

PEMODELAN PROSES PENYARINGAN RETROVIRUS ENDOGENOUS
MANUSIA DAN MENGIDENTITIKAN JUJUKAN ASID AMINO RETROVIRUS
ENDOGENOUS

YU SHUN AY

DISERTASI INI DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT MEMPEROLEHI IJAZAH SARJANA MUDA SAINS
DENGAN KEPUJIAN

PERPUSTAKAAN
UNIVERSITI MALAYSIA SABAH

PROGRAM MAEMATIK DENGAN EKONOMI
SEKOLAH SAINS DAN TEKNOLOGI
UNIVERSITI MALAYSIA SABAH

MAC 2007



UMS
UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN STATUS TESIS@

JUDUL: PEMODELAN PROSES PENYARINGAN RETROVIRUS ENDOGENOUS MANUSIA

DAN MENGIDEN TI TIKAN JVJUKAN ASID AMINU RETROVIRUS ENDOGENOUS

Ijazah: IJAZAH SARJANA MUDA SAINS

SESSI PENGAJIAN: 3 TAHUN

Saya YU SHUN AY

(HURUF BESAR)

mengaku membenarkan tesis (LPS/Sarjana/Doktor Falsafah)* ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:

1. Tesis adalah hak milik Universiti Malaysia Sabah.
2. Perpustakaan Universiti Malaysia Sabah dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. **Sila tandakan (/)

SULIT

TERHAD

TIDAK TERHAD

PERPUSTAKAAN
UNIVERSITI MALAYSIA SABAH

(Mengandungi maklumat yang berdarjah keselamatan atau
kepentingan Malaysia seperti yang termaktub di dalam
AKTA RAHSIA RASMI 1972)

Disahkan oleh

Y.S.Y. (TANDATANGAN PUSTAKAWAN)

(TANDATANGAN PENULIS)

Alamat Tetap: NO. 47, 2½ MILE MILES,
OLD ADOYA ROAD, 96000 SIBU,

SAWARAK

Nama Penyelia

Tarikh: 23 APRIL 2007

Tarikh:

CATATAN: * Potong yang tidak berkenaan.

** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi
berkenaan dengan menyatakan sekali sebab dan tempoh tesis ini perlu dikelaskan sebagai SULIT
dan TERHAD.

@ Tesis dimaksudkan sebagai tesis bagi Ijazah Doktor Falsafah dan Sarjana secara penyelidikan, atau
disertasi bagi pengajian secara kerja kursus dan penyelidikan, atau Laporan Projek Sarjana Muda
(LPSM).

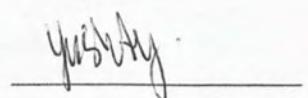


UMS
UNIVERSITI MALAYSIA SABAH

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya kecuali nukilan dan ringkasan yang setiap satunya telah dijelaskan sumbernya.

12 MAC 2007



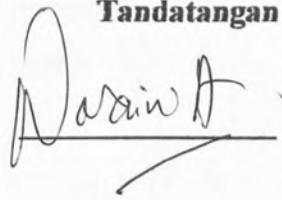
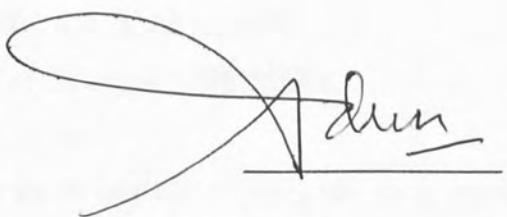
YU SHUN AY

HS2004-4588

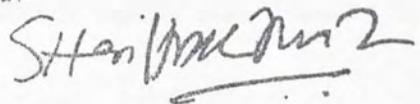
PERPUSTAKAAN
UNIVERSITI MALAYSIA SABAH



UMS
UNIVERSITI MALAYSIA SABAH

DIPERAKUKAN OLEH**1. PENYELIA****(PUAN NORAINI ABDULLAH)****Tandatangan****2. PEMERIKSA****(PROF. DR. ZAINODIN HAJI JUBOK)****3. DEKAN****(PROF. MADYA DR. SHARIFF A.K OMANG)**

PERPUSTAKAAN
UNIVERSITI MALAYSIA SABAH

**UMS**

UNIVERSITI MALAYSIA SABAH

PENGHARGAAN

Saya ingin mengambil kesempatan untuk mengucapkan terima kasih kepada semua individu yang terlibat dalam membantu saya menjayakan projek.

