UNIVERSA MEDICINA

Juli-September 2007 Vol.26 - No.3

Exon prediction on DNA-genes of Plasmodium falciparum based on coding sequence structure using hidden Markov model

Suhartati Agoes*[a], Dadang Gunawan**, Sardy S**, and Hoedojo***

ABSTRACT

BACKGROUND
A hidden Markov model (HMM) is used for exon prediction on DNA of genes Plasmodium falciparum that has a model structure based on exon region structure in coding sequence (CDS). The objective research was to develop a new structure model to predict exon on DNA-genes of Plasmodium falciparum based on CDS structure using the HMM system.

METHODS
Model design in CDS, between two exon regions can be found one intron region and the model state number is used for its region. Its state number is used by separating start codon from first exon region and stop codon from the last exon region up to 9. The Viterbi algorithm and the backward-forward method for transition as well as emission states are used for training process. Furthermore, Viterbi and Baum-Welch algorithms are used for the testing process. The correlation coefficient (CC) was used as performance indicator, as the ratio of the estimated state in the output and the original state in the input of the model.

RESULTS
The simulation results has shown that the CC values depend on the given of the backward-forward transition state values randomly. The model with state number 9 showed the highest average of CC values of 0.7289 for Viterbi algorithm, and is 0.7166 for Baum-Welch algorithm. However, the lowest average of CC values has been found for the model with state number five. Its values are 0.6735 by using Viterbi algorithm and 0.6661 by using Baum-Welch algorithm.

CONCLUSION
The new structure model based on HMM system was valid to predict exon on DNA-genes of Plasmodium falciparum.

Keywords: *Exon Prediction, DNA-gene, coding sequence, Hidden Markov Model*

* Electrical Engineering Department, Faculty of Industrial Technology, Trisakti University
** Electrical Engineering Department, Faculty of Engineering, Indonesia University
*** Department of Parasitology, Medical Faculty, Trisakti University

Korespondensi
[a]Ir. Suhartati Agoes, MT
Electrical Engineering Department, Faculty of Industrial Technology, Trisakti University
Jl. Kyai Tapa No.1, Grogol
Jakarta 11440
Telp. 08164836613
Email:
suhartati_agoes@yahoo.com

Universa Medicina 2007; 26: 129-36.

Prediksi ekson DNA-gen Plasmodium falciparum berdasarkan struktur *coding sequence* dengan menggunakan model *hidden* Markov

Suhartati Agoes*[a], Dadang Gunawan, Sardy S**, dan Hoedojo*****

* Jurusan Teknik Elektro
Fakultas Teknologi Industri
Universitas Trisakti
** Jurusan Teknik Elektro,
Fakultas Teknik
Universitas Indonesia
*** Bagian Parasitologi
Fakultas Kedokteran
Universitas Trisakti

Korespondensi
[a]Ir. Suhartati Agoes, MT
Jurusan Teknik Elektro
Fakultas Teknologi Industri
Universitas Trisakti
Jl. Kyai Tapa No.1, Grogol
Jakarta 11440
Telp. 08164836613
Email:
suhartati_agoes@yahoo.com

Universa Medicina 2007; 26: 129-36.

ABSTRAK

LATAR BELAKANG
Sebuah *hidden* Markov model (HMM) yang digunakan untuk memprediksi ekson gen DNA Palsmodium falciparum mempunyai struktur model berdasarkan struktur gen DNA pada *coding sequence* (CDS). Penelitian ini bertujuan untuk mengembangkan sebuah model stuktur baru untuk prediksi ekson gen DNA Plasmodium falciparum berdasarkan struktur CDS dengan menggunakan sistem HMM.

METODE
Rancangan model pada CDS, di antara dua lokasi ekson dapat diketahui sebuah lokasi intron dan jumlah *state* model dapat dilakukan pada lokasi tersebut. Jumlah *state* dilakukan dengan memisahkan *codon start* dari ekson pertama dan *codon stop* dari ekson terakhir hingga mencapai 9. Algoritma Viterbi dan metode *backward-forward* transisi *state* serta emisi *state* digunakan untuk proses *training*. Sedangkan untuk proses testing menggunakan algoritma Viterbi dan Baum-Welch. Sebagai kinerja model digunakan *Correlation Coefficient* (CC) yang didapat dari perbandingan *state* estimasi pada *output* dan *state* asli pada *input* model.

HASIL
Hasil simulasi menunjukkan bahwa nilai CC tergantung pada pemberian nilai acak state transisi *backward-forward*. Model dengan jumlah *state* 9, menunjukkan nilai CC rata-rata tertinggi adalah 0,7289 untuk algoritma Viterbi dan 0,7166 untuk algoritma Baum-Welch.

KESIMPULAN
Struktur model berdasarkan sistem HMM sahih untuk memprediksi ekson gen DNA Plamodium falciparum

Kata kunci : Prediksi ekson, gen-DNA, *coding sequence*, model *hidden* Markov

INTRODUCTION

In coding sequence (CDS), the exon prediction on DNA-gene is very important to find the protein structure. Genome *P. falciparum* belongs to genome eukaryotic and has a long DNA genome, consisting of several exons and introns alternately located. After the splicing process in DNA sequence, some regions of exon in CDS will be found to produce the protein.[(1)]

The hidden Markov model (HMM) structure is one of the model used to predict the exon in desoxyribonucleai acid (DNA) which is based on the exon region structure in CDS. The HMM structure of the model: start codon (Methyanine/ATG), exon, intron and stop codon (TAA, or TGA, or TAG) has been demonstrated by Rabiner, Henderson and Krogh.[2-4] This research developed a new model of structures for exon prediction on DNA of genes *Plasmodium falciparum was* based on the work of Nicorici and Anantharaman.[5,6] The significant difference of this model is it at least has two exon regions used for exon prediction like in CDS. Usually the 5' boundary of introns in most eukaryotes contains the dinucleotide Guanine-Thymine (GT), and the 3' boundary contains the dinucleotide Adenine-Guanine (AG). Furthermore, in intron regions of the dinucleotide GT were separated to be nucleotide Guanine (G) and Thymine (T), and dinucleotide AG to be nucleotide A and G.

