MODEL P-CALS BAGI MENYELESAIKAN PERMASALAHAN RALAT LATA PADA PEMBINAAN SEMULA RANGKAIAN PENGAWAL ATUR GEN

FARIDAH HANI BINTI MOHAMED SALLEH

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEHI
IJAZAH DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

26 November 2018

FARIDAH HANI BINTI
MOHAMED SALLEH
P62446

PENGHARGAAN

Pertama sekali, jutaan ucapan terima kasih diucapkan kepada penasihat saya Dr. Suhaila Zainudin, Prof.Madya Dr Firdaus Raih dan Dr Shereena Arif atas sokongan dan bimbingan yang diberikan sepanjang pengajian saya di Universiti Kebangsaan Malaysia. Mereka banyak memberi tunjuk ajar dan menyediakan persekitaran yang selesa untuk saya menjalankan kajian saya.

Saya juga ingin mengucapkan terima kasih kepada pihak Universiti Tenaga Nasional kerana memberi keizinan untuk saya bercuti belajar. Begitu juga dengan pihak Kementerian Pengajian Tinggi yang memberikan saya biasiswa MyBrain15 yang sudah tentunya merupakan antara perangsang yang kuat untuk saya menyambung pelajaran dan menyumbang kepada negara. Ucapan terima kasih juga ditujukan pada rakan sekerja di Universiti Tenaga Nasional, dan rakan sepengajian dan staf di Fakulti Teknologi dan Sains Maklumat, UKM.

Pengajian saya tidak akan berakhir tanpa sokongan berterusan daripada ibu-bapa dan seluruh ahli keluarga saya. Akhir sekali, kepada suami dan anak-anak, merekalah individu yang paling banyak mendorong saya bagi mencapai kejayaan ini.

ABSTRAK

Pembinaan semula RPAG (rangkaian pengawal atur gen) ditakrifkan sebagai proses mengenal pasti hubungan antara gen berdasarkan data eksperimen menggunakan analisis pengkomputeran. Pembinaan semula RPAG adalah suatu proses yang mencabar kerana melibatkan data yang dipengaruhi hingar dan saiz pemerhatian terhad. Ketidaktepatan ramalan pada motif lata merupakan sebab utama mengapa model pembinaan semula RPAG yang dicipta sebelum ini mengalami penurunan prestasi. Ralat lata ditakrifkan sebagai kesalahan ramalan pada motif lata, di mana hubungan tidak langsung disalahtafsirkan sebagai hubungan langsung. Walaupun kajian dalam bidang ini aktif dijalankan dari segi pelbagai aspek, perbincangan spesifik bagi menyelesaikan permasalahan berkaitan ralat lata adalah masih kurang. Malahan, eksperimen terdahulu tidak menjurus ke arah pembuktian kebolehupayaan sesuatu model itu dalam mengelak berlakunya ralat lata. Ralat lata berlaku kerana hubungan langsung yang wujud antara gen memberikan isyarat yang kuat sehingga mengaburi petanda lain yang mungkin wujud bagi menunjukkan hubungan sebenar gen. Kewujudan gen yang tidak berkaitan dalam pengiraan juga menyukarkan pengenalpastian hubungan gen sebenar. Model Hingar Gaussian (MHG), Pekali Kolerasi Pearson (PKP), Pearson & Gaussian (P&G) dan Regresi Linear Berbilang (RLB) telah diaplikasikan ke atas data sintetik. Penemuan utama eksperimen adalah jumlah positif palsu yang tinggi terhasil adalah disebabkan oleh ralat lata. P&G telah dikenal pasti sebagai model terbaik untuk digunakan ke atas data sintetik. Walaubagaimanapun, struktur data sebenar ekspresi gen yang berbeza daripada data sintetik menyebabkan aplikasi P&G ke atas set data sebenar tidak boleh dilakukan. Lalu, RLB dipilih sebagai model terbaik untuk ditambahbaik bagi menyelesaikan ralat lata. Oleh kerana bilangan pemerhatian data eksperimen nyata adalah jauh berkurangan, sebilangan peramal disingkirkan secara sistematik dengan mengekstrak sub-jaringan rawak daripada jaringan besar. Walaubagaimanapun, pengaplikasian sub-jaringan pada RLB adalah diragukan kerana hubungan antara gen boleh menjadi tidak tepat. Kekurangan yang didapati pada semua model yang diuji telah memberi motivasi kepada pengusulan P-CALS yang merupakan model berasakan KDTS dan terdiri daripada dua kaedah bermula dengan mengenalpasti gen peramal bererti oleh APK. Kemudian, gen peramal tersebut diumpukkan pada KDTS bersama-sama gen peramal yang dikenalpasti oleh UKMA dalam model KDTS. Prestasi P-CALS diuji dengan RPAG berskala genom *E. coli*, *S. cerevisiae* dan set data siliko. P-CALS mencatatkan keputusan yang baik dengan kesemua sub-jaringan mencatatkan nilai AUROC melebihi 0.5, dengan 82.31% pembolehubah bererti dan hubungan gen dikenalpasti pada jaringan paling kompleks. Eksperimen pada set data sebenar seterusnya diperluaskan untuk mengkaji keefisyenan P-CALS dalam menangani ralat lata dengan menggunakan prosedur eksperimen yang turut diusulkan dalam kajian ini.

P-CALS MODEL FOR SOLVING CASCADE ERRORS IN GENE REGULATORY NETWORKS RECONSTRUCTION

ABSTRACT

Gene regulatory network (GRN) reconstruction is the process of identifying regulatory gene interactions from experimental data through computational analysis. Reconstructing GRN is a challenge because the reconstruction has to be made from noise-affected data with a very limited number of observations. One of the main reasons for the reduced performance of previous GRN reconstruction model had been inaccurate prediction of cascade motifs. Cascade error is defined as the wrong prediction of cascade motifs, where an indirect interaction misinterpreted as a direct interaction. Despite the active research on various GRN reconstruction model, the discussion on specific model to solve problems related to cascade errors is still lacking. In fact, the experiments conducted by the past studies were not specifically geared towards proving the ability of GRN prediction methods in avoiding the occurrences of cascade errors. Cascade error occurs because the signal of direct interaction between genes is too strong till it affects the other signals that have potential to discover the true gene interactions. The presence of irrelevant genes in calculation increase the difficulty of identifying the true gene interactions. Gaussian Noise Model, Pearson Correlation Coefficient (PCC), Pearson & Gaussian (P&G), Mutual Information (MI) and Multiple Linear Regression (MLR) were implemented to infer GRN from the synthetic datasets. The most notable finding from these experiments is that, cascade error causes the high false positives. P&G has been identified as the best method to be implemented on synthetic data. Despite the good performance of P&G, the different structure of the gene expression data makes the implementation of that model to real datasets is not possible. Thus, MLR is chosen to be the best possible model to be further improved to solve cascade errors. Since the number of observations of the real experiment datasets was far less than the number of predictors, some predictors were eliminated systematically by extracting the random sub-networks from global interaction networks. However, performing MLR on sub-networks is dubious as the interrelationships between the regressors may be ignored. The limitations that we had discovered in our previous works have motivated us to propose P-CALS that is based on PLS(Partial Least Squares) and consists of two steps, starting from identifying the significant predictors variables using PCA (Principal Component Analysis). Then, the predictors are included into PLS together with the predictors identified from Marten Uncertainty Test (MUT) in PLS. The performance of P-CALS is assessed to the genome-scale GRN of *E. coli, S. cerevisiae* and an in-silico datasets. P-CALS achieved good results as all of the sub-networks from diverse datasets achieved AUROC values above 0.5, with 82.31% of the significant variables and gene relationships were discovered at the most complex network. The experiment for the real datasets was extended to assess the effectiveness of P-CALS in dealing with cascade error by using a novel experimental procedure that had been proposed in this work.

KANDUNGAN

PENGAKUAN		ii
PENGHARGAAN		iii
ABSTRAK		iv
ABSTRACT		v
KANDUNGAN		vi
SENARAI JADUAL		x
SENARAI ILUSTRASI		xiv
SENARAI RINGKASAN		vi
BAB I	**PENDAHULUAN**	
1.1	Latar Belakang Kajian	1
1.2	Pernyataan Masalah	2
1.3	Objektif Kajian	6
1.4	Skop Kajian	6
1.5	Sumbangan Utama Kajian	7
1.6	Metodologi Kajian	8
1.7	Organisasi Tesis	9
1.8	Rumusan	10
BAB II	**PEMBINAAN SEMULA RANGKAIAN PENGAWAL ATUR GEN**	
2.1	Pendahuluan	11
2.2	Rangkaian Pengawal Atur Gen (RPAG)	11
	2.2.1 Pengenalan	11
	2.2.2 Gambaran Keseluruhan Rangkaian Pengawal Atur Gen	12
	2.2.3 Ralat Lata	19

	2.2.4 Kajian Dalam Bidang Penyelesaian Permasalahan Berkaitan Motif Lata	22
2.3	Data Analitik dan Pengaplikasiannya Pada Domain Perkomputeran Biologi	25
	2.3.1 Pengurangan Data	34
	2.3.2 Pengurangan Dimensi	35
	2.3.3 Data Hingar	36
2.4	Kajian Pelbagai Bidang Pembinaan Semula Rangkaian Pengawal Atur Gen	38
	2.3.1 Kajian Model Berasaskan Analisis Regresi	38
	2.3.2 Kajian Model Berasaskan Teori Maklumat	56
	2.3.3 Kajian Model Berasaskan Tapisan *(Filter)*	57
	2.3.4 Kajian Model Berasaskan Kebarangkalian Dan Statistik	58
	2.3.5 Kajian Model Berasaskan Algoritma Berinspirasikan Alam	58
	2.3.6 Kajian Model Berasaskan Kolerasi Dan Kebergantungan	62
	2.3.7 Kajian Model Berasaskan Pembelajaran Mesin	63
	2.3.8 Kajian Model Berasaskan Meta Algoritma	70
	2.3.9 Analisis Kritis	71
2.4	Rumusan	75
BAB III	**METODOLOGI**	
3.1	Pendahuluan	76
3.2	Keseluruhan Proses Penyelidikan	76
	3.2.1 Langkah 1	76
	3.2.2 Langkah 2	83
	3.2.3 Langkah 3	83
	3.2.4 Langkah 4	89
3.4	Rumusan	93
BAB IV	**KAJIAN PENENTUAN MODEL RPAG**	
4.1	Pendahuluan	95
4.2	Model Hingar Gaussian (MHG)	95

	4.2.1	Rasional Pemilihan Anggaran Serentak Berbanding Anggaran Berfasa ke Atas Sisihan Piawai dan Nilai Jenis Liar	96
	4.2.2	Keputusan Eksperimen Menggunakan Model Hingar Gaussian	98
4.3		Pekali Kolerasi Pearson (PKP)	100
4.4		Pearson & Gaussian (P&G)	101
	4.4.1	Justifikasi Penggabungan PKP dan Gaussian	101
	4.4.2	Keputusan Eksperimen Menggunakan Integrasi P & G	105
4.5		Gabungan Keputusan Eksperimen PKP, Model Gaussian dan P&G	106
4.6		Regresi Linear Berbilang (RLB)	110
	4.6.1	Mengenalpasti Motif Lata	111
	4.6.2	Keputusan Eksperimen Menggunakan RLB	115
	4.6.3	Pengekstrakan Sub-Jaringan	115
	4.6.4	Keputusan Eksperimen Menggunakan RLB Dengan Pengaplikasian pada Set Data M3D	117
	4.6.5	Perbandingan Antara RLB Dengan Model Pembinaan Semula RPAG yang Lain	120
4.7		Perbincangan	121
4.8		Rumusan	125
BAB V		**P-CALS SEBAGAI MODEL BERASASKAN REGRESI BAGI MENYELESAIKAN PERMASALAHAN RALAT LATA**	
5.1		Pendahuluan	126
5.2		Kaedah Pembinaan Semula RPAG Menggunakan Teknik Berasaskan Analisis Regresi	126
5.3		Ujian Diagnostik Multikolinearan	127
5.4		Analisis Prinsip Komponen (APK)	129
	5.4.1	Mengenalpasti Pembolehubah Tidak Bererti Bagi Mengurangkan Dimensi Data	130
	5.4.2	Kaedah Berasaskan Kuadrasi	131
	5.4.3	Korelasi Gen Berbeban Tinggi Selari Pada PK Yang Sama	134

	5.4.4 Keputusan Eksperimen Menggunakan APK	137
5.5	Kuasa Dua Terkecil Separa (KDTS)	138
	5.5.1 Keputusan Eksperimen Menggunakan KDTS	141
5.6	P-CALS	144
	5.6.1 Keputusan Eksperimen Menggunakan P-CALS	148
5.7	Pengujian Tahap Multikolinearan	154
5.8	Pengujian Kesan Hingar	154
5.9	Perbincangan	160
5.10	Rumusan	163
BAB VI	**PERBINCANGAN AKHIR**	
6.1	Pendahuluan	164
6.2	Hasil Kajian	164
6.3	Arah Tuju Masa Hadapan	168
6.4	Rumusan	170
RUJUKAN		173
LAMPIRAN		
Lampiran A	AUROC dan AUPR dihasilkan oleh GeneNetWeaver untuk set data DREAM3 bersaiz 10.	187

SENARAI JADUAL

No Jadual		Halaman
Jadual 2.1	Contoh hubungan tidak langsung dalam sub-jaringan *E.coli2*	21
Jadual 2.2	Rumusan kajian utama dalam bidang hubungan tidak langsung atau motif lata	24
Jadual 2.3	Perisian data analitik	28
Jadual 2.4	Perisian data analitik (sambungan)	29
Jadual 2.5	Perisian data analitik tambahan	30
Jadual 2.6	Perisian data analitik tambahan (sambungan)	31
Jadual 2.7	Pilihan model berasaskan regresi (MathWorks 2015)	41
Jadual 2.8	Rumusan kajian terpilih berkaitan model berasaskan regresi (bahagian 1)	53
Jadual 2.9	Rumusan kajian terpilih berkaitan model berasaskan regresi (bahagian 2)	54
Jadual 2.10	Rumusan kajian terpilih berkaitan model berasaskan regresi (bahagian 3)	55
Jadual 2.11	Rumusan kajian terpilih berkaitan model berinspirasikan alam	61
Jadual 2.12	Rumusan kajian terpilih berasaskan pembelajaran mesin (machine learning)	69
Jadual 2.13	Jadual data untuk pemodelan ramalan (Rapidminer 2018)	73
Jadual 3.1	Contoh data pemotongan homozigot (Marbach, Schaffter et al. 2009	79
Jadual 3.2	Sebahagian daripada data jujukan masa pada sela masa yang berlainan. Dipaparkan data dengan 11 titik masa	80
Jadual 3.3	Sebahagian data M3D	81
Jadual 3.4	Ciri-ciri set data yang digunakan dalam eksperimen	82
Jadual 3.5	Sebahagian data hasil eksperimen kajian awal yang menggunakan Pearson ke atas set data sintetik	83

Jadual 3.6	Hubungan antara gen yang ditemui oleh model yang diusulkan dalam kajian ini untuk *E.coli1* (Saiz 10) dari DREAM3	93
Jadual 4.1	AUROC dan AUPR yang dicatatkan model Gaussian apabila diaplikasikan pada data homozigot dan heterozigot dari set data DREAM3 dan DREAM4	99
Jadual 4.2	Ramalan menggunakan data pemotongan homozigot dan heterozigot daripada set data DREAM 3 menggunakan model Gaussian	100
Jadual 4.3	Algoritma P&G	102
Jadual 4.4	Pengkategorian Dancey and Reidy's (2007)	102
Jadual 4.5	Pengkategorian Evans (1996)	103
Jadual 4.6	Nilai ramalan AUROC dan AUPR menggunakan P&G, dilakukan ke atas data pemotongan homozigot rangkaian saiz 10 dan 50 dari DREAM3	105
Jadual 4.7	Penilaian terperinci mengenai set data DREAM3 saiz 10 menggunakan P & G	105
Jadual 4.8	Model dengan prestasi tertinggi dalam AUROC dan AUPR	109
Jadual 4.9	Algoritma pengenalpastian ralat lata	112
Jadual 4.10	Mengenalpasti motif lata dan ralat lata	113
Jadual 4.11	Prestasi Gaussian dan RLB dilakukan ke atas data pemotongan dan data jujukan masa	115
Jadual 4.12	Proses RLB yang diaplikasikan pada set data ekspresi gen	117
Jadual 4.13	Keputusan daripada eksperimen yang menggunakan set data M3D dengan motif lata yang telah dikeluarkan	118
Jadual 4.14	Keputusan eksperimen yang menilai prestasi RLB apabila diaplikasikan pada set data dengan motif lata.	119
Jadual 4.15	Ciri-ciri set data yang diuji dalam eksperimen dan nilai AUROC yang diperolehi	120

Jadual 4.16	AUROC kaedah yang diaplikasikan ke atas set data M3D *E.coli*	121
Jadual 5.1	Indeks Syarat (IS) dan paras kekolinearan (Belsley et al. 2005)	128
Jadual 5.2	S dan PPV empat gen dari Set 3 sebagai contoh data yang dihasilkan dari ujian diagnostik	128
Jadual 5.3	Keputusan eksperimen menggunakan APK (dengan kaedah berbeza)	137
Jadual 5.4	Pembolehubah bererti yang dikenalpasti oleh APK	137
Jadual 5.5	Pengujian motif lata APK	138
Jadual 5.6	Keputusan eksperimen menggunakan KDTS	142
Jadual 5.7	Pengenalpastian pembolehubah bererti oleh KDTS-UKMA	143
Jadual 5.8	Pengujian motif lata KDTS-UKMA	143
Jadual 5.9	RMSE, R-Square dan Varians Tertakrif (VT) eksperimen KDTS	143
Jadual 5.10	Keputusan eksperimen menggunakan APK, KDTS dan P-CALS	149
Jadual 5.11	KDTS-UKMA dan APK	149
Jadual 5.12	Pembolehubah bererti yang dikenalpasti oleh APK, KDTS-UKMA dan P-CALS	150
Jadual 5.13	Pengujian motif lata	151
Jadual 5.14	Ramalan betul (PB) yang dilakukan oleh APK, KDTS-UKMA dan P-CALS	152
Jadual 5.15	Keputusan eksperimen menggunakan P-CALS dan perbandingan dengan kaedah lain	153
Jadual 5.16	Pengujian tahap multikolinearan P-CALS berdasarkan nilai IS	154
Jadual 5.17	Pengujian tahap multikolinearan P-CALS berdasarkan nilai PPV	154