Terlebih dahulu, saya ingin berterima kasih kepada penyelia projek saya iaitu Puan Noraini Abdullah yang sentiasa membimbing saya. Saya berasa suka hati menjalankan projek saya di bawah Puan Noraini, kerana beliau sentiasa memberi sokongan dan motivasi supaya saya membuat kerja dengan lebih dedikasi.

Selain itu, saya juga ingin berterima kasih kepada Dr. Roziah yang memberi tunjuk ajar yang berkaitan dengan topik projek saya. Saya juga ingin memberikan penghargaan kepada rakan-rakan yang sentiasa memberi sokongan moral kepada saya. Akhirnya, saya ingin mengucapkan terima kasih kepada rakan saya Nancy yang menolong saya dalam memahami pengaturcaraan C.



ABSTRAK

Penyaringan retrovirus endogenous adalah dengan menggunakan teknologi bioinformatik iaitu termasuk BLAST , CLUSTALW, pokok filogenetik. Menurut penyelidik-penyalidik yang telah menjalankan kajian penyaringan retrovirus endogenous, terdapat tiga jenis retrovirus iaitu HERV-E, ERV-9, dan MSRV telah menjalankan penyelidikan untuk mencari jujukan-jujukan yang mempunyai sifat kesamaan yang tinggi dengannya. Daripada pencarian dan padanan jujukan-jujukan, kita boleh memperolehi banyak jujukan yang mempunyai persamaan yang tinggi dengan merujuk kepada nilai skor bitnya. Daripada maklumat jujukan-jujukan yang diperolehi, seperti nombor akses nukleotida, jujukan asid amino dan jenis kromosom, kita boleh menyimpan maklumat tersebut dalam pengaturcaraan C jujukan-jujukan yang ingin dicari hanya dengan menaip nombor akses nukleotidanya. Akhirnya, semua maklumat jujukan tersebut akan dipaparkan.



MODELLING SCREENING PROCESS HUMAN RETROVIRUS ENDOGENOUS
AND IDENTIFY THE SEQUENCES OF RETROVIRUS ENDOGENOUS

ABSTRACT

Screening the retrovirus endogenous involved the technology of bioinformatics, including BLAST, CLUTALW, and phylogenetic tree. Previous researchers had been done shown that there were three types of retrovirus endogenous, HERV-E, ERV-9, MSRV. They searched the similarities of these viruses. From the bioinformatics searching and matching, we can achieve many sequences alignment which have similarities that have the highest score bit. From the information of sequences that they obtained, such as the accession number nucleotide, sequences of the amino acid and the type of chromosome, we can stored it in C programming and searched from there, just by key in its accession number. All information of the sequences want to be searched will come out.



KANDUNGAN

	Muka Surat
PENGAKUAN	ii
PENGESAHAN	iii
PENGHARGAAN	iv
ABSTRAK	v
ABSTRACT	vi
SENARAI KANDUNGAN	vii
SENARAI JADUAL	x
SENARAI RAJAH	xi
SENARAI SIMBOL	xii

BAB 1 PENDAHULUAN

1.1 PENGENALAN	1
1.2 BIOINFORMATIK	2
1.2.1 BLAST	2
1.2.2 CLUSTALW	4
1.2.3 Pokok Filogenetik	4
1.3 PENGENALAN RETROVIRUS ENDOGENOUS	7
1.4 PANGKALAN DATA BIOLOGI	9
1.5 OBJEKTIF KAJIAN	12
1.6 SKOP KAJIAN	12



BAB 2 ULASAN PERPUSTAKAAN

2.1 PENGENALAN	13
2.2 PEMODELAN	15
2.3 PENGATURCARAAN KOMPUTER	16
2.3.1 PERISIAN KOMPUTER YANG DIGUNAKAN	17
2.4 KEPENTINGAN PENGIRAAN SKOR BIT	18

