由 PubMed 文獻資料庫中自動搜尋與整合生化反應路徑的方法

Automatic Pathway Discovery and Integration from PubMed Literature Database

曾坤源、劉志俊*
中華大學資訊工程學系
ccliu@chu.edu.tw

林恩仲
台大動物科學技術學系
eclin@mail2000.com.tw

摘要

在功能性基因體中，生化反應路徑的資料是其中非常要一環。目前雖有一些生化反應路徑資料庫例如 KEGG、BioCyc 等，但其資料離完整仍有非常遙遠的距離，且以人工閱讀文獻的方式來描繪生化反應路徑，其速度難以跟上 DNA 定序資料產生的速度，所以如何以生物資訊的方法，由現有數以千萬計的相關參考文獻資料庫中，快速篩選出各物種可能的生化反應路徑草圖，成為目前生物資訊領域的重要研究主題。因此，本文提出一種由 PubMed 文獻中自動尋找生化反應路徑的方法，自動找出生化反應路徑，做為進行研究的基本資料。

1 概論

隨著人類基因體定序計劃的完成，生物科技進入以功能為主的後基因體時代。研究重點由各物種基因體的定序轉為這些定序資料的解讀，我們想要知道這些基因除了 DNA 序列資訊之外，究竟在生物體內扮演那些功能，以及我們又能如何利用這些資訊進行基因表現調整。

伴隨者科技不斷進步，人類在生物科技領域的發展也有長足的進步，使得人類對於未來的生活方式充滿了美好的憧景。也因為生物科技不斷的進步，生物學家也就累積了大量的相關文獻資料，例如：人類基因資料、蛋白質資料、生化反應路徑、各式遺傳疾病基因、酵素反應、藥物分子結構等等。而這些科技未來將實現人類夢寐以求的生活，例如：各種重大疾病可以得到有效的治療、直接培植人體重要器官、改變人類下一代的各項特質等等。因為生物科技的範圍實在非常浩瀚，各種分子生物知識仍有太多的未知等待發掘，例如：遇到新的蛋白質結構時，我們不可能毫無目標的盲目進行實驗，一一測試新蛋白質結構的特性，這樣成本太高，也太耗時了。此時若能從前人所寫的相關文獻中，找出可能的研究方向，避免錯誤的發生，提升研究速度。

在了解到生物科技文獻的重要性之後，如何能快速、精確的參考到相關的生物科技文獻則顯得相當重要。但我們知道全世界各國對生物科技均抱者高度的期待，也都投入了大量的人力、金錢去做研究，使得產生了大量的研究報告。

這些資料量非常的龐大，並且目前正以驚人的速度繼續成長，若仍然用人工的方式來閱讀，恐怕終其一生之力也無法將文件全部讀完，更徨論要在其中進行地毯式的搜索，完整地找出有價值的資訊，指導新研究的進行。由此可知若能讓電腦自動分析文獻中的資訊，找出具有高度相關的文獻提供給研究人員，是一件非常有意義的研究。

我們知道蛋白質是組成所有生物的基本元素，而蛋白質之間的交互作用則決定了生物內部各種複雜的反應。然而要探討蛋白質之間的交互作用，就必需要先了解每個蛋白質的相關文獻，在從中找出各蛋白質之間的關連，進而分析出完整的生化反應路徑。但這也意味者會涉及到多種蛋白質以及龐大的文獻資料，分析起來將更加不易。但若想要徹底了解生化反應路徑，這又是非做不可的工作。因此如何能有效地由現有文獻中，找出蛋白質之間的交互作用以及可能的生化反應路徑成為相當重要的研究主題。

目前文獻資料自動探勘技術[4]的研究主要可分為二類：1.語言學方式分析[5][6] 2.關鍵字搜尋分析[7][8][9][10]。第一類是將文獻用傳統語言學技術進行分析，分析文獻的文法、語意、句型，藉語言學的語意分析將文獻的內容轉換成電腦可以處理的規則。因為是使用語言學的方式分析，理論上分析的結果應較為準確，但是由於語言的規則太多，且又很多例外情形，再加上寫文章的人母語不一定是英語，可能會犯文法上的錯誤，或是錯用慣用字等等。程式很難包含所有的情況，反而較難自動找出文獻的正確內容。第二類是用蛋白質名稱及關鍵字(例如：催化)搜尋文獻，找出相關的句子，判定是否有關連性。若有則將它轉換成電腦可以處理的規則。這類方法雖然沒有嚴謹的語法分析，但是往往會有較佳的分析結果。

