Transition State Analysis of HMM for DNA Exon Controlling Using Bioinformatic Simulation

Suhartati Agoes[A], Alfred Pakpahan[B], Binti Solihah[A]
[A]Faculty of Industrial Technology Trisakti University, Jakarta, 11440, Indonesia
[B]Faculty of Dentistry Trisakti University, Jakarta 11440, Indonesia
Emails: sagoes@trisakti.ac.id; alfred@gmail.com; binti76@yahoo.com

Abstract

This paper describes the analysis of transition state value of HMM for DNA exon controlling using Bioinformatic simulation. Exon region in DNA is called a coding sequence (CDS) of genes in many regions at least two regions of exon. HMM model is generate using start and stop gene 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. HMM properties and parameters such as Markov chain, transition state, emission state, HMM training and HMM testing is used to identify original exon region with estimated exon. The performance of estimation result shown by Correlation Coefficient (CC). Random values of transition state used for HMM train makes many differences in the CC of the model. Furthermore, the analysis of transition state values is very important to finding optimum of CC. Several models with the parameters of HMM were simulated, trained and tested for the implementation of number of states with HMM method. The simulation result predicted that the CC value is very much influenced by the value of transition state and improved the number of states on the model makes increasing of CC.

Keywords: HMM, Transition state, Exon, Correlation Coefficient (CC).

1. Introduction

To controlling exon in DNA sequences of genes *Plasmodium falciparum* has the region of exon based on coding sequence (CDS) in database from Genbank, including the region of intron has been know on CDS. In principle, start codon, exon, intron and stop codon are becomes the state to use in the structure of models. Each state of the models has many bases DNA actually start and stop codons [Nicorici, *et al.* 2003]. Transition states of the models are depends on state number and its bases on each state and transition states values minimum is 0 and maximum is 1. Trial and error in simulation using transition states values are very important to have the optimal performance of the models. Additional states in the structure of the models in regions exon and intron of DNA sequences.

Plasmodium falciparum genome belongs to eukaryotic genome 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 alternately 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 [Samatova, 2003; Anantharaman, 2004]. 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.

HMM based finding usually gives more accurate results compared to other methods [Rabiner, 1989; Henderson, et al.1997]. In this paper, HMM's models was chosen based on HMM

method for several models with its number of states. One of the other performance parameters of HMM's models is shown by its Correlation Coefficient (CC) value for each models.

2. Materials and Methods

2.1. Materials

The materials have 152 DNA sequences of genes from genome *Plasmodium falciparum* in GenBank format with searching to: http://www.ncbi.nlm.nih.gov/entrez or www.ncbi.nlm.nih.gov/Web/Search/index.html [Alphey, 1997; Anastassiou, 2001]. Genbank format describes CDS and original DNA sequences of genes *Plasmodium falciparum*. In CDS contains at least two regions and maximum 10 regions of exons. Minimum length of sequences for this simulation is 684 base pairs (bp) and maximum length is 10095 bp. Matlab 7.0 Mathworks, Massachussets, USA was used for simulations and its hardware PC IBM Standard has specification: Processor Intel(R)Pentium(R) 4 CPU 2.8 GHz; Memory 1.99 GB of RAM; Harddisk 40 GB; Operating System Microsoft Windows XP, 2002 version.

2.2. Methods

HMM process consist of input, HMM training, HMM testing and output. Input for HMM training consists of DNA sequences, information of exon position of each sequence, transition matrix and emission distribution matrix. DNA sequence data and information of exon position of each sequence are used to set the state number for each base depend on model will be generated. Than transition matrix is defined and emission distribution matrix of each base state is calculated. Furthermore, HMM training has the both algorithms are Viterbi and Baum-Welch and needs the transition states and emission states for the process. The result of HMM training is the estimated transition states and emission states. The estimated transition and emission states are used for HMM testing process has both of the algorithms above and its result of HMM testing is the estimated states of the model. The performance of the model are the value of CC, which calculated by comparing the estimated state from HMM testing result with the original state of the input sequences [Vaisman, 1998; Samatova, 2003], the formula as equations (1).

$$CC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

where:

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

Several models in this simulation 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 general structure of the model like in figure 1 below.

