J M S K
Jurnal Matematika, Statistika, & Komputasi
http://journal.unhas.ac.id/index.php/jmsk
Vol. 13, No. 1, 55-65, Juli 2016
ISBN:1858-1382

Analisis *Hidden Markov Model* untuk Segmentasi Barisan DNA

Anisa, Andi Kresna Jaya, Sunarti*

Abstrak

Barisan DNA merupakan barisan yang terdiri dari basa Adenin (A), Sitosin (C), Timin (T), dan Guanin (G) yang diulang ribuan hingga jutaan kali dalam genom. Pada barisan ini, dilakukan analisis segmentasi DNA untuk mengidentifikasi dan memprediksikan pola kemunculan basa A, C, T, dan G. Pada penelitian ini, analisis segmentasi barisan DNA dilakukan dengan menggunakan model *Hidden Markov Model* (HMM) orde pertama yang merupakan sebuah model statistik dari sebuah sistem yang diasumsikan sebagai suatu proses Markov dengan parameter yang tak diketahui (Λ), dan tantangannya adalah menentukan parameter tersembunyi (*hidden*) dari parameter yang dapat diamati (ρ). HMM orde pertama berarti bahwa peluang munculnya suatu basa A, C, T, dan G, hanya dipengaruhi oleh satu basa sebelumnya. Dari hasil HMM pada penelitian ini yang merupakan suatu matriks peluang transisi, diperoleh nilai peluang terbesar adalah proses perpindahan basa Guanin (G) ke Adenin (A) dengan nilai peluang 0,434 dan nilai peluang terkecil adalah perpindahan basa Sitosin (C) ke Guanin (G) dengan nilai peluang 0,072. Data yang digunakan dalam penelitian ini adalah data barisan DNA *Homo sapiens* (manusia), yang merupakan salah satu database barisan DNA pada GenBank.

Kata Kunci: DNA, segmentasi DNA, Hidden Markov Model (HMM*).*

1. Pendahuluan

Istilah bioinformatik mulai dikemukakan pada pertengahan era 1980-an untuk mengacu pada penerapan komputer dalam biologi. Namun demikian, penerapan bioinformatika dalam bidang lain seperti pembuatan basis data dan pengembangan algoritma untuk analisis sekuens (barisan) biologis sudah dimulai sejak tahun 1960-an.

Kemajuan teknik biologi molekular dalam mengungkap sekuens biologis dari protein sejak awal 1950-an dan asam nukleat sejak 1960-an, mengawali perkembangan basis data dan teknik analisis sekuens biologis. Basis data sekuens protein mulai dikembangkan pada tahun 1960-an di Amerika Serikat, sementara basis data sekuens dari *Deoxyribonucleid Acid* (DNA) dikembangkan pada akhir 1970-an di Amerika Serikat dan Jerman, yang dipusatkan pada Laboratorium Biologi Molekular Eropa (*European Molecular Biology Laboratory*). Penemuan teknik sekuensing DNA yang lebih cepat pada pertengahan 1970-an dan menjadi landasan terjadinya ledakan jumlah sekuens DNA yang berhasil diungkapkan pada 1980-an dan 1990-an, menjadi salah satu pembuka jalan bagi proyek-proyek pengungkapan genom yang merupakan

* *Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin, Jl. Perintis Kemerdekaan Km.10 , Tamalanrea Makassar,*
email: nkalondeng@yahoo.com

keseluruhan kode genetik, meningkatkan kebutuhan akan pengelolaan dan analisis sekuens, sehingga mendorong berkembangnya bioinformatika.

Bioinformatika merupakan ilmu yang mempelajari penerapan teknik komputasional untuk mengelola dan menganalisis informasi biologis. Bidang ini mencakup penerapan metode-metode matematika, statistika, dan informatika untuk memecahkan masalah-masalah biologis, terutama dengan menggunakan sekuens DNA dan asam amino serta informasi yang berkaitan DNA tersebut. Contoh topik utama bidang ini meliputi basis data untuk mengelola informasi biologis, penyejajaran sekuens (*sequence alignment*), prediksi struktur untuk meramalkan bentuk struktur protein maupun struktur sekunder *Ribonucleid Acid* (RNA), analisis filogenetik untuk mengetahui hubungan evolusi, dan analisis ekspresi gen.

Istilah genom dipakai untuk menunjukkan keseluruhan kode genetik pada kromosom yang ada pada suatu organisme, dan barulah pada tahun 1944 diketahui materi dari kode genetik itu adalah DNA yang ada pada setiap organisme. DNA terdiri dari empat struktur kimia yang hampir sama susunannya yang dinamakan nukleotida, yaitu Adenin (A), Timin (T), Sitosin (C) dan Guanin (G). Empat basa ini diulang-ulang ribuan sampai jutaan kali jumlahnya dalam genom, mulai dari organisme sederhana bersel tunggal seperti bakteri, sampai yang kompleks dengan sel banyak seperti tanaman, hewan dan manusia.

Berbagai metode statistika dikembangkan untuk mengidentifikasi segmentasi barisan DNA. Beberapa diantaranya adalah metode Pendugaan Maksimum Likelihood (*Maximum Likelihood Estimation*) oleh Fu dan Curnow tahun 1990, pendekatan Bayesian statistik (*Bayesian Approach*) oleh Liu dan Lawrence tahun 1996, dan Model Markov Tersembunyi atau *Hidden Markov Model* (HMM) oleh Churchill tahun 1989 (Braun & Muller, 1998).

