國立臺灣師範大學生命科學系碩士論文

AryNet-以微陣列數據進行基因網絡視覺化的方式來比較化學物質與精神病關聯性之網路應用系統

AryNet- Web App for detecting the relationships between Mental Diseases and Chemicals through Gene Network Analysis based on Microarray data

研究生： 宋鴻青

Hung-Ching Sung

指導教授：沈林琥 博士

Sher Singh, Ph.D.

中華民國 105 年一月

致謝

過去大學、當兵服役期間，就不斷嚮往有朝一日可以參與學術研究，很感謝台灣師範大學生科系提供就學機會，讓我可以在碩士班生涯中圓夢實習。

由於缺乏研究經驗，初進實驗室時在修課、閱讀文獻過程中，常弄到一頭霧水難以找到研究方向。感謝指導教授沈林琥老師諸多教誨，老師時常一邊給予鼓勵、關鍵的方向指導，在我散漫疏失時又從不吝於嚴加提醒，這次研究過程中需要的一些程式撰寫、網路資源蒐集技術，老師都費盡心思幫助我找到最棒的學習資源。碩士研究過程中，我改正了不少基本的工作態度缺失、壞習慣，真的非常感謝沈老師的用心引導。

本論文承蒙口試委員俞松良博士和蘇家玉博士在百忙中抽空參加，感謝他們細心指證與寶貴意見，使論文能更趨完善，在此致上誠摯謝意。

感謝學弟育興耐心教導程式語言，並且協助我熟悉研究環境，在我成`撰寫遇到瓶頸，癱坐在地不知該如何是好時，真的很感謝學弟每一次即時出手相助，沒有這些幫助我根本無法完成這次研究。

感謝王昀珩學長提供我實驗中各種硬體、軟體支援，並且在我遭遇挫折心情沮喪時給予極大的關心與鼓勵，讓我能堅持下去。感謝國安、擎天、丹瑜伴我度過實驗室生活，為冰冷這冰冷的冷氣房添加不少溫暖的的風趣與歡笑。

感謝爸爸、媽媽，在自己經歷經濟困頓、病痛折磨時，仍然苦撐著生活支持我讀研究所，儘管我能力不足而延長修業傾家盪產，爸媽從來沒有任何怨言，

只有安慰和鼓勵，父母恩情我畢生難忘。感謝妹妹鴻蓮時常對我分享研究生活注意要項，讓原本迷迷糊糊的我仍能在各樣複雜行政程序中進入狀況。

感謝愛妻莉穎任勞任怨的陪伴與支持，一個人賺錢承擔兩人生活，還得時常聽我吐盡苦水，在我難過、氣憤時，妳的每一次相伴和忍耐是我最甜蜜的力量泉源。

最後，我真摯向遠在韓國用盡眼淚禱告和心情血汗來陪伴、守護我一生的鄭約書亞總裁牧師獻上最高致謝，真的很感謝您在這段日子中給我的所有幫助，和永遠的不離不棄，日後生涯中我都會紀念這份愛，並且把愛傳出去。

碩士生涯終於要告一段落，回首兩年多中自己的成長、學習，我真的對大家萬分感謝，將來不管身處何處，我一生中都不會忘記這段大家陪我走過的甜苦歷程，望將來能報效國家社會，不辜負上天透過所有人對我的付出。

民國一〇五年二月

宋鴻青於 師大生科系

摘要

生物晶片研究因為擁有再現性、樣本重複性，可藉由大量數據彙整進行統計分析來進行基因體交互作用相關研究。現今關於生物晶片數據分析與視覺化的平台，除市面上付費軟體外，網路上開源相關軟體套件和網站系統亦有提供一些免費資源，但這些免費平台大多容易因為功能限制而造成不便。

本研究針對幾種生物晶片平台建立一個MVC模式資訊整合資料庫--AryNet，資料庫收集來自Reactome的基因轉譯後蛋白質交互作用資訊，以及自GEO下載的常見精神疾病、神經退化疾病以及環境化學物質的相關晶片數據，特別是「基因表現」、「核糖核酸甲基化」兩種類型的晶片原始數據。資料庫後端透過R-Engine配置多種Bioconductor所提供的數據標準化、統計校正演算法功能。系統將這些演算法相關參數以JavaScript方式呈現於網頁上供使用者自由調整，並以互動式基因網絡圖形將分析結果即時回傳使用者。

我們使用AryNet收錄資料庫統計躁鬱症、思覺失調症、重度憂鬱症等疾病的可能相關基因，然後結合基因調控網絡資訊計算這些疾病與幾種環境化學物質造成相似外表型的可能相關性。

關鍵詞：生物晶片 、基因網絡、環境化學物質、精神病

Abstract

As the Internet develops, microarray data shows a powerful potential for detecting the novel genomic interaction with its powerful repeatability and reproducibility. There are a lot of web tools and open source packages for analysis and visualization of microarray data. But most of them are costly or too technical to use.

For the purpose of developing a tool to operate data processing, statistics, genomic comparison, and visualization on microarray data in a friendly usage. We built a web system named AryNet. Applying the MVC framework with an interactive controlling panel on web page, AryNet possesses a SQL database storing epigenetic and gene expression profiles of microarray data downloaded from GEO database including samples with mental diseases, neural diseases, and chemical exposure. The information of protein-protein interactions from Reactome is also installed. The Java-based controller was armed with a plugin named R-Engine to drive the R-package of data processing and statistics obtain from Bioconductor. The resulting analysis will be retrieved back to the user view and generates a gene networks diagram on time.

We obtain the differential expression genes (DEGs) profiles from bipolar disorder, schizophrenia, major depression and chemical exposure by AryNet. And then we generated the gene network diagrams combining the DEGs with genes which have relationships in protein-protein interaction of each profiles. Comparison with the gene networks shows that there might be some resemblance between the phenotypes of diseases and endocrine disruptors.

Keywords:
biochip, microarray, gene network, chemical, mental disorder

目錄

圖目錄

表目錄

1 前言

1.1 研究動機

身為萬物之靈，人類最引以為傲的就是凌駕於所有生物之上的智慧大腦，在一定的秩序主管能力下，人腦跳躍思考、解決複雜問題的能力，成為每個人一生中適應各種生活環境的關鍵利器。然而精神病患在發病過程中喪失特定心智控管功能、大腦失去對外部訊息判讀、整頓能力，自古以來飽受無法正常交際、適應環境之苦，其中雙急性躁鬱症病人甚至因此擁有高達超過 20%的自殺率[1]。隨著科技日益革新，新儀器、新知識、以及複雜演算法研發，如今人類才得已擁有能力整合腦神經科學以及心理學，進一步進行精神疾病研究[2]。

針對思覺失調症以及躁鬱症，許多研究報告指出這些疾病兼具遺傳性[3]及環境誘導性[4] [5] [6]，且部分基因甲基化比例在病患體內有較一般健康對照組異常現象[7] [8]，加上患者常伴有內分泌變異的生理症狀[9]，於是有學者提出「部分精神疾病的致病機轉可能跟表觀遺傳有關」[10]。

研究指出雙酚 A、壬基酚等環境賀爾蒙在特定濃度下亦會透過表觀遺傳機制對人體內分泌造成干擾[11]，進而可能導致細胞癌化機率增高[12]，也有研究指出一些被甲基化的基因甚至影響腦神經發育而可能造成類似精神病變的個體行為變異[13] [14]。

無論是關於精神疾病或是環境賀爾蒙研究，皆已有不少生物晶片實驗數據收錄於免費網路資源中，可供進一步基因體關聯性研究。但因為精神病大多仍

無法確知其致病機轉，疾病分類上只能以臨床症狀做粗略樹狀分類，以至於在直接使用這些數據進行統計分析時，常因為診斷訊息不明確而出現嚴重數據雜訊[15]。為了克服這點，數據分析時還需要引用一些複雜的內部分群技術來分析雜訊。

本研究整合基因網絡關聯性計算、蛋白質交互作用資訊、多種生物晶片數據標準化及自動分群演算法，架設一個易於線上操作的生物晶片資訊視覺化系統，並內建收錄精神疾病與環境化學物質原始數據的資料庫，希望建立一個便於操作的「精神疾病與環境化學物質可能關聯性」網路研究工具。

1.2 研究目的

基於前述研究動機，本研究主要有以下研究目的：

A. 設計一套能夠簡易操作的生物晶片數據關聯性的視覺化工具，其中尤其著重於基因網絡分析操作方面的功能。

B. 收錄整理精神疾病與環境化學物質實驗數據，並嘗試從基因表現及甲基化相似度評估兩者可能關聯性。

2 文獻回顧表觀遺傳(Epigenetics)

過去人們對「遺傳」的概念建在親代對子代細胞中核糖核酸序列的傳承，認為生物體在透過親代遺傳獲得所屬基因型，然後展開生命週期時，除了強烈環境刺激造成其配子細胞 DNA 序列突變之外，大部分其個體行為結果都是自行承擔，無法透過「遺傳」影響下一代。在這樣的知識背景下，過去醫學在看待「遺傳性疾病」時亦定義為「由親代攜行突變基因透過繁衍過程遺傳給下一代的疾病」，有別於一般病原體接觸或環境傷害造成的疾病。因此，過去認為除非是妊娠期間透過母體垂直感染，或者是環境因子嚴重到改變了生殖細胞遺傳基因序列，不然親代所攜行的傳染性疾病、個體生理變化，皆無法透過遺傳影響下一代。

然而，1942 年英國生物學家 C. H. Waddington 於著作中取「後生論」(epigenesis)與「遺傳」(genetic)兩詞做複合，提出「表觀遺傳學」(epigenetics) (圖 一)，再次顛覆了傳統遺傳學界教條觀念[16]，生命體在其生命週期中的個別行為、環境互動，不只會影響自身生理變化，這些變化甚至可能透過「組織蛋白修飾」、「DNA 甲基化」、「microRNA」三種方式記錄於配子細胞基因訊號中，而在「DNA 序列」沒有改變的情況下進而影響其子代[17] [18]。

表觀遺傳說發表後開啟一系列相關研究，一些過去被認為並未涉及基因突變而只能由外在環境感染、誘導的疾病，近來也在研究數據中被發現跟表觀遺傳有關而可能傳承給下一代，這種結果看起來有點像「母體垂直感染疾病」的

世代傳承機制，不需要任何病原體參與，完全出自於生物體自身存在機能，而且這種透過配子攜行帶給下一代的變異訊息可來自父母雙方[19]。

表觀遺傳學概念發表後，對於「遺傳」與「環境誘導」兩種生物行為因子的分野開始漸漸趨向模糊[20]，除了癌症之外，目前還有一些行為學[21](圖 三)、化學物質毒害[22] (圖 二)等議題，都陸續被加入相關研究中。

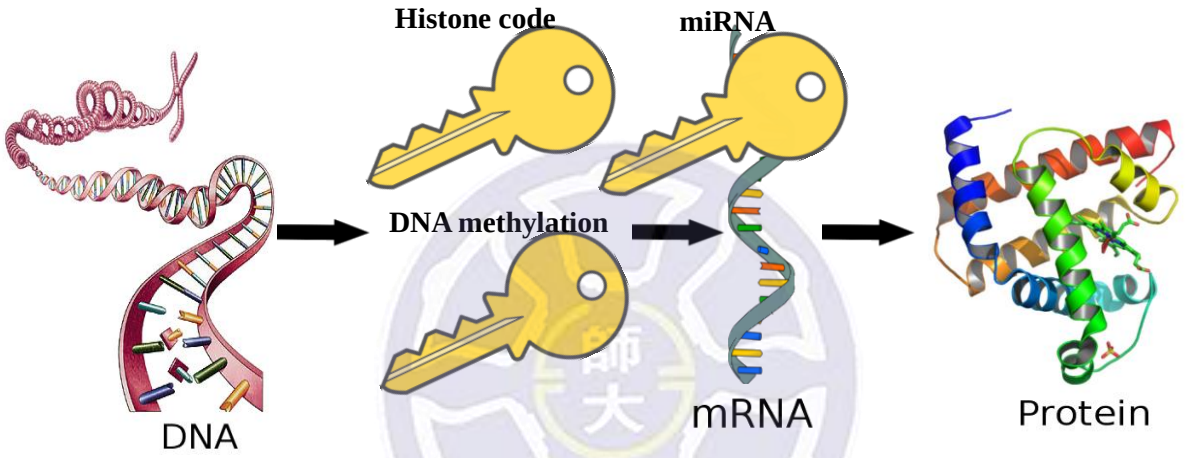

圖 一表觀遺傳學說明

泛指探討基因在轉錄、轉譯過程中，以DNA組織蛋白、DNA甲基化、microRNA三種途徑來調控基因最終表現量的遺傳分子生物學，狹義表觀遺傳學則指部分學者認為這些變異可以藉由配子攜行造成遺傳變異。

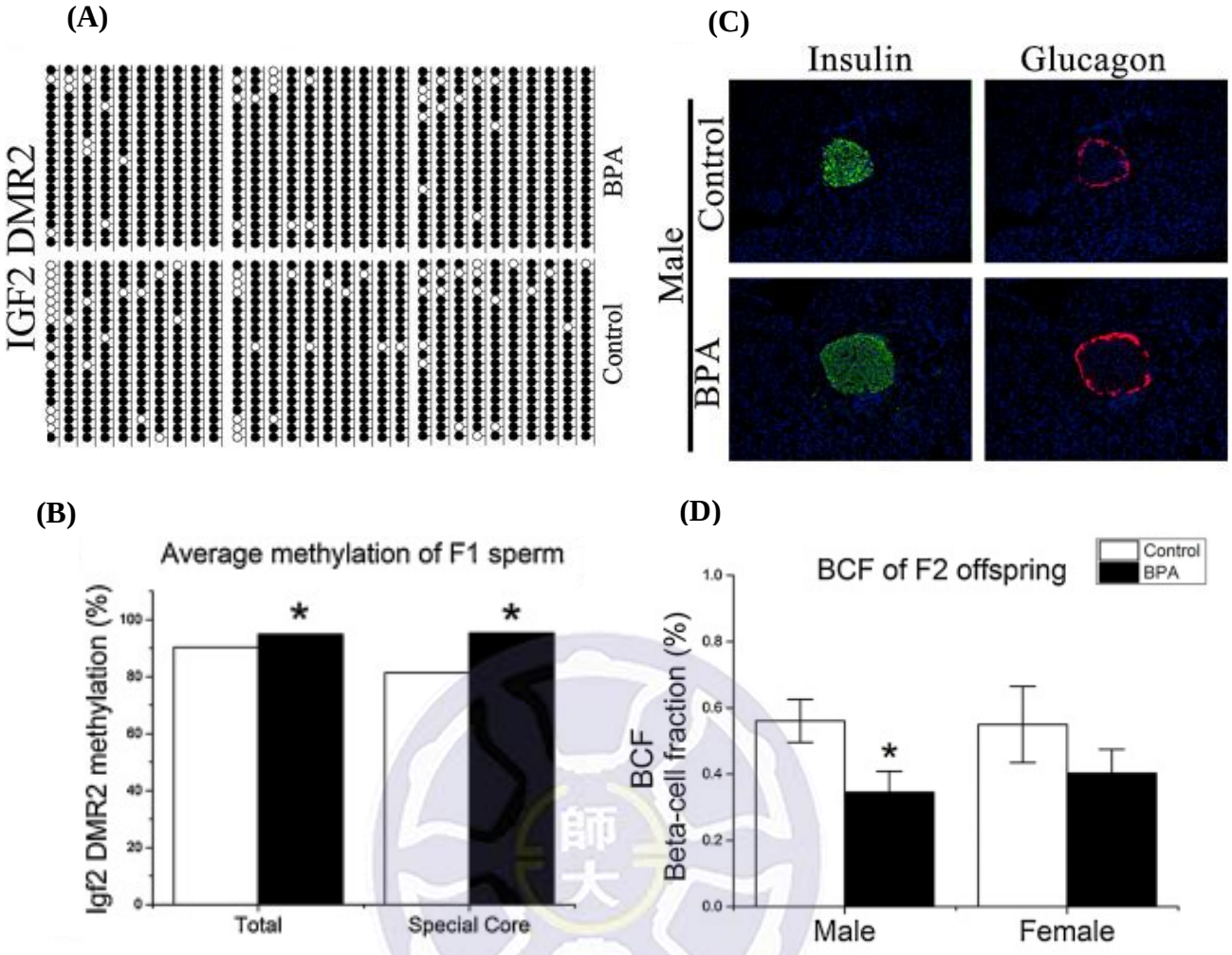

圖 二 暴露雙酚A的雄鼠可能透過精子中IGF2過度甲基化影響其雄性子代

(A)(B)為親代精子IGF2基因在DNA上甲基化程度比較，暴露於雙酚A的雄鼠其配子擁有較高的甲基化比例

(C)(D)暴露BPA的雄鼠其子代具有胰臟中beta細胞所占比例相對較對照組小的遺傳變異

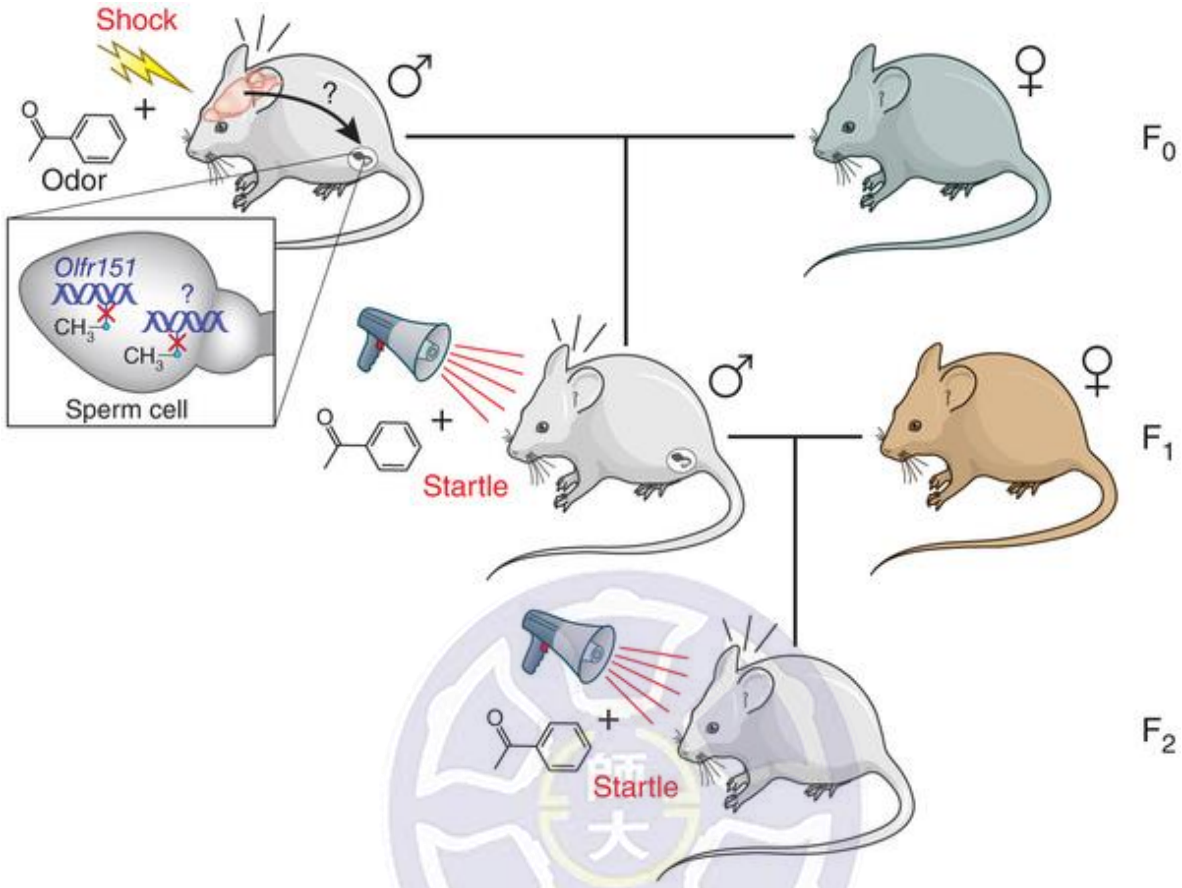

圖 三 恐怖記憶遺傳

Syzf M發現親代雄鼠透過後天訓練完成的恐怖記憶，可以透過親子教育以外的途徑造成其子代擁有一樣的恐怖記憶，Sxyf M推測可能跟表觀遺傳機制有關。

2.2 精神疾病(Mental Disorders)

