Folia Medica Indonesiana Vol. 46 No. 1 January – March 2010 : 24-29

THE IMPLEMENTATION OF HIDDEN MARKOV MODEL FOR EXON PREDICTION ON DNA-GENE *Plasmodium falciparum*

Suhartati Agoes[1], Dadang Gunawan[2], Sar Sardy[3]
[1]Department of Electrical Engineering, Faculty of Industrial Engineering, Trisakti University
[2,3]Department of Electrical Engineering, University of Indonesia

ABSTRACT

The Hidden Markov Model (HMM) implementation for exon prediction on DNA-gene Plasmodium falciparum *has various models based on exon region's structure in coding sequence (CDS). The increment of state number was done randomly up to 100, by using a backward-forward HMM with some transition and emission components on each state. Based on the gene basic structure on CDS, it is developed two expansion models for increasing the state number. The Viterbi algorithm is used for training process, and both Viterbi and Baum-Welch algorithms are used respectively for testing process. The correlation coefficient (CC) is used as a performance indicator for all expansion models. The result of simulation for the basic structure at state 9, shows that the best average CC value is 0.73 for Viterbi algorithm, and is 0.72 for Baum-Welch algorithm. In the expansion models, it is found that the best average CC value is 0.78 from the first expansion model for both algorithms at stage 100. The average processing time for training is faster at stage 20 and 30, but almost 15~20 times longer at stage 100, while the average processing time for testing by using Baum-Welch algorithm is twice slower than the Viterbi algorithm.*

Keywords**:** *HMM structure, coding sequence, correlation coefficient*

Correspondence: Suhartati Agoes, Department of Electrical Engineering, Faculty of Industrial Engineering, Trisakti University, Kampus A Usakti, Jl. Kyai Tapa Grogol, phone 62-21-5663232 ext. 429, fax 62-21-5605841, e-mail: suhartati_agoes@yahoo.com

INTRODUCTION

Despite more than a century of efforts to eradicate or control malaria, the disease remains a major and growing threat to the public health and economic development of countries in the tropical and subtropical regions of the world. Approximately 40% of the world's population lives in areas where malaria is transmitted. There are an estimated 300~500 million cases and up to 2.7 million deaths from malaria each year.

Human malaria is caused by infection with intracellular parasites of the genus Plasmodium that are transmitted by Anopheles mosquitoes. From four species of Plasmodium that infect humans, *Plasmodium falciparum* is the most lethal (Gardner et al 2002).

Genome *Plasmodium falciparum* belongs to the genome eukaryotic and has a long DNA genome which consists of some exons and introns, where both are alternately located. In coding sequence (CDS) of the DNA, -after splicing process-, it will be found some exon regions for producing the protein (Figure 1) (Vaisman 1989; Samatova 2003; Edwards 2005). Therefore, exon prediction on DNA-gene is very important to find the protein exactly.

The structure model of HMM implementation for exon prediction on DNA-gene based on exon region's structure in CDS has start gene (codon ATG), exon, intron and stop gene (codon: TAA, or TGA, or TAG). The model was designed based on detailed description of HMM system in gene finding (Rabiner 1989; Henderson, Salzberg, Fasman 1997; Krogh, 1998). This experiment proposes a new structure expansion model for exon prediction on DNA-genes *Plasmodium falciparum* based on the HMM system (Nicorici, Astola, Tobus 2003; Edwards 2005).

The significant different of the model is at least have two exon regions used for exon prediction like in CDS. Usually the 5' (reading start) boundary of introns in most eukaryotes contains the dinucleotide GT, and the 3' (reading stop) boundary contains the dinucleotide AG. Expansion of the model is also examined by increasing the state number randomly on exon and intron regions.

A genome sequence of *Plasmodium falciparum* using 3D7 clone has 23-megabase nuclear genomes consist of

14 chromosomes, encodes about 5,300 genes, and it has the most (A+T)-rich genome sequenced to date (Gardner et al 2002). The data for this simulation has 152 DNA sequences of genes from genome *Plasmodium falciparum* at chromosome 1 (Locus: NC_004325), chromosome 2 (Locus: NC_000910), chromosome 3 (Locus: NC_000521), chromosome 4 (Locus: NC_004318), chromosome 5 (Locus: NC_004326), and chromosome 9 (Locus: NC_004330), in Genbank format with searching to: http://www.ncbi.nlm.nih.gov/entrez or www.ncbi.nlm.nih. gov/Web/Search/index.html (Anastassiou 2001).

