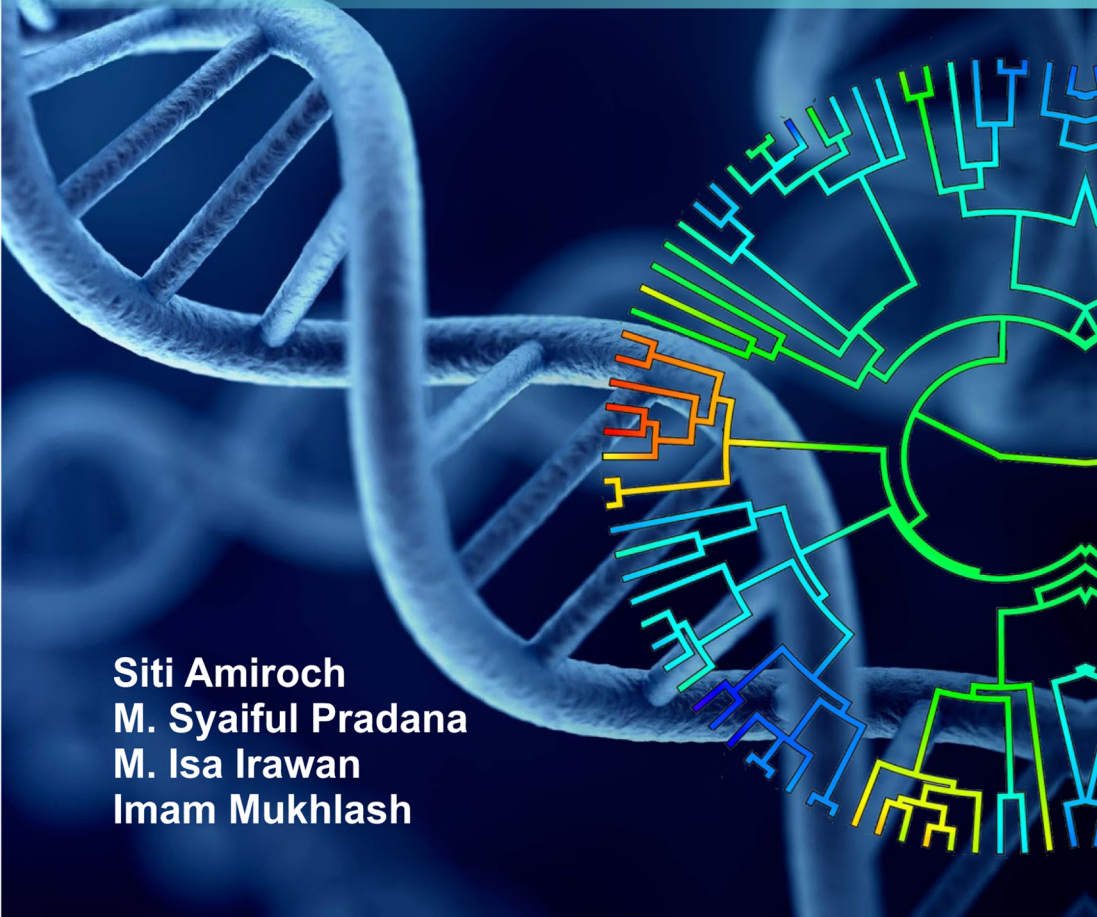
Pustaka ILALANG

Studi Kasus Penyebaran Epidemi SARS

BIOINFORMATIKA

Perspektif Matematika Pada Analisis Sekuen dan Filogenetika



Siti Amiroch
M. Syaiful Pradana
M. Isa Irawan
Imam Mukhlash

BIOINFORMATIKA

Perspektif Matematika Pada Analisis Sekuen dan Filogenetika

Oleh
Siti Amiroch
M. Syaiful Pradana
M. Isa Irawan
Imam Mukhlash

Pustaka ILALANG

BIOINFORMATIKA

Perspektif Matematika
Pada Analisis Sekuen dan Filogenetika

Penulis:
Siti Amiroch
M. Syaiful Pradana
M. Isa Irawan
Imam Mukhlash

Penyunting: **Novita Eka Chandra**
Lay out: **SixmidArt**
Desain Sampul: **M. Syaiful Pradana**

Diterbitkan Oleh:
CV. Pustaka Ilalang Group
Jalan Airlangga No. 3 Sukodadi – Lamongan
Jalan Raya Lamongan – Mantup 16 km Kedungsari
Kembangbahu – Lamongan – Jawa Timur - Indonesia
e-mail: pustaka_ilalang@yahoo.co.id
Hp. 081 330 501 724

Cetakan pertama : Desember 2018
Halaman: x + 134 hlm
Ukuran: 14 x 20,5 cm

ISBN : 978-602-6715-95-1

Daftar Isi

PROLOG

Bukan Sekadar Filogenetika, Bukan Sekadar Epidemi

Dalam epidemiologi, **epidemi** berasal dari bahasa Yunani yaitu ***epi*** *(yang artinya "pada") dan* ***demos*** (yang artinya "rakyat") adalah sebuah penyakit yang timbul sebagai kasus baru pada populasi tertentu pada manusia, dalam suatu periode waktu tertentu, dengan laju yang melampaui "ekspektasi" (dugaan), dan didasarkan pada pengalaman mutakhir. Dengan kata lain, epidemi adalah wabah yang terjadi secara lebih cepat daripada yang diduga.

Dalam hal kaitannya dengan epidemi, merupakan sesuatu yang menarik karena matematika juga bisa membahas masalah Epidemi. Hal ini bisa dipelajari pada bidang Bioinformatika, yaitu salahsatu disiplin ilmu yang menerapkan teknik komputasional untuk mengelola dan menganalisis informasi biologi yang mencakup penerapan metode matematika, statistika, dan informatika untuk memecahkan masalah-masalah biologis, terutama dengan menggunakan informasi yang terdapat pada *sequence* DNA maupun sequence protein.

Memang secara umum contoh kasus yang disajikan di buku ini "bukan sekedar epidemi", namun banyak sisi keilmuan yang bisa digali dalam kasus epidemi tersebut. Pada sebuah epidemi, bisa diulas mengenai pensejajaran sekuen, mutasi, dan pohon filogenetik yang menggambarkan sebuah epidemi bermula maupun proses penyebarannya. Semua itu bukan hal yang mustahil untuk dipelajari oleh mahasiswa matematika.

Sebagaimana buku ini yang ditujukan untuk mahasiswa matematika, terutama semester VII yang mengambil mata kuliah Bioinformatika.

Sementara itu, di bagian ini sekilas kita mengenalkan virus SARS. Yaitu virus yang kita jadikan sample baik untuk analisis sekuen maupun untuk konstruksi pohon filogenetik. SARS (*Severe Acute Respiratory Syndrome*: Sindrom pernafasan akut parah) adalah bentuk serius dari pneumonia yang disebabkan oleh virus corona. Virus SARS ini menyebabkan gangguan pernafasan akut (kesulitan bernapas berat) dan kadang-kadang kematian. SARS adalah contoh dramatis betapa cepatnya perjalanan dunia dapat menyebarkan penyakit. Hal ini juga merupakan contoh seberapa cepat sistem kesehatan terhubung dapat merespon ancaman kesehatan yang baru.

Karena SARS menyebar dengan cepat dan menginfeksi ribuan orang di seluruh dunia, termasuk Asia, Australia, Eropa, Afrika, Amerika Utara dan Selatan. Sampai-sampai WHO mengumumkan SARS sebagai ancaman kesehatan global, dan mengeluarkan *travel advisory*. Menurut data WHO, pada tahun 2003 wabah diperkirakan mencapai 8000 kasus dengan 750 angka kematian (http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0004460).

Penyebaran kasus ini sangat menarik, penyebarannya juga bisa diibaratkan sebagai sebuah pohon yaitu pohon filogenetik. Kaitannya dengan buku ini, bahwa proses pembentukan pohon filogenetik merupakan salahsatu contoh penerapan metode matematika maupun statistika pada Bioinformatika. Sebagaimana telah dibahas pada Irawan dan Amiroch (2015), Amiroch dan Rohmatullah (2017), Amiroch, dkk (2017), Andriani dan Irawan (2017) dan Amiroch, dkk (2018) tentang beberapa metode yang digunakan dalam pembentukan pohon filogenetik, diantaranya metode jarak dan metode *maximum*

likelihood yang diterapkan pada kasus penyebaran virus MERS, virus EBOLA, dan virus SARS dengan penyebaran yang begitu cepat dan menyita perhatian publik di berbagai negara. Pembentukan pohon filogenetik tersebut dimaksudkan sebagai sebuah pola yang menggambarkan sejarah evolusi maupun pola penyebaran sebuah virus di sejumlah negara.

Terkait dengan pendalaman dari metode maximum likelihood sebagaimana dibahas pada Amiroch, dkk (2018) serta mengingat kajian filogenetik dengan metode Bayes untuk penyebaran suatu virus belum banyak dilakukan, maka masalah ini menjadi hal yang sangat penting untuk dikaji dan dianalisa guna menambah wawasan dan pengembangan ilmu ke depan. Buku ini juga membahas metode Bayes pada inferensi filogenetik epidemi SARS dengan menggunakan data sequence DNA pasien penderita SARS dari berbagai negara. Tujuannya adalah agar rekam jejak penyebaran virus SARS dapat tergambar dalam sebuah pohon filogenetik, memudahkan pencatatan sejarah dan meningkatkan kewaspadaan akan bahaya serupa.

Maka dari itulah, dalam sebuah kasus penyebaran virus, bukan sekedar epidemi, bukan sekedar filogenetika.

Ucapan Terima Kasih

Buku ini lahir karena peran serta sejumlah pihak, baik langsung ataupun tidak langsung. Untuk itu, saya berterima kasih kepada: Direktorat Riset dan Pengabdian Masyarakat, Direktorat Jenderal Penguatan Riset dan Pengembangan, Kementrian Riset, Teknologi dan Pendidikan Tinggi atas Hibah Penelitian Kerjasama antar Perguruan Tinggi (PKPT) yang diberikan. Juga kepada Bapak ketua Senat Dr. M. Afif Hasbullah, S.H, ibu rektor Ainul Masruroh, S.H, M.Hi, Bapak ketua LPPM UNISDA Bapak Ir. Choirul Anam, M.P, serta ibu Novita Eka Chandra, S.Si, M.Sc. Terima kasih juga untuk tim

admin departemen Matematika ITS dan petugas laboratorium Ilmu Komputer yang ada di Departemen Matematika ITS.

Semoga buku ini bermanfaat untuk semua.

Lamongan, November 2018

Siti Amiroch
M. Syaiful Pradana
M. Isa Irawan
Imam Mukhlash

BAB I

PENYEJAJARAN SEKUEN

Penyejajaran sekuen digunakan untuk melihat homologi baik secara keseluruhan maupun parsial yang nantinya dari data tersebut dapat digunakan untuk melihat kekerabatan antarspesies. Kekerabatan antarspesies ini dapat dilihat melalui pohon filogenetik. Untuk memahami hal tersebut, diperlukan beberapa pengertian yang dijelaskan berikut ini.

1.1 Sekuen

Istilah sekuen biologi pada umumnya digunakan untuk menyatakan sekuen DNA, sekuen RNA, dan sekuen protein. Dalam pengertian biologi molekuler, sekuen biologi terdiri dari banyak makromolekul, dimana semua makromolekul memiliki fungsi-fungsi yang spesifik dalam kondisi tertentu. Makromolekul tersebut dapat dibagi kedalam sejumlah mikromolekul dengan fungsi-fungsi tertentu. Pada umumnya, sekuen DNA/RNA didasarkan pada empat nukleotida, sedangkan sekuen pada protein didasarkan pada 20 asam amino (Shen, 2007).

Banyak cara dapat dilakukan untuk merepresentasikan struktur dari sekuen biologi. Cara yang paling sering digunakan dengan mendeskripsikan sekuen tersebut kedalam bentuk struktur primer, sekunder, dan tersier (struktur tiga dimensi). Untuk sekuen protein, struktur primernya mendeskripsikan komponen-komponen nukleotida. Struktur sekundernya mendeskripsikan sifat lokal, sedangkan struktur tiga dimensi atau struktur

tersiernya mendeskripsikan susunan tiga-dimensi (posisi koordinat) dari atom konstituen dalam molekul.

Struktur sekunder dari sekuen protein menunjukkan struktur khusus (motif) dari masing-masing segmen protein, bisa berupa struktur *helix*, untai atau struktur lainnya. Super struktur sekunder juga sering digunakan untuk mendeskripsikan suatu keadaan antara struktur sekunder dan struktur tersier, yang terdiri dari sebagian besar kelompok molekul kompak (domain).

Dalam ilmu biologi molekuler modern dijelaskan bahwa sekuen DNA/RNA dan sekuen protein merupakan unit dasar yang terlibat dalam fungsi biologi khusus, sehingga bisa dikatakan bahwa sekuen biologi hanyalah kombinasi dari unit-unit dasar. Karakteristik fungsional dari sekuen tersebut tidak hanya melibatkan struktur utamanya, tetapi juga bentuk tiga dimensinya. Misalnya, suatu kantong pengikat protein memiliki peranan penting dalam mengontrol fungsinya. Dengan demikian, bentuk yang terbentuk oleh sekuen asam amino dalam ruang dimensi tiga menjadi sangat relevan untuk perawatan klinis penyakit-penyakit yang melibatkan mutasi genetik. Konfigurasi protein digunakan untuk menggantikan bentuk protein dalam ruang tiga-dimensi. Proses mutasi pada sekuen dapat mengubah konfigurasi dan mengubah fungsi. Untuk mengetahui mutasi dapat dilakukan penyejajaran antar sekuen.

Menurut Shen (2007), untuk mendeskripsikan sekuen biologi, digunakan notasi berikut:

$$X = (x_1, x_2, \ldots, x_{n_a}), \quad Y = (y_1, y_2, \ldots y_{n_b}), \quad Z = (z_1, z_2, \ldots z_{n_b})$$

dengan X, Y, Z menyatakan sekuen x_i, y_i, z_i adalah unit-unit dasar dari sekuen pada posisi ke-i, di mana

elemen-elemen tersebut diperoleh dari himpunan $V_q = \{0, 1, \ldots, q-1\}$. Panjang dari *X, Y* dan *Z* dinyatakan oleh n_x, n_y, n_z. Jika *X,Y,Z* merupakan sekuen DNA/RNA maka $V_4 = \{a, c, g, t\}$ atau $\{a, c, g, u\}$, sedangkan jika sekuen protein, maka q = 20 dan V_q = {A,R,N,D,C,Q,E,G,H,I,L ,K,M,F,P,S,T,W,Y,V}, yang secara langsung mewakili 20 molekul asam amino. Untuk jenis asam amino dan simbol standarnya tercantum dalam tabel di bawah ini:

Tabel 1.1 Jenis asam amino dan simbol standarnya

A	Alanin	**L**	Leusin
R	Arginin	**K**	Lisin
N	Asparagin	**M**	Metionin
D	Asam Aspartik	**F**	Penilalanin
C	Cistein	**P**	Prolin
Q	Glutamin	**S**	Serin
E	Asam Glutamik	**T**	Treonin
G	Glisin	**W**	Triptopan
H	Histidin	**Y**	Tirosin
I	Isoleusin	**V**	Valin

Sumber: Nello C, Computational Genomics, 2006

Simbol asam amino sebagaimana tercantum pada Tabel 2.1 merupakan suatu urutan unik yang ditentukan oleh pengkodean nukleotida gen. Kode genetik adalah kumpulan tiga nukleotida yang disebut kodon dan setiap kombinasi tiga nukleotida menunjuk asam amino, misalnya ATG (Adenin-Timin-Guanin) adalah kode untuk metionin. Kode standar genetik lainnya bisa dilihat pada tabel di bawah ini:

Tabel 1.2 Kode standar genetik

A			G		C		T	
A	AAA	K	AGA	R	ACA	T	ATA	I
	AAG	K	AGG	R	ACG	T	ATG	M
	AAC	N	AGC	S	ACC	T	ATC	I
	AAT	N	AGT	S	ACT	T	ATT	I
G	GAA	E	GGA	G	GCA	A	GTA	V
	GAG	E	GGG	G	GCG	A	GTG	V
	GAC	D	GGC	G	GCC	A	GTC	V
	GAT	D	GGT	G	GCT	A	GTT	V
C	CAA	Q	CGA	R	CCA	P	CTA	L
	CAG	Q	CGG	R	CCG	P	CTG	L
	CAC	H	CGC	R	CCC	P	CTC	L
	CAT	H	CGT	R	CCT	P	CTT	L
T	TAA	*	TGA	*	TCA	S	TTA	L
	TAG	*	TGG	W	TCG	S	TTG	L
	TAC	Y	TGC	C	TCC	S	TTC	F
	TAT	Y	TGT	C	TCT	S	TTT	F

Sumber: Nello C, Computational Genomics, 2006

Dalam penyejajaran sekuen, homologi berarti bahwa sekuen berasal dari satu nenek moyang dan tidak hanya sama secara kebetulan. Jadi setelah dilakukan penyejajaran sekuen, barulah diketahui berapa tingkat homologinya.

1.2 Penyejajaran Dua Sekuen

Dalam Shen (2006), penyejajaran sekuen merupakan metode yang penting dalam analisis posisi dan tipe mutasi yang tersembunyi didalam sekuen biologi dan memungkinkan dilakukan perbandingan yang tepat. Hal yang terpenting dalam penyejajaran sekuen adalah menentukan perpindahan mutasi. Diberikan dua sekuen

A_1 dan A_2 yang didefinisikan sebagai berikut:

$$A_1 = (a_{11}, a_{12}, \dots, a_{1n_a}) \text{dan } A_2 = (a_{21}, a_{22}, \dots, a_{2n_a}). \quad (2.1)$$

Penyisipan simbol "-" ke dalam sekuen A_1 dan A_2 bertujuan untuk membentuk dua buah sekuen yang baru yaitu sekuen A_1' dan A_2'. Selanjutnya, elemen-elemen dari sekuen A_1 dan A_2 memiliki range dari $V_5 = \{0,1,2,3,4\}$ atau $\{a, c, g, t, -\}$.Definisi dari penyejajaran ganda adalah kumpulan dari sekuen yang dinyatakan sebagai:

$$\mathcal{A} = \{A_1, A_2, \dots, A_m\}. \quad (2.2)$$

Untuk setiap A_s merupakan sekuen terpisah yang didefinisikan pada V_q, dan dinyatakan sebagai berikut:

$$A_s = (a_{s,1}, a_{s,2}, \dots, a_{s,n_s}), \qquad s = 1,2, \dots, m \quad (2.3)$$

dengan n_s adalah panjang sekuen A_s dan m adalah banyaknya sekuen pada masing-masing kelompok.

1.3 Pemrograman Dinamik

Pemrograman dinamik (DP) adalah prosedur rekursif yang membagi masalah menjadi sekumpulan sub-masalah yang saling bergantung dimana solusi intermediate berikutnya merupakan fungsi dari sub-masalah sebelumnya dan hanya tergantung pada tetangga terdekatnya. Untuk masalah penyejajaran berpasangan DP dimulai pada akhir sekuen dengan upaya untuk mencocokkan semua kemungkinan pasangan residu menurut skema penilaian untuk kecocokan, ketidakcocokan dan *gap* yang menghasilkan matriks nilai skor untuk semua kemungkinan penyejajaran antara dua sekuen. Skor tertinggi mengidentifikasi penyejajaran optimal. Matriks skor memiliki dimensi (n,

m) - di mana n dan m adalah panjang dari kedua sekuen yang dibangun dari atas ke bawah; untuk mencapai posisi (i, j) dalam langkah sebelumnya, ada tiga kemungkinan jalan: pergerakan diagonal tanpa penalti gap dari posisi (i-1, j-1); bergerak dari posisi (i-1, j) ke (i, j), dan bergerak dari posisi(i, j-1) sampai (i, j) dengan penalti gap.

Penyejajaran sebenarnya diperoleh dari matriks kedua, matriks trace-back yang menyimpan informasi gerakan melalui matriks dengan mengulangi langkah yang dilakukan untuk mendapatkan skor tertinggi (Durbin et al., 1998). Algoritma berbasis DP yang dikenal untuk penyejajaran berpasangan adalah algoritma *Needleman Wunchs* dan algoritma *Smith Waterman*. Keduanya merupakan algoritma penyejajaran global dan penyejajaran lokal.

1.4 Algoritma Needleman Wunchs

Algoritma *Needleman Wunchs* merupakan algoritma penyejajaran global untuk sekuen yang berpasangan. Langkah-langkah pada algoritma ini sebagai berikut: (Shen, 2006).

1. Buatlah tabel dua dimensi dari dua sekuen.
   Jika diberikan sekuen $A = (a_1, a_2, \dots, a_n)$, dan $B = (b_1, b_2, \dots, b_n)$, maka tabel dari dua sekuen tersebut terdapat pada Tabel 1.3 dengan sebagai berikut:

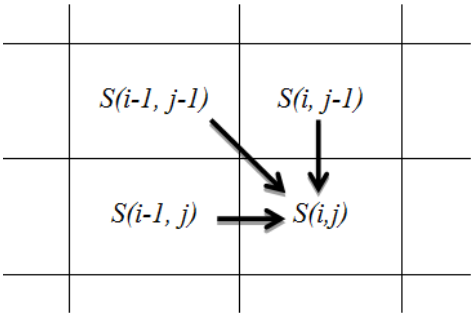
Tabel 1.3. Tabel Dua Dimensi Sekuen A,B

	a_1	a_2		.....	a_n
	s(0,0)	s(1,0)	s(2,0)	.....	s(n,0)
b_1	s(0,1)	s(1,1)	s(2,1)	.....	s(n,1)
b_2	s(0,2)	s(1,2)	s(2,2)	.....	s(n,2)
.....	.....	.....	.....		.....
b_m	s(0,m)	s(1,m)	s(2,m)	.....	s(n,m)

2. Hitung elemen *s(i,j)* dari tabel dua dimensi
   Masing-masing elemen *s(i,j)* dari tabel dua dimensi ditentukan oleh tiga elemen; *s(i-1,j-1)* pada pojok kiri atas, *s(i-1,j)* pada sisi kiri dan *s(i, j-1)* pada bagian atas. Pertama-tama tentukan skor marjinal *s(i,0)* dan *s(0,j)*. Sedangan skor virtual simbol adalah *d*, dimana virtual simbol merupakan panjang dari virtual symbol itu sendiri. Sehingga *s(0,j) = -jxd, s(i,0)= -ixd*, dan *s(0,0)=0*.

   Maka elemen dapat dihitung menggunakan rumus berikut:

$$s(i,j) = \max\{(i-1,j-1) + s(a_i, b_j),\ \ s(i-1,j) - d, s(i,j-1) - d\}.$$



Gambar 1.1. *Perhitungan s(i,j)*

3. Algoritma *Traceback*
   Kemudian, untuk elemen *s(i,j)* pada alur mundur adalah:
   - Notasikan pasangan dari sekuen sebagai a_i, b_i jika alur mundurnya dimulai dari a_i, b_i ke sudut kiri atas.
   - Sisipkan suatu virtual simbol pada sekuen vertikal dan menotasikannya sebagai $(a_i, -)$ jika alur mundurnya horisontal.
   - Sisipkan suatu virtual simbol pada sekuen horisontal dan menotasikannya sebagai $(-, a_i)$ jika alur mundurnya vertikal.

- Terakhir, diperoleh suatu penyejajaran optimal dari dua sekuen.

Contoh:
Misalkan sekuen A= aaattagc, dan B=gtatatact. Kita akan gunakan algoritma berbasis pemrograman dinamik untuk memperoleh penyejajaran. Jika ditentukan skor 5 untuk nucleotide yang cocok, -3 bila tidak cocok, dan -7 untuk simbol virtual. Maka:

1. Buatlah tabel dua dimensi dan hitung nilai masing-masing elemen. Nilai pada elemen baris pertama didefinisikan *s(i,0) =-ixd*, sedangkan nilai pada elemen kolom pertama didefinisikan *s(0,j) = -jxd.* Contoh, s(1,1) = max(0-3, -7, -7, -7, -7) = -3 dan alur mundurnya (1,1)→(0,0).
2. *Traceback*: kita mulai alur mundur dari s(8,9). Nilai s(8,9)=1 diperoleh dari elemen sebelah kiri atas s(7,8). Penjelasannya, s(8,9) = s(7,8) + s(c,t) = 4-3 = 1. Jadi alur mundur pertama kali dari (8,9)→(7,8). Hal ini diulang terus sampai jalur mundur mencapai s(0,0).

Tabel 1.4. Tabel dua dimensi sekuen A,B

		a	a	a	t	t	a	g	c
	0	-7	-14	-21	-28	-35	-42	-49	-56
g	-7	-3	-10	-17	-24	-31	-38	-37	-44
t	-14	-10	-6	-13	-12	-19	-26	-33	-40
a	-21	-9	-5	-1	-8	-15	-14	-21	-28
t	-28	-16	-12	-8	4	-3	-10	-17	-24
a	-35	-23	-11	-7	-3	1	2	-5	-12
t	-42	-30	-18	-14	-2	2	-2	-1	-8
a	-49	-37	-25	-13	-9	-5	7	0	-4
c	-56	-44	-32	-20	-16	-12	0	4	5
t	-63	-51	-39	-27	-15	-11	-7	-3	1

		a	a	a	t	t	a	g	c
	0	-7	-14	-21	-28	-35	-42	-49	-56
g	-7	**-3**	-10	-17	-24	-31	-38	-37	-44
t	-14	-10	**-6**	-13	-12	-19	-26	-33	-40
a	-21	-9	-5	**-1**	-8	-15	-14	-21	-28
t	-28	-16	-12	-8	**4**	-3	-10	-17	-24
a	-35	-23	-11	-7	**-3**	1	2	-5	-12
t	-42	-30	-18	-14	-2	**2**	-2	-1	-8
a	-49	-37	-25	-13	-9	-5	7	0	-4
c	-56	-44	-32	-20	-16	-12	0	**4**	5
t	-63	-51	-39	-27	-15	-11	-7	-3	**1**

Alur mundurnya adalah (8,9) → (7,8) → (6,7) → (5,6) → (4,5) → (4,4) → (3,3) → (2,2) → (1,1) → (0,0). Sehingga dengan mengikuti alur mundur tersebut, diperoleh hasil penyejajaran seperti dibawah ini:

A' = (aaat-tagc)

B' = (gtatatact)

1.5 Algoritma Smith Waterman

Algoritma Smith Waterman merupakan jenis Algoritma *alignment* lokal. Meskipun terlihat sederhana untuk pengembangan algoritma yang berbasis program dinamik dengan *alignment* lokal yang sesuai, algoritma ini sangat berperan dalam bioinformatika. *Software* yang sangat terkenal dalam bioinformatika, BLAST, dikembangkan berdasarkan algoritma ini. Dua aspek penting dalam algoritma Smith Waterman antara lain:

a. Menghitung nilai pada tabel dua dimensi

Algoritma Smith Waterman menambahkan 0 ketika menghitung $s(i,j)$ sehingga skor negatif tidak akan pernah terjadi pada Algoritma ini. Keuntungannya akan memperjelas lintasan *backward.*

$$s(i,j) = \max \begin{cases} 0 \\ s(i-1,j-1) + s(x_i, y_i) \\ s(i-1,j) - d \\ s(i,j-1) - d \end{cases}$$

b. Algoritma ***Traceback***

Titik awal dan akhir dari metode *backtrace* pada Algoritma Smith Waterman dipilih elemen dengan skor maksimal. Titik akhirnya adalah elemen pertama dengan nilai 0 pada proses *backtrace*. Titik awal dengan skor maksimal akan menjamin skor maksimal pada *alignment sequence* lokal, dan titik akhirnya adalah elemen pertama dengan nilai 0 menjamin bahwa bagian tersebut tidak terlampaui. Bagian yang berhubungan dengan lintasan backward merupakan bagian yang memiliki skor penalti minimum.

1.6 Penyejajaran Ganda

Penyejajaran Ganda (*Multiple Alignment*) adalah metode untuk mensejajarkan tiga atau lebih sekuen secara bersamaan. Tentu saja yang dibandingkan adalah sequence yang sama, yaitu sekuen nukleotida dibandingkan dengan sekuen nukleotida lainnya, sedangkan sekuen asam amino dibandingkan dengan sekuen asam amino lainnya. Hal ini dilakukan dengan harapan terdapat kemiripan atau kesamaan lebih lanjut antara sekuen yang dibandingkan, karena multiple alignment digunakan untuk membandingkan sekuen yang homolog dengan tujuan menentukan hubungan antara sekuen yang bermutasi.

1.6.1 *Progressive Alignment*

Metode penyejajaran progresif bersifat heuristik yang menghasilkan penyejajaran sekuen ganda dari sejumlah penyejajaran berpasangan. Skema umum metode tersebut adalah: pertama, dua sekuen (sekuen 1 dan sekuen 2) dipilih dan disejajarkan, lalu sekuen ketiga dipilih dan disejajarkan dengan sekuen pertama. Demikian proses berlanjut sampai semua sekuen habis (Isaev, 2006).

CLUSTALW (Thompson et al, 1994; Cenna et al., 2003) adalah program penyejajaran sekuen ganda yang sangat terkenal. Program tersebut menggunakan metode penyejajaran progresif. Adapun proses penyejajarannya adalah:

1. Penyejajaran sekuen dilakukan dengan menggunakan pemrograman dinamik.
2. Skor dari penyejajaran berpasangan digunakan untuk membentuk matriks jarak dari jarak genetik, sedangkan untuk membentuk pohon filogenetik digunakan metode *Neighbor Joining* (Saitou and Nei, 1987).
3. Pemrograman dinamik digunakan untuk mensejajarkan sekuen dari hubungan yang terdekat atau jarak yang terdekat dari pohon.