METHOD

Research design

The HMM framework in this experiment used Viterbi algorithm for the training process and both Viterbi and Baum-Welch algorithms for the testing process. The same sequence has been done for the training and testing process. The program is written in Matlab 7.0, and Bio-informatics' toolbox to generate DNA sequences in Genbank format and has the functions of HMM training and testing.

The first experiment for the model based on CDS has been done with separated start codon (codon: ATG) from first exon and stop codon (codon: TAA, or TAG, or TGA) from the last exon, its model has the state number of 5 (Figure 1). Furthermore, the second experiment used separated dinucleotide GT and dinucleotide AG from the intron region, its model has the state number of 7 (Figure 2). The last experiment in the model also used separated dinucleotide GT into G and T states and dinucleotide AG into A and G states; the state number of the model is 9 (Figure 3). The emission values of the models can be performed as the matrix, the columns are the bases numbered on DNA sequences and the rows are the states number of the models designed. The state transition values of the models can be shown also in the figures.

Samples

A genome sequence of *P. falciparum* using 3D7 clone has 23-megabase nuclear genomes consisting of 14 chromosomes, encodes about 5,300 genes, and it has the most (A+T)-rich genome sequenced to date.[8-10] The data set of experiments 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.[11,12] Genbank format describes the CDS and original DNA sequences of genes *P. falciparum*. In CDS of genes contains at least 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.

HMM provides a good probability method for discrete sequences model of data like DNA sequences (alphabet of four letters: A, C, G and T). Prediction of exon in DNA of genes *P. falciparum* is based on exon region in CDS and it has at least two exon regions. The structure of the model based on CDS structure as shown in Figure 4, which separated is start codon from the first exon and separated stop codon from the last exon.

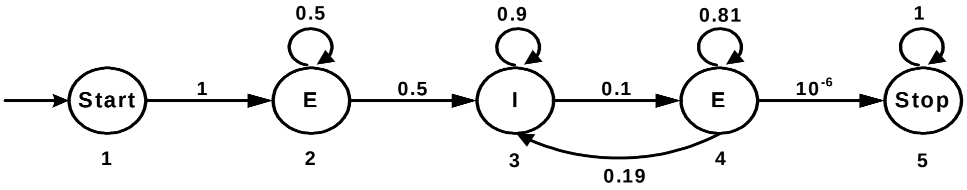


Figure 1. The HMM structure with 5 states

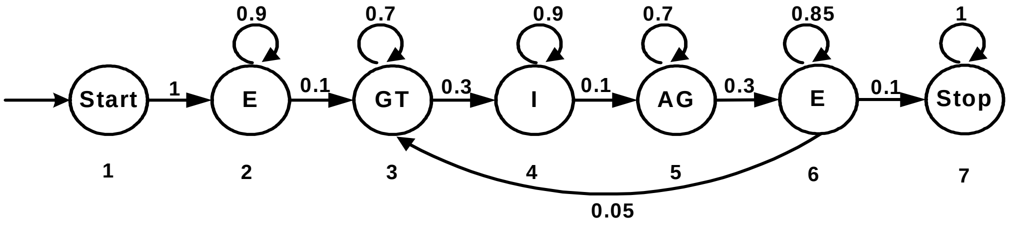


Figure 2. The HMM structure with 7 states

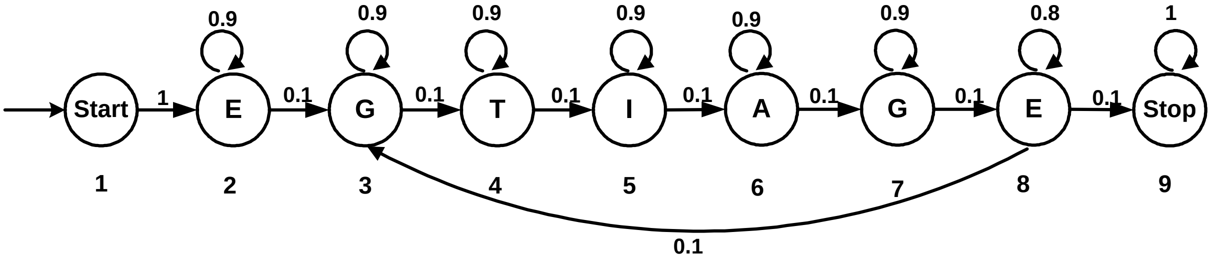


Figure 3. The HMM structure with 9 states

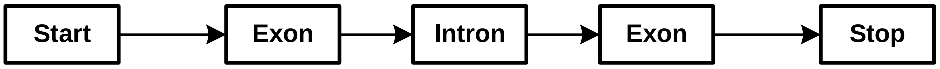


Figure 4. The HMM structure based on CDS

In the inputs are DNA sequences and set the state number of the sequences depends on the structure model. Based on HMM method, the Viterbi algorithm used for the training and needs the transition state values and emission values for the process. The value of transition state is random and the value of emission can be found with the distribution number of DNA nucleotide in each states. The result of HMM training is the estimated transition state and emission state values. The estimated transition and emission values were used for HMM testing process with Viterbi and Baum-Welch algorithms. The results of HMM testing process are the estimated states of the model.(2) The parameter of performance indicator used the correlation coefficient (CC), as a ratio of the estimated states from the output and the original states from the input of the model[1,7] and the formula is the following:

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

Where: TP= True Positive; FP= False Positive; TN= True Negative; FN= False Negative

Statistical analysis

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 ]
```

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 ]
```

The transition state values are randomly, but the first state on the model has the minimum transition state value of 0 and the last state has the maximum transition state value of 1. On the other hand, the emission state values were found from the distribution number of DNA nucleotide in each state depending on the models. Calculation of the CC is by using the above equation with the assumptions that the exon is positive and intron is negative.