在關鍵字搜尋分析的相關研究方面，Friedman 等人，在 2001 年提出一種叫 GENIES 的自然語言處理(Natural Language Processing)系統[8]，可以從文獻資料中找出生化反應路徑。該文提出一種叫"type-value pair"的資料儲存格式來儲存文獻中蛋白質與基因之間的關連性資料(relationship)，將一般的文字敘述轉成易於

*通訊作者

電腦處理的資料格式。但是該篇文章僅止於提出儲存蛋白質及基因之間關連的格式，然而真正的困難點是在於將蛋白質及基因之間的關連接起來，建構出完整的生化反應路徑，這篇文章卻沒有提出具體的做法，更遑論要自動維護目前已存在的生化反應路徑。

在生化反應路徑找尋方面，Nathan 等人提出了一種由基因組找尋生化反應路徑的方法[18]。此種方法主要是用基因組方面的資料來確認已找出的生化反應路徑是否正確以及其中是否有多餘的部份，但是並不能主動的找出生化反應路徑。

Ng 及 Wong[7]提出了一種半自動尋找生化反應路徑的方法。其方法首先在網際網路上用使用者輸入的關鍵字找出相關的文獻，用現有的蛋白質名稱配合特定的樣版，找出可能的關連，最後由使用者以手動方式找出生化反應路徑。這種方式將決定權交給使用者，使用者還是要花費時間閱讀文獻原文，對自動找尋生化反應路徑而言，幫助還是有十分有限。

Stefan 等人提出了一種稱為基本通量模式(elementary flux modes)的方式[11][12][13][14][15][16][17]，可以將輸入的片斷化學反應方程式，經由化簡規則合併後，得到整合的生化反應路徑。這個方法成功的提供一個快速合併產生生化反應路徑的方法，並且分析的結果也得到生物方面專家的認同。同時也解決了目前生化反應路徑分析出來的結果都是星狀的問題。該方法可以有效找出線狀的生化反應路徑，同時也可以有效的分析目前已經存在的生化反應路徑，加入新發現的關連特性，重新加以合併化簡，並產生出更加完整的生化反應路徑。這對於目前已經現存的生化反應路徑的維護而言是非常重要的一點。但是要使該方法必需要先找出所有相關的化學反應方程式，而在目前現有的生物醫學文獻中，大部份都只是文字的描述居多，並且還有相同的物質卻有多種名稱的問題，不能輕易的從中找出完整的化學反應方程式。

本篇論文結構如下：第 2 節我們將針對系統整體架構圖作介紹，第 3 節是說明如何由文獻中找出長度為 1 的生化反應路徑，第 4 節為生化反應路徑之整合，第 5 節為系統實作及實驗，第 6 節為結論與未來工作。

2 自動找尋生化反應路徑

在本章節中，首先我們將說明整個系統的架構，接著說明如何在 PubMed 中找出與某一使用者指定生化反應路徑名稱相關的文獻，及其中所有可能相關之生化反應的方法。

2.1 系統架構

我們所提出之生化反應路徑自動探勘系統架構圖如圖 1 所示，包含了 MeSH 資料來源(MeSH) 、PubMed 文獻資料來源、候選生化反應尋找(Candidate Path Finding)、生化反應路徑搜尋(Pathway Discovery)、生化反應路徑整合(Pathway Integration)等六個部分。前四個部分涵義說明如下文。而最後兩個部分分別於第 3 節與第 4 節中說明。