1.6.2 Matriks Penalti

Matriks penalti didefinisikan dalam Shen (2006), misal *C* adalah matriks alignment yang diinduksi oleh sekuen ganda *A*. Jika fungsi penalti *W=w(a,b)* terdefinisi pada V_5, maka untuk sebarang $s, t \in M$, diperoleh dua ekspansi $C_{s,t}$ dan $C_{t,s}$ berdasarkan pada pasangan sekuen A_s dan A_t. Skor penalti untuk

pasangan $C_{s,t}$ dan $C_{t,s}$ didefinisikan oleh:

$$w_{s,t}(\bar{C}) = w(C_{s,t}, C_{t,s}) = \sum_{j=1}^{n_{s,t}} w(C_{s,t;j}, C_{t,s;j}) \qquad (2.4)$$

dengan matrıks

$$\bar{W}(\bar{C}) = [W_{s,t}(\bar{C})]_{s,t=1,2,\ldots,m} \qquad (2.5)$$

adalah matriks penalti yang diinduksi oleh penyejajaran berpasangan dari sekuen ganda A dan disederhanakan sebagai **matriks penalti**.

1.6.3 Analisis Penyejajaran Ganda

Di antara berbagai pohon topologi yang dihasilkan oleh output penyejajaran ganda (MA) adalah graf dan pohon yang digunakan untuk mengekspresikan hubungan antara mutasi dan evolusi. Sebuah sistem jaringan yang dihasilkan oleh mutasi dari sekuen ganda digunakan untuk menjelaskan struktur mutasi dari hasil MA melalui graf busur berwarna. Menurut Shen (2006), sistem jaringan yang dihasilkan oleh *output* MA adalah:

a. Sistem jaringan topologi
   Pada sistem ini dihasilkan output MA: $G(W) = \{M, V, W\}$, di mana W adalah fungsi penalti dari output MA (A') didefinisikan oleh $W = (w_{s,t})$ berdasarkan A'.

b. Sistem jaringan daerah mutasi
   Pada sistem ini divisualisasi pohon filogenetik yang diperoleh dari perhitungan jarak minimum pohon G_1 yang terbentuk.

c. Sistem jaringan mode mutasi
   Pada sistem ini, berdasarkan jarak minimum

pohon G_1 dan daerah mutasi matrix Δ, dibentuklah jaringan dekomposisi orthogonal mode mutasi.

1.7 Studi Kasus: Epidemi SARS

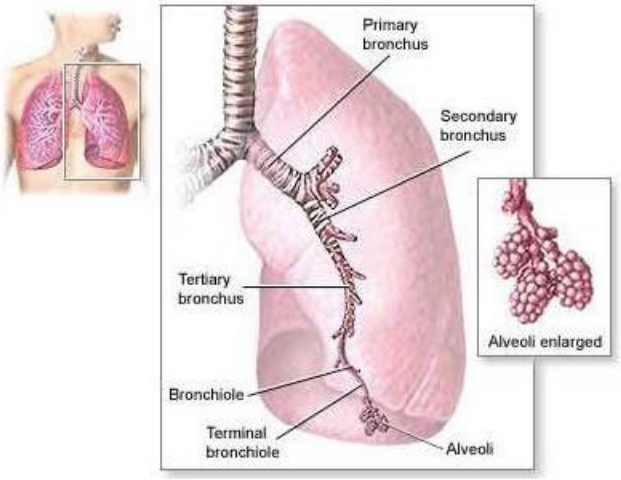
Sindrom pernapasan akut parah (*Severe Acute Respiratory Syndrome: SARS*) adalah bentuk serius dari pneumonia yang disebabkan oleh virus corona. Virus SARS ini menyebabkan gangguan pernapasan akut (kesulitan bernapas berat) dan kadang-kadang kematian. SARS adalah contoh dramatis betapa cepatnya dunia dapat menyebarkan penyakit. Hal ini juga merupakan contoh seberapa cepat sistem kesehatan terhubung dapat merespon ancaman kesehatan yang baru.

Identifikasi SARS bermula pada tahun 2003 ketika seorang dokter dari WHO, Dr Carlo Urbani mendiagnosis penyakit tersebut pada seorang pengusaha yang telah melakukan perjalanan dari provinsi Guangdong China, melalui Hong Kong, ke Hanoi Vietnam. Selang beberapa waktu kemudian pasien dan dokter yang menanganinya meninggal akibat penyakit tersebut.

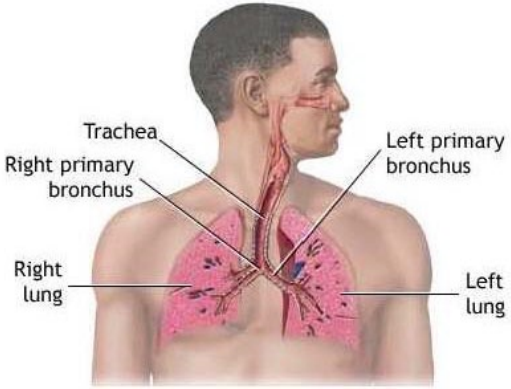
Karena SARS menyebar dengan cepat dan menginfeksi ribuan orang di seluruh dunia, termasuk Asia, Australia, Eropa, Afrika, Amerika Utara dan Selatan. Sampai-sampai WHO mengumumkan SARS sebagai ancaman kesehatan global, dan mengeluarkan *travel advisory*. Menurut data WHO, pada tahun 2003 wabah diperkirakan mencapai 8000 kasus dengan 750 angka kematian (http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0004460).

Ketika seseorang yang terinfeksi SARS batuk atau bersin, virus yang menyebar ke udara dapat terhirup atau terkena partikel lain sehingga menular ke individu kontak. Virus SARS ini dapat hidup di tangan, jaringan, dan permukaan lainnya hingga 6 jam dan bertahan sampai

3 jam setelah tetesan dikeringkan. Bahkan pada tinja orang yang terinfeksi SARS ditunjukkan virus bertahan hidup sampai 4 hari. Virus ini mungkin dapat hidup selama berbulan-bulan atau bertahun-tahun ketika suhu di bawah titik beku.



Gambar 1.3 Paru-paru manusia
Sumber: http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0004460



Gambar 1.4 Sistem pernapasan
Udara dihirup melalui hidung, perjalanan melalui trakea dan bronkus ke paru-paru.
Sumber: http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0004460

Gejala biasanya terjadi sekitar 2 sampai 10 hari setelah individu terkontak dengan virus. Gejala SARS menyerupai gejala penyakit influenza. Pasien penderita SARS biasanya mengalami demam yang tinggi (di atas 38°C atau lebih), dan kadang-kadang disertai bercak merah, rigors, kepala pusing, limbung, nyeri otot atau bahkan diare (beberapa pasien mengalami kesulitan pernapasan). Setelah beberapa hari, gejala awal tersebut diikuti dengan infeksi saluran pernapasan yang ringan, termasuk batuk tanpa dahak (sputum) dan kesulitan bernapas. Pada sekitar 10% pasien, penyakit SARS dapat menyebabkan terganggunya pernapasan yang memerlukan perawatan medis intensif. Gejala SARS tidak sama kondisinya pada pasien lanjut usia. Masa Inkubasi biasanya terjadi sekitar 2 sampai 10 hari setelah kontak dengan virus.

Kasus SARS ini sangat menarik, yang pertama karena identifikasi host (barrier, pembawa virus) epidemi tersebut yang semula disinyalir dari beberapa binatang yang dicurigai. Yang kedua karena penyebarannya begitu cepat dari satu negara ke negara lain, bisa diibaratkan dalam sebuah pohon, yaitu pohon filogenetik yang nantinya kita bahas di Bab selanjutnya. Sedangkan pada bagian ini dicontohkan proses penyejajaran dari masing-masing sekuen DNA dari sampel pasien yang terinfeksi virus SARS.

1.8 Latihan Soal

1. Pertimbangkan skema penskoran berikut untuk sekuen DNA:
   - Jika $a = b$ skor = 1
   - Jika $a \neq b$ skor = -1
   - Jika ada *gap* skor = -2

   Untuk semua $a,b \in Q = \{A, C, G, T\}$.

   Misal $x = ACTGTCCA$ dan $y = CTGAATCAGA$. Carilah skor dari penyejajaran berikut:

   $x = ACTG - - - T - CCA$

   y = C- TGAAT – CAGA

   Apakah ada skor yang lebih tinggi antara x dan y?

2. Pertimbangkan skema penskoran dengan model *gap* linier untuk sekuen DNA:
   - Jika a = b skor = 5
   - Jika a ≠ b skor = -4
   - Jika ada *gap* skor = -2

   Gunakan algoritma pemrograman dinamik yang sesuai dengan skema penskoran tersebut untuk menemukan:

   (i) Semua penyejajaran global optimal dari sekuen

   x = A*AGTTCGT*

   y = *CAGT*AAT,

   Semua penyejajaran lokal optimal dari sekuen

   x = AGTGGCATT

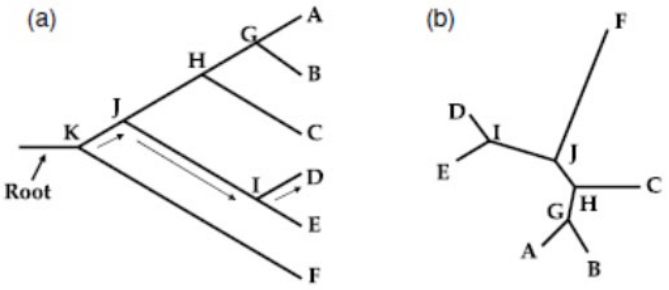
   y = TGTCGCAT.

BAB II

POHON FILOGENETIK

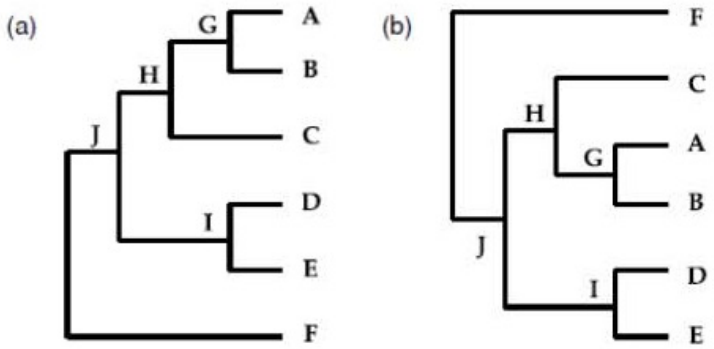
Pohon filogenetik adalah diagram yang digunakan untuk menggambarkan hubungan evolusi antara gen dan organisme dalam suatu hubungan kekerabatan yang erat. Disebut pohon filogenetik karena bentuknya menyerupai struktur pohon, sedangkan istilah yang digunakan pada diagram ini merujuk ke berbagai bagian dari pohon (yaitu akar, cabang, *node*, dan daun). *External node* atau daun merepresentasikan taksa dan disebut *Operational Taxonomic Units* (OTUs), istilah tersebut juga mewakili berbagai jenis taksa yang sebanding, misalnya sebuah keluarga organisme, individu, atau strain virus dari satu spesies atau dari spesies yang berbeda. Node internal dapat disebut *Hipothetical Taxonomic Units* (HTU) untuk menekankan bahwa mereka adalah leluhur hipotetis OTUs. Sekelompok taksa yang berbagi cabang yang sama memiliki asal monofiletik dan disebut sebuah cluster.

Dalam Gambar 2.1, taksa A, B, dan C membentuk cluster, memiliki leluhur bersama H, karena asalnya monofiletik. C, D , dan E tidak membentuk cluster tanpa memasukkan strain tambahan dan tidak berasal dari monofiletik, maka disebut *paraphyletic*. Percabangan pola tersebut dinamakan topologi pohon (Lemey et al, 2009).



Gambar 2.1 (a) pohon filogenetik *rooted* dan (b) *unrooted*.

Pada Gambar 2.1, kedua pohon memiliki topologi yang sama. Pada pohon berakar (rooted tree) A, B, C, D, E, dan F merupakan node eksternal atau OTU. Sedangkan G, H, I, J, dan K adalah node internal atau HTU, dengan K sebagai simpul akar. Panah menunjukkan arah evolusi (misalnya dari akar K ke node eksternal D). Sedangkan Unrooted tree tidak memiliki simpul akar, hanya garis antara node cabang. Sebuah unrooted tree hanya memposisikan sekelompok individu tanpa menunjukkan arah proses evolusi. Dalam sebuah unrooted tree, tidak ada indikasi yang mewakili nenek moyang dari semua OTU.



Gambar 2.2 Struktur dari pohon filogenetik berakar

Pada Gambar 2.2 menunjukkan pohon yang sama seperti pada Gambar 2.1, namun dalam bentuk yang berbeda. Cabang di internal node dapat diputar tanpa mengubah topologi pohon. Kedua pohon pada gambar diatas memiliki topologi yang identik.

Dalam satu kumpulan organisme diharapkan setiap gen yang terbagi akan menunjukkan diagram pohon yang sama atau sangat mirip. Tiap-tiap gen mungkin bermutasi dan berkembang pada tingkat yang berbeda, tetapi semua gen akan diwariskan sebagai sebuah kelompok dan akan diteruskan pada keturunannya secara bersama-sama, sehingga pada diagram pohon yang sama, rekombinasi antara sekuen-sekuen dalam suatu spesies menyebabkan dua gen (atau bagian-bagian berbeda dari gen yang sama) mempunyai sejarah yang berbeda. Gen yang berbeda berkembang pada tingkat yang berbeda, dan spesies yang berbeda memiliki tingkat mutasi yang berbeda pula. Akibatnya, meskipun semua taksa eksternal pada diagram pohon adalah jarak yang sama dari node akar, mutasi berarti bahwa kenodean diagram pohon bisa mencakup beberapa cabang yang sangat panjang dan beberapa cabang yang sangat pendek. Semua taksa eksternal belum tentu sama jarak genetiknya diukur dari akar (Christianini, 2006).

Ada banyak metode untuk membangun pohon filogenetik. Metode-metode tersebut adalah sebagai berikut:

1. Metode berbasis jarak (misalnya, *Neighbor Joining*). Setiap hasil penyejajaran dapat digunakan untuk menghitung matriks jarak antar sekuen. Berdasarkan pada matriks jarak, akan dapat dihasilkan pohon filogenetik yang sesuai. Metode yang paling populer disebut UPGMA dan *Neighbor Joining*.

   UPGMA (*Unweighted Pair Group Method with Aritmatic*) digunakan untuk membangun pohon filogenetik dengan

cara yang mirip dengan metode *clustering*, perbedaan utamanya adalah formula yang digunakan untuk menghitung jarak kelas. Jika banyaknya sekuen dalam dua kelas berbeda, maka harus menghitung jarak dari *cluster* baru untuk semua cluster lain sebagai rata-rata bobot jarak dari komponennya.

2. Metode berbasis fitur (misalnya, metode *Maximum Parsimony*). Metode jenis ini menggunakan fitur (karakteristik) dari output penyejajaran untuk membangun pohon filogenetik.
3. Metode berbasis probabilitas (misalnya, metode *Maximum Likelihood* dan metode Bayes). Penggunaan metode ini untuk membangun pohon filogenetik dimulai dengan membangun suatu model probabilitas untuk mutasi sekuen, kemudian membangun pohon filogenetik didasarkan pada output dan model probabilitas.

Masing-masing penjelasannya nanti akan dipaparkan lebih lanjut di BAB III, BAB IV, dan BAB V.

BAB III

PEMBENTUKAN POHON FILOGENETIK DENGAN METODE JARAK

Metode jarak adalah salah satu metode pembentukan pohon filogenetik dari sekumpulan jarak antar setiap pasang sekuen yang telah disejajarkan. Sebelum pembahasan lebih lanjut, berikut diuraikan tentang matriks jarak, model evolutioner dan algoritma yang berkaitan.

3.1 Matriks Jarak

Sekumpulan jarak yang dituliskan dalam bentuk matriks disebut matriks jarak (Isaev A, 2006). Contoh:

M_d	x^1	x^2	x^3	x^4	x^5
x^1	0	8	3	14	10
x^2	8	0	9	10	6
x^3	3	9	0	15	11
x^4	14	10	15	0	10
x^5	10	6	11	10	0

Gambar 3.1. Matriks jarak 5 OTU

Gambar 3.1 menunjukkan matriks jarak dari lima sekuen (OTU) dengan himpunan sekuen {x^1, x^2, x^3, x^4, x^5} berasal dari lima spesies virus yang berbeda. Setiap elemen matriks tersebut merepresentasikan jarak genetik antar sekuen yang terlibat. Misalnya, jarak antara OTU x^2 dan x^3 adalah 9, artinya perbedaan genetik sekuen x^2dan x^3

sebesar 9 satuan. Perbedaan tersebut terjadi karena proses evolusi yang terjadi di dalam struktur genetiknya. Angka-angka tersebut dapat dikatakan sebagai waktu evolusi atau perbedaan banyaknya gen akibat evolusi.

Secara formal, suatu matriks jarak dibentuk berdasarkan *distance function* yang didefinisikan sebagai berikut:

Definisi 3.1 Misalkan M adalah sebuah himpunan dan $d: MxM \rightarrow R$ adalah sebuah fungsi, d dikatakan sebagai fungsi jarak pada M jika:

(i) $d(u,v) > 0$ untuk setiap $u, v \in M, u \neq v$,

(ii) $d(u,u) = 0$ untuk setiap $u \in M$,

(iii) $d(u,v) = d(v,u)$ untuk setiap $u, v \in M$,

(iv) memenuhi ketaksamaan segitiga $d(u,v) \leq d(u,w) + d(w,v)$ untuk setiap $u, v, w \in M$

Jika d adalah fungsi jarak pada M, maka untuk $u, v \in M$,bilangan $d(u,v) > 0$ disebut sebagai jarak antara u dan v. Himpunan yang akan dipakai di sini adalah himpunan terhingga $M = \{x1, x2, \ldots., xN\}$ yang merupakan himpunan sekuen (OTU) yang akan dibentuk pohon filogenetiknya. Diasumsikan bahwa fungsi jarak d terdefinisi di M dan d relevan secara biologi, maksudnya adalah d sesuai dengan informasi genetik yang ada pada sekuen (OTU) di M. Sebagai contoh $d(x1, x2) > d(x3, x4)$ berarti OTU $x1$ dengan $x2$ lebih jauh hubungan evolusi atau kekerabatannya dibanding OTU $x3$ dengan $x4$. Untuk menyederhanakan penulisan, $d(xi, xj)$ ditulis sebagai d_{ij} dengan $i, j \in \{1,2, \ldots., N\}$. Berdasarkan fungsi jarak tersebut dapat diperoleh matriks jarak (*distance matrix*), $M_d = (d_{ij})$ dengan definisi formal berikut.

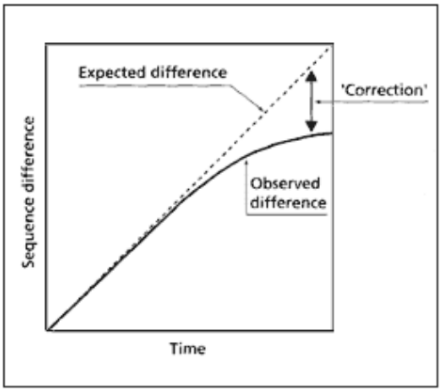
Definisi 3.2 Misalkan *d* adalah suatu fungsi jarak, *Md* disebut sebagai matriks jarak yang didefinisikan oleh:

$$Md = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1j} & \cdots & d_{1N} \\ d_{21} & d_{22} & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ d_{i1} & \ddots & \ddots & d_{ij} & \ddots & d_{iN} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{Nj} & \cdots & d_{NN} \end{pmatrix}$$

dengan $i,j = 1,2,3,.....N$ dan N adalah jumlah OTU yang terlibat (Isaev, 2006).

3.2 Model Evolutioner Jukes Cantor

Mengingat bahwa pada jarak genetik yang diamati mungkin tidak memperhatikan jumlah sebenarnya dari perubahan evolusioner, telah banyak penelitian yang mengembangkan metode untuk mengubah jarak tersebut menjadi jarak evolusi yang sebenarnya. Teknik ini sering disebut *metode koreksi jarak*; tujuannya adalah untuk 'mengkoreksi' jarak yang diamati dengan memperkirakan jumlah perubahan evolusioner yang telah terjadi.



Gambar 3.2 Koreksi jarak
Sumber: Molecular Evolution, Page and Holmes

Dibutuhkan koreksi untuk memperbaiki perbedaan sekuen. Tingkat perbedaan antara dua sekuen yang diamati tidak linier terhadap waktu tetapi melengkung karena beberapa hits. Tujuan dari koreksi metode jarak adalah untuk memulihkan jumlah perubahan evolutioner beberapa hits yang terjadi dan untuk 'mengkoreksi' jarak untuk hits yang tidak teramati. Sehingga, metode ini bertujuan untuk 'meluruskan' garis yang merepresentasikan perbedaan yang diamati.

Sebagian besar metode untuk 'mengoreksi' jarak saling terkait, hanya berbeda dalam cara memasukkan parameter. Sebagai contoh, beberapa metode memungkinkan untuk variasi frekuensi nukleotida; metode yang lebih kompleks memungkinkan berbagai jenis substitusi terjadi dengan probabilitas yg berbeda, sementara yang lain memperhitungkan variasi banyaknya tingkat substitusi antara posisi. Namun secara umum dapat digunakan kerangka umum tunggal untuk menunjukkan bagaimana model ini saling terkait. Dalam kerangka ini, probabilitas substitusi nukleotida yang diberikan tetap konstan dari waktu ke waktu, dan komposisi dasar sekuen berada dalam kesetimbangan. Pada penelitian ini digunakan model evolutioner Jukes Cantor sebagai koreksi pada metode jarak.

Pada tahun 1969 Thomas Jukes dan Cantor mengusulkan sebuah model probabilistik untuk memperbaiki jumlah perbedaan yang diamati untuk memperhitungkan kemungkinan beberapa substitusi. Model Jukes Cantor ini merupakan salah satu model evolusi sekuen yang paling sederhana yang mengasumsikan bahwa empat nucleotide memiliki frekuensi yang sama, dan bahwa semua substitusi sama-sama mungkin. Pada model ini jarak antara dua sekuen nukleotid diberikan oleh:

$$d = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right)$$

dengan p adalah proporsi nucleotida yang berbeda pada dua sekuen, sedangkan untuk sekuen protein, model yang digunakan adalah:

$$d = -\frac{19}{20}\ln\left(1-\frac{20}{19}p\right).$$

(http:/evolution-textbook.org/content/free/contens/ch27.html)

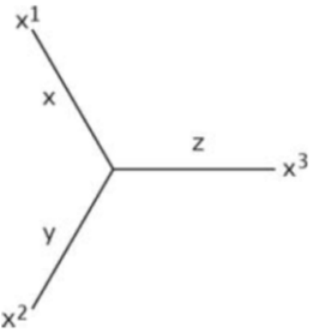
3.3 Algoritma *Neighbor Joining*

Metode *Neighbor-Joining* adalah sebuah metode berbasis jarak yang digunakan untuk membangun pohon filogenetik (Saitou N dalam Amiroch, S dan Rohmatullah A, 2017). Algoritma ini membutuhkan *input* berupa matriks jarak, dimana matriks jarak diperoleh dengan mensejajarkan masing-masing *sekuen* untuk mencari *similaritasnya*, sedangkan jarak adalah *dissimilaritas* dari *sekuen* yang dimaksud.

Sebelum membahas algoritma ini secara umum, perhatikan ilustrasi berikut. Diasumsikan jumlah OTU yang terlibat adalah 3 (tiga) atau N=3. Akan dicari tiga buah bilangan positif *x,y,z* sehingga memenuhi

$$\begin{aligned} x + y &= d_{12} \\ x + z &= d_{13} \\ y + z &= d_{23} \end{aligned} \qquad (3.1)$$

di mana hanya ada satu pohon yang bersesuaian, sebagaimana pada Gambar 3.2.



Gambar 3.3 Pohon untuk 3 sekuen (OTU)

Dengan melakukan substitusi dan eliminasi pada persamaan (3.1), diperoleh solusi dari sistem persamaan tersebut, yaitu:

$$x = \frac{1}{2}(d_{12} + d_{13} - d_{23})$$
$$y = \frac{1}{2}(d_{12} + d_{23} - d_{13}) \qquad (3.2)$$
$$z = \frac{1}{2}(d_{13} + d_{23} - d_{12})$$

Nilai *x, y, z* tersebut merepresentasikan panjang *edge* pada Gambar 3.3, sehingga untuk N = 3, diperoleh pohon dengan panjang *edge* seperti pada persamaan (3.2). (Isaev, 2006)

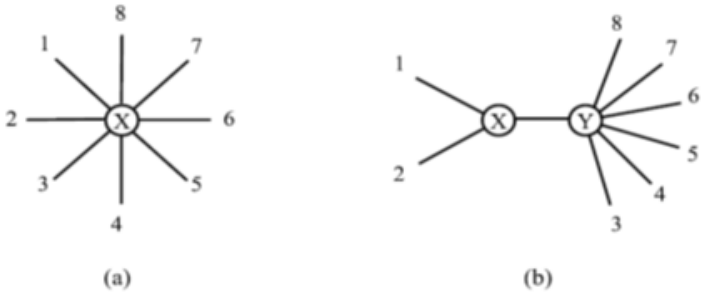
Metode *Neighbor Joining* dimulai dari *starlike structure*, dan mengumpulkan semua "*neighbors*" bersama untuk membentuk sebuah pohon tanpa akar sebagai *output*. Untuk himpunan N sekuen, langkah-langkah komputasi diberikan sebagai berikut:

1. Tentukan matriks jarak N sekuen.
2. Asumsikan sebuah pohon dengan semua OTU dalam matriks sebagai percabangan dari titik pusat, kemudian

bentuk dalam pola *star-like* seperti dalam representasi skematik pada Gambar 3.4a.

3. Untuk masing-masing OTU, hitunglah S, dimana S adalah penjumlahan dari jarak (D) antara OTU satu dengan OTU yang lain, dibagi (N-2), dengan N adalah total banyaknya OTU.

$$S_i = \frac{1}{N-2}\sum_{k=1}^{N} D_{ik}$$



Gambar 3.4a, b. *Neighbor-Joining*. a. Struktur awal *starlike,* b. struktur pohonlike

4. Identifikasi pasangan OTU dengan nilai minimum:

$$M_{ij} = D_{ij} - S_i - S_j.$$

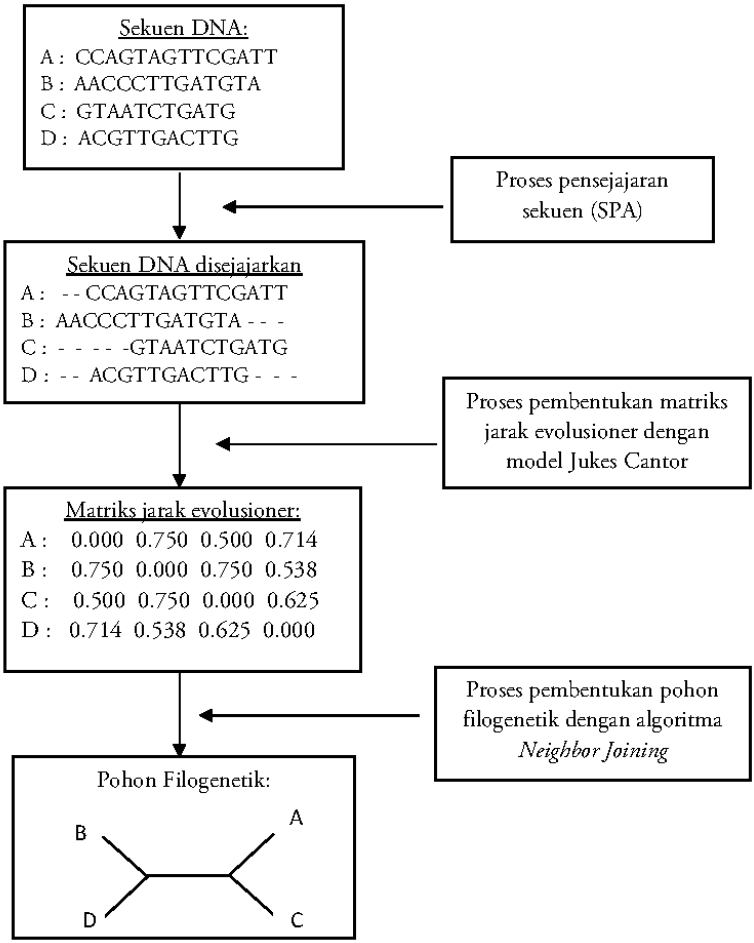
5. Gabungkan dua taksa pada suatu node dalam sebuah subpohon.
6. Hitung panjang cabang:

$$D_{Xi} = \frac{(D_{ij} + S_i - S_j)}{2}, \qquad D_{Xj} = \frac{(D_{ij} + S_j - S_i)}{2}.$$

7. Hitung jarak matriks yang baru dengan menghubungkan *i* dan *j* dan menggantinya dengan *node* (*x*) yang menghubungkannya.

$$D_{xk} = \frac{D_{ik} + D_{jk} - D_{ij}}{2}$$

8. Ulangi langkah di atas!



Gambar 3.5 Alur penyejajaran sekuen sampai terbentuknya pohon filogenetik

3.4 Studi Kasus: Pembentukan Pohon Filogenetik Untuk Menentukan Host Dari Virus SARS

Pada kasus ini, setelah dilakukan penyejajaran dari 12 sekuen yang dituliskan dalam sebuah matriks jarak dan dikonversi ke dalam matriks jarak evolutioner Jukes Cantor, diperoleh hasil sebagai berikut :

d_new =

Columns 1 through 9

0	0.1191	2.2059	1.1882	1.3609	1.4032	1.4052	1.4113	0.4398
0.1191	0	1.2006	1.1920	1.4438	1.3995	1.4540	1.3791	0.4527
2.2059	1.2006	0	0.1314	1.4012	1.3872	1.4090	1.3786	1.2622
1.1882	1.1920	0.1314	0	1.3923	1.3914	1.4332	1.3583	1.2430
1.3609	1.4438	1.4012	1.3923	0	0.7985	0.8363	1.3257	1.3898
1.4032	1.3995	1.3872	1.3914	0.7985	0	0.8688	1.3699	1.3702
1.4052	1.4540	1.4090	1.4332	0.8363	0.8688	0	1.3079	1.4290
1.4113	1.3791	1.3786	1.3583	1.3257	1.3699	1.3079	0	1.3258
0.4398	0.4527	1.2622	1.2430	1.3898	1.3702	1.4290	1.3258	0
0.4351	0.4514	1.1996	1.2013	1.4715	1.3762	1.4012	1.3464	0.1947
0.4363	0.4502	1.1971	1.1988	1.4675	1.3762	1.3974	1.3464	0.1938
0.4568	0.4596	1.2188	1.2127	1.3869	1.4155	1.3700	1.3405	0.2115

Columns 10 through 12

0.4351	0.4363	0.4568
0.4514	0.4502	0.4596
1.1996	1.1971	1.2188
1.2013	1.1988	1.2127
1.4715	1.4675	1.3869
1.3762	1.3762	1.4155
1.4012	1.3974	1.3700
1.3464	1.3464	1.3405
0.1947	0.1938	0.2115
0	0.0051	0.0901
0.0051	0	0.0901
0.0901	0.0901	0

Matrik jarak evolusioner yang terbentuk diatas, merupakan inputan yang digunakan untuk proses pembentukan pohon filogenetik. Adapun langkah-langkah pembentukan pohon filogenetik dengan algoritma *Neighbor Joining* adalah sebagai berikut:

Cycle 1

1. Input: Matriks jarak evolusioner

   Pada matriks jarak evolusioner tersebut A, B, C.....L menunjukkan nama OTU yang mewakili masing-masing sekuen dimana:

   a. A mewakili sekuen 1 Murine HVI,
   b. B mewakili sekuen 2 Murine HV2,
   c. C mewakili sekuen 3 Human SARS Co-V,
   d. D mewakili sekuen 4 Palm Civet,
   e. E mewakili sekuen 5 Canine Co-V1,
   f. F mewakili sekuen 6 Feline Co-V,
   g. G mewakili sekuen 7 Porcine PEDV,
   h. H mewakili sekuen 8 IBV 3,
   i. I mewakili sekuen 9 Porcine HEV 3,
   j. J mewakili sekuen 10 Bovine CoV1 ,
   k. K mewakili sekuen 11 Bovine CoV2,
   l. L mewakili sekuen 12 Human coronavirus OC 43.