Jadual 5.18 Pengujian tahap hingar pada KDTS dan KDTS-UKMA menggunakan varians tertakrif. 157

SENARAI ILUSTRASI

No Rajah		Halaman
Rajah 1.1	Metodologi kajian	8
Rajah 2.1	Aliran daripada DNA kepada protein (Alon 2007)	13
Rajah 2.2	Elemen rangkaian transkripsi (Sumber: (Alon 2007))	14
Rajah 2.3	Gen pada RPAG dan istilah berkaitan	16
Rajah 2.4	Tingkah laku 'rangkaian pengawal atur gen', sumber: (Endy & Brent 2001)	16
Rajah 2.5	RPAG e.coli yang diperolehi daripada pengkalan data RegulonDB (Allen et al. 2012). Setiap bulatan merah ialah gen dan garisan biru yang menghubungkan gen ialah hubungan antara gen.	17
Rajah 2.6	RPAG e.coli	18
Rajah 2.7	Multikolineariti yang menunjukkan kaitan antara peramal yang tinggi	19
Rajah 2.8	Jenis motif RPAG (Marbach et al. 2010), tidak termasuk motif arah sendiri. Bulatan yang dinomborkan ialah gen.	20
Rajah 2.9	Ralat lata. Hubungan antara gen dalam rajah ini adalah berdasarkan data pada Jadual 2.1	22
Rajah 2.10	Pengaplikasian data analitik pada perkomputeran biologi	32
Rajah 2.11	Perkomputeran biologi	33
Rajah 2.12	Peringkat pengurangan data (Sumber: IBM, 2018)	34
Rajah 2.13	Teknik pengurangan data (Sumber: Packt 2014)	35
Rajah 2.14	Data pelbagai dimensi (Sumber: Peter Gleesen, 2017)	36
Rajah 2.15	Data hingar (Source: (Schiffermuller & Jungnickel 2006))	37
Rajah 2.16	RLB dalam konteks RPAG	39

Rajah 2.17	Regresi Linear	40
Rajah 2.18	Sampel pada ruang pelbagai dimensi(CAMO 2003)	47
Rajah 2.19	PK 1 dan 2 pada ruang multidimensi (CAMO 2003)	49
Rajah 2.20	Proses Pengoptimuman Kerumunan Zarah Berinspirasikan Serigala Kelabu	60
Rajah 2.21	Konsep asas RNB (Bullinaria 2013)	64
Rajah 2.22	Lembaran cheat (Cheat Sheet) algoritma pembelajaran mesin (Sumber: (Hui Li 2017).	66
Rajah 2.23	Garis masa kemunculan Pembelajaran Mendalam dan algoritma pembelajaran mesin yang biasa digunakan. Sumber: (Cao, Liu et al. 2018)	67
Rajah 2.24	Perbandingan AUROC	75
Rajah 3.1	Metodologi kajian secara umum	77
Rajah 3.2	Perisian The Unscambler X	82
Rajah 3.3	Langkah kajian pengusulan model yang mengambilkira peramal dan pembolehubah sambutan secara serentak	84
Rajah 3.4	Langkah kajian pengusulan model berasaskan regresi multivariat	85
Rajah 3.5	Langkah kajian mengenalpasti hubungan antara gen	88
Rajah 3.6	Langkah kajian penggabungan model berasaskan unjuran dengan regresi multivariat	90
Rajah 3.7	Skrin GeneNetWeaver	92
Rajah 3.8	Metodologi kajian dan kaitannya dengan objektif kajian	94
Rajah 4.1	Perbandingan prestasi antara anggaran berfasa (kaedah 1) dan anggaran serentak (kaedah 2) sisihan piawai dan nilai jenis liar	98
Rajah 4.2	Proses penghasilan Pearson-Gaussian	104
Rajah 4.3	Prestasi model diuji pada data pemotongan homozigot dari set data DREAM3 saiz 10	106

Rajah 4.4	Prestasi model diuji pada data pemotongan homozigot dari DREAM3 saiz 50	107
Rajah 4.5	Prestasi model diuji pada data pemotongan homozigot dari DREAM4 saiz 10	107
Rajah 4.6	Motif lata dari Jadual B (Jadual 4.10(b)) .Garis putus-putus menunjukkan ralat lata	114
Rajah 4.7	Senarai hubungan gen terarah dan motif lata. Garis putus-putus mewakili ralat lata	114
Rajah 4.8	Keseluruhan perjalanan eksperimen awal bagi pembinaan model RPAG.	124
Rajah 5.1	Hubungan FT dengan gen sasaran. Garis putus-putus mewakili FT	127
Rajah 5.2	Gen yang dibulatkan adalah berkolerasi positif dengan FT yang ditunjukkan dengan anak panah	132
Rajah 5.3	Algoritma pembinaan semula RPAG menggunakan APK dengan kaedah pemilihan pembolehubah berasaskan kuadrasi	133
Rajah 5.4	Bebanan kolerasi menunjukkan pembolehubah berpembebanan tinggi selari pada PK yang sama dipilih dan pembolehubah menghampiri 0 direndahkan pemberatnya	135
Rajah 5.5	Proses APK	136
Rajah 5.6	Skor, pemberat X dan Y, varians tertaktif dan nilai R-square yang dicatatkan oleh KDTS yang dilaksanakan ke atas Jaringan 4	140
Rajah 5.7	Proses keseluruhan KDTS	141
Rajah 5.8	Model proses P-CALS	146
Rajah 5.9	Algoritma P-CALS	147
Rajah 5.10	Pengujian kesan hingar pada APK menggunakan nilai residual. PC 0 ialah keadaan hingar sebelum APK diaplikasikan.	155
Rajah 5.11	Pengujian kesan hingar pada APK menggunakan varians tertakrif.	156

Rajah 5.12	Pembolehubah di bahagian pusat atau tengah disingkirkan	159
Rajah 5.13	Senarai data yand diimport berada dalam petak garis merah	159
Rajah 5.14	Plot yang digunakan untuk membantu analisa data dalam petak garis merah	160
Rajah 6.1	Keseluruhan komponen yang terlibat dalam kajian ini termasuk pengetahuan yang diperolehi	172

SENARAI RINGKASAN

APK	Analisis Prinsip Komponen *(Principal Component Analysis)*
DREAM	Dialogue for Reverse Engineering Assessments and Methods
FT	Faktor Transkripsi
GNW	GeneNetWeaver
IS	Indeks Syarat
KDTS	Kuasa Dua Terkecil Separa *(Partial Least Squares)*
KPB	Kadar Positif Benar
KPB	Kadar Positif Benar
KPP	Kadar Positif Palsu
KPP	Kadar Positif Palsu
MHG	Model Hingar Gaussian *(Gaussian Noise Model)*
NB	Negatif Benar
NP	Negatif Palsu
P&G	Pearson & Gaussian
PB	Positif Benar
P-CALS	**P**rincipal **C**omponent **A**nalysis & Partial **L**east **S**quares

PE	Piawai Emas *(Gold standard)*
PK	Prinsip Komponen
PKR	Prinsip Komponen Regresi *(Principal Component Regression)*
PKS	Pengoptimuman Kerumunan Semut
PKZ	Pengoptimuman Kerumunan Zarah
PP	Positif Palsu
PPV	Perkadaran Penguraian Varians
pRNA	polimerase RNA
RLB	Regresi Linear Berbilang *(Multiple Linear Regression)*
RNB	Rangkaian Neural Berulang
RPAG	Rangkaian Pengawalatur Gen *(Gene regulatory networks)*
TM	Teori Maklumat
TMB	Teori Maklumat Bersyarat
UKB	Ujian Kolineariti Belsley
UKMA	Ujian Ketidakpastian Marten

BAB I

PENDAHULUAN

1.1 LATARBELAKANG KAJIAN

Kajian berkaitan pembinaan semula RPAG telah banyak menyumbang kepada kejayaan dalam mencari sasaran dadah (*drug target*) untuk rawatan penyakit manusia, termasuk kanser (Ao & Palade 2011) (Mitra et al. 2011; Ahmad et al. 2012) (Wang & Michoel 2017). Pembinaan semula rangkaian pengawalatur gen (RPAG) daripada profil ekspresi gen telah membolehkan pelbagai penemuan penting merentasi pelbagai domain seperti biologi molekul, biokimia, biokejuruteraan dan farmaseutikal. Atas faktor RPAG yang bersifat kompleks, maka kaedah berkomputer digunakan bagi mengenalpasti interaksi antara faktor transkripsi (FT) dan gen sasaran. Walaubagaimanapun, lima permasalahan penting berkaitan kaedah berkomputer tidak dikaji dengan mendalam.

Pertama sekali, kebanyakan ralat positif palsu (PP) yang dilakukan oleh kajian sebelum ini adalah disebabkan oleh ralat lata (Marbach et al. 2012). Ketepatan model pembinaan semula RPAG akan meningkat dengan ketara jika ralat lata dapat dielakkan. Selain dari itu, prestasi kebanyakan model pembinaan semula RPAG direncatkan oleh permasalahan multikolinearan (Belsley et al. 2005). Multikolinearan merupakan masalah kritikal kerana ianya boleh meningkatkan varians anggaran koefisien *(coefficient estimates)*, lalu mengakibatkan anggaran menjadi sensitif terhadap perubahan kecil yang berlaku pada kaedah pengiraan (Belsley et al. 2005). Hasilnya, anggaran koefisien menjadi tidak stabil dan susah untuk ditafsir. Seterusnya, kaedah pengurangan dimensi diperlukan kerana dimensi yang berkurangan menjadikan komputeran kurang kompleks, lalu menjadikan penggunaan algoritma yang kompleks seperti NIPALS *(NonLinear Iterative Partial Least Squares)* boleh digunakan pada

model yang diusulkan dalam kajian ini, iaitu P-CALS *(**P**rincipal **C**omponent **A**nalysis & Partial **L**east **S**quares).*

Bagi menyelesaikan permasalahan yang dinyatakan sebelum ini, kajian ini mensasarkan untuk 1) mengusulkan model yang mengambilkira kedua-dua peramal dan pembolehubah sambutan secara serentak, 2) mengusulkan model berasaskan regresi multivariat yang berkeupayaan mengelak hubungan tidak langsung (A→ B → C) ditakrif sebagai sebagai hubungan langsung (A → C) (ralat lata) pada data dengan bilangan pemerhatian terhad, 3) mengusulkan model berasaskan unjuran untuk mengenalpasti pembolehubah bererti dengan menyingkirkan kesan hingar dan 4) mengintegrasikan model berasaskan unjuran dengan regresi multivariat bagi meningkatkan tahap keefisyenan model pembinaan semula RPAG.

1.2 PERNYATAAN MASALAH

RPAG terdiri daripada jutaan peramal dan pembolehubah sambutan yang mempunyai hubungan antara satu sama lain, kompleks dan bersifat tidak menentu *(stochastic) (Alon 2007; Das et al. 2010)*. Perubahan keadaan pada suatu gen boleh memberi tindakbalas kepada gen lain. Kewujudan kolerasi yang tinggi di antara pembolehubah tak bersandar pada model regresi dikenali sebagai multikolinearan. Isu multikolinearan perlu diselesaikan kerana apabila kolineariti berlaku pada peramal, koefisyen-b (*b-coefficients*) tidak lagi boleh dipercayai dan model menjadi tidak stabil dan tidak diyakini (Freund & Wilson 1998). Penentuan hubungan antara peramal dengan pembolehubah sambutan tidak boleh dilakukan satu persatu. Maka, oleh kerana kesemua gen dalam RPAG berhubung antara satu sama lain, **suatu model yang mengambilkira kedua-dua peramal dan pembolehubah sambutan secara serentak** diusulkan dalam kajian ini. Antara beberapa kategori model seperti statistik, pembelajaran mesin, teori graf, didapati model berasaskan regresi multivariat adalah paling sesuai.

Salah satu sebab utama model pembinaan semula RPAG sebelum ini tidak mencatatkan prestasi ramalan yang baik adalah kerana kegagalan meramal motif lata dengan tepat (Prill et al. 2010). Walaupun pelbagai model pembinaan semula RPAG

telah diterbitkan oleh pelbagai jurnal sebelum ini (Zuo et al. 2017), (Singh & Vidyasagar 2016), (Guo et al. 2016), (Chan et al. 2016), (Chan et al. 2012), (Geeven et al. 2012), (Küffner et al. 2012), (Gregoretti et al. 2010), perbincangan mengenai kaedah bagi menyelesaikan masalah berkaitan ralat lata adalah masih berkurangan. Istilah motif lata *(cascade motif)* dan ralat lata *(cascade error)* digunakan dalam keseluruhan dokumen. Selain daripada istilah ralat lata, istilah lain yang turut digunakan bagi mewakili perkara yang sama dalam penulisan tesis ini ialah hubungan tidak langsung (*indirect effects)* (Küffner et al. 2012; Feizi et al. 2013). Eksperimen menggunakan Pearson , Gaussian, Regresi Linear Berbilang (RLB) dan Pearson & Gaussian (P&G) telahpun dijalankan sebelum ini oleh (Salleh et al. 2015) dan (Faridah Hani et al. 2013) yang menyimpulkan bahawa kebanyakan PP adalah diakibatkan oleh ralat lata. Ini turut disokong oleh (Pinna et al. 2010), (Yip et al. 2010), (Marbach et al. 2010), (Marbach et al. 2012) , (Küffner et al. 2012), (Zhang et al. 2013) dan (Feizi et al. 2013). Kajian berkaitan ralat lata walaupun ada dinyatakan dalam beberapa kajian, namun, kajian khusus bagi mengurangkan ralat lata masih diperlukan. Perisian GNW (GeneNetWeaver) yang dibangunkan oleh (Schaffter et al. 2011) telah memberikan impak positif yang besar dalam bidang sistem biologi, khasnya pembinaan semula RPAG. GNW melakukan analisis motif rangkaian termasuk ralat lata. Analisis motif rangkaian berkemampuan menunjukkan pada motif jenis manakah kesalahan ramalan banyak dilakukan oleh sesuatu model pembinaan semula RPAG. Kekurangan GNW ialah tidak berkeupayaan untuk menganalisis data eksperimen kompleks.

Model pembinaan semula RPAG menghadapi permasalahan data mempunyai bilangan pemerhatian yang kurang berbanding dengan bilangan gen, $n \leq p$ (n = pemerhatian dan p = pembolehubah /gen) (Chan, Zhang et al. 2015) (Shi et al. 2016).Senario seperti ini sering dihadapi oleh eksperimen bidang biologi atau perubatan kerana kos eksperimen yang melibatkan benda hidup seperti manusia atau haiwan adalah tinggi. Isu berkaitan saiz sampel yang sangat rendah berbanding dengan bilangan gen adalah suatu isu penting yang sangat perlu diberi perhatian (Zhang et al. 2010), (Xu et al. 2007) (Singh & Vidyasagar 2016) (Guo et al. 2017). Bagi mengatasi isu ini, (Salleh et al. 2017) menggunakan RLB dan menyingkirkan sebahagian peramal secara sistematik dengan mengekstrak sub-jaringan rawak daripada jaringan global menggunakan kaedah pengekstrakan sub-jaringan oleh (Marbach, Schaffter et al. 2009).

(Marbach 2009) pula mengusulkan kaedah menghasilkan jaringan siliko dengan mengekstrak modul daripada jaringan global (sumber jaringan). Walaubagaimanapun, ketepatan pembinaan semula RPAG menggunakan kaedah ini kurang diyakini kerana tidak menjamin hubungan antara peramal diambilkira secara keseluruhan. (Tenenhaus et al. 2010) menyatakan bahawa data dengan jumlah pemerhatian terhad telah menjadikan anggaran tidak dapat dilakukan dengan cara menentukan pemberat kepada keseluruhan set pengatur *(regulators)*. Jika semua set pengatur yang lengkap jumlahnya dalam sesebuah RPAG tidak dapat digunakan dalam pengiraan, suatu langkah perlu dilaksanakan bagi memikirkan kaedah terbaik bagi memilih hanya sebilangan set pengatur yang terlibat dalam pengiraan dan dalam masa yang sama tidak menjejaskan ramalan lengkap keseluruhan RPAG. **Maka, kajian ini mengusulkan Kuasa Dua Terkecil Separa (KDTS) yang merupakan suatu model berasaskan regresi multivariat yang berkeupayaan mengelak hubungan tidak langsung (A→ B → C) ditakrif sebagai sebagai hubungan langsung (A → C) (ralat lata) dalam keadaan data bilangan pemerhatian terhad.**

Pembinaan semula RPAG berhadapan beberapa cabaran tersendiri termasuk sel-sel hidup yang berdimensi tinggi, di mana puluhan ribu gen bertindak pada kombinasi temporal dan ruang berbeza (Das et al. 2010). Interaksi antara gen yang pelbagai, sama ada secara langsung atau tidak langsung, menjadikan jaringan RPAG sangat dinamik dan tidak linear. Jaringan yang dinamik dan bersaiz besar ini menyukarkan pembinaan semula tanpa menggunakan sebarang teknik pengkomputeran. Pertambahan bilangan dimensi yang mendadak penting untuk ditangani memandangkan pada era teknologi masa kini, bilangan data mengalami lonjakan tinggi kerana kewujudan pelbagai perkakasan eksperimen moden yang berterusan merekod data dan menyimpannya untuk analisis pada masa akan datang. Pengambilan data eksperimen oleh perkakasan ini terdedah kepada pelbagai jenis ralat dan kesan hingar yang mengganggu ketepatan ramalan (Barzel & Barabási 2013; Zhang et al. 2013). Kewujudan nod tersembunyi pada RPAG juga antara cabaran yang wujud (Barzel & Barabási 2013). Menggunakan kes mudah lata $i \rightarrow k \rightarrow j$, apabila nod perantaraan k tersembunyi, nod i dan j menjadi terasing antara satu sama lain. Kewujudan hubungan tidak langsung dalam RPAG yang tersembunyi mengganggu keseluruhan model RPAG. Berdasarkan semua permasalahan

ini**, APK yang berasaskan kaedah unjuran diusulkan untuk mengenalpasti pembolehubah bererti dengan menyingkirkan kesan hingar.**

Kekuatan utama APK adalah dari segi keupayaan mengenalpasti pembolehubah bererti dengan memfokus pada peramal dengan cara mendapatkan unjuran *(projection)* terbaik peramal. Namun, APK mempunyai kelemahan menentukan hubungan antara gen dengan tepat kerana pengiraannya tidak merangkumi kedua-dua pembolehubah sekaligus. KDTS yang diaplikasikan bersama Ujian Ketidakpastian Marten (UKMA) mempunyai kekuatan menentukan hubungan gen, menangani kolineariti dan pengenalpastian pembolehubah bererti. Namun, penentuan pembolehubah bererti KDTS tidak setanding APK. Maka, **model unjuran (APK) diusulkan untuk bergabung dengan regresi multivariat (KDTS) supaya kedua-duanya melengkapi antara satu sama lain bagi meningkatkan tahap keefisyenan model pembinaan semula RPAG** yang didapati menunjukkan prestasi terbaik pada Net4 yang merupakan jaringan paling kompleks.