「精神疾病」在美國醫學主題詞表(MeSH)中定義為「知覺、感覺、行為等調適機能障礙而造成的痛苦或傷害」，並被歸在 Disease 項目以外的「精神心理疾病學」一大項類中(MeSH ID:D001523)。常見精神疾病包含情感性疾患(mood disorders)、焦慮性疾患(anxiety disorder)、精神分裂症(schizophrenia)，其中情感性疾患又包括著名的憂鬱症、雙極性情感失調症(躁鬱症)。古時精神病被認為是一種人格心智混亂或是靈干擾現象[23]，直到十七世紀由 Felix Plater 等醫學家對精神病狀加以描繪並提出大腦生理相關理論後，精神醫學的基礎才漸漸奠定[24]。二十世紀中期隨著鋰鹽、第一代抗精神病藥的發現，人類開始能擁有更具系統性的醫學理論來對精神病進行控制與治療[25]。

精神病患常因為不同程度的異常行為，造成無法正常社會交際，不僅影響病患生活品質，有些精神疾病甚至導致攻擊性[26]、自殺行為[1]。其中尤以重度憂鬱症(Major depression)、躁鬱症(bipolar disorder)、思覺失調症(schizophrenia)患者自殺率最高[27]。

生理學實驗檢查證實不少精神病患在異常的外顯行為上，也都伴隨著部分內分泌量異常，尤其是一些跟情緒、大腦興奮有關的賀爾蒙途徑[9] [28]。然而是什麼致病機轉造成這些生理狀況進而影響行為，至今仍然無法確知，無論是遺傳或環境誘導理論，兩方皆有實驗數據支持。以躁鬱症為例，醫學統計發現躁鬱症病患常伴有家族病史，雙親中若有一人患病躁鬱症，子女患病機率明顯

高於沒有家族史的子女[29]，科學家也找到諸如 BDNF、DAOA、ANK3、DISC1 等基因上的突變率在病患與對照組身上有顯著差異[30]。但是臨床上卻也有一部份病患是沒有家族病史，反而是經歷重大環境變故或衝擊後而開始產生和具有家族史病患一樣的躁鬱症狀[31] [32]。另外，也有學者發現躁鬱症病患在一些基因位點上有程度較高的甲基化現象[8]。

因此漸漸有人認為，像躁鬱症、思覺失調症這類看似兼具遺傳及環境誘導性的精神疾病，其致病原因有可能跟表觀遺傳有關[10](圖 四)。

Heterochromatin **Euchromatin**

Residue	Mark/Effect	Region	Stress induction	Age of stress	Specific genes regulated	Refs.
General H3	Ac ↑	HPC	Maternal separation	Postnatal	global	80
	Ac ↓	NAc	Chronic mild stress in stress-sensitive mice, reversed by imipramine; human depression with medication	Adult	global; *CaMKIIa*	20, 30
	Ac ↑	PFC	Social stress	Adult	global	28
	Ac ↓	HPC	Social stress, LR/HR rats	Adult	global	23, 25, 27
	Ac ↑	HPC	Social stress	Adult	*Bdnf*	13
H3K27	2Me ↑↓	NAc	Broad changes with social stress, reversed by imipramine	Adult	global	32
	2Me ↑	HPC	Social stress		*Bdnf*	13
	3Me ↑	NAc	Social stress	Adult	*Rac1*	31
	3Me ↑	PFC	Human depression/suicide		*TRKB*	40, 41
	3Me ↓	HPC	restraint stress		global	33
	3Me ↓	PFC	Human depression with medication		*BDNF*	38
H3K14	Ac ↑	AMY	Social stress	Adult	global	25
	Ac ↑↓	HPC	Temporally modulated by social stress; regulation in LR/HR rats		global	23, 25
	Ac ↑↓	NAc	Temporally modulated by social stress		global	14
H3K9	Ac ↓	HPC	Low maternal care	Postnatal	*Nr3c1, Grm1, Gad1*	17, 81-83
	1Me ↓	HPC	Restraint stress	Adult	global	33
	2Me ↓	NAc	Social stress	Adult	global	22
	2Me ↑	NAc	Fluoxetine; human depression with medication		*Camkiia*	30
	3Me ↑	HPC	Restraint stress	Adult	global; transposable elements	33, 34
H3K4	2Me ↑	HPC	Social stress and imipramine	Adult	*Bdnf*	13
	3Me ↓	HPC	Low maternal care	Postnatal	*Grm1*	81
	3Me ↓	NAc, HPC	Social stress and chronic mild stress, reversed with imipramine	Adult	global	20, 33
	3Me ↑	PFC	Human depression/suicide	Adult	*SYN1, OAZ1*	36, 37
General H2B	Ac ↓	HPC	Regulation in LR/HR rats	Adult	global	23
General H4	Ac ↑	HPC	Maternal separation	Postnatal	global, *Arc, Egr1*	80
	Ac ↓	HPC	Regulation in LR/HR rats	Adult	global	23
H4K12	Ac ↑	forebrain	Maternal separation	Postnatal	global	79

圖 四 DNA 組織蛋白圖譜相關的精神病症案例以及抗精神藥物

2.3 環境賀爾蒙(Environmental Hormone)

1979 年由 McLachlan 所領導研究團隊發表一個小鼠實驗數據，提出某些廣泛散播在環境的化學物質，具有跟脊椎動物體內固醇類激素擁有類似的化學性質[33]，自此開啟了科學家們開始陸陸續續一系列的「環境內分泌干擾素」(Endocrine Disrupter)研究，文獻中常簡稱為 EDS，1997 年日本橫濱市立大學教授 Iguchi 提出「環境賀爾蒙」一詞，並更具體描述這群具備干擾賀爾蒙作用而造成生理危害的環境化學物質[34]。

這些化學物質大多為生活用品中人工合成材料或是副產物，廣泛存藏於生活環境中，雖然來自於外在環境，卻能在生物體內造成內分泌系統干擾，包括賀爾蒙的合成、分泌、傳輸、受體結合、半衰期等[35]。2013 年世界衛生組織公布環境干擾素研究報告列舉名單中，其中部分 EDS 甚至因為分子結構和生物體內賀爾蒙相仿，而能夠直接在生物體中引起本來應該由賀爾蒙來正常調控的生理反應，造成生物體在環境暴露下產生內分泌失調的異常生理現象[36]。

環境賀爾蒙彷彿一個隨意亂發布訊號的遠端遙控器一般，直接以外來因子方式干擾人體內本來應該由自身調控的生理核心訊息，因此容易造成生物體各種疾病危害，影響胎兒發育[37]、性腺發育[38]、免疫力降低[39]、致癌風險[12]、行為變異[40]。

生物資訊方面，美國國家醫學圖書館建立並歸檔的醫學圖書館標題表(Medical Subject Heading)在長期收錄各種人體健康相關化學物質資料時，其中

亦包含 WHO 所公布環境賀爾蒙名單中所有項目，成為後續研究的參照目標。比較毒理基因體學資料庫(Comparative Toxicogenomics Database, CTD)於 2015 年所更新系統中亦收錄超過 70 種環境賀爾蒙物質名單，併入其化學物質、基因、疾病交互關聯作用資料庫，這些資料庫提供了一些研究化學物質對人體可能健康影響的預測及剖析功能，成為目前生物資訊學毒物研究中廣泛引用的利器。幾例來說，2011 年 Singh 等人即引用 CTD 資料庫，藉由生物資訊方法發現磷苯二甲酸類(phthalates)對人體多種疾病具有潛在威脅，再次凸顯了現今生物資訊資源對環境荷爾蒙研究發展的潛在價值[41]。

近年有越來越多實驗證明各類 EDS 在干擾生物體內分泌作用中，亦包含了 miRNA 表現變化、基因甲基化程度增減等表觀遺傳相關現象。例如 2010 年 Elyakim 等人發現人體肝細胞中的 miR-191 會受到戴奧辛 Dioxin 調控而過量表現，進而提高肝癌風險[42]。2013 年 Kundakovic 從小鼠實驗中發現懷孕母鼠體內的雙酚 A(Bisphenol A)含量，會對子代腦部以及血液中部分基因甲基化程度造成影響，進而造成子代行為上與健康母鼠子代差異，而且這些差異還會受到子代性別影響[43]，加上這些被甲基化的基因當中，有些在人類精神疾病患者體內也有被甲基化的現象，Kundakovic 甚至大膽假設這類環境賀爾蒙透過表觀遺傳對小鼠造成的影響，也有可能和人類一些精神失調有關[14](圖 六)。

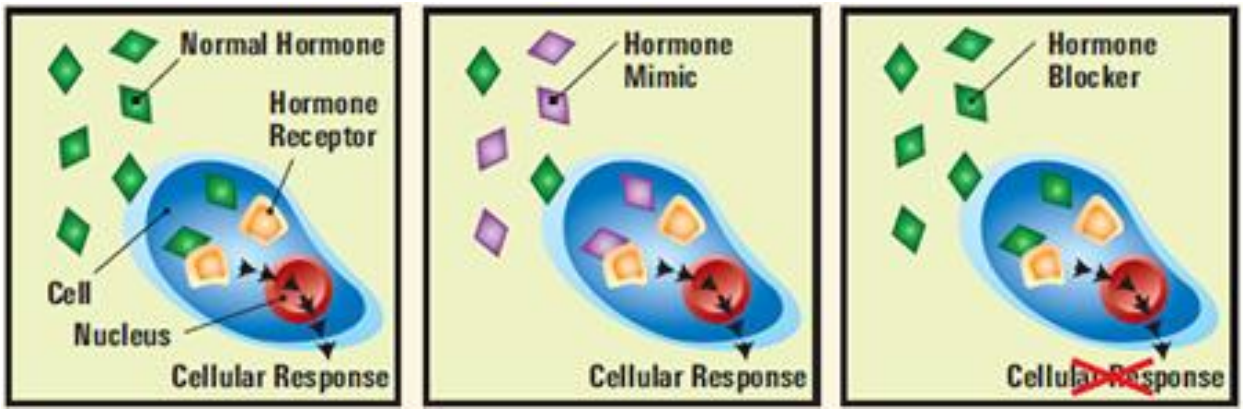

圖 五 環境賀爾蒙的作用機制之一

透過其相似於生物體內正常賀爾蒙的部分分子結構，參與並干擾細胞內訊息傳遞路徑，造成生物體內細胞訊息錯亂，而表現異常。

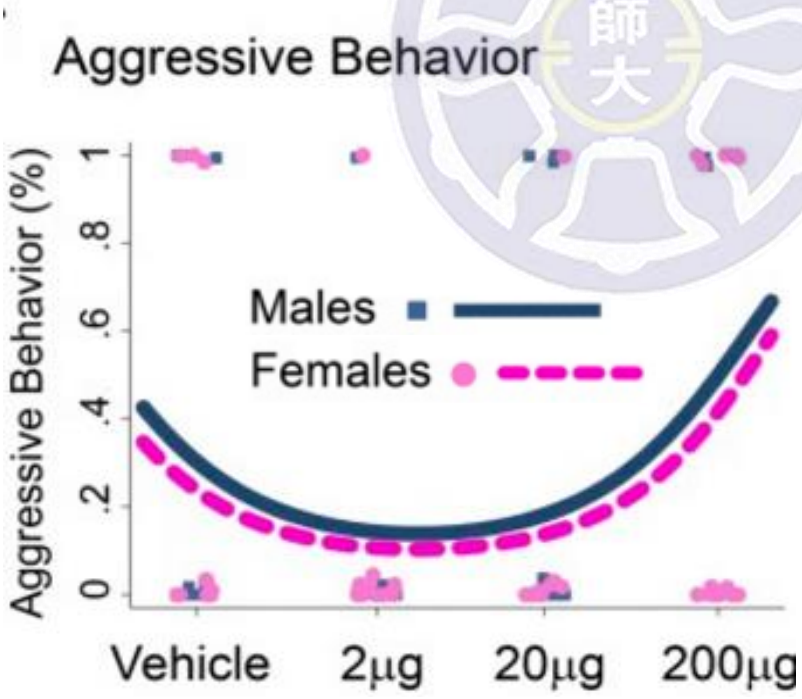

圖 六 母鼠懷孕期間暴露給予口服 BPA 劑量與其子代攻擊行為關係

文獻中還提到，這些子代鼠輩中，雄鼠體內部分基因亦有隨著母親 BPA 劑量濃度增加而過甲基化的現象。

2.4 生物途徑(Biological Pathway)

在後基因體時代，許多研究在處理複雜的疾病與生理問題時，一種常見的實驗方式就是比較實驗組與對照組的基因體或蛋白體狀態，找出表現量具顯著差異的基因或蛋白質[44]，這種實驗方法往往得到大量基因/蛋白質列表名單，但基因彼此關聯性強弱不一，彼此牽動機制可能無干，也有可能是多路徑牽動，甚至有些基因實際上是所探討因子的關鍵基因，卻因為在實驗控制變因下遭多重上游基因調控而無法通過顯著差異統計篩選。上述種種原因使得在分析這類實驗結果時時，可能獲得散亂無意義的資料而難以解讀。這時需要利用生物途徑資料來進行分門別類整理，並適度引入一些沒有顯著差異表現的基因[45]。

生物體是一個高度複雜性的化學反應系統，其中各樣可能只是出現片刻或長時穩定的化學物質彼此在分子層次上共同參與各式複雜化學反應，建立起多樣化的網狀關聯性。生物途徑就是將這些分子化合物間關聯性組合起來的網路。

日本京都大學建立的基因組百科全書(Kyoto Encyclopedia of Genes and Genomes,KEGG)的生物途徑資料庫中收集最新文獻紀錄的細胞內分子交互作用，並建立 479 張可執行元件搜尋的生物途徑網絡圖，供使用者免費查詢，以及付費下載資料庫[46]。另外也有免費提供下載資料庫的 Reactome 網站，內儲有經人工閱讀校正過的基因轉譯蛋白質交互作用生物途徑[47](圖 七)。

除了上述由特定研究團隊在維護的資料庫外，也有像 BioCarta 與

WikiPathways 這類由使用者形成社群共同建置的資料庫。BioCarta 為私人企業提供的無償生物途徑開發工具，提供研究人員上傳合乎格式規範的自行繪製生物途徑圖，並開放免費使用

(網址：http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways) [48]

WikiPathways 為荷蘭生物資訊中心(Netherlands Bioinformatics Centre,NBIC)資助建立，提供使用者可以在其中輕易搜尋特定生物途徑、相關基因，其由 Wikimedia 基金會提供的 WikiMedia 系統亦提供讓使用者可以看到這些生物途徑的修改紀錄[49]。

在分析高通量基因名單實驗結果時，使用這些生物途徑資料庫來搜尋名單中基因的可能相關生物途徑，可以幫助研究者找出可能的分子反應機制，也能方便比較基因名單中每個基因和操作因子的調控相關度。

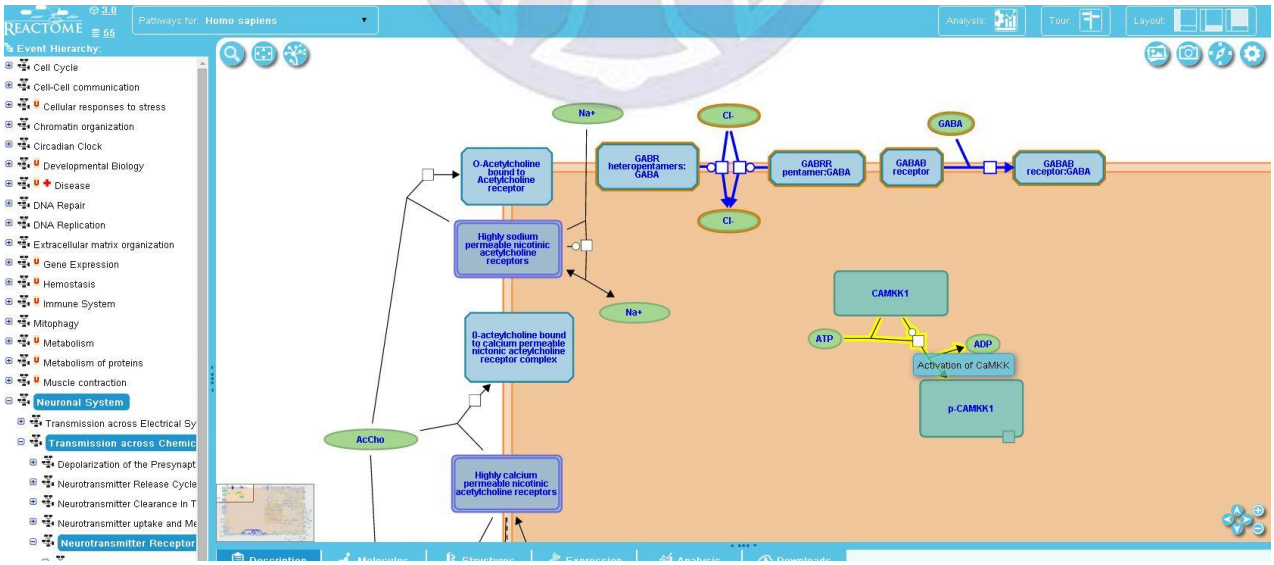

圖 七 Reactome 的網頁操作介面，

AryNet 目前收錄蛋白質交互作用資料即來自此網站資料庫

2.5 基因晶片 (Gene Microarray)

基因體研究可用以獲得更多交互作用的資料，藉由高通量篩選實驗流程，其實驗結果具備更全面性的細胞生理資料，亦成為各實驗室在進行篩藥、病理機制等研究時輔以縮小研究範圍的重要依據[50] [51]。

基因晶片是一種高通量並具有高度重複性的基因體研究平台。藉由晶片中核糖核酸探針雜合技術，使用者可於每張晶片實驗操作中快速獲得每個實驗樣本之大量基因表現資料，其優點主要有下列四個[52]：

