Analisis Kinerja Model Pengontrol Ekson DNA Menggunakan Metode Model Hidden Markov

Suhartati Agoes[1], Binti Solihah[2], Alfred Pakpahan[3]
[1]Dosen tetap Jurusan Teknik Elektro, FTI Usakti. Tlp 021-566-3232 ext 8429;
[2]Dosen tetap Jurusan Teknik Informatika, FTI Usakti. Tlp 021-566-3232 ext 8436;
[3]Dosen tetap Biologi FKG Usakti. Tlp 021-5672731 ext 6104;
[1]sagoes@ trisakti.ac.id, [2]binti_76@ yahoo.com, [3]alfred@ trisakti.ac.id

Abstrak *– urutan Deoxyribo asam nukleat (DNA) yang memiliki beberapa bagian ekson dalam urutan coding (cd) adalah bagian penting dalam proses biologis untuk menghasilkan protein. Tujuan dari penelitian ini adalah untuk mengontrol ekson DNA yang ada di CD dengan menggunakan Hidden Markov Model (HMM) sehingga protein yang dihasilkan tidak berubah. HMM metode memiliki parameter misalnya; negara, nilai keadaan transisi, negara emisi dasar dan algoritma yang digunakan untuk pelatihan dan proses pengujian. Nilai dari negara transisi secara acak berbagai ditentukan nilai antara 0 ~ 1. Pelaksanaan HMM di ekson kontroler memiliki struktur model 20-negara dan tes simulasi dilakukan dengan menggunakan nilai negara transisi dan jumlah urutan yang berbeda. Proses simulasi dengan struktur model 20-negara adalah menghasilkan nilai kinerja model dengan Koefisien Korelasi (CC) adalah 0,7571 dengan menggunakan 220 urutan. Penelitian ini meningkatkan nilai CC dengan cara mengelompokkan data dan hasilnya adalah 0,8808 untuk sub model dengan 69 urutan dan 0,8183 dengan 157 urutan.*

Kata kunci : *DNA, Exon, Coding urut, Korelasi koeffisien*

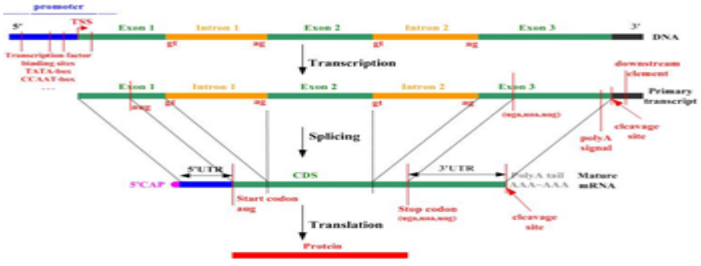
Abstract *– Deoxyribo nucleic acid (DNA) sequences which has several sections exons in the coding sequence (cds) is an important part in the biological process to produce the protein. The aim of this study is to control the exons of DNA that are on cds by using Hidden Markov Models (HMM) so that protein produced is not changed. HMM methods have parameters for example; state, the value of the transition state, the base emission state and the algorithm that are used for training and testing process. The value of the transition state is randomly determined range of values between 0~1. The implementation of the HMM in exon controller has a 20-state model structure and simulation tests performed using the value of the transition state and the number of different sequences. The simulation process with the 20-state model structure is produces value of performance model with Correlation Coefficient (CC) is 0.7571 by using of 220 sequences. This study increasing the CC value by clustering the data and the result are 0.8808 for sub model with 69 sequences and 0.8183 with 157 sequences.*

Keywords : *DNA, Exon, Coding sequence, Correlation Cofficient*

I. PENDAHULUAN

Deoxyribo Nucleic Acid (DNA) memiliki 4 macam basa nitrogen yaitu Adenin (A), Sitosin (C), Guanin (G) dan Timin (T). Keempat macam basa nitrogen ini menyusun DNA secara berpasangan, Guanin hanya dapat berpasangan dengan Sitosin atau sebaliknya, sedangkan Adenin dengan Timin atau sebaliknya, struktur ini dalam dogma DNA di kenal sebagai DNA double helix.

