行政院國家科學委員會專題研究計畫 成果報告

資料探勘技術在基因體與蛋白體資料庫方面之應用(2/2)
研究成果報告(完整版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 95-2221-E-216-051-
執 行 期 間 ：95 年 08 月 01 日至 96 年 07 月 31 日
執 行 單 位 ：中華大學資訊工程學系

計 畫 主 持 人：劉志俊
共 同 主 持 人：林恩仲、孟子青、張慧玫
計畫參與人員：博士班研究生-兼任助理：楊為宏、林怡伸、吳明家、曾坤源、蒲宗賢
碩士班研究生-兼任助理：黃柏諺、游傑泰、黃裕翔

處 理 方 式 ：本計畫可公開查詢

中 華 民 國 96 年 10 月 31 日

行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※ 資料探勘技術在基因體與蛋白體資料庫方面之應用 ※
※　　　　　　　　　　(2/2)　　　　　　　　　　　※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：■個別型計畫　　□整合型計畫

計畫編號：　NSC 95-2221-E-216 -051

執行期間：　95 年 08 月 01 日 至 96 年 07 月 31 日

計畫主持人：劉志俊

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

□出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

執行單位：中華大學資訊工程學系

中　華　民　國　96　年　10　月　30　日

行政院國家科學委員會專題研究計畫成果報告

資料探勘技術在基因體與蛋白體資料庫方面之應用(2/2)

The Applications of Data Mining Technologies in Genome and Proteome Databases (2/2)

計畫編號：NSC 95-2221-E-216 -051
執行期限：95 年 08 月 01 日 至 96 年 07 月 31 日

主持人：劉志俊　中華大學資訊工程學系
計畫參與人員：張慧玫、楊為宏、林怡伸、吳明家、曾坤源、蒲宗賢、黃裕翔、黃柏諺、游傑泰 中華大學資訊工程學系

一、中文摘要

本計劃的主要研究目的在於將資料探勘的技術應用在基因體與蛋白體資料庫之上，並結合以往我們與其他學術單位合作所建立的基因體與蛋白體資料庫，豬與土雞功能性基因體資料庫、脂肪細胞功能性基因體資料庫，以及結合共同主持人對基因體與蛋白體的專業知識，以實際的資料進行資料探勘的分析，解決在基因體與蛋白體資料庫方面的重要問題,包含由 cDNA 探針的最佳化選取問題、定序原始序列中徹底移除選殖轉接器(clone adapter)序列的方法、以及 DSP 蛋白質家族的結合專一性模型。生物反應路徑的資料在功能性基因體中是極其重要的一環。我們提出一種從 PubMed 文獻資料庫發現並整合反應路徑的方法。此方法不僅能簡化基因調控路徑的尋找工作，其分析出的豬生殖與胚胎發育的基因調控路徑可能供給研究人員研究利用。

關鍵詞：PubMed、MeSH、豬胚胎、反應路徑、反應路徑整合、反應路徑探勘

二、緣由與目的

隨著生物科技不斷的進步，產生了龐大的定序資料與研究文獻資料。對這些文獻，必需快速有效地從中整理綜合出有用的結論來加以應用。

對於生物研究人員而言，這些大量文獻資料實際上埋藏許多線索，足以使得生物實驗中亟需突破的有趣問題，得以找出新的切入點來尋找出生物問題的答案。例如：豬的繁殖力，以經濟因素，極需改進。由於豬肉為高比例人口取用的日常肉食，增加豬隻的繁殖力便能大幅提高其養殖業經濟效益。本研究正是以豬的繁殖力作為生物主題，來做資訊工具應用的對象來研究探討。

文獻探勘技術，主要是對大量論文文字或初始數據做自動化分析處理，以便萃取出有用資料，再形成一個知識庫。目前的研究主流有二種，第一種是以語言學方式進行分析[1][2][3][4][5]，第二種是以同時出現(co-occurence)的關鍵字搜尋再以統計作分析[6][7]。第一種語言學方式的分析是將文獻以傳統的語言學技術來進行分析，使之轉成可依規則處理的形態，而其分析結果理論上會是較準確的。可是由於語言的規則極多，使得程式的設計難以涵蓋所有的可能情況。另一種關鍵字搜尋的分析是依基因或蛋白質的名稱或關鍵字來搜尋文獻，以兩名稱或關鍵字的共同出現率來判定是否有關聯性，若有便將之轉成電腦可處理的規則，而這樣較簡化一點的方式，其分析往往會有較佳的結果。

