ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

Optimization Structure of Hidden Markov Model for Plasmodium Falciparum Gene Prediction

Binti Solihah, Suhartati Agoes, Alfred Pakpahan

Abstract*— Optimized Structure of Hidden Markov Model for exon controlling can be built by attending several aspects, i.e. sequence characteristics and the parameters of HMM Structure. The major sequence characteristics of HMM structure are number of exon in coding sequence and mean of the exon length of each sequence data. The parameters of HMM structure are state definition, length of the state, transition matrix definition, and log likelihood. To predict the accuracy of HMM structure with specific value of the parameters, a back propagation neural network is proposed to bind the searching space of combination problem on the HMM structure for exon controlling. In the experiment, relation between likelihood and result accuracy was identified. Effect of using pseudo transition on training phase to the result of accuracy during data leakage was also identified. The influence of gene data characteristic for development of HMM structure for exon controlling was also identified. The experiment result shows that optimized structure of HMM shown by high correlation coefficient (CC) value of the state was affected by data set choose for training phase, state definition, and initiation of transition matrix. The number of iteration to reach convergent and fluctuation of log likelihood value represent the accuracy of model.

***Index Terms*— Back Propagation, Exon Controlling, Hidden Markov Model, Neural Network, Plasmodium Falciparum.**

I. INTRODUCTION

Eukaryotic gene is composed of a transcribed region. It is divided into coding region (known as exon) and non coding region called intron. Intron position is in between exons. Exons compose an Open Reading Frame (ORF). ORF composition is known as splicing. Exon begins with a start codon (ATG) and the end is marked with a stop codon (one of TAA, TAG, or TGA). To adapt with environment exchange, organism might be mutate on gene level. Mutation process can be insertion, deletion or substitution on the organism genome. When the sequencing result from the mutated genome is provided, one of the approaches to identify gene is by computational method to control exons. The computational method to control exon can be developed with the same principle with computational model for gene finding. Gene finding methods are categorized into two approaches [1], intrinsic method (such as homology) and extrinsic method or prediction method (such as HMM). Homology can determine up to 50% of gene area when the similar sequence is in database. To determine the remaining 50% or more, we can use the prediction method. HMM is one of the approaches to predict. First use of HMM for gene finding appears in [2] to find gene in E. coli DNA. This method has been developed and has been improved for better accuracy by modifying the HMM such as geneMark.hmm [3], Evolutionary HMM [4] that is improved by [5], and class HMM [6]. The standard methods for applying HMM to biological problem is to find a Viterbi path through HMM graph. In [7] sampling from the posterior distribution is shown as a natural way to compute probabilities for predicted exons and gene structures being correct under the assumed model. Formerly, HMM for exon controlling for *Plasmodium falciparum* has been developed in [8] using HMM structure similar to model that developed in [9]. In [8], it is concluded that the more state numbers the larger correlation coefficient (CC). But, larger number of state leads to larger searching space and longer time in finding optimal structure. To save searching time for state composition, the neural network back propagation is proposed to predict a specified HMM structure accuracy. Then, in the experiment and discussion section, effect of several parameters to the accuracy of model is discussed.

II. THE MODEL

A. Hidden Markov Model

In a simple manner, Hidden Markov Model is a statistical approach to determine the hidden parameters of a set of observable parameters. In the exon controlling problem, the observable parameter is the sequence of nucleotides (A, C, G, and T) and the hidden parameter is the position of exon and intron. To build HMM for exon controlling, there are four major steps: 1) defining the structure of model, 2) preparing data for training, 3) training phase and 4) testing phase. The objective of the training phase is to generate model by estimating the observation sequence. To generate the observation sequence, HMM use number of states, state transition probability distribution, emission distribution of the states and initial state distribution [10]. State transition probability distribution is represented in a square matrix n x n which n is the number of state transition, and have to satisfy the stochastic constrain that the sum of the overall column in each row equals to 1 and each elements must be greater than 0. HMM structure developed in this research depicted in Fig 1. During the state definition phase, coding sequence area is divided in three areas, i.e. first exon, introns and other exons area. Each area is populated with a numbers of states defined in first stage. For example, if the number of state is *m* and the number of state in first exon is *i*, in intron area is *j* states, and in others exon is *k* states, then *i+j+k = m*. Intron area is divided into three areas, front (consists of *x* states), middle (consists of *y* states) and back (consists of z states) and *x+y+z = j*. The rules for state numbering of this model are as follows:

1. The first three nucleotides known as start codon are encoded as state 1. The next state and so on are encoded as 2, 3, ..., i-3, respectively and the rest are encoded as i-2.