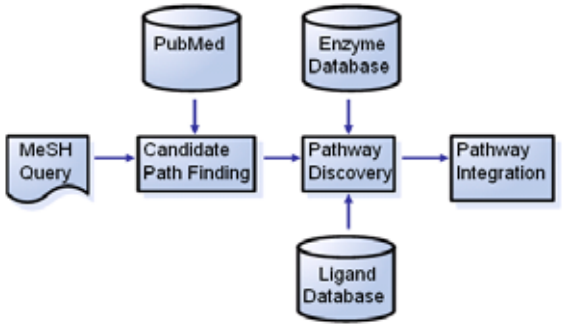

圖 1 生化反應路徑自動探勘系統架構圖。

2.2 PubMed 文獻資料庫

PubMed 是目前生物醫學方面資料最完整的資料庫，到目前為止仍以極快的速度增加中，至 2004 年 11 月為止，PubMed 的文獻總數已達 1 千 2 百多萬筆，其中包含了醫藥、醫療護理、臨床醫學經驗、各式醫療文獻、流行傳染疾病、公共衛生政策等等的各式的文件。

而要自動產生生化反應路徑的首要條件就是要有一個資料完整，數量龐大的資料庫做為資料來源，而 PubMed 剛好就很符合這個條件。但是在 PubMed 中，因為資料包含的時間範圍很長久，資料的來源範圍也非常的廣，而且來自世界各地，很容易發生相同的一個概念，不同的人卻使用不同的專有名詞，不同的說法來描述同一件事的不一致狀況。

目前 PubMed 提供了網路查詢介面，但這些查詢介面都是以某一個關鍵字做查詢，這樣的查詢方式得到的結果十分的局部，沒有辦法找出與關鍵字高度相關卻又沒有提到該關鍵字的文獻。若是直接用 PubMed 的查詢結果進行生化反應分析，會遺失許多重要的關連，同時也會包含了錯誤的資料在內，不適合用來直接進行分析。必需要藉由 MeSH 專有名詞定義的幫助來找出與特定生化反應路徑關鍵字有高度相關的文獻。

2.3 MeSH 專有名詞定義

MeSH 的全名是 Medical Subject Headings (MeSH®)，它是由美國國家醫學圖書館(National Library of Medicine's)所定義與維護，用以統一文獻內容註解與分類的專有名詞系統。目前包含了 11 個階層，22,568 個專有名詞的描述，139,000 個概念文獻補充，及 1,000 個專有名詞之間的相關參考連結。

MeSH 主要包含了生物醫學相關字彙使用及名稱描述，最主要的目地是要提供一個能有效、快速、有階層架構的專有名詞查詢方式。因為在生物醫學領域中有些關鍵字包含了太大的範圍，若不用階層分類方式查詢，很容易會發散開來，找到完全不是使用者想要的資料。相反的，有些關鍵字包含的範圍又太小，很容

易會遺失掉重要的訊息。MeSH 開放給全世界的生物醫學專家提出建議，由美國國家醫學圖書館做最後決議。藉由集合全世界專家的建議，來找出最合適的關鍵字組，在由 MeSH 將關鍵字組與 PubMed 相關文獻做連結，提供一個正確、完整的關鍵字組查詢系統。

藉由 MeSH 專有名詞系統的幫助，可以給使用者所要找尋的關鍵字一個良好的建議，並且找到與這些關鍵字組高度相關的 PubMed 文獻(以下稱為反應路徑論文集)。

2.4 尋找所有可能的生化反應路徑

候選生化反應尋找是將使用者輸入的特定生化反應路徑關鍵字輸入到 MeSH 專有名詞系統，列出它建議使用的關鍵字組，讓使用者選擇正確的關鍵字，再利用 MeSH 專有名詞系統與 PubMed 之間的文獻連結，找出反應路徑論文集，最後在將文獻的摘要取回存成檔案。

3 生化反應路徑探索

一個生化反應路徑是由一組相關的生化反應連接而成，所以要找出特定的生化反應路徑，必需先找到其中所有化學反應，也就是長度為 1 的生化反應路徑集合。

3.1 找尋所有長度為 1 之生化反應路徑

Ligand 資料庫收集了目前已知的小分子資料庫。將 Ligand 資料庫中所有的小分子任取 2 個做排列組合，找出所有長度為 1 之生化反應路徑所有可能出現的組合。把這些組合當成關鍵字，送到反應路徑論文集中找出在同一篇文獻中同時出現這 2 個小分子的文獻，如果同一篇文獻中同時出現了這 2 個關鍵字，即認定為這 2 個小分子是有相關的，如果文獻的篇數愈高，就表示這 2 個小分子的相關度愈高。