RESULTS

Based on exon regions structure in CDS, the simulation results for the model with state number 5 above and transition state values randomly has found the average of CC value in Table 1. Table 2 and Table 3 shows the average of CC values of the model with state number 7 and state number 9. The transition state value in the first state of all structures on the models is 0 and the last state is 1. The other transition state values except in Table 2 are 0.1 and the highest average of CC values for this experiments are shown in Figure 5.

The model based on exon regions structure in CDS for state number nine resulted in the highest average of CC. Its values are 0.7289 by using Viterbi algorithm and 0.7166 by using Baum-Welch algorithm. However, the lowest average of CC values has been found for the model with state number five. Its values are 0.6735 by using Viterbi algorithm and 0.6661 by using Baum-Welch algorithm.

DISCUSSION

All these were based on the model of exon regions structure in CDS, at the state number nine resulted in the highest average of CC values. It is also shown that the emission state value of the models have the (A+T)-rich genome sequences of genes *Plasmodium falciparum*.

Table 1. The average of CC values for the model with state number 5

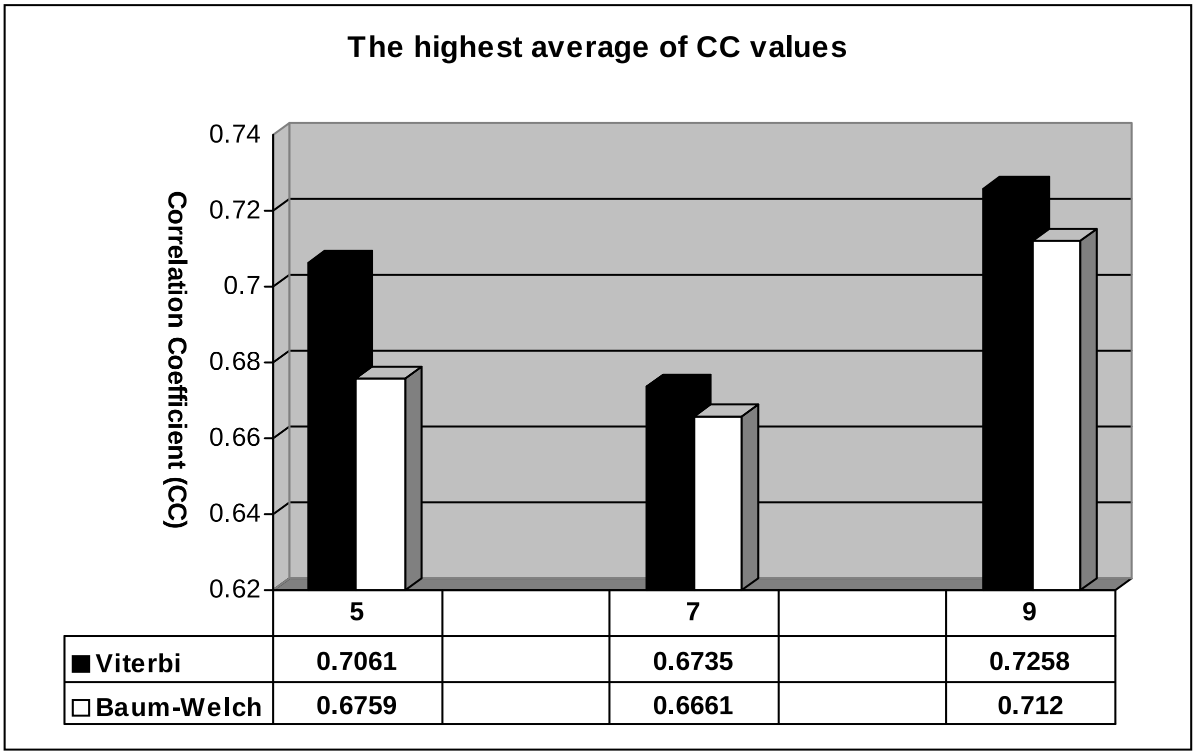
Exps	**Transition state values**					**Corr.Coef.(CC)**	
	tr$_{11}$	**tr$_{22}$**	**tr$_{33}$**	**tr$_{44}$**	**tr$_{55}$**	**Viterbi**	**B-Welch**
1	0	0.5	0.9	0.80	1	0.7054	0.6750
2	0	0.5	0.9	0.81	1	0.7061	0.6759
3	0	0.5	0.9	0.82	1	0.7054	0.6750
4	0	0.5	0.9	0.83	1	0.7054	0.6750
5	0	0.5	0.9	0.84	1	0.7045	0.6741
6	0	0.5	0.9	0.85	1	0.7054	0.6750
7	0	0.5	0.9	0.86	1	0.7062	0.6744
8	0	0.5	0.9	0.87	1	0.0012	0.0015
9	0	0.5	0.9	0.88	1	0.0011	0.0013
10	0	0.5	0.9	0.89	1	0.0011	0.0013

Table 2. The average of CC values for the model with state number 7

Exps	**Transition state values**							**Corr.Coef.(CC)**	
	tr$_{11}$	**tr$_{22}$**	**tr$_{33}$**	**tr$_{44}$**	**tr$_{55}$**	**tr$_{66}$**	**tr$_{77}$**	**Viterbi**	**B-Welch**
1	0	0.4	0.2	0.8	0.10	0.7	1	0.6713	0.6643
2	0	0.4	0.2	0.8	0.20	0.7	1	0.6713	0.6643
3	0	0.4	0.2	0.8	0.30	0.7	1	0.6733	0.6660
4	0	0.4	0.2	0.8	0.40	0.7	1	0.6731	0.6660
5	0	0.4	0.2	0.8	0.50	0.7	1	0.6735	0.6661
6	0	0.4	0.2	0.8	0.60	0.7	1	-0.0642	-0.0729
7	0	0.4	0.2	0.8	0.59	0.7	1	-0.0568	-0.0639
8	0	0.4	0.2	0.8	0.58	0.7	1	0.6735	0.6661
9	0	0.4	0.2	0.8	0.56	0.7	1	0.6735	0.6661
10	0	0.4	0.2	0.8	0.55	0.7	1	0.6735	0.6661