Pada DNA terdapat rangkaian basa-basa penyandi yang disebut dengan ekson dan rangkaian basa-basa bukan penyandi protein atau disebut dengan intron, dimana ekson atau kodon dapat ditranslasi menjadi protein atau asam amino, sedangkan intron harus dihilangkan saat dilakukan proses translasi menjadi protein. Struktur gen eukariot memiliki rangkaian-rangkaian penyandi atau ekson yang diselingi oleh rangkaian¬-rangkaian bukan penyandi atau intron. Ekson atau kodon dapat ditranslasi menjadi protein atau asam amino, sedangkan intron harus dihilangkan saat dilakukan proses ditranslasi menjadi protein. Struktur gen eukariot memiliki rangkaian-rangkaian penyandi atau ekson yang diselingi oleh rangkaian¬-rangkaian bukan penyandi atau intron seperti dijelaskan pada Gambar 1 [1].



Gambar 1. Struktur gen eukariot.

Pada Gambar 1 diatas dapat pula diketahui letak ekson dan intorn sekuen DNA yang letaknya bergantian (alternatively location) dan melalui proses transkripsi, dan translasi maka ekson pada cds dapat menjadi protein. Penelitian ini untuk mengontrol ekson cds sekuen DNA karena bila ekson cds berubah maka protein yang dihasilkan akan berubah pula.

Pemrosesan sinyal genom seperti DNA dan basa-protein dengan menggunakan teknologi digital membuat basa DNA dan asam amino yang membentuk protein dapat di asumsikan sebagai karakter string (huruf-huruf alpabet) sehingga dapat di manipulasi menjadi bit-bit 1 dan 0. Pemrosesan sinyal genom seperti DNA dengan berbantuan komputer menyebabkan terjadinya overload data yang mengakibatkan dikembangkannya berbagai metode untuk memprediksi langsung daerah penyandi (ekson).

Sampai saat ini telah dikenal dua puluh macam asam amino sebagai bahan dasar untuk protein yang terbentuk dari urutan tiga basa yang memberikan kode untuk satu asam amino, maka terjadi 43 = 64 kemungkinan kombinasi dari nukleotida sehingga menghasilkan 64 macam kelompok nukleotida [2].

II. METODE MODEL HIDDEN MARKOV

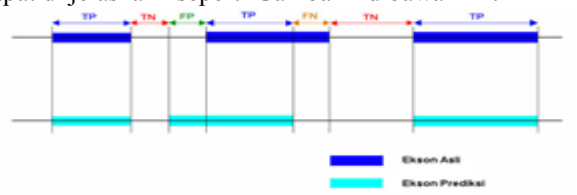
Hidden Markov Model (HMM) merupakan suatu model statistik yang digunakan untuk membuat karakteristik suatu frame sinyal DNA yang dapat di karakterisasi sebagai suatu representasi proses random parametrik. Beberapa garis besar dalam penggunaan metode HMM adalah seperti rantai Markov, elemen HMM, training dan testing HMM [3,4,5]. Pada penelitian ini menggunakan metode HMM untuk menganalisis dan mengontrol ekson DNA yang terdapat pada coding sequence (cds) sehingga perubahan urutan nukleotida pada ekson dapat diketahui [6]. Oleh karena itu metode ini sesuai dengan karakteristik suatu sekuen DNA dan sebagai kinerja dari model hidden Markov ini adalah nilai Correlation Coefficient (CC) dan diperoleh dari persamaan (1).

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

dimana:

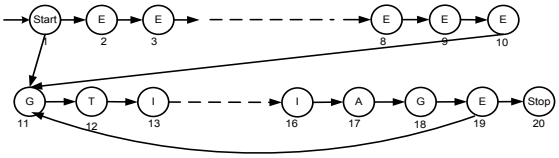
TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative.

Sebagai ilustrasi parameter-parameter TP (True Positive), TN (True Negative), FP (False Positive) dan FN (False Negative) di dalam mendapatkan nilai CC dapat di jelaskan seperti Gambar 2 dibawah ini.



Gambar 2. Ilustrasi TP, TN, FP dan FN.