BAB 3 METODOLOGI

3.1 PENGENALAN	20
3.2 DATA DARIPDA PENCARIAN BLAST	20
3.3 MENGIRA SKOR BIT	21
3.4 MENYUSUN JUJUKAN HERVS	22
3.5 PEMBENTUKAN PENGATURCARAAN C	24

BAB 4 ANALISIS DAN PERBINCANGAN

4.1 PENGENALAN	27
4.2 ARAHAN BAGI PENGATURCARAAN C	27
4.3 ARAHAN BAGI PENGGUNAAN PENGATURCARAAN	29

BAB 5 KEPUTUSAN DAN CADANGAN

5.1 PENGENALAN	33
5.2 CADANGAN UNTUK PENYELIDIKAN MASA DEPAN	34
RUJUKAN	36



LAMPIRAN A	39
LAMPIRAN B	40
LAMPIRAN C	41
LAMPIRAN D	43



SENARAI JADUAL

No. Jadual	Muka Surat
1.1 Famili-famili Retrovirus Endogenous yang Tergolong Dalam Kelas I, II, III	8
1.2 Senarai Asid Amino Dengan Nama dan Simbol	10
1.3 Asid-asid Amino Yang Ditulis Dalam Kod IUB Dengan Sifat Fisiokimia	11



SENARAI RAJAH

No. Rajah	Muka Surat
1.1 Contoh Keputusan yang Diperolehi daripada Hasil Pencarian BLAST	3
1.2 Pokok Filogenetik Jenis Bebas	5
1.3 Pokok Filogenetik Jenis Radial	5
1.4 Pokok Filogenetik Jenis Kladogram	6
1.5 Pokok Filogenetik Jenis Filogram	6
1.6 Pokok Filogenetik Berakar	6
1.7 Pokok Filogenetik Jenis Serong	7
1.8 Motif Terpelihara Pada Domain I hingga V Gen Transkriptase Berbalik Dalam Retrovirus Endogenous Manusia	9
3.1 Carta Alir Pemodelan Proses Penyaringan	23
3.2 Carta Alir Bagi Menentukan Nombor Akses Nukleotida dan Genom	25
4.1(a) Contoh Output yang Dikeluarkan	31
4.1(b) Contoh Output yang Dikeluarkan	32



SENARAI SIMBOL

S = Skor jujukan

B = Skor Bit

λ = lambda /fungsi bagi matriks penskoran

K = Fungsi bagi saiz pangkalan data

\sum = Hasil tambah



BAB 1

PENDAHULUAN

1.1 PENGENALAN

Pada hari ini, sains komputer semakin berkembang dan telah memanfaatkan orang ramai kerana apa-apa saja boleh disimulasikan dengan menggunakan komputer. Mengikut prinsip Matematik dan Sains, simulasi tersebut dikenali sebagai “Pemodelan secara pengkomputeran”. Pemodelan secara pengkomputeran dapat menolong menyelesaikan pelbagai permasalahan termasuk masalah biologi.

Komputer merupakan alat pemodelan yang paling berguna, dan pemodelan berkomputer ialah menulis komputer program atau dikenali sebagai “*computer code*” berserta dengan data berangka untuk mensimulasikan sistem. Program komputer adalah susunan perintah-perintah dan operasi-operasi yang mengimplementasikan algoritma tertentu disertai yang ditulis dalam bahasa komputer tertentu. Bahasa komputer adalah teknik komunikasi piawai untuk memerintah komputer.



Penyaringan retrovirus endogenous manusia melibatkan teknologi bioinformatik. Tujuan penyaringan retrovirus endogenous manusia adalah untuk menyusun jujukan-jujukan hasil daripada penyaringan dan menganalisis hubungan di antara retrovirus endogenous manusia daripada autosom dan kromosom seks dengan retrovirus endogenous perumah lain contohnya ikan dan tikus.

1.2 BIOINFORMATIK

Bioinformatik ialah bidang sains yang menggabungkan biologi, sains komputer dan perisian dalam simulasi proses biokimia dan analisis data biologi molecular. Bioinformatik adalah pangkalan data yang digunakan untuk membentuk, mengurus dan menganalisis data-data biologi. Penyelidikan dalam bioinformatik adalah pembentukan cara untuk menyimpan, mendapat dan menganalisis data. Data-data adalah jujukan DNA atau asid amino, dan struktur protein tiga dimensi. Teknologi bioinformatik ini termasuk penggunaan program BLAST, Clustalw, dan pembinaan pokok filogenetik.