Table 3. The average of CC values for the model with state number 9

Exps	**Transition state values**									**Corr.Coef.(CC)**	
	tr$_{11}$	**tr$_{22}$**	**tr$_{33}$**	**tr$_{44}$**	**tr$_{55}$**	**tr$_{66}$**	**tr$_{77}$**	**tr$_{88}$**	**tr$_{99}$**	**Viterbi**	**B-Welch**
1	0	0.3	0.2	0.2	0.80	0.10	0.10	0.70	1	0.7242	0.7124
2	0	0.3	0.1	0.1	0.79	0.10	0.10	0.70	1	0.7242	0.7124
3	0	0.3	0.1	0.1	0.78	0.09	0.09	0.70	1	0.7242	0.7124
4	0	0.3	0.1	0.1	0.78	0.20	0.20	0.70	1	0.7258	0.7120
5	0	0.3	0.1	0.1	0.80	0.20	0.20	0.80	1	0.7258	0.7120
6	0	0.3	0.1	0.1	0.80	0.15	0.15	0.74	1	0.7258	0.7120
7	0	0.3	0.1	0.1	0.80	0.15	0.15	0.80	1	0.7258	0.7120
8	0	0.3	0.1	0.1	0.80	0.15	0.15	0.70	1	0.7258	0.7120
9	0	0.8	0.2	0.2	0.80	0.10	0.10	0.70	1	0.6787	0.6755
10	0	0.9	0.2	0.2	0.80	0.10	0.10	0.70	1	0.0451	0.0437



	5		7		9
■ Viterbi	0.7061		0.6735		0.7258
□ Baum-Welch	0.6759		0.6661		0.712

Figure 5. The highest average of CC values of the experiments

This can also be showed in their matrix values described above. In order to match with Markov's chain in eukaryotic gene structure for connecting the exon and intron, then the backward-forward method are used in this HMM system. The experiments were done by Viterbi algorithm for the training process and the both Viterbi and Baum-Welch algorithms for the testing process.

A set of 152 sequences is then used as data to train the model and a different HMM is produced for each set of sequences. HMM have been previously used in sequence analysis to produce an HMM that represents a sequence profile (a profile HMM), to analyze sequence composition and patterns, to locate genes by predicting open reading frame or coding sequence, and to produce protein structure predictions. This HMM generates sequences with various combinations of matched, mismatched, insertions and deletions, and gives these a probability, depending on the values of the various parameters in the model.

A total of 424 predicted protein-coding genes and 3 tRNA genes of *Plasmodium falciparum* have been identified on chromosomes 2 and 3, the two chromosomes for which sequencing have been completed.(13,14) Approximately 37% (158 genes) of three genes have a readily identifiable homologue in another species. Such similarity allows a function to be implied, with many of those identified being involved in parasite metabolism. Interestingly, a comparison of the predicted protein sequences with their homologues in other species showed that the majority of *Plasmodium falciparum* proteins have insertion of low complexity sequences, often runs of a single amino acid residue (typically asparagines, lysine or glutamine acid), or tandem arrays of a short peptide repeat sequence. Examples of such regions have been identified previously, and are polymorphic between different parasite isolates. Perhaps it will open up new avenues for drug and vaccine development. Perhaps it will make a genetically modified mosquito that cannot act as vector a reality once processes are in place to deal with ethical, legal and social issues. But any new developments are unlikely to happen before the 2010 deadline set for Roll Back Malaria. New drugs based on knowledge about the genome will probably not be affordable to those countries that need it the most.(15)

CONCLUSION

This research with a new structure model was using for exon prediction on DNA-genes of Plasmodium falciparum based on CDS structure by using the HMM system.

ACKNOWLEDGMENT

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

References

1. Samatova N.F. Computational gene finding using HMMs. UT-Battelle Information Center, 1201 Oak Ridge Turnpike, Suite 100, Oak Ridge, TN 37830. Institute Oak Ridge National Laboratory; 2003.
2. Lawrence R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceeding of The IEEE. 1989; 77: 257-68.
3. Henderson J, Salzberg S, Fasman K.H. Finding genes in DNA with a hidden Markov model. J Comput Biol 1997; 4: 127-42.
4. Krogh A. An introduction to hidden Markov models for biological sequences. In Computational methods in molecular biology. Salzberg SL, Searls DB, Kasif S. editors. Denmark. Center for Biological Sequence Analysis, Technical University of Denmark. 1998. p. 45-63.
5. Nicorici D, Astola J, Tobus I. Computational identification of exons in DNA with a hidden Markov model. Tampere International Center for Signal Processing, Tampere University of Technology, Finland. Available at: http://www.gensips.gatech.edu/processings/contributed/CP2-06.pdf. Accessed May 12, 2005.
6. Anantharaman T. Finding genes in genomic DNA The GENESCAN System. Available at : http://www.biostat.wisc.edu/bmi776. Accessed June 6, 2005.
7. Vaisman I. Bioinformatics and gene discovery. Bioinformatics Tutorial. North Carolina, United State: University of North Carolina at Chapel Hill. 1998.
8. Hall N, Gardner M.J, Hyman RW, Lasonder E, Wilson RJM, Scherf A, et al. Sequence of plasmodium falciparum chromosomes 1, 3-9 and 13. Nature 2002; 419: 527-31.
9. Gardner M.J, Hall N, Hyman RW, Hinterberg K, Mattei D, Wellem TE, et al. Sequence of plasmodium falciparum chromosomes 2, 10, 11 and 14. Nature 2002; 419: 531-4.
10. Hyman RW, Gardner M.J, Hall N, De Bruin D, Scherf A, Day KP, et al. Sequence of plasmodium falciparum chromosome 12. Nature 2002; 419: 534-7.
11. Anastassiou D. Genomic signal processing. IEEE Signal Processing Magazine 2001; 18: 4.
12. Alphey L. DNA sequencing. Manchester UK : Bios Scientific Publishers Limited; 1997.
13. Gardner M.J, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum. Science 1998; 282: 1126-32.
14. Bowman S, Lawson D, Basham D, Brown D, Chillingworth T, Churcher C. M, et al. The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum. Nature 1999; 400: 532-8.
15. Anonymous. Malaria after the genomes. The Lancet 2002; 360: 1107.