1.樣本數高

2.同時分析大量基因表現

3.自動化分析

4.再現性高

然而由於這類技術高度自動化、規模化及微型化的特性，使其實驗結果具有極龐大的數據量和複雜的資料型態，再加上其實驗操作過程中產生的雜訊、系統及非系統變異等干擾因子，生物晶片的數據處理分析已超過傳統統計分析方法能及，因此統計學家們投入許多研究開發各種生物晶片平台的前置處理、數據校正演算法模型，以 Affymetrix 公司所出產的 Human Genome 系列基因表現晶片為例，除了 Affymetrix 公司本身所提供的 MAS5(Affymetrix's MicroArray Suite 5)全套演算法之外，尚有 RMA(Rubust Multi-array Analysis)、GCRMA(Guanine Cytosine RMA)、dCHIP 等常被引用的演算法可用來對晶片原

始數據做前置處理及校正，網路上很多探討這些演算法使用時機的文獻，需評估實驗情況來選擇最適合的演算法進行數據處理[53] [54]。

基因晶片包含 DNA 微陣列、基因表現晶片、染色體晶片(aCGH chip)、SNP 點突變晶片、miRNA 晶片等，另外一些用於檢測 DNA 甲基化圖譜的生物晶片，近年在各分子生物學實驗研究中，有使用上日益頻繁的趨勢[55]。

目前網路上有許多資料庫收錄各種平台的基因晶片研究數據，美國 NCBI 網站內部建置的基因體學資料庫 GEO(Gene Expression Omnibus)中收錄超過 160 萬包括各類平台及模式生物的基因體實驗數據，幾乎所有收錄在 pubmed 系統中文獻所引用的基因體數據都會被收錄於此，使用者可以使用 NCBI 所提供網頁系統從同一個實驗組(GEO dataset)中挑選樣本(GEO sample)重新分組來看分析結果，如果想要跨實驗組分析、調整細部運算參數，GEO 也開放使用者自行下載原始資料檔來重新分析[56]。TCGA(The Cancer Genome Atlas)資料庫收錄 34 種癌症細胞樣本的基因體實驗數據，提供使用者自由下載原始數據或已經標準化數據，且裡面每個樣本同時擁有基因表現、染色體複製、基因甲基化、基因突變的多種晶片實驗結果，以及詳盡的病患臨床診斷書，是探討癌細胞分生機制時極為高價值且多用途的著名重要資源，Sebastain 即利用 TCGA 中每個腦癌樣本的甲基化、染色體、CNV 三種晶片數據比較，進一步發現了腦癌幹細胞中指標基因 HOXA10 在基因表現量上無法精準辨別腦癌細胞種類的原因，是因為表觀遺傳機制參與影響了 mRNA 表現量[57](圖 八)。

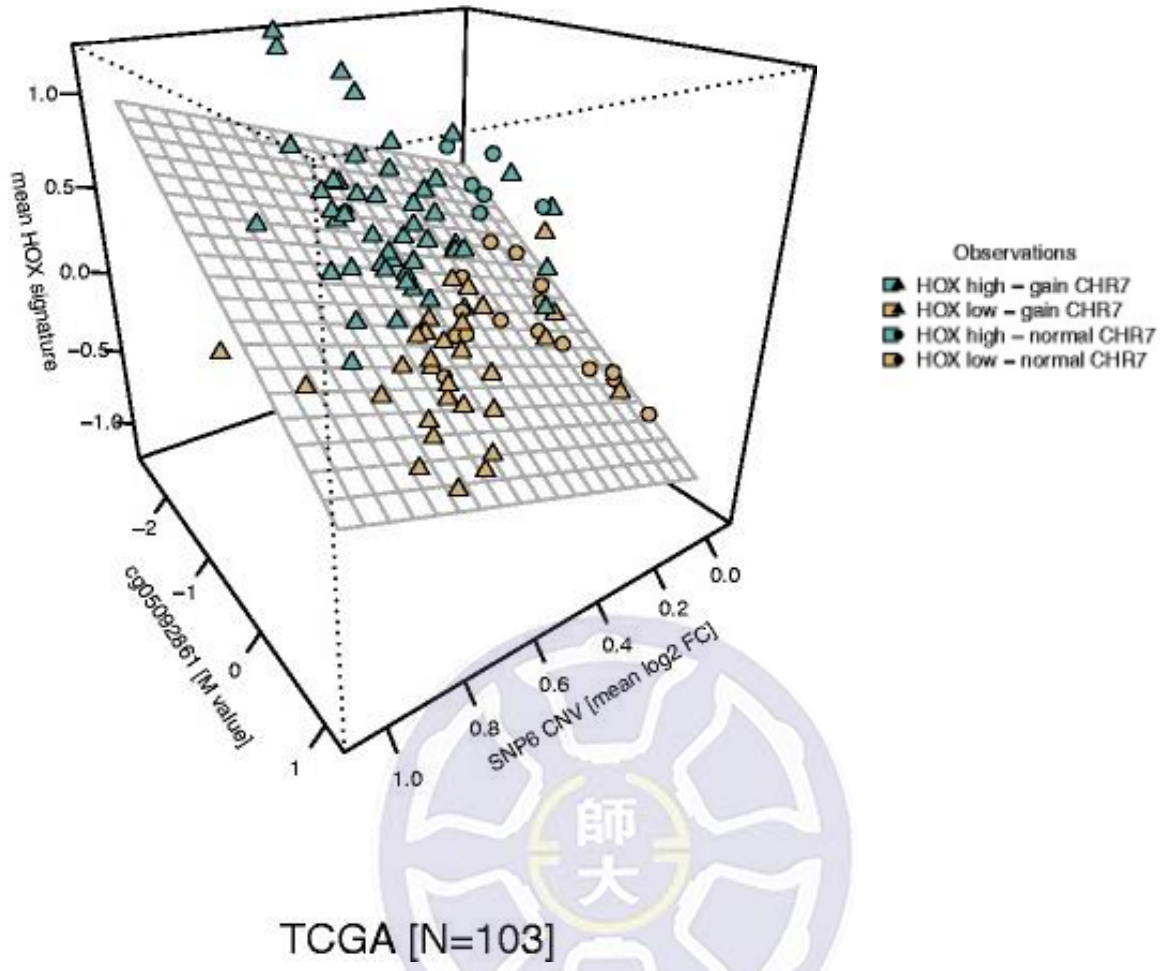

圖 八 TCGA 資料庫運用

Sebastain 等人使用三種不同類型晶片數據推測腦癌幹細胞指標基因 HOXA10 受到 DNA 甲基化調控，圖中每個點為各個腦癌病患檢體，三角形表示檢體中 HOXA10 所在的第七對染色體有異常複製增殖現象，圓形表示沒有染色體增值現象低，綠色和黃色分別代表 HOXA10 表現量相較健康檢體為高或低。

每個點的 3 個座標參數分別為相對於健康檢體的 7p15.2 處 DNA 增值、cg05092861 處甲基化、mRNA 表現量。

作者發現黃色三角形節點，皆坐落於高甲基化座標的位置。

2.6 基因網絡視覺化(Visualization by Gene Networks)

隨著科技發達，人類在各領域研究成果資訊漸漸邁向高通量數據，擁有大量的資料筆數、資料維度。想要在這種龐大複雜資料中歸納系統、趨勢時，基本的表格呈現方式已經無法提供人員閱讀，資料科學家需要輔以各種「視覺化」系統來對資料進行降維、壓縮、統計分析後，以較能閱讀的圖形呈現，才有辦法做後續判別分析[51, 58]。

生物資訊界所面對的統整資料，每一筆數維度可達上萬，各筆資料間彼此還有複雜的交互關聯，而這些關聯往往又是絕不可遺漏的寶貴資訊，因此常使用「節點網絡分析」。將各筆資訊完成維度統整、降低後，在圖形上以一個「節點」(node)代表一筆資料，節點的位置、顏色、大小、形狀呈現該筆資料的各參數。節點之間具有「節線」(edge)，用以呈現各筆資料的交互關聯，以節線的粗細、顏色、方向、樣式、曲度來呈現各種交互關聯性資訊。節點與節線在平面、空間上交織擺放後，呈現一個實體的網狀構造，讀者便可以目視方式一覽所有資料的整體資訊[59]，這些在生物研究界所使用的網絡圖統稱為生物網絡(biological networks)。

生物網絡中每個節點的連接節線數稱為其鄰居數(direct neighbors)，研究顯示生物網絡大多遵守點鄰居數 power-low 分布現象[60](圖 九)，表示在生物網絡中，常只有少數點具有高鄰居數，若刪除這些具有高鄰居數的節點，便有可能破壞網絡整體結構。然而並非所有高鄰居數節點都具有這種影響整體網絡構

造的能力，也有些節點雖然不俱影響整體網絡結構能力，卻位於最能夠最快速擴散訊息的交通要道，因此拓樸學定義了連接中間度指標(degree centrality)、近距中間度指標(closeness centrality)、參與度中間指標(betweenness centrality)等參數，用以量化評估每個節點在網絡中各種角色[61]。

其中參與度中間指標(圖 十)計算方式為該節點出現在網絡中任兩個節點間最短路徑上的次數，其公式如下：

$C_B(t) = \sum_{i \neq t} \sum_{j \neq i \neq t} \frac{\sigma_{i,j}(t)}{\sigma_{i,j}}$ (公式一)

其中 $\sigma_{i,j}$為網絡中第 i 個節點與第 j 個節點最短路徑數，$\sigma_{i,j}(t)$為$\sigma_{i,j}$中經過第 t 個節點的路徑數，代入(公式一)得$C_B(t)$即為第 t 個點的參與度中間指標。

基因體研究常用的基因網絡圖，即是以每個節點代表一個基因或轉譯蛋白，節點之間連線表示基因或蛋白彼此調控、交互作用[62]，其中所搭配的節點座標演算法極為重要，選擇適當的座標演算法能讓基因網絡圖展現出更多隱藏資訊，讀者可以很方便從節點分群散步結果一目了然其中的關鍵基因、參與生物途徑。目前常用的節點座標計算方式包含主成分分析[63]、圓形圖一維排列、重力導向[64]等演算法。

網路免費資源中包含許多如 Cytoscape[65]、Gephi、Pajek 等開源視覺化軟體供研究者輸入基因表現資料以繪製基因網絡圖。另外像 STITCH[66]、BioGrid[67]、Reactome 等網站提供使用者輸入基因名單後，根據其內建生物途徑資料庫搜尋相關基因或化學物質回傳基因網絡圖。Array Mining 提供使用者

上傳生物晶片原始檔數據，然後根據數據中基因表現狀況顯示對應的基因網絡圖[68]。

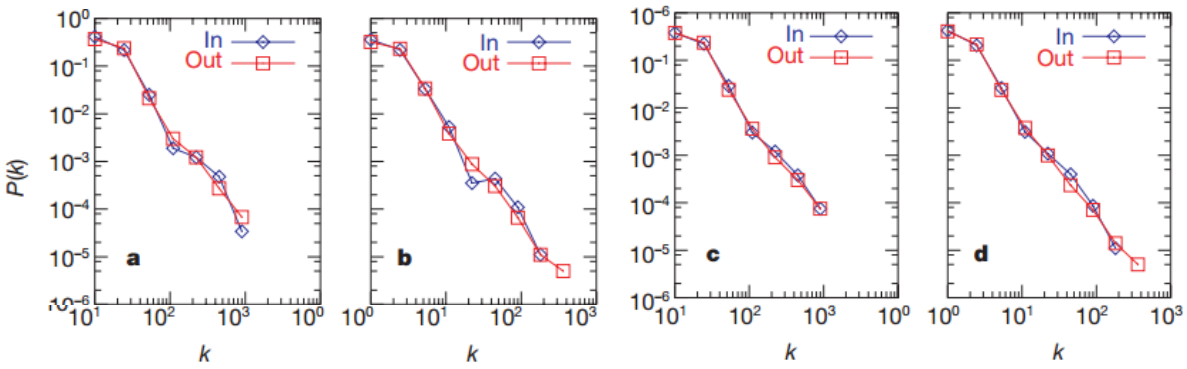

圖 九 生物網絡的 power-law 現象

a,b,c,d 為四種不同微生物細胞代謝途徑繪製出的四個不同網絡圖中各節點鄰居數計算分布圖，不同於電腦隨機產生網絡圖的鐘型分布結果，生物網絡中具有越高度鄰居數的點數量越少

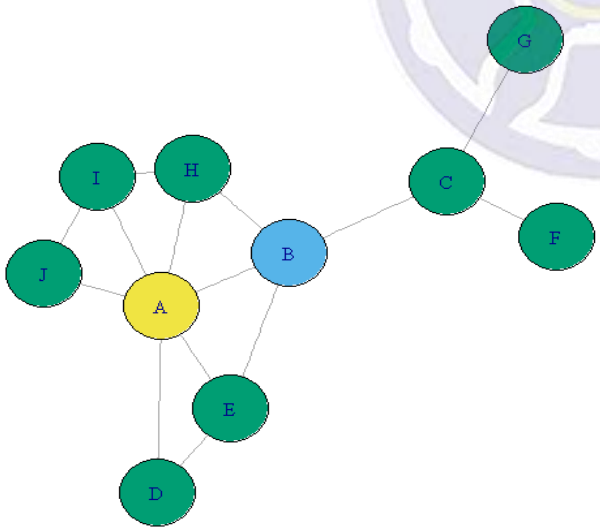

圖 十 中心度指標說明圖

圖中 A 節點擁有最多鄰居與最高連接中間度指標，鄰居數較少卻位居兩大群落節點必經之路的 B 節點則擁有最高的參與度中間指標

2.7 R統計軟體(R Language)

R 語言是統計程式語言，其開發環境和運作引擎皆由 C 語言撰寫而成，其內部採用的特殊記憶體參照共用配置演算法，而能在有限硬體資源內快速完成許多大型矩陣運算。R 提供使用者極為直觀簡潔的矩陣變數型態和邏輯運算語法，又內建許多統計上常用的繪圖套裝指令，加上其完全免費開源性質，使 R 成為全球學術研究上廣泛使用的後端數據處理工具，也因此獲得極為龐大的套件開發者社群。

Bioconductor 計畫於 2001 年由美國福瑞德·哈金森癌症研究中心主持並實行，建置一個具高度擴充性的 R 生物資訊套件集，其中包含各式生物研究界常用的進階統計、繪圖套件、註解資料庫以及基因體研究相關數據處理演算法。其目標為促進分析工具開發、減少生物資訊學研究成本門檻。目前 2015 年 3.1 版 Bioconductor 共收錄 1024 種投稿經認證的套件，供 R 軟體使用者免費安裝使用。

2.8 MVC架站結構(Model-View-Controller)

MVC 為軟體工程中一種設計結構概念，將軟體系統分成內建模型(Model)、外部檢視(View)、中間控制(Controller)器三個部分(圖 十一)，不僅大幅降低程式設計時複雜度，且使得一般公司行號在撰寫大型程式時，可以依照此種設計架構讓程式撰寫人員各自發揮所長，用自己最熟悉的程式語法和系統結構來進行分工合作[69]。

MVC 概念若套用在網路資料庫架站，更成為一種安全性、擴充性更強的資料庫架站模式。MVC 清楚分明的三段式程式設計，使資料庫維護人員在交接管理工作時變得非常方便，大幅增強資料庫網站永續經營維護的機動性。而其「必須經過 Controller 來控制資料庫」的特性，也較能保障內部資料庫免於來自外部指令的攻擊與破壞。

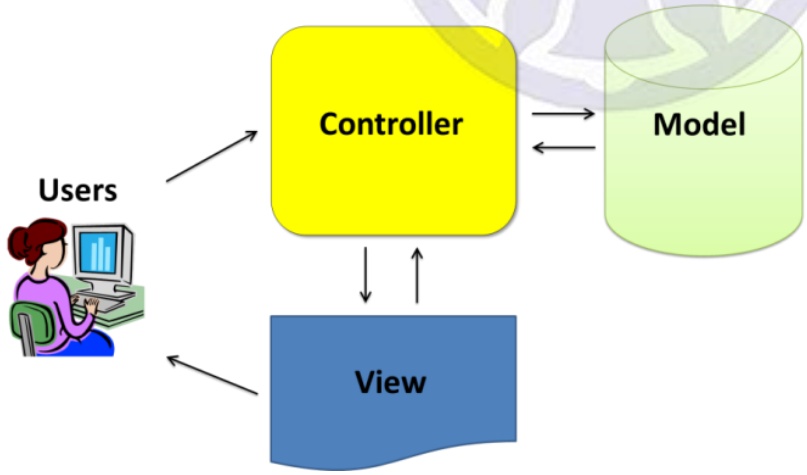

圖 十一 MVC 架構

將軟體系統分成內建模型(Model)、外部檢視(View)、中間控制(Controller)器三個部分，其中透過 Controller 與 View 與使用者互動。

3 研究方法

3.1 架設MVC資料庫

採用 MVC 結構架設一個用於分析生物晶片的網頁資料庫系統，命名為 AryNet(圖 十二)

- 資料模型(Model)方面主要以 postgreSQL 軟體管理並存放資料庫，有些維度太高的資料則以 Rdata 檔方式存放於硬碟資料夾中，由 R 統計軟體予以管理。
- 用戶檢視端(View)以 ZK 網頁模組作為主要網頁框架，並輔以現成 JavaScript 函式庫—Sigma.js 進行網絡視覺化呈現。
- 中央控制器以 Java 語言進行開發，分別透過 REngine、ZK、Mybatis、sigma 四種介面套件來分別對用戶端和資料庫進行控制，四種套件將會在以下實驗方法敘述中詳細說明。

主機型號為 optiplex 755，搭配 Intel 公司出廠 Q6600 微處理器，實體記憶體 8 GB，平台使用支援 64 位元系統的 Windows 10，以 Netbeans 內建 Apache Tomcat 8 負責伺服器回應。系統網路位址則採用 port-directed 技術占用實驗室配置 ip 中的 7777 連接阜。

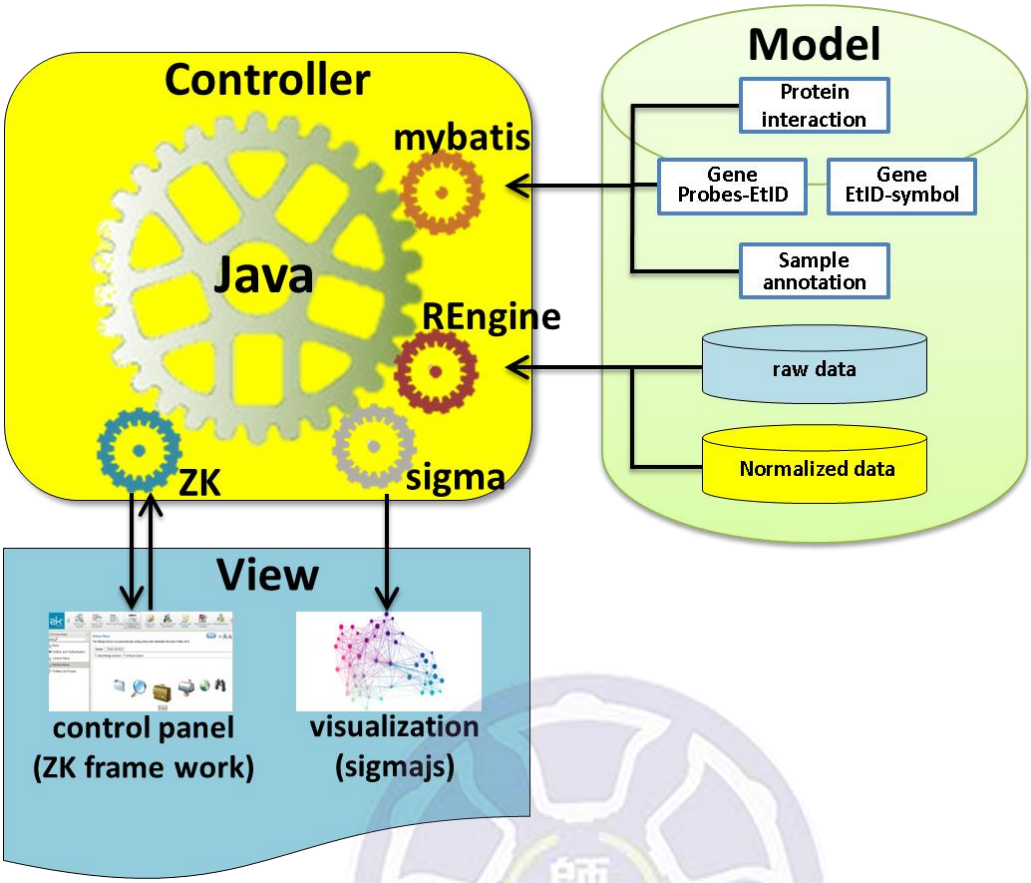

圖 十二 AryNet 的系統架構

AryNet 以 Java 為中央控制器(Controller)，透過 ZK、sigma 兩個套件連接以 JavaScript 語法為主的用戶檢視端(View)，透過 Mybatis 與 REngine 兩個套件分別連接到 postgreSQL 的四種資料表和硬碟中的 Rdata 檔案