Stapley 等人於 2000 年提出了一篇文

獻，其內容提出一個BioBibliometric distance的計算公式，藉由此公式的計算可以得到兩個基因間的關聯性，再對這關聯性的系統設定一個門檻值來剔除較弱的基因組合，從而推測出生化反應路徑[8]。

Famili [9]等人於2003年提出一篇文獻，其中提到一種化學計量矩陣（Stoichiometric matrix）的方法，將化學反應式以矩陣的方式表達，並可將之做簡化合併，使得到整合過的生化反應路徑。目前生物醫學方面還有相同物質卻有不同命名的問題存在，以致於很難輕易的從中取出完整的化學反應式。

三、研究成果

(一) 反應路徑資料探勘系統

一個反應路徑是由一組相關的反應連接而成，所以要找出特定的反應路徑，必需先找到其中所有反應，也就是長度為1的反應路徑集合。以下將介紹本系統的反應路徑探勘技術使用的方法，稱之為「關聯度」方法及其驗證。

一般而言，研究論文多呈現正面的結果，較少提出負面的結論，例如：研究的結果可顯示某兩個蛋白分子具相關性，甚少報導兩個蛋白分子無相關的研究結果，於是本研究假設若同一篇文獻裡出現兩個蛋白分子，則兩者應具有正向的關聯性。

任取兩個蛋白分子稱之為A與B，將兩者在反應路徑論文集中出現的次數，分別定義為α及β，$(\alpha \cap \beta)$表示A、B同時出現在同一篇文獻中的數量，δ表示A、B之間的關聯度，計算方式如下：

$$\delta(\alpha,\beta)=\frac{2(\alpha\cap\beta)}{(\alpha+\beta)} \qquad (1)$$

本研究以實際已知的反應路徑驗證關聯度的計算是否符合實際的情況。KEGG的全名是 Kyoto Encyclopedia of Genes and Genomes，係為了利用基因訊息對更高層次和更複雜細胞活動和生物體行為計算推測而設計的。KEGG 中的生化反應路徑資料庫(PATHWAY database)整合分子間交互作用及相關反應的所有知識。因為係採用人工閱讀的方式整裡，因此資料的正確性最高。本研究選擇 KEGG 的 pathway 資料庫中的反應路徑圖做為驗證的資料來源，首先將 KEGG 裡 pathway 資料庫的圖依照其分類，取出各大分類中的小分類之任兩張圖中的兩個蛋白分子做為關聯度的計算依據，接著以不同的反應路徑距離進行關聯度的計算與比較。驗證之結果顯示出反應路徑距離越近的兩個蛋白分子之關聯度越大，此與本研究之假設相符合。

利用反應路徑論文集找出與使用者輸入的關鍵字有關的基因，則稱之為「候選反應基因尋找」。依據關聯度的計算找到反應路徑長度為一的兩兩基因對。利用關聯度與支持度之公式進行長度為一之反應路徑篩檢。其中支持度之計算方式如下所示：

$$\varepsilon(\alpha,\beta)=\frac{\log(\alpha\cap\beta)}{\log(Max(\alpha,\beta))} \qquad (2)$$

支持度可用於判定之前計算出之關聯度的可信度高低。藉由設定關聯度δ及支持度ε的門檻值，可將一些誤判為相關具長度為一之反應路徑剔除，以提昇系統的可信度。此外需限定出現蛋白分子的文獻篇數的下限限制，因此將出現於文獻篇數中不足 5 篇的蛋白分子剔除，以避免剛好被順便提到但真正出現於文獻中之篇數極少的蛋白分子之干擾。

(二) 反應路徑整合

延續找到的長度為一之基因對的集合，利用反應方程式平衡計算矩陣，將長度為一之反應路徑轉成矩陣的形式，再將反應方程式平衡計算矩陣進行初始化的動作，最後利用反應路徑整合條件，進行反應矩陣的整合。

首先列出全部要分析的長度為一之反應路徑的集合，如圖 2 所示，輸入的反應方程式共有 3 個，因為進行的反應路徑並非一般的生化反應，所以在此不考慮酵素。在本範例中參與反應的蛋白分子共有{ p300 , GADD45 , p53 , PCNA }4 個蛋白分子，故其反應方程式平衡計算矩陣的行數為 4，於是得到一個 4 * 3 的反應方程式平衡計算矩陣。