2. The first, second, and so on of introns are defined as state *i-1*, *i,i+1*, respectively until the *(i+k)*th nucleotide. The next *k*-nucleotides are encoded as state *i+j-(k-1)*. The others are labeled as state *i+j-z*.

3. The first, second, and so on of exons in the back area are encoded as state *i+j+k -1* and the last 3-nucleotides known as stop codon are encoded as state *i+j+k*.

With the specification of HMM state as described above, then start codon will be state 1, stop codon will be state *m*, several nucleotides in the first exon will have the same state number, and n numbers of nucleotide in the same position in exons will have same state number, and *n-1* numbers of nucleotide in the same position in introns will have same state number, where n is the numbers of exon in gene. How to define population of state in each area on the proposed model and how to initialize transition matrix will have significant effect on model accuracy. The influence of transition matrix initialization to model accuracy is shown in Table 1. The same models with different initialization value on state transition yields different accuracy. Defining the population of state in the three areas leads to problem of large searching space and longer finding time due to increasing of searching time. The Neural Network to reduce searching time on the mentioned problem will be discussed in the next chapter.

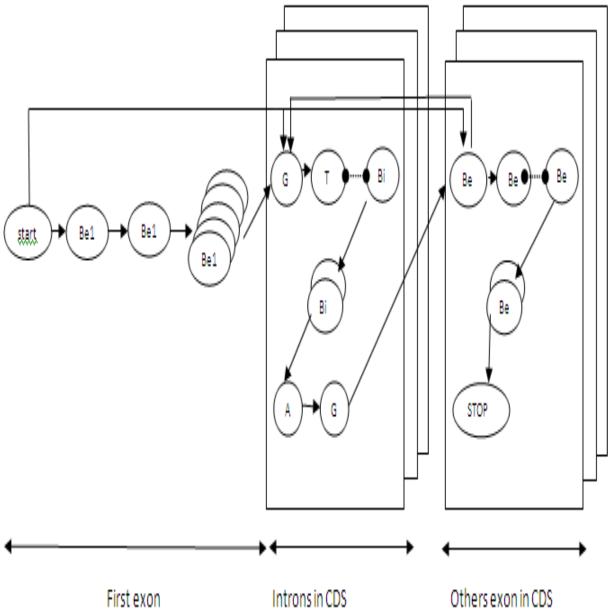


Fig. 1 HMM Structure to Control Exon of CDS Plasmodium Falcipharum

Table 1.Influence of Transition Matrix (Tr) Initialization To CC

No	state		tr on Row 1	tr on Row n	CC
1	20	[10 8 2]	0.999981 0.000001	0.050001 0.849981 0.100001	0.7481
2	20	[10 8 2]	0.849981 0.150001	0.100001 0.799981 0.100001	0.6994
3	20	[10 8 2]	0.899981 0.100001	0.050001 0.849981 0.100001	0.7066

Table 2.Mean Time of Training Phase for HMM Development

No	Number of state	Training time (hours)
1	20	0.191297222
2	30	0.738
3	50	1.236138889
4	150	13.86

B. *Neural Network Back Propagation*

Although it is proven that the larger the state numbers of HMM structure the larger CC of model, it turns out that finding composition of population in each area is a time wasting activity. Table 2 contains a list of mean time of training phase to find optimal structure. Fig. 2 shows that the significant increase on the required time in that phase. Table 3 shows that correlation between percentages of population of state in each area and accuracy is not linear.