(Singh & Vidyasagar 2016) adalah merupakan kajian paling hampir dengan P-CALS memandangkan kedua-duanya menggunakan kaedah berasaskan regresi. Kedua-dua P-CLAS dan bLARS mendapatkan gen pengawalselia paling berpotensi yang berhubung dengan gen sasaran melalui hasil gabungan beberapa proses sebelumnya. Perbezaan antara P-CALS dan bLARS (kaedah diusulkan oleh Singh & Vidyasagar 2016)) ialah P-CALS mengenalpasti gen pengawalselia yang berpotensi menggunakan kaedah tertentu manakala bLARS memilih gen pengawalselia berdasarkan sub-jaringan yang didapati secara rawak. Penggabungan beberapa fungsi regresi pada bLARS memberi ketepatan yang tinggi, namun, proses pengiraan berulang yang diusulkan oleh bLARS menjadikan modelnya kompleks.

1.3 OBJEKTIF KAJIAN

Objektif utama kajian ini adalah mengusulkan model pembinaan semula RPAG berasaskan gabungan kaedah unjuran dan analisis multivariat dengan menyelesaikan masalah ralat lata, jumlah pemerhatian terhad, multikolineran dan kesan hingar. Model yang diusulkan haruslah boleh diaplikasikan pada data sebenar ekspresi gen. Objektif utama ini disokong oleh beberapa sub-objektif berikut:

(1) Mengusulkan model yang mengambilkira kedua-dua peramal dan pembolehubah sambutan secara serentak.

(2) Mengusulkan model berasaskan regresi multivariat yang berkeupayaan mengelak hubungan tidak langsung (A→ B → C) ditakrif sebagai sebagai hubungan langsung (A → C) (ralat lata) pada data dengan bilangan pemerhatian terhad.

(3) Mengusulkan model berasaskan unjuran untuk mengenalpasti pembolehubah bererti dengan menyingkirkan kesan hingar.

(4) Mengintegrasikan model berasaskan unjuran dengan regresi multivariat bagi meningkatkan tahap keefisyenan model pembinaan semula RPAG.

1.4 SKOP KAJIAN

Hasil akhir kajian ini meliputi pembinaan semula RPAG berskala genom bagi organisma *e.coli* dan *s.serevisae* menggunakan data ekspresi gen. Kajian ini melakukan pembinaan semula RPAG menggunakan set data eksperimen yang kemudiannya dibandingkan dengan piawai emas *(gold standard)*. Metrik pengujian yang digunakan adalah AUROC *(Area Under ROC Curve)* dan AUPR *(Area Under Precision-Recall)*. Program yang dihasilkan bagi menterjemahkan algoritma yang diusulkan tidak mengambilkira aspek kelajuan pemprosesan secara khusus.

1.5 SUMBANGAN UTAMA KAJIAN

Sumbangan utama kajian adalah pengusulan P-CALS yang dihasilkan dari penggabungan model unjuran dengan regresi multivariat dan berkeupayaan menunjukkan prestasi yang baik menggunakan set data dengan bilangan pemerhatian yang terhad. P-CALS bermatlamat untuk mengelakkan berlakunya berlakunya kesalahan meramal hubungan tidak langsung (A→ B → C) sebagai hubungan langsung (A → C). Kajian ini turut mengusulkan model berasaskan unjuran untuk mengenalpasti pembolehubah bererti dengan menyingkirkan kesan hingar. Model yang mengambilkira kedua-dua peramal dan pembolehubah sambutan secara serentak serta mampu menangani isu multikolineariti turut diusulkan dalam kajian ini. Satu algoritma ringkas dan mudah digunakan bagi menguji kemampuan model pembinaan semula RPAG dalam menangani ralat lata turut diusulkan dalam kajian ini. Kebanyakan model utama dalam kajian ini menggunakan algoritma tersebut untuk pengujian. Kajian awal turut menyumbang pengetahuan baru iaitu P & G yang didapati merupakan kaedah terbaik untuk diaplikasikan pada data pemotongan homozigot dan heterozigot, iaitu data yang memerlukan perbezaan jelas antara nilai asal dengan nilai selepas tindakbalas berlaku.

1.6 METODOLOGI KAJIAN

Rajah 1.1 menunjukkan metodologi kajian secara umum. Langkah spesifik bagi mencapai setiap objektif yang tersenarai diterangkan pada Bab III.

Kajian Awal

- Menjalankan kajian literatur dan mengenalpasti masalah.
- Mengenalpasti MHG, PKP, P&G dan RLB bagi pembinaan semula RPAG.
- Menjalankan eksperimen dan membandingkan prestasi.
- Mengenalpasti RLB sebagai fokus seterusnya.
- **Hasil:** Jadual perbandingan model, Pernyataan masalah, Analisis kritis

Analisis Regresi — *Penubuhan Objektif 1*

- Mengusulkan model melibatkan peramal dan pembolehubah sambutan iaitu RLB.
- Mengusulkan algoritma pengujian ralat lata.
- Menjalankan eksperimen dan membandingkan prestasi.
- Mengenalpasti penambahbaikan dan mengenalpasti KDTS untuk terus ditambahbaik.
- **Hasil:** Keputusan eksperimen, Analisis penambahbaikan model, algoritma pengujian ralat lata

Analisis Multivariat — *Penubuhan Objektif 2*

- Mengusulkan KDTS bagi menyelesaikan ralat lata.
- Menjalankan eksperimen dan membandingkan prestasi.
- Mengenalpasti keperluan model bagi mengenalpasti pembolehubah bererti.
- **Hasil:** Keputusan eksperimen, Analisis penambahbaikan model

Kaedah Unjuran — *Penubuhan Objektif 3*

- Mengusulkan model berasaskan kaedah unjuran (APK) untuk mengenalpasti pembolehubah bererti dan menyingkirkan kesan hingar.
- Menjalankan eksperimen dan membandingkan prestasi.
- **Hasil:** Keputusan eksperimen, Analisis penambahbaikan model

Penggabungan — *Penubuhan Objektif 4*

- Mengenalpasti APK dan KDTS untuk terus dikembangkan.
- Mengenalpasti kaedah terbaik bagi menggabungkan kekuatan APK dan KDTS.
- Menjalankan eksperimen dan membandingkan prestasi.
- **Hasil:** Keputusan eksperimen, Analisis keseluruhan

Rajah 1.1 Metodologi kajian

1.7 ORGANISASI TESIS

Tesis ini dibahagikan kepada enam bab termasuk Pendahuluan (Bab I) dan (Bab VI) Perbincangan Akhir.

- **Bab II: Pembinaan Semula Rangkaian Pengawal Atur Gen**

  Bab ini menjelaskan teori RPAG dan istilah biologi yang berkaitan. Penyelidikan terkini berkaitan pembinaan semula RPAG terutama isu kritikal seperti permasalahan berkaitan motif lata, multikolinearan, data hingar dan pelbagai dimensi serta data dengan jumlah pemerhatian terhad dibentangkan. Metrik pengujian prestasi model yang paling sesuai digunakan turut diusulkan. Model sedia ada yang terdiri daripada pelabagai kategori seperi pembelajaran mesin, kebarangkalian, model berinspirasikan alam dan regresi dibincangkan dengan menerangkan konsep asas, kelemahan dan kelebihan serta kajian terkini serta pengaplikasiannya pada mana-mana domain.

- **Bab III: Metodologi**

  Bab ini membentangkan aktiviti kajian dan kaitannya dengan objektif penyelidikan. Turut dibentangkan ialah 4 langkah utama dalam metodologi kajian ini yang meliputi penerangan terperinci mengenai 1) pengusulan model yang mengambil kira peramal dan pembolehubah sambutan secara serentak, (2) pengusulan model berasaskan regresi multivariat, (3) pengusulan model berasaskan unjuran dan (4) penggabungan model berasaskan unjuran dengan regresi multivariat.

- **Bab IV: Kajian Penentuan Model RPAG**

  Bab ini mendemonstrasikan kajian awal terhadap MHG, PKP, P&G dan RLB. Hasil kajian dalam bab ini memberikan maklumat penting berkenaan kewujudan ralat lata dan kesannya terhadap prestasi model serta pemilihan satu model fokus untuk terus diterokai pada fasa kajian seterusnya.

- **Bab V: P-CALS Sebagai Model Berasaskan Regresi Bagi Menyelesaikan Permasalahan Ralat Lata**

  Bab ini telah memfokus pada model berasaskan regresi. Kajian ini memilih satu lagi jenis set data sintetik dan tiga set data nyata untuk diuji dalam eksperimen. Akhir sekali, bab ini memperlihatkan keberkesanan P-CALS yang terhasil daripada dua model yang saling melengkapi. P-CALS adalah merupakan model terbaik antara semua model yang telah diuji sebelum ini. Hasil eksperimen UKB dan algoritma mudah bagi pengenalpastian motif lata turut dibentangkan.

- **Bab VI Perbincangan Akhir**

  Bab ini memperihalkan perkara yang wajar dilaksanakan bagi terus mengembangkan kajian yang telah dilengkapkan ini. Beberapa saranan diberikan berdasarkan trend masa kini. Bab ini juga merumuskan keseluruhan kajian.

  Program dan jurnal yang telah diterbitkan berdasarkan tesis ini boleh didapati di: http://metalab.uniten.edu.my/~faridahhani/RPAG

1.8 RUMUSAN

Permasalahan kajian yang dihuraikan dalam bab ini ialah ralat lata, multikolineariti, bilangan pemerhatian terhad, kesan hingar dan keperluan terhadap kaedah pengenalpastian pembolehubah bererti. Setiap permasalahan kajian dihuraikan bersama objektif kajian yang mensasarkan untuk menyelesaikan setiap permasalahan yang dihuraikan. Set data dan metodologi kajian secara umum turut dibentangkan. Langkah kajian khusus bagi mencapai setiap objektif yang disenaraikan diperihalkan dalam Bab III Metodologi. Bab seterusnya membentangkan kajian kesusasteraan yang telah dijalankan bagi mengenalpasti ruang penambahbaikan.

BAB II

PEMBINAAN SEMULA RANGKAIAN PENGAWAL ATUR GEN

2.1 PENDAHULUAN

Bab ini membentangkan kajian terdahulu dalam menangani ralat lata, kelemahan serta kekuatan setiap model. Kajian terkini 8 kategori model iaitu teori maklumat, tapisan, kebarangkalian dan statistik, algoritma berinspirasikan alam, kolerasi dan kebergantungan, pembelajaran mesin dan meta algoritma dibentangkan. Jadual analisis semua kajian terkini berkaitan analisis regresi yang merupakan bidang fokus kajian turut dibentangkan dalam bab ini.

2.2 RANGKAIAN PENGAWAL ATUR GEN (RPAG)

2.2.1 Pengenalan

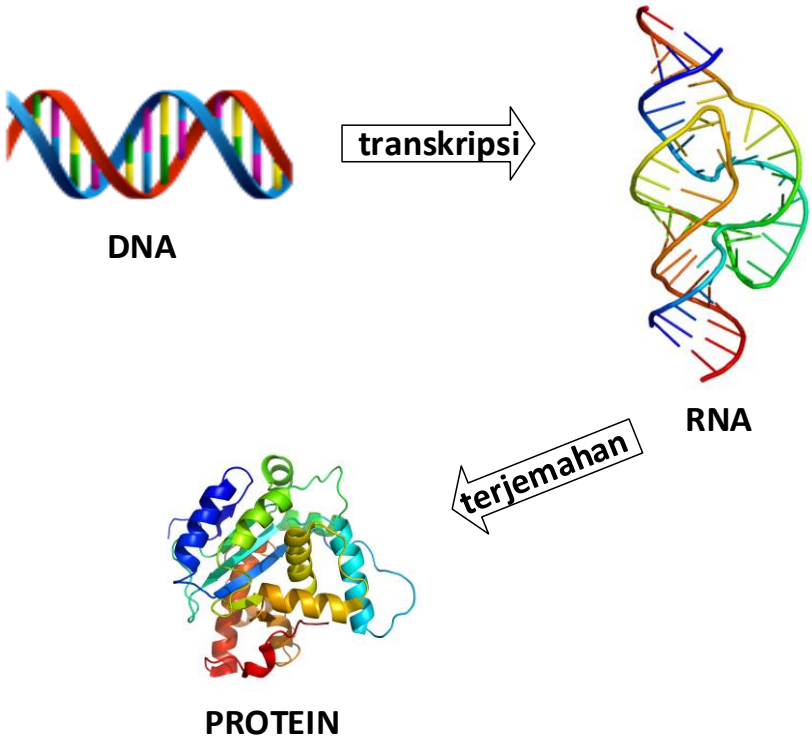
Pembinaan semula RPAG adalah bidang penyelidikan yang aktif dengan bilangan algoritma dan penerbitan yang semakin meningkat dari masa ke semasa (Chai et al. 2014; Wang & Huang 2014). Malah, terdapat pertandingan khas yang dianjurkan bagi penyelidik mengemukakan algoritma mereka untuk diuji dan dibandingkan dengan model lain. Pertandingan yang dinamakan DREAM *(Dialogue for Reverse Engineering Assessments and Methods)* dianjurkan dengan objektif untuk menjadi pemangkin interaksi antara teori dan eksperimen, khususnya dalam bidang pembinaan semula RPAG dan pembangunan model kuantitatif (Guo et al. 2016) . Keputusan eksperimen yang melibatkan komuniti penyelidik dari seluruh dunia seperti DREAM menunjukkan pencarian model pembinaan semula RPAG yang boleh dipercayai masih lagi merupakan masalah yang tidak dapat diselesaikan sepenuhnya (Marbach et al. 2010).

2.2.2 Gambaran Keseluruhan Rangkaian Pengawal Atur Gen

RPAG adalah satu set gen-gen yang berinteraksi antara satu sama lain untuk mengawal fungsi sel tertentu. RPAG adalah penting dalam pembangunan, pembezaan dan tindak balas terhadap isyarat persekitaran. RPAG terdiri daripada dua komponen utama: nod dan tepian *(edges)*. Nod yang terdiri daripada gen dan pengawal selia *(regulator)* manakala tepian adalah hubungan antara nod. RPAG mentakrifkan sifat utama struktur dan fungsi program kawalan genom dalam haiwan.