反應路徑1：p300 → p53

反應路徑2：GADD45 → PCNA

反應路徑3：p53 → GADD45

圖 2 輸入的反應路徑

下一步要在矩陣中填入數值，將反應路徑的係數填入矩陣中，反應路徑中的反應物以負數表示，反應出來的產物則以正數表示，若是與此列的反應無關則以 0 表示。圖 3 為圖 2 轉換完成後的結果。

$$S = \begin{array}{c} \begin{matrix} p300 & GADD45 & p53 & PCNA \end{matrix} \\ \begin{bmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 1 & -1 & 0 \end{bmatrix} \end{array}$$

圖 3 反應方程式平衡計算矩陣

接著將進行反應方程式平衡計算矩陣初始化的動作，即在反應方程式平衡計算矩陣的右側加上一個 n * n 的單位矩陣，如式子 3 所示：

$$T(0) = [\, S \mid I \,] \qquad (3)$$

式子 3 中 S 表示反應方程式平衡計算矩陣，I 表示單位矩陣。Rk 表示為反應路徑 k，k = 1 , 2 , 3 , … ,n，如圖 4 所示，R1 表示反應路徑 1，R2 則表示為反應路徑 2。

$$T(0) = \begin{array}{c} \begin{matrix} p300 & GADD45 & p53 & PCNA & R1 & R2 & R3 \end{matrix} \\ \left[\begin{array}{cccc|ccc} -1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 1 \end{array}\right] \end{array} \begin{matrix} \\ R1 \\ R2 \\ R3 \end{matrix}$$

圖 4 初始化完成的反應方程式平衡計算矩陣

其中單位矩陣其意義是用來記錄由哪些長度為一之反應路徑可組成最後合併後的反應路徑，而矩陣中的數值則表示要組成最後合併完成的反應路徑需要由多少個長度為一之反應路徑組成。

反應路徑整合的條件是：第一，兩個反應路徑，此兩個反應路徑至少要有一個相同的蛋白分子，第二，相同的蛋白分子要在反應路徑的不同側，必須一個屬於反應物，另一個則屬於產物，這樣反應路徑才能合併。

如圖 4 所示，假設共有 n 個整合矩陣的步驟，每個步驟用 T(p)表示，p 是指目前進行運算的步驟數，p = 0 , 1 , 2 , … , n-1，用 T(p)Ir,Jc 表示在第 p 個步驟時第 r 列第 c 行的數值，其中 r = 0 , 1 , 2 , … , n-1，而 c = 0 , 1 , 2 , … , n-1，Z(p)Ir 表示第 p 個步驟時第 r 列所有值為 0 的元素。而合併的步驟如下：

1. 先依序搜尋 T(p)Jc，找出第一個有兩個以上非 0 值的行，並找出非 0 值是出現在哪幾列。

2. 為了避免合併時的一些陷阱，合併前必須要符合下列兩個條件才可合併：

- $z(p)Iu \cap z(p)Iw \not\subset z(p+1)Ik$ 。其意義為 2 個要合併的列，其中它們值為 0 的元素所在位置的集合，不可以完全包含於其它任一列值為 0 元素位置的集合。
- $z(p+1)Iu \cap z(p)Iw \not\subset z(p)Ik$ 。其公式的含義係同一列在做合併後，其中它們值為 0 的元素所在位置的集合，不可以完全包含於其它任一列值為 0 元素位置的集合。

3. 確認可以進行合併後，將 1.所找出的列記錄起來並做合併。

4. 重複 1.~3.的步驟，把左側矩陣中能變成 0 的值皆合併使其變為 0，直到不能合併為止。

5. 把合併出來的反應路徑進行字串比對，將字串相同的部份整合以產生出完整的反應路徑。

(三) 豬生殖與胚胎發育反應路徑探勘個案探討

在豬生殖與胚胎發育反應路徑探勘個案探討中，其關鍵字經過 MeSH 專有名詞系統的建議後選擇以”Pig embryo development”為本個案探討的關鍵字，而找

出與之相關的文獻截止於 2007 年 5 月 28 日共計 487 篇，其中的 65 篇為 review paper。再以具有統整性與權威性的 review paper 來當作本研究的反應路徑論文集。之後以人工閱讀文獻本文的方式進行找尋可能的反應基因共計 203 個。另外將這些基因的全名作 MeSH 專有名詞的校正。

本個案探討的關聯度門檻值定為 0.013，支持度設為 0.01，其中關聯度的閥值是將驗證 KEGG 裡的各分類 pathway 所算出來的 3126 基因對的 distance 1 平均值除以 2 來設定的。另外，為了尋求生物實驗突破，我們將高研究集中率的基因加以遮蓋，從基因的文獻提及數觀察出大約 11000 筆的文獻提及數為較明顯的文獻提及數之分界，於是將文獻提及數小於 11000 筆之基因做反應路徑探勘與整合的動作，並與未做遮蓋的反應圖作比較。