3.2 安裝R運算引擎套件

由於分析生物晶片數據時，需要引用很多生物統計專用的複雜演算法，加上使用者欲進行晶片樣本品質監控、基因表現量觀察時所需要的基因體學常用圖表，這些細節運算處理若直接用系統中央控制器 Java 語言撰寫，將會耗用龐大撰寫人力，而且無效率。

AryNet 伺服器引用由 rforge 套件開發社群網站下載的 Java-R-interface 以及 R-Java 兩種套件個別安裝於伺服器的 Java 引擎和 R 軟體中，並按照設計者提供方法在 AryNet 的中央控制器 Java 程式碼中建置一個 REngine 物件。REngine 物件允許程式在運行 Java 引擎時，以虛擬主機方式另外開啟一個 R 語言運作引擎。程式設計者可以在 Java 語言中對運作引擎下 R 指令，並且獲取運作引擎所回傳的運算結果圖片、矩陣值，甚至還可以透過 R 運作引擎直接攫取或輸出實體主機硬碟中的檔案來做處理(圖 十三)。

由於 REngine 實作方式是直接將系統原安裝的 R 軟體以 Java 虛擬主機方式開啟並呼叫，因此程式設計師可以在 Java 中使用之前 R 軟體中安裝的所有引用套件，同時享受到 R 軟體本身優化過的記憶體分配運算，且不會影響到原系統安裝的 R 軟體設定。目前網路免費 REngine 尚未提供平行運算處理功能，故一台伺服器僅能同時發動一個引擎同時服務所有連線使用者，AryNet 將所有使用者命令和使用物件以使用者 ID 標示分別，以避免共用引擎造成的資料衝突。

AryNet 使用 REngine 主要來進行下列工作

1. 生物統計運算(包括生物晶片前置作業、標準化運算)
2. 多維矩陣運算
3. 生物晶片數據檔案存讀
4. 生物晶片平台註解讀取
5. 靜態統計圖繪製

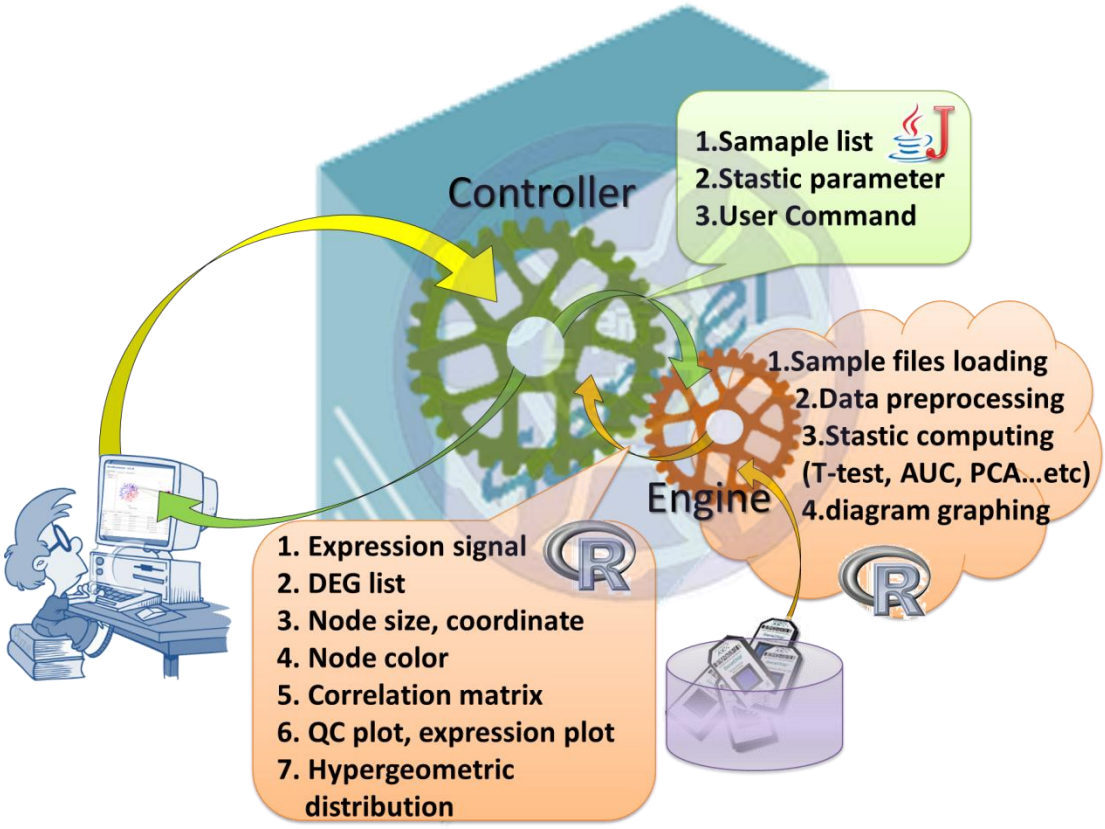

圖 十三 AryNet 中央控制器與 R 運算引擎架構示意圖

綠色齒輪表示系統中央控制器，橘色齒輪表示 R 運算引擎，中央控制器依據使用者要求將控制命令傳送給 REngine，使用 REngine 處理複雜、大型的統計運算以及部分繪圖，然後接收資料並準備處理視覺化。

PostgreSQL 是一種關聯性資料庫管理系統，以加密檔形式將數張使用者設計的虛擬表格存放於伺服器硬碟中，並提供安全防護的 SQL 類型指令供使用者提取、修改資料表格內容。 AryNet 中央控制器所安裝的 MyBatis 套件，是由 Ibatis 創作團隊開發，使 AryNet 能透過 Java 語言控制其外接 PostgreSQL 資料庫模型。

目前 AryNet 資料庫中包含 4 張不同平台生物晶片註解表(Probes_to_EntrizID)、1 張人體基因註解表(gene_annotation)、1 張蛋白質交互作用關聯性表(protein_interaction)以及 2 張樣本註解表(sample_annotation)，這些表格都是以 PostgreSQL 加密檔存放，透過 Mybatis 操控 PostreSQL 管理。

另外 AryNet 還有一個占用大部分硬碟的專用目錄資料夾，裡面存放所有收錄生物晶片數據經 R 軟體讀取過後輸出檔(Rdata)，以及晶片數據檔(CEL、idat)，這些檔案存放於資料夾中透過 REngine 管理。

3.4 資料庫來源

系統內建資料庫包括生物晶片實驗數據、晶片樣本註解、晶片探針註解、人體基因註解和蛋白質交互作用五種資料 (圖 十四) 。

A. 晶片實驗數據：

目前 AryNet 所收錄晶片實驗數據皆來自 NCBI 網站的 GEO 資料庫，研究人員以手動操作方式搜尋 GEO 中 hgu95av2、 hgu133a、 hgu133aPlus2、450k 4 種平台的環境荷爾蒙、精神疾病相關原始數據並下載到伺服器硬碟專用目錄中，使用 Bioconductor 提供的 Affy 和 450k 兩種套件載入 R 程式後，先將原始數據(raw data)以 R 程式暫存檔(Rdata)輸出存放硬碟中，然後使用 MAS5 或 Illumina Preprocess 兩種演算法將載入 R 程式中的原始數據個別進行均一標準化，再次以 Rdata 檔輸出。以 Rdata 暫存檔形式存放，讓系統能夠在每一次使用者要求讀取晶片時，直接快速載入 REengine。綜合上述，每個晶片樣本會有三個對應檔案硬碟中，原始檔(CEL,idat)、原始數據快取檔(Rdata)、MAS5 或 Illumina 方式標準化後數據快取檔(Rdata)。

B. 晶片樣本註解：

自 GEO 下載的樣本註解檔案格式為 CSV 表格，研究人員於註解中獲取樣本的晶片平台、GEO 編號、細胞組織、作者、採樣日期、添加化學藥物以及樣本提供者的年齡、性別、生活特性、診斷病症等資訊，手動轉存為 SQL 語法後輸入 PostgreSQL 中的 sample_annotation 表格存放。

C. 晶片基因註解：

晶片數據經標準化後僅輸出各個探針組的代表螢光值及探針組編號，無法得知其探針組對應序列為哪一個基因所有，此時需要特定晶片基因註解資訊將探針組編號(Probe name)翻譯為對應基因。

AryNet 內建四張晶片註解表格資料皆來自 Bioconductor 網站下載的 R 套件，目前僅安裝了 hgu95av2.db、hgu133a.db、hgu133aPlus2.db 和 IlluminaHumanMethylation450k.db 四種套件。我們擷取其中所有探針編號以及對應基因的 Entrez ID 後按照各平台製成不同 PostgreSQL 表格輸入 AryNet。

另外，因為 Affymetrix 系列晶片探針組是根據其內部每個探針序列對應 mRNA 精準度標上_x、_a 等符號來命名(圖 十五)，我們在 AryNet 中撰寫一段 R 語言函數，負責將指向同一個基因的不同探針組按照其專一性進行「優先權排序」，其排序演算法大致如下：

I. 若探針組名稱中含「_at」的一律先加一分。

II. 名稱中含「_x_at」扣分，扣除分數大小視「_ 任何字母_at」總數而定。

III. 名稱中含「_s_at」扣分，扣分大小視「_s_at」和「_ 任何字母_at」總數而定。

IV. 名稱中含「_a_at」扣分，扣分大小視「_a_at」、「_s_at」、「_ 任何字母_at」總數而定。

演算法結果會將各探針按照「_at」>「_a_at」>「_s_at」>「_x_at」的方式排序優先權，編號 1 到 n，這些計算結果也都併入到 affymetrix 序列的晶片基因註解表格中。

D. 人體全基因註解：

因為每一種基因可能會有兩個以上不同的 symbol name，因此針對上述晶片基因註解資訊，AryNet 皆只擷取探針組編號以及 Entrez ID 來存放，要對使用者顯示視覺化圖形時，AryNet 會根據一張「人體全基因註解表格」統一顯示對應的基因名稱。 目前 AryNet 內建人體全基因註解表格內容來自美國 NCBI 網站 FTP 空間中的 human_gene_info.csv，除了 Entrez ID 和 symbol name 之外，表格內還包括基因的染色體位置、轉譯蛋白質形式、更新日期資訊。

E. 蛋白質交互作用訊息：

來自 Reactome 網站下載頁面的 Human protein-protein interaction pairs.csv 檔案 2014 版，其中蛋白質交互作用資訊皆通過 Reactome Team 校對審核，包含蛋白質複合體(complex, binding)、磷酸化(phosphorylation)、去磷酸化(dephosphorylation)、抑制(inhibition)、激活(activation)、催化(catalyze)、基因表現調控(expression regulation)、泛素化(ubiquitination)等資訊，每一筆交互作用資料皆各有 target protein 和 source protein 兩種參數。

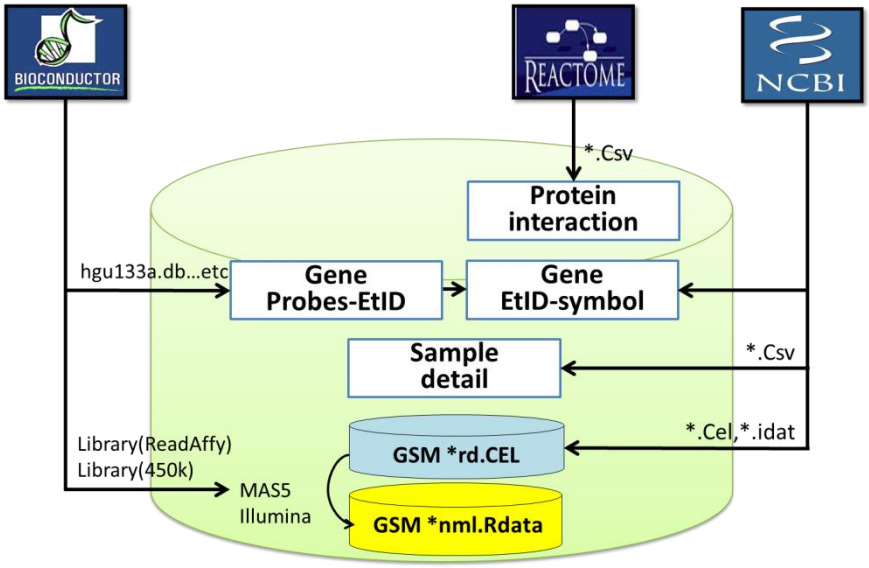

圖 十四 AryNet內建資料庫資料來源

方塊格子表示postgreSQL資料表，小圓盤圖表示存放於硬碟目錄中的Rdata檔案

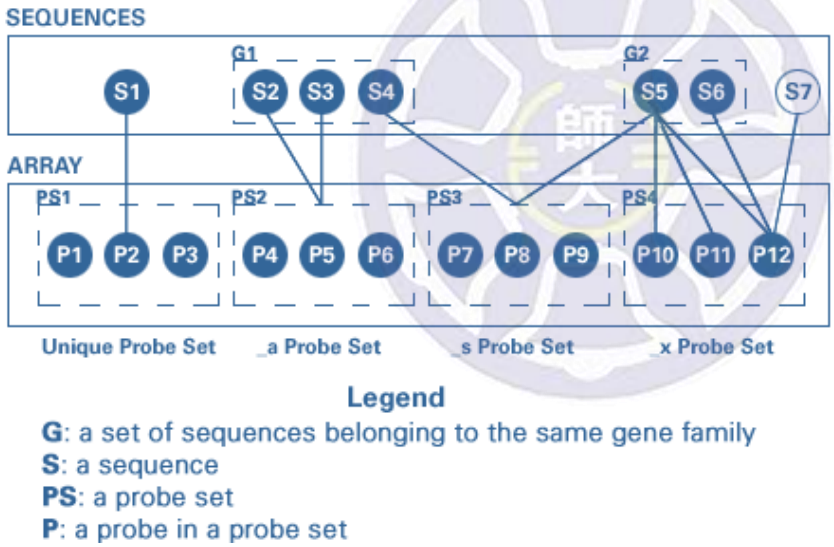

圖 十五 Affymetrix公司提供的探針組命名法則

若探針組名是純粹以_at結尾(ex: 2343_at)，則其序列只會專一偵測到一種基因mRNA，一般生物晶片研究會優先採用此種探針組作為訊號依據。若以_a_at(ex: 2343_a_at)結尾，表示其探針會黏到同基因家族中的不同基因。_s_at表示其探針序列會黏上不同基因家族的mRNA片段。

3.5 使用者操作介面與中央控制器設定

AryNet 採用 ZK 框架來建立動態網頁式的使用者控制介面。ZK 是台灣人開發的一套網頁應用程式框架，藉由類似 xml 格式的 zul 檔作為 Java 與 JavaScript 溝通橋梁，因此程式撰寫人員可以在完全不懂 JavaScript 的狀況下透過 Java 語法建立物件來控制 zul 檔中所宣告的 JavaScript 物件。

在 AryNet 架構中，幾乎整個首頁和內部子頁面中所有網頁元件，都有相對應的 Java 物件在中央控制器中負責操控，當來自遠端使用者進行連線請求時，Apache Tomcat 會為了該使用者將整個網頁元件、中央控制器的所有對應物件和變數全部完整重新製作一份開放給他，因此，使用者可以各自享受到幾乎完整的主機端服務。

AryNet 的首頁分成上、下、中、左三個框架，其中上框架為訊息監控儀；左框架為主功能開關列；下框架為網絡圖的「群落控制面板」；中間為顯示基因網絡視覺圖的主要畫面顯示區，另外還有一個以浮動式窗方式顯示的「節點控制面板」(圖 十六)。

AryNet 透過 ZK 套件部署共 63 個事件聆聽函數(Event Listener)於中央控制器，全程監控使用者在網頁上對任何 JavaScript 元件的滑鼠拖曳、點按動作，並即時啟動中央控制器內部的負責應對函數改變網頁輸出圖形。此種架設方式讓 AryNet 可即時回應使用者每一次參數調整動作，成為具有互動式調控介面的動態型網頁系統。

圖 十六 AryNet 使用者介面

綠色矩形範圍為「主功能開關列」；褐色矩形範圍為「群落控制面板」；紅色矩形範圍為「主要畫面顯示區」，藍色矩形範圍為一浮動式窗「節點控制面板」

3.6 互動式網絡圖功能

由於 ZK 框架本身並沒有提供網絡圖繪製功能，因此我們在 AryNet 的主要顯示框架中安插一個 div 元件，引用 Sigma.js 函式庫來繪製基因網絡圖於該元件中。Sigma.js 是一個專門繪製網絡圖型的 JavaScript 函式庫，僅提供使用者網頁 JavaScript 語法撰寫操作。為了讓 Sigma.js 的繪圖功能能夠和 AryNet 的 Java 中央控制器更具縝密性來整合在一起，我們為 Sigma.js 撰寫一組 Java-JavaScript Interface 套件，命名為 sigma，必且在其中增加了一些生物基因網絡需要用到的加強功能。

透過 sigma 套件，AryNet 的中央控制器不僅能對使用者端的基因網絡圖做即時性動畫修改，還能監聽使用者對圖中節點的點按、拖曳動作來予以回應。

3.7 系統運算流程

當使用者選好對照組和實驗組的晶片編號並且下達繪圖指令時，AryNet 會載入對應的晶片數據並按照使用者選取演算法進行去背景值、數據標準化；然後根據使用者所選擇的差異統計量公式以及門檻值，挑選出在實驗組和對照組之間表現量具有顯著差異的基因(Differential Expression Gene ,DEG)；接著系統按照過濾過的 DEG 名單根據其表現量計算出相關係數矩陣。