Basa-basa ekson pada cds di dalam implementasi model hidden Markov dapat di bentuk menjadi state di dalam struktur model [6,7,8]. Pada penelitian ini struktur model yang digunakan berjumlah 20 stste dengan state pertama berisi basa-basa ATG dari bagian ekson pertama dan state terakhir adalah salah satu dari ketiga kodon stop yaitu bisa TAA atau TAG atai TGA. Struktur model hidden Markov dengan 20 state tersebut dapat di gambarkan secara umum seperti pada Gambar 3.



Gambar 3. Struktur model HMM untuk 20 state.

III. METODOLOGI PENELITIAN

Identifikasi model berbasis HMM yang menghasilkan kinerja model lebih baik untuk pengontrol ekson, diharapkan dapat berlaku umum untuk semua sekuen, Pada penelitian ini dilakukan ujicoba lebih lanjut pada model yang sudah dikembangkan pada penelitan sebelumnya yaitu dengan melakukan pengelompokkan data [9]. Pengujian kehandalan model dilakukan dengan cara menambahkan data dan mengidentifikasi kehandalan model yang dibuat dengan menghitung nilai CC. Hasil evaluasi pada tahap 1 menghasilkan sebuah hipotesa, keragaman data menyebabkan kegagalan proses generalisasi pada model yang terbentuk sehingga pengelompokan data akan menghasilkan kehandalan yang lebih baik. Berdasarkan hipotesa tersebut, pada langkah selanjutnya dilakukan pembentukan sub-sub model dengan cara membagi data dengan memperhatikan luaran model sebelumnya. Selanjutnya sub-sub model tersebut dihitung nilai CC nya untuk menunjukkan performa sub model.

Perangkat ujicoba simulasi ini terdiri dari input sekuen DNA Plasmodium falciparum, Plasmodium vivax dan Plasmodium knowlesi dengan panjang basa sekuen minimum 684 pasang basa atau basepair (bp) dan maksimum 10095 bp yang di unduh dari situs http://www.ncbi.nlm.nih.gov/entrez/ [2] dengan kriteria sekuen yang dapat di proses untuk ujicoba yaitu hanya memiliki satu cds yang dimulai dengan satu kodon start pada bagian ekson pertama dan di akhiri dengan salah satu kodon stop pada akhir bagian ekson terakhir (bukan merupakan sekuen partial) dan sekuen bukan pseudogene. Sedangkan sebagai perangkat keras adalah sebuah notebook dengan RAM 4GB, processor AMB A8, sedangkan Bahasa pemrograman menggunakan Matlab versi 2012 a. Proses training dan testing menggunakan algorithma Viterbi dan Baum-Welch.