因為分析出的反應路徑子圖太多，故沒有全部繪出，只將反應路徑最長的 10 個子圖組合並繪出，如下所示：

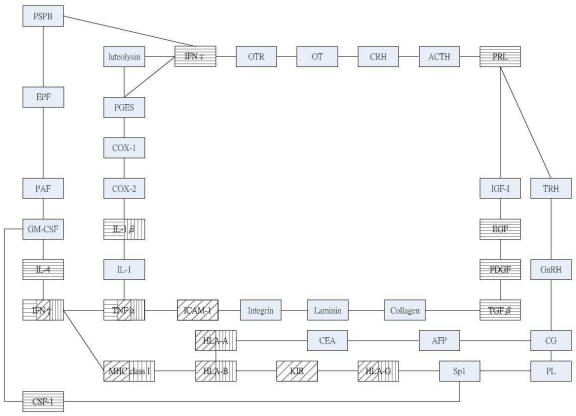

圖 5 包含高文獻提及數之基因的部份反應路徑
(關聯度設為 0.013，支持度設為 0.01)

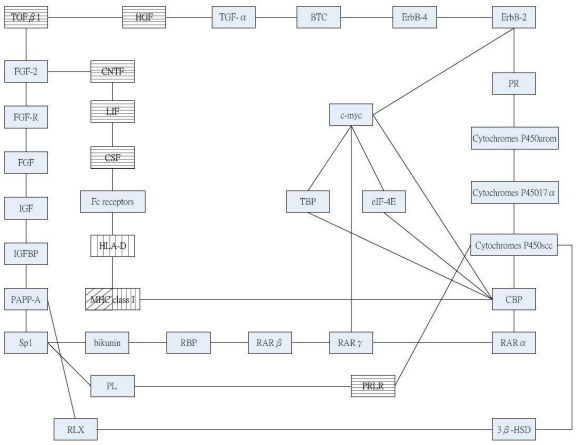

圖 6 不包含高文獻提及數之基因的部份反應路徑
(關聯度設為 0.013，支持度設為 0.01)

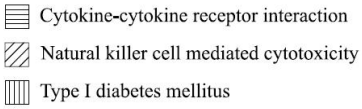

圖 7 分析出之反應路徑的線條所對應於 KEGG pathway 的名稱

圖 5 的線條標示部份，對應於 cytokine-cytokine receptor interaction 的橫線標示佔全部基因的 25%為最高比例，對應於斜線的 Natural killer cell mediated cytotoxicity(佔 20%)與對應於直線的 Type I diabetes mellitus(佔 17.5%)也佔有不小的比例。

圖 5 中佔第二大比例的線條標示為斜線，其所對應於 KEGG 的反應路徑為 natural killer cell mediated cytotoxicity，其中的 HLA-A、HLA-B、HLA-G 等為 natural killer cell mediated cytotoxicity 反應路徑的起始端。而本研究結果 TNFα、IFNγ 相鄰近的 IL-1 不屬於 KEGG 裡的 natural killer cell mediated cytotoxicity 反應路徑，我們推測 IL-1 與 natural killer cell 的反應機制有某種關聯，這部分與 Leonard 等人的假設相符合[11]。

在圖 5 中有一個無線條標示的 COX-2，它介於標示為橫線、斜線、直線的部份連結之間，可是它卻沒有出現在那些線條所對應的 KEGG 的反應路徑裡，在 KEGG 裡與之較有關聯的反應路徑則是 VEGF signaling pathway，於是本研究假設 COX-2 也許是能調節對應於橫線、斜線、直線標示的反應路徑，而在文獻裡有研究

顯示在子宮裡前列腺素(prostaglandin)和埋植、控制細胞激素的釋放、細胞生長等反應有關[12]，而圖 5 中無線條標示的 COX-2(cyclooxygenase-2)則能調控前列腺素的活化與否[13]，此與本研究的假設相符合。

綜合前面所述，本個案探討所分析出來的反應路徑圖會比較偏向於細胞內較表層的交互作用，這可能是因為目前所擷取的文獻較少討論到細胞內的調控機制。在分析出來的反應路徑圖上線條分布方面，同種線條的基因會較趨近於同區域甚至彼此間的距離會比較相近，顯示本研究結果與 KEGG pathway 資料庫結果相呼應。另外，本研究提出一種剔除高文獻提及數之基因的方式來進行反應路徑的探勘與整合，期望其結果可能供給生物研究人員作為實驗研究的新切入點。

四、計畫成果自評

本計劃為兩年期計畫第二期，目前的研究成果至已投稿論文四篇[14][15][16][17]。

五、參考文獻

[1] Sekimizu,T., Park,H.S., Tsujii,J. “Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts.” , Genome Inform. Ser. Workshop Genome Inform. 1998 , 9:62-71.