1.2.1 BLAST

Penyaringan jujukan-jujukan HERVs daripada Genbank adalah dengan menggunakan BLAST (*Basic Local Alignment Search Tool*). Genbank merupakan pangkalan data DNA daripada *National Center for Biotechnology Information* (NCBI). BLAST merupakan alat yang digunakan untuk mencari jujukan pangkalan data untuk dibuat penyusunan dan untuk mengenalpasti jujukan protein yang diperolehi. Ia juga boleh dikatakan adalah satu

algoritma yang digunakan untuk membandingkan jujukan protein asid-asid amino protein yang berbeza atau juga jujukan-jujukan nukleotida DNA. BLAST mempunyai algoritma pemadanan untuk mencari asid-asid amino yang sepadan dalam kecenderungan dua jujukan.

Melalui internet, laman web BLAST boleh dilayari dengan alamat laman <http://www.ncbi.nlm.nih.gov/blast/> dan <http://www.ncbi.nlm.nih.gov>.

Related Structures

Sequences producing significant alignments:

Score (bits)	E Value	2	3
419	e-117		
419	e-117		
419	e-117		
412	e-115		
366	e-101		
345	7e-95		
303	2e-82		
247	1e-65		
173	4e-43		
171	1e-42		

1

gi 5754190 ref NP_034554.1 hemochromatosis [Mus musculus] ...
gi 26354116 dbj BAC40688.1 unnamed protein product [Mus mu...]
gi 12844463 dbj BAB26373.1 unnamed protein product [Mus mu...]
gi 2574263 ref NP_445753.1 hemochromatosis [Rattus norveg...]
gi 2624957 gb AAB86597.1 hereditary hemochromatosis protei...
gi 2072657 emb CAA73197.1 MFE (HLA-H) [Mus musculus]
gi 5734363 gb AAD49965.1 AF176534.1 hemochromatosis gene pr...
gi 1930010 gb AAB51504.1 hereditary haemochromatosis prote...
gi 2225995 emb CAA74333.1 HNC class I alpha chain [Rattus ...]
gi 2851391 sp P16391 HA12 RAT RT1 class I histocompatibilit...

Petunjuk:

- Link untuk rujukan kepada jujukan.
- Skor bits yang mewakili tahap persamaan dalam penyusunan jujukan dari satu lokasi ke satu lokasi pada jujukan tersebut
- Jangkaan bagi peluang mencari pangkalan data terhadap jujukan tertentu