完成初步運算後，系統會於使用者端的主畫面視窗繪製節點，一個節點代表一個顯著差異基因，系統會根據相關係數矩陣將彼此表現量具高度相關的基

因上虛線表示而構成網絡圖型，綠色和紫色虛線分別代表正、負相關。

上述以一個對照組和實驗組比對所產生的一個 DEG 基因網絡，我們在系統中命名為一個「群落」(Crowd) (圖 十七)， AryNet 針對甲基化晶片資料視覺化也另外提供類似的節點群落圖，節點代表每個顯著差異的基因甲基化熱點，形狀改以菱形代替(圖 十八)。使用者可以選取多對實驗組與對照組數據而在畫面中產生好幾個群落，系統會要求使用者對每個群落命名以方便區別。

在使用者操作介面中的下方控制面板即為「群落控制儀」，可隨時調整每個群落的圖形設定。

若使用者開啟「known linkage」功能，則系統會根據內建資料表搜尋一個群落中各個基因是否彼此有轉譯蛋白質交互作用，並分別以不同實心箭號予以標示，若使用者點選了其中「client involved」選項，則系統會根據內建資料庫將群落中每個節點代表基因的相關蛋白質基因全部加入群落中，然後重新繪製實心箭號，這些因為交互作用資訊而被加入的附加節點(Client Node)，和其所連結的 DEG 節點會納入同一個群落(圖 十九)。

當使用者要求進行兩組群落基因名單交集分析時(2 Crowds overlaps analysis)(圖 二十)，AryNet 會統計兩個群落名單中是否有相同 Gene Symbol，然後回傳兩個參數，一個是超幾何分布公式計算的 p-value，另一個則是比較兩群 DEG 相較於各自對照組的顯著差異方向是否具同質性的參考數值。

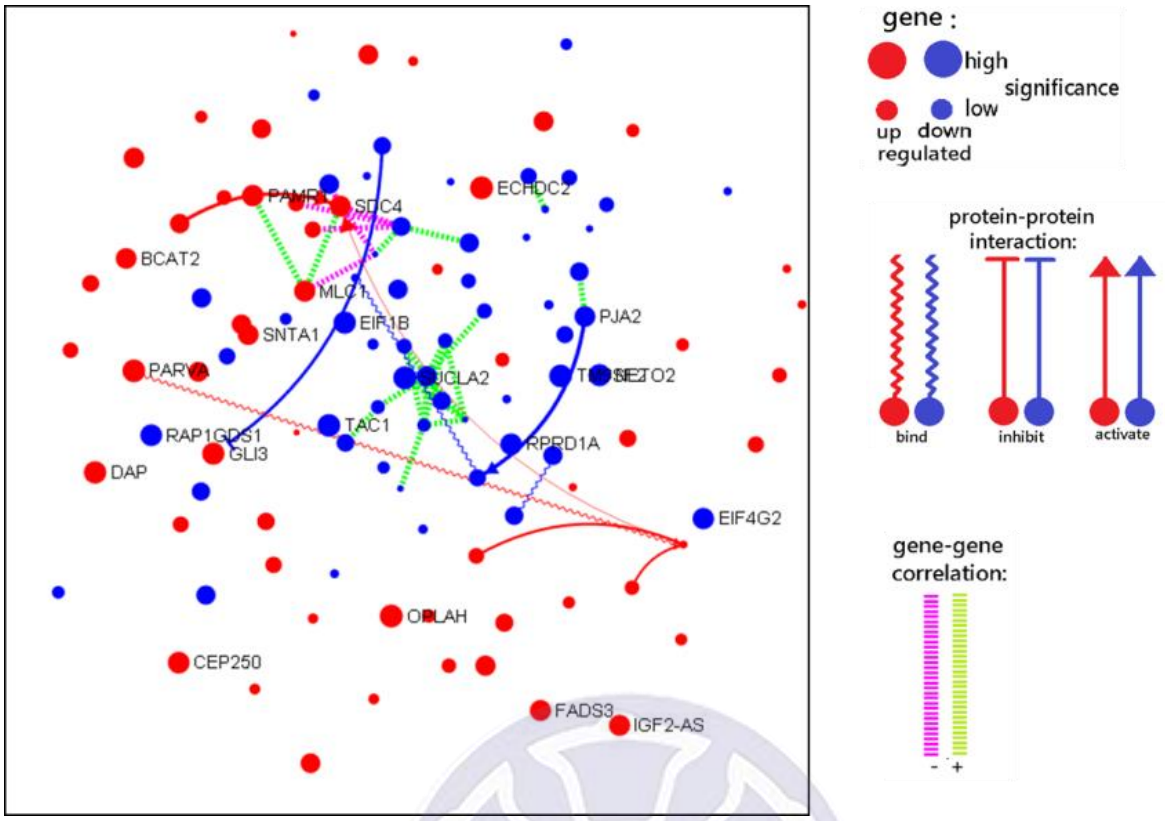

圖 十七 基因群落

由一對對照組和實驗組晶片數據分析而成的基因群落，其中紅色節點為正調控基因，藍色節點為負調控基因，綠色虛線表示高度正相關，實心箭頭為蛋白質激化作用，實心丁字箭頭為蛋白質抑制作用，z型連線表示蛋白質複合體。

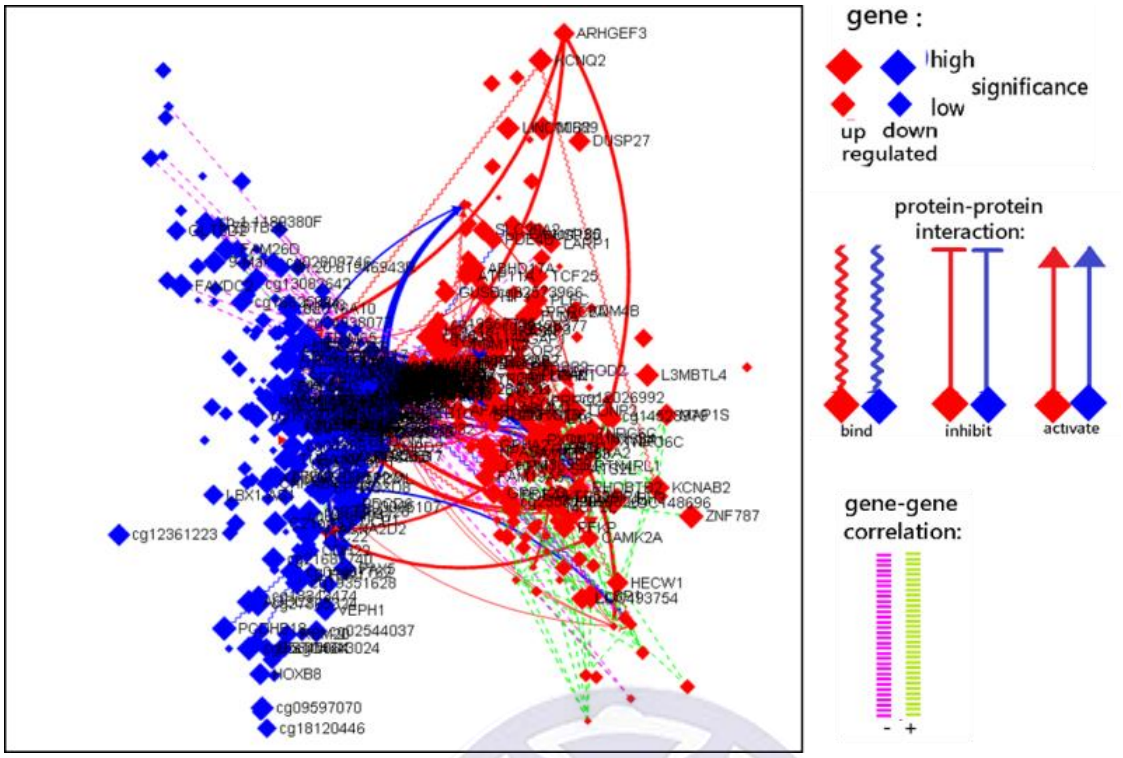

圖 十八 甲基化晶片的熱點群落圖

其中每個菱形節點為實驗組與對照組樣本中甲基化量具顯著差異的熱點，若此熱點為某個基因所有，則標上基因名稱，否則標上DNA位址，其餘圖形表示方式雷同DEG群落。

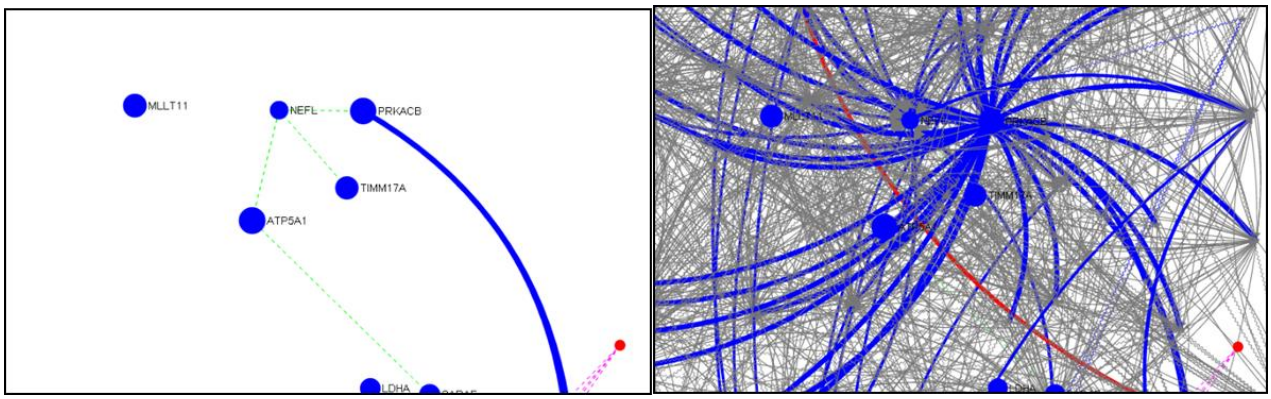

圖 十九 透過Reactome尋找附加節點功能

左圖為尚未開啟「尋找附加節點」功能的基因網絡，其中NEFL和PRKACB有綠色虛線連接，表示表現量有高度相關，但因為沒有直接的蛋白質交互作用，看不出為何會相關。右圖為對NEFL開啟「尋找附加節點」功能的結果，每個灰色節點是跟NEFL有直接蛋白質交互作用而被重新貼上來的基因，可以看到這些附加基因的蛋白質大部分都受到PRKACB蛋白質的正調控(從PRKACB發射出去的眾多藍色箭頭)，然後進而正調控NEFL(指向NEFL的眾多灰色箭頭)，從此可以看出NEFL和PRKACB表現量正相關原因

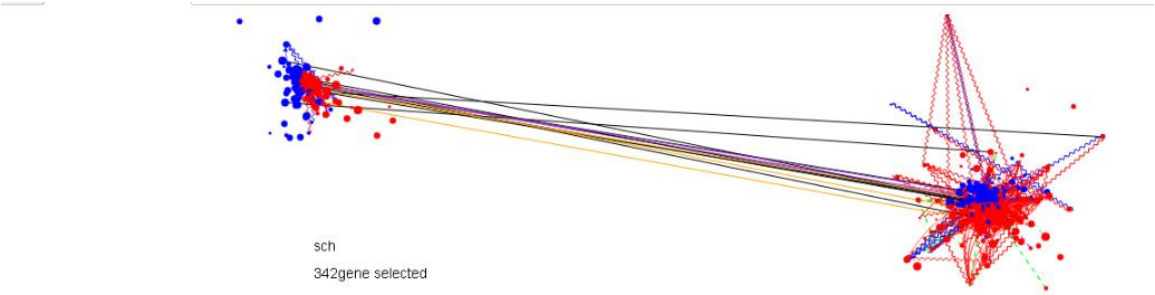

圖 二十 雙基因群落交集分析

兩個群落做交集分析，群落間連線代表系統找到的兩群落共同重複基因，紫色線表示該基因在兩群落皆為正調控，橘色線表示皆為負調控，黑色線表示調控方向相反

3.8 相關演算法實作與引用統計公式

3.8.1顯著差異基因

AryNet 提供三種差異統計量公式來過濾用來產生結點群落的基因名單，使用者可自由調整其門檻值。

A. T-test(Student's t test):

AryNet 使用變異數樣本數皆不相等的雙尾數 t 檢定公式，回傳每個基因在實驗組與對照組之間拒絕虛無假設的 p-value，然後使用 R Engine 中的 qvalue 套件以 FDR 方式換算 q -value，越小表示顯著差異度越高。

$$t = \frac{\overline{x}_1 - \overline{x}_2 - \mu_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)} \quad \text{(公式一)}$$

B. AUC(Area Under The Curve):

針對一個基因訂定各種閾值，並計算該基因在對照組與實驗組數據中以每個閾值作為鑑定基準時敏感度(Sensitivity)和專一性(Specificity)各為多少，然後將專一性對敏感度作圖後積分計算曲線下面積，此即為 AUC 值，AryNet 以 AUC-0.5 的絕對值來比較顯著差異度。

C. SVM(Support vector machine):

AryNet 會先記錄前一個差異統計演算法所選出的 DEG 名單，並按照其顯著差異度進行排序，並定義 m 初始值為 1，接著 AryNet 會向使用者索取 2 個值，分別為「平均分數最小值」和「分數標準差最大值」，然後在 DEG 名單中選出顯著差異度最大的前 m 個基因。

將要進行過濾的樣本組(包含對照組與實驗組)中每個晶片數據設為一個資料點，前述 m 個顯著差異基因表現量即為該資料點的 m 個參數。將資料點按照其參數散播於 m 維空間中，並標上組別編號(對照組或是實驗組)。

在空間中以重複修正的方式求出可以將對照組和實驗組資料點分開的最佳超平面，然後計算其劃分結果的準確性(accuracy)、專一性(specificity)、敏感度(sensitivity)將這三個數值取平均數及標準差，比較是否符合使用者訂定閾值，若符合，則此 m 個基因即為新的 DEG 名單，否則 m 值加一，繼續重複以上步驟。

3.8.2基因表現相關性

AryNet 在確認一個群落欲顯示的 DEG 名單後，會依據名單裡基因在每個樣本中表現量，於 REngine 中製作一分相關係數矩陣表格，其內容為名單中每個基因在表現量上兩兩皮爾森相關係數(Pearson Correlation)。

$$\mathrm{r}(geneA, geneB) = \frac{\sum[(expA_i - \overline{expA})(expB_i - \overline{expB})]}{\sqrt{\sum(expA_i - \overline{expA})^2 \sum(expB_i - \overline{expB})^2}} \quad \text{(公式二)}$$

若一共有 n 個 DEG，則相關性矩陣

$\mathrm{corM}_{n\times n}$為

$$\mathrm{M}_{i,j} = [r(gene_i, gene_j)] \quad \text{(公式三)}$$

3.8.3節點座標計算

AryNet 會按照提供下列三種預設方法讓使用者決定一個群落內部每個節點的初始擺設方式，節點擺設完畢後，使用者仍然能以滑鼠拖曳方式來改變每個節點的座標。

A. 組成分分析(Principal Component Analysis, PCA)：

一種將多維度資料進行降維然後以散播圖進行視覺化的表示法。

舉例來說，假設一個群落引用 6 張生物晶片數據並挑選出 10 個 DEG，一個 DEG 會有 6 個表現量，即 6 個參數。若希望以視覺化方式呈現每個 DEG 在 6 個參數中數值是否相似，則需要六維度空間來繪製 10 個節點。

AryNet 提供使用者使用 PCA 主成分分析方式將這 6 種參數壓所成 2 種參數，然後作為每個 DEG 節點在平面圖形上的橫坐標和縱座標，達成降維視覺化目的。圖形中若兩個節點距離相近，即表示該兩個基因的 6 個樣本中表現值高低相仿，在此舉上述例子概略說明實作內容：

10 個 DEG 在 6 個樣本中表現量 xM=$\begin{bmatrix} expr_{1,1} & \cdots & expr_{1,10} \\ \vdots & \ddots & \vdots \\ expr_{6,1} & \cdots & expr_{6,10} \end{bmatrix}$

找出一個最佳旋轉矩陣 rM=$\begin{bmatrix} r_{1,1} & \cdots & r_{1,6} \\ \vdots & \ddots & \vdots \\ r_{6,1} & \cdots & r_{6,6} \end{bmatrix}$並計算投影點

得 pcM=$\begin{bmatrix} PC1_1 & \cdots & PC1_{10} \\ \vdots & \ddots & \vdots \\ PC6_{1,} & \cdots & PC6_{10} \end{bmatrix}$ =rM × xM

即$\left[pcM_{i,j} = \sum_{k=1}^{6} r_{i,k} \times expr_{k,j}\right]$ (公式四)

(其中$pcM_{i,j}$即為第 j 個 DEG 的第 i 個主成分)

此時 pcM 含有最大變異數，且 var(PC1)> var(PC2)>…var(PC6)

B. 圓形圖排列：

先將一個群落的中所有 DEG 各自在對照組的表現量平均值和實驗組表現量平均值相減，產生一個正負號陣列

d=[-1,+1,+1,-1……]

其中 di=+1 或 -1 分別代表第 i 個 DEG 將較於對照組為「正調控基因」或「負調控基因」

設定每個節點在圓周的對應角度$\theta_i = d_i \times a_i \times \pi$ (公式五)

因此每個 DEG 節點座標為$(\cos\theta_i, \sin\theta_i)$

C. 相關度重力導向圖(Force directed graph)：

此為 AryNet 目前配置節點座標演算法中最耗記憶體資源的一種。目的在於呈現一個節點間距離反比於基因表現量相關係數的網絡圖，因此所呈現圖中節點若距離相近，表示此兩基因表現量具高度相關，無論正、負。

其原理為先依據群落中 DEG 的「相關係數矩陣」換算出「理想距離矩陣」

$\text{disM}_{n\times n}$=1-$\text{corM}_{n\times n}$=$[d_{ideal}(N_i, N_j)]$ (公式六)

然後將此矩陣輸入網絡中各節點的「自我調整函數」中，每個節點會同時於每瞬間自我檢查與周遭其他節點的距離是否與理想距離相等，若太遠，會產生一個朝目標節點移動的「拉力向量」，若太近，則產生一個遠離目標節點的「推力向量」(圖 二十一)，將這些「拉力向量」和「推力向量」加總成為每個節點的「校正移動向量」(圖 二十一)

$$\overrightarrow{m_i} = \sum_{j=0}^{num} \frac{\overrightarrow{N_iN_j}}{|\overrightarrow{N_iN_j}|} \cdot (d(N_i, N_j) - d_{ideal}(N_i, N_j)) \quad \text{(公式七)}$$