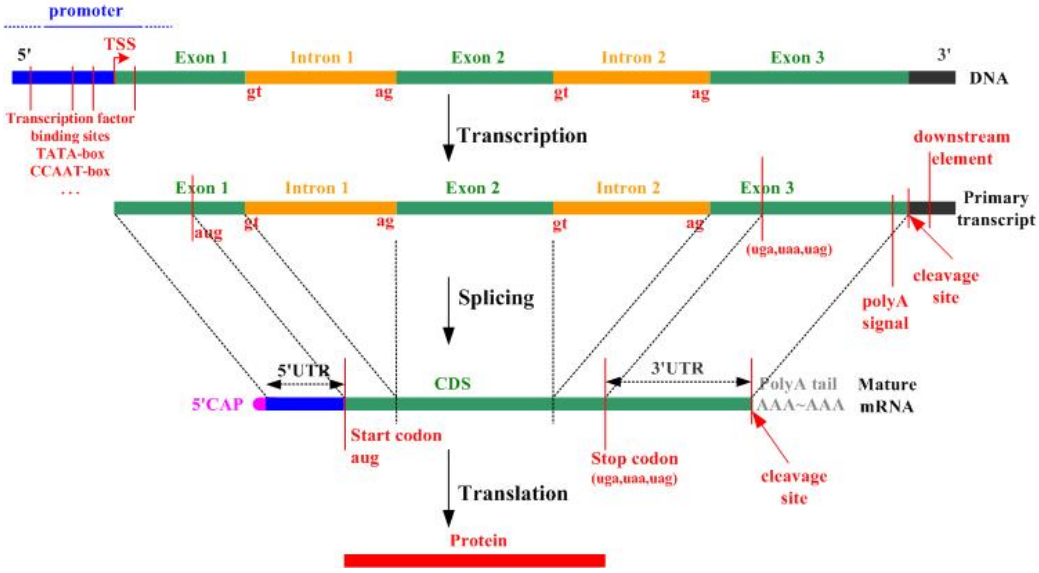

Figure 1. Eukaryotic gene structures

Genbank format describes the coding sequences (CDS) and original DNA sequences of genes *Plasmodium falciparum*. In this experiment, CDS of genes contains at least has two exon regions and maximum 10 exon regions. The minimum length sequence of genes is 684 base pairs (bp) and maximum length is 10095 bp.

Figure 2. The basic structure model, based on CDS

Hidden Markov Model (HMM) provides a good probability method for discrete sequences data like DNA-gene sequences (alphabet of four letters: A, C, G and T). For the exon prediction such in *Plasmodium falciparum*, at least two exon regions are needed in CDS.

The basic structure model based on these exon regions was designed (Figure 2), where the minimum state number of the model should have 5 states. The further state numbers was found by separating start gene from the first exon and stop gene from the last exon. Furthermore, the intron regions of the dinucleotide GT will be separated to become G and T states, dinucleotide AG to become A and G states. The increasing of state number in the model can randomly be used on each exon region in CDS, including intron regions in its location between two exon regions. The inputs are DNA sequences and the set of state number in the sequences depends on the model. Based on HMM method, the Viterbi algorithm is used for the training and it needs the transition and emission state values for the process. The value of emission state can be found with the distribution number of DNA nucleotide in each state. The result of the HMM training process is the estimated transition state and emission state values, and their values then to be used for testing process with Viterbi and Baum-Welch algorithms. The results of testing process are the estimated states of the model (Rabiner 1989). The correlation coefficient (CC) is used as a performance indicator parameter, which calculated by the ratio of the estimated states output and the original states input of the model (Samatova 2003), as the equation below:

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

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

Therefore, CC values will indicates the number of exon and intron states which have the matched location in output and input sequence.

MATERIALS AND METHODS

The basic structure model

The implementation of HMM's structure based on CDS can be explained as the basic structure of the models. The separation of start gene from first exon and stop gene from the last exon, causes the initial state number of the model becomes 5 (Figure 3). Furthermore, the separation of dinucleotide GT and dinucleotide AG from the intron region, causes the state number of the model becomes 7 (Figure 4).