Rajah 1.1 Contoh Keputusan yang Diperolehi daripada Hasil Pencarian BLAST

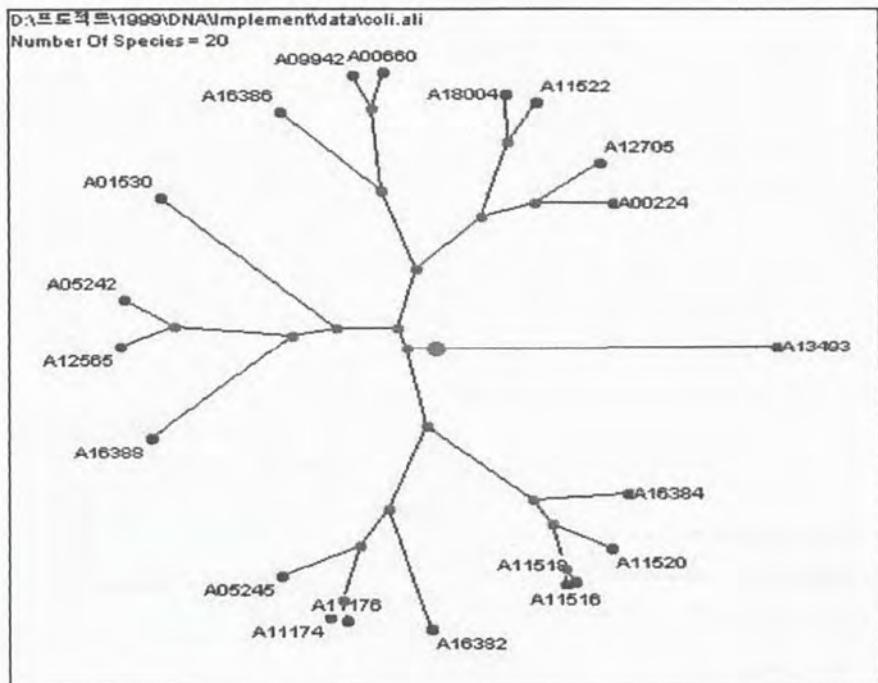
1.2.2 CLUSTALW

Selain pencarian BLAST, program Clustalw telah digunakan untuk menyusun jujukan berganda untuk DNA atau protein. Tujuan program ini digunakan adalah untuk menghitung pasangan yang sesuai bagi jujukan yang dipilih, dan menyusunnya supaya persamaan dan perbezaan dapat dilihat dengan jelas. Jujukan yang mempunyai persamaan yang tinggi akan disusun terlebih dahulu diikuti oleh yang rendah persamaannya.

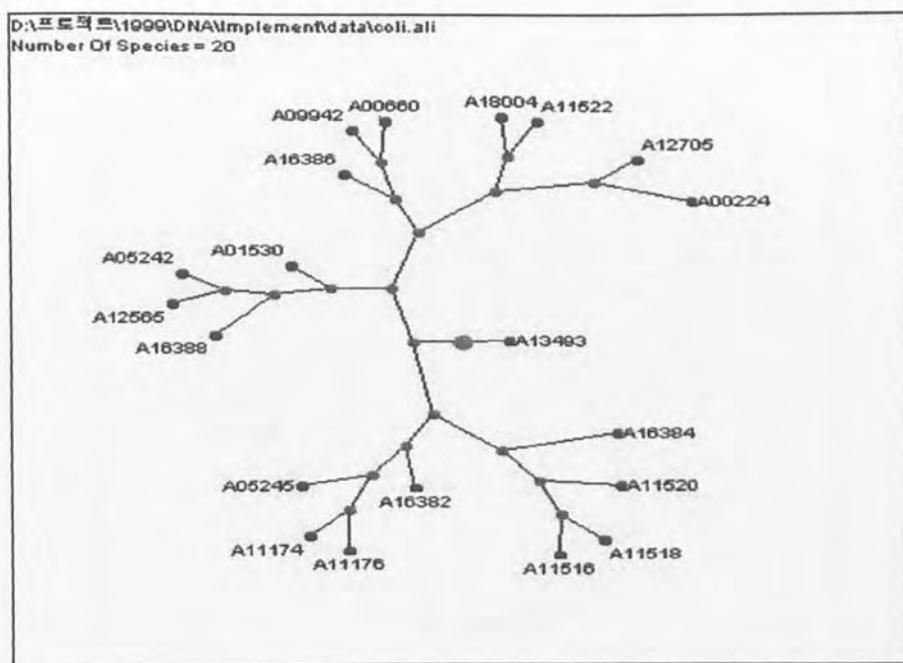
1.2.3 POKOK FILOGENETIK

Pokok filogenetik merupakan diagram yang mempersebahkan hubungan evolusi yang dijangka antara satu organisma dengan organisma yang lain. (<http://www.bscls.org/onco/glossary.htm>). Setelah penyusunan daripada Clustalw, pokok filogenetik dibina berdasarkan maklumat yang diperolehi untuk mengenalpasti dan mengklasifikasikan famili retrovirus endogenous manusia dalam genom manusia dengan virus dari genera yang lain. Terdapat banyak jenis pokok filogenetik iaitu Rajah 1.2 pokok bebas, Rajah 1.3 pokok radial, Rajah 1.4 pokok berakar, Rajah 1.5 segiempat tepat kladogram, Rajah 1.6 filogram, dan Rajah 1.7 kladogram serong (sumber: <http://pearl.cs.pusan.ac.kr/phylodraw/#test>)