Metode statistik yang diusulkan oleh Churchill mendeskripsikan struktur rangkaian DNA dengan sebuah HMM yang merupakan sebuah model statistik dari sebuah sistem yang diasumsikan sebuah proses Markov dengan parameter yang tak diketahui, dan tantangannya adalah menentukan parameter tersembunyi (*hidden*) dari parameter yang dapat diamati. Pada model Markov umum, segmentasi dalam hal ini disebut sebagai state, langsung dapat diamati, sehingga probabilitas transisi state menjadi satu-satunya parameter. Di dalam model Markov yang tersembunyi (HMM), state-nya tidak dapat diamati secara langsung, akan tetapi yang dapat diamati adalah variabel-variabel yang terpengaruh oleh state. Setiap state memiliki distribusi probabilitas atas barisan yang mungkin muncul. Oleh karena itu rangkaian barisan yang dihasilkan oleh HMM memberikan sebagian informasi tentang sekuens state.

Berdasarkan hal tersebut di atas, maka penelitian ini akan mengkaji dan menganalisis segmentasi barisan DNA. Sebelumnya, untuk penentuan jumlah segmen yang akan digunakan pada proses segmentasi tersebut, dilakukan dengan metode pengelompokan (*Cluster Analysis*).

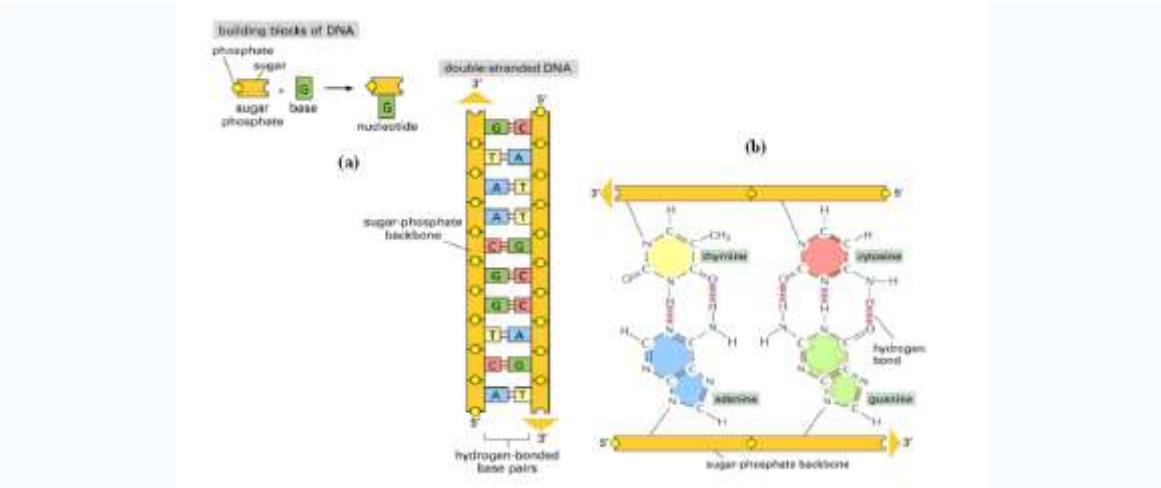
2. Landasan Teori

2.1. Deoxyribonuclead Acid (DNA)

Struktur DNA

Molekul *Deoxyribonucleid Acid* (DNA) adalah blok yang membangun kehidupan karena DNA merupakan komponen kimia utama kromosom dan merupakan bahan yang menghasilkan gen. DNA kadangkala disebut sebagai molekul warisan, karena DNA memuat materi genetik yang membawa sifat-sifat organisme induk yang diwariskan kepada generasi selanjutnya. Di samping itu, DNA juga berfungsi untuk proses sintesis protein. Pada tahun 1953, Watson dan Crick telah membuka wawasan baru tentang penemuan model struktur DNA, dimana struktur ini membentuk suatu ikatan yang disebut *double helix* DNA. Basa nukleid terdiri dari sub unit kecil nukleotida yang merupakan kumpulan fosfat, gula pentosa, gula ribosa dan deoksiribosa, serta basa nitrogen purin dan pirimidin. Komponen basa nitrogen pada DNA terdiri dari Adenin (A),