UNIVERSA MEDICINA

Juli-September 2007 Vol.26 - No.3

Exon prediction on DNA-genes of Plasmodium falciparum based on coding sequence structure using hidden Markov model

Suhartati Agoes*[a], Dadang Gunawan, Sardy S**, and Hoedojo*****

ABSTRACT

BACKGROUND

A hidden Markov model (HMM) is used for exon prediction on DNA of genes Plasmodium falciparum that has a model structure based on exon region structure in coding sequence (CDS). The objective research was to develop a new structure model to predict exon on DNA-genes of Plasmodium falciparum based on CDS structure using the HMM system.

METHODS

Model design in CDS, between two exon regions can be found one intron region and the model state number is used for its region. Its state number is used by separating start codon from first exon region and stop codon from the last exon region up to 9. The Viterbi algorithm and the backward-forward method for transition as well as emission states are used for training process. Furthermore, Viterbi and Baum-Welch algorithms are used for the testing process. The correlation coefficient (CC) was used as performance indicator, as the ratio of the estimated state in the output and the original state in the input of the model.

RESULTS

The simulation results has shown that the CC values depend on the given of the backward-forward transition state values randomly. The model with state number 9 showed the highest average of CC values of 0.7289 for Viterbi algorithm, and is 0.7166 for Baum-Welch algorithm. However, the lowest average of CC values has been found for the model with state number five. Its values are 0.6735 by using Viterbi algorithm and 0.6661 by using Baum-Welch algorithm.

CONCLUSION

The new structure model based on HMM system was valid to predict exon on DNA-genes of Plasmodium falciparum.

Keywords: *Exon Prediction, DNA-gene, coding sequence, Hidden Markov Model*

* Electrical Engineering Department, Faculty of Industrial Technology, Trisakti University
** Electrical Engineering Department, Faculty of Engineering, Indonesia University
*** Department of Parasitology, Medical Faculty, Trisakti University

Korespondensi
[a]Ir. Suhartati Agoes, MT
Electrical Engineering Department, Faculty of Industrial Technology, Trisakti University
Jl. Kyai Tapa No.1, Grogol
Jakarta 11440
Telp. 08164836613
Email:
suhartati_agoes@yahoo.com

Universa Medicina 2007; 26: 129-36.

Prediksi ekson DNA-gen Plasmodium falciparum berdasarkan struktur *coding sequence* dengan menggunakan model *hidden* Markov

Suhartati Agoes*[a], Dadang Gunawan, Sardy S**, dan Hoedojo*****

ABSTRAK

* Jurusan Teknik Elektro
Fakultas Teknologi Industri
Universitas Trisakti
** Jurusan Teknik Elektro,
Fakultas Teknik
Universitas Indonesia
*** Bagian Parasitologi
Fakultas Kedokteran
Universitas Trisakti

Korespondensi
[a]Ir. Suhartati Agoes, MT
Jurusan Teknik Elektro
Fakultas Teknologi Industri
Universitas Trisakti
Jl. Kyai Tapa No.1, Grogol
Jakarta 11440
Telp. 08164836613
Email:
suhartati_agoes@yahoo.com

Universa Medicina 2007; 26: 129-36.

LATAR BELAKANG
Sebuah *hidden* Markov model (HMM) yang digunakan untuk memprediksi ekson gen DNA Palsmodium falciparum mempunyai struktur model berdasarkan struktur gen DNA pada *coding sequence* (CDS). Penelitian ini bertujuan untuk mengembangkan sebuah model stuktur baru untuk prediksi ekson gen DNA Plasmodium falciparum berdasarkan struktur CDS dengan menggunakan sistem HMM.

METODE
Rancangan model pada CDS, di antara dua lokasi ekson dapat diketahui sebuah lokasi intron dan jumlah *state* model dapat dilakukan pada lokasi tersebut. Jumlah *state* dilakukan dengan memisahkan *codon start* dari ekson pertama dan *codon stop* dari ekson terakhir hingga mencapai 9. Algoritma Viterbi dan metode *backward-forward* transisi *state* serta emisi *state* digunakan untuk proses *training*. Sedangkan untuk proses testing menggunakan algoritma Viterbi dan Baum-Welch. Sebagai kinerja model digunakan *Correlation Coefficient* (CC) yang didapat dari perbandingan *state* estimasi pada *output* dan *state* asli pada *input* model.

HASIL
Hasil simulasi menunjukkan bahwa nilai CC tergantung pada pemberian nilai acak state transisi *backward-forward*. Model dengan jumlah *state* 9, menunjukkan nilai CC rata-rata tertinggi adalah 0,7289 untuk algoritma Viterbi dan 0,7166 untuk algoritma Baum-Welch.

KESIMPULAN
Struktur model berdasarkan sistem HMM sahih untuk memprediksi ekson gen DNA Plamodium falciparum

Kata kunci : Prediksi ekson, gen-DNA, *coding sequence,* model *hidden* Markov

INTRODUCTION

In coding sequence (CDS), the exon prediction on DNA-gene is very important to find the protein structure. Genome *P. falciparum* belongs to genome eukaryotic and has a long DNA genome, consisting of several exons and introns alternately located. After the splicing process in DNA sequence, some regions of exon in CDS will be found to produce the protein.[(1)]

The hidden Markov model (HMM) structure is one of the model used to predict the exon in desoxyribonucleai acid (DNA) which is based on the exon region structure in CDS. The HMM structure of the model: start codon (Methyanine/ATG), exon, intron and stop codon (TAA, or TGA, or TAG) has been demonstrated by Rabiner, Henderson and Krogh.(2-4) This research developed a new model of structures for exon prediction on DNA of genes *Plasmodium falciparum was* based on the work of Nicorici and Anantharaman.(5,6) The significant difference of this model is it at least has two exon regions used for exon prediction like in CDS. Usually the 5' boundary of introns in most eukaryotes contains the dinucleotide Guanine-Thymine (GT), and the 3' boundary contains the dinucleotide Adenine-Guanine (AG). Furthermore, in intron regions of the dinucleotide GT were separated to be nucleotide Guanine (G) and Thymine (T), and dinucleotide AG to be nucleotide A and G.

