDS of DNA *Plasmodium falciparum* contains exon in many regions at least two exon regions. The properties of HMM such as Markov chain, transition state, emission state, HMM training and HMM testing algorithms. Random values of transition state used for HMM training makes many differences in the performance of the model. Furthermore, the transition state value is very important to find the optimum performance of the models. The improvement of the models of HMM structure is using the increasing number of states on CDS with HMM method. The simulation result predicted that the performance parameters values are depends on the value of transition state and the number of states on the model.

Index Terms—Coding Sequence, Performance of Hidden Markov Model, *Plasmodium falciparum*

I. INTRODUCTION

To controlling exon in DNA *Plasmodium falciparum* based on coding sequence (CDS) in application the structure of models are similar to the HMM structure in Daniel Nicorici’s paper [1]. Genome *Plasmodium falciparum* belongs to genome eukaryotic and has a long DNA genome and intron or splicing process. Biological model of DNA structure from gene eukaryotic consists of some exon and intron which alternate located. CDS is a result from splicing process of intron inside the DNA and consists of some region of exon. The first region of exon in CDS starts with start codon which is ATG bases and the last region of exon, there’s one of the three stop codons such as TAA, TAG and TGA bases [2] [3]. In the minimum CDS, there are two region of exon which enable us to find out that there’s minimum one region of intron and usually region of intron starts with G and T bases and ends with A and G bases.

Many methods can be used to identify exon and intron or HMM based finding, for example are like GENESCAN [4] and VEIL [5]. This research is using exon condition established on CDS, therefore it needs controlling on exon where there’s possibility of mutation like dilation, insertion, inversion, translocation of exon on CDS. HMM based finding usually gives more accurate results compared to other methods [6]. In this paper, HMM’s models was chosen based on HMM method for several models with its number of states. Performance parameters of HMM’s models are shown by its value for each models.

II. HIDDEN MARKOV MODEL

HMM process can be described that the inputs are the data or DNA sequences and set the number of states to the sequences depends on the model. Furthermore, HMM training has the both algorithms Viterbi and Baum-Welch and needs the transition state and emission state values for the process. The result of HMM training is the estimated transition state and emission state. The estimated transition and emission state used for HMM testing process has both of the algorithms above. The result of HMM testing is the estimated states of the model. The performance of the model is calculated by comparing the estimated state from HMM testing result with the original state of the input sequences [2] [4], and the parameters can be explained as follows:

a. Sensitivity (Sn)

Sensitivity is an accuracy measurement parameter that shows the comparison between the correctly predicted numbers of nucleotide character from exon/intron in an actual gene with total numbers of nucleotide character in that exon/intron. Mathematically, sensitivity parameter equation is as per below:

$$S_n = \frac{TP}{TP + FN}$$  \hspace{1cm} (1)

b. Specificity (Sp)

Specificity is a comparison between numbers of nucleotide character that correctly predicted from an exon/intron in an actual gene with total numbers of nucleotide character in an exon/intron based on prediction. Specificity parameter equation is as per below:
\[ S_p = \frac{TP}{TP + FP} \]  
(2)

c. Correlation Co-efficiency (CC)
Correlation co-efficiency shows the correlation level between locations of exon based on prediction with the actual location. Correlation co-efficiency parameter equation is as per below:

\[ CC = \frac{(TP \cdot TN) - (TP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \]  
(3)

d. Approximate Correlation (AC)
In the case where coding region in a gene segment is not available or the gene segment is identified as non-coding, therefore CC parameter is un-identifiable because the price of TP and FP will be equal to zero. In this case, Approximate Correlation (AC) is identified for accuracy prediction calculation with broader usage where AC can always be calculated by using below equation:

\[ AC = \frac{1}{2} \left( \frac{TP}{TP + FP} + \frac{TP}{TP + FN} + \frac{TN}{TN + FN} + \frac{TN}{TN + FP} \right) - 1 \]  
(4)

where:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

The illustration for TP, TN, FP and FN used to calculated the HMM performance parameters above as Figure 1.