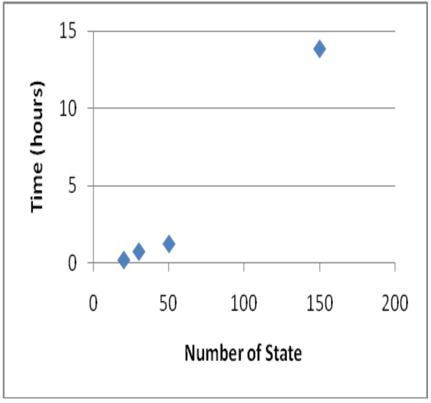


Fig. 2 Mean Time of Training Phase For HMM Development

To prune the spending time in searching of optimal structure of HMM, a neural network defined by back propagation algorithm is proposed to predict the accuracy of a proposed structure. The architecture of proposed neural network is shown in Fig. 3. The neural network have 13 nodes in input layer, two hidden layer composed by 17 and 7 nodes, and one output. Input parameters consist of number of state, number of state in front area, number of state in the middle area, number of state in the back area, the element of transition matrix in the first row, in diagonal, in the predefined row in the same column with the diagonal element, and in the backward element of the model, maximum of exon number in the data training, minimum length of exon, mean length of exon and length of minimum exon.

Table 3. Best Composition for Population of State in Model

No	Number of state	Best composition for Population of state in front, middle and back of model	CC
1	20	[10 8 2]	0.7481
2	30	[13 15 2]	0.7651
3	50	[18 30 2]	0.7727
4	100	[18 80 2]	0.7845
5	150	[45 103 2]	0.7253

C. Data Set Preparation and Implementation

The Data set use in training phase of HMM have a large impact on the built model. Good data that satisfactory from number of aspects will lead to good model. For this exon controlling application, there are four criteria for the experiment data item: sequence just have one CDS, sequence that is not partial sequence (just have one start codon in the beginning of exon position and one stop codon in the end of exon position), the sequence that is not pseudo gene, and there isn't unknown gene in CDS (sequencing process not complete yet).

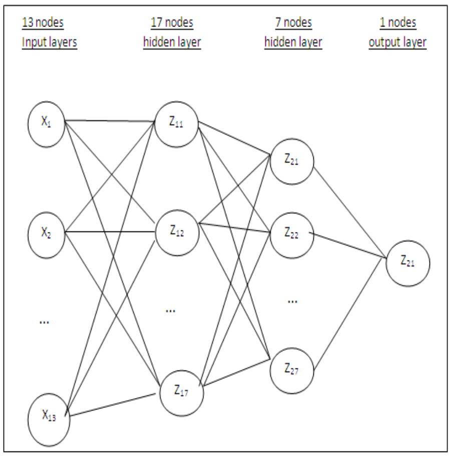


Fig. 3 Neural Network Back Propagation Architecture

In experiment, 152 CDS data of *Plasmodium falciparum* were used as training and testing data. Data set contains sequence with number of exon are variations between 2 until 10. Mean of exon length in sequence is between 34 until 2352. The length of the shortest exon in sequence is between 2 until 1238 and length of longest exon in sequence is between 112 until 9359. The proposed model is implemented with MATLAB R2010a from Math-work and run in PC. The specification of PC are Intel® Core™2 Duo CPU F8500 @3.16GHz; 1.99Gb of RAM; Operating system Micro-soft Windows XP sp 2.

III. EXPERIMENT AND DISCUSSION

A. Scenario

To identify how to get an optimal structure of HMM on n number of states, the experiment was run as follow:

1. Comparing the accuracy of three models that were built with different training data. The first model was built with specific number of exon and the others models were built with the number of exons less than five and more than five.
2. Using neural network back propagation to predict the accuracy of HMM structure to prune the time spending on searching the optimal solution.
3. Identifying the effects of log likelihood fluctuation on training phase to the model accuracy.
4. Identifying the advantages of pseudo transition when leak of data training was occurred.