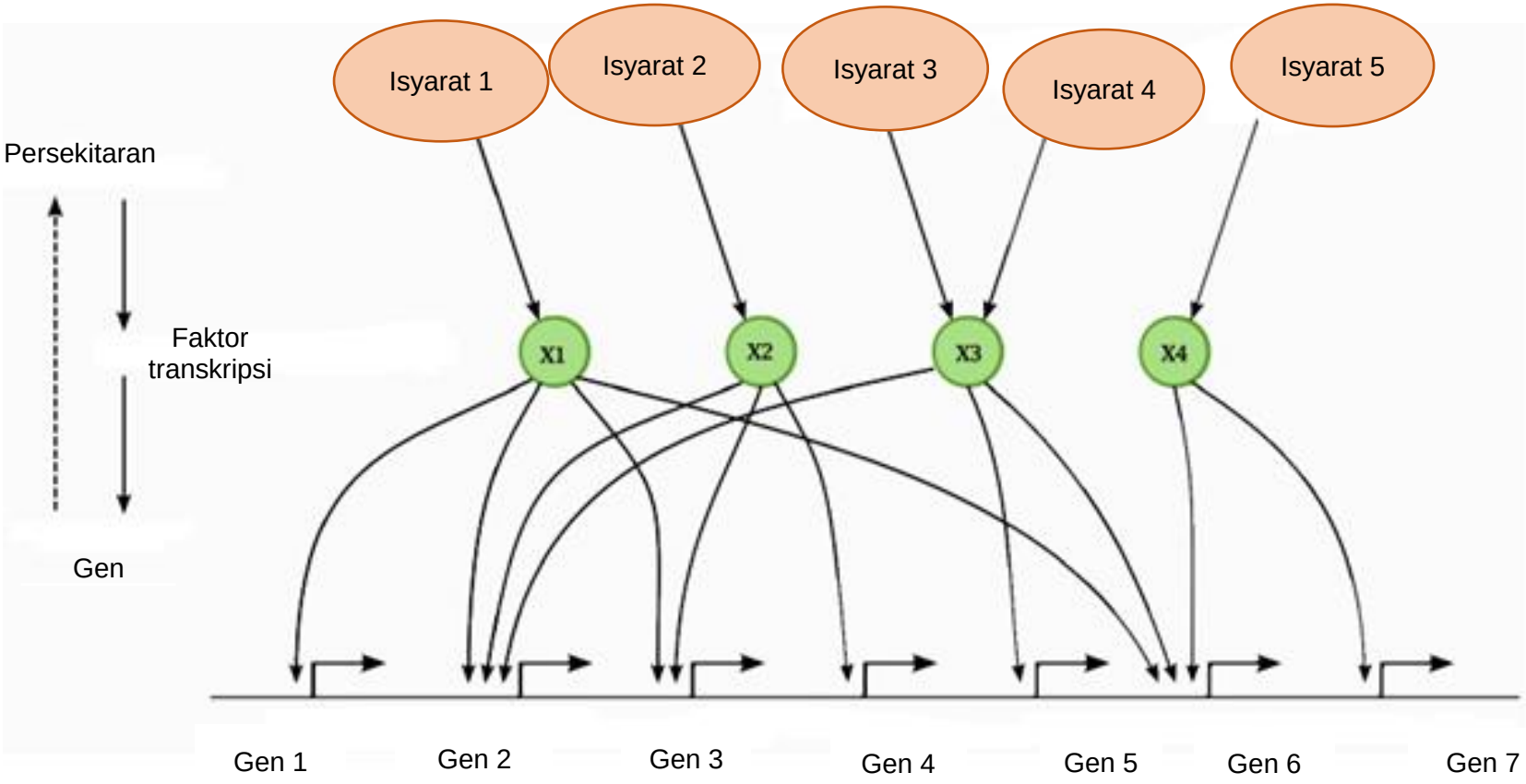
Pengawal atur gen digambarkan sebagai satu rangkaian yang nodnya ialah gen dan FT, serta mengandungi interaksi antara gen, di mana gen saling mempengaruhi antara satu samal lain(Salgado et al. 2012). Gen terhasil daripada DNA yang terdiri daripada empat asas nukleotida; adenina (A), guanina (G), sitosina (S) dan timina (T). Asas nukleotida berpasangan bersama (A-T dan S-G) memberikan DNA bentuk heliks ganda dua (Alon 2007). Kemudian, transkripsi gen berlaku. Permulaan transkripsi ialah langkah pertama dalam ekspresi gen dan merupakan titik penting kawalan dalam organisma. Peristiwa permulaan tersebut berlaku apabila enzim polimerase RNA (pRNA) mengenal pasti dan mengikat kepada jujukan DNA tertentu yang dikenali sebagai penggalak (Das et al. 2010).

Polimerase RNA (pRNA) ialah enzim yang menghasilkan mRNA yang sepadan dengan jujukan pengekodan gen. pRNA menguraikan bebenang DNA heliks ganda dua untuk menghasilkan bebenang mRNA tunggal (Alon 2007). Sejenis protein dipanggil FT mempengaruhi kadar pRNA memulakan transkripsi gen. pRNA mengikat penggalak yang merupakan kawasan DNA yang mengawal kadar gen yang disalin dan bilangan mRNA yang dihasilkan per unit masa. mRNA kemudiannya diterjemahkan dan menghasilkan protein (Das et al. 2010). Rajah 2.1 menunjukkan gambarajah ringkas aliran daripada DNA kepada protein. Gen-gen DNA diekspresikan dengan cara disalin kepada RNA dan kemudian diterjemahkan kepada protein.



Rajah 2.1 Aliran daripada DNA kepada protein (Alon 2007)

Bukan semua protein menjadi FT. Terdapat protein yang turut melaksanakan fungsi lain seperti membina struktur dan memangkinkan tindak balas. FT ialah protein yang mengubah kadar transkripsi set gen sasaran. FT bertindak sebagai pengaktif atau represor. Pengaktif meningkatkan kadar transkripsi sementara represor mengurangkan kadar transkripsi. Setelah protein dihasilkan, kandungan maklumat yang terdapat dalam protein dipindahkan apabila protein tersebut berinteraksi dengan protein lain. Interaksi ini mewujudkan rangkaian yang kompleks. Sukar untuk seseorang mengesan interaksi tersebut, namun komputer boleh digunakan untuk memodelkan perkara yang mungkin berlaku apabila beribu-ribu protein berinteraksi (Gopal et al. 2008; Das et al. 2010). Rajah 2.2 menunjukkan pemetaan antara isyarat persekitaran, FT dalam sel, dan gen yang dikawal atur. Isyarat tersebut memperoleh maklumat daripada persekitaran yang mempengaruhi FT, seterusnya mengubah cara protein dikawal atur.

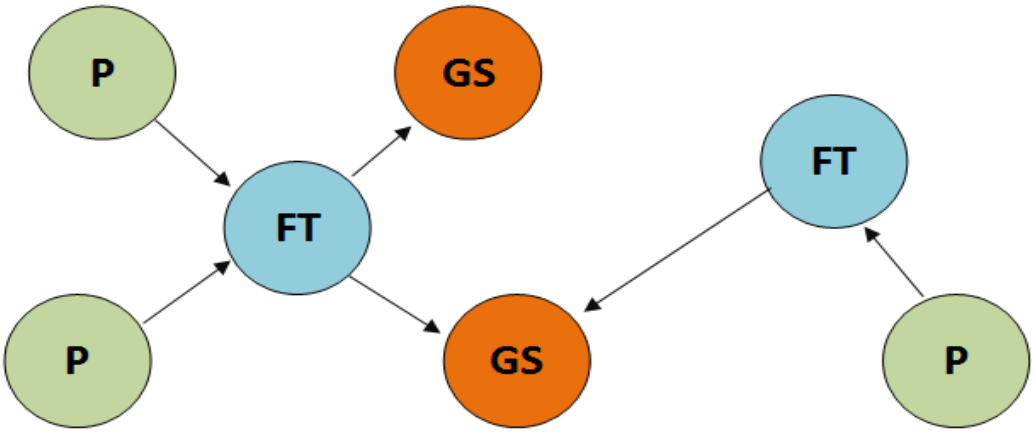


Rajah 2.2 Elemen rangkaian transkripsi (Sumber: (Alon 2007))

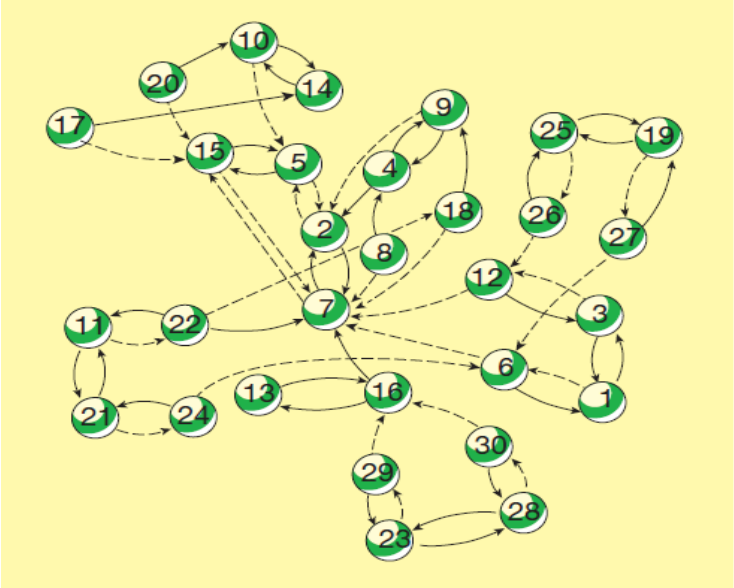
Persekitaran membawa isyarat seperti suhu, tekanan atau deria terhadap nutrien. Contoh senario protein yang bertindak balas dengan persekitaran adalah apabila protein tunggal yang digelar TRPM8 mengesan suhu sejuk dan mentol (sebatian yang memberikan rasa sejuk yang menyegarkan pada daun pudina)(Ledford 2007). Kesejukan boleh melegakan kesakitan dan mengurangkan keradangan. Tanpa TRPM8, kita akan kekurangan sensitiviti terhadap rasa sejuk (Dhaka et al. 2007). TRPM8 yang wujud pada tikus melaksanakan tugas yang sama pada manusia. Secara ringkas, rangkaian tersebut menggambarkan satu sistem dinamik yang mengawal isyarat masuk, menjadikan terjadinya perubahan aktiviti FT, lalu menyebabkan perubahan turut berlaku dalam kadar pengeluaran protein.

Umum mengetahui yang ekspresi gen ialah satu fenomena stokastik yang menjadikan rangkaian sangat rumit (Murphy & Mian 1999). Dalam konteks rangkaian transkripsi, stokastik melibatkan tahap rawak tertentu atau tidak dapat diduga. Sebagai contoh, penggunaan pemodelan stokastik untuk rangkaian transkripsi banyak bergantung pada model stokastik yang digunakan untuk meramal laluan interaksi protein. Ini adalah kerana interaksi protein adalah sesuatu yang sukar dijangka dan penggunaan model ramalan interaksi protein memang boleh membantu. Gangguan yang boleh mengubah dinamik rangkaian boleh berlaku dan model yang dibangunkan perlu mengambil kira fenomena ini. Rajah 2.3 menunjukkan gen dan istilah rasminya pada RPAG. Istilah ini digunakan pada keseluruhan tesis. Pengatur (P) adalah gen yang berhubung dengan FT dari arah keluar P. Hubungan keluar dari FT adalah kepada GS (Gen sasaran).

Rajah 2.4 menunjukkan contoh kawal atur daripada tindakan-rentas pengaktif dan represor. Setiap satu daripada beberapa ratus gen dikawal atur oleh produk pengaktif atau represor. Rajah tersebut menunjukkan peralihan antara 30 keadaan yang stabil dalam rangkaian ini. Rangkaian transkripsi ialah rangkaian elemen berinteraksi yang dari masa ke masa mempengaruhi keadaan antara satu sama lain. Dinamik rangkaian bergantung pada corak sambungan dan peraturan pengemaskinian bagi setiap elemen (Vohradsky 2001). Bilangan pengawal atur bagi setiap gen menunjukkan tahap kompleksiti sesebuah RPAG (De Jong 2002).

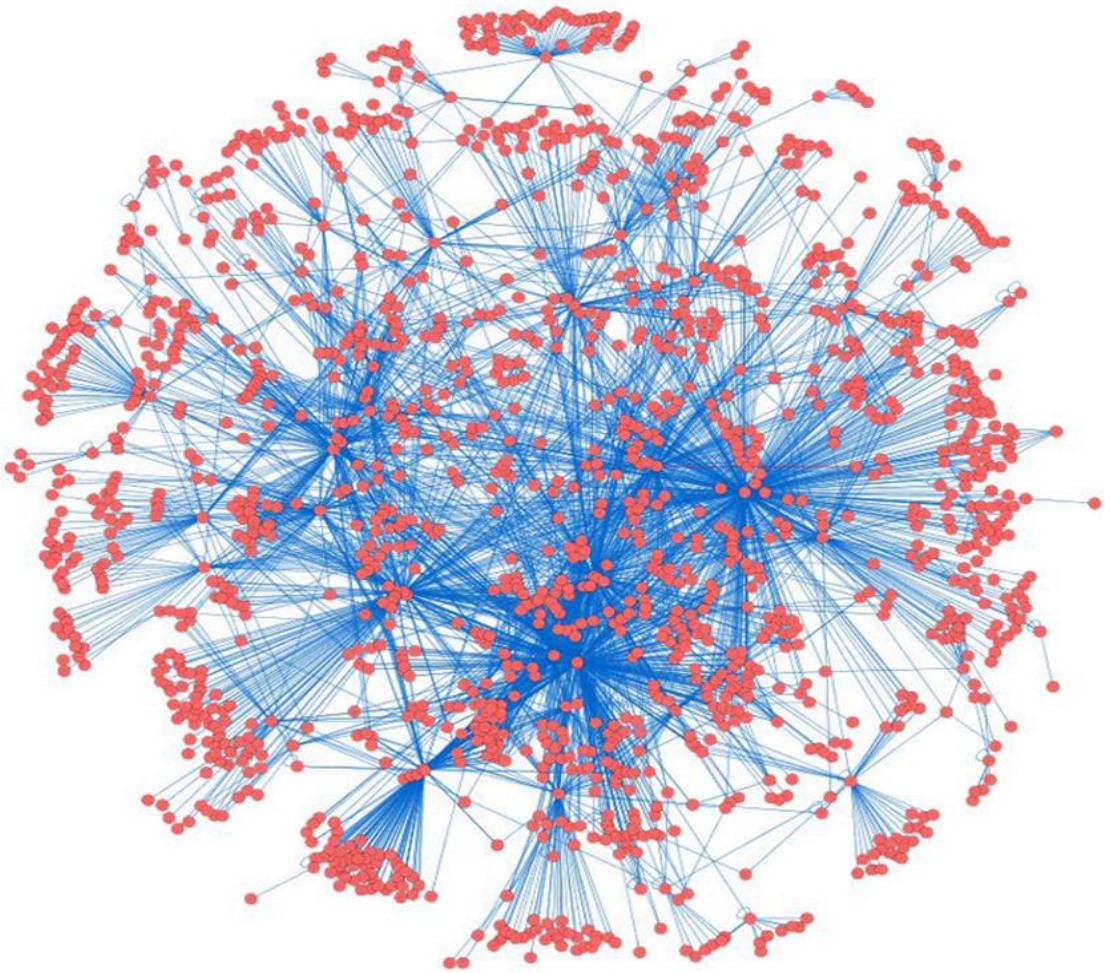


Rajah 2.3 Gen pada RPAG dan istilah berkaitan

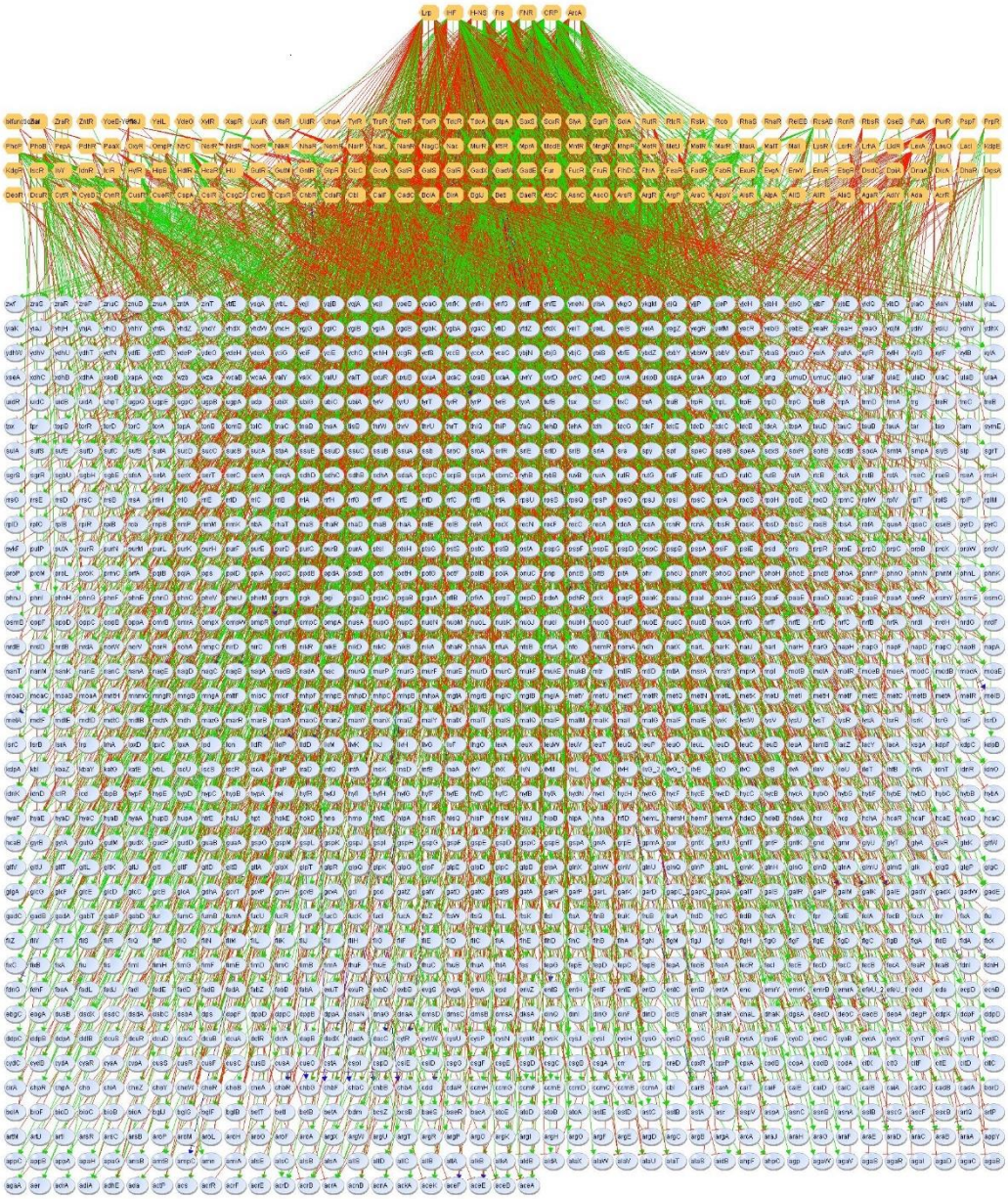


Rajah 2.4 Tingkah laku 'rangkaian pengawal atur gen', sumber: (Endy & Brent 2001)

Meningkatkan pemahaman terhadap RPAG bukan sahaja akan meningkatkan kefahaman entang sistem biologi, tetapi juga akan menyediakan asas pengetahuan menangani asal-usul penyakit manusia yang kompleks. Memahami bagaimana algoritma genom mengawal fenotip selular memerlukan penyelesaian fungsi pengawalseliaan pada skala luas genom. Rajah 2.5 menunjukkan rangkaian lengkap *e.coli* yang mempunyai jumlah gen sekitar 4,377. Dalam jaringan besar RPAG, terdapat 4 jenis motif rangkaian termasuk motif lata. Rajah 2.6 turut memaparkan RPAG *e.coli* dalam bentuk gambaran yang berbeza.