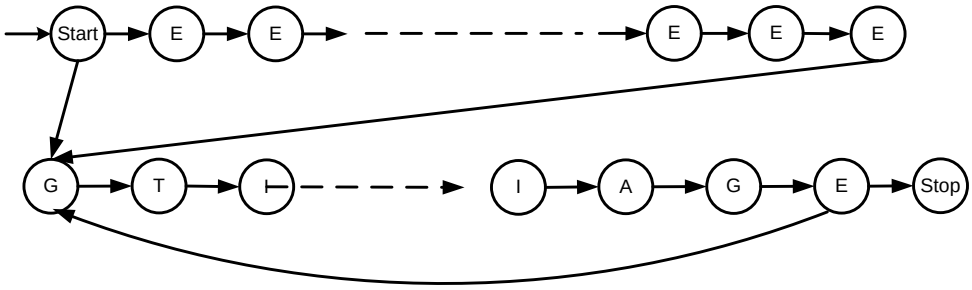

Figure1. General structures of HMM

The HMM parameters were set for the HMM training and testing method and its algorithm [Rabiner, 1989; Anastassiou, 2001].
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, there's toolbox Bio-informatics to generate DNA sequences in GenBank format and has functions of HMM training and HMM testing.

3. Result

Several models have been simulated from HMM implementation for controlling exon region in DNA sequences of genes *Plasmodium falciparum*.
The models were formed with a number of state randomly until the model is formed into 20 states, 30 states and 50 states following the general stucture of HMM as seen in figure 1. Each model simulation performed three times by using the values of the different state transition, can even use the same exon and intron structure with a number of different states.
Random value of transition state can be the analysis to find CC values optimum but the emission of each state has the distribution constant value of bases DNA sequences depends on the models. To calculate CC is by using the equation (1) and the assumption of exon is positive and intron is negative.
The simulation results with random values of transitions states for each model like as Table 1 below.

Table1. CC values for each model performed by number of state

Transition State Values (Performed by 20 states)											CC	
1	2-9	10	11-15	16	17-18	19	20	19-11	19-20	1-11	Vit	B-W
0	0.1	0.1	0.1	0.9	0.1	0.8	1	0.1	0.1	0.1	0.7077	0.6911
0	0.1	0.1	0.1	0.9	0.1	0.85	1	0.05	0.1	0	0.7131	0.6939
0	0.1	0.9	0.1	0.9	0.1	0.85	1	0.05	0.1	0	0.7481	0.7414
Transition State Values (Performed by 30 states)											CC	
1	2-12	13	14-25	26	27-28	29	30	29-14	29-30	1-24	Vit	B-W
0	0.1	0.1	0.1	0.9	0.1	0.8	1	0.1	0.1	0.1	0.6278	0.6179
0	0.1	0.9	0.1	0.9	0.1	0.8	1	0.1	0.1	0	0.7651	0.7551
0	0.1	0.9	0.1	0.9	0.1	0.85	1	0.05	0.1	0	0.7217	0.7189
Transition State Values (Performed by 50 states)											CC	
1	2-22	23	24-45	46	47-48	49	50	49-24	49-50	1-24	Vit	B-W
0	0.1	0.9	0.1	0.9	0.1	0.8	1	0.1	0.1	0	0.7520	0.7494
0	0.1	0.9	0.1	0.9	0.1	0.8	1	0.1	0.1	0.1	0.7436	0.7016
1	2-17	18	19-45	46	47-48	49	50	49-19	49-50	1-19	Vit	B-W
0	0.1	0.9	0.1	0.9	0.1	0.8	1	0.1	0.1	0	0.7727	0.7661

The simulation result of several HMM structure above, the values of CC has been influenced by transition state value and the improved states on the model. The graphic optimum CC values in this simulations result as figure 2; where HMM testing using algorithm Viterbi is better than using Baum-Welch algorithm.