3.2 長度為 1 之生化反應路徑篩檢

因為是用所有的排列組合去找尋長度為 1 之生化反應路徑，所以很有可能在同一篇文獻中只是湊巧同時提到那 2 個小分子而已，並不是它們之間真的有相關，故還需要做篩檢的動作，將相關度較低的組合去除。

任取 2 個小分子稱為 M 及 N，先找出每個小分子在反應路徑論文集中出現的次數，稱之為 α 及 β，(α∩β)表示 M、N 同時出現在同一篇文獻中的數量，δ 表示 M、N 之間的關連度，Max 表示單一小分子在所有反應路徑論文集中出現的最大次數，δ 計算方式如下：

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

藉由設定 δ 的門檻值，可以將一些誤判為有相關的組合去除，以提升系統的準確度。

4 生化反應路徑整合

在本章節中，我們說明如何將前一章節中所得到長度為 1 之生化反應路徑，組成完整的生化反應路徑的方法。

4.1 化學方程式平衡計算

首先要將找出的長度為 1 之生化反應路徑集合轉成矩陣的形式，以便進行合併化簡運算。我們先列出所有要分析長度為 1 之生化反應路徑集合，再找出所有參與反應的小分子，將小分子放在行的位置，並令它的個數為 m。接下來列出所有參與反應的酵素，將它放在列的位置，令酵素的個數為 n，可以得到一個大小為 m * n 的矩陣，稱為 Ms。接下來要在矩陣中填入值，將生化反應路徑的平衡系數放入陣列中，若是反應物值的量會隨反應時間的增加而增加，則以正數表示。若是反應物值的量會隨反應時間的增加而減少，則以負數表示。若是二者無關則以 0 表示。

接下來舉一個範例來說明如何由長度為 1 之生化反應路徑集合轉換成*化學方程式平衡計算矩陣(Stoichiometry Matrix)*。有二個生化反應路徑：glux + ATP → G6P + ADP 及 G6P → G1P，而參與反應的酵素分別為 HK 及 PGM，因為參與反應的小分子共有 5 種，而參與的酵素有 2 種，故可以得到一個 5 * 2 的矩陣，在將長度為 1 之生化反應路徑的平衡參數加入，就可以得到化學方程式平衡計算矩陣，如圖 3 所示：

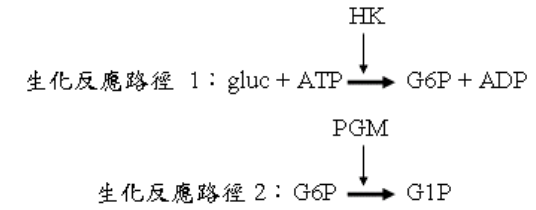

圖 2 原始輸入生化反應路徑。

$$S = \begin{array}{c} \begin{matrix} gluc & G6P & G1P & ATP & ADP \end{matrix} \\ \begin{bmatrix} 0 & -1 & 1 & 0 & 0 \\ -1 & 1 & 0 & -1 & 1 \end{bmatrix} \end{array} \begin{matrix} \text{PGM} \\ \text{HK} \end{matrix}$$

圖 3 化學方程式平衡計算矩陣

4.2 組成初始反應矩陣

我們用化學方程式平衡計算矩陣將生化反應路徑轉成矩陣的目的，是要用矩陣的列運算將相關的生化反應路合併。在圖 2 的範例中，2 個生化反應路徑合併的方法是將生化反應路徑 1 右手邊的 G6P 與生化反應路徑 2 左手邊的 G6P 做代入，並將生化反應路徑 2 右手邊的 G1P 代入，就可以得到合併後的生化反應路徑。在這個範例中可以發現若 2 個生化反應路徑要合併必需要有 2 個條件

第一：必需要有相同的小分子同時參與這 2 個生化反應路徑。化學方程式平衡計算矩陣在行的部份就是小分子，所以 2 個生化反應路徑

在同一行就表示有相同的小分子參與反應。

第二：這個小分子必需要在 2 個生化反應路徑的不同側，若小分子在同一側是不可能合併的。化學方程式平衡計算矩陣用負號表示在生化反應路徑左邊的小分子，用正號表示在生化反應路徑右邊的小分子，所以如果 2 個生化反應路徑在同一行中的值是一正一負，表示這 2 個小分子是在反應的不同側，可以相加合併。