   Berikut adalah matriks jarak evolusinya:

Tabel 3.1 Matriks jarak evolusioner *cycle* 1

	A	B	C	D	E	F	G	H	I	J	K	L
A	0	0,1191	2,2059	1,1882	1,3609	1,4032	1,4052	1,4113	0,4398	0,4351	0,4363	0,4586
B	0,1191	0	1,2006	1,1920	1,4438	1,3995	1,4540	1,3791	0,4527	0,4514	0,4502	0,4596
C	2,2059	1,2006	0	0,1314	1,4012	1,3872	1,4090	1,3786	1,2622	1,1996	1,1971	1,2188
D	1,1882	1,1920	0,1314	0	1,3923	1,3914	1,4332	1,3583	1,2013	1,2013	1,1988	1,2127
E	1,3609	1,4438	1,4012	1,3923	0	0,7985	0,8363	1,3257	1,3898	1,4715	1,4675	1,3869
F	1,4032	1,3995	1,3872	1,3914	0,7985	0	0,8688	1,3699	1,3702	1,3762	1,3762	1,4155
G	1,4052	1,4540	1,4090	1,4332	0,8363	0,8688	0	1,3079	1,4290	1,4012	1,3974	1,3700
H	1,4113	1,3791	1,3786	1,3583	1,3257	1,3699	1,3079	0	1,3258	1,3464	1,3464	1,3405
I	0,4398	0,4527	1,2622	1,2430	1,3898	1,3702	1,4290	1,3258	0	0,1947	0,1938	0,2115
J	0,4351	0,4514	1,1996	1,2013	1,4715	1,3762	1,4012	1,3464	0,1947	0	0,0051	0,0901
K	0,4363	0,4502	1,1971	1,1988	1,4675	1,3762	1,3974	1,3464	0,1938	0,0051	0	0,0901
L	0,4586	0,4596	1,2188	1,2127	1,3869	1,4155	1,3700	1,3405	0,2115	0,0901	0,0901	0

2. Step 1:
   Hitung S_i mengikuti rumus:

$$S_i = \frac{1}{N-2}\sum_{k=1}^{N} D_{ik}$$

   Dengan N adalah banyaknya OTU, D_{ik} adalah jarak dari i ke k pada matriks evolusinya. Sedangkan N pada Cycle ini adalah 12.

   Hasil perhitungan masing-masing S_i adalah:

   S_A = 1.08618 ; S_B = 1.0002 ; S_C = 1.3991 ; S_D = 1.2942
   S_E = 1.42744 ; S_F = 1.4156 ; S_G = 1.4312 ; S_H = 1.4889
   S_I = 0.95125 ; S_J = 0.9172 ; S_K = 0.9158 ; S_L = 0.9252

3. Step 2:
   Dicari nilai minimum untuk masing-masing pasangan sekuen:

   $M_{ij} = D_{ij} - S_i - S_j$.

   Apabila dituliskan secara lengkap, matriks -nya sebagai berikut:

	A	B	C	D	E	F	G	H	I
A	0	-1.9673	-0.2794	-1.1923	-1.1528	-1.0986	-1.1122	-1.1638	-1.5976
B	-1.9673	0	-1.1987	-1.1025	-0.9839	-1.0164	-0.9774	-1.1101	-1.4988
C	-0.2794	-1.1987	0	-2.5620	-1.4254	-1.4276	-1.4214	-1.5096	-1.0882
D	-1.1923	-1.1025	-2.5620	0	-1.3293	-1.3185	-1.2923	-1.4249	-1.0026
E	-1.1528	-0.9839	-1.4254	-1.3293	0	-2.0446	-2.0223	-1.5907	-0.9889
F	-1.0986	-1.0164	-1.4276	-1.3185	-2.0446	0	-1.9781	-1.5347	-0.9967
G	-1.1122	-0.9774	-1.4214	-1.2923	-2.0223	-1.9781	0	-1.6123	-0.9534
H	-1.1638	-1.1101	-1.5096	-1.4249	-1.5907	-1.5347	-1.6123	0	-1.1144
I	-1.5976	-1.4988	-1.0882	-1.0026	-0.9889	-0.9967	-0.9534	-1.1144	0
J	-1.5683	-1.4661	-1.1168	-1.0102	-0.8732	-0.9568	-0.9473	-1.0599	-1.6738
K	-1.5658	-1.4658	-1.1180	-1.0114	-0.8758	-0.9554	-0.9497	-1.0585	-1.6733
L	-1.5546	-1.4659	-1.1056	-1.0068	-0.9658	-0.9254	-0.9864	-1.0738	-1.6650

	H	I	J
A	-1.5683	-1.5658	-1.5546
B	-1.4661	-1.4658	-1.4659
C	-1.1168	-1.1180	-1.1056
D	-1.0102	-1.0114	-1.0068
E	-0.8732	-0.8758	-0.9658
F	-0.9568	-0.9554	-0.9254
G	-0.9473	-0.9497	-0.9864
H	-1.0599	-1.0585	-1.0738
I	-1.6738	-1.6733	-1.6650
J	0	-1.8280	-1.7524
K	-1.8280	0	-1.7510
L	-1.7524	-1.7510	0

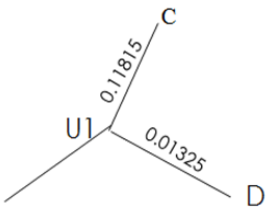
Diperoleh pasangan M terkecil adalah M_{CD} = -2.5620

4. Step 3:
   Definisikan OTU baru yaitu U_1 yang menggantikan pasangan terkecil (C dan D). Selanjutnya taksa tersebut digabungkan sebagai U_1 mengikuti rumus:

   $S_{CU_1} = 0.5(D_{CD} + S_C - S_D) = 0.11815$

   $S_{DU_1} = 0.5(D_{CD} + S_D - S_C) = 0.01325$

5. Step 4:
   Hubungkan taksa U_1 dengan C dan U_1 dengan D, masing-masing dengan mengikuti panjang *edge* atau jarak sebagaimana hasil perhitungan pada step 3.



Gambar 3.6 Pohon pada *cycle* 1

Pada pohon diatas, panjang cabang menggambarkan jarak evolusinya.

6. Step 5:
   Gabungkan jarak baru dari semua taksa ke U_1

$$D_{AU_1} = 0.5(D_{CA} + D_{DA} - D_{CD}) = 1.6323$$

$$D_{BU_1} = 0.5(D_{CB} + D_{DB} - D_{CD}) = 1.1306$$

$$D_{EU_1} = 1.3310 \qquad D_{FU_1} = 1.3236 \qquad D_{GU_1} = 1.3554$$

$$D_{HU_1} = 1.3027 \qquad D_{IU_1} = 1.1869 \qquad D_{JU_1} = 1.1347$$

$$D_{KU_1} = 1.1322 \qquad D_{LU_1} = 1.1500$$

Hasil dari jarak baru U1 ke semua taksa untuk selanjutnya dimasukkan dalam matriks jarak evolusioner yang baru.

Langkah perhitungan yang berikutnya sama seperti pada cycle 1, nilai berbeda karena N berbeda, *Mij* terkecil berbeda, yang akhirnya jarak baru ke masing-masing taksa juga berbeda.

Cycle 2

1. Matriks jarak evolutioner yang baru:

	A	B	U1	E	F	G	H	I	J	K
A	0									
B	0,1191	0								
U1	1,6313	1,1306	0							
E	1,3609	1,4438	1,3310	0						
F	1,4032	1,3995	1,3236	0,7985	0					
G	1,4052	1,4540	1,3554	0,8363	0,8688	0				
H	1,4113	1,3791	1,3027	1,3257	1,3699	1,3079	0			
I	0,4398	0,4527	1,1869	1,3898	1,3702	1,4290	1,3258	0		
J	0,4351	0,4514	1,1347	1,4715	1,3762	1,4012	1,3464	0,1947	0	
K	0,4363	0,4502	1,1322	1,4675	1,3762	1,3974	1,3464	0,1938	0,0051	0
L	0,4586	0,4596	1,1500	1,3869	1,4155	1,3700	1,3405	0,2115	0,0901	0,0901

2. Step 1:
   Dengan rumus yang sama seperti *cycle* 1, perhitungan diperoleh dengan N = 11 karena adanya dua taksa

terkecil C dan D yang bergabung menjadi satu taksa baru (U_1).

$S_A = 1.011$; $S_B = 0.9711$; $S_{U1} = 1.4087$;

$S_E = 1.4235$; $S_F = 1.4113$; $S_G = 1.4250$; $S_H = 1.4950$

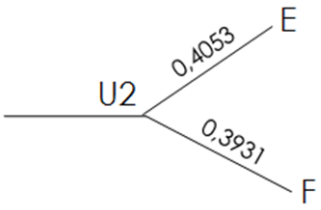
$S_I = 0.9105$; $S_J = 0.8785$; $S_K = 0.8772$; $S_L = 0.8856$

3. Step 2:
   Diperoleh pasangan M terkecil adalah M_{EJ} = -2.0363
4. Step 3:
   Gabungkan taksa E dan F sebagai U_2 dan hitung jaraknya:

   $S_{EU_2} = 0.5(D_{EF} + S_E - S_F) = 0.40535$

   $S_{FU_1} = 0.5(D_{CD} + S_D - S_C) = 0.39315$

5. Step 4:
   Hubungkan taksa U_2 dengan E dan U_2 dengan F, masing-masing dengan mengikuti panjang *edge* atau jarak sebagaimana hasil perhitungan pada step 3.



Gambar 3.7 Pohon pada *cycle* 2

Panjang cabang dari EU_2 mewakili jarak evolusinya, begitu pula dengan FU_2.

6. Step 5:
   Diperoleh jarak baru dari semua taksa ke U_2

$D_{AU_2} = 0.9828$ $D_{BU_2} = 1.0224$ $D_{U_1U_2} = 0.9280$

$D_{GU_2} = 0.4533$ $D_{HU_2} = 0.9485$ $D_{IU_2} = 0.9807$

$D_{JU_2} = 1.0246$ $D_{KU_2} = 1.0226$ $D_{LU_2} = 1.0019$

   Hasil dari jarak baru U_2 ke semua taksa untuk selanjutnya dimasukkan dalam matriks jarak evolusioner yang baru.

Cycle 3

1. Matriks jarak evolutioner yang baru

	A	B	U1	U2	G	H	I	J	K
A	0								
B	0,1191	0							
U1	1,6313	1,1306	0						
U2	0,9828	1,0224	0,92805	0					
G	1,4052	1,4540	1,3554	0,4533	0				
H	1,4113	1,3791	1,3027	0,9485	1,3079	0			
I	0,4398	0,4527	1,1869	0,9807	1,4290	1,3258	0		
J	0,4351	0,4514	1,1347	1,0246	1,4012	1,3464	0,1947	0	
K	0,4363	0,4502	1,1322	1,0226	1,3974	1,3464	0,1938	0,0051	0
L	0,4586	0,4596	1,1500	1,0019	1,3700	1,3405	0,2115	0,0901	0,0901

2. Step 1:
   Dengan $N=10$ diperoleh:

$S_A = 0.9147$; $S_B = 0.8649$; $S_{U1} = 1.3689$;

$S_{U2} = 1.0456$; $S_G = 1.4466$; $S_H = 1.4636$;

$S_I = 0.8018$; $S_J = 0.7604$; $S_K = 0.7593$; $S_L = 0.7713$.

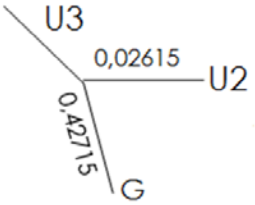
3. Step 2:
   Diperoleh pasangan M terkecil adalah = -2.0389

4. Step 3:
   Gabungkan taksa U_2 dan G sebagai U_3

   $S_{U_2U_3} = 0.5(D_{U_2G} + S_{U_2} - S_G) = 0.02615$

   $S_{GU_3} = 0.5(D_{U_2G} + S_G - S_{U_2}) = 0.42715$

5. Step 4:
   Hubungkan taksa U_3 dengan G dan U_3 dengan U_2, masing-masing dengan mengikuti panjang *edge* atau jarak sebagaimana hasil perhitungan pada step 3.



Gambar 3.8 Pohon pada *cycle* 3

6. Step 5:
   Diperoleh jarak baru dari semua taksa ke U_3

   $D_{AU_3} = 0.9673$ $D_{BU_3} = 1.0115$ $D_{U_1U_3} = 0.91507$

   $D_{HU_3} = 0.90157$ $D_{IU_3} = 0.97822$

   $D_{JU_3} = 0.98625$ $D_{KU_3} = 0.98335$ $D_{LU_3} = 0.95932$

   Hasil dari jarak baru U_3 ke semua taksa untuk selanjutnya dimasukkan dalam matriks jarak evolusioner yang baru.

Cycle 4

1. Matriks jarak evolutioner yang baru:

	A	B	U1	U3	H	I	J	K
A	0							
B	0,1191	0						
U1	1,6313	1,1306	0					
U3	0,9673	1,0115	0,9150	0				
H	1,4113	1,3791	1,3027	0,9015	0			
I	0,4398	0,4527	1,1869	0,9782	1,3258	0		
J	0,4351	0,4514	1,1347	0,9862	1,3464	0,1947	0	
K	0,4363	0,4502	1,1322	0,9833	1,3464	0,1938	0,0051	0
L	0,4586	0,4596	1,1500	0,9593	1,3405	0,2115	0,0901	0,0901

2. Step 1:
   Dengan N=9 diperoleh:

   S_A = 0.8424 ; S_B = 0.7792 ;

   S_{U1} = 1.3691 ; S_{U3} = 1.1004 ; S_H = 1.4791 ;

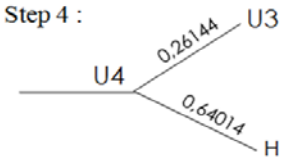
   S_I = 0.7119 ; S_J = 0.6634 ; S_K = 0.6625 ; S_L = 0.6797.

3. Step 2:
   Diperoleh pasangan M terkecil adalah M_{U_3H} = -1.67792
4. Step 3:
   Gabungkan taksa U_3 dan H sebagai U_4

   $$S_{U_3U_4} = 0.5\left(D_{U_3H} + S_{U_3} - S_H\right) = 0.26144$$

   $$S_{HU_4} = 0.5\left(D_{U_3H} + S_H - S_{U_3}\right) = 0.64014$$

5. Step 4:
   Hubungkan taksa U_4 dengan H dan U_4 dengan U_3, masing-masing dengan mengikuti panjang *edge* atau jarak sebagaimana hasil perhitungan pada step 3.



Gambar 3.9 Pohon pada *cycle* 4

6. Step 5:
   Diperoleh jarak baru dari semua taksa ke U_4

   $D_{AU_4} = 0.73851$ $D_{BU_4} = 0.74454$ $D_{U_1U_4} = 0.6581$

   $D_{IU_4} = 0.70122$ $D_{JU_4} = 0.71554$ $D_{KU_4} = 0.71408$

   $D_{LU_4} = 0.69912$

   Hasil dari jarak baru U_4 ke semua taksa untuk selanjutnya dimasukkan dalam matriks jarak evolusioner yang baru.

Cycle 5

1. Matriks jarak evolutioner yang baru

	A	B	U1	U4	I	J	K
A	0						
B	0,1191	0					
U1	1,6313	1,1306	0				
U4	0,7385	0,7445	0,6581	0			
I	0,4398	0,4527	1,1869	0,7012	0		
J	0,4351	0,4514	1,1347	0,7155	0,1947	0	
K	0,4363	0,4502	1,1322	0,7140	0,1938	0,0051	0
L	0,4586	0,4596	1,1500	0,6991	0,2115	0,0901	0,0901

2. Step 1:
   Dengan N=8 diperoleh:

   $S_A = 0.7095$; $S_B = 0.6347$; $S_{U1} = 1.3373$; $S_{U4} = 0.8285$

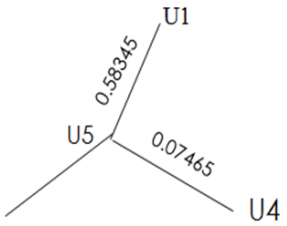
   $S_I = 0.5634$; $S_J = 0.5044$; $S_K = 0.50363$; $S_L = 0.5262$

3. Step 2:
   Diperoleh pasangan M terkecil adalah $M_{U_1U_4}$= -1.5077
4. Step 3:
   Gabungkan taksa U1 dan U4 sebagai U5

   $S_{U_1U_5} = 0.5(D_{U_1U_4} + S_{U_1} - S_{U_4}) = 0.58345$

   $S_{U_4U_5} = 0.5(D_{U_1U_4} + S_{U_4} - S_{U_1}) = 0.07465$

5. Step 4:
   Hubungkan taksa U_5 dengan U_1 dan U_5 dengan U_4, masing-masing dengan mengikuti panjang *edge* atau jarak sebagaimana hasil perhitungan pada step 3.



Gambar 3.10 Pohon pada *cycle* 5

6. Step 5:
   Diperoleh jarak baru dari semua taksa ke U_5

   $D_{AU_5} = 0.85585$ $D_{BU_5} = 0.6085$ $D_{IU_5} = 0.61501$

   $D_{JU_5} = 0.59607$ $D_{KU_5} = 0.59409$ $D_{LU_5} = 0.59551$

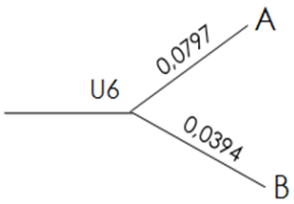
   Hasil dari jarak baru U_5 ke semua taksa untuk selanjutnya dimasukkan dalam matriks jarak evolusioner yang baru.

Cycle 6

1. Matriks jarak evolutioner yang baru

	A	B	U5	I	J	K
A	0					
B	0,1191	0				
U5	0,8558	0,6085	0			
I	0,4398	0,4527	0,6150	0		
J	0,4351	0,4514	0,5961	0,1947	0	
K	0,4363	0,4502	0,5941	0,1938	0,0051	0
L	0,4586	0,4596	0,5955	0,2115	0,0901	0,0901

2. Step 1:
   Dengan N=7 diperoleh:
   $S_A = 0.54858$; $S_B = 0.5083$; $S_{U5} = 0.7730$;
   $S_I = 0.4215$; $S_J = 0.3545$; $S_K = 0.3539$; $S_L = 0.3807$
3. Step 2:
   Diperoleh pasangan M terkecil adalah M_{AB} = -0.9378
4. Step 3:
   Gabungkan taksa A dan B sebagai U_6
   $S_{AU_6} = 0.5(D_{AB} + S_A - S_B) = 0.0797$
   $S_{BU_6} = 0.5(D_{AB} + S_B - S_A) = 0.0394$
5. Step 4:
   Hubungkan taksa U_6 dengan A dan U_6 dengan B, masing-masing dengan mengikuti panjang *edge* atau jarak sebagaimana hasil perhitungan pada step 3.



Gambar 3.11 Pohon pada *cycle* 6

6. Step 5:
   Diperoleh jarak baru dari semua taksa ke U6
   $D_{U_5U_6} = 0.6726$ $D_{IU_6} = 0.3867$
   $D_{JU_6} = 0.3837$ $D_{KU_6} = 0.3837$ $D_{LU_6} = 0.3987$

   Hasil dari jarak baru U_6 ke semua taksa untuk selanjutnya dimasukkan dalam matriks jarak evolusioner yang baru.

Cycle 7

1. Matriks jarak evolutioner yang baru

	U6	U5	I	J	K
U6	0				
U5	0,6726	0			
I	0,3867	0,6150	0		
J	0,3837	0,5961	0,1947	0	
K	0,3837	0,5941	0,1938	0,0051	0
L	0,3987	0,5955	0,2115	0,0901	0,0901

2. Step 1:
   Dengan N=6 diperoleh:
   S_{U6}= 0.55635 ; S_{U5}= 0.76832 ; S_I = 0.4004 ;
   S_J = 0.3174 ; S_K= 0.3167 ; S_L = 0.3465.

3. Step 2:
   Diperoleh pasangan M terkecil adalah $M_{U_6U_5}$ = -0.652

4. Step 3:
   Gabungkan taksa U5 dan U6 sebagai U7
   $$S_{U_6U_7} = 0.5\left(D_{U_5U_6} + S_{U_6} - S_{U_5}\right) = 0.2303$$
   $$S_{U_5U_7} = 0.5\left(D_{U_5U_6} + S_{U_5} - S_{U_6}\right) = 0.4423$$

5. Step 4:
   Hubungkan taksa U6 dengan U7 dan U5 dengan U7, masing-masing dengan mengikuti panjang edge atau jarak sebagaimana hasil perhitungan pada step 3.

Gambar 3.12 Pohon pada cycle 7

6. Step 5:
   Diperoleh jarak baru dari semua taksa ke U7

   $D_{IU_7} = 0.1645$ $D_{JU_7} = 0.1536$

   $D_{KU_7} = 0.1526$ $D_{LU_7} = 0.1608$

   Hasil dari jarak baru U7 ke semua taksa untuk selanjutnya dimasukkan dalam matriks jarak evolusioner yang baru.

Cycle 8

1. Matriks jarak evolutioner yang baru

	U7	I	J	K
U7	0			
I	0,1645	0		
J	0,1536	0,1947	0	
K	0,1526	0,1938	0,0051	0
L	0,1608	0,2115	0,0901	0,0901

2. Step 1:
   Dengan N=5 diperoleh:

   $S_{U7} = 0.2105$; $S_I = 0.2548$;

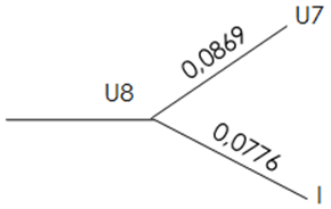
   $S_J = 0.1478$; $S_K = 0.1472$; $S_L = 0.1841$

3. Step 2:
   Diperoleh pasangan M terkecil adalah $M_{U_7 I}$ = -0.3008
4. Step 3:
   Gabungkan taksa U7 dan I sebagai U8

   $S_{U_7 U_8} = 0.5(D_{U_7 I} + S_{U_7} - S_I) = 0.0869$

   $S_{I U_8} = 0.5(D_{U_7 I} + S_I - S_{U_7}) = 0.0776$

5. Step 4:
   Hubungkan taksa U7 dengan U8 dan I dengan U8, masing-masing dengan mengikuti panjang edge atau jarak sebagaimana hasil perhitungan pada step 3.



Gambar 3.13 Pohon pada cycle 8

6. Step 5:
   Diperoleh jarak baru dari semua taksa ke U8

   $D_{JU_8} = 0.0919$; $D_{KU_8} = 0.0909$; $D_{LU_8} = 0.1039$

   Hasil dari jarak baru U8 ke semua taksa untuk selanjutnya dimasukkan dalam matriks jarak evolusioner yang baru.

Cycle 9

1. Matriks jarak evolutioner yang baru

	U8	J	K
U8	0		
J	0,0919	0	
K	0,0909	0,0051	0
L	0,1039	0,0901	0,0901

2. Step 1:
   Dengan N=4 diperoleh:

   $S_{U8} = 0.14335$; $S_J = 0.09355$; $S_K = 0.09305$; $S_L = 0.14205$

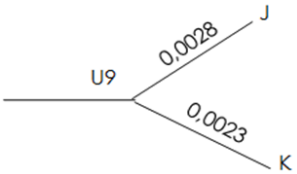
3. Step 2:
   Diperoleh pasangan M terkecil adalah M_{JK} = -0.1814

4. Step 3:
   Gabungkan taksa J dan K sebagai U9

   $S_{JU_9} = 0.5(D_{JK} + S_J - S_K) = 0.0028$

   $S_{KU_9} = 0.5(D_{JK} + S_K - S_J) = 0.0023$

5. Step 4:
   Hubungkan taksa J dengan U9 dan K dengan U9, masing-masing dengan mengikuti panjang *edge* sebagaimana hasil perhitungan step 3.



Gambar 3.14 Pohon pada *cycle* 9

6. Step 5:
   Diperoleh jarak baru dari semua taksa ke U_9

   $D_{U_9U_8} = 0.0888\,;\qquad D_{U_9L} = 0.0875$

   Hasil dari jarak baru U_9 ke semua taksa untuk selanjutnya dimasukkan dalam matriks jarak evolusioner yang baru.

Cycle 10

1. Matriks jarak evolutioner yang baru

	U8	U9
U8	0	
U9	0,0888	0
L	0,1039	0,0875

2. Step 1:
   Dengan N=3 diperoleh:

   $S_{U8} = 0.1927$; $S_{U9} = 0.1763$; $S_L = 0.1914$

3 Step 2:
   Diperoleh pasangan M semua nilainya sama yaitu:

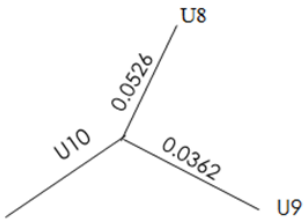
   $M_{U_8U_9} = -0.2802$; $M_{U_8L} = -0.2802$; $M_{U_9L} = -0.2802$;

   Dipilih salah satu yaitu U_8U_9.
4. Step 3:
   Gabungkan taksa U_8 dan U_9 sebagai U_{10}

   $S_{U_8U_{10}} = 0.5(D_{U_8U_9} + S_{U_8} - S_{U_9}) = 0.0526$

   $S_{U_9U_{10}} = 0.5(D_{U_8U_9} + S_{U_9} - S_{U_8}) = 0.0362$

5. Step 4:
   Hubungkan taksa U8 dengan U10 dan U9 dengan U10, masing-masing dengan mengikuti panjang edge sebagaimana hasil perhitungan step 3.



Gambar 3.15 Pohon pada cycle 10

6. Step 5:
   Diperoleh jarak baru dari semua taksa terakhir U10 ke L

   $D_{U_{10}L} = 0.0513$

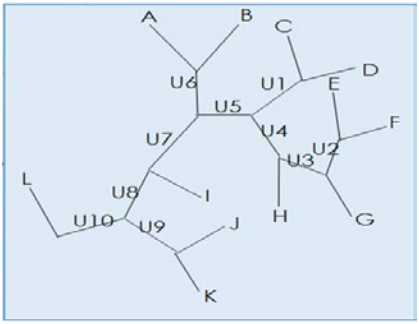
Cycle 11

1. Matriks jarak evolutioner yang baru

	U10
U10	0
L	0,0513

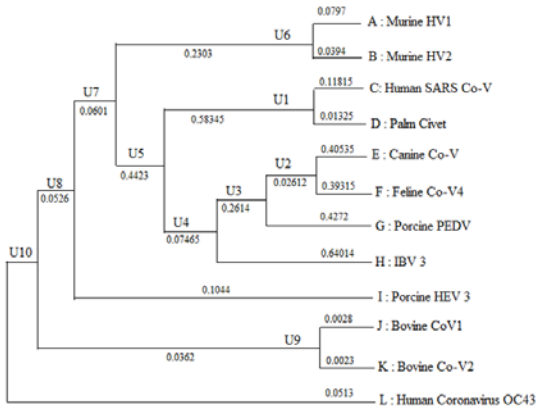
2. Sampai di sini iterasi berakhir karena (N-2) = 0
   Sehingga percabangan terakhir adalah U_{10} ke L.