The next separation of dinucleotide GT into G and T states and dinucleotide AG into A and G states, causes the state number of the model becomes 9 (Figure 5). The emission values of the components model can be considered as a matrix, where its columns indicate DNA nucleotide and its rows indicate state number of the designed model. The transition state values of the models can also be shown in the figure.

The emission matrix of the model with state number 5 is:

e = [0.3333 0 0.3333 0.3333 ;
0.4342 0.1094 0.1507 0.3056 ;
0.4210 0.0599 0.0703 0.4487 ;
0.4215 0.1173 0.1622 0.2990 ;
0.5592 0 0.1075 0.3333]

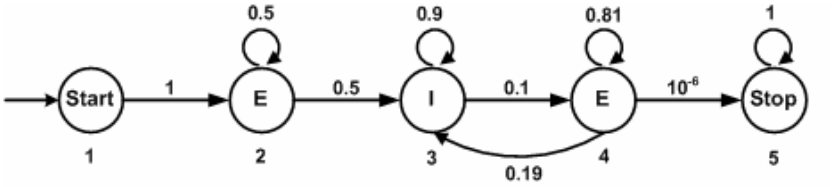

Figure 3. The structure model with state number 5

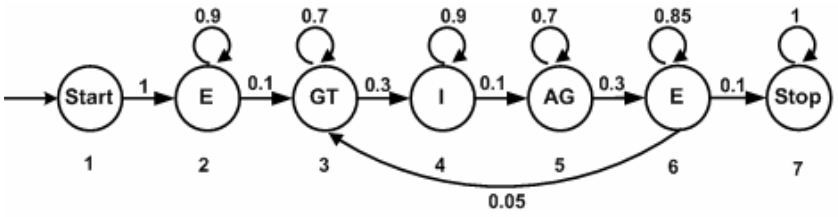

Figure 4. The structure model with state number 7

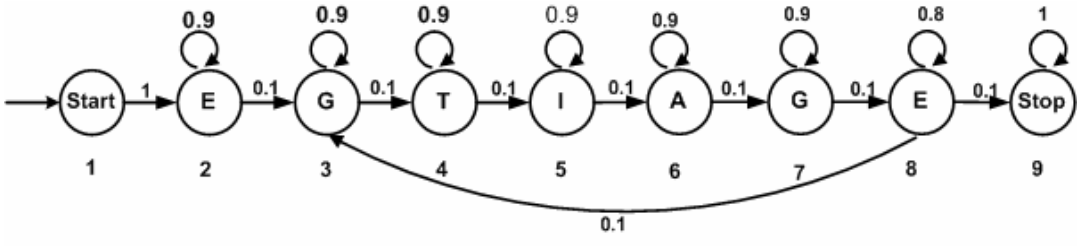

Figure 5. The structure model with state number 9

The emission matrix of the model with state number 7 is:

```
e = [0.3333    0       0.3333  0.3333  ;
     0.4342    0.1094  0.1507  0.3056  ;
     0         0       0.5000  0.5000  ;
     0.4210    0.0599  0.0703  0.4487  ;
     0.5000    0       0.5000  0       ;
     0.4215    0.1173  0.1622  0.2990  ;
     0.5592    0       0.1075  0.3333  ]
```

The emission matrix of the model with state number 9 is:

```
e = [0.3333    0       0.3333  0.3333  ;
     0.4342    0.1094  0.1507  0.3056  ;
     0         0       1.0000  0       ;
     0         0       0       1.0000  ;
     0.4210    0.0599  0.0703  0.4487  ;
     1.0000    0       0       0       ;
     0         0       1.0000  0       ;
     0.4215    0.1173  0.1622  0.2990  ;
     0.5592    0       0.1075  0.3333  ]
```