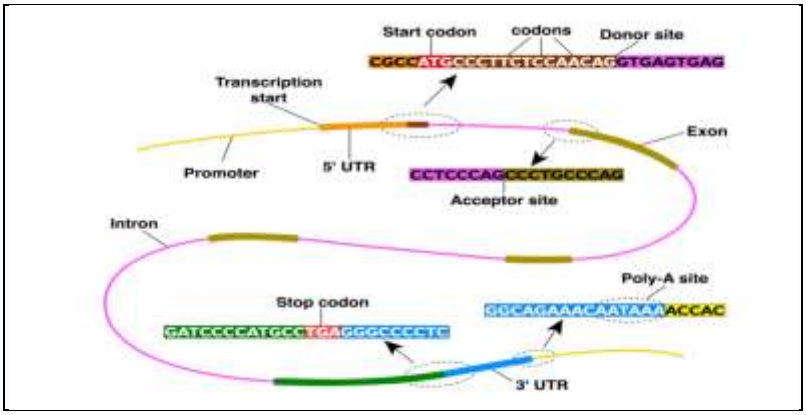
Sitosin (C), Timin (T), dan Guanin (G). DNA terbentuk dari empat tipe nukleotida, yang berikatan secara kovalen membentuk rantai polinukleotida (rantai DNA atau benang DNA), dengan tulang punggung gula-fosfat tempat melekatnya basa-basa. Dua rantai polinukleotida saling berikatan melalui ikatan hidrogen antara basa-basa nitrogen pada rantai yang berbeda. Semua basa berada di dalam *double helix*, sedangkan tulang punggung gula-fosfat berada di bagian luar. Tulang punggung gula fosfat terdiri dari ujung 5' dan 3'. Basa purin (A-T) selalu berpasangan dengan pirimidin (G-C). Perpasangan secara komplemen tersebut memungkinkan pasangan basa dikemas dengan susunan yang paling sesuai. Hal ini bisa terjadi bila kedua rantai polinukleotida tersusun secara terpilin (*anti paralel*). Untuk lebih jelasnya, pembentukan struktur DNA dari gula fosfat sebagai tulang punggung (*backbone*) dan basa-basa nukleotida diberikan pada Gambar 1 berikut.



Gambar 1. (a) Pembentukan Secara Skematik Struktur DNA dari Gula Fosfat Sebagai '*backbone*' dan Basa Nukleotida.
(b) AT Dihubungkan oleh 2 Ikatan Hidrogen dan GC Dihubungkan oleh 3 ikatan hidrogen.

Segmentasi Barisan DNA

Metode pengkajian segmentasi barisan DNA diawali dengan mengetahui intron, yang merupakan wilayah dari gen yang tidak membawa kode genetik (*non-coding*). Nishio (1995) menyatakan bahwa struktur intron ini sangat menarik karena memiliki bentuk transposisi yang tidak dapat diubah lagi, sehingga lebih meyakinkan untuk menentukan pedoman waktu untuk menguraikan filogenik (hubungan evolusi). Sedangkan exon merupakan wilayah dari gen yang dapat membawa kode genetik. Suatu daerah khusus dari exon yang membawa kode genetik disebut dengan kodon. Daerah ini terdiri atas titik inisiasi yaitu start kodon dengan kombinasi basa ATG, yang berguna pada proses sintesis protein, serta stop kodon dengan tiga kemungkinan kombinasi basa yaitu TAA, TAG, dan TGA. Untuk itu, dasar dari analisis segmentasi DNA lebih dikembangkan pada daerah exon serta barisan yang terdiri dari intron dan exon (Braun dan Muller, 1998). Suatu gambaran sederhana mengenai struktur intron dan exon diberikan pada gambar berikut.



Gambar 2. Struktur Intron dan Exon pada DNA.

2.2. Rantai Markov (*Markov Chain*)

Konsep dasar Rantai Markov (Markov Chain), yang akan disingkat dengan MC, diperkenalkan sekitar tahun 1907 oleh Andrei A. Markov (1856-1922). Model ini berhubungan dengan suatu rangkaian proses, dimana kejadian akibat suatu eksperimen hanya bergantung pada kejadian yang langsung mendahuluinya dan tak bergantung pada rangkaian kejadian sebelumnya. MC merupakan proses Markov dengan ruang keadaan diskrit.

Matriks Peluang Transisi

Matriks peluang transisi adalah matriks yang memuat semua informasi yang mengatur perpindahan sistem dari suatu state ke state lainnya (Purnaba, 2001). Suatu matriks dikatakan matriks transisi atau matriks stokastik jika semua peluang transisi P_{ij} adalah tetap, dan tidak bergantung pada waktu *t*, dimana P_{ij} adalah peluang transisi satu langkah yang bergerak dari keadaan *i* ke keadaan *j*. Penjelasan mengenai matriks peluang transisi diuraikan berikut ini.

Misalkan $\{Y_n\}$ adalah suatu Rantai Markov dengan parameter diskrit.

Fungsi massa peluang dari $\{Y_n\}$ adalah :

$$P_j(n) = P[Y_n = j] \tag{1}$$

Sedangkan fungsi kepadatan peluangnya dituliskan dalam bentuk:

$$P_{i,j}(m,n) = P\left[Y_n = j \,|\, Y_m = i\right] \tag{2}$$

dimana $n \geq m \geq 0$, n dan m adalah waktu, dan i dan j adalah state. Selanjutnya $P_{i,j}(m,n)$ disebut fungsi peluang transisi dari MC $\{Y_n\}$, dimana $\{Y_n\}$ dikatakan homogen jika $P_{i,j}(m,n)$ bergantung pada $n-m$,

$$P_{i,j}(n) = P[Y_{n+t} = j | Y_z = i] \text{ untuk suatu } z \geq 1,\ z \in \text{Bilangan Bulat.} \tag{3}$$

Persamaan (3) adalah peluang transisi Rantai Markov $\{Y_n\}$ homogen pada *n*-state. $P_{i,j}(n)$ adalah peluang bersyarat Rantai Markov $\{Y_n\}$ homogen di state *i* akan berpindah ke state *j* dalam *n-tahap*. Peluang transisi 1-*tahap*, dapat dituliskan dalam bentuk sederhana, yaitu

$$P_{i,j} = P[Y_{n+1} = j | X_n = i], \text{ untuk setiap bilangan bulat } t \geq 0 \tag{4}$$