接下來要將化學方程式平衡計算矩陣轉成初始化矩陣 T(0)。將化學方程式平衡計算矩陣放在 T(0)的左手邊，再將現有矩陣的右側產生一個單位矩陣(identity matrix)令為 Mi，它的大小為 M，將 Ma 放在 T(0)矩陣的左手邊，將 Mi 放在 T(0)的右手邊，中間用線隔開，形成初始矩陣 T(0)。

圖 4 是圖 3 轉成初始化矩陣的結果。

$$\left[\begin{array}{ccccc|cc} 0 & -1 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 & -1 & 0 & 1 \end{array}\right]$$

圖 4 初始化矩陣

在初始化矩陣的右手邊是表示最後反應方程式是由多少初始反應方程式組成的，它主要是做合併後的反應方程式之化學平衡用。

4.3 生化反應路徑的整合

建立初始化矩陣後的下一個階段便是要進行矩陣的合併。 假設共有 n 個矩陣化簡步驟，每個步驟用 T(j)表示，j=0,1,2…n-1，j 表示目前進行運算的步驟數，用 T(j)i,h 表示在第 j 個步驟第 i 列第 h 行個元素，T(j)k 表示第 j 個步驟中的第 k 列。Z(j)m 表示第 j 個步驟第 m 列所有值為 0 的元素所在位置，例如：設第 T (3)第 5 列的值為(0 0 1 0 0 1 0 0)，則 Z(3)5=(1,2,4,5,7,8,9)。合併步驟如下:

1.將固定會發生反應的反應方程組，組合成單一的化學反應方程式組(非必要)，如此一來可以減少矩陣的大小，同時也可以避免不必要的錯誤合併。

2.如果第(j+1)行的值為 0，則整列的數字可以直接由 T(j)的矩陣直接複製到 T(j+1)的矩陣。

3.其它在列的值則必需要進行列運算將它第(j+1)行的值化為 0，例如給 2 個列 T(j)1 及 T(j)2，它們第(j+1)行的值分別為-1 及 2 則必需要進行下列的列運算，做化簡：2 * T(j)1 + T(j)2。

4.為了避免合併過程中的一些陷阱，合併前必需要符合下列二個條件才能進行合併：

- $z(j)i \cap z(j)k \notin z(j+1)n$。它的含義就是 2 個要合併的列，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。例如：T(j)2={0,0,2,-1,0,0,0,0,0} ， Z (j)2={1,2,5,6,7,8,9} ， T(j)6={0,0,0,1,0,2,0,0,0} ， Z (j)6={1,2,3,5,7,8,9} ， Z (j+1)3={0,0,1,0,0,1,0,0,0} ， Z (j)2∩Z (j)6={1,2,5,7,8,9} ，Z (j)2∩Z (j)6 被 Z (j+1)3 所包含，故這 2 列就不可以合併。
- $z(j+1)i \cap z(j)k \notin z(j)n$。公式的含義就是同一列在做合併後，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。

5.確定可以合併以後，將在第幾個步驟那 2 個長度為 1 之生化反應路徑合併的資訊記錄下來。

6.重覆步驟 2~5，直到在左手邊的矩陣全部變成 0 為止，在實際的案例中，幾乎都能將左手邊的矩陣都化成 0。

4.4 生化反應路徑取得

到目前為止已經成功的將左手邊的矩陣合併完成。左手邊完成合併後，右手邊的矩陣就是這個生化反應路徑的組成要素，說明這個生化反應路徑是由那些長度為 1 之生化反應路徑整合完成。接下來將在合併過程中所記錄的合併過程與右手邊矩陣的資料依記錄的過程畫出完整的生化反應路徑。