Several models in this experiment are using start and stop gene like as a state and consist of three bases in each states. Furthermore, the region of intron in the model is able to increase the states by separating bases GT and bases AG from the length of intron. The HMM parameters were set for the HMM training and test method and its algorithms [6] [7].

III. MATERIALS

Matlab R2010a Mathworks, Massachusetts, USA was used for simulations and its hardware PC has specification: Intel(R) Core(TM)2 Duo CPU E8500 @ 3.16 GHz; 3.18 GHz.1.99 GB of RAM; Operating System Microsoft Windows XP, 2002 version, Service Pack 3.

IV. SIMULATIONS AND RESULTS
Based on HMM method for the training uses Viterbi algorithm and testing uses both algorithms are Baum-Welch and Viterbi. HMM training and HMM testing use the same sequences. The programming is written in Matlab R2010a, there’s toolbox bioinformatics to generate DNA sequences in GenBank format and has functions of HMM training and HMM testing.

Dogma’s DNA alternately location for exon and intron on coding sequence for controlling exon DNA Plasmodium falciparum like in Figure 2 below.

![Figure 2. Exon and intron are alternately location on CDS.](image)

General structure design of the model based on HMM method and dogma’s DNA was used for the simulations like as Figure 3, to develop of the model with random increasing the number of state in exon and intron regions on CDS.

![Figure 3. General structures design of the model.](image)

Random value of transition state can be the analysis to find the optimum performance parameters and emission state value has the constant value depends on the models. To calculate HMM performance parameters using the assumption of exon is positive and intron is negative.

Simulation result of this research which the number of states model structure design are 20, 30, 50 and 100, have the HMM performance parameters as Table 1 below.

<table>
<thead>
<tr>
<th>Number of State</th>
<th>Viterbi Algorithm</th>
<th>Baum-Welch Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN</td>
<td>SP</td>
<td>CC</td>
</tr>
<tr>
<td>20</td>
<td>0.9512</td>
<td>0.9330</td>
</tr>
<tr>
<td>30</td>
<td>0.9488</td>
<td>0.9415</td>
</tr>
<tr>
<td>50</td>
<td>0.9569</td>
<td>0.9351</td>
</tr>
<tr>
<td>100</td>
<td>0.9768</td>
<td>0.9334</td>
</tr>
</tbody>
</table>

Sensitivity parameter graph from Table1 using both of algorithms as Figure 4 below is showed that the increasing the

![Table 1. HMM Performance Parameters](image)
number of states on the models becomes $S_n$ value is higher except for number of state 30 is lower.

Figure 4. Sensitivity parameter model charts

Specificity parameter graph from Table 1 using both of algorithms as Figure 5 is showed that the increasing the number of state on the models becomes $S_p$ value is lower except for number of state 30 is higher.

Figure 5. Specificity parameter model charts

Correlation Coefficient parameter graph from Table 1 using both of algorithms as Figure 6 is showed that the increasing the number of state on the models becomes CC value is higher.

Figure 6. Correlation Coefficient parameter model charts

Approximate Correlation parameter graph from Table 1 using both of algorithms as Figure 7 is showed that the increasing the number of state on the models becomes AC value is higher.

Figure 7. Approximate Correlation parameter model charts

5. CONCLUSIONS

Based on the simulation results, we can draw some conclusions as follows:

1. In general, if number of states is increasing, the value of performance parameters model is increasing as well.
2. The HMM performance parameters are better using Viterbi algorithm for HMM testing than Baum-Welch algorithm.
3. The value of random transition state is very influential in determining the optimum value of performance parameter model.
4. It’s better to separate GT bases into G and T states as well as AG bases into A and G states in intron region becomes HMM performance parameters.
ACKNOWLEDGMENTS
The authors would like to thank the Directorate of Higher Education (DIKTI) for fundamental research grant 2011 and Trisakti University Research Institute for all valuable discussions and suggestions.

REFERENCES