Fungsi peluang transisi dalam Rantai Markov $\{Y_n\}$ ini memenuhi persamaan Chapman-Kolmogorov. Untuk suatu waktu $n > u > m \geq 0$ dan state *i* dan *j*, maka persamaan Chapman-Kolmogorov dituliskan dalam bentuk:

$$P_{i,j}(m,n) = \sum_{state} P_{i,t}(m,u) P_{t,j}(u,n) \tag{5}$$

Dari persamaan (1) dan (4) diperoleh :

$$P[Y_n = j | Y_m = i] = \sum_{state} P[Y_n = j | Y_u = t, Y_m = i] P[Y_u = t | Y_m = i] \tag{6}$$

Misalkan $\{Y_n\}$ suatu Rantai Markov dengan state {1,2,.....,*n*}, maka peluang transisi dari Rantai Markov tersebut dapat dinyatakan dalam bentuk matriks peluang transisi dengan lambang ***P*** (*m*,*n*).

$$\mathbf{P}(m,n) = \begin{bmatrix} P_{1,1}(m,n) & P_{1,2}(m,n) & \dots & P_{1,j}(m,n) \\ P_{2,1}(m,n) & P_{2,2}(m,n) & \cdots & P_{2,j}(m,n) \\ \vdots & \vdots & \ddots & \vdots \\ P_{i,1}(m,n) & P_{i,2}(m,n) & \cdots & P_{i,j}(m,n) \end{bmatrix} \tag{7}$$

dimana semua elemen matriks peluang transisi **P**(*m*,*n*) memenuhi kondisi :

1. $P_{i,j}(m,n) \geq 0$ *untuk setiap i,j* (8)
2. $\sum_j P_{i,j}(m,n) = 1$ *untuk setiap i* (9)

Kedua kondisi inilah yang menggambarkan bahwa proses terjadinya suatu transisi adalah melalui proses acak.

2.3. *Hidden Markov Model* (*HMM*)

Model Markov Tersembunyi (*Hidden Markov Model*) merupakan sebuah model statistik dari sebuah sistem yang diasumsikan sebuah proses Markov dengan parameter yang tak diketahui, dan tantangannya adalah menentukan parameter-parameter tersembunyi (*hidden*) dari parameter-parameter yang dapat diamati.

Churchill (1992) mengusulkan sebuah model rantai Markov tersembunyi untuk model segmentasi dari barisan DNA dan berusaha memprediksikan lokasi segmen-segmen yang mungkin dalam mitokondria yang merupakan bagian dari sel yang berfungsi untuk respirasi. Asumsi model Markov tersembunyi adalah bahwa segmen yang berbeda dapat diklasifikasikan ke dalam suatu himpunan berbeda dari state. Dalam setiap state, data nukleotida diasumsikan mengikuti sebuah distribusi probabilitas, sebagai contoh Rantai Markov orde pertama. State diasumsikan berpindah dari state yang satu ke state yang lain secara acak sesuai dengan konsep probabilitas. State yang tersembunyi ini merupakan kejadian acak. Kedua asumsi ini merupakan asumsi dasar sebuah Rantai Markov tersembunyi, dan model ini pertama kali diperkenalkan oleh Rabiner tahun 1989 (Braun & Muller, 1998).

Pada analisis DNA, sebuah barisan DNA $y = y_1, y_2, y_3, \cdots, y_n$ dapat dipertimbangkan sebagai suatu realisasi dari proses acak $Y_1, Y_2, Y_3, \cdots, Y_n$, dengan $Y_t \in \{a, c, t, g\}$, t = 1, 2, 3,…,*n*, yang merupakan representasi dari 4 nukleotida yaitu Adenin, Sitosin, Guanin dan Timin. Segmentasi barisan DNA merupakan suatu hal yang tidak dapat ditentukan secara pasti. Oleh karena itu, segmen tersebut dapat diasumsikan sebagai suatu state yang tersembunyi (*hidden*) pada Rantai Markov. Proses acak yang mungkin terjadi merupakan rangkaian kode genetik dari DNA yaitu A, C, G, dan T. Karena hanya ada empat kemungkinan kode yang bisa muncul, maka

kemunculan tersebut diasumsikan mengikuti distribusi Dirichlet, dimana distribusi ini merupakan generalisasi dari distribusi Beta-Binomial untuk data Multinomial (Congdon, 2003). Boys *et al.* (2000) juga menyatakan bahwa distribusi Dirichlet adalah distribusi yang dapat menggambarkan secara luas mengenai peluang transisi dari suatu perpindahan lokasi. Sehingga distribusi Dirichlet inilah yang digunakan sebagai asumsi untuk distribusi dari state.

Bentuk umum dari distribusi Dirichlect adalah:

$$D(\theta|\alpha) = \frac{1}{Z(\alpha)} \prod_{k=1}^{N} \theta_k^{\alpha_k - 1} \tag{10}$$

dengan parameter $\theta = (\theta_1, \theta_2, \cdots, \theta_N)$ adalah distribusi probabilitas, dan domain Θ berada pada

$$\Theta = \{\theta | \theta_k \geq 0, \forall k = 1,2,\cdots,N, dan \sum_{k=1}^{N} \theta_k = 1\} \tag{11}$$

Jika diasumsikan terdapat *r* segmen dalam rangkaian DNA. Tipe segmen tersembunyi pada lokasi *t* akan ditunjukkan oleh $S_t \in \{1,2,3,\cdots,r\}$ untuk $t = 1,2,3,\cdots,n$. Agar proses segmentasi bisa dilakukan dengan lebih sederhana, maka nilai *r* akan ditentukan langsung berdasarkan data yang digunakan, yang akan dijelaskan pada sub bab berikutnya.