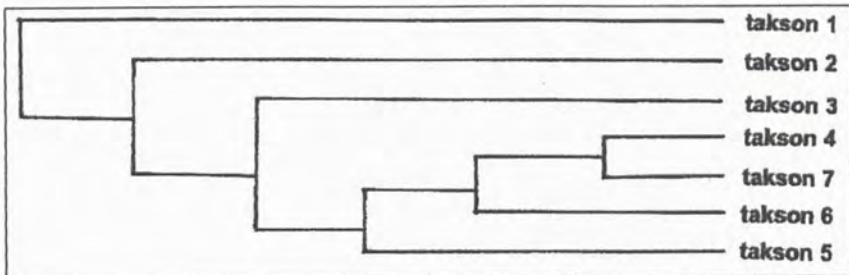




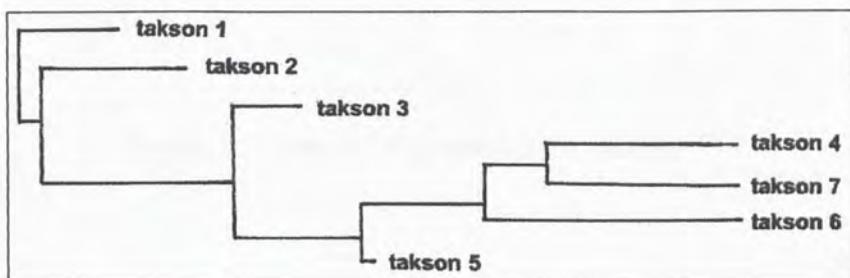
Rajah 1.2 Pokok Filogenetik Jenis Bebas



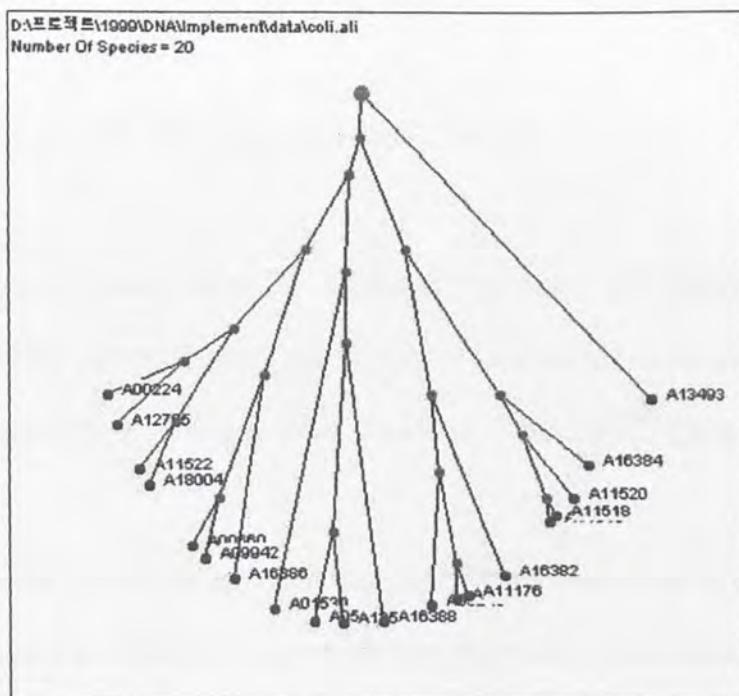
Rajah 1.3 Pokok Filogenetik Jenis Radial



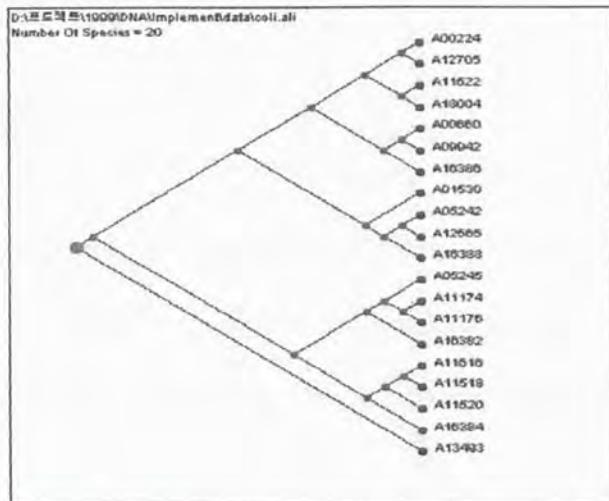
Rajah 1.4 Pokok Filogenetik Jenis Kladogram



Rajah 1.5 Pokok Filogenetik Jenis Filogram



Rajah 1.6 Pokok Filogenetik Berakar



Rajah 1.7 Pokok Filogenetik Jenis Serong

Kepanjangan dahan filogram menunjukkan aman perubahan (proses evolusi) yang telah berlaku. Kepanjangan dahan kladoram adalah sama panjang dan tidak dapat melihat jumlah masa evolusi yang digunakan untuk pembahagian taksa.