5 系統實作

在本節中我們舉一個實際的範例說明整個自動找尋生化反應路徑的系統實際運作方式。

5.1 尋找所有可能的生化反應路徑

我們實做了一個網路版的候選生化反應尋找程式。發展環境選用 J2EE(Java 2Eenterprise Edit)架構，網頁伺服用的是 BEA Weblogic 8.1，作業系統是 Windows 2000 Professional，資料庫是用 My SQL 4.0。這個程式會自動將使用者輸入的 MeSH 關鍵字送到 PubMed 網站進行查詢，並且自動將所有 MeSH 連結到的反應路徑論文集文獻自動取回，再將資料自動新增到資料庫中，同時也會將文獻用純文字方式存成文字檔。圖 5 為系統自動查詢的畫面，圖 6 為查詢結果單筆記錄範例。

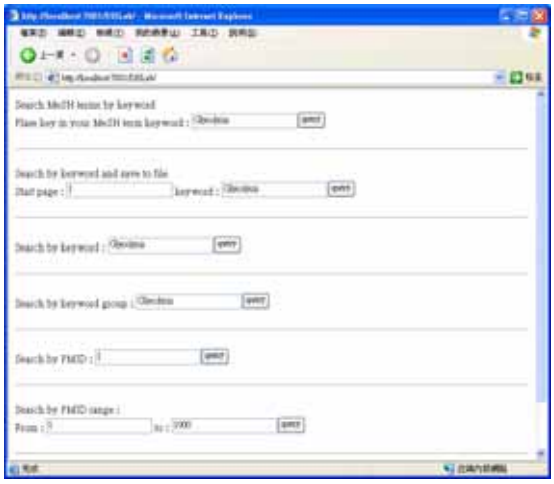

圖 5 MeSH 專有名詞定義查詢系統。

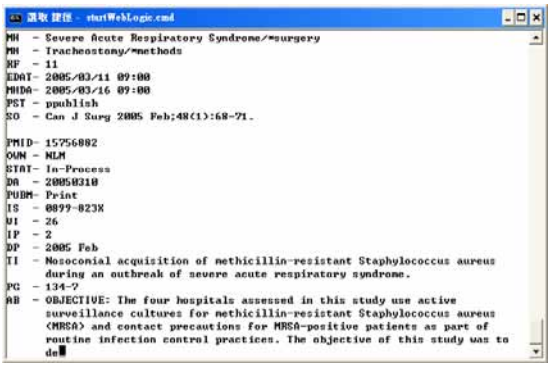

圖 6 候選生化反應尋找程式執行結果

5.2 生化反應路徑探索

我們將 ENZYME 酵素資料庫以反向工程技術重新建構於本地端 MySQL 資料庫中，以方便資料分析。它記錄了所有的酵素相關資料，同時它也記錄了每個酵素所具有高度專一性的化學反應方程式。我們同時將與某個 MeSH 專有名詞相關的論文集合存在資料庫中。圖 7 是 MySQL 中的酵素資料庫資料及關連度計算結果。

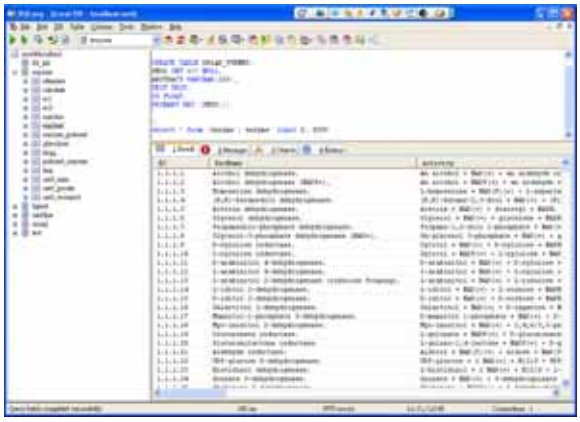

圖 7 酵素資料庫資料及關連度計算結果

5.3 生化反應路徑整合

這部份的實作因為不需使用到網路的資源，且為了提供較佳的使用者操作介面，故我們採用 J2SE 開發。程式需要輸入的參數有：可逆反應方程式、不可逆反應方程式、與外部有接觸的反應物質、與外部沒有接觸的反應物質，輸入完成後選執行頁面，在按"開始執行按鈕"後，程式就會開始執行，並將結果顯示於按鈕下方的文字方塊中，如圖 8 所示：