Parameter ρ

Diasumsikan bahwa transisi dasar $Y_t \rightarrow Y_{t+1}$ mengikuti orde pertama rantai Markov dengan pilihan dari matriks transisi ditentukan oleh state tersembunyi. Matriks transisi ditunjukkan oleh $\rho = \{P^1, P^2, P^3, \cdots, P^r\}$ dimana $P^k = (P_{ij}^k)$. Sehingga bentuk struktur matriks peluang transisi orde pertama rantai Markov adalah.

Nukleotida pada *(i +1)*

Nukleotida pada *i*	*a*	*c*	*t*	*g*
a	P_{11}	P_{12}	P_{13}	P_{14}
c	P_{21}	P_{22}	P_{23}	P_{24}
t	P_{31}	P_{32}	P_{33}	P_{34}
g	P_{41}	P_{42}	P_{43}	P_{44}

P_{ij} ↑

peluang nukleotida *i* berpindah ke nukleotida *j*

Persamaan terbaru untuk transisi dasar adalah :

$$P(Y_t = y_t | S_t = s_t, y_1, y_2, y_3, \ldots, y_{t-1}, \rho) = P(Y_t = y_t | S_t = s_t, y_{t-1}, \rho) = P_{y_{t-1}, y_t}^{(s_t)} \tag{12}$$

Parameter Λ

Proses state tersembunyi dari tipe segmen diasumsikan homogen pada orde pertama rantai Markov dan berdistribusi Dirichlet dengan matriks transisi *r x r*.

Distribusi probabilitas transisi state $\Lambda = \{\lambda_{ij}\}$ dengan $i = j = 1,2,3,\cdots,r$ adalah

$$\lambda_{ij} = P(S_{i+1} = s_j | S_i = s_i) \tag{13}$$

Bentuk matriks transisi state tersembunyi :

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1j} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{i1} & \lambda_{i2} & \cdots & \lambda_{ij} \end{bmatrix}$$

Persamaan terbaru untuk transisi segmen adalah :

$$P(S_t = s_t | s_1, s_2, s_3, \ldots, s_{t-1}, \Lambda) = P(S_t = s_t | s_{t-1}, \Lambda) = \lambda_{s_{t-1}, s_t} \quad (14)$$

Diasumsikan bahwa y_1 dan s_1 mengikuti distribusi seragam Diskrit yang independen. Jika diberikan observasi rangkaian DNA *y* dan tipe segmen tersembunyi *s* maka fungsi Likelihood untuk model parameter ρ dan Λ adalah

$$L(\rho, \Lambda | y, s) = \prod_{t=1}^{n} P(y_t, s_t | \rho, \Lambda) = P(y_1, s_1 | \rho, \Lambda) * \prod_{t=2}^{n} P(y_t | y_{t-1}, s_t, \rho) P(s_t | s_{t-1}, \Lambda)$$

$$= \frac{1}{4r} \prod_{t=2}^{n} P_{y_{t-1}, y_t}^{(s_t)} \lambda_{s_{t-1}, s_t} \quad (15)$$

2.4. Metode Penentuan Jumlah Segmen

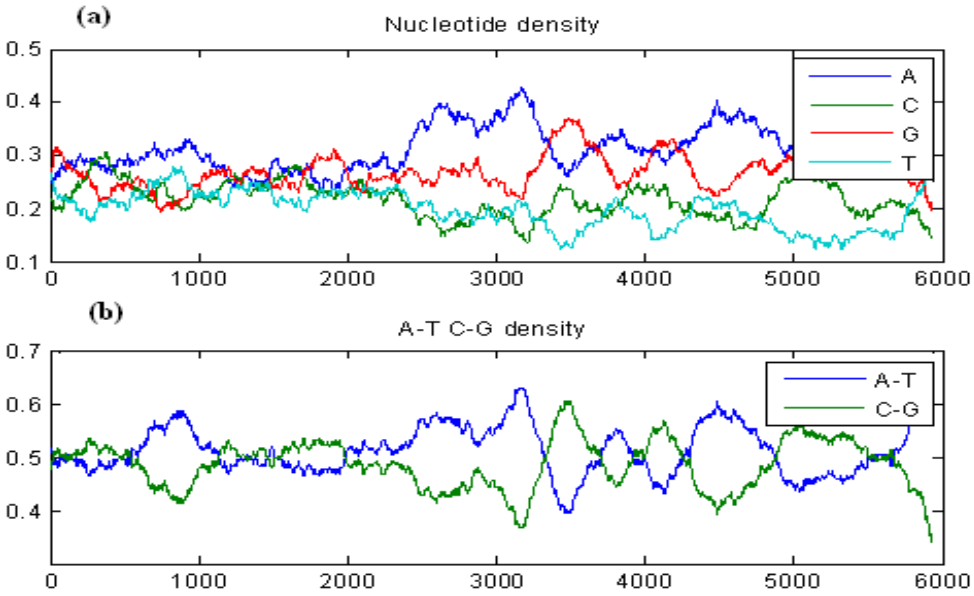
Data yang digunakan untuk segmentasi DNA biasanya berjumlah ribuan, sehingga untuk penyederhanaan proses segmentasi, maka jumlah *r* sebaiknya diketahui sebelumnya. Salah satu metode yang dapat diadopsi untuk menentukan berapa jumlah segmen relatif terhadap data yang digunakan adalah dengan melakukan pengelompokan atau *cluster*. Untuk itu, penelitian ini akan mengadopsi Metode *Cluster Analysis* yang merupakan bagian dari Analisis Multivariat (Johnson & Wichern, 2002).