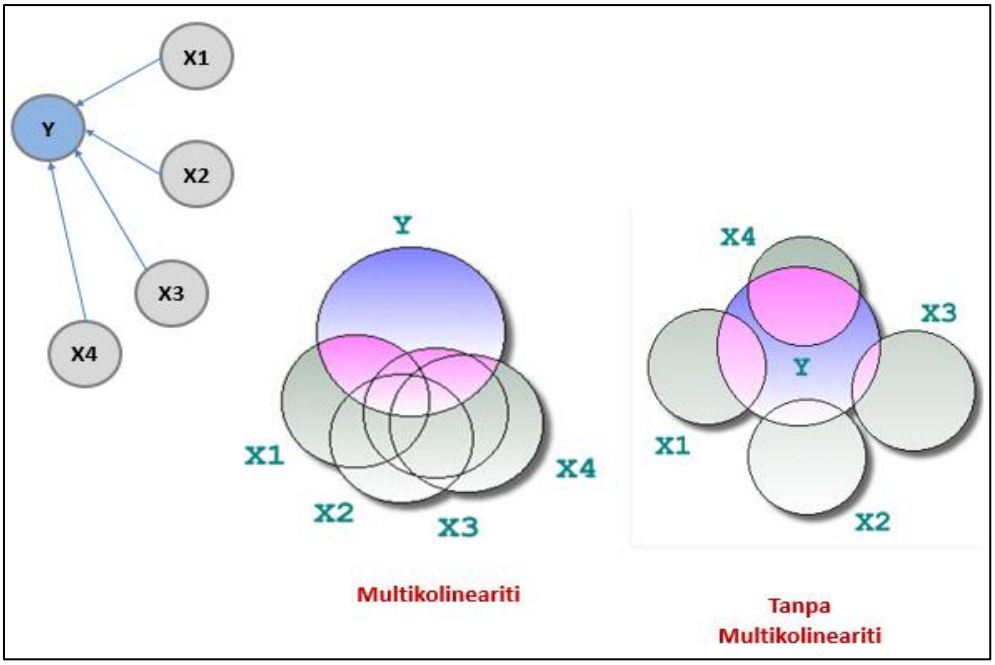
Rajah 2.5 RPAG e.coli yang diperolehi daripada pengkalan data RegulonDB (Allen et al. 2012). Setiap bulatan merah ialah gen dan garisan biru yang menghubungkan gen ialah hubungan antara gen



Rajah 2.6. RPAG *e.coli*

Multikolineariti ialah perkaitan yang tinggi antara dua atau lebih peramal (Freund & Wilson 1998), (Panik 2009), (Belsley et al. 2005). Mengambil contoh peramal X1, X2, X3 dan X4 yang bertindak ke atas gen sasaran Y, Rajah 2.7 menunjukkan senario kewujudan multikolineariti dan tanpa multikolineariti. Multikolinearan melemahkan analisis kuasa statistik, menyebabkan koefisien menukar isyarat, lalu menjadikannya lebih sukar untuk memfokus pada kaedah pengkomputeran yang betul (Miles & Shevlin 2001). Mulikolineariti juga menyebabkan pengiraan menjadi sensitif terhadap sebarang

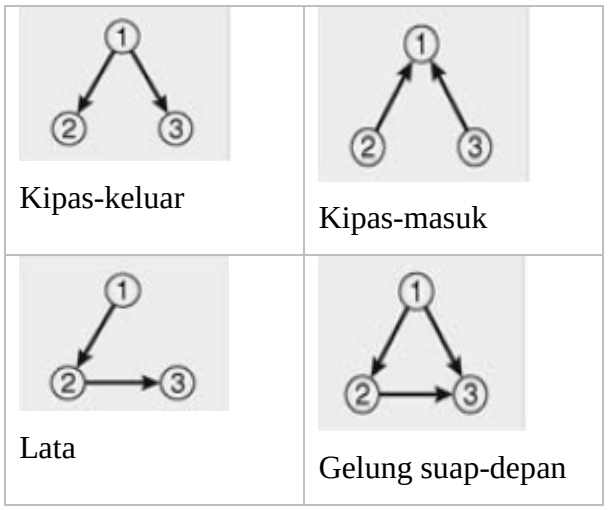
perubahan kecil (Belsley et al. 2005). Penggunaan kaedah pemusat bermakna (mean centering) yang lazimnya dilakukan atau menyingkirkan pembolehubah tanpa sebarang kaedah sistematik bukanlah cara terbaik untuk menangani multikolineariti.(Assaf et al. 2019). Kajian oleh (Assaf et al. 2019), (Yu et al. 2015), (Roozbeh et al. 2018) (Jadhav et al. 2014) mengusulkan kaedah pengekstrakan subset tersendiri. Mengatasi multikolineariti dengan menyingkirkan ralat daripada pemboleh ubah bersandar (Kemalbay & Korkmazoglu 2012). (Katrutsa & Strijov 2017) membentangkan pendekatan baru untuk mengelakkan multikolineariti dalam pemilihan ciri *(feature selection)*.



Rajah 2.7 Multikolineariti yang menunjukkan kaitan antara peramal yang tinggi

2.2.3 Ralat Lata

Perbincangan mengenai ralat lata dimulakan dengan Rajah 2.8 menunjukkan empat jenis motif rangkaian yang terdapat pada RPAG (tidak termasuk motif arah sendiri) iaitu kipas-keluar, kipas-masuk, lata dan gelung suap-depan (Markowetz & Spang 2007; Zhang et al. 2012).



Rajah 2.8 Jenis motif RPAG (Marbach et al. 2010), tidak termasuk motif arah sendiri. Bulatan yang dinomborkan ialah gen.

Ralat lata ditakrifkan sebagai kesalahan meramal "jalan pintas" atau hubungan tidak langsung disalah tafsir sebagai hubungan langsung. Menggunakan *motif lata* A -> B -> C sebagai contoh, ramalan salah yang sering dilakukan adalah dengan meramal A mempunyai hubungan terus dengan C (A -> C) dengan mengabaikan B. (Pinna et al. 2010; Swain et al. 2010; Wang & Zhou 2012; Zhang et al. 2012; Zhang et al. 2012; Barzel & Barabási 2013) (Küffner et al. 2012) (Tenenhaus et al. 2010) membincangkan mengenai ralat lata dalam kajian mereka. Walau bagaimanapun, eksperimen yang dijalankan dalam semua kajian ini tidak menjurus khusus ke arah membuktikan keupayaan model pembinaan semula RPAG dalam mengelak berlakunya ralat lata. (Friedman, 2004) menyatakan bahawa membezakan antara hubungan langsung dan tidak langsung (ralat lata) adalah suatu permasalahan yang telah diketahui kesukaran dalam menyelesaikannya, namun tidak pernah dinilai secara kuantitatif.

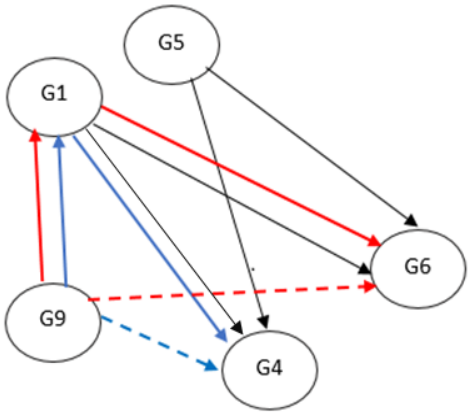
Jadual 2.1 menunjukkan contoh sebahagian daripada data ekspresi gen *E.coli2* (Marbach et al. 2009), digunakan untuk menunjukkan contoh hubungan tidak langsung. Rajah 2.9 memaparkan hubungan gen secara grafik berdasarkan data dalam Jadual 2.1. Pada Rajah 2.9, garis putus-putus mewakili ramalan yang salah, garis biru mewakili G_4^a , garis merah mewakili G_6^a , di mana *a* ialah gen yang dipadam dan garis hitam mewakili hubungan gen yang sebenar. Berdasarkan nilai sisihan antara data gen dan jenis-liar *(wild-type)*, terdapat tiga gen yang berkemungkinan mempunyai hubungan dengan G4 iaitu G1, G5 dan G9. Jenis-liar ialah bentuk biasa atau genotip satu organisma seperti

yang terdapat dalam alam semulajadi. Walaubagaimanapun, hanya G1->G4 dan G5->G4 yang mempunyai hubungan sebenar dengan G4. Kewujudan gen lain yang turut berpotensi menyukarkan proses mengenalpasti hanya satu gen yang benar-benar mempunyai hubungan dengan gen tertentu. Sebagai contoh, kajian ini meramal G9->G4, lalu menghasilkan ralat PP. G9 mempunyai hubungan dengan G4, namun secara tidak langsung iaitu (G9 -> G1 -> G4). Kes yang sama berlaku pada G9, di mana mungkin boleh disalah ramal mempunyai hubungan dengan G6, sedangkan G9 mempunyai hubungan dengan G6 melalui hubungan tidak langsung (G9 -> G1 -> G6). Daripada tiga yang mungkin, terdapat dua yang mempunyai hubungan yang benar iaitu G1->G6 dan G5->G6. Kebarangkalian untuk melakukan kesalahan ramalan ini semakin meningkat apabila data terkesan dengan hingar dan melibatkan jumlah gen yang banyak dan jaringan yang kompleks.

Jadual 2.1 Contoh hubungan tidak langsung dalam sub-jaringan *E.coli2*

strain	**G4**	**G6**
jenis-liar (*wild-type*)	0.552872	0.999801
G1(-/-)	**0.764693 ****	**0.004259 ****
G2(-/-)	0.552872	0.999801
G4(-/-)	0	0.999801
G5(-/-)	**0.314304 ****	**1 ****
G6(-/-)	0.552872	0
G8(-/-)	0.552872	0.999801
G9(-/-)	**0.572319**	**0.636537**

Merujuk pada Jadual 2.1 data gen ekspresi yang digelapkan menunjukkan gen yang berkemungkinan mempunyai hubungan dengan G4 dan G6. Gen yang sebenarnya berhubung dengan G4 dan G6 telah ditanda dengan asterisk (**). Data yang ditunjukkan adalah bebas hingar. Permasalahan dalam melakukan ramalan lebih ketara apabila data mempunyai hingar.



Rajah 2.9 Ralat lata. Hubungan antara gen dalam rajah ini adalah berdasarkan data pada Jadual 2.1

2.2.4 Kajian dalam Bidang Penyelesaian Permasalahan Berkaitan Motif Lata

(Barzel & Barabási 2013) merupakan kajian utama yang boleh dijadikan sebagai tanda aras bagi melihat kebolehupayaan model pembinaan semula RPAG dalam membezakan antara hubungan langsung dengan tidak langsung bagi gen-gen yang bergabung dalam satu rangkaian besar. (Barzel & Barabási 2013) mengusulkan beberapa formula yang menonjolkan hubungan langsung yang wujud antara gen berbanding hubungan tidak langsung, sekaligus menyebabkan ramalan ke atas hubungan langsung lebih mudah dilakukan tanpa sebarang bayangan daripada hubungan tidak langsung. Melihat pada senario yang dihadapi (Barzel & Barabási 2013) dalam menjalankan eksperimen ke atas data dengan bilangan nod sebanyak 4,511 dan 805 bilangan pemerhatian, kekurangan data pemerhatian menyebabkan (Barzel & Barabási 2013) melakukan seperti protokol DREAM5 yang memfokus pada kolerasi yang berlaku di 141 FT yang telah diketahui. (Kim et al. 2011) mengambil pendekatan yang berbeza iaitu dengan membahagikan RPAG kepada sub-jaringan yang setiap satunya terdiri daripada bilangan gen yang lebih kecil dan meyingkirkan pautan antara kluster bagi mengecilkan saiz jaringan. (Kim et al. 2011) mengusulkan tiga algoritma pengklusteran iaitu MPBN-MJB (Matrik Pemfaktoran Bukan Negatif untuk Motif Jaringan Biologi), MPBNGO-MJB (Matrik Pemfaktoran Bukan Negatif dengan terma GO bagi Motif Jaringan Biologi) dan VOLTAGE-MJB(Pengklusteran Voltage untuk Motif Jaringan Biologi). VOLTAGE-MJB telah digunakan untuk pengklusteran jaringan sebelum ini, tetapi bukan untuk mengenalpasti motif jaringan. Jika kajian ini mengusulkan penilaian ke atas kebolehan

sesuatu model pembinaan semula RPAG dalam meramal motif lata, (Barzel & Barabási 2013) melakukan sesuatu yang berbeza dengan menguji modelnya pada RPAG yang mempunyai nod tersembunyi. (Tenenhaus et al. 2010) menyentuh mengenai ralat lata dalam kajian mereka. Namun, tiada eksperimen khusus dijalankan bagi kajian tersebut membuktikan keupayaan KDTS dalam mengelakkan terjadinya ralat lata. (Feizi et al. 2013) dan (Kim et al. 2011) menyentuh aspek motif lata walaupun hasil kajiannya tidak menjurus terus kepada pengaplikasian ke atas pembinaan semula RPAG. Dua kajian ini juga dipilih untuk terus diselidiki kerana menghasilkan sesuatu berpotensi untuk diaplikasikan bagi menambahbaik kajian ini pada masa akan datang.

(Feizi et al. 2013) mengusulkan model pembinaan semula RPAG dengan mengenalpasti hubungan langsung gen daripada matrik kolerasi yang mengandungi gabungan hubungan langung dan tidak langsung. Secara amnya, gen yang berhubung langsung adalah lebih mudah diramal daripada yang tidak berhubung langsung. Maka, (Feizi et al. 2013) mengusulkan teknik penyahkonvolusi jaringan yang merupakan teknik yang digunakan secara meluas dalam bidang sains jaringan daripada pelbagai bidang. Penyahkonvolusi telah mengurangkan jumlah PP pada hubungan tidak langsung dengan cara masih mengekalkan jumlah PB pada gelung suap depan *(feed-forward loop).*

Kebanyakan model pembinaan semula RPAG mengalami penurunan prestasi ramalan disebabkan oleh jumlah bilangan PP yang banyak pada ramalan mereka (Wang & Michoel 2017). PP terhasil apabila hubungan antara gen yang diramalkan oleh model pembinaan semula RPAG tidak wujud dalam jaringan sebenar. Tidak seperti kelaziman, penemuan kajian (Feizi-Khankandi et al. 2013) tidak hanya diaplikasikan pada domain tertentu secara spesifik, tetapi dibuktikan boleh diaplikasi secara meluas pada sebarang domain lain. Jadual 2.2 memaparkan merumuskan 2 kajian utama yang menyentuh mengenai isu ralat lata.

Jadual 2.2 Rumusan kajian utama dalam bidang hubungan tidak langsung atau motif lata

Pengarang dan tahun terbitan	(Tenenhaus et al. 2010)	(Barzel & Barabási 2013)
Objektif/ Persoalan Kajian	1. Mengenalpasti langkah-langkah yang perlu diambil untuk membezakan antara hubungan terus dan hubungan tidak terus dalam data berdimensi tinggi seperti mikrotatasusunan 2. Membina RPAG yang mempunyai kadar positif palsu rendah	Membina kaedah bagi melenyapkan kesan hubungan tidak langsung
Model dan kaedah	KDTS	Kaedah senyap diterbitkan daripada prinsip matematik kolerasi dinamik pada jaringan
Set data benar	Data masa perjalanan mikrotatasusunan *e.coli*	Set data *e.coli* DREAM5
Set data sintetik	ODE	Tiada
Kaedah pengujian	AUROC, Kepekaan, Nilai Ramalan Positif	AUROC, Nisbah Diskriminasi
Kelebihan	Menangani isu n « p secara spesifik, mensimulasikan hasil kajian pada kedua-dua data jenis statik dan dinamik	Kaedah pelenyapan hubungan tidak langsung meningkatkan keupayaan pengekstrakan hubungan langsung yang terdapat pada RPAG yang diuji.
Kekurangan	Tiada pembuktian khusus bagi mengkaji keberkesanan kebolehan membezakan antara hubungan terus dengan tidak terus	Model yang diusulkan akan gagal berfungsi sekiranya peratusan gen tersembunyi melebihi 60%.
Perbincangan mengenai ralat lata	Dibincangkan, namun tiada penilaian berkaitan ralat lata dilakukan	Dibincangkan dan terdapat penilaian berkaitan ralat lata dilakukan. Namun, kaedah terperinci tentang bagaimana kaedah yang diusulkan menangani ralat lata tidak dibentangkan.
Bagaimana kajian ini menangani data pemerhatian terhad	Anggaran Penghitungan Kolerasi Separa berdasarkan PLS-R	Memfokus kepada kolerasi antara 141 FT yang telah diketahui.

2.3 DATA ANALITIK DAN PENGAPLIKASIANNYA PADA DOMAIN PERKOMPUTERAN BIOLOGI

Kemajuan pesat dalam teknologi genomik menjadikan data analitik merupakan suatu keperluan dalam bidang biologi. Data analitik telah menjadi suatu keperluan mendesak sejak beberapa tahun kebelakangan ini. Beberapa profesion baru seperti saintis data dan penganalisa data telah muncul kesan daripada kebangkitan era data analitik. Data analitik telah digunakan dan sedang rancak diaplikasikan pada domain industri seperti perkilangan (Zhang et al. 2017), penerbangan (Chen et al. 2017), pertanian (Pham & Stack 2018) (Kamilaris et al. 2017) (Devalkar et al. 2018), pelancongan (Assaf et al. 2019) (Liu et al. 2018), penjagaan kesihatan (Wang et al. 2018). Kebanyakan aplikasi data analitik ini dilakukan dengan tujuan meningkatkan produktiviti dan keuntungan, menambahbaik proses sedia ada, membantu dalam membuat keputusan, merangka strategi dan meramalkan prospek masa hadapan (Vassakis et al. 2018). Bahagian ini membincangkan secara khusus mengenai data analitik dan pengaplikasiannya pada domain perkomputeran biologi. Perkomputeran biologi dibincangkan kerana ianya berkait rapat dengan kajian dalam tesis ini. Selain perkomputeran biologi terdapat bidang lain dalam biologi seperti sistem biologi, biokimia, genomik, zoologi, genetik, immunologi, biofizik dan mikrobiologi.