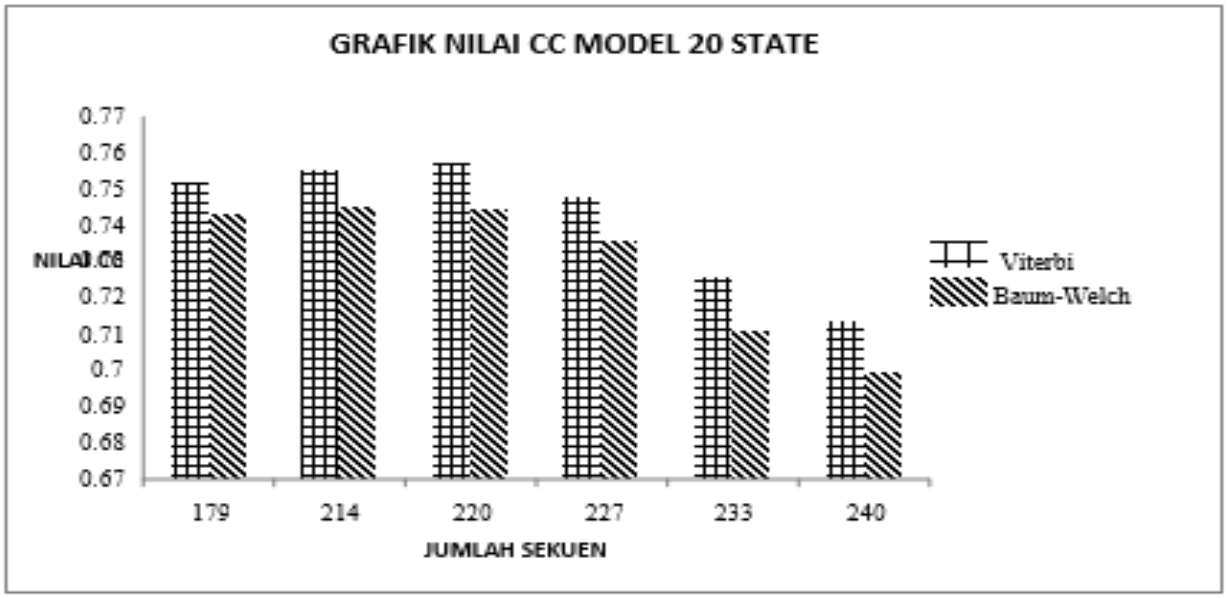
IV. HASIL SIMULASI

Proses simulasi dilakukan untuk beberapa kali ujicoba dengan menggunakan nilai-nilai transisi state yang berbeda-beda dan jumlah sekuen DNA sebagai input ditambahkan tahap demi tahap dengan maksimal berjumlah 240 sekuen. Struktur model yang

diimplementasikan terhadap metode HMM ini mempunyai jumlah state 20 dan hasil kinerja model adalah nilai-nilai CC seperti yang terdapat pada Tabel 1 dan untuk nilai-nilai CC yang lebih tinggi nilainya sesuai dengan bertambahnya jumlah sekuen di gambarkan secara grafik seperti pada Gambar 3 di bawah ini.

Tabel 1.Nilai CC Model PengontrolEksondenganstruktur 20 state

No	∑Sekuen	NilaiTransisi State					Iterasi	Nilai CC	
		State 1	State 2	State 11	State 19	State 20		Vit	Baum
1.	**179**	**0.1**	**0,9**	**0,1**	**0,81**	**0,1**	**24**	**0,7515**	**0,7433**
2.	214	0.1	0,9	0,1	0,81	0,1	22	0,7462	0,7341
3.	**214**	**0.1**	**0,9**	**0,1**	**0,85**	**0,05**	**69**	**0,7553**	**0,7452**
4.	216	0.1	0,9	0,1	0,85	0,05	48	0,7544	0,7452
5.	220	0.1	0,9	0,1	0,85	0,05	36	0,7539	0,7423
6.	220	0,05	0,95	0,1	0,85	0,05	29	0,7527	0,7468
7.	**220**	**0,01**	**0,99**	**0,1**	**0,85**	**0,05**	**43**	**0,7571**	**0,7447**
8.	227	0,01	0,99	0,05	0,80	0,15	19	0,7397	0,7321
9.	**227**	**0,05**	**0,95**	**0,05**	**0,80**	**0,15**	**22**	**0,7478**	**0,7358**
10.	227	0,1	0,9	0,05	0,80	0,15	21	0,7390	0,7278
11.	233	0,01	0,99	0,05	0,80	0,15	17	0,7243	0,7136
12.	**233**	**0,01**	**0,99**	**0,05**	**0,85**	**0,1**	**31**	**0,7254**	**0,7109**
13.	**240**	**0,1**	**0,9**	**0,1**	**0,80**	**0,1**	**31**	**0,7139**	**0,6991**
14.	240	0,01	0,99	0,1	0,80	0,1	37	0,7089	0,6971
15.	240	0,1	0,9	0,2	0,75	0,05	34	0,7084	0,6968



Gambar 3.Grafik Nilai CC Model 20

Pada Gambar 3 tampak bahwa model tidak mampu merepresentasikan kondisi semua data, terlihat saat jumlah data dinaikkan, nilai CC model turun. Ini menunjukkan bahwa model gagal untuk melakukan generalisasi terhadap data.

Tabel 2 berikut ini menunjukkan hasil nilai CC apabila data dibagi dua dan masing-masing digunakan untuk membangun submodel maka terjadi peningkatan nilai CC

Tabel 2. Nilai CC dengan submodel untuk 20 state.