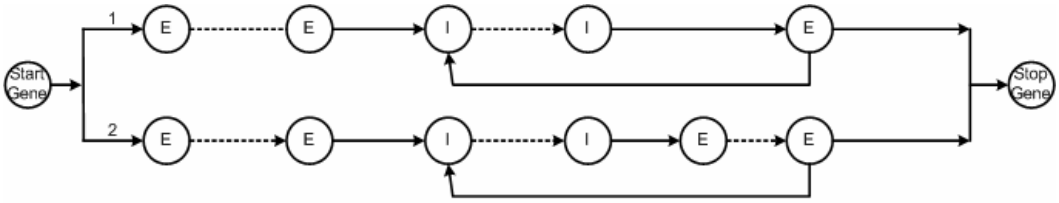

Figure 6. The general structures of expansion model

The expansion model

In order to expand the basic structure model, then two expansion models for increasing state number are randomly created. The experiment has selected the state numbers as 20, 30, 50 and 100, and the general structure of expansion model is depicted (Figure 6). The first expansion model for increasing state number is on first exon region after it is separated from start gene and in intron region, the last exon region still has state number 2 (path1). The second expansion for increasing state number is at the all exon and intron regions (path2). The first state in all structure models that has the minimum transition state value is given by 0, and the last state that has the maximum transition state value is given by 1. The emission values have been found from the DNA nucleotide's distribution on each state. The CC value calculation is done by using the above equation, by using the assumptions that the exon is positive and the intron is negative. The above experiments were done by the simulation, where the used the program is written in Matlab 7.0, with Bio-informatics' toolbox to generate DNA sequences in Genbank format and, where it also has the functions of HMM training and HMM testing.

RESULTS

The basic structure model based on CDS, the simulation results has found the average of CC values shown in Table 1 and its graphic (Figure 7). Furthermore, simulation result of expansion model has found the average of CC values by using Viterbi and B-Welch algorithms shown in Table 2 and Table 3, and its graphics (Figure 8 and Figure 9). The average processing time for training and testing are shown in Table 4.

Table 1. The average of CC values for the models based on CDS

	State number		
HMM Testing	5	7	9
CC of Viterbi	0.7061	0.6870	**0.7289**
CC of B-Welch	0.6759	0.6905	**0.7166**

Table 2. The average of CC values by using the Viterbi algorithm for expansion model

Expansion	State number			
Model	20	30	50	100
1	0.7481	0.7651	0.7727	**0.7820**
2	0.7330	0.7469	0.7671	**0.7378**

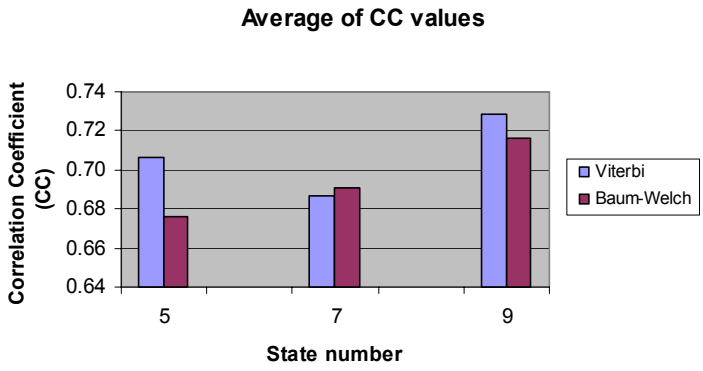

Figure 7. The graph of average CC values, for the models based on CDS

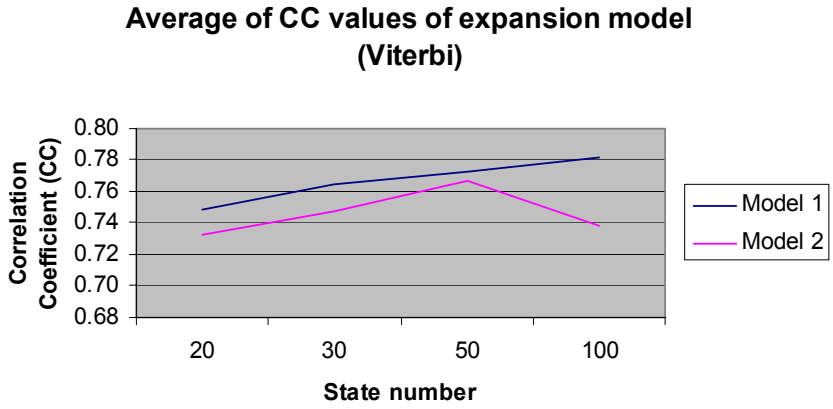

Figure 8. The graph of the average of CC values, by using the Viterbi algorithm for expansion model