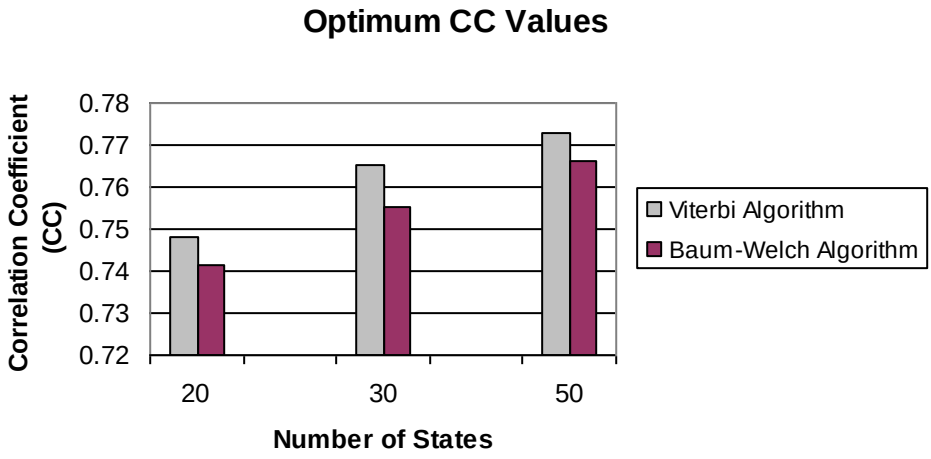

Figure 2. Optimum CC Values from simulation results.

The random value of transition state became the CC values negative, its means that the assumption of exon to be negative and intron to be positive. Meanwhile, emission state can be used to find out the distribution of emission of DNA bases in each state in accordance with HMM structure that has been used.

4. Discussion

The HMM model that has been used is developed from above model which will be tested by adding states in each exon and intron regions. The forward and backtracking directions between states in the used algorithm must be matched with one of the nature of HMM is Markov's chain and eukaryotic gene structure. Algorithm Viterbi was used at the HMM training and algorithm Viterbi and Baum-Welch was used at HMM testing.

Simulation result shows that the highest optimum CC value was for the model with 50 states with HMM testing by using Viterbi algorithm is 0.7727 and by using Baum-Welch algorithm is 0.7661.

Furthermore, improvement states of the model of this study can be used by using a number of state and sequence DNA more than that already simulated because it predicted to produce higher value of CC.

Acknowledgements

The authors would like to thanks to Directorate of Higher Education (DIKTI) for fundamental research grant 2011 and Trisakti University Research Institute for all valuable discussions and suggestions.

References

Alphey, Luke., DNA sequencing, Bios Scientific Publishers Limited, 1997.

Anantharaman, Thomas., Finding Genes in Genomic DNA The GENESCAN System,http://www.biostat.wisc.edu/bmi776, February 2004.

Anastassiou, Dimitris., *Genomic signal processing,* IEEE Signal Processing Magazine, Vol. 18, No.4, July 2001.

Henderson J, Salzberg S, Fasman KH, *Finding gene in DNA with a Hidden Markov Model,* Journal Computational Biology, Vol. 4, Issue 2, pp 127-141, 1997.

Nicorici, Daniel., Jaakko Astola, Ioan Tobus, Computational identification of exons in DNA with a Hidden Markov Model, Tampere International Center for Signal Processing, Tampere University of Technology, 2003

Rabiner, R Lawrence., *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceeding of The IEEE, Vol. 77. No.2, February 1989.

Samatova, F Nagiza., *Computational gene finding using HMMs*, Computational Biology Institute Oak Ridge National Laboratory, 2003.

Vaisman, Iosif., Bioinformatics and Gene Discovery, Bioinformatics Tutorial, University of North Carolina at Chapel Hill, 1998.