Rujuk Rajah 2.10 yang memaparkan keseluruhan komponen dalam data analitik pada perkomputeran biologi. Membina model bagi data analitik adalah kompleks, mempunyai pelbagai langkah, dan melibatkan proses rekursif yang terdiri daripada tiga langkah asas. Pertama, data mentah diproses untuk menghasilkan satu set data dengan perwakilan dan kualiti yang lebih baik (data preprocessing). Proses awal data termasuk normalisasi, transformasi, pengendalian data hilang, dan pemilihan ciri (feature selection) (Ma, Zhang et al. 2014). Perkara asas yang penting adalah kefahaman terhadap konsep biologi, statistik dan matematik, pengaturcaraan dan kemahiran penggunaan perisian. Antara semua cabang kaedah, analisa multivariat dan pembelajaran mesin adalah dua kategori utama yang penyelidik wajar fokuskan untuk mengaplikasikan data analitik dalam mana-mana kajian dalam bidang biologi. Bidang

fokus ini perlu ditentukan awal supaya mudah bagi penyelidik untuk memilih jenis program latihan yang ingin disertai.

Statistik deskriptif adalah lebih mudah difahami dan melibatkan konsep seperti Mean, Maksimum, Minimum, Median, Kuartil, Julat (Maksimum-Minimum, Varians), Taburan Kepencongan (Skewness), Kurtosis, Korelasi silang (Cross correlations) dan graf serak (scatter graph). Statistik deskriptif adalah koefisien deskriptif ringkas yang meringkaskan set data dan boleh berfungsi samada sebagai perwakilan pada keseluruhan atau sampel populasi. Contoh maklumat statistik deskriptif yang dipaparkan adalah mean, median, mod, kuartil, sisihan piawai dan varians. Jadual 2.3, Jadual 2.4, Jadual 2.5 dan Jadual 2.6 memaparkan senarai perisian data analitik. Penerangan mengenai ini dimasukkan kerana sejak beberapa tahun kebelakangan ini kewujudan pelbagai aplikasi komersil dilihat sebagai salh satu aspek yang perlu diambil kira bagi menganalisis data. Bagi kajian berkaitan biologi, tidak semua perisian komersil sesuai digunakan kerana tidak semua perisian ini menyediakan kemudahan untuk memanipulasi algoritma bagi memenuhi keperluan penyelidik. Perisian seperti R dan MATLAB adalah contoh perisian yang fleksibel di mana penyelidik boleh mengubah algoritma. Namun, keburukan perisian seperti ini adalah memerlukan seseorang untuk mempunyai asas pengaturcaraan. Ingin ditegaskan bahawa bagi bidang biologi, kebolehupayaan mengubahsuai algoritma adalah merupakan suatu ciri penting dalam pemilihan perisian. Pemilihan perisian perlu dilakukan dengan teliti kerana perisian komersil pada masa kini memerlukan kos yang tinggi untuk mendapatkannya. Satu lagi ciri penting untuk kajian bidang biologi adalah manipulasi matrik. Sebagai contoh, jika satu gen ingin diketahui samada mempunyai hubungan dengan setiap satu gen dalam satu senarai yang panjang, manipulasi matrik sangat berguna.

Dari Jadual 2.3, Jadual 2.4, Jadual 2.5 dan jadual 2.6 juga dapat dilihat bahawa model pepohon keputusan *(decision tree)*, semua model berasas regresi dan APK adalah model yang sering ditawarkan oleh perisian analisa data pada masa kini. Ini juga memberi petanda bahawa model tersebut mempunyai potensi yang luas untuk terus diterokai. Pemprosesan awal adalah melibatkan tugas seperti mempertimbangkan tentang apa yang perlu dilakukan sekiranya mempunyai nilai hilang *(missing values)*, transformasi data (contoh dari deskriptif ke berangka *(numerical)*), penyingkiran

outlier, data terkeluar dari julat yang sepatutnya, dan data tidak konsisten seperti mengandungi percanggahan dalam kod atau nama.

Visualisasi juga merupakan suatu ciri penting diperlukan oleh penyelidik untuk melihat suatu ciri ketara seperti kesan hingar dan kepencongan (skewness). Selari dengan teknologi dunia moden yang menjadikan proses mendapatkan data dalam jumlah yang tinggi semakin mudah, keperluan mengendalikan pangkalan data raya semakin memuncak. Tidak semua perisian mempunyai integrasi terus dengan pangkalan data tersebut seperti RapidMiner dan Radoop, namun, kebanyakan perisian membenarkan integrasi dilakukan dengan pangkalan data sumber terbuka *(open source)* seperti MATLAB dan Hadoop. Pengujian prestasi yang diperlukan bagi sesebuah kajian adalah berbeza antara satu sama lain. Perbandingan prestasi lazimnya diketahui melalui pengkajian terhadap hasil kajian penyelidik lain. Varians tertaktif, Nilai reja *(residual)*, AUROC, AUPR, RMSE dan PPV adalah antara pengujian prestasi model. Rajah 2.11 menunjukkan cabang bidang kajian perkomputeran biologi yang khusus. Jika bidang lain mengaplikasikan data analitik lebih kepada keperluan terhadap memaksimakan keuntungan, memperbaiki mutu perkhidmatan atau komersil, perkomputeran biologi pula bertujuan lebih kepada pembentukan pengetahuan asas untuk kajian masa hadapan.

Pemprosesan awal *(pre-processing data)* adalah satu tugasan yang penting dilakukan pada semua jenis data sebelum memulakan sebarang analisa data. Sesetengah proses ini tidak boleh dilakukan secara manual dan memerlukan perisian sebagai contoh, menggantikan nilai hilang *(missing values)* dengan nilai mod untuk atribut nominal dan min bagi atribut angka berterusan, menggantikan nilai yang hilang dengan nombor rawak, atau menggantikan nilai yang hilang dengan anggaran nilai menggunakan regresi linear (Rapidminer 2018). Satu lagi aspek penting untuk dipertimbangkan semasa pemprosesan data adalah sama ada perlu menormalkan *(normalize)* mana-mana sifat. Normalisasi diperlukan untuk sifat-sifat berangka yang berterusan dan mempunyai skala yang berbeza. Terdapat algoritma hanya menerima atribut diskrit dan bukan berterusan *(continuous)*. Maka, atribut tersebut perlu dibah kepada atribut diskret *(discretized)*. Proses awal lain yang perlu dipertimbangkan adalah keperluan pengenalpastian nilai ekstrim yang berada di luar taburan data umum *(outlier)* dan pengenalpastian atribut yang berkait *(correlated attribute)*.

Jadual 2.3. Perisian data analitik

	The Unscrambler X (Camo Analytics 2018)	**RapidMiner (Rapidminer 2018)**	**R (R 2018)**
Bahasa pengaturcaraan utama	Tidak dilaporkan	JAVA	R
Pengujian prestasi	Dibekalkan, bergantung pada jenis model	Dibekalkan, bergantung pada jenis model	Dibekalkan, bergantung pada jenis model
Manipulasi algoritma	Tidak diberikan	Tidak diberikan	Diberikan dan algoritma ditulis dalam bentuk arahan
Manipulasi matrik	Tidak diberikan	Tidak diberikan	Diberikan
Pembelajaran mesin	Tidak disokong oleh perisian	Rangkaian neural *(Neural network)*, Mesin Vektor Sokongan *(Support Vector Machines)*, Pepohon keputusan *(decision tree)*,	Rangkaian neural, Algoritma Genetik *(Genetic Algorithm)*, Bayesian, Pepohon keputusan *(decision tree)*, pembelajaran dalam *(deep learning)*
Regresi	RLB, APK, Regresi Prinsip Komponen *(Principal Component Regression)*, KDTS, Regresi Mesin Vektor Sokongan *(Support Vector Machines Regression)*	Regresi Linear, Logistik Kernel *(Kernel Logistic)*, Analisis Diskriminan Linear *(Linear Discriminant Analysis)*, Analisis Diskriminan Kuadratik *(Quadratic Discriminant Analysis)*, Analisi Diskriminan Sekata *(Regularized Discriminant Analysis)*, Berperingkat Kehadapan *(Stepwise forward)*	Regresi Linear, KDTS, Kernel, Bent-Cable, LASSO, Cox, Canonical,

Jadual 2.4. Perisian data analitik (sambungan)

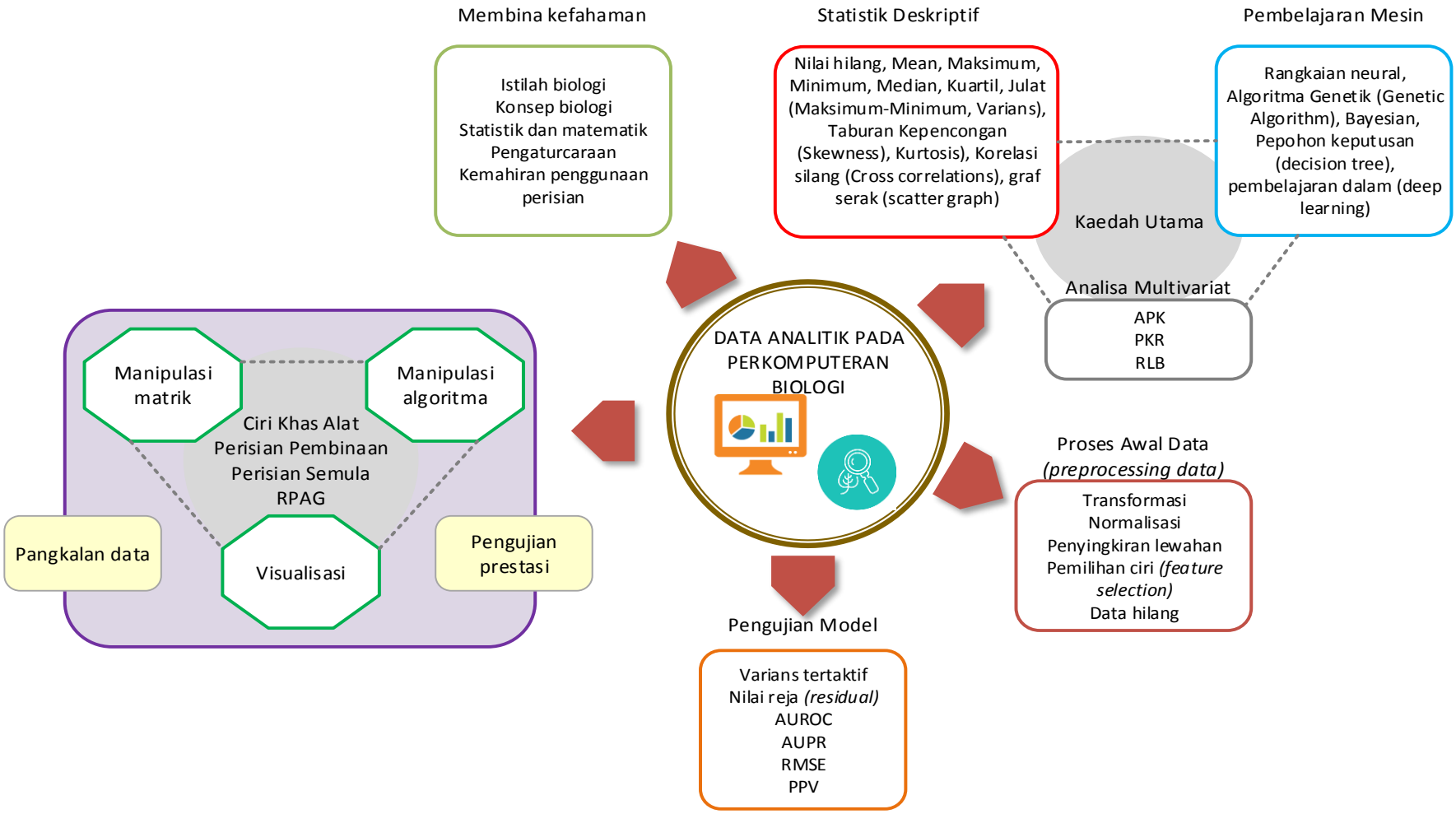
	The Unscrambler X (Camo Analytics 2018)	**RapidMiner (Rapidminer 2018)**	**R (R 2018)**
Statistik deskriptif *(Descriptive Statistics)*	Nilai hilang, Mean, Maksimum, Minimum, Median, Kuartil, Julat (Maksimum-Minimum, Varians), Taburan Kepencongan *(Skewness)*, Kurtosis, Korelasi silang *(Cross correlations)*, graf serak *(scatter graph)*	Pemberat setiap atribut *(attribute)*, matrik MS *(Mutual Information matrix)*, matrik ANOVA, matrik korelasi, Nilai hilang, Mean, Maksimum, Minimum,	Mean, Maksimum, Minimum, Median, Kuartil, Julat (Maksimum-Minimum, Varians), sisihan piawai
Pangkalan data data raya *(big data)*	Tidak	Ya Radoop	Tidak
Menyokong pengaturcaraan	Tidak	Tidak	Ya
Perlesenan	Komersil	Komersil	Sumber terbuka
Ciri antara muka	Capaian fungsi diperolehi dengan mencari dari menu. Banyak visualisasi grafik diberikan.	Konsep seret dan letak *(drag and drop)*. Fungsi *(function)* yang digunakan diwakili oleh ikon.	Tidak mesra pengguna kerana berorientasikan pengkompil *(compiler)*.
Proses awal data *(preprocessing data)*	Nilai hilang *(missing values)* , transformasi data (contoh dari deskriptif ke berangka *(numerical)*), penyingkiran outlier, tidak konsisten: mengandungi percanggahan dalam kod atau nama.	Nilai hilang *(missing values)* , transformasi data (contoh dari deskriptif ke berangka *(numerical)*), penyingkiran outlier, tidak konsisten: mengandungi percanggahan dalam kod atau nama.	Memerlukan program ringkas untuk beri arahan

Jadual 2.5. Perisian data analitik tambahan

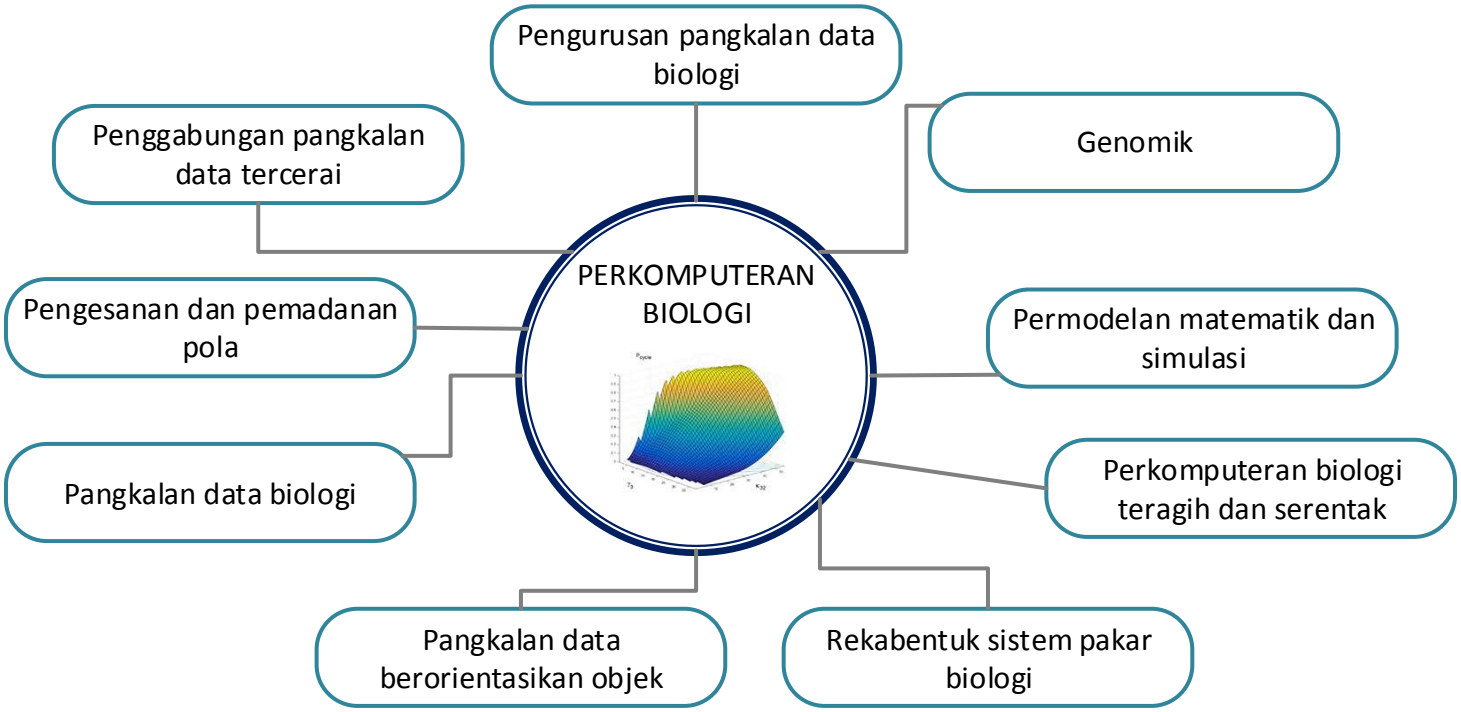
	SAS (SAS 2018)	**SAP (SAP 2018)**	**MATLAB (MATLAB 2018)**
Bahasa pengaturcaraan utama	Tidak dilaporkan	JAVA	C, C++, Java
Pengujian prestasi	Dibekalkan, bergantung pada jenis model	Dibekalkan, bergantung pada jenis model	Dibekalkan, bergantung pada jenis model
Manipulasi algoritma	Dibekalkan, bergantung pada jenis model	Dibekalkan, bergantung pada jenis model	Diberikan dan algoritma ditulis dalam bentuk arahan
Manipulasi matrik	Tidak diberikan	Tidak diberikan	Diberikan
Pembelajaran mesin	Tidak disokong oleh perisian	Rangkaian neural *(Neural network)*, Mesin Vektor Sokongan *(Support Vector Machines)*	Rangkaian neural, Algoritma Genetik *(Genetic Algorithm)*, Bayesian, Pepohon keputusan *(decision tree)*, pembelajaran dalam *(deep learning)*
Regresi	RLB, APK, Regresi Prinsip Komponen *(Principal Component Regression)*, KDTS, Regresi Mesin Vektor Sokongan *(Support Vector Machines Regression)*	Regresi Linear, Logistik Kernel *(Kernel Logistic)*, Analisis Diskriminan Linear *(Linear Discriminant Analysis)*, Analisis Diskriminan Kuadratik *(Quadratic Discriminant Analysis)*, Analisi Diskriminan Sekata *(Regularized Discriminant Analysis)*, Berperingkat Kehadapan *(Stepwise forward)*	Regresi Linear Mudah *(Simple Linear Regression)*, Regresi Berperingkat *(Stepwise Regression)*, Regresi Tak Linear *(Nonlinear Regression)*, Regresi Jujukan Masa *(Time Series Regression)*