[2] Ono,T., Hishigaki,H., Tanigami,A., Takagi,T. “Automated extraction of information on protein: Protein interactions from the biological literature. “ , Bioinformatics. 2001 , 17(2): 155 – 161.

[3] Daraselia,N., Yuryev,A., Egorov,S., Novichkova,S., Nikitin,A., Mazo,I. “Extracting human protein interactions from MEDLINE using a full-sentence parser.” , Bioinformatics. 2004 , 20(5):604-11.

[4] Santos,C., Eggle,D., States,D.J. “Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction.” , Bioinformatics. 2005 , 21(8):1653-8.

[5] Joshi-Tope,G., Vastrik,I., Gopinath,G.R., Matthews,L., Schmidt,E., Gillespie,M., D'Eustachio,P., Jassal,B., Lewis,S., Wu,G., Birney,E., Stein,L. “The Genome Knowledgebase: a resource for biologists and bioinformaticists.” , Cold Spring Harb. Symp. Quant. Biol. 2003 , 68:237-43.

[6] Weeber,M., Vos,R., Klein,H., De Jong-Van Den Berg,L.T., Aronson,A.R., Molema,G. “Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide.” , J. Am. Med. Inform. Assoc. 2003 , 10(3):252-9.

[7] Wren,J.D., Garner,H.R. “Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network.” , Bioinformatics. 2004 , 20(2):191-8.

[8] Stapley,B.J., Benoit,G. “information retrieval and visualization from co-occurrences of gene names in Medline abstracts.” , Pac. Symp. Biocomput. 2000:529-40.

[9] Famili,I., Palsson,B.O. “Systemic metabolic reactions are obtained by singular value decomposition of genome-scale stoichiometric matrices.” , J. Theor. Biol. 2003 , 224(1):87-96.

[10] Przyłęcka,J., Gregoraszczuk,E.L., Kotwica,G., Stefańczyk-Krzymowska,S., Ziecik,A.J., Blitek,A., Ptak,A., Rak,A., Wójtowicz,A., Kamiński,T., Siawrys,G., Smolińska,N., Franczak,A., Kurowicka,B., Oponowicz,A., Wasowska,B., Chłopek,J., Kowalczyk,A.E., Kaczmarek,M.M., Wacławik,A. “Mechanisms ensuring optimal conditions of implantation and embryo development in the pig.” , Reprod. Biol. 2006;6 Suppl. 1:59-87.

[11] Leonard,S., Murrant,C., Tayade,C., van den Heuvel,M., Watering,R., Croy,B.A. “Mechanisms regulating immune cell contributions to spiral artery modification -- facts and hypotheses -- a review.” , Placenta. 2006 Apr;27 Suppl. A:S40-6. Epub 2006 Jan 4.

[12] Kelly,R.W., King,A.E. Critchley,H.O. “Cytokine control in human endometrium.” , Reproduction. 2001 , 121 3–19.

[13] Bracken,K.E., Elger,W., Jantke,I., Nanninga,A., Gellersen,B. “Cloning of guinea pig cyclooxygenase-2 and 15-hydroxyprostaglandin dehydrogenase complementary deoxyribonucleic acids: steroid-modulated gene expression correlates to prostaglandin F2 alpha secretion in cultured endometrial cells.” , Endocrinology. 1997 , 138 237–247.

[14] Yu-Shiang Huang, Chih-Chin Liu, Whei-Meih Chang, A Method for Predicting the Active Sites of Dual-specific Protein Phosphotases Based on Protein Surface Patches, submitted for publication.

[15] Wei-Hung Yang, Chih-Chiu Liu and Whei-Meih Chang, Detection of core promoter elements for dual specificity phosphatase genes, submitted

for publication.
[16] 林怡伸、劉志俊*、劉世華、林恩仲, 通用於 cDNA 微陣列晶片之快速探針挑選方法, 論文投稿中.
[17] 黃柏諺、劉志俊、張慧玫, 整合 PubMed 文獻資料庫中基因相關性成為基因調控路徑的方法之探討, 論文投稿中.