METHOD

Research design

The HMM framework in this experiment used Viterbi algorithm for the training process and both Viterbi and Baum-Welch algorithms for the testing process. The same sequence has been done for the training and testing process. The program is written in Matlab 7.0, and Bio-informatics' toolbox to generate DNA sequences in Genbank format and has the functions of HMM training and testing.

The first experiment for the model based on CDS has been done with separated start codon (codon: ATG) from first exon and stop codon (codon: TAA, or TAG, or TGA) from the last exon, its model has the state number of 5 (Figure 1). Furthermore, the second experiment used separated dinucleotide GT and dinucleotide AG from the intron region, its model has the state number of 7 (Figure 2). The last experiment in the model also used separated dinucleotide GT into G and T states and dinucleotide AG into A and G states; the state number of the model is 9 (Figure 3). The emission values of the models can be performed as the matrix, the columns are the bases numbered on DNA sequences and the rows are the states number of the models designed. The state transition values of the models can be shown also in the figures.

Samples

A genome sequence of *P. falciparum* using 3D7 clone has 23-megabase nuclear genomes consisting of 14 chromosomes, encodes about 5,300 genes, and it has the most (A+T)-rich genome sequenced to date.(8-10) The data set of experiments 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.(11,12) Genbank format describes the CDS and original DNA sequences of genes *P. falciparum*. In CDS of genes contains at least 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.

HMM provides a good probability method for discrete sequences model of data like DNA sequences (alphabet of four letters: A, C, G and T). Prediction of exon in DNA of genes *P. falciparum* is based on exon region in CDS and it has at least two exon regions. The structure of the model based on CDS structure as shown in Figure 4, which separated is start codon from the first exon and separated stop codon from the last exon.

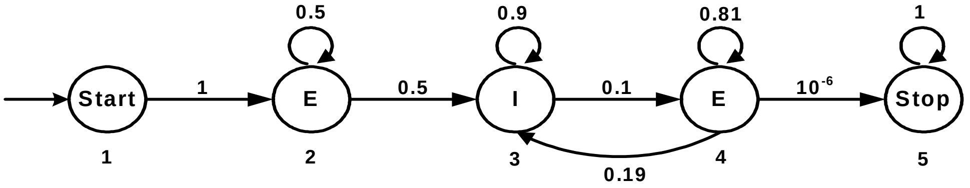


Figure 1. The HMM structure with 5 states

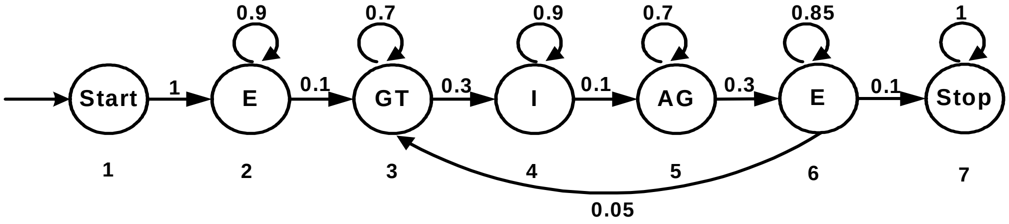


Figure 2. The HMM structure with 7 states

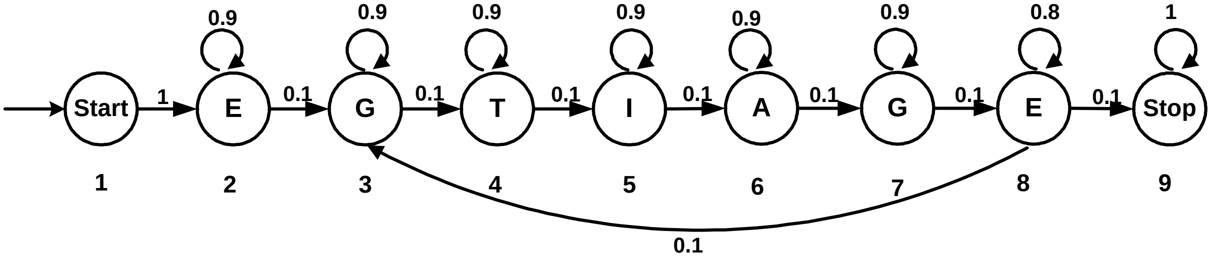


Figure 3. The HMM structure with 9 states

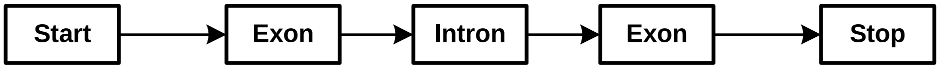


Figure 4. The HMM structure based on CDS

In the inputs are DNA sequences and set the state number of the sequences depends on the structure model. Based on HMM method, the Viterbi algorithm used for the training and needs the transition state values and emission values for the process. The value of transition state is random and the value of emission can be found with the distribution number of DNA nucleotide in each states. The result of HMM training is the estimated transition state and emission state values. The estimated transition and emission values were used for HMM testing process with Viterbi and Baum-Welch algorithms. The results of HMM testing process are the estimated states of the model.[2] The parameter of performance indicator used the correlation coefficient (CC), as a ratio of the estimated states from the output and the original states from the input of the model[1,7] and the formula is the following:

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

Where: TP= True Positive; FP= False Positive; TN= True Negative; FN= False Negative

Statistical analysis

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 ]
```

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 ]
```

The transition state values are randomly, but the first state on the model has the minimum transition state value of 0 and the last state has the maximum transition state value of 1. On the other hand, the emission state values were found from the distribution number of DNA nucleotide in each state depending on the models. Calculation of the CC is by using the above equation with the assumptions that the exon is positive and intron is negative.