Setelah semua percabangan digabungkan, diperoleh pohon filogenetik berikut:



Gambar 3.16 Pohon filogenetik Host SARS Co-V

Apabila digambarkan dalam bentuk kladogram sebagai berikut:



Gambar 3.17 Pohon filogenetik Host SARS Co-V dalam bentuk cladogram

Pohon filogenetik pada Gambar 3.15 menjelaskan tentang kekerabatan antar spesies yang diperkirakan sebagai host dari coronavirus. Masing-masing *leaf* (A,B,C,D,E,F,G,H,I,J,K,L) pada *pohon* disebut juga OTU (*Operasional Taxonomy Unit*) yang mewakili spesies yang disejajarkan. Pada gambar nampak percabangan yang paling dekat adalah antara OTU A dengan B, OTU C dengan D, OTU E dengan F, dan OTU J dengan K.

OTU A yang mewakili spesies Murine HVI berada pada satu cabang yang sama dengan OTU B yaitu Murine HV2. Keduanya merupakan virus Hepatitis pada tikus yang berbeda strainnya. Jarak evolusi keduanya sebesar 0.1191. Pada gambar diatas nilai tersebut tidak nampak karena sudah dipecah sebagai jarak *branch* U_6 ke spesies A (0.0797) dan jarak *branch* U_6 ke spesies B (0.0394). Bila dijumlahkan maka diketahui jarak evolusi A ke B (0.1191).

OTU C yang mewakili Human SARS Co-V berada pada percabangan yang sama dengan OTU D yaitu Palm Civet. Human SARS Co-V adalah Coronavirus yang menyerang manusia, sedangkan Palm Civet adalah musang kelapa Himalaya. Jarak evolusi keduanya 0.1314. Angka tersebut bisa dilihat dari penjumlahan jarak percabangan Palm Civet ke *branch* U_1 (0.01325) dengan jarak Human SARS Co-V ke *branch* U_1 (0.11815). Human SARS Co-V adalah coronavirus pada manusia, dengan melihat hasil analisa pohon filogenetiknya ternyata SARS Co-V hubungan evolusinya paling dekat dengan Palm Civet. Ini menandakan bahwa Palm Civet adalah host yang dicari pada kasus ini.

OTU E yang mewakili Canine Co-V1 dan OTU F yang mewakili Feline Co-V4. Canine Co-V adalah Canine Enteric Coronavirus K378 yang menyerang anjing, sedangkan Feline Co-V4 adalah Feline Infectious Peritonitis Virus pada babi.

Meskipun berada pada satu percabangan yang sama, akan tetapi hubungan kekerabatan keduanya terbilang jauh karena jarak evolusi yang cukup besar 0.7985. Namun apabila dibandingkan dengan spesies yang lain, jarak evolusi keduanya lebih dekat. Hal ini terbukti apabila dilihat dari jarak evolusinya pada Tabel 3.1.

OTU J yang mewakili Bovine Co-V1 dan OTU K yang mewakili Bovine Co-V2. Keduanya adalah Bovine Coronavirus dengan strain berbeda yang menyerang babi. Jarak evolusi keduanya sangat kecil yaitu 0.0051. Artinya, kekerabatan Coronavirus pada babi sangat erat sekali, terbukti dari jumlah protein yang bermutasi pada keduanya sangat kecil. Dari situ bisa disimpulkan bahwa pada spesies virus yang sama meski strain berbeda apabila menyerang host yang sama bisa dikatakan mempunyai hubungan kekerabatan yang erat karena banyaknya protein yang berevolusi pada keduanya sangat sedikit.

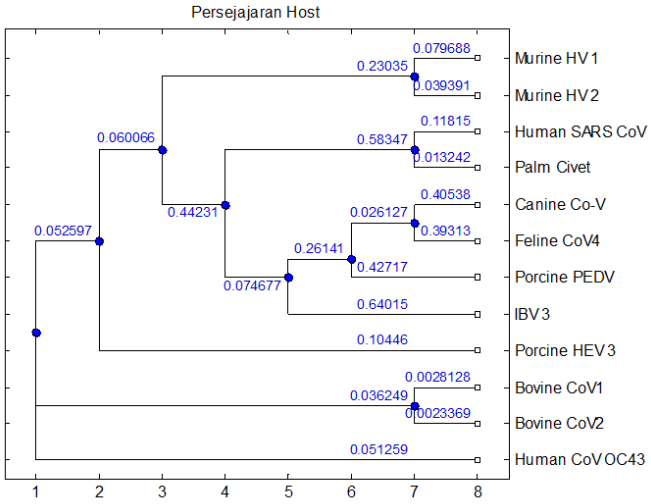
Berdasarkan analisa pohon filogenetik pada tingkat *leaf* diatas bisa disimpulkan bahwa jarak evolusioner yang paling dekat dengan human SARS Co-V adalah Palm Civet. Hal ini dikarenakan sekuen protein pada Palm Civat bila disejajarkan dengan sekuen protein human SARS Co-V bermutasi sebanyak 0.1314 yang menandakan bilangan paling kecil bila dibandingkan dengan mutasi pada sekuen protein host yang lain. Dan jarak evolusioner terkecil itu menandakan host yang dicari.

Berdasarkan *branch* (percabangan), pada Gambar 3.15 *branch* adalah (U_1, U_2, U_3, U_4, U_5, U_6, U_7, U_8, U_9, U_{10}). *Branch* terdekat dengan *leaf* E dan F adalah U_2, sedangkan *branch* U_3 menghubungkan U_2 dan G, *branch* U_4 menghubungkan *branch* U_3 dan H. Hal ini menunjukkan bahwa Canine Co-V1, Feline Co-V4, Porcine PEDV, dan IBV 3 mempunyai

hubungan kekerabatan yang dekat dibandingkan dengan spesies yang lain.

3.5 Pohon Filogenetik Penentuan Host SARS Hasil Matlab dan Clustal W

Setelah diperoleh hasil perhitungan secara manual, selanjutnya dibuatlah program untuk memperoleh hasil yang lebih cepat dan akurat. Berikut adalah pohon filogenetik hasil running program Matlab:



Gambar 3.18 Pohon filogenetik hasil komputasi yang disimulasikan dalam Matlab

Pohon filogenetik yang disimulasikan dalam Matlab memberikan hasil yang sama dengan hasil perhitungan manual. Tampak semua daun maupun cabang yang terbentuk berada pada posisi yang sama. Nilai jarak evolusi antar masing-masing spesies juga menunjukkan nilai yang sama. Hal ini dikarenakan pada perhitungan manual maupun komputasi,

keduanya menggunakan algoritma *Neighbor Joining* dengan model koreksi Jukes Cantor.

Sebagai software pembanding digunakan Clustal W. Clustal W merupakan algoritma Heuristik yang melakukan pemecahan masalah penyejajaran ganda ke dalam masalah penyejajaran berpasangan secara komputasional. Berikut adalah tampilan pohon filogenetik hasil *running software* Clustal W2.

Gambar 3.19 Pohon Filogenetik hasil running *software* Clustal W2

Bila dibandingkan dengan hasil *running software* Clustal W, pohon filogenetik hasil *running* program Matlab meskipun sama, namun jarak genetik masing-masing spesiesnya berbeda. Penyebabnya adalah karena pada *software* Clustal W2, meski algoritma pembentuk pohon filogenetiknya menggunakan algoritma *Neighbor Joining*, namun tidak ada proses koreksi pada jarak genetiknya, dalam hal ini karena *software* tidak dilengkapi dengan *tool* untuk pemakaian koreksi jarak baik dalam model Jukes Cantor maupun dalam model yang lain.

Dari hasil kedua gambar, nilai jarak masing-masing spesiesnya ke percabangan terdekat apabila ditabelkan sebagai berikut:

Tabel 3.2 Perbandingan jarak masing-masing spesies ke percabangan terdekat antara hasil Matlab dan Clustal W.

No	Nama Virus	Kode akses GenBank	Jarak Evolusi	Jarak Genetik
			Matlab	Clustal W
1.	Murine HV1	YP_209233.1	0.079688	0.03974
2.	Murine HV2	NP_045300.1	0.039391	0.03805
3.	Human SARS CoV	AAP 41037.1	0.11815	0.00665
4.	Palm Civet	AAV 49723.1	0.013242	0.00609
5.	Canine CoV	Q65984.1	0.40538	0.26168
6.	Feline CoV4	BAA06805.1	0.39313	0.26516
7.	Porcine PEDV	NP_598310.1	0.42717	0.2855
8.	IBV 3	NP_040831.1	0.64015	0.36444
9.	Porcine HEV3	AAL80031.1	0.10446	0.08499
10.	Bovine CoV1	AAL57308.1	0.0028128	0.00267
11.	Bovine CoV2	AAK 83356.1	0.0023369	0.00247
12.	Human CoV OC43	NP_937950.1	0.051259	0.04042

Sedangkan pada Gambar 3.17, bagian yang dilingkari adalah pada spesies Human SARS Co-V dan Palm Civet. Subject id ditunjukkan oleh gi|30795145|gb|AAP41037.1 yang mewakili Human SARS Co-V dengan jarak genetik 0.00665 terhadap cabang, sedangkan subjek id gi|55275769|gb|AAV49723.1 mewakili Palm Civet sebagai host dengan jarak genetik 0.00609 terhadap cabang. Jarak genetik antara Human SARS Co-V dengan Palm Civet sebesar nilai penjumlahan keduanya yaitu 0.01274.

3.6 Soal Latihan

Perhatikan matriks jarak berikut:

M_d	x^1	x^2	x^3	x^4	x^5	x^6	x^7
x^1	0	13	9	6	13	13	17
x^2	13	0	4	15	16	16	20
x^3	9	4	0	11	12	12	16
x^4	6	15	11	0	15	15	19
x^5	13	16	12	15	0	16	6
x^6	13	16	12	15	16	0	20
x^7	17	20	16	19	6	20	0

Bentuklah pohon filogenetik dengan menggunakan algoritma *neighbor joining*.

BAB IV

PEMBENTUKAN POHON FILOGENETIK DENGAN METODE MAXIMUM LIKELIHOOD

Metode Maximum Likelihood merupakan metode yang berbasis pada perkalian fungsi densitas probabilitas. Metode ini dapat diaplikasikan untuk mengkonstruksi pohon filogenetik DNA. Untuk itu perlu dipelajari beberapa dasar yang digunakan pada metode ini.

4.1 Proses Stokastik

Proses stokastik adalah kumpulan dari variabel random $\{X(t), t \in T\}$. Indeks t diinterpretasikan sebagai waktu, dan $X(t)$ adalah variabel random yang menunjukkan keadaan dari proses pada waktu t. Jika T dapat dihitung yaitu $T = \{0,1,2,...\}$, maka proses stokastik ini disebut proses waktu diskrit. Sementara itu, jika T berupa interval dari garis Riil $(T = \{t \mid -\infty < t < \infty\})$, maka disebut proses waktu kontinu (Praptono, 1986). Dengan demikian, untuk membedakannnya dapat dituliskan $\{X_n, n = 0,1,2,...\}$ sebagai proses stokastik dengan waktu diskrit, sedangkan $\{X(t), t \geq 0\}$ merupakan proses stokastik untuk waktu kontinu.

4.2 Rantai Markov

Rantai Markov merupakan bentuk khusus dari suatu proses stokastik. Rantai Markov didefinisikan sebagai proses stokastik dari variabel random $\{X_n, n = 0,1,2,...\}$ yang membentuk suatu deret yang memenuhi sifat

Markov. Sifat Markov menyatakan bahwa jika diberikan suatu peristiwa yang telah berlalu yaitu $X_0, X_1, ..., X_{n-1}$, dan peristiwa yang sedang berlangsung yaitu X_n, maka peristiwa yang akan datang yaitu X_{n+1} bersifat independen dari peristiwa masa lalu. Dengan kata lain, peristiwa yang akan datang hanya bergantung pada peristiwa yang sedang berlangsung.

Untuk pengamatan yang prosesnya sampai waktu ke n, maka distribusi nilai proses dari waktu ke $n+1$ hanya bergantung pada nilai dari proses pada waktu ke n. Secara umum, sifat Markov dapat ditulis

$$P(X_{n+1} = j \mid X_0 = i_0, X_1 = i_1, ..., X_n = i) = P(X_{n+1} = j \mid X_n = i)$$

.

(Viniotis, 1998).

Probabilitas bersyarat $P(X_{n+1} = j \mid X_n = i)$ disebut probabilitas transisi dari keadaan i ke keadaan j. Selanjutnya, probabilitas transisi (p_{ij}) dapat didefinisikan sebagai $p_{ij} = P(X_{n+1} = j \mid X_n = i)$.

4.3 Matriks Probabilitas Transisi

Matriks probabillitas transisi dari suatu rantai Markov adalah suatu matriks bujur sangkar berukuran $n \times n$, dengan n bergantung pada banyak peristiwa atau *state* pada rantai Markov tersebut. Elemen pada matriks probabilitas transisi adalah probabilitas perubahan keadaan pada peristiwa j yang pada peristiwa sebelumnya berada pada keadaan i. Tetapi, pada saat rantai Markov mencapai situasi stasioner, maka probabilitas tersebut tidak lagi bergantung pada n, sehingga dituliskan besar probabilitas sebagai p_{ij}. Dengan demikian, probabilitas proses berpindah dari keadaan i ke keadaan j homogen dalam waktu. Jadi, dapat didefinisikan

seluruh probabilitas proses dalam bentuk matriks P yang disebut sebagai matriks probabilitas transisi dari rantai Markov, dituliskan dalam bentuk matriks berikut (Viniotis, 1998):

$$P = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0j} \\ p_{10} & p_{11} & \cdots & p_{1j} \\ \vdots & \vdots & \ddots & \vdots \\ p_{i0} & p_{i1} & \cdots & p_{ij} \end{bmatrix}$$

Untuk banyak peristiwa atau *state* adalah n berhingga, maka matriks probabilitas transisi berukuran $n \times n$. Setiap elemen matriks adalah positif $\left(p_{ij} \geq 0 \text{ untuk } i, j = 0,1,2,...\right)$. Total probabilitas dalam setiap baris adalah satu, yaitu $\sum_{j=0}^{\infty} p_{ij} = 1$ untuk setiap baris $i = 0,1,2,...$.

4.4 Persamaan Chapman Kolmogorov

Probabilitas transisi satu langkah didefinisikan dengan p_{ij}. Sekarang akan didefinisikan probabilitas transisi $n-$langkah, dinotasikan p_{ij}^{n}, yaitu , probabilitas proses berpindah dari keadaan i ke keadaan j setelah n langkah dituliskan

$$p_{ij}^{n} = P(X_{n+k} = j \mid X_k = i), n \geq 0, i, j \geq 0$$

Untuk menyelesaikan probabilitas transisi $n-$langkah dapat menggunakan persamaan Chapman Kolmogorov, yaitu

$$p_{ij}^{n+m} = \sum_{k=0}^{\infty} p_{ik}^{n} p_{kj}^{m}, \text{ untuk setiap } n, m \geq 0, \text{ setiap } i, j \geq 0$$

Bukti:

$$
\begin{aligned}
p_{ij}^{n+m} &= P(X_{n+m} = j \mid X_0 = i) \\
&= \sum_{k=0}^{\infty} P(X_{n+m} = j, X_n = k \mid X_0 = i) \\
&= \sum_{k=0}^{\infty} P(X_{n+m} = j \mid X_n = k, X_0 = i) P(X_n = k \mid X_0 = i) \\
&= \sum_{k=0}^{\infty} p_{kj}^{m} p_{ik}^{n}.
\end{aligned}
$$

Dalam bentuk matriks, p_{ij}^{n} dituliskan sebagai $P^{(n)}$. Dengan demikian, persamaan Chapman Kolmogorov dapat dituliskan dalam bentuk matriks, yaitu

$P^{(n+m)} = P^{(n)} P^{(m)}$.

Misalnya, p_{ij}^{n} adalah probabilitas keadaan j pada waktu n yang berasal dari keadaan awal i waktu ke nol. Misalkan dinotasikan probabilitas awal saat keadaan i, yaitu $\pi_i = P(X_0 = i)$ dan vektor probabilitas awal $\pi = (\pi_0, \pi_1, ...)$. Dengan begitu,

$$
\begin{aligned}
P(X_0 = i_0, ..., X_n = i, X_{n+1} = j) &= P(X_0 = i_0, ..., X_n = i) \times P(X_{n+1} = j \mid X_0 = i_0, ..., X_n = i) \\
&= P(X_0 = i_0, ..., X_n = i) \times P(X_{n+1} = j \mid X_n = i) \\
&= P(X_0 = i_0) p_{i_0 i_1} p_{i_1 i_2} \cdots p_{ij},
\end{aligned}
$$

dan probabilitas tak bersyarat dapat dihitung dengan mensyaratkan pada keadaan awal, yaitu

$$
\begin{aligned}
P(X_n = j) &= \sum_{i=0}^{\infty} P(X_n = j \mid X_0 = i) P(X_0 = i) \\
&= \sum_{i=0}^{\infty} p_{ij}^{n} \pi_i.
\end{aligned}
$$

4.5 Rantai Markov Waktu Kontinu

Berdasarkan definisi rantai markov, rantai markov waktu kontinu merupakan proses stokastik dari variabel random $\{X(t), t \geq 0\}$ yang memenuhi sifat Markov. Dengan demikian, probabilitas transisi dari keadaan i ke keadaan j pada waktu $t \geq 0$,

$$p_{ij}(t) = P(X(t+s) = j \mid X(s) = i)$$

dengan $0 \leq p_{ij} \leq 1, \sum_j p_{ij}(t) = 1, \forall j$.

Matriks probabilitas transisi dinotasikan oleh:

$$P(t) = \begin{bmatrix} p_{00}(t) & p_{01}(t) & \cdots & p_{0j}(t) \\ p_{10}(t) & p_{11}(t) & \cdots & p_{1j}(t) \\ \vdots & \vdots & \ddots & \vdots \\ p_{i0}(t) & p_{i1}(t) & \cdots & p_{ij}(t) \end{bmatrix}$$

dengan $p_{ij}(t) \geq 0$ dan $\sum_j p_{ij}(t) = 1$.

Dinotasikan probabilitas pada keadaan j saat waktu ke t oleh

$$\pi_j(t) = P(X(t) = j),$$

vektor $\pi(t) = (\pi_0(t), \pi_1(t), \ldots)$ merupakan vektor probabilitas dari keadaan pada waktu t, dan $\pi(0)$ adalah vektor probabilitas awal. Dengan demikian,

$$\begin{aligned} \pi_j(t) &= \sum_i P(X(t+s) = j \mid X(s) = i) P(X(s) = i) \\ &= \sum_i p_{ij}(t) P(X(0) = i) \\ &= \sum_i p_{ij}(t) \pi_i(0). \end{aligned}$$

Bentuk persamaan Chapman Kolmogorov untuk rantai markov waktu kontinu untuk probabilitas transisi dari keadaan i ke keadaan j pada waktu $t+s$ sebagai berikut:

$$p_{ij}(t+s) = \sum_{k=0}^{\infty} p_{ik}(t)\, p_{kj}(s) \text{ untuk setiap } t, s \geq 0 .$$

Dalam bentuk matriks dapat ditulis sebagai $P(t+s) = P(t)P(s)$.

$$\begin{aligned} P(X(0)=i_0,\ldots,X(s)=i,X(t+s)=j) &= P(X(0)=i_0,\ldots,X(s)=i)\times P(X(t+s)=j \mid X(0)=i_0,\ldots,X(s)=i) \\ &= P(X(0)=i_0,\ldots,X(s)=i)\times P(X(t+s)=j \mid X(s)=i) \\ &= \pi_{i_0}(0)p_{i_0i_1}(t)p_{i_1i_2}(t)\ldots p_{ij}(t). \end{aligned}$$

DNA memiliki empat nukleotid, yaitu A, C, G, dan T. Keempat nukleotid tersebut digunakan sebagai himpunan dari keadaan atau *state* yang mungkin terjadi. Dengan demikian, $S = \{A, C, G, T\}$, sehingga bentuk matriks probabilitas transisinya yaitu:

$$P(t) = \begin{bmatrix} p_{AA}(t) & p_{AC}(t) & p_{AG}(t) & p_{AT}(t) \\ p_{CA}(t) & p_{CC}(t) & p_{CG}(t) & p_{CT}(t) \\ p_{GA}(t) & p_{GC}(t) & p_{GG}(t) & p_{GT}(t) \\ p_{TA}(t) & p_{TC}(t) & p_{TG}(t) & p_{GG}(t) \end{bmatrix}$$

dengan $t \geq 0$, $P(t) \geq 0$ dan nilai probabilitas jumlah elemen setiap barisnya sama dengan satu.

Substitusi nukleotid dari *state* i ke *state* j merupakan rantai Markov dengan waktu kontinu. Dengan begitu, perubahan nukleotid dari *state* i ke *state* j selama $t+s$ waktu itu berarti perubahan nukleotid dari *state* i ke *state* k selama t waktu, kemudian dilanjutkan dari *state* k ke

state j selama s waktu. Dengan demikian, berdasarkan persamaan Chapman Kolmogorov dapat dituliskan bahwa

$$p_{ij}(t+s) = \sum_{k=0}^{\infty} p_{ik}(t)\, p_{kj}(s),$$

atau dalam bentuk matriks dapat ditulis sebagai

$$P(t+s) = P(t)P(s) \text{ untuk } t, s \geq 0 .$$

Karena substitusi nukleotid tersebut merupakan rantai Markov dengan waktu kontinu, dapat dikatakan sebagai rantai Markov dengan waktu kontinu regular. Hal ini berarti bahwa untuk waktu ke $t = 0$, $P(0) = I$ yang mana I merupakan matriks identitas ukuran 4×4, karena

$$p_{ij}(t) = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases},$$

sehingga

$$P(0) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Dan juga, $P(t)$ diferensiabel untuk $t \geq 0$. Artinya, $P(t)$ memiliki turunan pertama untuk $t \geq 0$, sehingga ditulis $P'(t)$. Berdasarkan definisi turunan, untuk $h > 0$ dan $t \geq 0$

$$P'(t) = \lim_{h \to 0} \frac{P(t+h) - P(t)}{h}.$$

Berdasarkan persamaan Chapman Kolmogorov dan $P(0) = I$, diperoleh

$$\begin{aligned} P'(t) &= \lim_{h \to 0} \frac{P(t)(P(h) - P(0))}{h} \\ &= P(t) \lim_{h \to 0} \frac{P(0+h) - P(0)}{h} \\ &= P(t)P'(0). \end{aligned}$$

Untuk memperoleh $P(t)$, dilakukan dengan mengintegralkan nilai dari $P'(t)$. Sebelumnya, $P'(t)$ dapat ditulis sebagai $\frac{d(P(t))}{dt}$. Dengan demikian,

$$\frac{d(P(t))}{dt} = P(t)P'(0)$$
$$\frac{d(P(t))}{P(t)} = P'(0)dt.$$

Selanjutnya, dengan mengintegralkan kedua ruas persamaan dan menggunakan bilangan natural e, diperoleh

$$\begin{aligned} \int \frac{d(P(t))}{P(t)} &= \int P'(0)dt \\ \ln P(t) &= P'(0)t + c \\ P(t) &= e^{P'(0)t+c} \\ P(t) &= e^{P'(0)t}e^{c} \\ P(t) &= ce^{P'(0)t}. \end{aligned}$$

Karena $P(0)=1$, maka diperoleh nilai $c=1$. Akibatnya, $P(t)=e^{P'(0)t}$.

Misalkan $Q=P'(0)$. Dengan begitu, $P(t)=e^{Qt}$.

Rantai Markov dengan waktu kontinu harus memenuhi syarat memiliki distribusi probabilitas stasioner dan bersifat *time-reversible*. Berikut penjelasannya.

a. Distribusi probabilitas stasioner bersifat tunggal

Definisi 4.1 Vektor $\varphi=(\varphi_A,\varphi_C,\varphi_G,\varphi_T)$ dengan $\varphi_i \geq 0$ dan $\sum_{i\in S}\varphi_i=1$ disebut distribusi probabilitas stasioner dari rantai Markov jika $\varphi Q=0$. Hal ini ekuivalen dengan $\varphi P(t)=\varphi$.

Berikut beberapa sifat distribusi probabilitas stasioner.

1) Probabilitas saat berada di *state i* pada waktu t adalah $p_i(t)=\sum_{k\in S}\pi_k p_{ki}(t)$ dengan $\pi=(\pi_A,\pi_C,\pi_G,\pi_T)$ merupakan vektor probabilitas awal. Misalkan $\pi=\varphi$, dengan begitu $p_i(t)=\sum_{k\in S}\varphi_k p_{ki}(t)=\varphi_i$.
2) Matriks transisi probabilitas untuk $t\to\infty$,

$$P(t)\to\begin{bmatrix}\varphi_A & \varphi_C & \varphi_G & \varphi_T\\ \varphi_A & \varphi_C & \varphi_G & \varphi_T\\ \varphi_A & \varphi_C & \varphi_G & \varphi_T\\ \varphi_A & \varphi_C & \varphi_G & \varphi_T\end{bmatrix}.$$

Artinya, untuk nilai t yang semakin besar, $P(t)$ akan stabil dengan distribusi probabilitas stasioner yang berkaitan dengannya.

b. Bersifat *time-reversible*

Definisi 4.2 Misalkan π vektor probabilitas awal dari rantai Markov dan $p_i(t) \neq 0$ untuk $i \in S$ dan $t \geq 0$. Didefinisikan rantai Markov *reversed* sebagai rantai Markov dengan waktu kontinu yang diberikan oleh matriks probabilitas transisi $P^*(t)$ dengan

$$p_{ij}^*(t) = \frac{\pi_j p_{ji}(t)}{p_i} \text{ untuk setiap } i, j .$$

Misalkan setiap anggota distribusi probabilitas stasioner φ tidak sama dengan nol. Oleh karena $\pi = \varphi$, diperoleh untuk setiap i, j

Definisi 4.3 Rantai markov yang memenuhi asumsi pada Definisi 4.2 disebut *time-reversible* atau *simple reversible* jika $P^*(t) = P(t)$ untuk $t \geq 0$.

Apabila setiap anggota dari φ tidak sama dengan nol. Oleh karena $\pi = \varphi$, reversible artinya

$$\varphi_i p_{ij}(t) = \varphi_j p_{ji}(t)$$

4.6 Metode Maximum Likelihood

Pada evolusi, mutasi merupakan perubahan peristiwa. Probabilitas menemukan mutasi sepanjang satu cabang pada pohon filogenetik dapat dihitung dengan menggunakan metode maximum likelihood. Ide utama dari pembentukan pohon filogenik dengan metode maximum likelihood adalah untuk menentukan topologi pohon, panjang cabang, dan parameter dari model evolusi yang memaksimumkan probabilitas dari sekuen (Isaev, 2006). Dengan kata lain, fungsi likelihood adalah probabilitas bersyarat dari data yang diberikan. Dengan demikian,

$$L(\tau \mid \theta) = P(Data \mid \tau, \theta)$$
$$= P(\text{align sekuen} \mid \text{pohon, model evoolusi}).$$

MLE dari τ dan θ yaitu $\hat{\tau}$ dan θ, adalah membuat fungsi Likelihood sebisa mungkin menjadi:

$$\hat{\tau}, \hat{\theta} = \arg\max_{\tau,\theta} L(\tau, \theta).$$

4.7 Metode Maximum Likelihood Untuk Dua Sekuen

Misalkan terdapat dua sekuen, yang mana evolusi sekuen ini mengikuti model Jukes Cantor. Keduanya saling independen dan memiliki tingkat evolusi yang sama. *Alignment* memiliki panjang l untuk dua sekuen, yaitu $s_1 = (x_1, x_2, ..., x_n)$ dan $s_2 = (x_1, x_2, ..., x_n)$ dengan . Terdapat *root* antara setiap sekuen, dengan jarak d_1 untuk sekuen s_1 dan jarak d_2 untuk sekuen s_2, sehingga jarak antara sekuen s_1 dan s_2 yaitu $d = d_1 + d_2$.Fungsi likelihood untuk $1 \le k \le n$ dengan *time-reversible* sebagai berikut:

$$L_k = \sum_{j \in S} \varphi_j p_{jx_k}(d_1) p_{jy_k}(d_2)$$
$$= \varphi_{x_k} p_{x_k y_k}(d).$$

dengan φ_j merupakan elemen dari distribusi probabilitas stasioner. Dengan begitu,bentuk fungsi likelihoodnya sebagai berikut:

$$L = L_1 L_2 .. L_n$$
$$= \varphi_{x_1} p_{x_1 y_1}(d) \varphi_{x_2} p_{x_2 y_2}(d) ... \varphi_{x_n} p_{x_n y_n}(d).$$

Dalam model evolusi, d didefinisikan sebagai waktu t. Dengan demikian, apabila digunakan matriks probabilitas transisi, maka bentuk fungsi *likelihood*-nya adalah:

$$L = \varphi_{x_1} p_{x_1 y_1}(t) \varphi_{x_2} p_{x_2 y_2}(t) ... \varphi_{x_n} p_{x_n y_n}(t).$$

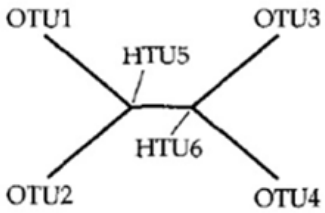
Probabilitas transisi dari sekuen satu ke sekuen yang lain merupakan jumlah probabilitas dari semua kemungkinan jalur yang menghubungkan kedua sekuen tersebut. Jalur khusus dari transisi satu sekuen ke sekuen yang lain dapat dinyatakan sebagai penyejajaran (Thorne, dkk, 1991).