3. Hasil dan Pembahasan

Deskripsi Data Barisan DNA

Data yang digunakan dalam penelitian ini merupakan data sekunder mengenai barisan DNA manusia (*homosapiens*), yang merupakan data yang diperoleh dari database GenBank. Dari database tersebut dapat diketahui bahwa jumlah data sebanyak *n* = 5920, dengan kombinasi jumlah basa Adenin A = 1869, Sitosin C = 1285, Guanin G = 1592, dan Thymin T = 1174, serta jumlah dari kombinasi kodon, yang merupakan gabungan dari tiga jenis basa tersebut, dan terdapat dalam barisan. Pada proses sintesis protein, kodon ini mengalami proses transkripsi dan translasi dengan bantuan mRNA (*RNA messenger*), serta berkorespondensi dengan jenis-jenis asam amino. Sebagai contoh, TTT merupakan kode asam amino *Phenylalanine*. Selain itu, terdapat tiga kode lain, yaitu TAA, TAG, dan TGA, yang tidak berkorespondensi dengan jenis asam amino karena merupakan kode *stop kodon.*

Gambar berikut merupakan grafik peluang munculnya A, C, T dan G, serta grafik basa *Purin* (A-T) dan *Pirimidin* (G-C) pada barisan DNA.



Gambar 4. (a) Grafik Kepadatan Peluang Basa Nukleotida.
(b) Grafik Kepadatan Peluang Basa *Purin* (A-T) dan *Pirimidin* (G-C).

Gambar 4(a) menunjukkan kepadatan peluang nukleotida A, C, T, dan G dalam barisan DNA. Pada interval barisan 1 sampai 2000, rata-rata peluang kemunculan A, C, T, dan G berkisar antara nilai 0,2 sampai 0,3. Hal ini berarti jumlah basa tersebut dalam barisan DNA memiliki proporsi yang relatif sama. Namun pada interval barisan 2000 sampai 6000, terlihat nilai peluang kemunculan basa Adenin lebih besar dibanding basa yang lain, artinya jumlah basa Adenin dalam barisan DNA relatif lebih banyak dibanding jumlah basa yang lain. Peluang kemunculan basa yang terkecil adalah Timin, artinya jumlah basa Timin dalam barisan DNA relatif paling sedikit dibanding basa yang lain. Sedangkan pada gambar 4(b) menunjukkan kepadatan peluang basa purin (A-T) dan primidin (G-C) dalam barisan DNA. Terlihat dari grafik tersebut, garis peluang purin (dalam warna biru) selalu berkomplemen dengan garis peluang pirimidin (dalam warna hijau) dengan sumbu simetri terdapat pada nilai peluang 0,5.

Analisis Pengelompokan Segmen (*Cluster Analysis*)

Pada barisan DNA *homosapiens* tersebut, dilakukan C*luster Analysis*, dengan menggunakan software Minitab 14, untuk menentukan jumlah dari segmen (state) serta barisan-barisan yang termasuk dalam segmen tersebut. *Cluster Analysis* adalah salah satu metode untuk mengelompokkan barisan DNA yang mempunyai kemiripan sifat sedemikian sehingga barisan DNA yang terdapat pada satu kelompok bersifat homogen, sedangkan antar kelompok yang satu dengan yang lainnya bersifat heterogen. Pengelompokan barisan DNA ini merupakan bentuk segmentasi, yang pada Rantai Markov diasumsikan sebagai state.

Berdasarkan hasil *Cluster Analysis* tersebut, diperoleh tiga kelompok (segmen) barisan. Dari hasil cluster, dapat diketahui bahwa jumlah barisan paling banyak terdapat pada *cluster* 1 dengan n_1 = 3790, kemudian *cluster* 2 dengan n_2 = 1550, serta cluster 3 dengan n_3 = 580. Di samping itu, dapat diketahui jumlah dari basa Adenin, Sitosin, Timin, dan Guanin pada masing-masing *cluster*. Dengan melakukan *Cluster Analysis*, dapat diketahui jumlah barisan dan jumlah kombinasi dari basa A, C, T, dan G dari masing-masing kelompok, serta akan memudahkan dalam proses analisis model Markov Tersembunyi (HMM).

Analisis *Hidden Markov Model* (HMM)

Pada data barisan DNA *homosapiens* akan dicari matriks peluang transisi untuk A, C, T, dan G pada ketiga state (ρ), dan estimasi matriks peluang transisi dari segmen (Λ). Berdasarkan hasil analisis HMM dengan menggunakan software Matlab 7.1, diperoleh matriks transisi observasi barisan DNA pada tiap segmen.