RESULTS

Based on exon regions structure in CDS, the simulation results for the model with state number 5 above and transition state values randomly has found the average of CC value in Table 1. Table 2 and Table 3 shows the average of CC values of the model with state number 7 and state number 9. The transition state value in the first state of all structures on the models is 0 and the last state is 1. The other transition state values except in Table 2 are 0.1 and the highest average of CC values for this experiments are shown in Figure 5.

The model based on exon regions structure in CDS for state number nine resulted in the highest average of CC. Its values are 0.7289 by using Viterbi algorithm and 0.7166 by using Baum-Welch algorithm. However, the lowest average of CC values has been found for the model with state number five. Its values are 0.6735 by using Viterbi algorithm and 0.6661 by using Baum-Welch algorithm.

DISCUSSION

All these were based on the model of exon regions structure in CDS, at the state number nine resulted in the highest average of CC values. It is also shown that the emission state value of the models have the (A+T)-rich genome sequences of genes *Plasmodium falciparum*.

Table 1. The average of CC values for the model with state number 5

Exps	**Transition state values**					**Corr.Coef.(CC)**	
	tr$_{11}$	**tr**$_{22}$	**tr**$_{33}$	**tr**$_{44}$	**tr**$_{55}$	**Viterbi**	**B-Welch**
1	0	0.5	0.9	0.80	1	0.7054	0.6750
2	0	0.5	0.9	0.81	1	0.7061	0.6759
3	0	0.5	0.9	0.82	1	0.7054	0.6750
4	0	0.5	0.9	0.83	1	0.7054	0.6750
5	0	0.5	0.9	0.84	1	0.7045	0.6741
6	0	0.5	0.9	0.85	1	0.7054	0.6750
7	0	0.5	0.9	0.86	1	0.7062	0.6744
8	0	0.5	0.9	0.87	1	0.0012	0.0015
9	0	0.5	0.9	0.88	1	0.0011	0.0013
10	0	0.5	0.9	0.89	1	0.0011	0.0013

Table 2. The average of CC values for the model with state number 7

Exps	**Transition state values**							**Corr.Coef.(CC)**	
	tr$_{11}$	**tr**$_{22}$	**tr**$_{33}$	**tr**$_{44}$	**tr**$_{55}$	**tr**$_{66}$	**tr**$_{77}$	**Viterbi**	**B-Welch**
1	0	0.4	0.2	0.8	0.10	0.7	1	0.6713	0.6643
2	0	0.4	0.2	0.8	0.20	0.7	1	0.6713	0.6643
3	0	0.4	0.2	0.8	0.30	0.7	1	0.6733	0.6660
4	0	0.4	0.2	0.8	0.40	0.7	1	0.6731	0.6660
5	0	0.4	0.2	0.8	0.50	0.7	1	0.6735	0.6661
6	0	0.4	0.2	0.8	0.60	0.7	1	-0.0642	-0.0729
7	0	0.4	0.2	0.8	0.59	0.7	1	-0.0568	-0.0639
8	0	0.4	0.2	0.8	0.58	0.7	1	0.6735	0.6661
9	0	0.4	0.2	0.8	0.56	0.7	1	0.6735	0.6661
10	0	0.4	0.2	0.8	0.55	0.7	1	0.6735	0.6661

Table 3. The average of CC values for the model with state number 9

Exps	**Transition state values**									**Corr.Coef.(CC)**	
	tr$_{11}$	**tr**$_{22}$	**tr**$_{33}$	**tr**$_{44}$	**tr**$_{55}$	**tr**$_{66}$	**tr**$_{77}$	**tr**$_{88}$	**tr**$_{99}$	**Viterbi**	**B-Welch**
1	0	0.3	0.2	0.2	0.80	0.10	0.10	0.70	1	0.7242	0.7124
2	0	0.3	0.1	0.1	0.79	0.10	0.10	0.70	1	0.7242	0.7124
3	0	0.3	0.1	0.1	0.78	0.09	0.09	0.70	1	0.7242	0.7124
4	0	0.3	0.1	0.1	0.78	0.20	0.20	0.70	1	0.7258	0.7120
5	0	0.3	0.1	0.1	0.80	0.20	0.20	0.80	1	0.7258	0.7120
6	0	0.3	0.1	0.1	0.80	0.15	0.15	0.74	1	0.7258	0.7120
7	0	0.3	0.1	0.1	0.80	0.15	0.15	0.80	1	0.7258	0.7120
8	0	0.3	0.1	0.1	0.80	0.15	0.15	0.70	1	0.7258	0.7120
9	0	0.8	0.2	0.2	0.80	0.10	0.10	0.70	1	0.6787	0.6755
10	0	0.9	0.2	0.2	0.80	0.10	0.10	0.70	1	0.0451	0.0437