For the first scenario, the accuracy of model is shown by correlation coefficient (CC). CC is calculated with equation (1):

$$CC = \frac{(TP.TN) - (FP.FN)}{\sqrt{(TP + FP)(TN + FP)(TN + FP)(TN + FN)}}$$

Where TP: True Positive, TN: True Negative, FP: False Positive, and FN: False Negative.

B. Result and Discussion

Accuracy of HMM structure that was developed with specific number of exon compare with the other that developed with various number of exon is shown in Table 4. Higher CC value can be achieved with the unified number of exon in data training.

Table 4. Accuracy of Various HMM Structure Developed By Various Data Training

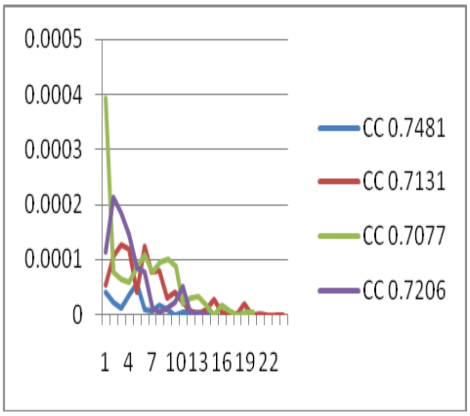
No	Number of Exon	CC		
		State 20	State 40	State 100
1	2	0.7911	0.7989	0.7885
2	< 5	0.7700	0.7768	0.7912
3	> 5	0.7481	0.7306	0.7727

Using back propagation neural network, as shown in Table 5, the CC of the proposed model can be predicted. Accuracy of neural network achieve mean of error maximum 0.1156 and minimum 0.0050. In the third scenario, the relation between log likelihood value and number of iteration needed to converge it into model structure accurate. This architecture still needs improvement to give good accuracy.

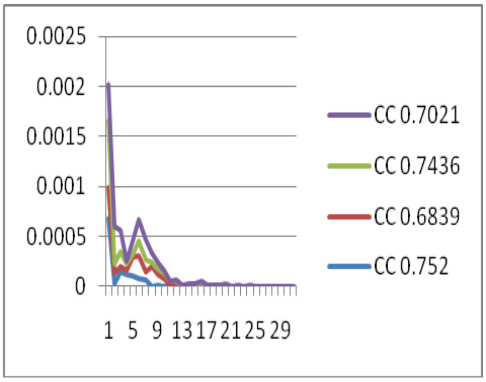
Table 5.Accuracy of Back Propagation Neural Network to Predict CC of Proposed Model

No	model	Mean of error
1	20	0.1156
2	30	0.0415
3	50	0.0050
4	100	0.0069
5	200	0.0116

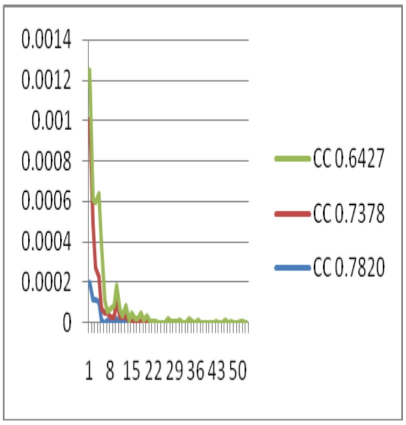
In the third scenario, the relation of log likelihood value and the number of iteration needed to convergent to model structure accuracy is identified. Figure 4 shows that from experiment it is concluded that fluctuation on likelihood value and longer iteration time will give the smaller CC value.



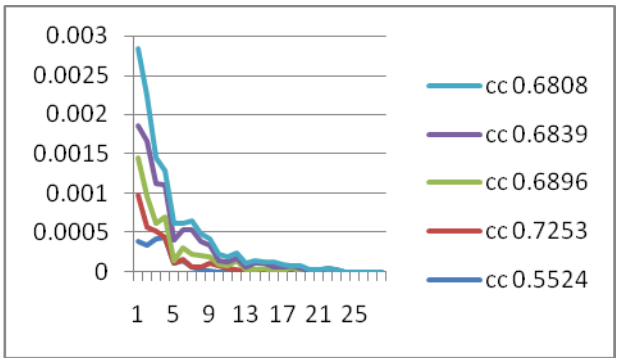
a. State 20



b. State 50



c. State 100



d. State 150

Fig 4. Log Likelihood on Training Phase of HMM Development

In the rest of this section, the advantage of pseudo transition is shown. In the Table 6, it is shown that pseudo transition can accelerate convergence of log likelihood of Hidden Markov model (i.e., iteration time faster).