4.8 Pohon Maximum Likelihood Untuk Empat Sekuen

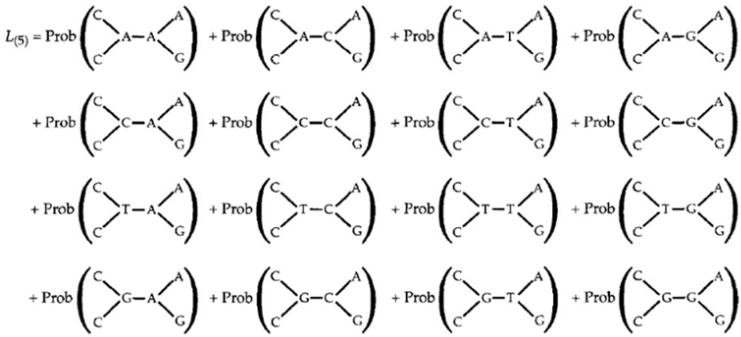
Berikut akan dijelaskan tahapan dalam pembentukan pohon *maximum likelihood* dengan menggunakan 4 taxa/ OTU. Sebagai contoh penyejajaran sekuen untuk ke-4 OTU tersebut:

OTU 1 = AACGCCTTT...N
OTU 2 = AACGCGTTA...N
OTU 3 = AACCAGTTT...N
OTU 4 = AACCGGTTT...N

Pada situs kelima, berturut-turut pada OTU 1, 2, 3, dan 4 memiliki nukleotida C, C, A, dan G. Dengan contoh situs ke-5 ini metode Maximum Likelihood merekonstruksi kekerabatan keempat OTU dengan membuat sebuah pohon *unrooted* dan kemudian memperhitungkan probabilitas nukleotida *ancestor* (5 & 6) yang menghasilkan state nukleotida seperti berikut:



Karena nucleotida ada 4 yaitu A, C, T, dan G. Dengan demikian, jumlah kombinasi nukleotida untuk dua ancestral state tersebut ada 16 (AA, AG, AC, AT, CA, CC, CT, CG, TA, TC, TT, TG, GA, GC, GT, GG).Probabilitas dari seluruh kombinasi tersebut dihitung dengan bantuan data penyejajaran sekuen yang ada dan kemudian diakumulasi menjadi nilai likelihood untuk situs nomor 5 ($\mathbf{L_5}$). Sehingga untuk *site* ke-5 nilai *likelihood*-nya adalah:



Selanjutnya metode Maximum Likelihood akan menghitung nilai *likelihood* untuk satu pohon berdasarkan seluruh situs dalam alignment. Dengan demikian nilai likelihood (L) dari satu pohon dirumuskan dengan:

$$L = L_{(1)} \times L_{(2)} \times .... \times L_{(n)} = \prod_{i=1}^{n} L_{(i)}$$

Pada umumnya nilai L ini sangat kecil sehingga dinyatakan dalam bentuk logaritmik (lnL). Dengan demikian, rumusnya menjadi:

$$\ln L = \ln L_{(1)} + \ln L_{(2)} + ... + \ln L_{(n)} = \sum_{i=1}^{n} \ln L_{(i)}$$

Rumus tersebut hanya mencari nilai *likelihood* dari satu pohon saja.

4.9 **Kemungkinan Pohon** Likelihood

Dalam rekonstruksi filogenetik, tugas yang berat adalah menemukan sebuah pohon di antara semua struktur pohon yang mungkin dengan memaksimalkan kemungkinan global. Sayangnya, belum ada algoritma yang efisien untuk menjamin lokalisasi pohon terbaik dari semua topologi pohon yang mungkin. Sejumlah pohon topologi biner tak berakar meningkat dengan sejumlah taksa (n), yang dapat dihitung menurut:

$$t_n = \frac{(2n-5)!}{2^{n-3}(n-3)!} = \prod_{i=1}^{n}(2i-5)$$

Saat menghitung pohon maximum likelihood, parameter model dan panjang cabang harus dihitung untuk setiap pohon, dan kemudian pohon yang menghasilkan nilai likelihood tertinggi akan dipilih. Nilai MLE yang paling tinggi menandakan bahwa pohon tersebut dapat menjelaskan penyejajaran sekuen dengan lebih baik. Karena banyaknya topologi pohon, uji coba semuapohon yang mungkin tidaklah efektif, dan juga secara komputasi tidak memungkinkan. Dengan demikian, berbagai metode heuristik digunakan untuk menyarankan pohon yang dipilih. (Lemey et al, 2009)

4.10 **Pohon Terbaik Dengan Metode Pencarian Heuristik**

Stepwise edition adalah heuristik pertama untuk mencari kemungkinan maximum likelihood dari sebuah pohon. Prosedur dimulai dari topologi pohon tidak berakar dari tiga taksa yang dipilih secara acak dari daftar n taxa. Kemudian salahsatunya merekonstruksi pohon maximum *likelihood* yang sesuai. Untuk memperpanjang pohon ini, secara acak dipilih salah satu dari n - 3 taksa yang tersisa. Takson ini

kemudian dimasukkan ke dalam masing-masing cabang pohon terbaik. Artinya, penyisipan cabang terjadi pada likelihood tertinggi. Dengan demikian, kriteria keputusan lokal memilih pohon dengan likelihood tertinggi dari daftar 2k - 3 pohon, jika k taksa sudah berada pada sub-pohon. Pohon yang dihasilkan kemudian akan digunakan untuk mengulangi prosedur. Setelah langkah n - 3, diperoleh pohon maximum likelihood, yang optimal secara lokal. Artinya, urutan penyisipan taksa dan kriteria keputusan lokal yang diberikan menunjukkan tidak ada pohon yang lebih baik. Akan tetapi pada *stepwise edition* ini hanya dihitung maximum *likelihood* untuk $\sum_{i=3}^{n}(2i-5) = (n-2)^2$ pohon. Dengan demikian, ada kemungkinan bahwa urutan penyisipan dari taksa yang lain akan memberikan pohon dengan likelihood yang lebih tinggi. Untuk mengurangi risiko terjebak dalam hal optimal lokal tersebut, disarankan untuk menggunakan *full tree rearrangement.*

Operasi *full-tree rearrangement* mengubah struktur pohon tertentu dengan n daun-daun. Prinsip dari operasi ini adalah: dari pohon awal, sejumlah pohondihasilkan sesuai dengan aturan yang ditentukan. Untuk setiap pohon yang dihasilkan, dihitung nilai maximum likelihood. Pohon dengan likelihood tertinggi kemudian digunakan untuk mengulangi prosedur. Operasi *full-tree rearrangement* ini berhenti jika tidak ditemukan pohon yang lebih baik. Pohon ini kemudian dikatakan sebagai pohon optimal lokal. Namun, peluang untuk benar-benar menentukan pohon yang optimal secara global tergantung pada data dan ukuran lingkungannya.

Tiga operasi *full-tree rearrangement* yang saat ini populer adalah: ***Nearest Neighbor interchange (NNI), sub-tree***

pruning and regrafting (SPR) dan ***tree-bisection and reconnection (TBR).***

4.11 DNAml DAN MBEToolbox

MBEToolbox, merupakan *toolbox* yang disajikan untuk memenuhi kebutuhan dalam manipulasi sekuen, estimasi jarak genetik dan inferensi filogeni di bawah lingkungan MATLAB. Selain itu, *toolbox* ini menyediakan kerangka fungsional yang diperluas untuk merumuskan dan memecahkan masalah dalam analisis data evolusioner. Fungsi utama yang diimplementasikan adalah: manipulasi sequence, perhitungan jarak evolusioner yang berasal dari model substitusi nukleotida, asam amino atau kodon, konstruksi pohon filogenetik, statistik sekuen dan fungsi grafik untuk memvisualisasikan hasil analisis.

4.12 Pohon Filogenetik Epidemi SARS dengan Metode Maximum Likelihood

Sebagaimana pada metode sebelumnya, data yang digunakan pada metode maximum likelihood ini juga sejumlah 14 sekuen DNA virus SARS. Data diambil secara online di genbank, database gen terbesar di dunia milik pemerintah Amerika Serikat. Adapun pengambilannya dengan mengakses National Center for Biotechnology Information (www.ncbi.nlm.nih.gov). Adapun tahapan untuk menentukan pohon filogenetik dengan MBBEToolbox adalah sebagai berikut:

1. Mencari jarak genetik dengan model evolutioner Jukes Cantor

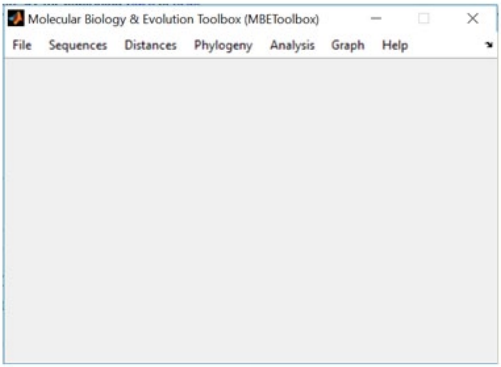
   Pada tahap ini dilakukan penyejajaran antar sekuen untuk mengetahui jarak genetik dari sekuen satu ke sekuen lain. Jarak genetik tersebut selanjutnya diubah menjadi jarak evolutioner dengan model Jukes Cantor.

2. Menghitung nilai Likelihood
   Berdasarkan keseluruhan site dari 14 sekuen tersebut selanjutnya dihitung kemungkinan nilai likelihood untuk masing-masing pohon. Artinya, akan ada banyak kemungkinan pohon yang terjadi. Sehingga agar proses komputasi tetap berjalan, digunakan metode heuristik *Stepwise Addition* untuk membatasi jumlah pohon yang dihitung.
3. Diperoleh Nilai Likelihoood tertinggi
   Nilai likelihood yang paling maksimum tersebut mengindikasikan pohon mana yang bisa menjelaskan pohon filogenetik yang terbaik.
4. Kesimpulan
   Diperoleh pohon filogenetik epidemi SARS menggunakan metode maximum likelihood beserta nilai likelihood, panjang masing-masing percabangan serta aproksimasi interval konfidensi.

Hasil sementara yang telah dicapai dalam proses inferensi filogenetik menggunakan metode maximum likelihood ini dengan memanfaatkan *Molecular Biology and Evolution Toolbox* (MBEToolbox) yang ditulis dalam Matlab (Amiroch, dkk, 2018). Sebagaimana paket program dalam Phylip, ada paket DNAml yang mengestimasi pohon filogenetik dari sekuen nukleotida dengan maximum likelihood. Model yang digunakan memungkinkan frekuensi yang diharapkan tidak sama dari empat nukleotida, tingkat transisi dan transversi yang tidak sama, dan rata-rata tingkat perubahan yang berbeda dalam berbagai kategori situs, serta penggunaan Hidden Markov Model.Model ini juga memungkinkan penggunaan distribusi gamma dan distribusi gamma-plus-invariant sites. Sedangkan MBEToolbox_DNMAL merupakan versi modifikasi DNAml pada Phylip untuk MBEToolbox yang ditulis dalam Matlab. (James J Cai etal, 2005).

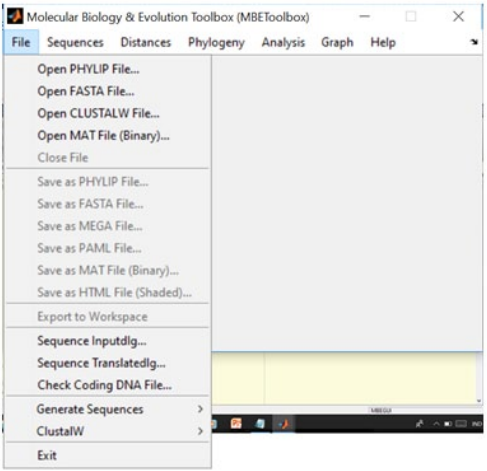
Berikut adalah hasil pembentukan pohon filogenetik dengan metode maximum *likelihood*.

1. Pada saat program *dirunning*, muncul menu tampilan *Molecular Biology and Evolution Toolbox* seperti berikut:



Gambar 7.5 Tampilan Menu MBEToolbox

Selanjutnya klik menu ***File*** dan pilih ***Open FASTA file.*** Pada layar *command window* akan muncul nama-nama file yang telah teridentifikasi pada data user.



Gambar 7.6 Tampilan menu pop-up pada File

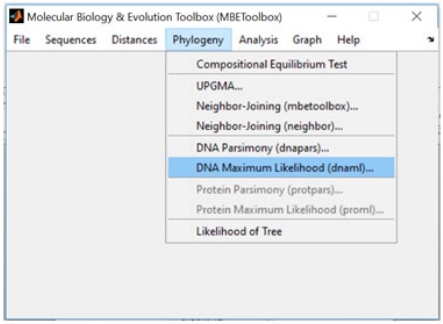
2. Berikutnya klik menu ***Distance* dan pilih JC69.** Dari sini pada *command window* akan ditampilkan jarak evolutioner Jukes Cantor sebagaimana yang diinginkan.

Tabel 7.1 Jarak evolutioner Jukes Cantor

	A	B	C	D	E	F	G	H	I	J	K	L	M
A													
B	4.683												
C	4.025	3.809											
D	3.961	2.588	4.071										
E	2.715	4.815	531.297	5.035									
F	5.301	0.049	4.208	2.608	4.880								
G	5.263	0.049	4.217	2.611	4.837	0.000							
H	4.952	2.670	3.334	2.500	3.839	2.682	2.677						
I	4.720	2.671	3.340	2.495	3.856	2.683	2.670	0.000					
J	0.049	531.297	4.875	3.923	2.556	4.248	4.248	5.516	5.465				
K	4.965	2.672	3.343	2.497	3.856	2.684	2.678	0.000	0.000	5.516			
L	5.263	0.049	4.198	2.610	4.880	0.000	0.000	2.682	2.675	4.248	2.682		
M	5.494	0.045	3.912	2.589	4.802	0.043	0.043	2.684	2.685	4.961	2.686	0.042	
N	531.297	2.537	3.677	2.818	3.881	2.552	2.552	3.252	3.252	531.297	3.232	2.552	2.552

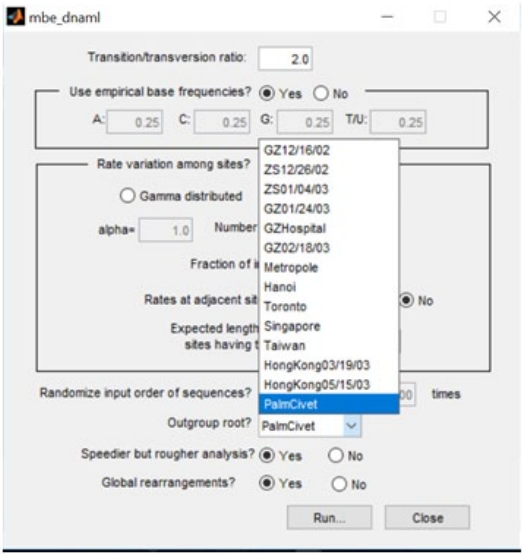
Dimana A mewakili sekuen 1 yakni GuangZhao12/16/02, B mewakili sekuen 2 yaitu ZongSan12/26/02, C mewakili sekuen 3 ZongSan01/04/03, D mewakili sekuen 4 GuangZhao 01/04/03, E mewakili sekuen 5 GuangZhou Hospital, F mewakili sekuen 6 GuangZhou02/18/03, G mewakili sekuen 7 Metropole, H mewakili sekuen 8 Hanoi, I mewakili sekuen 9 Toronto, J mewakili sekuen 10 Singapore, K mewakili sekuen 11 Taiwan, L mewakili sekuen 12 HongKong03, M mewakili sekuen 13 HongKong05, N mewakili sekuen 14 yaitu Palm Civet.

3. Langkah selanjutnya klik menu Phylogeny dan pilih DNAml.



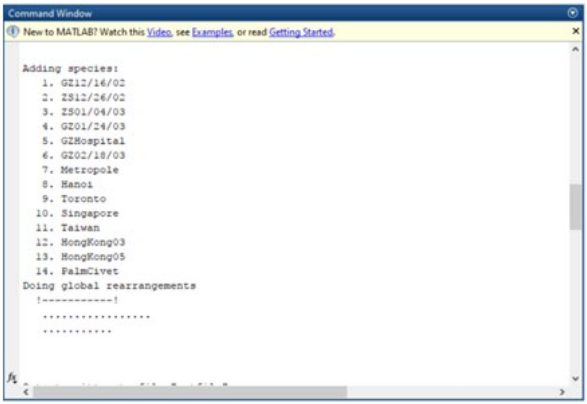
Gambar 7.7 Tampilan menu *pop-up* Phylogeny

Tampak tampilan GUI pada MBE untuk DNAml sebagai berikut:



Gambar 7.8 Tampilan GUI untuk MBE_DNAml

Pada Gambar 7.8 rasio transisi/tranversi diisi 2 sesuai default sedangkan untuk outgroup root dipilih PalmCivet karena Palm Civet merupakan host dari epidemi SARS. Proses berawal dengan identifikasi ke-14 sekuen.

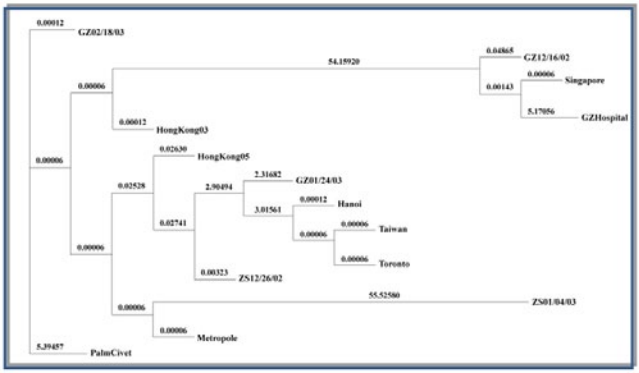


Gambar 7.9 Tampilan identifikasi seluruh sekuen

Selanjutnya muncul frekuensi dasar empiris untuk sekuen nucleotide pada metode maximum likelihood tersebut yaitu:

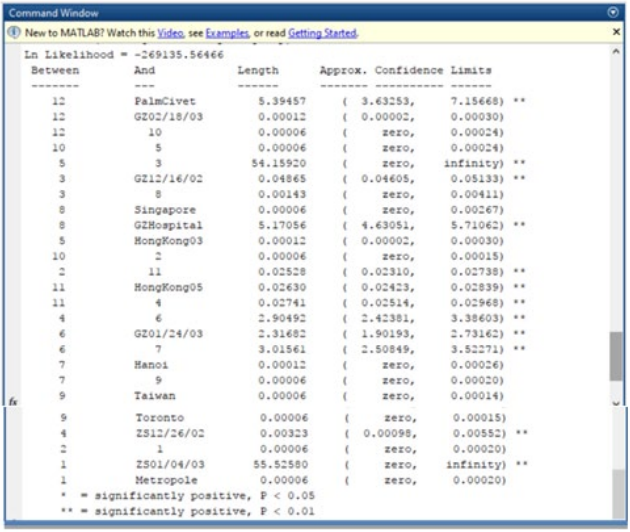
Nucleotide	A	C	G	T
frequencies	0.28476	0.19982	0.20794	0.30748

Sedangkan untuk topologi pohon, diperoleh pohon tak berakar berikut ini:



Gambar 7.10 Pohon Filogenetik maximum likelihood

Dengan nilai ln Likelihood = – 269135,56466.



Gambar 7.11 Panjang cabang dan aproksimasi interval konfidensi

Nilai ln likelihood sebesar – 269135,56466 menunjukan nilai estimasi optimal (maximum) likelihood yang dilakukan sebesar 269135,56466. Nilai ini menunjukan besarnya nilai estimasi yang paling maksimum untuk panjang cabang. Selanjutnya nilai ini digunakan untuk menentukan pohon yang dapat menjelaskan pohon fiogenetik yang lebih baik. Meskipun nilai ln likelihoodnya negatif, hal itu berarti probabilitas yang sesuai kurang dari 1 karena yang dilihat adalah nilai logaritmanya.

Pada tabel di atas, disebutkan antara dua percabangan terdapat panjang cabang (misalnya 12 dengan palm civet, panjang cabang 5.39457) berada pada approximasi confidence limits tertentu, artinya jarak tersebut berada dalam interval konfidensi 3,63253 – 7,15668. Nilai positif pada interval konfidensi berarti pada percabangan tersebut tidak perlu di arrangement, karena nilai intervalnya masih berada dibawah estimasi, dalam arti interval yang sempit menunjukkan panjang cabang lebih akurat. Keakuratan tersebut juga diperkuat dengan signifikansi positive dengan nilai $p < 0,01$. Meski demikian, pada beberapa percabangan, aproksimasi konfidensi limitnya mempunyai signifikansi positif dengan $p < 0,05$.

4.13 Soal-soal Latihan

1. Tunjukkan dengan skema, konstruksi HMM order-1 untuk merepresentasikan informasi dari sekuen berikut:

   GCCGCGCTTG
   GCTTGGTGGC
   TGGCCGTTGC

2. Dengan sekuen yang sama pada nomer 1, hitunglah nilai estimasi likelihood dari A, C, G, dan T.
3. Konstruksikan pohon filogenetik dengan MEGA X, gunakan metode neighbor-joining dan metode maximum likelihood dengan 8 data sekuen SARS dari 8 negara yang berbeda. Bandingkan!

BAB V

PEMBENTUKAN POHON FILOGENETIK DENGAN METODE BAYESIAN

Metode berbasis probabilitas selain Maximum Likelihood adalah metode Bayesian. Metode Bayesian merupakan metode yang menggunakan rasio likelihood untuk menentukan hipotesis (pohon) mana yang lebih mampu menjelaskan data sequence yang ada. Sebelum mempelajari metode Bayesian ini, perlu dikenal beberapa definisi dan teorema yang mendasarinya.

5.1 Variabel Random

Variabel random merupakan fungsi berharga real yang didefinisikan pada suaatu ruang sampel. Variabel random dibedakan menjadi dua, yaitu variabel random diskret dan variabel kontinu.

Definisi 5.1 Variabel random X disebut variabel random diskret jika nilai dari ruang sampelnya berhingga atau tak terhingga terhitung (*countable*). (Wackerly, Madenhall, dan Scheaffer, 2008)

Definisi 5.2 Misalkan X variabel random diskret dengan ruang sampel S. Fungsi massa probabilitas (f.m.p) dari X adalah

$$p(x) = P(X = x), x \in S.$$

(Hogg, McKean, Craig, 2005)

Teorema 5.1 Fungsi $p(x)$ adalah fungsi massa probabilitas dari variabel random diskret X jika hanya jika memenuhi (Wackerly, Madenhall, dan Scheaffer, 2008).

1. $0 \le p(x) \le 1, x \in S$ dan
2. $\sum_{x \in S} p(x) = 1$. .

Definisi 5.3 Variabel random X disebut variabel random kontinu apabila fungsi distribusi kumulatif (CDF) dari X, yaitu $F(x)$ merupakan fungsi kontinu untuk setiap $x \in R$. (Hogg, McKean, Craig, 2005).

Definisi 5.4 Misalkan $F(x)$ merupakan fungsi distribusi kumulatif untuk variabel random kontinu X, maka $f(x)$ diberikan oleh

$$f(x) = \frac{dF(x)}{dx} = F'(x)$$

disebut fungsi densitas probabilitas (p.d.f) dari variabel random kontinu X. (Wackerly, Madenhall, dan Scheaffer, 2008)

Berdasarkan Definisi 5.3 dan 5.4, fungsi distribusi kumulatif dari variabel random kontinu X dapat ditulis

$$F(x) = \int_{-\infty}^{x} f(t)dt .$$

Teorema 5.2 Jika $f(x)$ merupakan fungsi densitas probabilitas dari variabel random kontinu X, maka (Wackerly, Madenhall, dan Scheaffer, 2008)

1. $f(x) \ge 0, x \in R$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

Definisi 5.5 Misalkan X_1 dan X_2 merupakan variabel random diskret. Fungsi massa probabilitas gabungan dari X_1 dan X_2 yaitu

$$p(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$$

untuk $-\infty < x_1 < \infty$ dan $-\infty < x_2 < \infty$. (Wackerly, Madenhall, dan Scheaffer, 2008)

Teorema 5.3 Jika X_1 dan X_2 merupakan variabel random diskret dan $p(x_1, x_2)$ adalah fungsi massa probabilitas gabungannya, maka (Wackerly, Madenhall, dan Scheaffer, 2008)

1. $0 \le p(x_1, x_2) \le 1$
2. $\sum_{x_1, x_2} p(x_1, x_2) = 1$

Definisi 5.6 Diberikan X_1 dan X_2 merupakan variabel random kontinu dengan fungsi distribusi gabungan $F(x_1, x_2)$. Jika terdapat fungsi non negatif $f(x_1, x_2)$ sedemikian sehingga

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(t_1, t_2) dt_2 dt_1,$$

untuk $-\infty < x_1 < \infty$ dan $-\infty < x_2 < \infty$, maka X_1 dan X_2 disebut variabel random kontinu gabungan. Fungsi $f(x_1, x_2)$ disebutfungsi densitas probabilitas gabungan. (Wackerly, Madenhall, dan Scheaffer, 2008).

Teorema 5.4 Jika X_1 dan X_2 merupakan variabel random kontinu gabungan dengan fungsi densitas gabungan $f(x_1, x_2)$, maka (Wackerly, Madenhall, dan Scheaffer, 2008).

1. $f(x_1, x_2) \ge 0$ untuk setiap x_1, x_2
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 dx_1 = 1.$

Definisi 5.7 (Wackerly, Madenhall, dan Scheaffer, 2008)

a. Diberikan X_1 dan X_2 variabel random diskret gabungan dengan fungsi probabilitas $p(x_1, x_2)$, maka fungsi probabilitas marginal dari X_1 dan X_2 yaitu

$$p(x_1) = \sum_{x_2} p(x_1, x_2) \quad \text{dan} \quad p(x_2) = \sum_{x_1} p(x_1, x_2).$$

b. Diberikan X_1 dan X_2 variabel random kontinu gabungan dengan fungsi probabilitas $f(x_1, x_2)$, maka fungsi probabilitas marginal dari X_1 dan X_2 yaitu

$$f(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \quad \text{dan} \quad f(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1.$$

Definisi 8 Misalkan variabel random X_1 dan X_2 mempunyai fungsi probabilitas gabungan $p(x_1, x_2)$ atau $f(x_1, x_2)$ dan fungsi probabilitas marginal $p(x_1)$ atau $f(x_1)$ dan $p(x_2)$ atau $f(x_2)$. Variabel random X_1 dan X_2 disebut independen jika hanya jika
$p(x_1, x_2) = p(x_1)p(x_2)$ atau $f(x_1, x_2) = f(x_1)f(x_2)$.
(Hogg, McKean, Craig, 2005).

5.2 Distribusi Probabilitas Poisson

Variabel random diskret X dikatakan berdistribusi Poisson dengan parameter $\theta > 0$ jika memiliki fungsi massa probabilitas dengan bentuk (Wackerly, Madenhall, dan Scheaffer, 2008)

$$p(x) = \frac{e^{-\theta}\theta^x}{x!}, x = 0, 1, 2, \ldots$$

Notasi khusus untuk variabel random X berdistribusi Poisson dengan parameter θ adalah

$$X : POI(\theta).$$

Karena $\theta > 0$, maka $p(x) > 0$ untuk $x = 0,1,2,...$ dan $p(x) = 0$ untuk yang lainnya. Selanjutnya,

$$\begin{aligned}\sum_{x=0}^{\infty} p(x) &= \sum_{x=0}^{\infty} \frac{e^{-\theta}\theta^x}{x!} \\ &= e^{-\theta} \sum_{x=0}^{\infty} \frac{\theta^x}{x!} \\ &= e^{-\theta} e^{\theta} \\ &= 1.\end{aligned}$$

Dengan begitu, Teorema 5.1 telah terpenuhi, artinya $p(x)$ merupakan fungsi massa probabilitas.