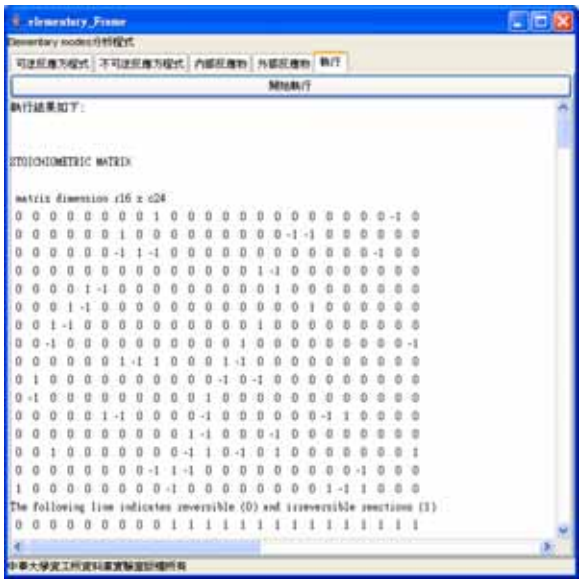

圖 8 生化反應路徑分析執行結果

5.4 實驗結果

為了方便表示，先將常用到的小分子名稱用簡寫表示，如表一所示：

表 1 反應物或產物全名與代號

小分子反應物/產物全名	代號
alpha-D-Glucose 6-phosphate	G6P
beta-D-Fructose 6-phosphate	F6P
beta-D-Fructose-1,6P2	FP2
Glyce ralde hyde-3P	GAP
Glyce rate-1,3P2	1.3BPG
Glyce rage-3P	3PG
Glyce rage-2P	2PG
Phasphoenal-pyruvate	PEP
Pyruvate	Pyr
D-Ribose-5P	R5P
D-Xylulose-5P	Xyl5P
6-Phospho-D-gluconate	6PG
Citrate	CIT
Oxaloacetate	OAA
Acetyl-CoA	AcCoA
4-amino-4-deoxychorismate	4ADeo