Table 6.Advantage of Pseudo Transition (Pt) On Training Phase

No	Model	Convergent after	
		Iteration Without pt/ With pt	Time Without pt/ With pt
1	20	21 /20	552.1875 /271.4063
2	30	24/28	1.3664e+003 /657.7813
3	50	23/26	3.53118e+003 /1.2374e+003

IV. Conclusion

The structure of HMM for exon controlling consists of three areas, first exon, introns and other exons. The proposed model can be optimized by choosing appropriate number of state that populate in each model components. The back propagation neural network is proposed to predict the accuracy of specified composition of each component to prune the search space on finding the appropriate composition. Accuracy of model can be identified by log likelihood and spending time to achieve convergence on training process.

V. ACKNOWLEDGMENT

The authors wish to thank Trisakti University Research Institute for all the valuable suggestions. This work was supported in part by a fundamental research grant from Directorate of Higher Education (DIKTI).

REFERENCES

[1] Mathe C. Sagot, M.F.,Schiex T.,Rouze P., "Survey and Summary Current methods of gene prediction, their strengths and weaknesses", Nucleid Acid Research, Vol 30, No 19, 2002.

[2] Krogh A., "A Hidden Markov Model that finds genes in E. coli DNA ", Nucleic Acids Research, v. 22 (1994) p. 4768-4778.

[3] Lukashin A.V. and Borodovsky M., "Genemark.hmm: new solution for gene finding", Nucleic Acids Research, 1998, Vol. 26, No. 4, p 1107 – 1115.

[4] Pedersen J.S. and Hein J., "Gene finding with hidden Markov Model of genome Structure and Evolution", Bioinformatics, 2002.

[5] Siepel A., and Haussler D., "Combining phylogenetic and hidden Markov models in biosequence analysis", J Comput Biol. 2004;11(2-3):413-28.

[6] Krogh A., "Two methods for improving performance of an HMM and their application for gene finding", Proc Fifth Conf. Intelligent Systems for Molecular Biology, p 179-186, 1997

[7] Simon L.C., and Lior P., HMM sampling and applications to gene finding and alternative splicing, Bioinformatics, Vol 2 Suppl 2 2003, DOI: 10.1093/bioinformatics/btg 1057, 2003.

[8] Agoes S., Pakpahan A., Solihah B., "Performance of Hidden Markov Model Structure on DNA Coding Sequence of plasmodium falciparum", ATST Vol 01 Issue 05, 2011.

[9] Nicorici D., Astola J., Tobus I., "Computational identification of exons in DNA with a Hidden Markov Model" Tampere International Center for signal Processing , 2003

[10] Rabiner L.R., "A Tutorial n Hidden Markov Models and Selected Applications in Speech Recognition, Proceeding of The IEEE, Vol 77 No 2, February 1989.

AUTHOR'S PROFILE

Binti Solihah, lecturer of Informatics department on Faculty of Industrial Technology, Trisakti University. Focus research on bioinformatics and image processing.

Suhartati Agoes lecturer of Electrical engineering on Faculty of Industrial Technology, Trisakti University. Obtain Doctoral degree on telecommunication engineering of Indonesia University in 2008. She has 10 publication on peer reviewed international journal and proceedings. She is a researcher on genomic signal processing main in digital signal processing.

Alfred Pakpahan lecture of Biology Department of Faculty of Dentistry, Trisakti University, Campus B, Grogol, Indonesia E-mail: alfred@trisakti.ac.id., focus research on Dentist and Biology.