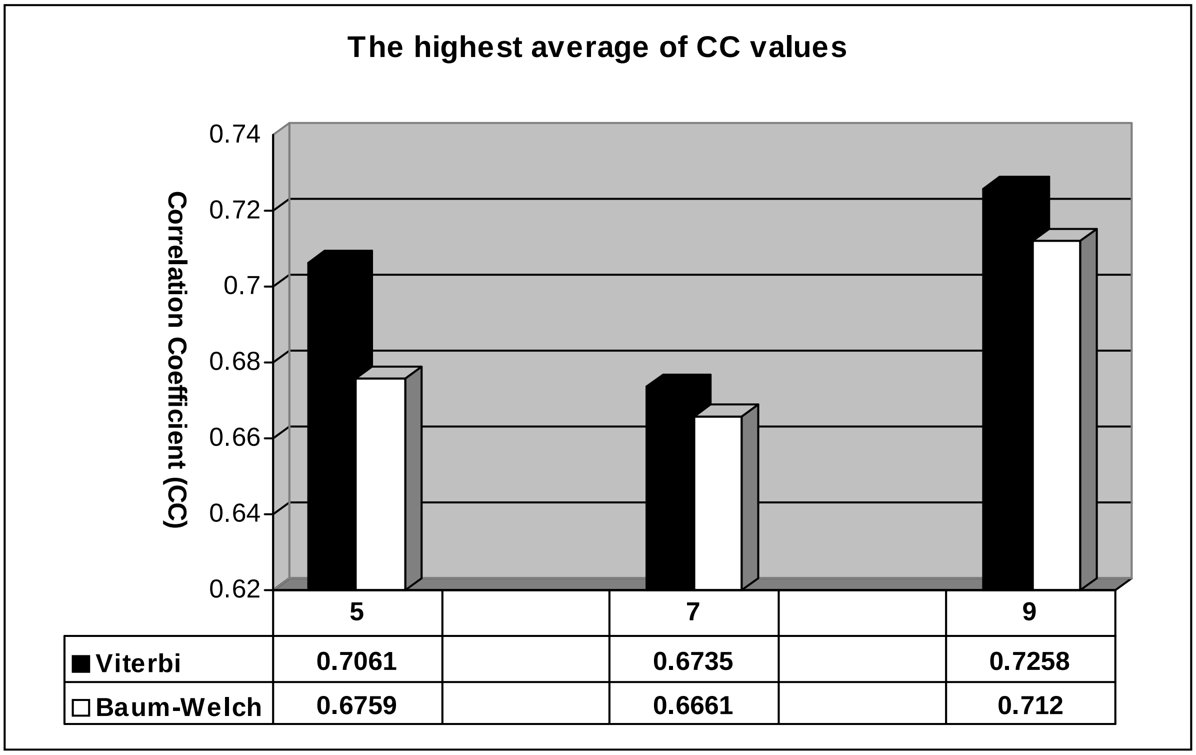


Figure 5. The highest average of CC values of the experiments

This can also be showed in their matrix values described above. In order to match with Markov's chain in eukaryotic gene structure for connecting the exon and intron, then the backward-forward method are used in this HMM system. The experiments were done by Viterbi algorithm for the training process and the both Viterbi and Baum-Welch algorithms for the testing process.

A set of 152 sequences is then used as data to train the model and a different HMM is produced for each set of sequences. HMM have been previously used in sequence analysis to produce an HMM that represents a sequence profile (a profile HMM), to analyze sequence composition and patterns, to locate genes by predicting open reading frame or coding sequence, and to produce protein structure predictions. This HMM generates sequences with various combinations of matched, mismatched, insertions and deletions, and gives these a probability, depending on the values of the various parameters in the model.

A total of 424 predicted protein-coding genes and 3 tRNA genes of *Plasmodium falciparum* have been identified on chromosomes 2 and 3, the two chromosomes for which sequencing have been completed.(13,14) Approximately 37% (158 genes) of three genes have a readily identifiable homologue in another species. Such similarity allows a function to be implied, with many of those identified being involved in parasite metabolism. Interestingly, a comparison of the predicted protein sequences with their homologues in other species showed that the majority of *Plasmodium falciparum* proteins have insertion of low complexity sequences, often runs of a single amino acid residue (typically asparagines, lysine or glutamine acid), or tandem arrays of a short peptide repeat sequence. Examples of such regions have been identified previously, and are polymorphic between different parasite isolates. Perhaps it will open up new avenues for drug and vaccine development. Perhaps it will make a genetically modified mosquito that cannot act as vector a reality once processes are in place to deal with ethical, legal and social issues. But any new developments are unlikely to happen before the 2010 deadline set for Roll Back Malaria. New drugs based on knowledge about the genome will probably not be affordable to those countries that need it the most.[15]

CONCLUSION

This research with a new structure model was using for exon prediction on DNA-genes of Plasmodium falciparum based on CDS structure by using the HMM system.

ACKNOWLEDGMENT

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

References

1. Samatova N.F. Computational gene finding using HMMs. UT-Battelle Information Center, 1201 Oak Ridge Turnpike, Suite 100, Oak Ridge, TN 37830. Institute Oak Ridge National Laboratory; 2003.
2. Lawrence R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceeding of The IEEE. 1989; 77: 257-68.
3. Henderson J, Salzberg S, Fasman K.H. Finding genes in DNA with a hidden Markov model. J Comput Biol 1997; 4: 127-42.
4. Krogh A. An introduction to hidden Markov models for biological sequences. In Computational methods in molecular biology. Salzberg SL, Searls DB, Kasif S. editors. Denmark. Center for Biological Sequence Analysis, Technical University of Denmark. 1998. p. 45-63.
5. Nicorici D, Astola J, Tobus I. Computational identification of exons in DNA with a hidden Markov model. Tampere International Center for Signal Processing, Tampere University of Technology, Finland. Available at: http://www.gensips.gatech.edu/processings/contributed/CP2-06.pdf. Accessed May 12, 2005.
6. Anantharaman T. Finding genes in genomic DNA The GENESCAN System. Available at : http://www.biostat.wisc.edu/bmi776. Accessed June 6, 2005.
7. Vaisman I. Bioinformatics and gene discovery. Bioinformatics Tutorial. North Carolina, United State: University of North Carolina at Chapel Hill. 1998.
8. Hall N, Gardner M.J, Hyman RW, Lasonder E, Wilson RJM, Scherf A, et al. Sequence of plasmodium falciparum chromosomes 1, 3-9 and 13. Nature 2002; 419: 527-31.
9. Gardner M.J, Hall N, Hyman RW, Hinterberg K, Mattei D, Wellem TE, et al. Sequence of plasmodium falciparum chromosomes 2, 10, 11 and 14. Nature 2002; 419: 531-4.
10. Hyman RW, Gardner M.J, Hall N, De Bruin D, Scherf A, Day KP, et al. Sequence of plasmodium falciparum chromosome 12. Nature 2002; 419: 534-7.
11. Anastassiou D. Genomic signal processing. IEEE Signal Processing Magazine 2001; 18: 4.
12. Alphey L. DNA sequencing. Manchester UK : Bios Scientific Publishers Limited; 1997.
13. Gardner M.J, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum. Science 1998; 282: 1126-32.
14. Bowman S, Lawson D, Basham D, Brown D, Chillingworth T, Churcher C. M, et al. The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum. Nature 1999; 400: 532-8.
15. Anonymous. Malaria after the genomes. The Lancet 2002; 360: 1107.