Jadual 2.6. Perisian data analitik tambahan (sambungan)

	SAS (SAS 2018)	**SAP (SAP 2018)**	**MATLAB (MATLAB 2018)**
Statistik deskriptif *(Descriptive Statistics)*	Nilai hilang, Mean, Maksimum, Minimum, Median, Kuartil, Julat (Maksimum-Minimum, Varians), Taburan Kepencongan *(Skewness)*, Kurtosis, Korelasi silang *(Cross correlations)*, graf serak *(scatter graph)*	Pemberat setiap atribut *(attribute)*, matrik MS *(Mutual Information matrix)*, matrik ANOVA, matrik korelasi, Nilai hilang, Mean, Maksimum, Minimum,	Mean, Maksimum, Minimum, Median, Kuartil, Julat (Maksimum-Minimum, Varians), sisihan piawai
Pangkalan data data raya *(big data)*	Tidak	Ya Radoop	Tidak
Menyokong pengaturcaraan	Tidak	Tidak	Ya
Perlesenan	Komersil	Komersil	Komersil
Ciri antara muka	Capaian fungsi diperolehi dengan mencari dari menu. Banyak visualisasi grafik diberikan.	Konsep seret dan letak *(drag and drop)*. Fungsi *(function)* yang digunakan diwakili oleh ikon.	Berorientasikan pengkompil *(compiler)* namun masih mempunyai sedikit antara muka ringkas.
Proses awal data *(preprocessing data)*	Nilai hilang *(missing values)*, transformasi data (contoh dari deskriptif ke berangka *(numerical)*), penyingkiran outlier, tidak konsisten: mengandungi percanggahan dalam kod atau nama.	Nilai hilang *(missing values)*, transformasi data (contoh dari deskriptif ke berangka *(numerical)*), penyingkiran outlier, tidak konsisten: mengandungi percanggahan dalam kod atau nama.	Memerlukan program ringkas untuk beri arahan



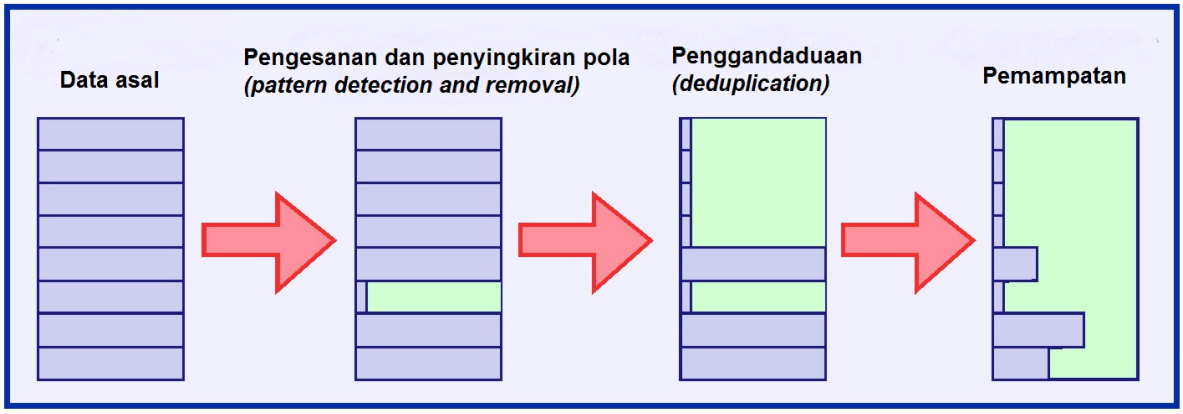
Rajah 2.10. Pengaplikasian data analitik pada perkomputeran biologi



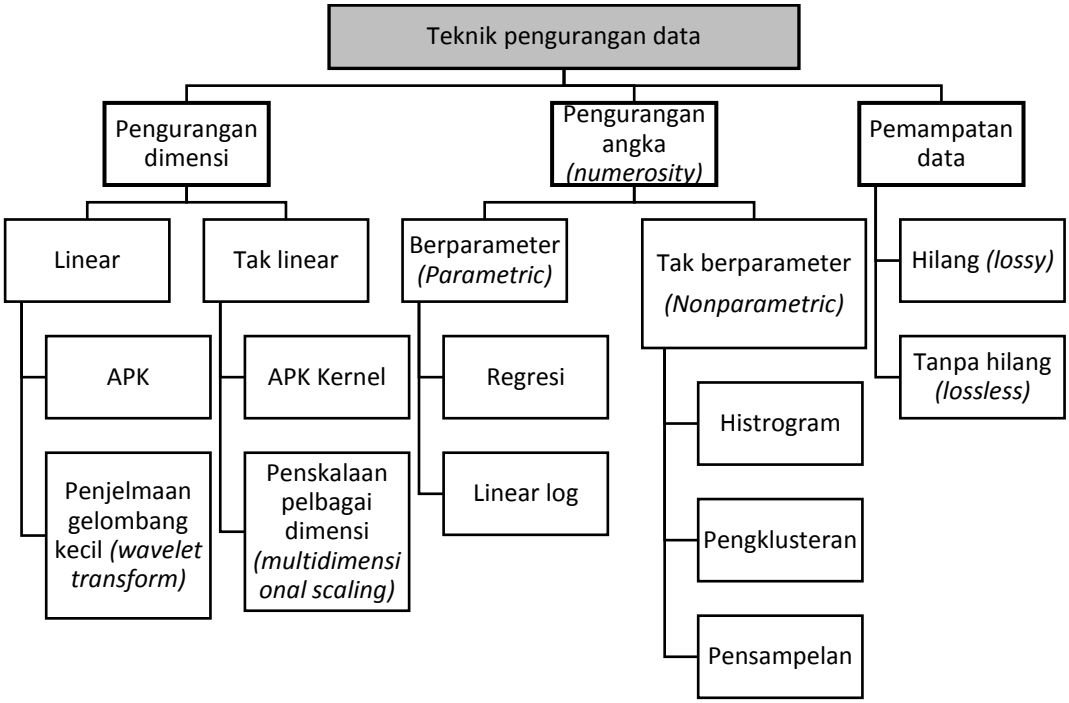
Rajah 2.11. Perkomputeran biologi

2.3.1 Pengurangan Data

Terdapat beberapa isu berkaitan data yang lazim ditemui dalam data analitik iaitu keperluan dalam mengurangkan data yang diproses, pengurangan dimensi dan jumlah sampel data yang terhad. Teknik pengurangan data sangat berguna untuk mendapatkan set data dalam jumlah yang lebih kecil, dan dalam masa yang sama mengelakkan kehilangan maklumat penting yangterdapat pada data asal. Kaedah pengurangan data termasuk pengurangan dimensi, pengurangan angka *(numerosity)*, dan pemampatan data (Arabnia and Tran 2016) (Packt 2014). (IBM 2018) mengusulkan teknik pengurangan data yang sedikit berbeza yang turut boleh diguna pakai oleh analisis data biologi terutama jika melibatkan pemprosesan kompleks yang mungkin mengambil masa yang terlalu lama. Kajian oleh (Xenitidis et al. 2017) juga telah membuktikan bahawa jumlah set data adalah penentu utama prestasi model pembinaan semula RPAG. Rajah 2.12 memaparkan peringkat pengurangan data dan Rajah 2.13 menyenaraikan jenis teknik pengurangan data.



Rajah 2.12. Peringkat pengurangan data (Sumber: IBM, 2018)

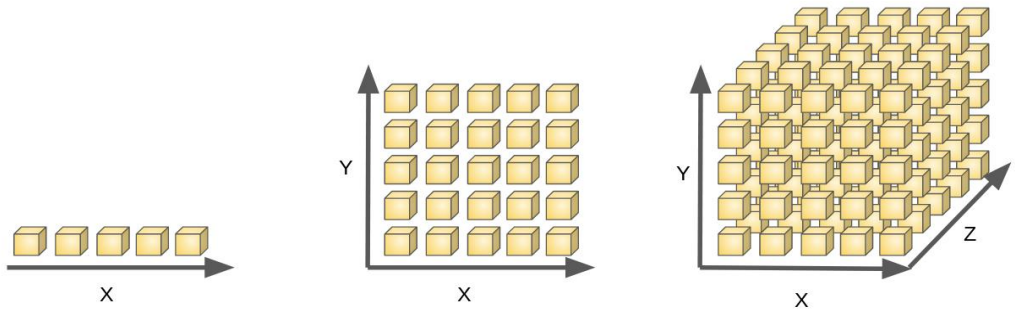


Rajah 2.13. Teknik pengurangan data (Sumber: Packt 2014)

2.3.2 Pengurangan Dimensi

Pengurangan dimensi adalah merupakan salah satu topik penyelidikan yang aktif dijalankan dalam kajian analisis data pada masa kini. Kewujudan pelbagai teknologi tinggi menjadikan proses penghasilan jutaan data dari eksperimen boleh dilakukan. Walau bagaimanapun, bilangan data adalah yang sangat tinggi ini perlu dikurangkan dengan cara tertentu kepada bilangan pemboleh ubah yang jauh lebih kecil tanpa kehilangan maklumat yang berguna untuk analisis. Antara teknik yang boleh digunakan adalah Unjuran Rawak *(Random Projection)* (Xie, Li et al. 2016), APK (Shi and Luo 2010), Analisis Pembeza Linear *(Linear Discriminant Analysis)*, Pemilihan Ciri *(Feature Selection)* (Arabnia and Tran 2016) dan Analisa Komponen Bebas *(Independent Component Analysis)* dan Koloni Lebah Buatan *(Artificial Bee Colony)* (Aziz, Verma et al. 2017). Rajah 2.14 menunjukkan data pelbagai dimensi. Rajah 2.14 digambarkan dengan situasi perlu mencari sesuatu barang dalam kiub yang tersusun. Dengan n = 1, terdapat 5 kiub yang perlu dicari. Sekiranya n = 2, terdapat 25 kiub dan jika n = 3, terdapat 125 kiub yang perlu dicari. Semakin besar nilai n, semakin sukar untuk mensampelkan kesemua kiub. Tambahan lagi, pencarian barang menjadi semakin

sukar apabila kiub-kiub adalah kosong. Secara umumnya, dengan n dimensi yang membenarkan m keadaan, terdapat m ^ n kombinasi yang mungkin. Perisian R mempunyai beberapa pakej khas seperti dimRed, DRR, imputeMDR, mbmdr yang boleh digunakan untuk mengurangkan dimensi. Selain daripada pengurangan dimensi data, turut terdapat keperluan menangani saiz sampel terhad (Li et al. 2015) (Zhang et al. 2015) (Zhang et al. 2017) yang sangat lazim ditemui pada data biologi dan perubatan.



Rajah 2.14. Data pelbagai dimensi (Sumber: Peter Gleesen, 2017)

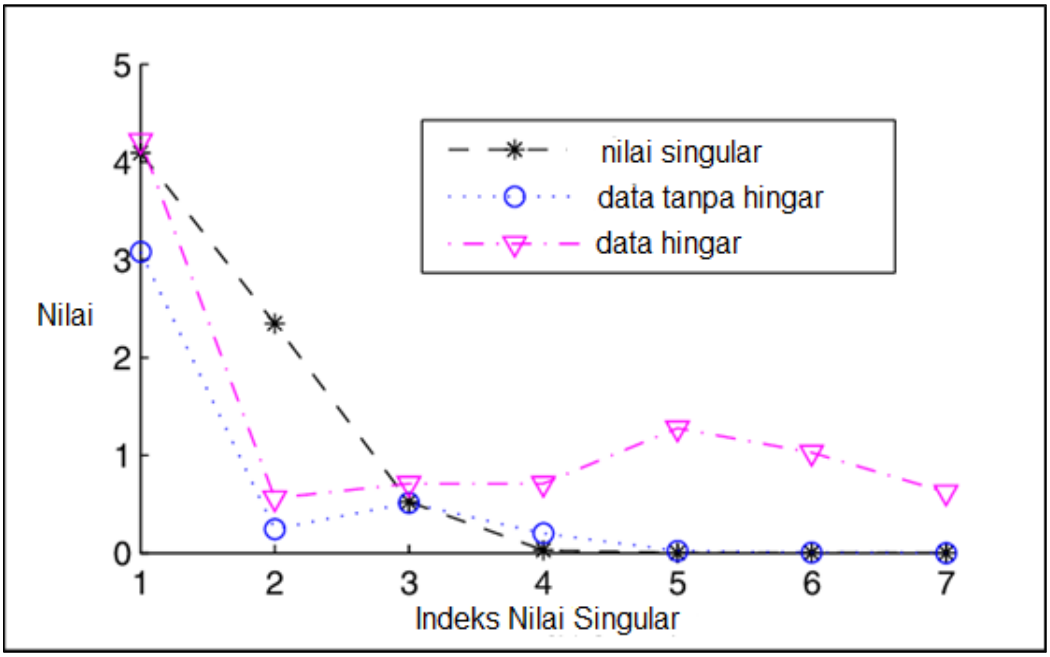
2.3.3 Data Hingar *(noisy data)*

Data hingar adalah data yang tidak bermakna. Sebarang data yang telah diterima, disimpan, atau diubah sehingga tidak dapat dibaca atau digunakan oleh program yang mula-mula membuatnya akan dikatakan sebagai data hingar. Data hingar tidak diperlukan kerana meningkatkan jumlah ruang storan yang diperlukan dan juga boleh menjejaskan keputusan analisis data. Secara khususnya, data hingar pada data ekspresi gen adalah disebabkan oleh sifat ketidaktentuan atau sifat rawak dari tindak balas biokimia terhadap ungkapan gen (Raser and O'shea 2005). Data biologi hingar mungkin disebabkan oleh ralat yang berlaku semasa pengumpulan data, seperti pencemaran dalam sampel makmal. Bagi data ekspresi gen, seringkali peralatan adalah penyebab data hingar (Libralon, de Leon Ferreira et al. 2009).

Menyedari keperluan menangani data hingar, (Ishigami & Furukawa 2018) mengusulkan masa-domain penapis pengurangan hingar Gaussian *(time-domain Gaussian-weighted noise reduction filter)* dan Extended Kalman Filter (Karunasingha & Liong 2018). Teknik yang digunakan untuk mengkaji keberkesanan pengurang

hingar dilakukan menggunakan Pearsons chi-kuadrat *(Pearsons chi- squared)* (Ishigami & Furukawa 2018) dan penngukuran statistic seperti entropi, min, varians, ralat mean persegi *(mean squared error)* (Sifakis et al. 2012). Rajah 2.15 menunjukkan contoh data hingar yang berbeza dengan data tanpa hingar. Sekiranya sesuatu kajian itu ingin mengkaji samada model yang diusulkan boleh menangani hingar, data sintetik akan ditambahkan dengan kesan hingar pada peratusan yang diingini. Semakin tinggi peratusan yang ditetapkan bermakna kesan hingarnya adalah pada tahap tinggi dan semakin sukar untuk ditangani. GeneNetWeaver mempunyai fungsi penambahan kesan hingar pada data sintetik yang dihasilkannya.

Algoritma Pembelajaran Mesin telah berjaya digunakan dalam analisis data ekspresi gen. Walaupun banyak algoritma Pembelajaran Mesin yang dapat menangani hingar, keupayaan mengesan dan membuang hingar dari set data latihan pada peringkat awal dapat membantu model mencatatkan tahap keefisyenan yang lebih tinggi. (Libralon, de Leon Ferreira et al. 2009) mengusulkan penggunaan teknik pra-pemprosesan jarak jauh *(distance-based pre-processing techniques)* untuk pengesanan hingar dalam data ekspresi gen. Dibuktikan bahawa, kelebihan kaedah yang diusulkannya ialah dapat mengurangkan masa pemprosesan data latihan kerana kompleksiti set data telah dikurangkan selepas pemprosesan awal data.