5.3 Distribusi Probabilitas Gamma

Variabel random X dikatakan berdistribusi Gamma dengan parameter skala $\alpha > 0$ dan parameter bentuk $\beta > 0$ jika hanya jika fungsi densitas probabilitas X yaitu (Wackerly, Madenhall, dan Scheaffer, 2008)

$$f(x) = \begin{cases} \dfrac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, 0 \le x < \infty \\ 0 \qquad\qquad , \text{yang lainnya} \end{cases}$$

dengan fungsi gamma $\Gamma(\alpha) = \int_0 x^{\alpha-1}e^{-x}dx.$

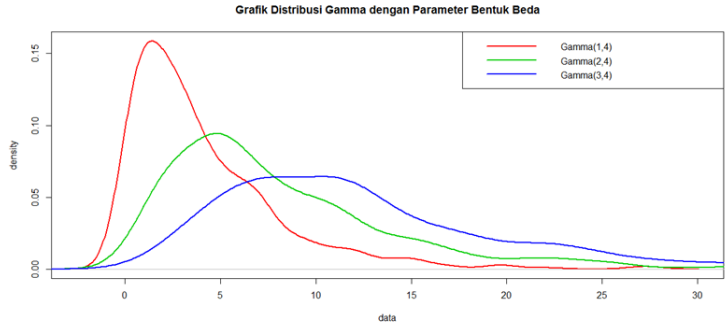
Dilakukan simulasi dengan membangkitkan data berdistribusi Gamma secara acak dengan software R dimana data dibangkitkan sebanyak 500 data dengan parameter bentuk berbeda dan parameter skala sama. Dibandingkan 3 buah distribusi Gamma, yaitu Gamma (1,4), Gamma

(2,4), dan Gamma (3,4). Dari ketiga distribusi tersebut dihitung mean dan standar deviasinya, diperoleh

Tabel 5.1. Perbandingan Distribusi Gamma dengan Parameter Bentuk Berbeda

Pembanding	**Gamma(1,4)**	**Gamma(2,4)**	**Gamma(3,4)**
Mean	4,102332	8,10976	12,32064
Standar Deviasi	3,983597	6,028978	7,208599

Berdasarkan hasil dari Tabel 5.1, diperoleh bahwa semakin besar nilai parameter bentuknya, maka nilai standar deviasinya akan semakin besar, begitu juga dengan nilai meannya. Selanjutnya, apabila dilihat dari hasil plot grafik ketiganya, diperoleh bahwa semakin besar nilai parameter bentuknya, maka kurva akan semakin bergeser ke kanan (Gambar 5.1)



Gambar 5.1. Grafik Distribusi Gamma dengan Parameter Bentuk Beda

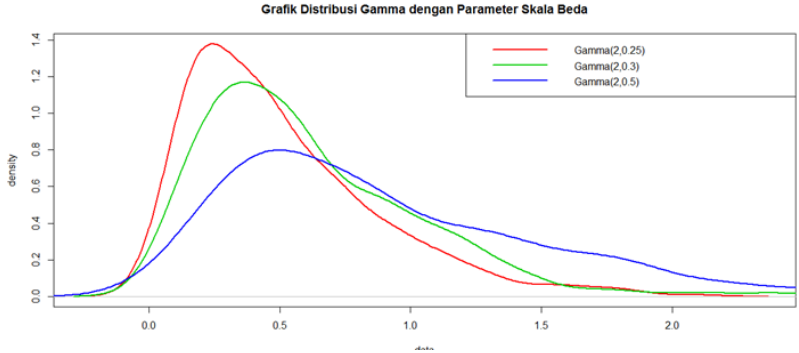
Selanjutnya, dilakukan simulasi dengan membangkitkan data berdistribusi Gamma secara acak dengan software R dimana data dibangkitkan sebanyak 500 data dengan parameter skala beda dan parameter bentuk sama. Dibandingkan 3 buah distribusi Gamma, yaitu Gamma (2;0,25), Gamma (2;0,3),

dan Gamma (2;0,5). Dari ketiga distribusi tersebut dihitung mean dan standar deviasinya, diperoleh

Tabel 5.2 Perbandingan Distribusi Gamma dengan Parameter Skala Berbeda

Pembanding	**Gamma(2;0,25)**	**Gamma(2;0,3)**	**Gamma(2;0,5)**
Mean	0,5099508	0,6244007	0,9406486
Standar Deviasi	0,3667272	0,447058	0,6812037

Berdasarkan hasil dari Tabel 5.2, diperoleh bahwa semakin besar nilai parameter skala, maka nilai standar deviasinya akan semakin besar, begitu juga dengan nilai meannya. Selanjutnya, apabila dilihat dari hasil plot grafik ketiganya, diperoleh bahwa semakin besar nilai parameter skalanya, maka kurva akan semakin melebar (Gambar 5.2)



Gambar 5.2. Grafik Distribusi Gamma dengan Parameter Skala Beda

5.4 Metode Bayesian

Metode Bayesian dalam statistika berbeda dengan metode Maximum Likelihood. Pada metode Maximum Likelihood parameter θ merupakan nilai tetap yang tidak diketahui. Sebaliknya, pada metode Bayesian, θ merupakan besaran yang variasinya digambarkan dengan

distribusi probabilitas, sehingga θ disebut distribusi prior. Distribusi prior diperoleh berdasarkan keyakinan seseorang dan dirumuskan sebelum data diambil. Kemudian, sampel diambil dari populasi dengan parameter θ dan distribusi prior disesuaikan dengan informasi sampel. Prior yang telah disesuaikan disebut distribusi posterior. Penyesuaian ini dilakukan dengan Teorema Bayes berikut: (Hogg, McKean, dan Craig, 2005)

$$P(A_k \mid B) = \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)} \quad (5.1)$$

Misalkan variabel random X memiliki distribusi probabilitas yang bergantung pada θ, $\theta \in \Omega$. Dengan demikian, variabel random X memiliki fungsi densitas probabilitas $f(x \mid \theta)$ dan $\theta \in \Omega$. Apabila distribusi θ pada Ω dinyatakan dengan $\pi(\theta)$, sehingga $\pi(\theta)$ merupakan distribusi prior dari θ. Dengan begitu, dapat dituliskan

$$X \mid \theta : \ f(x \mid \theta) \text{ dan } \theta : \ \pi(\theta).$$

Diberikan $X_1, X_2, ..., X_n$ sampel random dari distribusi bersyarat X diberikan θ dengan fungsi densitas probabilitas $f(x \mid \theta)$. Dengan demikian, fungsi densitas probabilitas gabungan dari $\mathbf{X} = (X_1, X_2, ..., X_n)$ diberikan θ yaitu

$$f(x \mid \theta) = f(x_1 \mid \theta) f(x_2 \mid \theta) ... f(x_n \mid \theta) \quad (5.2)$$

Persamaan (5.2) merupakan fungsi likelihood, sehingga

$$\begin{aligned} L(\mathbf{x} \mid \theta) &= f(\mathbf{x} \mid \theta) \\ &= f(x_1 \mid \theta) f(x_2 \mid \theta) ... f(x_n \mid \theta) \\ &= \prod_{i=1}^{n} f(x_i \mid \theta). \end{aligned}$$

Selanjutnya, fungsi densitas probabilitas gabungan $\mathbf{X}$ dan θ, yaitu

$$g(\mathbf{x},\theta) = L(\mathbf{x}\,|\,\theta)\pi(\theta).$$

Apabila θ adalah variabel random kontinu, maka fungsi densitas probabilitas marginal dari $\mathbf{X}$ yaitu

$$m(\mathbf{x}) = \int_{-\infty}^{\infty} g(\mathbf{x},\theta)d\theta$$

sedangkan bila θ adalah variabel random diskret, maka fungsi densitas probabilitas marginal dari $\mathbf{X}$ yaitu

$$m(\mathbf{x}) = \sum_{\theta} g(\mathbf{x},\theta).$$

Dengan begitu, fungsi densitas probabilitas bersyarat θ diberikan $\mathbf{X}$ yaitu

$$\pi(\theta\,|\,\mathbf{x}) = \frac{g(\mathbf{x},\theta)}{m(\mathbf{x})} = \frac{f(\mathbf{x}\,|\,\theta)\pi(\theta)}{m(\mathbf{x})}. \quad (5.3)$$

Distribusi yang didefinisikan oleh fungsi densitas probabilitas bersyarat θ diberikan $\mathbf{X}$ $\pi(\theta\,|\,\mathbf{x})$ merupakan distribusi posterior. Distribusi prior merefleksikan kepercayaan subyektif θ sebelum sampel diambil, sedangkan distribusi posterior adalah distribusi bersyarat θ setelah sampel diambil.

5.5 Proses Markov Chain Monte Carlo (Mcmc)

Inti dari inferensi Bayesian adalah analisis probabilitas. Jika *likelihood* didefinisikan sebagai probabilitas dari suatu data jika diberikan sebuah hipotesis (pohon), maka inferensi Bayesian ini menggunakan rasio *likelihood* tersebut untuk menentukan hipotesis (pohon) mana yang lebih mampu menjelaskan data sequence yang ada. Proses MCMC akan menghasilkan serangkaian hipotesis dalam rantainya (*chain*), masing-masing dengan nilai *likelihood*-nya. Nilai-nilai *likelihood* ini dimasukkan ke dalam Teorema Bayes

sebagaimana ditulis pada persamaan (5.1). Sebagai contoh, anggaplah kita telah memiliki dua nilai *likelihood* (L1 dan L2) untuk dua hipotesis dan ingin membandingkannya menggunakan Teorema Bayes. Dalam perbandingan ini, *state* yang baru diperoleh (*new state* / L2) dibandingkan dengan *state* awalan (*current state* / L1) untuk diperoleh nilai rasio r nya:

$$r = \frac{P(pohon \mid data)2}{P(pohon \mid data)1} = \frac{P(data \mid pohon)2 \times P(pohon)2}{P(data)2} \times \frac{P(data)1}{P(data \mid pohon)1 \times P(pohon)1}.$$

Jika diperhatikan masing-masing penyebut dan pembilang dari persamaan (P[data]) saling menghilangkan satu sama lain sehingga nilai r setara dengan perbandingan kedua nilai likelihhod (L2/L1) dari kedua hipotesis. Jika nilai r lebih besar dari 1 yang berarti state baru memiliki nilai likelihood yang lebih tinggi daripada current state, maka state baru tersebut akan dijadikan sebagai *current state* yang baru. Apabila nilai r lebih kecil dari 1 maka state yang baru akan diterima dengan probabilitas tertentu. Apabila masih tidak mungkin untuk diterima, maka state baru tersebut ditolak dan *current state* tetap menjadi current state untuk dibandingkan lagi dengan state baru lainnya.

Algoritma MCMC akan terus menyampling pohon hingga mencapai keadaan dimana tidak diperoleh lagi perbedaan nilai likelihood yang signifikan antar satu pohon dengan lainnya. Pada keadaan tersebut dapat dikatakan bahwa MCMC telah mencapai titik konvergensi. Pada titik konvergensi inilah pohon dengan distribusi posterior didapatkan.

Markov Chain mempunyai sifat konvergen menuju keadaan kesetimbangan terlepas dari titik awal. Kita hanya butuh mengatur rantai markov yang menyatu ke distribusi probabilitas posterior. Hal itu dapat dicapai dengan

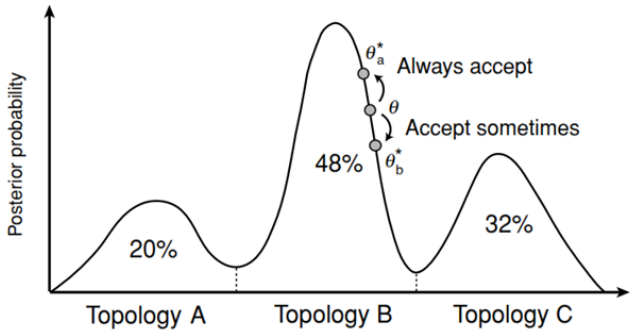
beberapa metode yang berbeda, metode yang paling flexibel dikenal sebagai metode metropolis-Hastings. (Lemey, Salemi, dan Vandamme, 2009).

Ide utamanya adalah membuat sedikit perubahan acak ke beberapa parameter nilai-nilai yang berlaku saat ini, kemudian menerima atau menolak perubahan tersebut sesuai dengan probabilitas yang sesuai. Prosedur dimulai pada rantai di titik sembarang θ (Gambar 5.1). Pada rantai generasi berikutnya, dipertimbangkan titik baru θ^* yang diambil dari sebuah distribusi $f(\theta^*|\theta)$. Kemudian rasio probabilitas posterior dihitung di dua titik. Kemungkinannya adalah apakah dihitung pada poin baru yang menanjak, di mana selalu dianggap sebagai titik awal untuk siklus berikutnya dalam rantai, atau pada posisi menurun, yang dianggap sebagai probabilitas yang sebanding dengan rasio tinggi.

Berikut adalah langkah-langkah Markov Chain Monte Carlo:

1. Mulailah pada titik sembarang (θ)
2. Buatlah sedikit gerakan acak (to θ')
3. Hitung rasio tertinggi (r) dari *state baru* (θ') ke *state* lama (θ)
   a. $r>1$: *state* baru diterima
   b. $r<1$: *state* baru diterima dengan probabilitas r
      jika *state baru ditolak*, tetaplah pada *state* lama
   c. Ulangi *step* 2

Sebagai ilustrasi dari algoritma di atas, berikut divisualisasikan dalam Gambar 5.3 berikut:



Gambar 5.3 Prosedur Markov chain Monte Carlo (MCMC)

Prosedur Markov Chain Monte Carlo (MCMC) digunakan untuk menghasilkan sampel yang valid dari posterior. Langkah pertama dengan membentuk rantai Markov yang memiliki posterior sebagai distribusi stasionernya. Rantai tersebut kemudian dimulai pada titik acak dan berjalan sampai menyatu dengandistribusi tersebut. Di setiap langkah (generasi) rantai, dilakukan perubahan kecil menuju nilai-nilai parameter model (sebagaimana pada langkah 2). Selanjutnya dihitung rasio r dari probabilitas posterior yang baru dan status saat ini. Jika r> 1, maka bergerak ke atas dan pergerakan tersebut selalu diterima (3a). Jika r <1, maka bergerak menurun dan menerima keadaan baru dengan probabilitas r (3b).

5.6 Pohon Filogenetik Epidemi SARS dengan Metode Bayesian

Pada inferensi Bayesian, parameter θ dianggap sebagai sebuah variabel random yang mengikuti sebuah distribusi tertentu. Distribusi ini disebut distribusi prior. Sebaliknya, distribusi posterior untuk variabel random θ memiliki

nilai yang sebanding dengan hasil kali antara distribusi prior setiap parameter dalam θ dan distribusi dari data (likelihood). Distribusi posterior digunakan untuk menentukan besarnya peluang penerimaan bagi kandidat yang terpilih sebagai anggota sampel random untuk θ.

Inferensi Bayesian pada filogenetik didasarkan pada perhitungan probabilitas posterior dari pohon filogenetik. Berdasarkan Teorema Bayes pada Persamaan (5.1), diperoleh perhitungan probabilitas posterior dari pohon filogenetik (θ) menurut (Huelsenback dan Ronquist, 2001) (Huelsenback, dkk, 2001) jika diketahui penyejajaran dari sekuen DNA $(\mathbf{X})$ yaitu

$$\pi(\theta \mid \mathbf{X}) = \frac{f(\mathbf{X} \mid \theta)\pi(\theta)}{m(\mathbf{X})} \qquad (5.4)$$

dengan $f(\mathbf{X} \mid \theta)$ merupakan probabilitas dari data penyejajaran sekuen DNA jika diketahui pohon filogenetik (*likelihood*), $\pi(\theta)$ merupakan probabilitas prior dari pohon filogenetik, dan $m(\mathbf{X})$ adalah probabilitas dari data penyejajaran dari sekuen DNA.

Untuk memodelkan penyejajaran digunakan proses stokastik. Menurut Shen dan Tuszynski (2007), mekanisme mutasi pada masing-masing tempat dalam suatu sekuen harus mengikuti aturan Poisson. Dengan demikian, data penyejajaran dari sekuen DNA mengikuti distribusi Poisson $(\mathbf{X} : POIS(\theta))$. Selanjutnya, Lemey, Salemi dan Vandamme (2009) mengatakan bahwa distribusi Gamma merupakan pilihan yang baik untuk digunakan sebagai distribusi prior $(\theta : Gamma(\alpha, \beta))$. Dalam hal ini,

$$f(\mathbf{X}\mid\theta)=\prod_{i=1}^{n} f(x_i\mid\theta)$$
$$=\prod_{i=1}^{n}\frac{\theta^{x_i}e^{-\theta}}{x_i!}$$
$$=\frac{\theta^{\sum_{i=1}^{n}x_i}e^{-n\theta}}{\prod_{i=1}^{n}x_i!}$$

dan

$$\pi(\theta)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\theta^{\alpha-1}e^{-\theta/\beta}.$$

Distribusi probabilitas gabungan dari $\mathbf{X}$ dan θ adalah

$$g(\mathbf{X},\theta)=f(\mathbf{X}\mid\theta)\pi(\theta)$$
$$=\left[\frac{\theta^{\sum_{i=1}^{n}x_i}e^{-n\theta}}{\prod_{i=1}^{n}x_i!}\right]\left[\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}}\right]. \quad (5.5)$$

Dari Persamaan (5.5), dapat diperoleh distribusi marginal dari $\mathbf{X}$ yaitu

$$m(\mathbf{X})=\int_0^{\infty} g(\mathbf{X},\theta)\,d\theta$$
$$=\int_0^{\infty}\frac{\theta^{\sum_{i=1}^{n}x_i}e^{-n\theta}\theta^{\alpha-1}e^{-\theta/\beta}}{\prod_{i=1}^{n}x_i!\,\Gamma(\alpha)\beta^{\alpha}}\,d\theta$$
$$=\frac{1}{\prod_{i=1}^{n}x_i!\,\Gamma(\alpha)\beta^{\alpha}}\int_0^{\infty}\theta^{\sum_{i=1}^{n}x_i}e^{-n\theta}\theta^{\alpha-1}e^{-\theta/\beta}\,d\theta$$
$$=\frac{1}{\prod_{i=1}^{n}x_i!\,\Gamma(\alpha)\beta^{\alpha}}\int_0^{\infty}\theta^{\sum_{i=1}^{n}x_i+\alpha-1}e^{-\theta(n+1/\beta)}\,d\theta$$
$$=\frac{\Gamma\left(\sum_{i=1}^{n}x_i+\alpha\right)\left((n+1/\beta)^{-1}\right)^{\sum_{i=1}^{n}x_i+\alpha}}{\prod_{i=1}^{n}x_i!\,\Gamma(\alpha)\beta^{\alpha}}\int_0^{\infty}\frac{\theta^{\sum_{i=1}^{n}x_i+\alpha-1}e^{-\theta(n+1/\beta)}}{\Gamma\left(\sum_{i=1}^{n}x_i+\alpha\right)\left((n+1/\beta)^{-1}\right)^{\sum_{i=1}^{n}x_i+\alpha}}\,d\theta$$
$$=\frac{\Gamma\left(\sum_{i=1}^{n}x_i+\alpha\right)\left((n+1/\beta)^{-1}\right)^{\sum_{i=1}^{n}x_i+\alpha}}{\prod_{i=1}^{n}x_i!\,\Gamma(\alpha)\beta^{\alpha}}$$
$$=\frac{\Gamma\left(\sum_{i=1}^{n}x_i+\alpha\right)}{\prod_{i=1}^{n}x_i!\,\Gamma(\alpha)\beta^{\alpha}\left(n+1/\beta\right)^{\sum_{i=1}^{n}x_i+\alpha}}.$$

Akibatnya, dengan menggunakan Persamaan (5.4) diperoleh distribusi posterior $\pi(\theta \mid \mathbf{X})$ berikut

$$\begin{aligned}
\pi(\theta \mid \mathrm{X}) &= \frac{\left[\dfrac{\theta^{\sum_{i=1}^{n} x_i} e^{-n\theta}}{\prod_{i=1}^{n} x_i!}\right]\left[\dfrac{1}{\Gamma(\alpha)\beta^{\alpha}}\theta^{\alpha-1}e^{-\theta/\beta}\right]}{\dfrac{\Gamma\left(\sum_{i=1}^{n} x_i+\alpha\right)}{\prod_{i=1}^{n} x_i!\,\Gamma(\alpha)\beta^{\alpha}\left(n+1/\beta\right)^{\sum_{i=1}^{n} x_i+\alpha}}} \\
&= \frac{\theta^{\sum_{i=1}^{n} x_i+\alpha-1} e^{-\theta(n+1/\beta)}}{\dfrac{\Gamma\left(\sum_{i=1}^{n} x_i+\alpha\right)}{\left(n+1/\beta\right)^{\sum_{i=1}^{n} x_i+\alpha}}} \\
&= \frac{\theta^{\sum_{i=1}^{n} x_i+\alpha-1} e^{-\theta\left((n+1/\beta)^{-1}\right)^{-1}}}{\Gamma\left(\sum_{i=1}^{n} x_i+\alpha\right)\left(\left(n+1/\beta\right)^{-1}\right)^{\sum_{i=1}^{n} x_i+\alpha}} \\
&= \frac{\theta^{\sum_{i=1}^{n} x_i+\alpha-1} e^{-\theta/(n+1/\beta)^{-1}}}{\Gamma\left(\sum_{i=1}^{n} x_i+\alpha\right)\left(\left(n+1/\beta\right)^{-1}\right)^{\sum_{i=1}^{n} x_i+\alpha}} \\
&= \frac{\theta^{\sum_{i=1}^{n} x_i+\alpha-1} e^{-\frac{\theta}{(\beta/(n\beta+1))}}}{\Gamma\left(\sum_{i=1}^{n} x_i+\alpha\right)\left(\beta/(n\beta+1)\right)^{\sum_{i=1}^{n} x_i+\alpha}}.
\end{aligned} \tag{5.6}$$

Perhatikan bahwa distribusi posterior pada Persamaan (5.6) merupakan distribusi Gamma dengan parameter $\alpha^* = \sum_{i=1}^{n} x_i + \alpha$ dan $\beta^* = \dfrac{\beta}{n\beta+1}$.

Untuk mendapatkan pohon filogenetik yang sesuai dengan posteriornya diperlukan integrasi probabilitas dari data penyejajaran sekuen DNA. Namun dalam

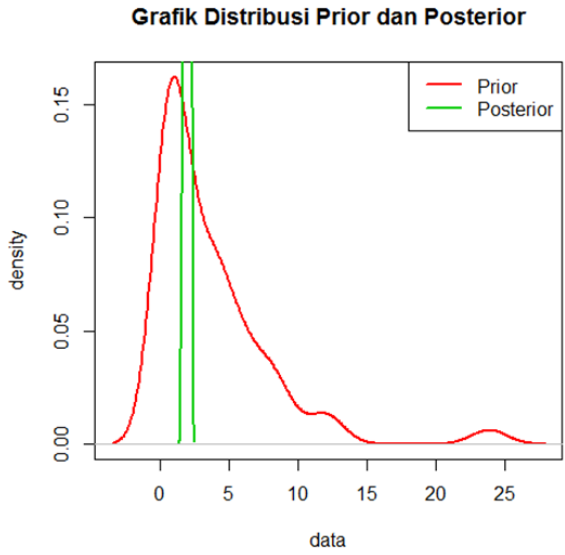
pengerjaannya untuk menentukan probabilitas dari data penyejajaran sekuen DNA sangatlah rumit, sehingga perlu dilakukan simulasi menggunakan proses MCMC.

Dilakukan simulasi dengan membangkitkan data random sebanyak $n=100$. Pertama, membangkitkan data berdistribusi Poisson dengan parameter $\theta=2$, diperoleh nilai $\sum_{i=1}^{100} x_i = 201$. Kedua, membangkitkan data distribusi prior dengan menggunakan distribusi $Gamma(\alpha,\beta)$ dengan $\alpha=1$ dan $\beta=4$. Kemudian, membangkitkan data posterior dengan distribusi $Gamma(\alpha^*,\beta^*)$ dengan $\alpha^* = \sum_{i=1}^{100} x_i + \alpha = 202$ dan $\beta^* = \frac{\beta}{n\beta+1} = 0,00998$. Akibatnya, diperoleh

Tabel 8.1 Perbandingan Distribusi Prior dan Posterior

Pembanding	**Prior**	**Posterior**
Mean	3,697739	2,003964
Standar Deviasi	4,261449	0,1487471

Berdasarkan hasil dari Tabel 8.1, diperoleh bahwa nilai standar deviasi untuk posterior lebih kecil dibandingkan prior. Selanjutnya, apabila dilihat dari hasil plot grafik Gambar 8.1, diperoleh bahwa grafik posterior lebih menyempit dibandingkan prior. Artinya, distribusi posterior yang dihasilkan sudah sesuai.



Gambar 8.1 Grafik Distribusi Prior dan Posterior

Simulasi MCMC digunakan dalam inferensi Bayesian. Simulasi ini digunakan apabila bentuk dari distribusi posterior tidak standar dan rumit. Berdasarkan bentuk hasil distribusi posterior yang diperoleh pada Persamaan (8.3), dilakukan simulasi MCMC dengan menggunakan *software* MRBAYES. Pada simulasi ini digunakan dua distribusi prior yang berbeda yaitu distribusi Invgamma (Bayes 1) dan distribusi Gamma (Bayes 2). Berikut hasil perbandingan dari kedua simulasinya.

Tabel 8.2. Perbandingan Hasil Bayesian

Pembanding	**Hasil Bayes 1**	**Hasil Bayes 2**
Rates	Invgamma	Gamma
Likelihood run 1 (L1)	-41867.03	-41869.6
Likelihood run 2 (L2)	-41867.05	-41869.34
Waktu analisis (detik)	6.67	6.94
Rata-rata standar deviasi	0.019806	0.012059
Maksimum standar deviasi	0.046136	0.033896
Rasio (L2/L1)	1.00000048	0.99999379
Rata-rata PSRF	1.009	1.007

Berdasarkan perbandingan pada Tabel 8.2 terlihat bahwa nilai rata-rata standar deviasi mendekati nol dan nilai rata-rata PSRF (*Potential Scale Reduction Factor*) mendekati satu, artinya proses MCMC telah mencapai titik konvergen. Pada saat konvergen inilah pohon filogenetik dengan distribusi posterior didapatkan. Dari kedua model yang diperoleh, nilai rata-rata standar deviasi terkecil yaitu pada hasil Bayes 2 (prior:Gamma). Hal ini menunjukan bahwa semakin kecil standar deviasinya maka akan semakin mendekati sebenarnya. Pada nilai log likelihoodnya yang paling maksimum pada hasil Bayes 2. Untuk nilai rasio perbandingan antara *Likelihood* 1 dan *Likelihood* 2 yang memiliki rasio kurang dari satu yaitu hasil Bayes 2. Ini berarti bahwa model yang menggunakan distribusi prior Gamma dapat menunjukkan pohon filogenetik terbaik berdasarkan distribusi probabilitas posterior tertinggi diantara pohon filogenetik yang lainnya.

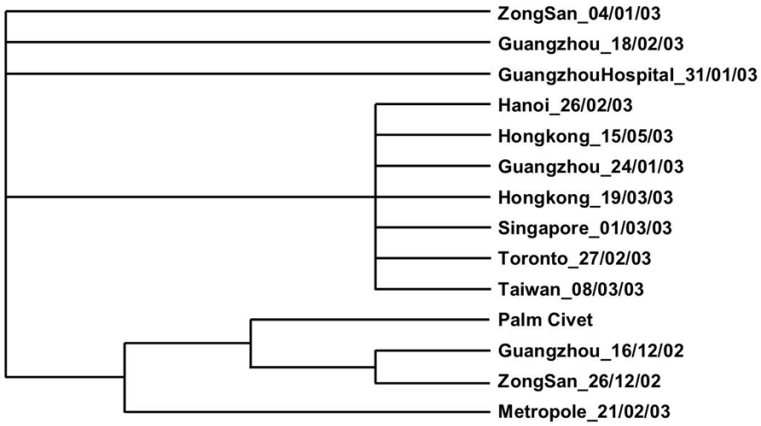
Untuk menentukan nilai minimum dan maksimum dari probabilitas posterior yang diperoleh digunakan interval kepercayaan 95% dengan rumus

$$\left(x - 1{,}96\frac{s}{\sqrt{n}}, x + 1{,}96\frac{s}{\sqrt{n}} \right), x = \text{probabilitas}.$$

Dengan begitu, misalkan $x = 0,992676$, $s = 0,010357$ dan $n = 2$, maka nilai minimum dan maksimumnya yaitu

$$\left(0,992676 - 1,96\frac{0,010357}{\sqrt{2}}; 0,992676 + 1,96\frac{0,010357}{\sqrt{2}}\right) = (0,985353; 1,007).$$

Adapun bentuk pohon filogenetik hasil inferensi Bayes, yaitu



Gambar 8.2 Pohon filogenetik Bayesian

Dapat disimpulkan bahwa distribusi prior Gamma cocok digunakan untuk mengkonstruksi pohon filogenetik pada metode Bayesian. Hal ini terlihat dari hasil simulasi yang dilakukan bahwa jika digunakan distribusi prior Gamma akan diperoleh nilai standar deviasi dan nilai rasio likelihood yang minimum. Selanjutnya, distribusi posterior yang diperoleh merupakan distribusi Gamma dengan parameter $\alpha^* = \sum_{i=1}^{n} x_i + \alpha$ dan $\beta^* = \frac{\beta}{n\beta + 1}$.

5.7 Soal Latihan

Konstruksilah pohon filogenetik dengan metode maximum likelihood dan metode Bayes. Gunakan paket software Phylip, PAML, dan MR.BAYES. Bandingkan hasilnya dan analisa. Gunakan 8 sekuen data SARS dari 8 negara yang berbeda.