使每個節點按照「校正移動向量」移動，到各節點幾乎都安定不動時，終止演算法，然後重新記錄節點座標。

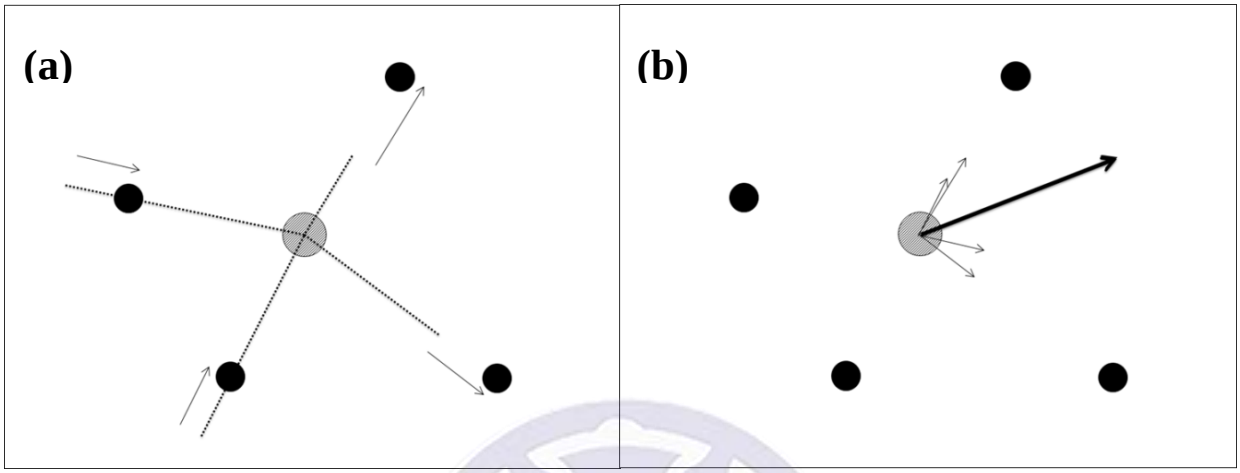

圖 二十一 相關度重力導向圖原理

(a)為利用實際歐幾里得距離與理想距離計算出推拉向量示意圖，**(b)**為將推拉向量加總成為移動校正向量示意圖。其中灰色節點為待校正節點，黑色節點為周遭其他節點，點狀虛線為理想距離，細箭號為推力向量或拉力向量，粗箭號為校正移動向量

3.8.4雙群落交集分析

兩個群落做交集分析時，因為兩個群落可能來自不統版本平台的晶片，一般費雪檢定(Fisher exact test)(圖 二十二)(公式八)會因為兩群落母群體不同而造成誤差，因此 AryNet 採用 Kupershmidt 於 2010 提出的雙基因群分析演算法(公式九)，是一種從費雪檢定法再做改良的超幾何分布公式，用來計算兩個集合的交集重疊是否具統計意義(圖 二十三)。

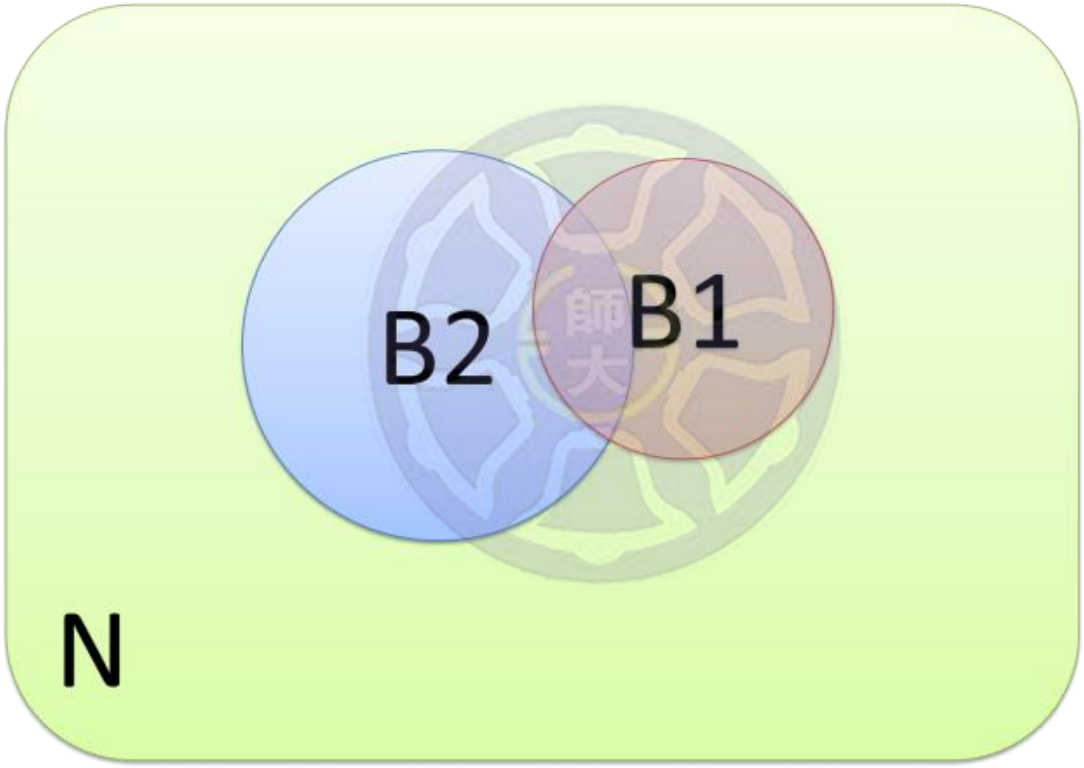

圖 二十二 費雪檢定法式意圖

利用超幾何分布公式計算 B2 和 B1 兩個集合交集是否具有統計意義。

$$p(B1,B2)=\sum_{B1\cap B2}^{B2}\frac{C_{B1\cap B2}^{B1}C_{N-B1\cap B2}^{N-B1}}{C_{B2}^{N}} \quad \text{(公式八)}$$

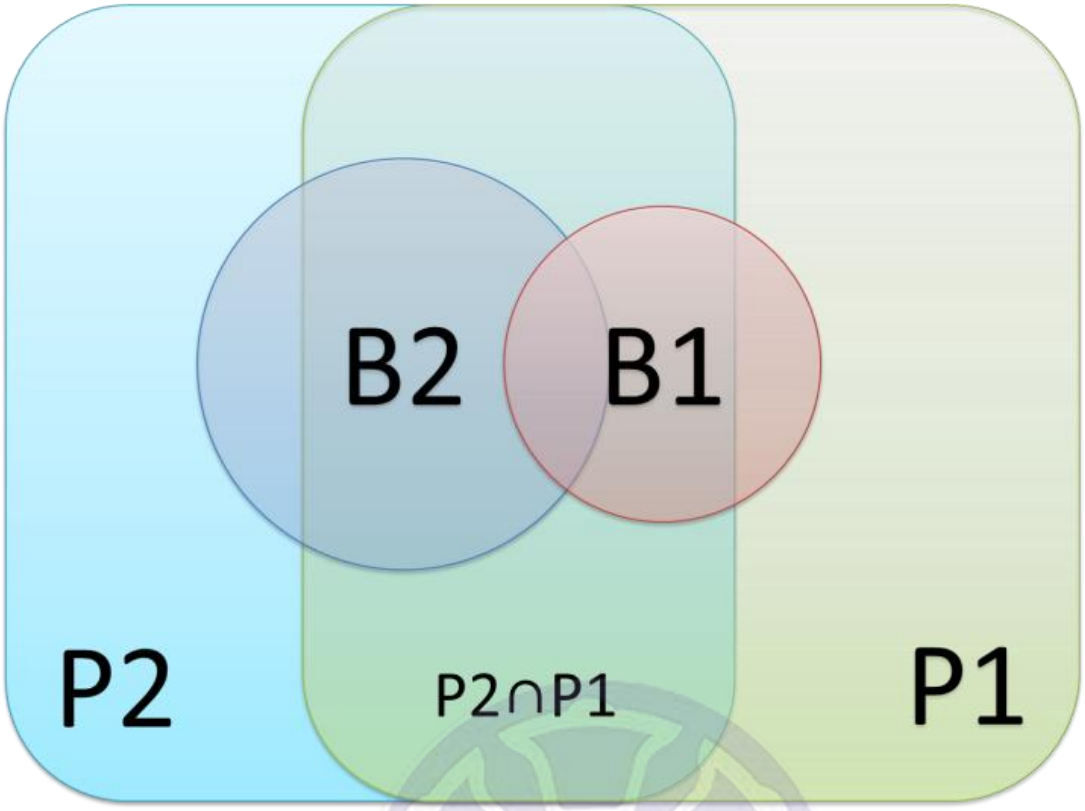

圖 二十三 改良過費雪檢定

當兩個集合各自擁有不一樣的母集合時(例如兩組 dataset 分別來自不同平台)，會因為過度估算總母集合而低估 p-value，修正過的公式只取中間兩平台交集來做運算。

$$p(B1,B2)=\sum_{B1\cap B2}^{B2\cap P1}\frac{C_{B1\cap B2}^{P2\cap B1}C_{P1\cap P2-B1\cap B2}^{P1\cap P2-B1}}{C_{P1\cap B2}^{P1\cap P2}} \qquad \text{(公式九)}$$

3.9 環境賀爾蒙與精神疾病相關性分析

我們將 35 個化學物質(表 一)相關基因體實驗數據載入系統 ，顯著差異基因(DEG, differential expression gene)過濾門檻設定為

1. T-test: p value 經過 FDR 校正後 q-value 需小於等於 0.05
2. AUC<0.3 或是 AUC>0.7

然後產生共 53 個基因群落(DEG 群落)(表 二)，通過統計門檻而偵測到表現量具顯著差異的基因群落共有 28 個，其中有 17 個群落中包含超過 10 種顯著差異基因。

另外將躁鬱症、帕金森氏症、思覺失調症、重度憂鬱症的基因體實驗數據載入系統，以同樣的顯著差異基因過濾門檻值產生共 19 個基因群落(表 三)，其中只有 4 個帕金森氏症的相關樣本組產生群落擁有較多顯著差異基因，另外還有 3 個躁鬱症相關的樣本組個別只找到 1 個顯著差異基因。

接著將通過統計門檻而擁有顯著差異基因的 28 化學物質 DEG 群落與 8 個疾病 DEG 群落做「化學物質對疾病」雙群落交集分析(圖 二十四)得到共 224 對分析結果，其中根據系統回傳的修正型費氏檢定 p 值(<0.01)重疊範圍具顯著意義的分析結果共有 31 對，分析其中包含的疾病與化學物質名單如(表 四)。

這些顯著差異基因和帕金森症有高度重合的化學物質中，其中雙酚 A(BPA)為目前已有部分研究顯示和一些精神疾病、腦神經退化疾病相關的環境干擾素。在這次統計中數據顯示其對 MCF10 細胞株影響的基因差異和帕金森氏症病人

部分腦切片的基因表現變異有高度重合(圖 二十六)。我們將「雙酚 A 樣本組基因群落」中和「帕金森氏症黑質組織樣本組基因群落」重合的 132 個基因名稱列出來並輸入 AryNet，其中表現量數據來源選擇「雙酚 A 10e-5M」，建立一個指定基因名單群落(圖 二十七)。這個基因群落中共有 12 筆被 Rectome 資料庫收入的蛋白質調控關係(圖 二十八)，其中尤以包括 ATR Serine/Threonine Kinase (ATR)的細胞分裂 S phase 的相關蛋白質反應途徑最多，另外我們將「基因表現量相關度」節線顯示功能打開(圖 二十九)，將「雙酚 A 10e-5M」樣本組中表現量相關度|r|>0.9 的基因連上虛線，然後找出網絡中中心度指標最高的前 15 名基因，其中包括 PIK3R3、TUBB3、PSMD11 等(圖 三十)。

表 一 35種化學物質

Chemicals			
Ampicillin	CyclosporineA	Genistein	Propanil
Aresnic trioxide	Daidzein	Lindane	Rapamycin
Azathioprine	Deoxynivalenol	Mannitol	Silver nitrate
benzo[a]pyrene	Diazinon	Methylmercury	Sodium citrate
Bis(2-ethylhexyl) phthalate	Dibutyltin dichloride	Mono-2-ethylhexyl phthalate	Tributyltin chloride
Bis(tributyltin) oxide	Diethylstilbestrol	Mycophenolic acid	Urethane
Bisphenol A	Fingolimod	Nonylphenol	17alpha-ethynylestradiol
Cobalt2chloride	Fluoxetine	Ochratoxin A	17beta-estradiol
Cyclophosphamide	Furosemide	Prednisolone	

表 二 35 種化學物質做 DEG 分析的 53 個基因群落內名單個數

↑、↓為表現量(或甲基化)比對照組高或低的基因數

Chemicals	↑	↓	Total	Chemicals	↑	↓	Total
Ampicillin	0	0	0	**Sodium citrate**	0	0	0
AresnicTrioxide	127	86	213	**Tributyltin chloride**	0	0	0
Azathioprine	0	0	0	**Urethane**	0	0	0
Benzo[a]pyrene	0	0	0	**17α-ethynylestradiol_dose1**	1745	1903	3648
Bis(2-ethylhexyl) phthalate	0	1	1	**17α-ethynylestradiol_dose3**	0	0	0
Bis(tributyltin) oxide	0	0	0	**17α-ethynylestradiol_dose2**	1211	1114	2325
Cobalt2chloride	0	1	1	**17β-estradiol_dose1**	1127	1368	2495
Cyclophosphamide	0	0	0	**17β-estradiol_dose3**	0	0	0
Cyclophosphamide_S9	0	1	1	**17β-estradiol_dose2**	44	39	83
CyclosporineA	0	0	0	**BPA_dose1**	18	11	29
Deoxynivalenol	0	0	0	**BPA_dose3**	0	0	0
Diazinon	0	0	0	**BPA_MCF10_10^-5**	693	304	997
Dibutyltin dichloride	0	1	1	**BPA_MCF10_10^-6**	1	4	5
Fingolimod	0	0	0	**BPA_dose2**	0	0	0
Fluoxetine	0	0	0	**Daidzein_dose1**	461	467	928
Furosemide	0	0	0	**Daidzein_dose3**	6	3	9
Lindane	796	651	1447	**Daidzein_dose2**	19	62	81
Mannitol	0	0	0	**Diethylstilbestrol_dose1**	1126	1537	2663
Methylmercury	0	0	0	**Diethylstilbestrol_dose3**	0	1	1
Mono-2-ethylhexyl	0	0	0	**Diethylstilbestrol_dose2**	2109	1992	4101
Mycophenolic acid	38	34	72	**Genistein_dose1**	2296	2427	4723
Ochratoxin A	1203	1523	2726	**Genistein_dose3**	0	3	3
Ochratoxin A S9	1	1	2	**Genistein_dose2**	50	48	98
Prednisolone	0	0	0	**Nonylphenol_dose1**	1	0	1
Propanil	0	0	0	**Nonylphenol_dose3**	0	0	0
Rapamycin	12	9	21	**Nonylphenol_dose2**	2	1	3
Silver nitrate	0	0	0				

表 三 4 種疾病做 DEG 分析的 19 個基因群落內名單個數

BA46,BA10為Boardman area編號

Disease	↑	↓	Total
Bipolar disorder_dorsolateral prefrontal cortex	1	0	1
Bipolar disorder _dorsolateral prefrontal cortex_male	0	0	0
Bipolar disorder_dorsolateral prefrontal cortex_female	0	0	0
Bipolar disorder_orbital prefrontal cortex	0	1	1
Bipolar disorder_orbital prefrontal cortex_male	1	0	1
Bipolar disorder_orbital prefrontal cortex_female	0	0	0
Bipolar disorder_prefrontal cortex_BA46	0	0	0
Bipolar disorder_prefrontal cortex_BA10	0	0	0
Bipolar disorder_lymphocytes	0	0	0
Schizophrenia_prefrontal cortex_BA46	0	0	0
Schizophrenia_prefrontal cortex_BA10	0	0	0
Major depression_prefrontal cortex_BA10	0	0	0
Major depression_dorsolateral prefrontal cortex	0	0	0
Parkinson's disease_lateral substantia nigra	1595	1179	2774
Parkinson's disease_ lateral substantia nigra _male	1362	1140	2502
Parkinson's disease_ lateral substantia nigra _female	89	54	143
Parkinson's disease_prefrontal cortex_BA9	85	87	172
Parkinson's disease_prefrontal cortex_BA9_male	1	1	2
Parkinson's disease_prefrontal cortex_BA9_female	0	0	0

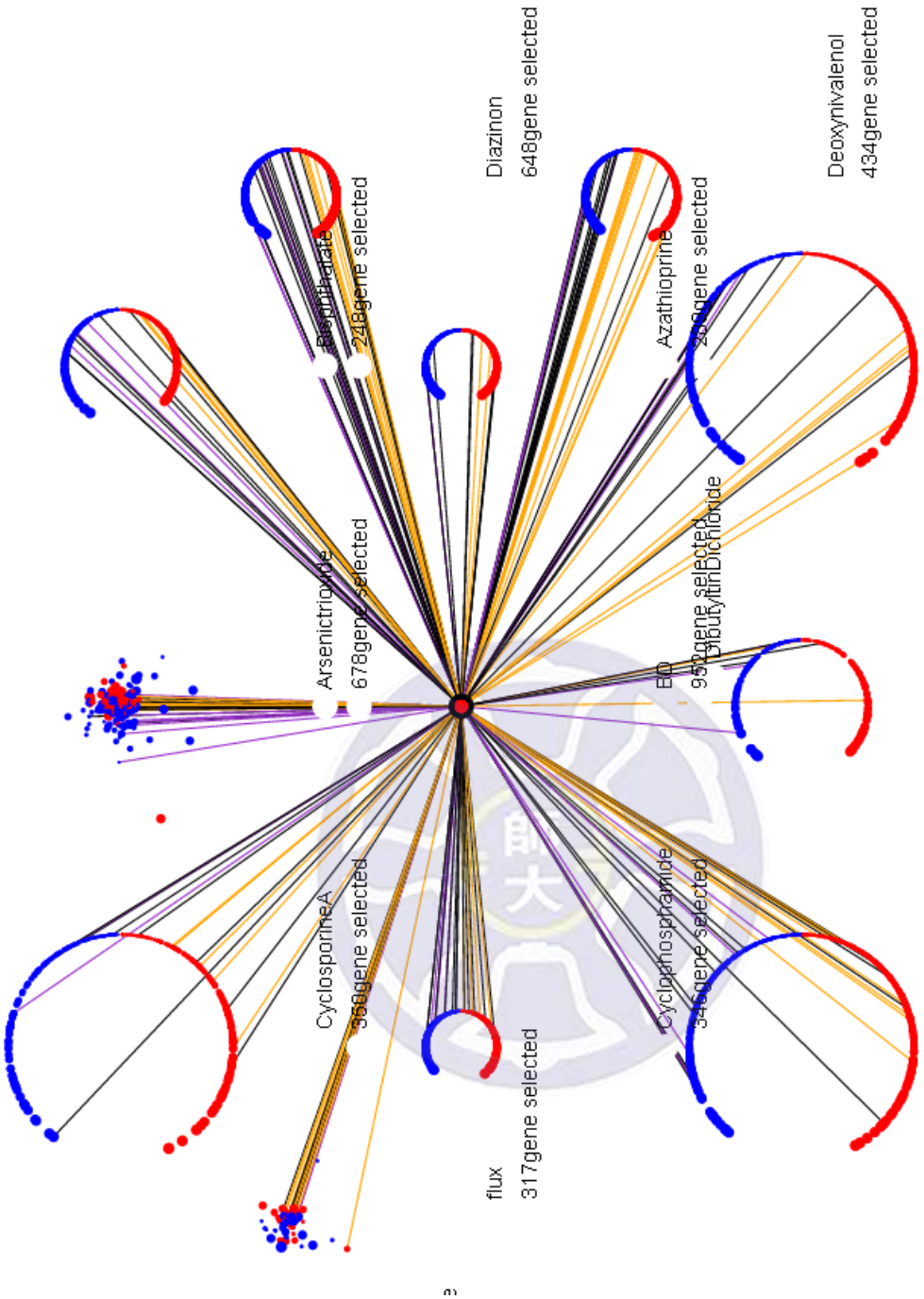

圖 二十四 化學物質與疾病基因群落交集分析過程圖

圖中最中間全部節點位置皆設為同一點的基因群落為躁鬱症的DEG群落，周邊依序為Arsenictrioxide、Bisphathlate、Diazinon等11種化學物質的DEG群落，群落間連線代表交集基因