1.3 PENGENALAN RETROVIRUS ENDOGENOUS

Retrovirus endogenous manusia adalah dikatakan berhubung kait dengan penyakit barah (Lower, 1999) dan penyakit sklerosis ganda. Ciri yang membolehkan retrovirus dikenalpasti daripada virus-virus lain ialah kehadiran enzim transkriptase berbalik.

Nama famili retrovirus endogenous ialah *Retroviridae* dan ia boleh dikelaskan kepada tujuh genera berdasarkan nukleotida dan organisasi genomnya. Genera-genera terdiri daripada genus Alfaretrovirus, Betaretrovirus, Deltaretrovirus, Epsilonretrovirus, Gammaretrovirus, Lentivirus dan Spumavirus.

Secara umumnya, retrovirus boleh dibahagi kepada 4 kumpulan mengikut morfologinya iaitu partikel jenis A, B, C dan D. Partikel jenis A tidak bersampul dan menunjukkan bentuk intrasel. Jenis B pula bersampul, partikel ekstrasel dengan hujung sampul yang menonjol dan terasnya berasentrik. Partikel jenis C juga sama dengan jenis B iaitu partikel ekstrasel tetapi terasnya di tengah-tengah dan hujungnya terdedah. Akhirnya partikel jenis D adalah lebih besar iaitu bersaiz 120nm dan mempunyai sampul yang kurang menonjol.

Selain itu, retrovirus endogenous boleh dikelaskan kepada tiga kelas iaitu kelas I, II dan III (Nelson *et al.* 2003), seperti yang ditunjukkan oleh Jadual 1.1.

Jadual 1.1 Famili-famili Retrovirus Endogenous yang Tergolong dalam Kelas I, II dan III (Nelson *et al.*, 2003)

Kelas	Contoh famili
Kelas I	HERV-H
	HERV-F
	HERV-W
	HERV-P
	HERV-R
	HERV-T
	HERV-IP
	HERV-I
	HERV-E

Kelas II	HERV-K
Kelas III	HERV-L HERV-S

Retrovirus endogenous mengandungi motif terpelihara pada domain I hingga IV gen transkriptase berbalik. Motif terpelihara adalah jujukan bes DNA bebas atau jujukan asid amino dalam protein yang kekal tidak berubah setelah evolusi berlaku (Attwood dan Parry-Smith, 1999). Rajah 1.8 menunjukkan motif terpelihara pada domain I hingga domain IV gen transkriptase berbalik dalam retrovirus endogenous manusia.

Domain I	Domain II	Domain III	Domain IV	Doamin V
WNTP	QDLR	KDAFF	LPQG	YV/YM/YIDD

Rajah 1.8 Motif terpelihara pada domain I hingga V gen transkriptase berbalik dalam retrovirus endogenous manusia

1.4 PANGKALAN DATA BIOLOGI

National Center for Biotechnology Information (NCBI) ataupun genbank telah menyimpan banyak jujukan protein dan genom yang terdiri daripada jujukan nukleotida dan asid amino. Jujukan asid amino dibentukkan oleh banyak abjad. Setiap abjad



mewakili asid amino yang berlainan, contohnya huruf A mewakili Alanine, C mewakili Cysteine, dan sebagainya. Jadual 1.2 merupakan senarai asid amino dengan nama dan simbolnya. Dalam tempoh evolusi, jujukan-jujukan protein yang berasal dari famili yang sama yang mengalami penggantian asid-asid amino mempunyai ciri-ciri fisiokimia yang sama. Sesetengah daripada fisiokimia asid-asid amino ditunjukkan pada Jadual 1.3.