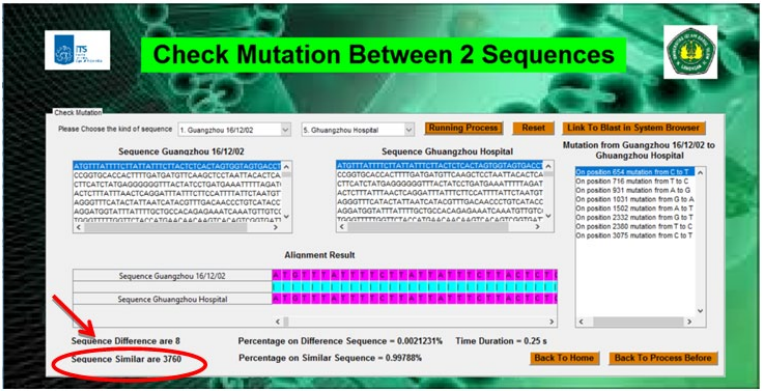
BAB VI

HASIL ANALISIS PENYEJAJARAN GANDA

Dari hasil analisis penyejajaran ganda pada 14 sekuen DNA pasien terinfeksi virus SARS yang telah dilakukan, diperoleh analisis sistem jaringan topologi, sistem jaringan daerah mutasi, dan sistem jaringan mode mutasi yang secara rinci dijelaskan sebagai berikut:

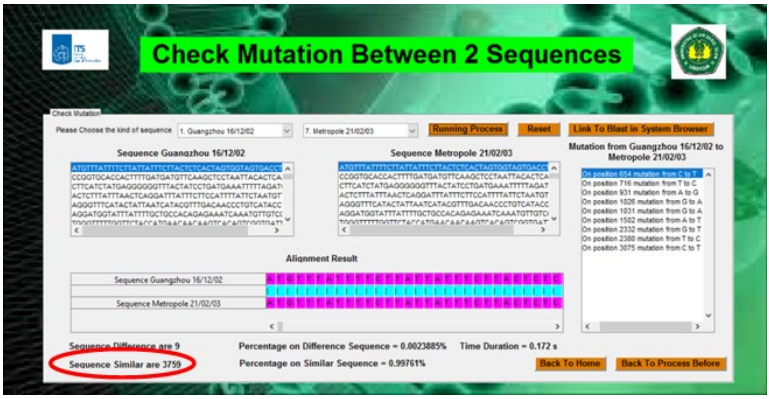
6.1 Analisis Sistem Jaringan Topologi

Sistem jaringan topologi yang dihasilkan oleh output penyejajaran ganda, yaitu $G(W) = \{M,V,W\}$ dimana W adalah fungsi penalty dari output penyejajaran ganda. Dan matriks penalti diperoleh dari penyejajaran antar 2 sekuen dengan menerapkan algoritma penyejajaran *Needleman Wunch* yang disimulasikan dalam Matlab. Dengan simulasi tersebut, diperoleh hasil perbedaan sekuen pada masing-masing penyejajaran sekuen, prosentase jumlah sekuen yang berbeda, banyaknya mutasi yang terjadi serta durasi waktu yang dibutuhkan untuk simulasi. Semuanya tampak sebagaimana tampilan pada menu *user interface* berikut:

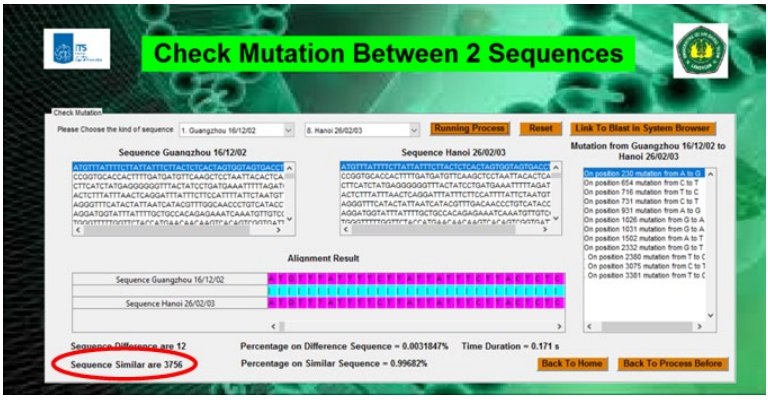


Gambar 6.1 Output penyejajaran sekuen antara sekuen GuangZhou dengan sekuen Guangzhou hospital

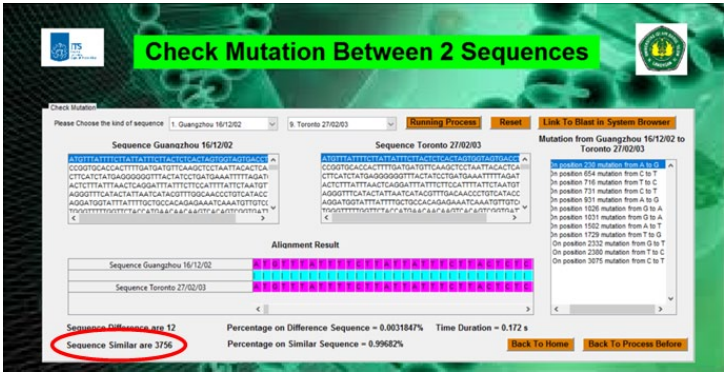
Pada bagian yang dilingkari merah menunjukkan perbedaan sekuen yang terjadi antara sekuen Guangzhou dengan Guangzhou hospital sebanyak 8 nukleotid dari 3768 panjang sekuen. Atau dalam prosentasenya sebesar 0,0021231%. Hasil tersebut nantinya dituliskan dalam matriks penalti pada baris 1 kolom 2 sebagaimana penjelasan berikutnya.



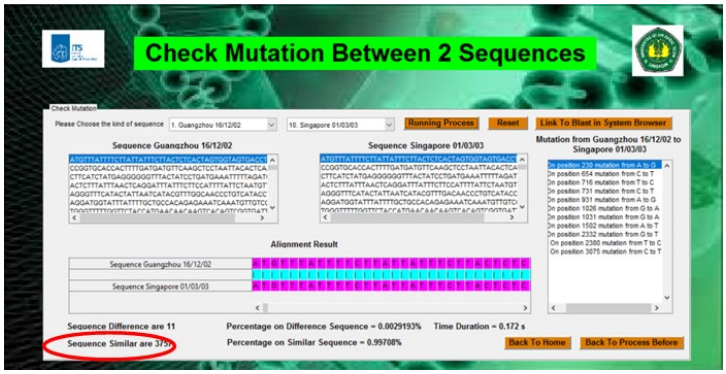
Gambar 6.2 Output penyejajaran sekuen antara sekuen GuangZhou dengan sekuen Metropole



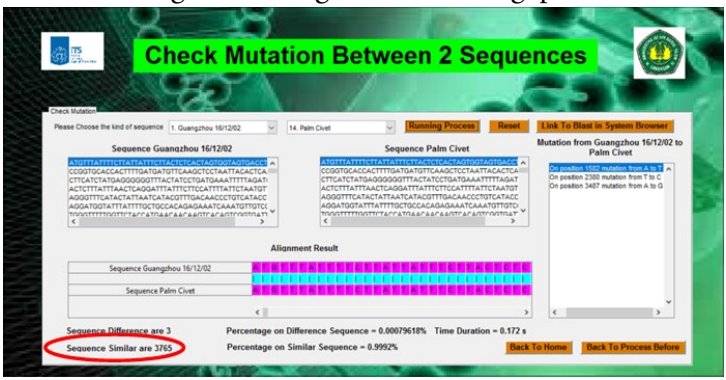
Gambar 6.3 Output penyejajaran sekuen antara sekuen GuangZhou dengan sekuen Hanoi



Gambar 6.4 Output penyejajaran sekuen antara sekuen GuangZhou dengan sekuen Toronto



Gambar 6.5 Output penyejajaran sekuen antara sekuen GuangZhou dengan sekuen Singapore



Gambar 6.6 Output penyejajaran sekuen antara sekuen GuangZhou dengan sekuen Palm Civet

Dari hasil masing-masing penyejajaran di atas dan seterusnya yang tidak semuanya ditampilkan pada laporan ini, akhirnya diperoleh matriks penalty sebagai berikut:

$$\bar{W}(\bar{C}) = \begin{array}{c} \begin{matrix} A & B & C & D & E & F & G & H & I & J & K & L & M & N \end{matrix} \\ \begin{bmatrix} 0 & 4 & 9 & 12 & 8 & 9 & 9 & 12 & 12 & 11 & 11 & 11 & 13 & 3 \\ 4 & 0 & 7 & 10 & 6 & 7 & 7 & 10 & 10 & 9 & 9 & 9 & 11 & 3 \\ 9 & 7 & 0 & 3 & 1 & 2 & 0 & 3 & 3 & 2 & 2 & 2 & 4 & 8 \\ 12 & 10 & 3 & 0 & 4 & 5 & 3 & 2 & 2 & 1 & 1 & 1 & 3 & 11 \\ 8 & 6 & 1 & 4 & 0 & 3 & 1 & 4 & 4 & 3 & 3 & 3 & 5 & 7 \\ 9 & 7 & 2 & 5 & 3 & 0 & 2 & 5 & 5 & 4 & 4 & 4 & 6 & 8 \\ 9 & 7 & 0 & 3 & 1 & 2 & 0 & 3 & 3 & 2 & 2 & 2 & 4 & 8 \\ 12 & 10 & 3 & 2 & 4 & 5 & 3 & 0 & 2 & 1 & 1 & 1 & 3 & 11 \\ 12 & 10 & 3 & 2 & 4 & 5 & 3 & 2 & 0 & 1 & 1 & 1 & 3 & 11 \\ 11 & 9 & 2 & 1 & 3 & 4 & 2 & 1 & 1 & 0 & 0 & 0 & 2 & 10 \\ 11 & 9 & 2 & 1 & 3 & 4 & 2 & 1 & 1 & 0 & 0 & 0 & 2 & 10 \\ 11 & 9 & 2 & 1 & 3 & 4 & 2 & 1 & 1 & 0 & 0 & 0 & 2 & 10 \\ 13 & 11 & 4 & 3 & 5 & 6 & 4 & 3 & 3 & 2 & 2 & 2 & 0 & 12 \\ 3 & 3 & 8 & 11 & 7 & 8 & 8 & 11 & 11 & 10 & 10 & 10 & 12 & 0 \end{bmatrix} \end{array}$$

Di mana A mewakili sekuen Guangzhou 16/12/02, B mewakili sekuen Zhongsan 26/12/02, C mewakili sekuen Zhongsan 04/01/03, D mewakili sekuen Guangzhou 24/01/03, E mewakili sekuen Guangzhou Hospital, F mewakili sekuen Guangzhou 18/02/03, G mewakili sekuen Metropole 21/02/03, H mewakili sekuen Hanoi 26/02/03, I mewakili sekuen Toronto 27/02/03, J mewakili sekuen Singapore 01/03/03, K mewakili sekuen Taiwan 08/03/03, L mewakili sekuen Hong Kong 19/03/03, M mewakili sekuen HongKong 15/05/03, dan N mewakili sekuen Palm Civet. Palm Civet adalah seekor musang yang disinyalir sebagai host dari epidemic SARS (Guan et all, 2003), sekuen palm civet dimunculkan dengan tujuan agar diketahui sekuen DNA dari pasien mana yang lebih dekat dengan host tersebut. Dan sekuen yang paling dekat dengan host berarti sebagai awal mula tempat penyebaran epidemik SARS. Dalam analisis sistem jaringan topologi

ini juga diperoleh daerah stabil yang menunjukkan posisi nukleotid yang sama pada penyejajaran ganda dan daerah tidak stabil yang menunjukkan posisi nukleotid yang berbeda. Pada daerah tidak stabil inilah diketahui mutasi antar sekuen tersebut berada. Berikut adalah tabel daerah stabil dan tidak stabil pada penyejajaran ganda epidemi SARS:

Tabel 6.1 Daerah stabil dan tidak stabil pada penyejajaran ganda sekuen SARS

	Posisi nucleotide ke-	**Jumlah**	**Prosentase**
Daerah stabil	1 – 80, 82 – 145, ...3624 – 3768	3749	99,5 %
Daerah tidak stabil	81, 146, 224, 230, 654, 716, 731, 931, 1026, 1031, 1502, 1729, 2332, 2380, 3075, 3381, 3487, 3505, 3623	19	0,5 %
	Panjang sekuen	3768	100 %

Adapun posisi masing-masing nukleotid pada daerah tidak stabil ditunjukkan dalam Tabel 6.1. Pada tabel tersebut, terlihat bahwa hanya pada 19 posisi saja dari 3768bp sekuen yang merupakan daerah tidak stabil, terjadi perubahan nukleotid penyebab mutasi. Pada masing-masing posisi, jumlah nukleotid yang berubah tidaklah sama. Contoh pada posisi 81, jumlah nukleotid T (Timin) sebanyak 13, nukleotid G (Guanin) ada 1. Namun jumlah berbeda terdapat pada posisi daerah tidak stabil yang berbeda pula. Sehingga apabila diringkas dalam satu tabel, akan dapat diketahui secara kseluruhan untuk banyaknya nucleotide pada daerah tidak stabil dan posisinya sebagaimana tabel berikut:

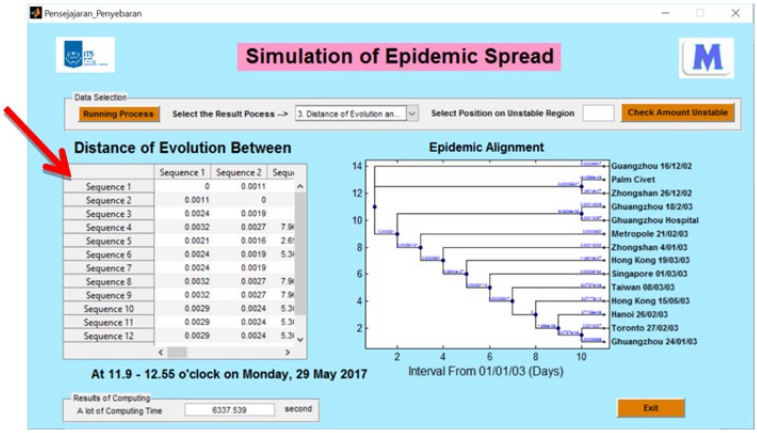
Tabel 6.2 Banyaknya nukleotid pada daerah tidak stabil dan posisinya

No	Posisi nukleotid	Jumlah nukleotid				
		A	C	G	T	-
1	81	0	0	1	13	0
2	146	0	13	0	1	0
3	224	0	13	1	0	0
4	230	7	0	7	0	0
5	654	0	2	0	12	0
6	716	0	11	0	3	0
7	731	0	7	0	7	0
8	931	3	0	11	0	0
9	1026	10	0	4	0	0
10	1031	11	0	3	0	0
11	1502	1	0	0	13	0
12	1729	0	0	1	13	0
13	2332	0	0	4	10	0
14	2380	0	13	0	1	0
15	3075	0	3	0	11	0
16	3381	0	1	0	13	0
17	3487	13	0	1	0	0
18	3505	13	0	1	0	0
19	3623	0	13	0	1	0

Dari Tabel 6.2 tersebut dapat diartikan bahwa pada posisi daerah tidak stabil tersebut adalah posisi terjadinya mutasi antar sekuen DNA pasien terinfeksi SARS. Perbedaannya terlihat dari jumlah nucleotide yang tidak sama pada masing-masing posisi nukleotid pada 14 sekuen yang disejajarkan.

6.2 Analisis Sistem Jaringan Daerah Mutasi

Analisis yang kedua adalah sistem jaringan daerah mutasi pada multiple alignment epidemi SARS. Pada bagian ini secara garis besar adalah bagaimana membangun graf dan pohon yang dihasilkan oleh epidemi SARS. Graf tersebut berupa pohon filogenetik, yaitu pohon kekerabatan yang mencerminkan hubungan evolusi antar sekuen. Pada kasus ini pohon filogenetik menggambarkan hubungan evolusi yang menunjukkan penyebaran epidemi dari sekuen satu ke sekuen lain yang mana masing-masing sekuen diambil dari daerah yang berbeda. Sehingga output dari pohon filogenetik nantinya menunjukkan proses penyebaran epidemi SARS antar daerah/negara. Pada bab ini, input dalam pembentukan pohon filogenetik berupa matriks jarak. Matriks jarak diperoleh dari matriks penalty yang memenuhi fungsi jarak sebagaimana definisi sebelumnya. Berikut adalah matriks jarak yang dihasilkan dari matriks penalti denganmasing-masing elemen dibagi dengan panjang penyejajarannya dan telah dikoreksi dengan model evolutioner Jukes Cantor.



Gambar 6.7 Tampilan GUI dari jarak evolusi

Pada *output* menu GUI di Gambar 6.7 tersebut jarak evolusi antara sekuen 1 ke sekuen berikutnya tidak bisa ditampilkan utuh karena keterbatasan *space* pada menu yang tersedia. Hasil jarak antar sekuen yang tidak terlihat bisa dibuka dengan cara menggeser kursor ke arah kanan. Hasil tersebut merupakan inputan yang digunakan untuk proses pembentukan pohon filogenetik dengan metode jarak.

Beralih menuju proses pembentukan pohon filogenetik, karena data yang digunakan dalam penelitian ini sebanyak 14 sekuen, maka proses pembentukan pohon filogenetik berlangsung sebanyak 13 *cycle*, dimulai dari *cycle* 1:

a. Input: matriks jarak evolutioner

Tabel 6.4 Output matriks jarak evolutioner hasil simulasi Matlab

	A	B	C	D	E	F	G
A	0	0.0021	0.0024	0.0032	0.0032	0.0029	0.0008
B	0.0021	0	0.0003	0.0011	0.0011	0.0008	0.0019
C	0.0024	0.0003	0	0.0008	0.0008	0.0005	0.0021
D	0.0032	0.0011	0.0008	0	0.0005	0.0003	0.0029
E	0.0032	0.0011	0.0008	0.0005	0	0.0003	0.0029
F	0.0029	0.0008	0.0005	0.0003	0.0003	0	0.0027
G	0.0008	0.0019	0.0021	0.0029	0.0029	0.0027	0

Dari matriks jarak evolutioner tersebut proses pembentukan pohon filogenetik dimulai:

b. Step 1

Hitung S_i mengikuti rumus:

$$S_i = \frac{1}{N-2}\sum_{k=1}^{N} D_{ik}$$

Di mana N adalah banyaknya sekuen, D_{ik}adalah jarak dari i ke k pada matrix evolusinya. Hasil dari perhitungan masing-masing S_i adalah:

S_A = 0.0029; S_B = 0.0015; S_C = 0.0014; S_D = 0.0018; S_E = 0.0018; S_F = 0.0015; S_G = 0.0023

c. Step 2: Dicari nilai minimum untuk masing-masing pasangan sekuen:

$$M_{ij} = D_{ij} - S_i - S_j$$

Apabila dituliskan secara lengkap, diperoleh matrix berikut:

	A	B	C	D	E	F	G
A	0	-0.0022	-0.0019	-0.0015	-0.0015	-0.0015	-0.0048
B	-0.0022	0	-0.0026	-0.0021	-0.0021	-0.0021	-0.0022
C	-0.0019	-0.0026	0	-0.0023	-0.0023	-0.0023	-0.0019
D	-0.0015	-0.0021	-0.0023	0	-0.0030	-0.0030	-0.0015
E	-0.0015	-0.0021	-0.0023	-0.0030	0	-0.0030	-0.0015
F	-0.0015	-0.0021	-0.0023	-0.0030	-0.0030	0	-0.0015
G	-0.0048	-0.0022	-0.0019	-0.0015	-0.0015	-0.0015	0

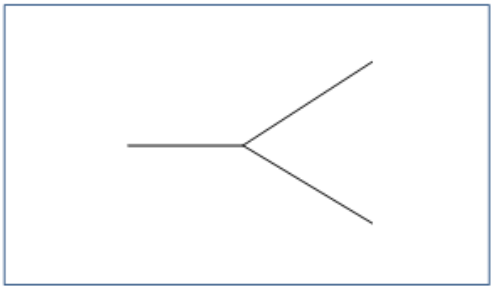
Diperoleh pasangan M terkecil adalah = -0.0048

d. Step 3: Definisikan sekuen baru yaitu U_i yang menggantikan pasangan terkecil (A dan G). Selanjutnya taxa tersebut gabungkan sebagai U_i mengikuti rumus:

$$S_{AU_1} = 0.5(D_{AG} + S_A - S_G) = 0.0005$$

$$S_{GU_1} = 0.5(D_{AG} + S_G - S_A) = 0.0003$$

e. Step 4: Hubungkan taxa U_1 dengan A dan U_1 dengan G masing-masing dengan mengikuti panjang edge yang mewakili jarak sebagaimana hasil pada perhitungan pada step 3.



Gambar 6.8 Tree pada cycle 1

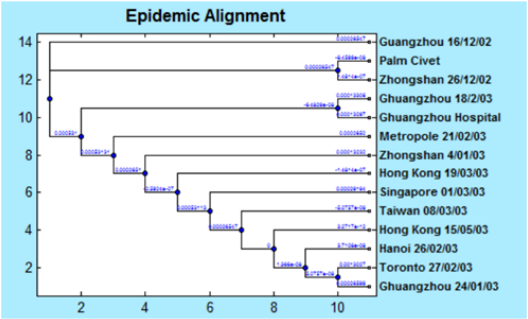
f. Step 5:
Gabungkan jarak baru dari semua taxa ke U_1, diperoleh:

$$D_{BU_1} = 0.5(D_{AB} + D_{GB} - D_{AG}) = 0.5\,(0.0021 + 0.0019 - 0.0008) = 0.0016$$

$$D_{CU_1} = 0.5(D_{AC} + D_{GC} - D_{AG}) = 0.5\,(0.0024 + 0.0021 - 0.0008) = 0.0019$$

$$D_{DU_1} = 0.0027; \qquad D_{EU_1} = 0.0027; \qquad D_{FU_1} = 0.0024$$

Hasil dari jarak baru U_1 ke semua taxa untuk selanjutnya dimasukkan dalam matrix evolutioner yang baru. Demikian berlanjut cycle 2 sampai cycle terakhir (cycle 6), sehingga diperoleh pohon filogenetik sebagai berikut:

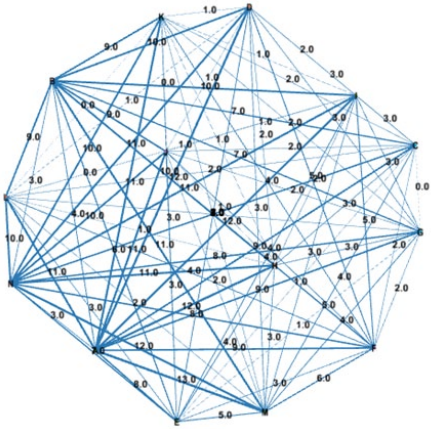


Gambar 6.9 Pohon filogenetik penyebaran epdemi SARS

Dari Gambar 6.9 tersebut, tampak bahwa yang terdekat dengan Palm Civet sebagai host (Irawan dan Amiroch, 2015) adalah Zhongsan 26/12/02. Namun apabila memperhatikan jarak genetiknya juga tidak lebih dari Guangzhou 16/12/02. Sehingga disimpulkan bahwa penyebaran epidemi SARS dari guangzhou 16/12/02, terus menyebar ke Zhongsan 26/12/02, kemudian hampir secara bersamaan ke guangzhou 18/2/03 dan Guangzhou hospital. Dari sana virus terus menyebar ke Metropole, Zhongsan, Hongkong, Singapore, Taiwan, Hongkong, Hanoi, GuangZhou 24/01/03 dan Toronto secara bersamaan.

6.3 Analisis Sistem Jaringan Mode Mutasi

Sebelum melakukan analisis jaringan mode mutasi, dari matriks penalty divisualisasi graf tidak berarah yang menunjukkan hubungan mutasi antar sekuen. Notasi pada simpul menunjukkan nama sekuen yang dikodekan sebagai huruf A, B, ...N dengan masing-masing kode mewakili nama sekuen sebagaimana tercantum pada pembahasan sebelumnya.



Gambar 6.10 Dekomposisi jaringan mutasi penyebaran epidemi SARS

Pada Gambar 6.10, label pada sisi menunjukkan banyaknya mutasi yang terjadi. Semakin tebal garis yang merepresentasikannya, menunjukkan semakin banyak mutasi yang terjadi antar simpul tersebut. Sebagaimana diketahui dari pembahasan sebelumnya, terdapat 19 mutasi pada daerah tidak stabil yang berbeda terhadap 14 sekuen DNA epidemik SARS ini. Terlihat dari gambar, beberapa mutasi orthogonal hanya terjadi pada busur order ke-1 saja, misalnya dalam Mode mutasi (mutasi pada sekuen Guangzhou 16/12/02 ke sekuen Toronto 27/03/03), mode mutasi (mutasi pada sekuen Guangzhou 16/12/02 ke sekuen Guangzhou Hospital), serta mode mutasi (mutasi pada sekuen Guangzhou Hospital ke sekuen Toronto 27/03/03). Pada berlaku: |AE| = |AB| + |BE| dan struktur modulusnya saling orthogonal. Representasi mutasi antara 2 sekuen dari semua mode mutasi dapat dilihat di Lampiran 3.

DAFTAR PUSTAKA

Andriani, T., dan Irawan, M. I., 2017, Application of Unweighted Pair Group Methods with Arithmetic Average (UPGMA) for Identification of Kinship Types and Spreading of Ebola Virus Through Establishment of Phylogenetic Tree, *AIP Conf. Proc.*, vol. 1867.

Amiroch, S., Rohmatullah, A., 2017, Determining Geographical Spread Pattern of Mers-CoV by distance methods using Kimura Model*, AIP Conference Proceedings* 1825 (1), 020001.

Amiroch, S., Pradana, M. S., Irawan, M. I., dan Mukhlash, I., 2017, Multiple Alignment Analysis on Phylogenetic Tree of The Spread of SARS Epidemic Using Distance Method, *Journal of Physics: Conference Series*, vol. 890, no. 1.

Amiroch, S., Pradana, M. S., Irawan, M. I., dan Mukhlash, I., 2018, Maximum Likelihood Method on The Construction of Phylogenetic Tree for Identification the Spreading SARS Epidemic, *Proceeding International Symposium on Advanced Intelligent Informatics* (SAIN), IEEE, No 1, 137-141.

Arkeman, Y., dkk, 2012. *Algoritma Genetika; Teori dan Aplikasinya untuk Bisnis dan Industri,* IPB Press, Bogor.

Christianini, N., Hahn, M.W, 2006, *Introduction to Computational Genomics A Case Studies Approach,* Cambridge University Press, New York.

Felsenstein, J., 1983, Statistical Inference of Phylogenesis, *Journal of the Royal Statitical Society Series A (General)* Vol. 146 No. 3 pg 246-272.

Hogg, R.V., McKean, J. W., dan Craig, A. T., 2005, *Introduction to Mathematical Statistics Sixth Edition*. Pearson Prentice Hall.

Huelsenbeck, J. P., dan Ronquist, F., 2001, MRBAYES : Bayesian Inference of Phylogenetic Trees, *Bioinforma. Appl. Note*, vol. 17, no. 8, pp. 754–755.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., dan Bollback, J. P., 2001, Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology, vol. 294, no. December, pp. 2310–2314.

Irawan, I., **Amiroch, S**, 2015, Construction Of Phylogenetic Tree Using Neighbor Joining Algorithms To Identify The Host And The Spreading Of SARS Epidemic, *Journal of Theoretical and Applied Information Technology* Vol.71 No.3 pg. 424-429.

Isaev, A., 2006, *Introduction to Mathematical Methods in Bioinformatics*, Springer.

Lemey P, Salemi M, Vandamme M, 2009, *The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press.

Marra, et al., 2003, The Genome Sequence of the SARS-Associated *Coronavirus.* Science 300, 1399 (2003), *The American Association for the Advancement of Science*, Washington, (www.sciencemag.org), diakses 29 Januari 2014.

Nascimento, F.F., Reis, M., dan Yang, Z., 2017, A Biologist's Guide to Bayesian Phylogenetic Analysis, *Nat. Ecol. Evol.*, vol. 1.

Page and Holmes, 1998. *Molecular Evolution*: *A Phylogenetic Approach,* Blackwell Publishing Ltd.

Praptono, 1986, *Pengantar Proses Stokastik I,* Karunika Jakarta.

Ross, S.M, 2010, *Introduction to Probability Model 10th Ed.,* Academic Press, USA.

Rota, et al., 2003, Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome.

Science 300, 1394 (2003), *The American Association for the Advancement of Science*, Washington, (www.sciencemag.org), diakses 29 Januari 2014.

Severe Acute Respiratory Syndrome (SARS), Pubmed Health, (http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0004460), diakses 29 Januari 2014.

Shen, S.N., dan Tuszynski, J. A., 2007, *Theory and Mathematical Methods for Bioinformatics, Biological and Medical Physics, Biomedical Engineering*, Springer.

Sivanandam & Deepa, 2008, *Introduction to Genetic Algorithm*, Springer.

Thorne, et al (1991), An Evolutionary Model For Maximum Likelihood Alignment of DNA Sequences, *Journal of Molecular Evolution* vol 33 pg 114-124.

Wackerly, D. D., Mendenhall, W., dan Scheaffer, R. L., 2008, *Mathematical Statistics with Applications*, 7th ed. Thomson.

Viniotis, Y, 1998, *Probability and Random Process,* Mc. Graw-Hill, Singapore.

Tentang Penulis

Siti Amiroch (siti.amiroch@unisda.ac.id) lahir di Lamongan, 11 Februari 1973. Setelah menamatkan S-1 di jurusan Matematika ITS pada tahun 1997, mulai mengajar di Program Studi Matematika Universitas Islam Darul 'Ulum Lamongan (UNISDA). Setelah lulus S-2 di jurusan dan kampus yang sama, ia menjadi Dekan di Fakultas Matematika dan Ilmu Pengetahuan Alam UNISDA. Ketertarikannya pada bidang Bioinformatika diawali dari tesis yang berjudul *Konstruksi Pohon Filogenetik menggunakan Algoritma Neighbor-Joining untuk Identifikasi Host dan Penyebaran Epidemi SARS* yang kemudian dipublikasikan dalam *Journal of Theoretical and Applied Information Technology.* Dari sana kemudian penelitian berlanjut dengan diperolehnya Hibah Penelitian Dosen Pemula (PDP) dari Kemenristekdikti pada tahun 2016 dan 2017, dan di tahun yang sama hingga tahun 2018 mendapatkan Hibah Penelitian Kerjasama Antar Perguruan Tinggi (PKPT) yang bermitra dengan ITS. Pernah mendapat penghargaan sebagai Dosen Berprestasi di UNISDA pada tahun 2017, dan di saat yang sama juga terseleksi untuk mengikuti *The 1st Indonesia-Taiwan Research Enhancement Workshop Series* di National University of Kaohsiung, atas biaya pemerintah Taiwan.

Mohammad Syaiful Pradana (syaifulp@unisda.ac.id), lahir di Gresik, 14 April 1990. Alumni magister matematika Institut Teknologi Sepuluh Nopember (ITS) Surabaya. Saat ini tercatat sebagai dosen tetap di Fakultas MIPA Universitas Islam Darul Ulum Lamongan. Memiliki *research interest* dalam bidang Pemodelan Matematika, Data Mining, dan Bioinformatika. Tertarik dengan menulis dan saat ini sedang mengembangkan komunitas menulis (*Writing Academy*) untuk menyalurkan pengetahuan menulis bagi mahasiswa, dosen dan umum

Mohammad Isa Irawan (mii@its.ac.id), lahir di Surabaya, 25 Desember 1963. Profesor di bidang Matematika dan Komputer Sains ini menempuh pendidikan S1 di Jurusan Matematika Universitas Airlangga Surabaya, S2 di Teknik Elektro Institut Teknlogi Bandung (ITB), dan S3 di bidang Ilmu Komputer di University of Technology Vienna – Austria. Memiliki *research interest* di bidang Bioinformatika, Jaringan Syaraf Tiruan dan *Decission Support System*. Saat ini menjabat sebagai Kepala Laboratorium Ilmu Komputasi di Departemen Matematika ITS dan aktif mengajar di Prodi S1 dan S2 Departemen Matematika di kampus yang sama, serta di Magister Management Teknik (MMT) ITS. Beliau aktif melakukan penelitian dan mendapatkan hibah penelitian dari Kemenristekdikti setiap tahunnya untuk skema yang berbeda-beda. Pernah menerima penghargaan Satya Lencana Karya Satya 10 tahun oleh Presiden RI pada tahun 2004 dan terpilih menjadi Dosen berprestasi peringkat I di FMIPA ITS tahun 2010.