Rajah 2.15. Data hingar (Source: (Schiffermuller & Jungnickel 2006))

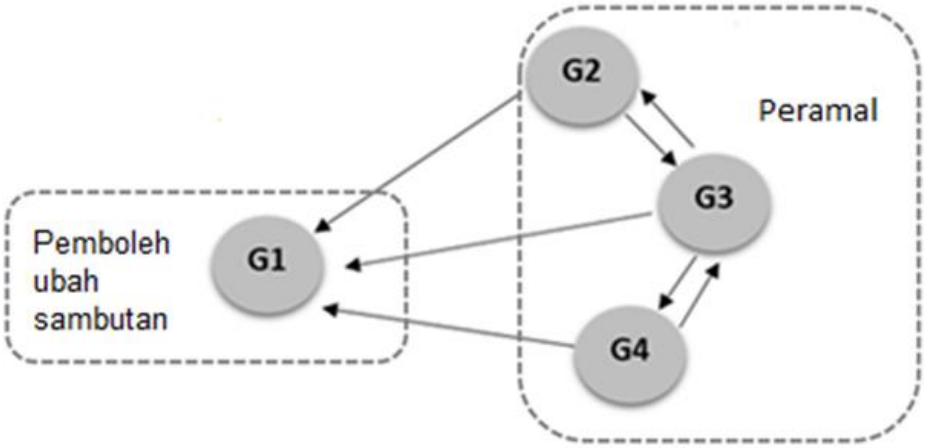
2.4 KAJIAN PELBAGAI BIDANG PEMBINAAN SEMULA RANGKAIAN PENGAWAL ATUR GEN

Pembinaan semula RPAG adalah merupakan bidang kajian yang aktif dijalankan dalam bidang bioinformatik, sering berevolusi dan hasil kajiannya sering mendapat perhatian penyelidik dan sering mengalami penambahbaikan dari masa ke semasa. Model pembinaan semula RPAG dibahagikan kepada 8 kategori model iaitu teori maklumat, berasaskan tapisan, berasaskan algoritma berinspirasikan alam, berasaskan analisis regresi, kebarangkalian dan statistik, kolerasi dan kesaling bergantungan, meta-algoritma dan pembelajaran mesin. Kategori model yang dikaji dikenalpasti melalui kajian kesusasteraan pembinaan semula RPAG.

2.4.1 Kajian Model Berasaskan Analisis Regresi

Penjelasan makna bagi setiap istilah yang digunakan dalam tesis diterangkan pada bahagian ini. Pembolehubah bersandar juga dikenali sebagai pembolehubah sambutan (*response variables*) atau pembolehubah sasaran. Pembolehubah tak bersandar pula dikenali sebagai regresor atau peramal (Montgomery et al. 2012). Bagi memastikan kekonsistenan tesis, istilah pembolehubah sambutan dan peramal digunakan pada keseluruhan manuskrip. Senario RLB dalam konteks RPAG digambarkan dalam Rajah 2.16. RPAG adalah senario di mana pembolehubah peramal berkemungkinan untuk mempunyai kaitan antara satu sama lain dan boleh mempengaruhi pembolehubah sambutan. Selain itu, persoalan seperti bagaimana untuk mengenalpasti pembolehubah yang berkaitan dan berapa besar peranan yang dimainkan oleh setiap pembolehubah boleh diselesaikan menggunakan analisis regresi.

Analisis regresi merupakan teknik memodelkan hubungan antara dua atau lebih pembolehubah (Miles & Shevlin 2001). Analisis regresi telah digunakan secara meluas untuk melakukan ramalan dan dijangkakan tahap penggunaannya semakin meningkat seperti pembelajaran mesin (Shi et al. 2016). Antara model berasaskan regresi terkini adalah RLB, APK, KDTS (Partial Least Squares Regression), Regresi Mesin Vektor Sokongan (Support Vector Machines Regression) dan L-Kuasa Dua Terkecil Separa (L-PLS Regression) (CAMO 2003).



Rajah 2.16 RLB dalam konteks RPAG

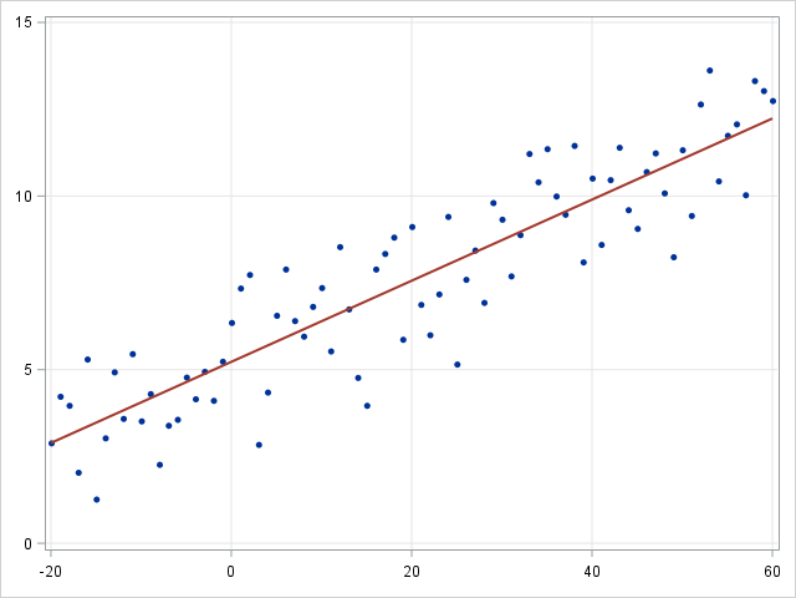
Walaupun teknik analisis regresi telah lama dipraktikkan pada domain lain, penggunaannya terhadap model pembinaan semula RPAG dilihat semakin meningkat sejak beberapa tahun kebelakangan ini (Bayar et al. 2017; Cheng & Sun 2017; Ferreira et al. 2017; Liley et al. 2017). Analisis regresi seperti yang diketahui umum adalah berasaskan matematik kompleks, lalu mengambil masa untuk diaplikasi pada mana-mana domain. Kini, dengan penambahbaikan yang banyak dilakukan pada teknik analisis data, penggunaan analisis regresi telah dipermudahkan sebahagian daripada prosesnya, walaupun tidak sepenuhnya. Model regresi linear terdiri daripada bahagian berketentuan dan bahagian rawak, secara amnya ditakrifkan sebagai:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2.1}$$

Bahagian berketentuan model,

$$\beta_0 + \beta_1 x \tag{2.2}$$

menyatakan bahawa bagi sebarang nilai pembolehubah tak bersandar, x , min populasi pembolehubah bersandar, y , digambarkan oleh fungsi garis lurus $\beta_0 + \beta_1 x$. ϵ ialah ralat *(error)* dan β adalah lereng *(slope).* Rajah 2.17 menunjukkan regresi linear.



Rajah 2.17. Regresi Linear

Mengenal pasti regresi sebagai salah satu model untuk diuji bagi menyelesaikan ralat lata hanyalah satu titik permulaan kepada keseluruhan proses kajian. Ini adalah kerana terdapat banyak lagi langkah ke hadapan yang perlu diambil, di mana salah satu daripadanya adalah mengenalpasti model spesifik regresi manakah yang perlu dipilih. Malah, terdapat lebih daripada 10 model berasaskan regresi yang perlu dipertimbangkan. Analisis regresi dikategorikan berdasarkan jenis pembolehubah yang boleh dikendalikan oleh setiap model regresi. Pengkategorian dibuat berdasarkan kajian literatur analisis regresi (Altland 1999; Panik 2009; Montgomery et al. 2012). Penentuan model regresi yang sesuai untuk digunakan adalah sangat bergantung kepada jenis data RPAG kerana setiap kajian objektif kajian yang berbeza.

Terdapat beberapa panduan penggunaan setiap daripada teknik regresi ini. Jadual 2.7 dibentangkan oleh MathWorks (2015) digunakan bagi mengenalpasti model regresi yang bersesuaian, di mana sel berwarna kelabu menunjukkan jenis pembolehubah yang sesuai dengan sifat RPAG dan model yang diusulkan ditanda dengan dua asterisk (**). (Zhang et al. 2010) (Zhang et al. 2013) mengandaikan bahawa hubungan antara gen ditakrifkan dengan model linear berdasarkan interaksi linear antara pembolehubah bersandar dan peramal; samada secara selanjar atau terhad. Kaedah selanjar dikenalpasti untuk digunakan dalam kajian ini. Berbilang pembolehubah adalah suatu keadaan di mana regresi multivariat perlu diaplikasikan. Bagi pembolehubah jenis tak bersandar, jenis selanjar adalah lebih bersesuaian.

Jadual 2.7 Pilihan model berasaskan regresi (MathWorks 2015)

	SIFAT DATA							JENIS MODEL YANG DIPILIH	MODEL YANG DIUSULKAN
	PEMBOLEHUBAH TAK BERSANDAR (PERAMAL)		PEMBOLEHUBAH BERSANDAR (PEMBOLEHUBAH SAMBUTAN)			MODEL			
	Selanjar	Kategori	Selanjar	Terhad	Berbilang Pembolehubah	Linear	Tidak-linear		
1.	✓ (ATAU)	✓ (ATAU)	✓			✓		Model suaian pekali	** Regresi linear
2.	✓ (ATAU)	✓ (ATAU)	✓			✓		Model suaian dan pekali suaian	** Regresi berperingkat
3.	✓ (ATAU)	✓ (ATAU)		✓		✓ (teritlak)		Model linear pekali suaian teritlak	Model linear teritlak
4.	✓		✓ (tak-linear)				✓	Model tak-linear pekali suaian	** Regresi tak-linear
5.	✓		✓			✓		Ridge/ LASSO/ Regresi Jaringan Anjal	** Ridge/ LASSO/ Regresi jaringan anjal
6.	✓ (kolerasi)		✓			✓		Model suaian dan pekali suaian	** Kuasa dua terkecil separa
7.	✓ (ATAU)	✓ (ATAU)	✓					Model tak-berparameter	** Pepohon klasifikasi, Pepohon regresi, Kaedah ensembel
8.		✓						ANOVA	ANOVA
9.	✓				✓	✓		Model regresi multivariat pekali suaian	
10.	✓		✓		✓	✓		Model pekali suaian campuran kesan	Model kesan-campuran

Beberapa kajian terbaik daripada sumber berkualiti dan bertahap tinggi telah dipilih untuk dibincangkan pada topik spesifik kajian analisis regresi ini. Oleh kerana analisis regresi melibatkan banyak teknik, dapat dilihat bahawa kajian-kajian yang dibincangkan ini menggunakan pelbagai teknik regresi yang berbeza (Das et al. 2010). Penentuan model yang tepat apabila menggunakan analisis regresi adalah penting dalam pembinaan semula RPAG (Freund & Wilson 1998). Fakta ini turut disokong oleh (Geeven et al. 2012) yang menyatakan kejayaan kaedah berasaskan regresi untuk untuk memodelkan ekspresi gen dan data jujukan DNA bergantung kepada pemilihan jenis model dan peramal yang digunakan sebagai input.

(Geeven et al. 2012) mengusulkan GEMULA yang mempunyai kaedah empat peringkat berasaskan LASSO dan digunakan untuk mengenalpasti dan mengutamakan interaksi sinergistik sesama peramal. (Chan et al. 2012) mengusulkan model baru mengenalpasti gen menggunakan model Kuasa Dua Terkecil Separa Ridge. Permasalahan dalam membuat anggaran dibaiki dengan menggabungkan Kuasa Dua Terkecil Separa Ridge dengan penyingkiran ciri berulang *(recursive feature elimination)* dan ralat Brier menggunakan pengesahan silang dua-bersarang *(two nested cross-validation)* (Chan et al. 2012). Model Ridge semakin mendapat perhatian daripada penyelidik berdasarkan keupayaannya dalam menangani permasalah berkaitan multikolineariti (Panik 2009). Cara menangani isu multikolinearan yang berpotensi untuk diuji; (1) mengumpul data tambahan, (2) model dispesifikasi semula *(respecification)*, dan (3) menggunakan Regresi Ridge (Montgomery et al. 2012). (Freund & Wilson 1998) mensarankan tiga pendekatan utama bagi kaedah pemulihan, iaitu (1) pemilihan pembolehubah, (2) mentakrif pembolehubah, dan (3) anggaran bias. Dalam konteks kajian ini, dipercayai bahawa antara semua model pemulihan, penggunaan Regresi Ridge untuk menggantikan kuasa dua terkecil (model pemadanan yang sedang digunakan) adalah berbaloi untuk dicuba kerana ia mengurangkan varians besar akibat daripada permasalahan data yang berdimensi tinggi dan bilangan sampel yang kecil (Chan et al. 2012). Ini disokong oleh (Tenenhaus et al. 2010) yang mencadangkan bukan sahaja Regresi Ridge tetapi beberapa model tambahan lain seperti Regresi Komponen Utama dan Lasso hendaklah diuji juga.

Salah satu perkara utama yang perlu diambilkira dalam mengaplikasikan analisis regresi ialah keperluan menangani bilangan pemerhatian terhad dan data berdimensi tinggi atau istilah khasnya *'curse of dimensionality'* (Chan et al. 2012). '*Curse of dimensionality'* ialah suatu keadaan yang melibatkan penggunaan data berdimensi tinggi yang menyukarkan proses pengiraan dilakukan. Dalam konteks ini, contoh data berdimensi tinggi ialah, katakan satu eksperimen gen ekspresi dijalankan menggunakan mikrotatasusunan, lalu datanya dibahagikan kepada dua kumpulan (kes dan kawalan) dengan 4 replikasi untuk setiap satu kumpulan. Setiap sampel dalam kumpulan tersebut akan mempunyai ribuan gen ekspresi yang didapati dari pengujian yang dilakukan berulang kali. (Chan et al. 2015) menyatakan bahawa bilangan sampel rendah yang terdapat pada setiap titik masa data yang diproses adalah salah satu isu utama yang perlu ditangani.

(Küffner et al. 2012) memberi penekanan pada keupayaan ANOVA bagi diaplikasikan pada data ekspresi gen tak linear tanpa perlu melakukan proses pendiskretan. Pendiskretan ialah proses yang digunakan untuk menukar persamaan selanjar ke dalam bentuk yang boleh digunakan untuk mengira penyelesaian berangka. (Gregoretti et al. 2010) mensasarkan untuk memperbaiki ketepatan ramalan jaringan bersaiz besar. Kajian ini menggunakan RLB dengan mengaplikasikan teknik pemprosesan selari. Namun demikian, kajian ini dijalankan ke atas semua set data yang sudah berada dalam keadaan ideal, iaitu jumlah pemerhatian yang tidak melebihi jumlah gen. Sebagai contoh, 1000 gen X 1000 eksperimen usikan *(perturbation)* dan 100 gen X 100 eksperimen usikan. Pemilihan data sebegini dilakukan mungkin kerana objektif eksperimen tersebut yang mahu memberi lebih fokus kepada keberkesanan pemproses selari yang diusulkan bagi memproses RPAG bersaiz besar yang terdiri daripada jutaan gen. Selain daripada (Tenenhaus et al. 2010) dan (Küffner et al. 2012), kesemua kajian yang diulas dalam bab ini tidak membincangkan secara khusus tentang bagaimana isu ralat lata ditangani.

(Singh & Vidyasagar 2016) mengutarakan pendapat yang berbeza dari kebanyakan kajian yang mengaplikasi model berasaskan analisis regresi. (Singh & Vidyasagar 2016) berpendapat andaian bahawa semua mekanisma peraturan RPAG adalah sama sebenarnya bertentangan dengan keadaan sebenar di mana terdapat

sesetengah RPAG yang mempunyai interaksi bukan linear. Oleh kerana belum terdapat kajian analisis regresi diusulkan untuk menampung kepelbagaian jenis mekanisma pengawal atur dalam RPAG, (Singh & Vidyasagar 2016) berfungsi dengan beberapa sub-jaringan yang terdiri daripada gen sasaran dan gen pengawalselia (*regulator*) diekstrak dari GRN secara rawak. Regresi Berperingkat Ke depan and Regresi 'Stagewise' dijalankan ke atas setiap sub-jaringan ini dengan menggunakan 5 fungsi berlainan iaitu linear, kubik, logistik kawalselia-atas (*logistic up-regulation*), logistik kawalselia-bawah (*logistics down-regulation*) dan punca kuasa dua (*square-root*). Gen pengawalselia berpemberat tinggi yang dihasilan dari setiap operasi digabungkan sebagai sekumpulan hasil akhir gen pengawalselia yang dilihat paling berpotensi untuk satu gen sasaran tersebut. Hasil akhir gen pengawalselia ini melalui proses skor berasaskan kawasan (*area based scoring*) bagi mengenalpasti gen pengawalselia terbaik untuk dipadankan dengan gen sasaran. Proses ini diulang ke atas setiap gen sasaran. Kekurangan bLARS adalah proses pengiraan yang berulang perlu dilakukan bagi mendapatkan hasil terbaik. Kebaikan bLARS adalah tidak memaksa setiap gen sasaran untuk diuji dengan setiap gen pengawalselia, sebaliknya gen sasaran diuji hubungannya dengan gen pengawalselia yang berada dalam sub-jaringan yang berbeza. Kaedah ini lebih menyerupai keadaan GRN sebenar yang mempunyai tahap ketidaktentuan yang tinggi.

a. Regresi Linear Berbilang (RLB)

Analisa multivariat semakin sering digunakan sejak akhir-akhir ini kerana kemampuannya dalam analisa pengkomputeran. Bahagian ini memperkenalkan konsep asas, prosedur berkaitan analisa multivariat dan menerangkan bagaimana analisa multivariat digunakan dalam penyelidikan sistem biologi. Analisa multivariat adalah bidang multidisiplin yang berkait dengan komputer dan statistik serta digemari oleh saintis data untuk mengeksploitasi maklumat yang tersembunyi pada set data. Ciri-ciri data pada masa kini yang dimensi tinggi, kompleks atau tidak berstruktur, tidak lengkap, hingar, dan mengandungi ralat telah tidak lagi sesuai bagi pendekatan statistik tradisi tradisional yang kebanyakannya direka untuk menganalisis sampel yang agak kecil. Sistem biologi masa kini boleh jadi sangat rumit sehingga tidak dapat digambarkan dengan tepat oleh kaedah statistik tradisional seperti regresi linear dan analisis statistik