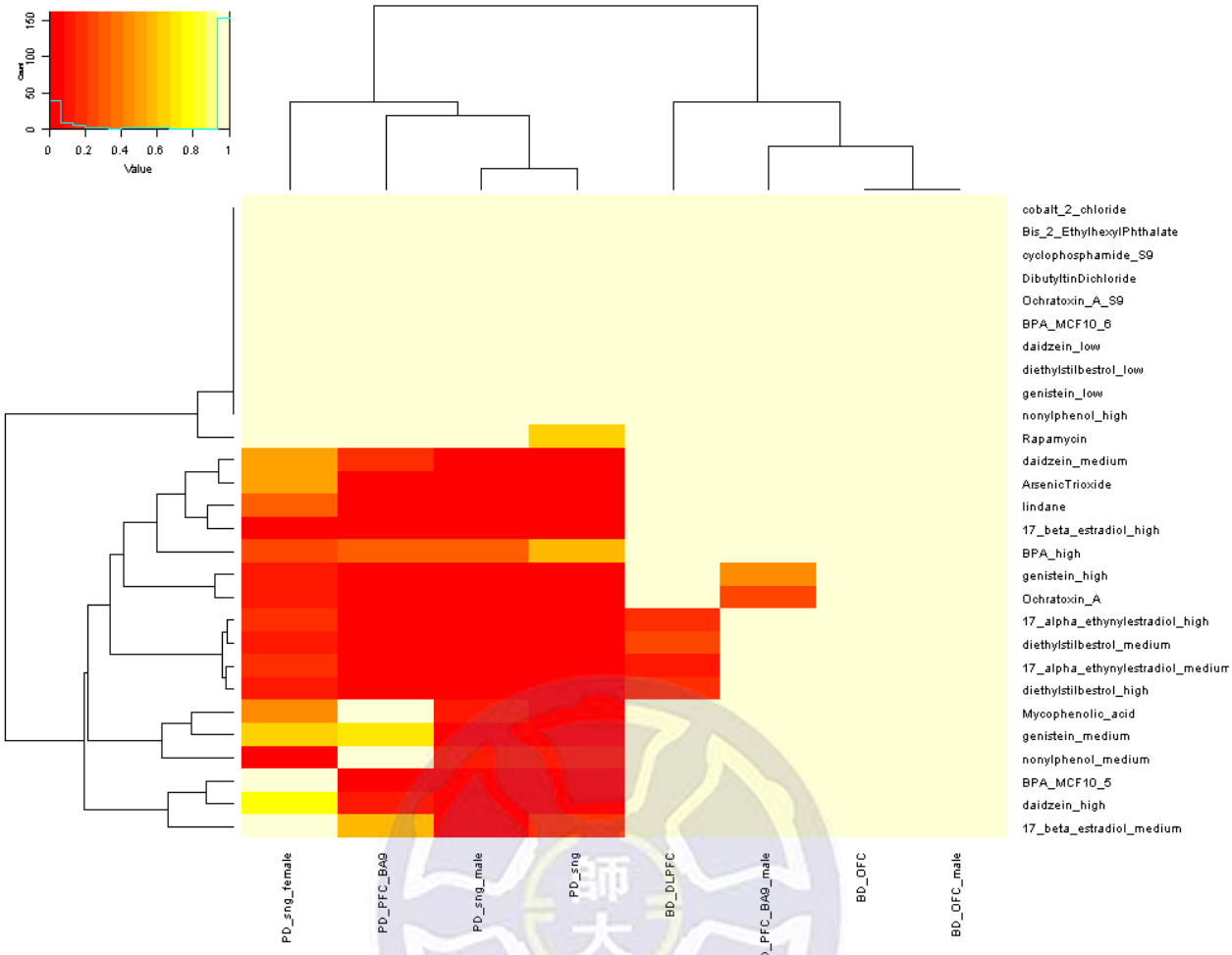

圖 二十五 化學物質和精神疾病做雙群落交集分析結果

橫座標為8組精神病基因群落，縱座標為28組化學物基因群落，中間28x8=224個色塊分別為所對應兩組基因群落交集分析後回傳的超幾何分布公式換算p-value，在這色塊矩陣中因為p值<0.01而成鮮紅色的色塊共有31塊。

表 四 通過篩選的 31 對雙群落分析結果中，其中包含的化學物質

Diseases 欄位中 PD 為帕金森氏症， PFC 為額葉皮質，sng 為黑質。

化學物質中包含部分環境賀爾蒙、工業排放廢料、內分泌調節藥物、植物次級代謝物等。

Diseases	**Chemical**
PD_sng	Arsenic trioxide, Lindane, Ochratoxin A, 17-alpha-ethynylestradiol, 17-beta-estradiol, BPA, Daidzein, Diethylstilbestrol, Genistein
PD_sng_male	Arsenic trioxide, Lindane, Ochratoxin A, 17-alpha-ethynylestradiol, 17-beta-estradiol, BPA, Daidzein, Diethylstilbestrol, Genistein
PD_PFC_BA9	Lindane, Ochratoxin A, 17-alpha-ethynylestradiol, BPA, Diethylstilbestrol, Genistein

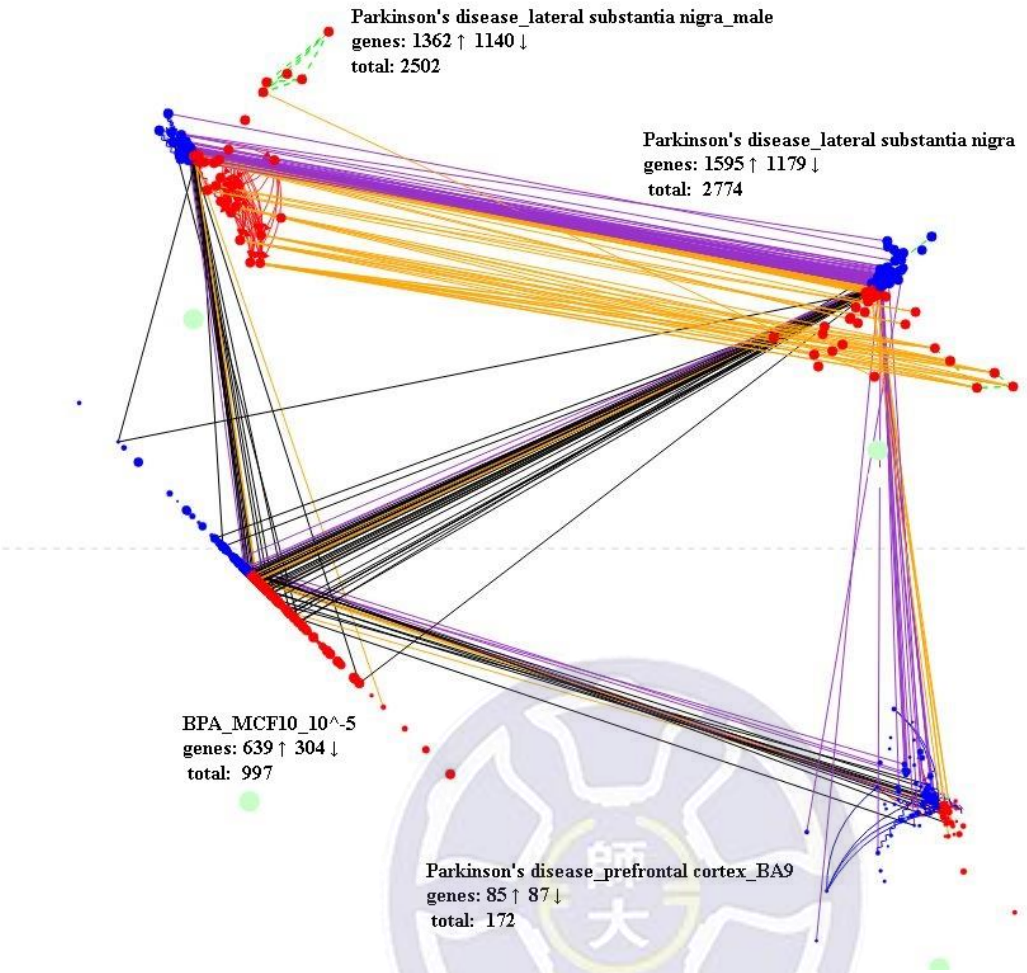

圖 二十六 雙酚A和帕金森氏症雙基因群落交集分析

其中左下角BPA_MCF10_10^5為雙酚A在10e-5M濃度下對MCF10細胞株造成的顯著差異基因群落，另外三個為帕金森式症不同腦切片樣本組所統計出來的顯著差異基因群落，群落間連線分別代表重合四個基因名單中彼此重合部分。紫色線表示該基因在兩群落中「皆為正調控」(在兩個樣本組中皆為實驗組表現量比對照組高)，橙色線表示「皆為負調控」(兩個樣本組中皆為對照組表現量較高)，黑色線表示在兩群落中「調控方向相反」，其中可以看見三組帕金森氏症相關基因群落彼此重合基因大多調控方向相同

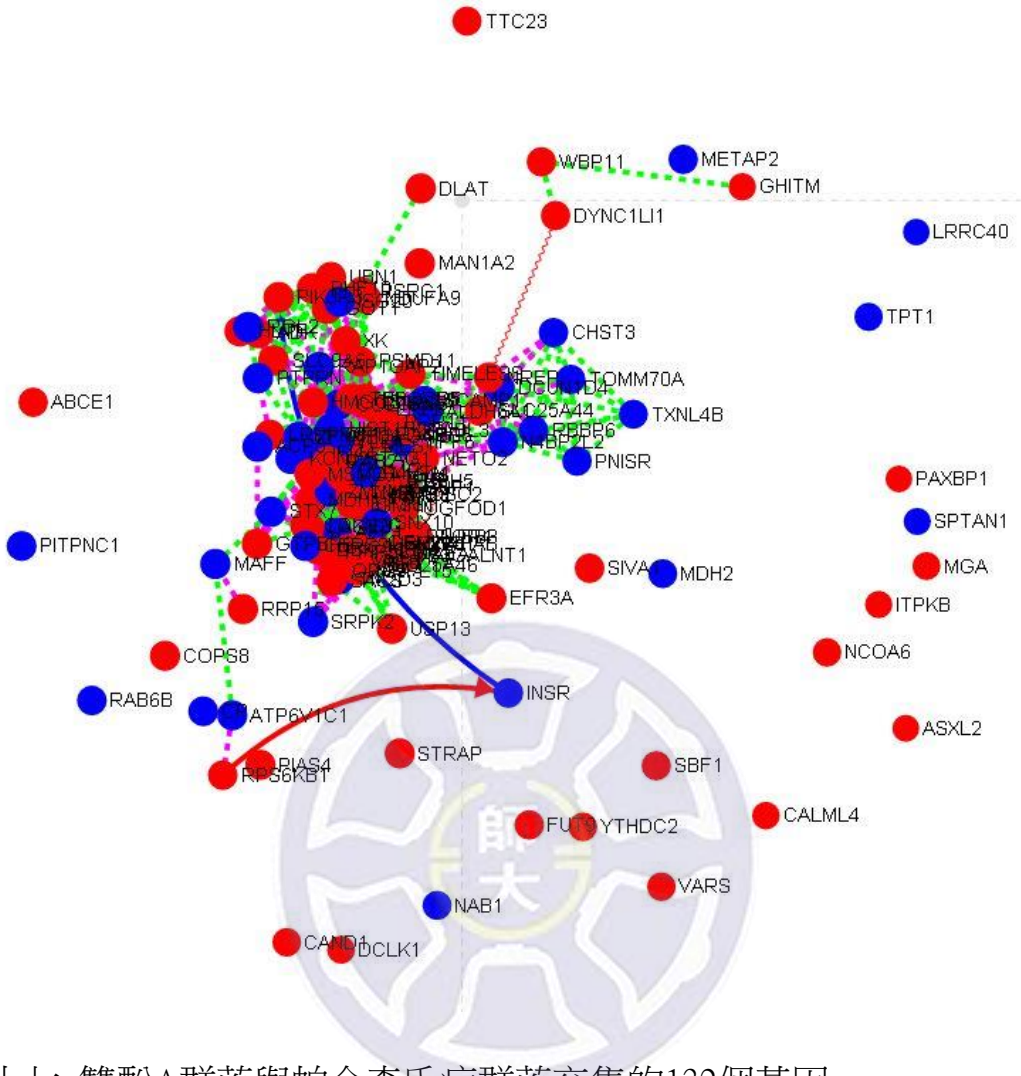

圖 二十七 雙酚A群落與帕金森氏症群落交集的132個基因

在「BPA_MCF10_10^-5基因群落」中挑出和「Parkinson’ s disease_lateral substansia nigra基因群落」重合的132個基因獨立建立一個新的基因群落，紅色、藍色分別表示基因在「BPA_MCF10_10^-5基因群落」中的調控方向。其中可看到中間有一群節點因為對應基因表現量彼此具有高度相關性(綠色、粉紅色虛線)而在圖中集結一起。其中也有一些基因已經被Reactome收錄紀載彼此具有蛋白質交互作用(實心箭頭、Z字線)。

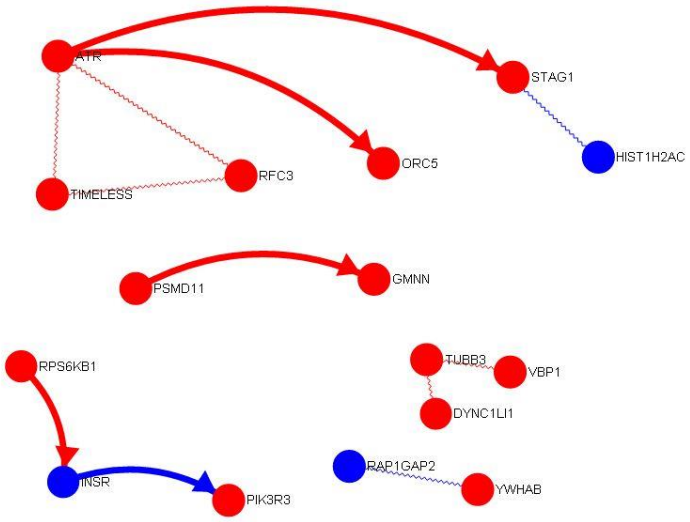

圖 二十八 雙酚A群落與帕金森氏症群落交集的蛋白質交互作用

其中以ATR為主的6個相關基因皆為細胞分裂S-Phase生化途徑中的參與基因。其他7筆蛋白質交互作用連線雖然尚無法找到明顯的共參與生化途徑，但仍能作為後續研究的參考目標。

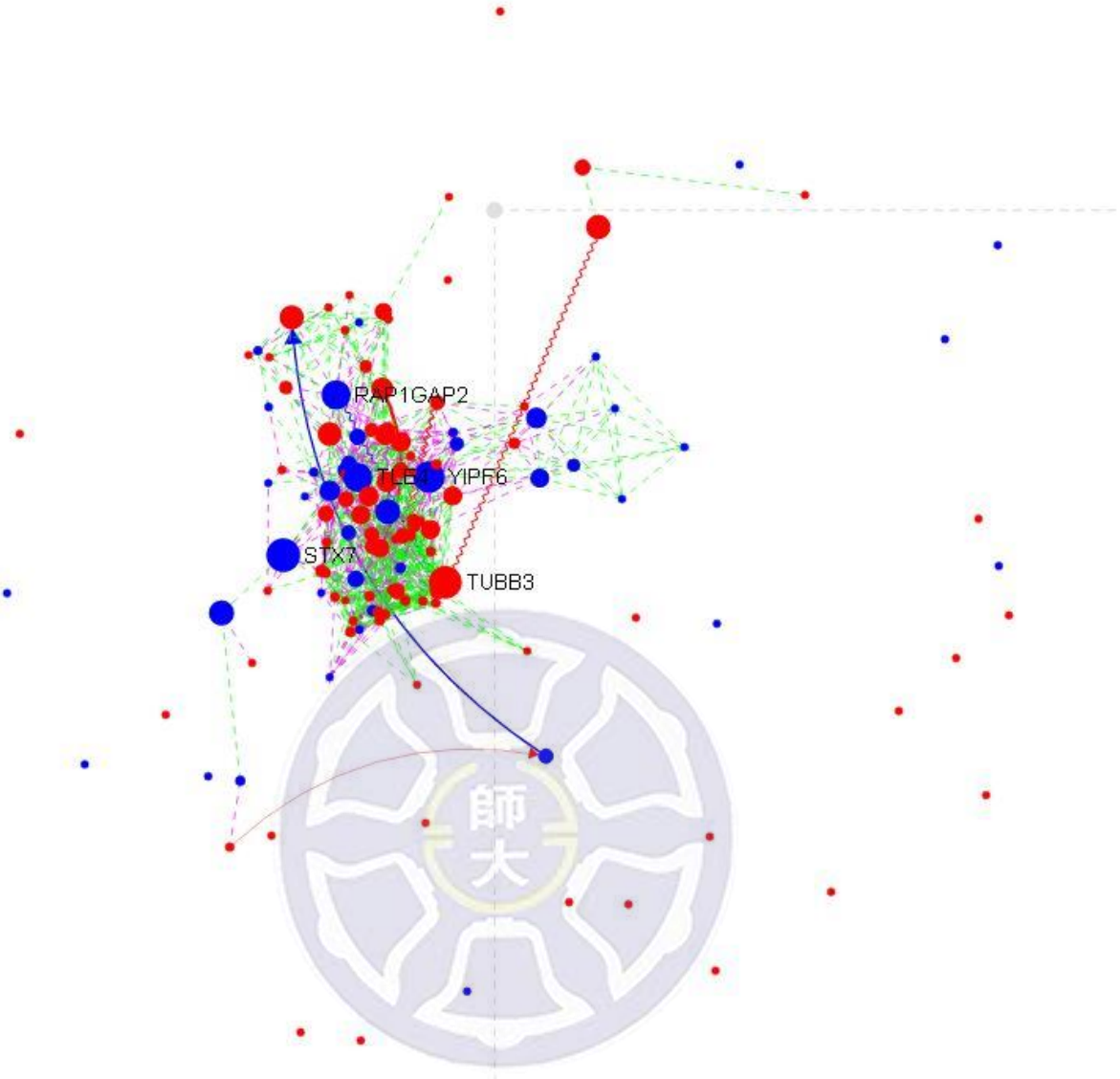

圖 二十九 132個重合基因於網絡中的中間度指標

「BPA_MCF10_10^-5基因群落」和「Parkinson’ s disease_lateral substansia nigra基因群落」重合的132個基因間開啟中間度指標功能後的圖表，以節點大小表示該節點於網絡中的最短路徑中間度指標，其中節點越大表示該基因於此132個基因組成網絡中越有可能扮演主要調控者的角色。