Imam Mukhlash adalah dosen di Departemen Matematika Fakultas Matematika, Komputasi dan Sains Data, Institut Teknologi Sepuluh Nopember Surabaya. Saat ini menjabat sebagai Kepala Departemen Matematika. Beliau menyelesaikan S1 di Jurusan Matematika ITS, S2 dan S3 di Jurusan Teknik Informatika ITB. Mempunyai ketertarikan riset di bidang Data Mining, Fuzzy Logic, dan Artificial Intelligent. Mendapatkan beberapa hibah penelitian dari Lembaga Penelitian dan Pengabdian Masyarakat Institut Teknologi Sepuluh Nopember Surabaya maupun hibah penelitian dari Kemenristekdikti. Beliau juga aktif dalam menulis artikel di jurnal maupun konferensi baik Nasional maupun Internasional.

Lampiran 1 Data Coronavirus

1. Murine HV1
   - Definition : Spike glycoprotein (Murine Hepatitis Virus strain JHM)
   - Kode akses : YP_209233
   - Panjang sequence : 1376 aa
   - Host : Tikus
2. Murine HV2
   - Definition : E2 Glycoprotein precursor (Murine Hepatitis Virus strain A59)
   - Kode akses : NP_045300.1
   - Panjang sequence : 1324 aa
   - Host : Tikus
3. Human SARS Co-V
   - Definition : Spike glycoprotein (SARS coronavirus Tor2)
   - Kode akses : AAP41037
   - Panjang sequence : 1255 aa
   - Host : Human
4. Palm Civet
   - Definition : Spike glycoprotein (SARS coronavirus PC4-241)
   - Kode akses : AAV49723
   - Panjang sequence : 1255 aa
   - Host : Musang
5. Canine Co-V1
   - Definition : Spike glycoprotein (Canine Enteric Coronavirus K378)
   - Kode akses : Q65984
   - Panjang sequence : 1453 aa
   - Host : Anjing

6. Feline Co-V4
   Definition : Peplomer Protein (Feline Infectious Peritonitis Virus)
   Kode akses : BAA06805
   Panjang sequence : 1464 aa
   Host : Kucing
7. Porcine PEDV
   Definition : Spike Protein (Porcine Epidemic Diarrhea Virus)
   Kode akses : NP_598310
   Panjang sequence : 1383 aa
   Host : Babi
8. IBV 3
   Definition : Spike Protein (Infectious Bronchitis Virus)
   Kode akses : NP_040831
   Panjang sequence : 1162 aa
   Host : Ayam
9. Porcine HEV 3
   Definition : Spike Glycoprotein (Porcine Hemagglutinating Encephalomyelitis Virus)
   Kode akses : AAL80031
   Panjang sequence : 1349 aa
   Host : Babi
10. Bovine CoV 1
   Definition : Spike Protein (Bovine Coronavirus Isolate BcoV-LUN)
   Kode akses : AAL57308

Panjang sequence : 1363 aa
Host : Babi

11. Bovine CoV 2
Definition : Spike Structural Protein (Bovine Coronavirus Isolate BcoV-ENT)
Kode akses : AAK83356
Panjang sequence : 1363 aa
Host : Babi

12. Human Coronavirus OC43
Definition : S Protein (Human Coronavirus OC43)
Kode akses : NP_937950
Panjang sequence : 1361 aa
Host : Human

Lampiran 2. Hasil Penyejajaran Sekuen Protein Host Coronavirus

Tabel 1. Hasil penyejajaran sekuen protein host Coronavirus

No	Pasangan sekuen	Kode akses	Panjang Penyejajaran	Penalti (beda)	Jarak genetik
1.	Sequence 1 dan Sequence 2	YP_2029233 dan NP_045300	1376	154	0.1119
2.	Sequence 1 dan Sequence 3	YP_2029233 dan AAP41037	1376	1179	0.8568
3.	Sequence 1 dan Sequence 4	YP_2029233 dan AAV49723	1410	956	0.6780
4.	Sequence 1 dan Sequence 5	YP_2029233 dan Q65984	1503	1087	0.7232
5.	Sequence 1 dan Sequence 6	YP_2029233 dan BAA06805	1495	1096	0.7331
6.	Sequence 1 dan Sequence 7	YP_2029233 dan NP_598310	1460	1071	0.7336
7.	Sequence 1 dan Sequence 8	YP_2029233 dan NP_040831	1396	1026	0.7350
8.	Sequence 1 dan Sequence 9	YP_2029233 dan AAL80031	1389	489	0.3521
9.	Sequence 1 dan Sequence 10	YP_2029233 dan AAL57308	1395	487	0.3491
10.	Sequence 1 dan Sequence 11	YP_2029233 dan AAK83356	1395	488	0.3498

11.	Sequence 1 dan Sequence 12	YP_2029233 dan NP_937950	1398	507	0.3627
12.	Sequence 2 dan Sequence 3	NP_045300 dan AAP41037	1366	931	0.6816
13.	Sequence 2 dan Sequence 4	NP_045300 dan AAV49723	1368	929	0.6791
14.	Sequence 2 dan Sequence 5	NP_045300 dan Q65984	1470	1091	0.7422
15.	Sequence 2 dan Sequence 6	NP_045300 dan BAA06805	1494	1094	0.7323
16.	Sequence 2 dan Sequence 7	NP_045300 dan NP_598310	1428	1063	0.7444
17.	Sequence 2 dan Sequence 8	NP_045300 dan NP_040831	1347	980	0.7275
18.	Sequence 2 dan Sequence 9	NP_045300 dan AAL80031	1369	493	0.3601
19.	Sequence 2 dan Sequence 10	NP_045300 dan AAL57308	1386	498	0.3593
20.	Sequence 2 dan Sequence 11	NP_045300 dan AAK83356	1386	497	0.3586
21.	Sequence 2 dan Sequence 12	NP_045300 dan NP_937950	1375	501	0.3644
22.	Sequence 3 dan Sequence 4	AAP41037 dan AAV49723	1255	16	0.0127
23.	Sequence 3 dan Sequence 5	AAP41037 dan Q65984	1470	1077	0.7327
24.	Sequence 3 dan Sequence 6	AAP41037 dan BAA06805	1482	1081	0.7294

25.	Sequence 3 dan Sequence 7	AAP41037 dan NP_598310	1412	1037	0.7344
26.	Sequence 3 dan Sequence 8	AAP41037 dan NP_040831	1284	934	0.7274
27.	Sequence 3 dan Sequence 9	AAP41037 dan AAL80031	1376	961	0.6984
28.	Sequence 3 dan Sequence 10	AAP41037 dan AAL57308	1393	949	0.6813
29.	Sequence 3 dan Sequence 11	AAP41037 dan AAK83356	1393	948	0.6805
30.	Sequence 3 dan Sequence 12	AAP41037 dan NP_937950	1385	951	0.6866
31.	Sequence 4 dan Sequence 5	AAV49723 dan Q65984	1470	1074	0.7306
32.	Sequence 4 dan Sequence 6	AAV49723 dan BAA06805	1480	1081	0.7304
33.	Sequence 4 dan Sequence 7	AAV49723 dan NP_598310	1403	1038	0.7398
34.	Sequence 4 dan Sequence 8	AAV49723 dan NP_040831	1287	930	0.7226
35.	Sequence 4 dan Sequence 9	AAV49723 dan AAL80031	1379	956	0.6933
36.	Sequence 4 dan Sequence 10	AAV49723 dan AAL57308	1392	949	0.6818
37.	Sequence 4 dan Sequence 11	AAV49723 dan AAK83356	1392	948	0.6810
38.	Sequence 4 dan Sequence 12	AAV49723 dan NP_937950	1387	950	0.6849
39.	Sequence 5 dan Sequence 6	Q65984 dan BAA06805	1509	815	0.5401

40.	Sequence 5 dan Sequence 7	Q65984 dan NP_598310	1471	818	0.5561
41.	Sequence 5 dan Sequence 8	Q65984 dan NP_040831	1472	1052	0.7147
42.	Sequence 5 dan Sequence 9	Q65984 dan AAL80031	1489	1087	0.7300
43.	Sequence 5 dan Sequence 10	Q65984 dan AAL57308	1485	1111	0.7481
44.	Sequence 5 dan Sequence 11	Q65984 dan AAK83356	1480	1106	0.7473
45.	Sequence 5 dan Sequence 12	Q65984 dan NP_937950	1489	1086	0.7293
46.	Sequence 6 dan Sequence 7	BAA06805 dan NP_598310	1493	850	0.5693
47.	Sequence 6 dan Sequence 8	BAA06805 dan NP_040831	1482	1075	0.7254
48.	Sequence 6 dan Sequence 9	BAA06805 dan AAL80031	1497	1086	0.7255
49.	Sequence 6 dan Sequence 10	BAA06805 dan AAL57308	1501	1091	0.7268
50.	Sequence 6 dan Sequence 11	BAA06805 dan AAK83356	1501	1091	0.7268
51.	Sequence 6 dan Sequence 12	BAA06805 dan NP_937950	1488	1095	0.7359
52.	Sequence 7 dan Sequence 8	NP_598310 dan NP_040831	1401	995	0.7102
53.	Sequence 7 dan Sequence 9	NP_598310 dan AAL80031	1444	1067	0.7389
54.	Sequence 7 dan Sequence 10	NP_598310 dan AAL57308	1455	1066	0.7326

55.	Sequence 7 dan Sequence 11	NP_598310 dan AAK83356	1454	1064	0.7318
56.	Sequence 7 dan Sequence 12	NP_598310 dan NP_937950	1453	1054	0.7254
57.	Sequence 8 dan Sequence 9	NP_040831 dan AAL80031	1374	982	0.7147
58.	Sequence 8 dan Sequence 10	NP_040831 dan AAL57308	1388	999	0.7197
59.	Sequence 8 dan Sequence 11	NP_040831 dan AAK83356	1388	999	0.7197
60.	Sequence 8 dan Sequence 12	NP_040831 dan NP_937950	1388	997	0.7183
61.	Sequence 9 dan Sequence 10	AAL80031 dan AAL57308	1363	240	0.1761
62.	Sequence 9 dan Sequence 11	AAL80031 dan AAK83356	1363	239	0.1753
63.	Sequence 9 dan Sequence 12	AAL80031 dan NP_937950	1366	259	0.1896
64.	Sequence 10 dan Sequence 11	AAL57308 dan AAK83356	1363	7	0.0051
65.	Sequence 10 dan Sequence 12	AAL57308 dan NP_937950	1373	118	0.0859
66.	Sequence 11 dan Sequence 12	AAK83356 dan NP_937950	1373	118	0.0859

Lampiran 3. Representasi Mutasi Antar Dua Sekuen

Tabel 2. Mutasi antara Dua Sekuen

Sekuen 1	Sekuen 2	Nucleotide		Position	Mutation	Prosentage	
		Diff.	Similar			Diff.	Similar
A. Guangzhou 16/12/02	B. Zhongsan 26/12/02	4	3764	224	C to G	0,1 %	99,9 %
				654	C to T		
				1502	A to T		
				2380	T to C		
	C. Zhongsan 4/01/03	9	3759	654	C to T	0,2 %	99,8 %
				716	T to C		
				931	A to G		
				1026	G to A		
				1031	G to A		
				1502	A to T		
				2332	G to T		
				2380	T to C		
				3075	C to T		
	D. Guangzhou 24/01/03	12	3756	230	A to G	0,3 %	99,7 %
				654	C to T		
				716	T to C		
				731	C to T		
				931	A to C		
				1026	G to A		
				1031	G to A		
				1502	A to T		
				2332	G to T		
				2380	T to C		
				3075	C to T		
				3505	A to G		
	E. GuangZhou Hospital	8	3760	654	C to T	0,2 %	99,8 %
				716	T to C		
				931	A to C		
				1031	G to A		
				1502	A to T		
				2332	G to T		
				2380	T to C		
				3075	C to T		
	F. GuangZhou 18/2/03	8	3760	654	C to T	0,2 %	99,8 %
				716	T to C		
				931	A to C		
				1031	G to A		
				1502	A to T		
				2332	G to T		
				2380	T to C		
				3075	C to T		
	G. Metropole 21/02/03	9	3759	654	C to T	0,2 %	99,8 %
				716	T to C		
				931	A to G		
				1026	G to A		
				1031	G to A		
				1502	A to T		
				2332	G to T		
				2380	T to C		

				3075	C to T		
	H. Hanoi 16/02/03	12	3756	230	A to G	0,3 %	99,7 %
				654	C to T		
				716	T to C		
				731	C to T		
				931	A to C		
				1026	G to A		
				1031	G to A		
				1502	A to T		
				2332	G to T		
				2380	T to C		
				3075	C to T		
				3381	T to C		
	I. Toronto 27/02/03	12	3756	230	A to G	0,3 %	99,7 %
				654	C to T		
				716	T to C		
				731	C to T		
				931	A to C		
				1026	G to A		
				1031	G to A		
				1502	A to T		
				1729	T to G		
				2332	G to T		
				2380	T to C		
				3075	C to T		
	J. Singapore 01/03/03	11	3757	230	A to G	0,3 %	99,7 %
				654	C to T		
				716	T to C		
				731	C to T		
				931	A to C		
				1026	G to A		
				1031	G to A		
				1502	A to T		
				2332	G to T		
				2380	T to C		
				3075	C to T		
	K. Taiwan 08/03/03	11	3757	230	A to G	0,3 %	99,7 %
				654	C to T		
				716	T to C		
				731	C to T		
				931	A to C		
				1026	G to A		
				1031	G to A		
				1502	A to T		
				2332	G to T		
				2380	T to C		
				3075	C to T		
	L. Hong Kong 19/03/03	11	3757	230	A to G	0,3 %	99,7 %
				654	C to T		
				716	T to C		
				731	C to T		
				931	A to C		
				1026	G to A		
				1031	G to A		
				1502	A to T		

				2332	G to T		
				2380	T to C		
				3075	C to T		
	M. Hong Kong 15/05/03	11	3757	81	T to G	0,3 %	99,7 %
				230	A to G		
				654	C to T		
				716	T to C		
				731	C to T		
				931	A to C		
				1026	G to A		
				1031	G to A		
				1502	A to T		
				2332	G to T		
				2380	T to C		
				3075	C to T		
				3623	C to T		
	N. Palm Civet	3	3775	1502	A to T	0,08 %	99,92 %
				2380	T to C		
				3487	A to G		
B. Zhongsan 26/12/02	C. Zhongsan 4/01/03	7	3761	224	G to C	0,2 %	99,8 %
				716	T to C		
				931	A to G		
				1026	G to A		
				1031	G to A		
				2332	G to C		
				3075	C to T		
	D. Guangzhou 24/01/03	10	3758	224	G to C	0,3 %	99,7 %
				230	A to G		
				716	T to C		
				731	C to T		
				931	A to G		
				1026	G to A		
				1031	G to A		
				2332	G to C		
				3075	C to T		
				3505	A to G		
	E. GuangZhou Hospital	6	3762	224	G to C	0,16 %	99,84 %
				716	T to C		
				931	A to G		
				1031	G to A		
				2332	G to C		
				3075	C to T		
	F. GuangZhou 18/2/03	7	3761	146	C to T	0,2 %	99,8 %
				224	G to C		
				716	T to C		
				931	A to G		
				1026	G to A		
				1031	G to A		
				3075	C to T		
	G. Metropole 21/02/03	7	3761	224	G to C	0,2 %	99,8 %
				716	T to C		
				931	A to G		
				1026	G to A		
				1031	G to A		
				2332	G to C		

				3075	C to T		
	H. Hanoi 26/02/03	10	3758	224	G to C	0,3 %	99,7 %
				230	A to G		
				716	T to C		
				731	C to T		
				931	A to G		
				1026	G to A		
				1031	G to A		
				2332	G to C		
				3075	C to T		
				3381	T to C		
	I. Toronto 27/02/03	10	3758	224	G to C	0,3 %	99,7 %
				230	A to G		
				716	T to C		
				731	C to T		
				931	A to G		
				1026	G to A		
				1031	G to A		
				1729	T to G		
				2332	G to C		
				3075	C to T		
	J. Singapore 01/03/03	9	3759	224	G to C	0,2 %	99,8 %
				230	A to G		
				716	T to C		
				731	C to T		
				931	A to G		
				1026	G to A		
				1031	G to A		
				2332	G to C		
				3075	C to T		
	K. Taiwan 08/03/03	9	3759	224	G to C	0,2 %	99,8 %
				230	A to G		
				716	T to C		
				731	C to T		
				931	A to G		
				1026	G to A		
				1031	G to A		
				2332	G to C		
				3075	C to T		
	L. Hong Kong 19/03/03	9	3759	224	G to C	0,2 %	99,8 %
				230	A to G		
				716	T to C		
				731	C to T		
				931	A to G		
				1026	G to A		
				1031	G to A		
				2332	G to C		
				3075	C to T		
	M. Hong Kong 15/05/03	11	3757	81	T to G	0,3 %	99,7 %
				224	G to C		
				230	A to G		
				716	T to C		
				731	C to T		
				931	A to G		
				1026	G to A		

				1031	G to A		
				2332	G to C		
				3075	C to T		
				3623	C to T		
	N. Palm Civet	3	3765	224	G to C	0,08 %	99,92 %
				654	T to C		
				3487	A to G		
C. Zhongsan 4/01/03	D. Guangzhou 24/01/03	3	3765	230	A to G	0,08 %	99,92 %
				731	C to T		
				3305	A to G		
	E. GuangZhou Hospital	1	3767	1026	A to G	0,02%	99,98%
	F. GuangZhou 18/2/03	2	3766	146	C to T	0,05%	99.95%
				2332	T to G		
	G. Metropole 21/02/03	0	3768			0%	100%
	H. Hanoi 26/02/03	3	3765	230	A to G	0,08%	99,92%
				731	C to T		
				3381	T to C		
	I. Toronto 27/02/03	3	3765	230	A to G	0,08%	99,92%
				731	C to T		
				1729	T to G		
	J. Singapore 01/03/03	2	3766	230	A to G	0,05%	99,95%
				731	C to T		
	K. Taiwan 08/03/03	2	3766	230	A to G	0,05%	99,95%
				731	C to T		
	L. Hong Kong 19/03/03	2	3766	230	A to G	0,05%	99,95%
				731	C to T		
	M. Hong Kong 15/05/03	4	3764	81	T to G	0,11%	99,89%
				230	A to G		
				731	C to T		
				3623	C to T		
	N. Palm Civet	8	3760	654	T to C	0,21%	99,79%
				716	C to T		
				931	G to A		
				1026	A to G		
				1031	A to G		
				2332	T to G		
				3075	T to C		
				3487	A to G		
D. Guangzhou 24/01/03	E. GuangZhou Hospital	4	3764	230	G to A	0,11%	99,89%
				731	T to C		
				1026	A to G		
				3505	G to A		
	F. GuangZhou 18/2/03	5	3763	146	C to T	0,13%	99,87%
				230	G to A		
				731	T to C		
				2332	T to G		
				3505	G to A		
	G. Metropole 21/02/03	3	3765	230	G to A	0,08 %	99,92 %
				731	T to C		
				3505	G to A		
	H. Hanoi 26/02/03	2	3766	3381	T to C	0,05 %	99,95 %
				3505	G to A		

	I. Toronto 27/02/03	2	3766	1729	T to G	0,05 %	99,95 %
				3505	G to A		
	J. Singapore 01/03/03	1	3767	3505	G to A	0,03 %	99,97 %
	K. Taiwan 08/03/03	1	3767	3505	G to A	0,03 %	99,97 %
	L. Hong Kong 19/03/03	1	3767	3505	G to A	0,03 %	99,97 %
	M. Hong Kong 15/05/03	3	3765	81	T to G	0,08 %	99,92 %
				3505	G to A		
				3623	C to T		
	N. Palm Civet	11	3757	230	G to A	0,30 %	99,70 %
				654	T to C		
				716	C to T		
				731	T to C		
				931	G to A		
				1026	A to G		
				1031	A to G		
				2332	T to G		
				3075	T to C		
				3487	A to G		
				3505	G to A		
E. GuangZhou Hospital	F. GuangZhou 18/2/03	3	3765	146	C to T	0,08 %	99,92 %
				1026	G to A		
				2332	T to G		
	G. Metropole 21/02/03	1	3767	1026	G to A	0,03 %	99,97 %
	H. Hanoi 26/02/03	4	3764	230	A to G	0,10 %	99,90 %
				731	T to C		
				1026	G to A		
				3381	T to C		
	I. Toronto 27/02/03	4	3764	230	A to G	0,10 %	99,90 %
				731	T to C		
				1026	G to A		
				3381	T to C		
	J. Singapore 01/03/03	3	3765	230	A to G	0,08 %	99,92 %
				731	T to C		
				1026	G to A		
	K. Taiwan 08/03/03	3	3765	230	A to G	0,08 %	99,92 %
				731	T to C		
				1026	G to A		
	L. Hong Kong 19/03/03	3	3765	230	A to G	0,08 %	99,92 %
				731	T to C		
				1026	G to A		
	M. Hong Kong 15/05/03	5	3763	81	T to G	0,13 %	99,87 %
				230	A to G		
				731	T to C		
				1026	G to A		
				3623	C to T		
	N. Palm Civet	7	3761	654	T to C	0,19 %	99,81 %
				716	C to T		
				931	G to A		
				1031	A to G		
				2332	T to G		
				3075	T to C		

				3487	A to G		
F. GuangZhou 18/2/03	G. Metropole 21/02/03	2	3766	146	T to C	0,05 %	99,95 %
				2332	G to T		
	H. Hanoi 26/02/03	5	3763	146	T to C	0,13 %	99,87 %
				230	A to G		
				731	C to T		
				2332	G to C		
				3381	T to C		
	I. Toronto 27/02/03	5	3763	146	T to C	0,13 %	99,87 %
				230	A to G		
				731	C to T		
				1729	T to G		
				2332	G to C		
	J. Singapore 01/03/03	4	3764	146	T to C	0,10 %	99,90 %
				230	A to G		
				731	C to T		
				2332	G to C		
	K. Taiwan 08/03/03	4	3764	146	T to C	0,10 %	99,90 %
				230	A to G		
				731	C to T		
				2332	G to C		
	L. Hong Kong 19/03/03	4	3764	146	T to C	0,10%	99,90%
				230	A to G		
				731	C to T		
				2332	G to T		
	M. Hong Kong 15/05/03	6	3762	81	T to G	0,16%	99,84%
				146	T to C		
				230	A to G		
				731	C to T		
				2332	G to T		
				3623	C to T		
	N. Palm Civet	8	3760	146	T to C	0,21%	99,79%
				654	T to C		
				716	C to T		
				931	G to A		
				1026	A to G		
				1031	A to G		
				3075	T to C		
				3487	A to G		
G. Metropole 21/02/03	H. Hanoi 26/02/03	3	3765	230	A to G	0,08%	99,92%
				731	C to T		
				3381	T to C		
	I. Toronto 27/02/03	3	3765	230	A to G	0,08%	99,92%
				731	C to T		
				1729	T to G		
	J. Singapore 01/03/03	2	3766	230	A to G	0,05%	99,95%
				731	C to T		
	K. Taiwan 08/03/03	2	3766	230	A to G	0,05%	99,95%
				731	C to T		
	L. Hong Kong 19/03/03	2	3766	230	A to G	0,05%	99,95%
				731	C to T		
	M. Hong Kong 15/05/03	4	3764	81	T to G	0,10%	99,90%
				730	A to G		
				731	C to T		
				3623	C to T		

	N. Palm Civet	8	3760	654	T to C	0,21%	99,79%
				716	C to T		
				931	G to A		
				1026	A to G		
				1031	A to G		
				2332	T to G		
				3075	T to C		
				3487	A to G		
H. Hanoi 26/02/03	I. Toronto 27/02/03	2	3766	1729	T to G	0,05%	99,95%
				3381	C to T		
	J. Singapore 01/03/03	1	3767	3381	C to T	0,03%	99,97%
	K. Taiwan 08/03/03	1	3767	3381	C to T	0,03%	99,97%
	L. Hong Kong 19/03/03	1	3767	3381	C to T	0,03%	99,97%
	M. Hong Kong 15/05/03	3	3765	81	T to G	0,08%	99,92%
				3381	C to T		
				3623	C to T		
	N. Palm Civet	11	3757	230	G to A	0,30%	99,70%
				654	T to C		
				716	C to T		
				731	T to C		
				931	G to A		
				1026	A to G		
				1031	A to G		
				2332	T to G		
				3075	T to C		
				3381	C to T		
				3487	A to G		
I. Toronto 27/02/03	J. Singapore 01/03/03	1	3767	1729	G to T	0,03%	99,97%
	K. Taiwan 08/03/03	1	3767	1729	G to T	0,03%	99,97%
	L. Hong Kong 19/03/03	1	3767	1729	G to T	0,03%	99,97%
	M. Hong Kong 15/05/03	3	3765	81	T to G	0,08 %	99,92 %
				1729	T to G		
				3623	C to T		
	N. Palm Civet	11	3757	230	G to A	0,30 %	99,70 %
				654	T to C		
				716	C to T		
				731	T to C		
				931	G to A		
				1026	A to G		
				1031	A to G		
				1729	G to T		
				2332	T to G		
				3075	T to C		
				3487	A to G		
J. Singapore 01/03/03	K. Taiwan 08/03/03	0	3768			0%	100%
	L. Hong Kong 19/03/03	0	3768			0%	100%
	M. Hong Kong	2	3766	81	T to G	0,05%	99,95%

	15/05/03			3623	C to T		
	N. Palm Civet	10	3758	230	A to G	0,26%	99,74%
				654	T to C		
				716	C to T		
				731	C to T		
				931	G to A		
				1026	A to G		
				1031	A to G		
				2332	T to G		
				3075	T to C		
				3487	A to G		
K. Taiwan 08/03/03	L. Hong Kong 19/03/03	0	3768			0%	100%
	M. Hong Kong 15/05/03	0	3768			0%	100%
	N. Palm Civet	10	3758	230	A to G	0,26%	99,74%
				654	T to C		
				716	C to T		
				731	C to T		
				931	G to A		
				1026	A to G		
				1031	A to G		
				2332	T to G		
				3075	T to C		
				3487	A to G		
L. Hong Kong 19/03/03	M. Hong Kong 15/05/03	2	3766	81	T to G	0,05%	99,95%
				3623	C to T		
	N. Palm Civet	10	3758	230	A to G	0,26%	99,74%
				654	T to C		
				716	C to T		
				731	C to T		
				931	G to A		
				1026	A to G		
				1031	A to G		
				2332	T to G		
				3075	T to C		
				3487	A to G		
M. Hong Kong 15/05/03	N. Palm Civet	12	3756	81	G to T	0,32	99,68
				230	G to A		
				654	T to C		
				716	C to T		
				731	T to C		
				931	G to A		
				1026	A to G		
				1031	A to G		
				2332	T to G		
				3075	T to C		
				3487	A to G		
				3623	T to C		

ISBN : 978-602-6715-95-1

BIOINFORMATIKA

Perspektif Matematika Pada Analisis Sekuen dan Filogenetika

Bioinformatika sebagai suatu disiplin ilmu yang baru berkembang di Indonesia, belum banyak dikenal masyarakat terutama dalam kaitannya dengan matematika. Padahal Bioinformatika sebagai ilmu yang menerapkan teknik komputasional untuk mengolah dan menganalisis informasi biologis memang mencakup penerapan metode-metode matematika, statistika dan informatika untuk memecahkan masalah-masalah biologi terutama dengan mengunakan sekuens DNA dan asam amino serta informasi yang berkaitan.

Buku ini menyajikan Bioinformatika ditinjau dari perspektif matematika, terutama pada bahasan analisis sekuens dan filogenetika, dengan mengambil contoh kasus pada epidemi SARS. Dengan membaca buku ini, akan membantu mahasiswa untuk memahami metode matematikayang diterapkan dibidang Bioinformatika terutama tentang kasus penyebaran Epidemi. Memang "bukan sekedar epidemi", karena dari sebuah epidemi akan banyak imu yang bisa dipelajari dan dikembangkan. Dari sebuah epidemi pula, banyak diterapkan algoritma-algoritma yang berbasis matematika.

Meskipun buku ini utamanya diperuntukkan bagi mahasiswa, namun buku ini dapat juga dibaca oleh khalayak umum dan praktisi yang tertarik dan berhubungan dengan Bioinformatika.

Buku ini merupakan luaran dari hasil penelitian kerjasama antar perguruan tinggi yang dibayai oleh Direktorat Riset dan Pengabdian Masyarakat, Direktorat Jenderal Penguatan Riset dan Pengembangan, Kementrian Riset, Teknologi dan Pendidikan Tinggi sesuai dengan kontrak penelitian tahun 2018.

Pustaka ILALANG

CV. PUSTAKA ILALANG Group®
Jl. Airlangga No. 3 (Depan BRI Unisda) Sukodadi Lamongan
Jalan raya Lamongan - Mantup 16 km
Kembangbahu Lamongan 62282
Email: pustaka_ilalang@yahoo.co.id

RISTEKDIKTI

ISBN 978-602-6715-95-1
9 786026 715951

Siti Amiroch, dkk

BIOINFORMATIKA - Perspektif Matematika Pada Analisis Sekuen dan Filogenetika