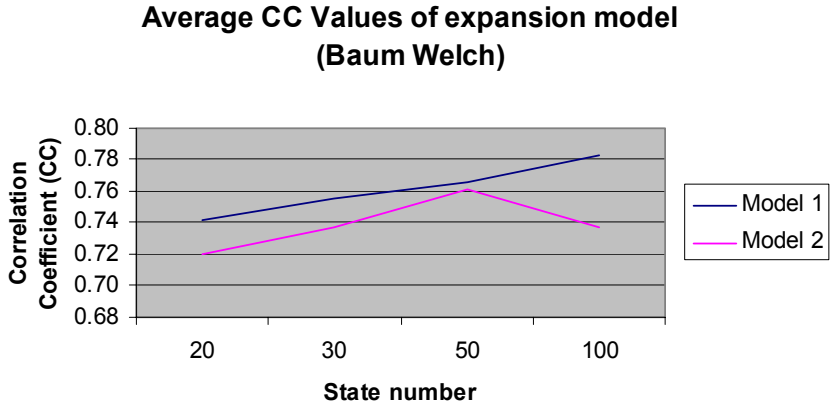

Figure 9. The graph of the average of CC values, by using the B-Welch algorithm for expansion model

Table 3. The average of CC values by using the B-Welch algorithm for expansion model

Expansion Model	State number			
	20	30	50	100
1	0.7414	0.7551	0.7661	**0.7827**
2	0.7195	0.7367	0.7611	**0.7375**

DISCUSSION

All of these works is based on basic structure model in CDS at nine states; due to this nine states give the highest CC values. In all the models have the (A+T)-rich genome sequences of genes *Plasmodium falciparum* showed in their matrix values for the basic structure model shows above. In order to match with Markov's chain in eukaryotic gene structure for connecting the exon and intron then the forward and backward are used in this HMM method. Therefore, the expansion model has been done by increasing the state number on exon and intron region.

At this simulation the increasing state numbers are chosen randomly i.e., 20, 30, 50 and 100. The experiments were done by Viterbi algorithm for training process and the both Viterbi and Baum-Welch algorithms for testing process. Probability of the sequence in Viterbi algorithm has the single path with the highest probability and the other hand Baum-Welch algorithm has the sum of the probabilities over all paths.

CONCLUSIONS

A new model of HMM structures has been proposed for exon prediction on DNA of genes *Plasmodium falciparum*. The simulation results show, that the model based on basic structure on exon regions in CDS for nine states provide the highest average of CC.

The average of CC values is 0.7289 by using the Viterbi algorithm and 0.7166 by using the Baum-Welch algorithm. The expansion model based on basic structure of the exon in CDS has shown that the second expansion model has the CC value higher than the first models. The higher CC value indicates that the second expansion model with state number 100 has the highest CC values = 0.78, by using the Viterbi and Baum-Welch algorithms. The average processing time for training is faster at state number 20 and 30, but almost 15 ~ 20 times slower at state number 100. Finally, the average processing time for testing by using the Baum-Welch algorithm is twice slower than the Viterbi algorithm.

ACKNOWLEDGEMENTS

The authors would like to thank dr. Herawati Sudoyo, PhD, dr. Helena MS, and Hidayat Trimarsanto B.Sc from the Eijkman Institute, Jakarta, for all the discussions and valuable suggestions.

REFERENCES

Anastassiou, D. 2001. Genomic signal processing, IEEE Signal Processing Magazine. 18(4), p 8-20.

Edwards N. 2005. Hidden Markov Models & Gene Finding. Lecture 22. Algorithms in Biosequence Analysis.

Gardner, MJ et al. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419, p 498-511.

Henderson, J, Salzberg, S, Fasman, KH. 1997. Finding Genes in DNA with a Hidden Markov Model. Journal of Computational Biology. 4(2), p127- 142.

Krogh, A. 1998. An Introduction to Hidden Markov Models for Biological Sequences. In Salzberg S. L, Searls D. B and Kasif S (ed). Computational Methods in Molecular Biology. Elsevier, p 45-63.

Nicorici, D, Astola, J, Tobus, I. 2003. Computational identification of exons in DNA with a Hidden Markov Model. Tampere International Center for Signal Processing. Tampere University of Technology. P.O. Box 553, FIN-33101 Tampere, Finland.

Rabiner, LR. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceeding of The IEEE, Vol. 77(2), p257-268.

Samatova, NF. 2003. Computational Gene Finding using HMMs. Computational Biology Institute Oak Ridge National Laboratory.

Vaisman, I. 1998. Bioinformatics and Gene Discovery, Bioinformatics Tutorials. University of North Carolina at Chapel Hill.