Tabel 1 merupakan matriks peluang transisi keempat basa nukleotida pada tiap segmen. Pada segmen I diperoleh peluang transisi basa Adenin ke basa Adenin barisan berikutnya, yaitu P_{11} = 0,334. Sedangkan peluang transisi basa Adenin ke basa Sitosin, yaitu P_{12} = 0,149, dan seterusnya. Jumlah nilai peluang dalam baris yang sama selalu sama dengan 1. Dari matriks peluang transisi barisan DNA pada segmen I, diperoleh peluang transisi terbesar $P_{41} = 0{,}434$. Hal ini menunjukkan bahwa peluang munculnya basa Adenin jika basa sebelumnya adalah Guanin sebesar 0,434. Di samping itu, diperoleh peluang transisi terkecil $P_{24} = 0{,}086$. Ini berarti, peluang munculnya basa Guanin jika basa sebelumnya adalah Sitosin sebesar 0,086. Pada segmen II, peluang transisi terbesar dan terkecil masing-masing $P_{21} = 0{,}379$ dan $P_{24} = 0{,}072$. Sedangkan pada segmen III, peluang transisi terbesar dan terkecil masing-masing $P_{34} = 0{,}395$ dan $P_{24} = 0{,}087$. Dari ketiga segmen, diperoleh peluang transisi terbesar, yaitu $P_{41} = 0{,}434$ terdapat pada segmen I. Sedangkan peluang transisi terkecil, yaitu $P_{24} = 0{,}072$ terdapat pada segmen II, yang artinya kemungkinan munculnya basa Guanin jika yang muncul sebelumnya adalah basa Sitosin memiliki peluang yang terkecil.

Tabel 1. Matriks Peluang Transisi ρ , Λ dan $\hat{\Lambda}$ pada Segmen I, II, dan III.

Segmen	ρ	Λ	$\hat{\Lambda}$
I	$\begin{bmatrix} 0{,}334 & 0{,}149 & 0{,}167 & 0{,}350 \\ 0{,}361 & 0{,}258 & 0{,}295 & 0{,}086 \\ 0{,}115 & 0{,}235 & 0{,}228 & 0{,}422 \\ 0{,}434 & 0{,}228 & 0{,}101 & 0{,}237 \end{bmatrix}$	$\begin{bmatrix} 0{,}75 & 0{,}15 & 0{,}10 \\ 0{,}15 & 0{,}80 & 0{,}05 \\ 0{,}12 & 0{,}07 & 0{,}81 \end{bmatrix}$	$\begin{bmatrix} 0{,}7311 & 0{,}1636 & 0{,}1053 \\ 0{,}0211 & 0{,}9135 & 0{,}0654 \\ 0{,}1163 & 0{,}0743 & 0{,}8094 \end{bmatrix}$
II	$\begin{bmatrix} 0{,}322 & 0{,}199 & 0{,}190 & 0{,}289 \\ 0{,}379 & 0{,}223 & 0{,}326 & 0{,}072 \\ 0{,}131 & 0{,}255 & 0{,}244 & 0{,}370 \\ 0{,}329 & 0{,}255 & 0{,}177 & 0{,}239 \end{bmatrix}$	$\begin{bmatrix} 0{,}50 & 0{,}25 & 0{,}25 \\ 0{,}20 & 0{,}60 & 0{,}20 \\ 0{,}25 & 0{,}25 & 0{,}50 \end{bmatrix}$	$\begin{bmatrix} 0{,}4969 & 0{,}2518 & 0{,}2513 \\ 0{,}1992 & 0{,}5994 & 0{,}2014 \\ 0{,}2432 & 0{,}2552 & 0{,}5016 \end{bmatrix}$
III	$\begin{bmatrix} 0{,}295 & 0{,}141 & 0{,}231 & 0{,}333 \\ 0{,}319 & 0{,}268 & 0{,}326 & 0{,}087 \\ 0{,}124 & 0{,}264 & 0{,}217 & 0{,}395 \\ 0{,}321 & 0{,}295 & 0{,}128 & 0{,}256 \end{bmatrix}$	$\begin{bmatrix} 0{,}34 & 0{,}33 & 0{,}33 \\ 0{,}33 & 0{,}34 & 0{,}33 \\ 0{,}33 & 0{,}33 & 0{,}34 \end{bmatrix}$	$\begin{bmatrix} 0{,}3496 & 0{,}2992 & 0{,}3511 \\ 0{,}3275 & 0{,}3184 & 0{,}3542 \\ 0{,}3283 & 0{,}3307 & 0{,}3409 \end{bmatrix}$

Selanjutnya, diberikan tiga nilai awal matriks peluang transisi Λ yang berbeda. Hal ini bertujuan untuk mengetahui besarnya peluang perpindahan dari satu segmen ke segmen yang lain. Pada tipe I, diberikan matriks Λ dengan nilai peluang yang besar untuk perpindahan ke segmen yang sama, yaitu λ_{11} = 0,75, λ_{22} = 0,8, dan λ_{33} = 0,81, serta nilai peluang yang kecil

untuk perpindahan ke segmen yang berbeda, misalnya λ_{12} = 0,15. Pada tipe II, diberikan nilai peluang yang sedang untuk perpindahan ke segmen yang sama, yaitu $\lambda_{11} = 0{,}50$, $\lambda_{22} = 0{,}60$, dan $\lambda_{33} = 0{,}50$. Sedangkan pada tipe III, diberikan nilai peluang yang kecil untuk perpindahan ke segmen yang sama. Perbedaan kombinasi nilai awal matriks Λ untuk ketiga tipe tersebut, dapat dilihat juga pada Tabel 1 di atas.