Jumlah Sekuen	Komposisi nilai transisi state	Nilai CC dengan algorithma Viterbi
69	State1 = 0.1 State 2 =0.9 State11=0.1 State19 =0.1 State20 =0.81	0.8808
157	State1 = 0.01 State 2 =0.99 State11=0.1 State19 =0.85 State20 =0.05	0.8183

V. KESIMPULAN

Uji coba simulasi penelitian ini menghasilkan sebagai berikut:

1. Pada umumnya nilai CC yang dihasilkan dengan menggunakan algorithma Viterbi lebih baik bila dibandingkan dengan menggunakan algorithma Baum-Welch.
2. Pada jumlah sekuen 220, nilai CC adalah 0,7571, hasil ini lebih baik dari hasil ujicoba simulasi dengan menggunakan jumlah data sekuen yang lainnya.
3. Besar nilai transisi state yang ditentukan secara acak dapat mempengaruhi nilai CC yang dihasilkan sehingga perlu dilakukan banyak kombinasi dan variasi dalam menentukan nilai transisi state ini.
4. Peningkatan nilai CC tidak linier terhadap penambahan jumlah data sekuen sehingga ujicoba perlu dilakukan dengan menggunakan pengelompokkan data agar karakteristik data dapat diketahui sehingga kinerja model optimal.

VI. DISKUSI

Peningkatan nilai CC yang dihasilkan dari proses simulasi penelitian ini tidak linier terhadap penambahan jumlah sekuen DNA, untuk itu perlu dilakukan pengelompokkan data sekuen yang memiliki karakteristik tertentu agar kinerja model menjadi optimal. Beberapa teknik pengelompokan data sekuen DNA sebagai input proses simulasi dapat di ujicoba pada penelitian lanjutan dengan struktur model yang berbeda-beda untuk mengontrol ekson DNA berbasis HMM.

UCAPAN TERIMA KASIH

Ucapan terima kasih Kami sampaikan kepada Direktorat Jendral Pendidikan Tnggi (DIKTI) yang telah memberikan dana penelitian hibah bersaing sehingga penelitian ini bisa dilakukan. Demikian pula ucapan terima kasih ini Kami sampaikan kepada Universitas Trisakti khususnya kepada Lembaga Penelitian (LEMLIT) Universitas Trisakti atas bantuan dan bimbingannya di dalam melaksanakan dan menyelesaikan penelitian ini.

DAFTAR PUSTAKA

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

[2] Anastassiou Dimitris, Genomic signal processing, IEEE Signal Processing Magazine, Vol. 18, No.4, pp 8-20, Juli 2001.

[3] Henderson John, S Salzberg teven, Fasman Kenneth H, Finding gene in DNA with a Hidden Markov Model, Journal Computational Biology, Vol. 4, Issue 2, pp 127-141, 199.

[4] Rabiner Lawrence R., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of The IEEE, Vol 77, No 2, pp 257-286, Februari 1989.

[5] Agoes Suhartati, Pakpahan Alfred, Solihah Binti, Performance of Hidden Markov Model Structure on Deoxyribo Nucleic Acid Coding Sequence of Plasmodium falciparum. International Journal Asian Transactions on Science & Technology (ATST), Volume 01, Issue 05, November 2011.

[6] Nicorici Daniel, Astola Jaakko, Tobus Ioan , Computational identification of exons in DNA with a Hidden Markov Model, Tampere International Center for Signal Processing, Tampere University of Tecnology, 2002.

[7] Yada Tetsushi, Hirosawa Makoto, Gene recognition in cyanobacterium genomic sequence data using the hidden Markov model, Proceeding International Conference Intell. Syst. Mol. Biol, Vol 4, pp 252-260, 1996.

[8] Yada Tetsushi, Hirosawa Makoto, Detection of short protein coding regions within the cyanobacterium genome: application of the hidden Markov model, DNA. Res. Vol 31, Issue 6, pp 355-361, 31 Desember, 1996.

[9] Solihah Binti, Agoes Suhartati, Pakpahan Alfred, Optimasi Model Pengontrol Ekson Berbasis HMM Dengan Preprosesing Data Menggunakan Fuzzy C-Mean. Seminar Nasional Teknologi Infrmasi 2013 (SNTI 2013), Vol.10, No.1 Tahun 2013, 16 November 2013.