Jadual 1.2 Senarai Asid Amino dengan Nama Dan Simbol

Sumber: (http://en.wikipedia.org/wiki/List_of_standard_amino_acids)

IUB	Singkatan	Nama penuh
A	ala	Alanin
C	cys	Sistin
D	asp	Asid aspartik
E	glu	Asid glutamik
F	phe	Fenilalanin
G	gly	Glisin
H	his	Histidin
I	ileu	Isoleusin
K	lys	Lisin
L	leu	Leusin
M	met	Metionin
N	asn	Asparagin
P	pro	Prolin



Q	gln	Glutamin
R	arg	Arginin
S	ser	Serin
T	thr	Threonin
V	val	Valin
W	trp	Triptofan
Y	tyr	Tirosin
U		Selenocysteine
-		Kodon perhentian

Jadual 1.3 Asid-asid Amino Yang Ditulis Dalam Kod IUB Dengan Sifat Fisiokimia

(Sumber: <http://wbiomed.curtin.edu.au/teach/biochem/tutorials/AAs/AA.html>)

Sifat Fisiokimia	Asid-asid Amino
Hidrofobik	A, G, P, I, L, V, C, M, W, F
Polar	S, T, N, Q, Y
Aromatik	W, F, Y, H
Basik	H, K, R
Asidik	D, E

Boleh diperhatikan bahawa beberapa asid amino mempunyai dua sifat. Contohnya, 'Y' mempunyai dua sifat polar dan aromatik.



RUJUKAN

- Alexander, B. & Arnaud, C. 2001. Modeling Biological Systems in Hybrid Concurrent Constraint Programming. *Proceedings of the 18th International Conference on Logic Programming*, Denmark.
- Attwood, T. K. & Parry Smith, D.J., 1999. *Introduction to Bioinformatics*, Prentice Hall, United Kingdom.
- Collin , F. S., Green, E. D., Guttmacher, A. E & Guyer, M.S., 2003. Avision for The Future of Genomics Research. *Nature*, **422**:835-847.
- Griffiths, J. D., 2001. Endogenous Retrovirus in the Human Genome Sequence. *Genome Biology*, **2**:1017.1-1017.5.
- Henikoff, S. & Henikoff, J. G., 1992. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci.*, Seattle, **89**:10915-10919.
- Ilya N. S. & Philip E. B., Helge, W., 1998. Macromolecular Structure Databases: Past Progress and Future Challenges. *Acta Cryst. International Union of Crystallography*, Great Britain, **D54**:1085-1094.
- Lower, R., 1999. The Pathogenic Potential of Endogenous Retroviruses : Facts and Fantasies *Trends Microbiol*, **7**(9):350-356.



Marolia, M., Khoja, R., Acharya, T. & Chakrabarti, C., 2005. A Coprocessor Architecture for Fast Protein Structure Prediction, *Pattern Recognition*, 39(12):413- 418.

Michael A. L. & Andy M. T., 2004. Modelling biological evolvability: Implicit Context and Variation Filtering in Enzyme Genetic Programming. *BioSystems*, 76(1):229-238.

Muir, A., Lever, A. & Moffett, A., 2004. Expression and Function of Human Endogenous Retrovirus In the Placenta, *Journal of General Virology*, 25:16-25.

Nelson, P. N., Carnegie, P. R., Martin, J., Ejtehadi, H.D., Hooley, P., Roden, D., RowlandJones, S.P., Warren, P., Astley, J., & Murray, P. G., 2003. Demystified: Human endogenous retrovirus Molecular Pathology, 56:11-18.

Peleg, M., Yeh, I., Altman, R. B., 2002. Modelling biological processes using workflow and Petri Net models, *Bioinformatics*, 18(6):825-837

Sandblad, B. & Meinzer, H. P., 1992. Modeling and Simulation of Complex Control Structures in Cell Biology. *Methods of Information in Medicine*, 31:36-43.

Selvin George, David Evans & Steven Marchette, 2002. A Biologically Inspired Programming Model for Self -Healing System. *Proceedings Of The First Workshop On Self-Healing System*, 102-104.



Tristem, M. 2000. Identification and Characterization of Novel Human Endogenous Retrovirus Families by Phylogenetic Screening of the Human Genome Mapping Project Database, *Journal of Virology*, 74(8):3715-3730.