Dari tabel di atas, terlihat bahwa untuk nilai awal I, diperoleh estimasi peluang transisi segmen I ke segmen I sendiri, yaitu $\lambda_{11} = 0{,}7311$. Sedangkan estimasi peluang transisi segmen I ke segmen II, yaitu $\lambda_{12} = 0{,}1636$, dan seterusnya. Di samping itu, diketahui bahwa peluang transisi terbesar terdapat pada $\lambda_{22} = 0{,}9135$, dan peluang transisi terkecil terdapat pada $\lambda_{21} = 0{,}0211$. Nilai λ_{21} artinya peluang perpindahan ke segmen I jika segmen sebelumnya adalah segmen II sebesar 0,0211. Pada tipe II, estimasi peluang transisi terbesar dan terkecil masing-masing $\lambda_{22} = 0{,}5994$ dan $\lambda_{21} = 0{,}1992$. Sedangkan pada tipe III, diperoleh estimasi peluang transisi terbesar dan terkecil masing-masing $\lambda_{23} = 0{,}3542$ dan $\lambda_{12} = 0{,}2992$. Pendugaan matriks transisi segmen dengan tiga tipe Λ dimaksudkan untuk melihat apakah terdapat perubahan yang signifikan dari nilai peluang transisi awal. Hasil penelitian menunjukkan bahwa perubahan nilai peluang transisi awal tidak terlalu signifikan.

Estimasi Parameter Λ dan ρ

Tabel berikut merupakan hasil pendugaan parameter Λ dan ρ, jika diberikan observasi barisan DNA dan tipe segmen tersembunyi.

Tabel 2. Nilai Pendugaan Parameter Λ dan ρ.

ρ	Λ		
	Tipe I :	Tipe II :	Tipe III :
0,246	0,2201	0,2156	0,2134

Nilai $\rho = 0{,}246$ menunjukkan bahwa rata-rata peluang transisi nukleotida Adenin, Sitosin, Timin dan Guanin dalam barisan DNA sebesar 0,246. Sedangkan $\Lambda = 0{,}2201$, $\Lambda = 0{,}2156$, dan $\Lambda = 0{,}2134$ menunjukkan nilai rata-rata peluang transisi antara ketiga segmen pada tipe I, II, dan III.

4. Kesimpulan

Berdasarkan hasil penelitian pada data barisan DNA *Homo sapiens*, dapat diambil kesimpulan sebagai berikut :

1. Peluang transisi observasi barisan adalah besarnya nilai peluang perpindahan ke observasi berikutnya yang hanya bergantung pada satu observasi sebelumnya. Antara ketiga segmen, diperoleh peluang transisi terbesar, yaitu $P_{41} = 0{,}434$ terdapat pada segmen I, dan peluang transisi terkecil, yaitu $P_{24} = 0{,}072$ terdapat pada segmen II.
2. Peluang transisi segmen adalah besarnya nilai peluang perpindahan ke segmen berikutnya yang hanya bergantung pada satu segmen sebelumnya. Pada tipe I, diperoleh estimasi peluang transisi terbesar dan terkecil masing-masing $\lambda_{22} = 0{,}9135$ dan $\lambda_{21} = 0{,}0211$. Pada

tipe II, estimasi peluang transisi terbesar dan terkecil masing-masing $\lambda_{22} = 0{,}5994$ dan $\lambda_{21} = 0{,}1992$. Sedangkan pada tipe III, diperoleh estimasi peluang transisi terbesar dan terkecil masing-masing $\lambda_{23} = 0{,}3542$ dan $\lambda_{12} = 0{,}2992$.

3. Nilai dugaan parameter ρ = 0,246 merupakan nilai estimasi rata-rata peluang transisi observasi barisan DNA. Sedangkan nilai dugaan parameter terbesar dan terkecil untuk segmen tersembunyi, masing-masing Λ = 0,2201**,** dan Λ = 0,2134 merupakan nilai estimasi rata-rata peluang transisi antar segmen pada tipe I dan III.

Daftar Pustaka

Boys, R.J, Henderson, D.A dan Wilkinson, D.J. 2000. *Detecting homogeneous segments in DNA segment by using Hidden Markov Models*. University of Newcastle, Australia.

Braun, J.V dan Muller, H.G. 1998. *Statistical Mehods for DNA Sequence Segmentation.* University of California.

Congdon, P. 2003. *Applied Bayesian Modelling.* University of London, Inggris.

Fatchiyah dan Amuringtyas. 2006. *Kromosom, Gen, DNA, Sintesis Protein dan Regulasi Gen.* Universitas Brawijaya, Malang.

Johnson, R.A, Wichern, D.W. 2002. *Applied Multivariate Statistical Analysis.* United State of America.

Nur, D, Allingham, D, dan Rousseau, J. 2000. *Bayesian Hidden Markov Model for DNA Sequence Segmentation: A Prior Sensitivity Analysis.* Australia. Oxford University Press.

Raychaudhuri, S. 2005. *Computational Test Analysis for Functional Genomics and Biounformatics.* USA. Oxford University Press.