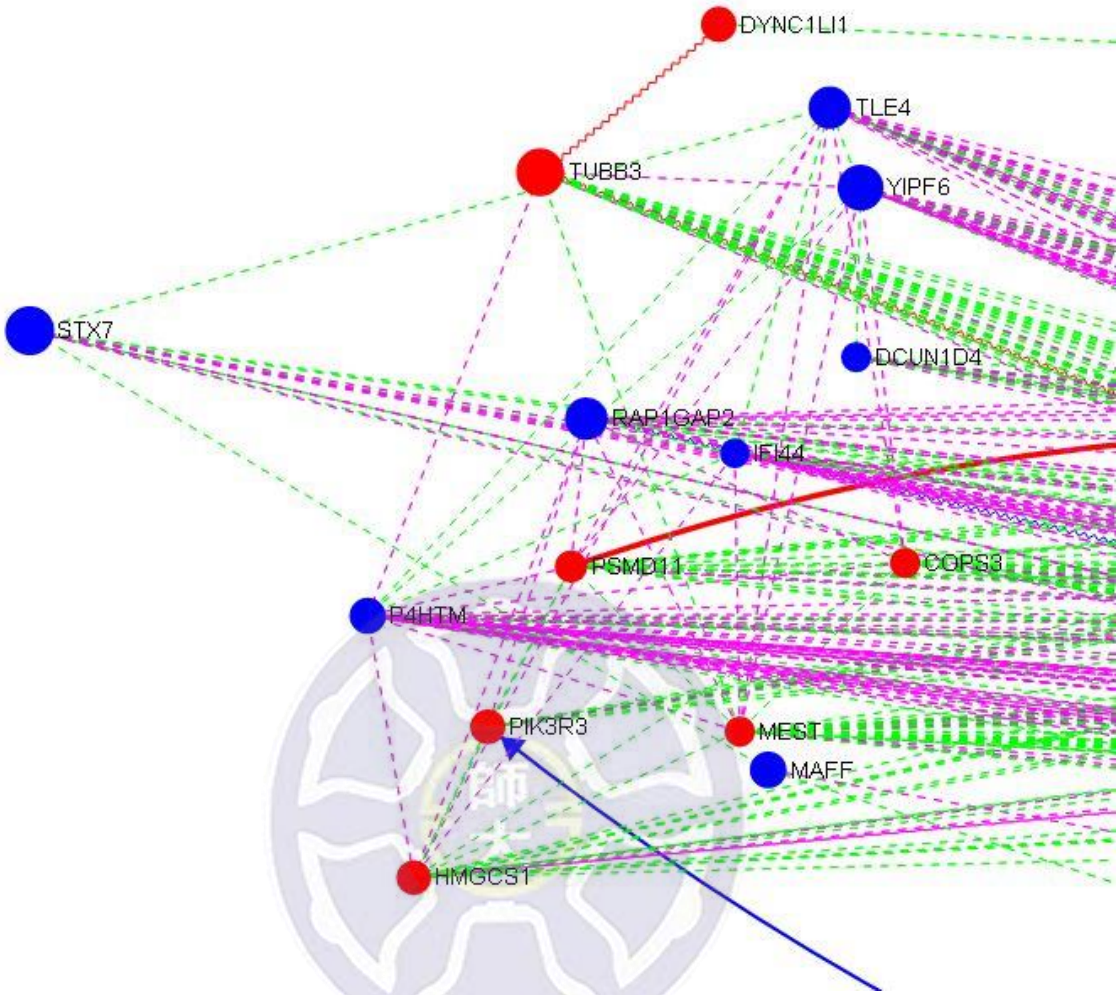

圖 三十132個重合基因中中間度指標最大的前15名基因

4 結果與討論

由於生物晶片實驗操作過程中，容易有許多人為操作以及設備本身造成的數據雜訊，加上晶片本身樣本註解資訊可能誤差，不同的樣本分組、生物晶片探針訊號換算方法、顯著表現基因統計量、超幾何分布公式選用，都會決定使用 AryNet 做病理因子關聯性分析時結果準確度。本次統計採用 FDR 方式校正晶片樣本組中的每個基因顯著差異統計值，造成許多樣本組找不到顯著差異基因。若將 FDR 校正過濾網移除，則 36 個化學物質與疾病的 DEG 基因交集分析結果中，Azathioprine、Furosemide、Fluoxetine 等已有前人研究顯示跟精神病具有關聯性的藥物，都能偵測出和躁鬱症、憂鬱症等精神病樣本顯著差異基因名單具有高度重合。但因為高通量基因篩選若採用原本為校正 P-value，又容易造成篩選誤差過大，固本實驗仍採用校正後 q-value 作為篩選門檻。

雙酚 A 為一種類雌激素環境賀爾蒙，其作用下游基因非常廣泛，若於 CTD 網站搜尋其相關基因數可達 18000 筆以上，其中包含至少超過 2000 個基因會參與於細胞分裂生物途徑。

本次分析採用數據來自不同樣本組織(雙酚 A 樣本來源為乳癌細胞株，帕金森氏症樣本則來自死去患者的大腦組織)，加上無法得知病患生前是否有接觸類似環境化學物質，這些都可能對本次實驗分析造成雜訊干擾。

本實驗結果尚無法證明帕金森氏症和雙酚 A 兩者之間因果關聯，僅能對兩者關連性研究提供可能研究對象參考。

5 結論

本研究以建置一個互動式線上工具系統的方式，將故有生物晶片分析處理技術、基因資訊、生物途徑資訊等前人研究加以統整併運用，預期將達到下列成果：

1. AryNet 資料庫收錄精神疾病和環境化學物質的相關晶片樣本，將樣本註解資訊加以分類歸納後並提供一個可依據年齡、性別等更明確資訊來進行搜尋的操作介面，可方便從事相關領域的研究者縮短搜尋網路開源數據的時間。
2. AryNet 提供即時性的參數調整結果輸出，幫助研究者進行晶片分析及基因群落分析時，可任意調整相關參數並迅速比對結果，減少數據分析時各種參數選擇的不確定性。
3. AryNet 為一個 HTML 通性協定為基礎的線上免費操作工具，使用者不用進行任何多餘的軟體安裝，靠垂手可得的一般網頁瀏覽器軟體即可正常使用其晶片分析數據功能，減少生物晶片分析的研究成本。
4. AryNet 收錄人類基因體及 DNA 甲基化數據共約 1100 個，搭配上內建已知生物途徑資訊，可幫助預測未知的基因交互關聯與表觀遺傳機制，可協助生化方面研究。且其提供的多基因網絡綜合分析的視覺化功能，能幫助尋找化學物質、精神疾病關聯性因子的研究人員縮小研究範圍。

6 未來研究方向

AryNet 目前僅提供人類基因體學相關資訊，且其收錄的生物晶片僅有 4 種常用平台，未來希望能夠朝跨物種、多種平台來做擴充。為達此目的，必須安裝更多生物晶片平台的相關演算法和晶片探針註解，並且需製作跨物種間同源基因的比對檢索表，此功能若完成，將大大增加 AryNet 收錄資料庫以及統整前人實驗的完整性。

另外 AryNet 內建的 REngine 計算引擎，目前尚無法進行平行運算，大幅限制了 AryNet 在服務多使用者時的運算效能，也浪費伺服器本身的 CPU 資源。目前市面上已有付費的 R 運算引擎-revolution R，即擁有平行運算能力，但其程式撰寫技巧目前仍為商業機密，期盼未來能引用此技術以增進 AryNet 的統計運算效能。

無論基因網絡的視覺化方式、使用者操作介面或是內部程式運作方法，AryNet 都還有很大的改善空間，隨著硬體設備升級，AryNet 也將能提供更大容量的計算。因此本研究系統以 MVC 方式分割撰寫，並且選用物件導向程式設計，使其在日後維護與更新上都能夠有較大的彈性，期盼 AryNet 能在未來基因體學以及病理因子研究領域中帶來正向幫助。

7 參考文獻

1. Tondo, L., et al., *Suicide attempts in bipolar disorders: comprehensive review of 101 reports.* Acta Psychiatr Scand, 2015.
2. Gabbard, G.O., *Psychotherapy in psychiatry.* Int Rev Psychiatry, 2007. **19**(1): p. 5-12.
3. Lee, S.H., et al., *Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs.* Nat Genet, 2013. **45**(9): p. 984-94.
4. Moutoussis, M., G.W. Story, and R.J. Dolan, *The computational psychiatry of reward: broken brains or misguided minds?* Front Psychol, 2015. **6**: p. 1445.
5. Etain, B., et al., *Beyond genetics: childhood affective trauma in bipolar disorder.* Bipolar Disord, 2008. **10**(8): p. 867-76.
6. Maniglio, R., *The impact of child sexual abuse on the course of bipolar disorder: a systematic review.* Bipolar Disord, 2013. **15**(4): p. 341-58.
7. Zhao, H., et al., *Genome-wide DNA methylome reveals the dysfunction of intronic microRNAs in major psychosis.* BMC Med Genomics, 2015. **8**(1): p. 62.
8. Dong, E., et al., *DNA-methyltransferase1 (DNMT1) binding to CpG rich GABAergic and BDNF promoters is increased in the brain of schizophrenia and bipolar disorder patients.* Schizophr Res, 2015. **167**(1-3): p. 35-41.
9. Belvederi Murri, M., et al., *The HPA axis in bipolar disorder: Systematic review and meta-analysis.* Psychoneuroendocrinology, 2015. **63**: p. 327-342.
10. Nestler, E.J., et al., *Epigenetic Basis of Mental Illness.* Neuroscientist, 2015.
11. Xu, Z., et al., *DNA-based hybridization chain reaction amplification for assaying the effect of environmental phenolic hormone on DNA methyltransferase activity.* Anal Chim Acta, 2014. **829**: p. 9-14.
12. Tang, W.Y., et al., *Neonatal exposure to estradiol/bisphenol A alters promoter methylation and expression of Nsbp1 and Hpcal1 genes and transcriptional programs of Dnmt3a/b and Mbd2/4 in the rat prostate gland throughout life.* Endocrinology, 2012. **153**(1): p. 42-55.
13. Kitraki, E., et al., *Developmental exposure to bisphenol A alters expression and DNA methylation of Fkbp5, an important regulator of the stress response.* Mol Cell Endocrinol, 2015. **417**: p. 191-9.
14. Kundakovic, M., et al., *DNA methylation of BDNF as a biomarker of early-life adversity.* Proc Natl Acad Sci U S A, 2015. **112**(22): p. 6807-13.
15. Hovenkamp-Hermelink, J.H., et al., *Low stability of diagnostic classifications of anxiety disorders over time: A six-year follow-up of the NESDA study.* J Affect Disord, 2015. **190**: p. 310-315.
16. Choudhuri, S., *From Waddington's epigenetic landscape to small noncoding RNA: some important milestones in the history of epigenetics research.* Toxicol Mech Methods, 2011.

21(4): p. 252-74.
17. Holliday, R., *Epigenetics: an overview.* Dev Genet, 1994. **15**(6): p. 453-7.
18. Pandian, G.N. and H. Sugiyama, *Strategies to modulate heritable epigenetic defects in cellular machinery: lessons from nature.* Pharmaceuticals (Basel), 2012. **6**(1): p. 1-24.
19. Vaiserman, A.M., *Epigenetic programming by early-life stress: Evidence from human populations.* Dev Dyn, 2014.
20. Liyanage, V.R., et al., *DNA modifications: function and applications in normal and disease States.* Biology (Basel), 2014. **3**(4): p. 670-723.
21. Szyf, M., *Lamarck revisited: epigenetic inheritance of ancestral odor fear conditioning.* Nat Neurosci, 2014. **17**(1): p. 2-4.
22. Mao, Z., et al., *Paternal BPA exposure in early life alters Igf2 epigenetic status in sperm and induces pancreatic impairment in rat offspring.* Toxicol Lett, 2015. **238**(3): p. 30-8.
23. Pierre, J.M., *Faith or delusion? At the crossroads of religion and psychosis.* J Psychiatr Pract, 2001. **7**(3): p. 163-72.
24. Belusa, L., A.M. Selzer, and B.D. Partecke, *[Description of Dupuytren disease by the Basel physician and anatomist Felix Plater in 1614].* Handchir Mikrochir Plast Chir, 1995. **27**(5): p. 272-5.
25. Jorgensen, K.N., et al., *First- and second-generation antipsychotic drug treatment and subcortical brain morphology in schizophrenia.* Eur Arch Psychiatry Clin Neurosci, 2015.
26. Edlinger, M., et al., *Risk of violence of inpatients with severe mental illness--do patients with schizophrenia pose harm to others?* Psychiatry Res, 2014. **219**(3): p. 450-6.
27. Shibre, T., et al., *Suicide and suicide attempts in people with severe mental disorders in Butajira, Ethiopia: 10 year follow-up of a population-based cohort.* BMC Psychiatry, 2014. **14**: p. 150.
28. Aydin, A., et al., *Mood and metabolic consequences of sleep deprivation as a potential endophenotype' in bipolar disorder.* J Affect Disord, 2013. **150**(2): p. 284-94.
29. Song, J., et al., *Bipolar disorder and its relation to major psychiatric disorders: a family-based study in the Swedish population.* Bipolar Disord, 2015. **17**(2): p. 184-93.
30. Szczepankiewicz, A., *Evidence for single nucleotide polymorphisms and their association with bipolar disorder.* Neuropsychiatr Dis Treat, 2013. **9**: p. 1573-82.
31. Etain, B., et al., *Childhood trauma is associated with severe clinical characteristics of bipolar disorders.* J Clin Psychiatry, 2013. **74**(10): p. 991-8.
32. Bratlien, U., et al., *Environmental factors during adolescence associated with later development of psychotic disorders - a nested case-control study.* Psychiatry Res, 2014. **215**(3): p. 579-85.
33. Korach, K.S., M. Metzler, and J.A. McLachlan, *Diethylstilbestrol metabolites and analogs. New probes for the study of hormone action.* J Biol Chem, 1979. **254**(18): p. 8963-8.
34. Iguchi, T., *[Environmental endocrine disruptors].* Nihon Rinsho, 1998. **56**(11): p. 2953-62.

35. Bazer, F.W., et al., *Environmental factors affecting pregnancy: Endocrine disrupters, nutrients and metabolic pathways.* Mol Cell Endocrinol, 2014.
36. Kucka, M., et al., *Atrazine acts as an endocrine disrupter by inhibiting cAMP-specific phosphodiesterase-4.* Toxicol Appl Pharmacol, 2012. **265**(1): p. 19-26.
37. Li, R., et al., *Effects of DEHP on endometrial receptivity and embryo implantation in pregnant mice.* J Hazard Mater, 2012. **241-242**: p. 231-40.
38. Multigner, L., et al., *[Environment and secular sperm trend. Stallion's semen quality during the last two decades].* Rev Epidemiol Sante Publique, 2000. **48 Suppl 2**: p. 2s72-8.
39. Xu, H., et al., *The impact of endocrine-disrupting chemicals on oxidative stress and innate immune response in zebrafish embryos.* Environ Toxicol Chem, 2013. **32**(8): p. 1793-9.
40. Perera, F., et al., *Prenatal bisphenol a exposure and child behavior in an inner-city cohort.* Environ Health Perspect, 2012. **120**(8): p. 1190-4.
41. Singh, S. and S.S. Li, *Epigenetic effects of environmental chemicals bisphenol A and phthalates.* Int J Mol Sci, 2012. **13**(8): p. 10143-53.
42. Elyakim, E., et al., *hsa-miR-191 is a candidate oncogene target for hepatocellular carcinoma therapy.* Cancer Res, 2010. **70**(20): p. 8077-87.
43. Kundakovic, M., et al., *Sex-specific epigenetic disruption and behavioral changes following low-dose in utero bisphenol A exposure.* Proc Natl Acad Sci U S A, 2013. **110**(24): p. 9956-61.
44. Kruppa, J. and K. Jung, *Set-Based Test Procedures for the Functional Analysis of Protein Lists from Differential Analysis.* Methods Mol Biol, 2016. **1362**: p. 143-56.
45. Mahajan, G. and S.C. Mande, *From System-Wide Differential Gene Expression to Perturbed Regulatory Factors: A Combinatorial Approach.* PLoS One, 2015. **10**(11): p. e0142147.
46. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Res, 1999. **27**(1): p. 29-34.
47. Vastrik, I., et al., *Reactome: a knowledge base of biologic pathways and processes.* Genome Biol, 2007. **8**(3): p. R39.
48. Mlecnik, B., et al., *PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W633-7.
49. Kutmon, M. and A. Riutta, *WikiPathways: capturing the full diversity of pathway knowledge.* 2015.
50. Viti, C., et al., *High-throughput phenomics.* Methods Mol Biol, 2015. **1231**: p. 99-123.
51. Delgado, S., et al., *A Novel Representation of Genomic Sequences for taxonomic clustering and visualization by means of Self-Organizing Maps.* Bioinformatics, 2014.
52. Zheng, Y., et al., *Standardization efforts enabling next-generation sequencing and microarray based biomarkers for precision medicine.* Biomark Med, 2015.
53. Lim, W.K., et al., *Comparative analysis of microarray normalization procedures: effects on*

reverse engineering gene networks. Bioinformatics, 2007. **23**(13): p. i282-8.

54. Millenaar, F.F., et al., *How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results.* BMC Bioinformatics, 2006. **7**: p. 137.
55. Zhao, X., et al., *[Whole genome methylation profiles of myelodysplastic syndrome and its diagnostic value].* Zhonghua Xue Ye Xue Za Zhi, 2014. **35**(10): p. 944-8.
56. Barrett, T., et al., *NCBI GEO: mining millions of expression profiles--database and tools.* Nucleic Acids Res, 2005. **33**(Database issue): p. D562-6.
57. Kurscheid, S., et al., *Chromosome 7 gain and DNA hypermethylation at the HOXA10 locus are associated with expression of a stem cell related HOX-signature in glioblastoma.* Genome Biol, 2015. **16**: p. 16.
58. Healy, K. and J. Moody, *Data Visualization in Sociology.* Annu Rev Sociol, 2014. **40**: p. 105-128.
59. Warner, J.L., et al., *Seeing the forest through the trees: uncovering phenomic complexity through interactive network visualization.* J Am Med Inform Assoc, 2014.
60. H. Jeong, B. Tombor, and R. Albert, *The large-scale organization of metabolic networks.* nature, 2000. **407**: p. 651.
61. Li, M., et al., *A Topology Potential-Based Method for Identifying Essential Proteins from PPI Networks.* IEEE/ACM Trans Comput Biol Bioinform, 2015. **12**(2): p. 372-83.
62. Davidson, E. and M. Levin, *Gene regulatory networks.* Proc Natl Acad Sci U S A, 2005. **102**(14): p. 4935.
63. Pedersen, F., et al., *Principal component analysis of dynamic positron emission tomography images.* Eur J Nucl Med, 1994. **21**(12): p. 1285-92.
64. Adai, A.T., et al., *LGL: creating a map of protein function with an algorithm for visualizing very large biological networks.* J Mol Biol, 2004. **340**(1): p. 179-90.
65. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
66. Kuhn, M., et al., *STITCH 4: integration of protein-chemical interactions with user data.* Nucleic Acids Res, 2014. **42**(Database issue): p. D401-7.
67. Stark, C., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.
68. Glaab, E., J.M. Garibaldi, and N. Krasnogor, *ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization.* BMC Bioinformatics, 2009. **10**: p. 358.
69. Maciel, A., et al., *Surgical model-view-controller simulation software framework for local and collaborative applications.* Int J Comput Assist Radiol Surg, 2011. **6**(4): p. 457-71.