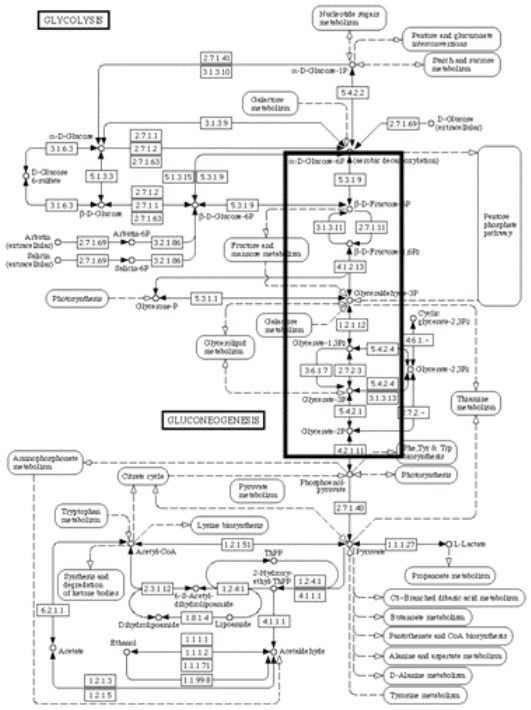

圖 9 KEGG 的 Glycolysis 生化反應路徑

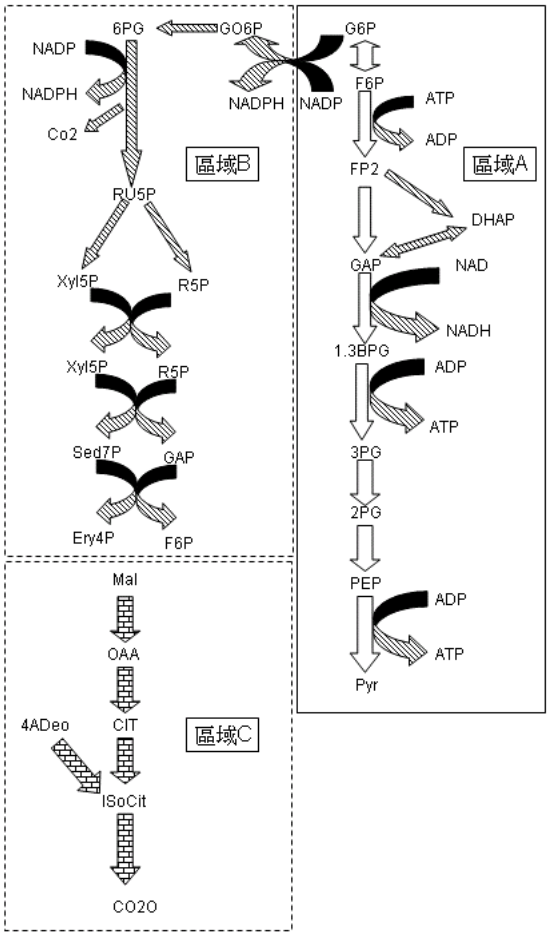

圖 10 程式分析執行結果

圖 10 是程式分析出的生化反應路徑，因為分析出的生化反應路徑片斷太多，故沒有全部繪出，只繪出部份。

我們針對程式分析找出的結果與 KEGG 的生化反應路徑兩者之間的差異部份進行分析：

圖 10 區域 A 的部份為 KEGG 有存在且我們分析程式也有找到的部份。這部份的生化反應路徑已經非常成熟，反應路徑論文集中相關討文獻也很多，可以分析出相同的結果。

圖 9 方框區域以外的部份為 KEGG 有而程式沒找到的部份。這部份發生的原因是因為程式在找出所有長度為 1 的生化反應路徑後，會設定門檻值，將一些相關度較低的去除，當門檻值的設定太高時就會把正確的生化反應路徑過濾掉。例如圖 9 中有一個 β-D-Glucose 與 α-D-Glucose 的路徑，它們各自出現在反應路徑論文集文獻的篇數是 27 與 47，共同出現在同一篇文獻的篇數為 1，很少文獻會同時出現這 2 個小分子，程式判定之間的相關度不高，這個路徑很容易就會被過濾掉了。

KEGG 沒有找到而程式有找到的部份：如圖 10 區域 B 及區域 C 的部份。區域 B 部份並不是錯誤的生化反應路徑，只是 KEGG 將這部份的生化反應路徑分類為 Pentose Phosphate，而它與 Glycolysis 相鄰。由於我們程式分類的方式與 KEGG 不同，故會將 KEGG 二相鄰的生化反應路徑合併放在一起。區域 C 的部份可能性就有比較多種可能，它可能是其它生化反應路徑中某一小段，剛好在反應路徑論文集文獻中也剛好有提到，故被分析出來。例如在區域 C 中有一個 Malate 及 Oxaloacetate 的路徑，它們各自出現在反應路徑論文集文獻的篇數是 726 次與 6948，共同出現在同一篇文獻的篇數為 57 篇，故程式將這 2 個小分子視為相關，雖然在 Glycolysis 中它是屬於雜訊，但在 Citrate Cycle(TCA cycle)中它卻是屬於正確的路徑。當然，區域 C 也很有可能是因為雜訊影響了程式的分析，產生的錯誤結果。

6 結論

生化反應路徑是生物領域研究人員很重要的參考資訊，它可以告訴研究人員在代謝反應進行時，它的過程是如何進展的，特定物質又會對這項行為造成什麼樣的影響，讓研究人員有足夠的資訊進行各式生化實驗。然而生化反應路徑的尋找卻是極度困難的，目前較佳的生化反應路徑資料庫如 KEGG，還是以人工的方式進行尋找，本篇論文希望提供一個自動尋找可能之生化反應路徑的機制，以簡化生化反應路徑的尋找工作。

未來的發展可以朝下列方向進行：

- 生化反應路徑自動繪圖部份：本論文目前分析出來的資料是用矩陣方式存在，必需要用手動方式將矩陣畫成生化反應路徑，當生化反應路徑較大時這個工作會耗時且麻煩，若這個工作可以交由程式自動轉換，可以節省

不必要時間浪費。

- 訊號傳導(Signal Transduction)路徑：本文所提之生化反應路徑是針對代謝反應路徑設計，我們將調整使其能尋找訊號傳導路徑。
- 此外，我們正在使用此項技術應用在新完成定序物種的基因體功能分析與註解，以加速對註解資訊不全的物種之功能解讀。

7 誌謝

本論文為農委會與國科會農業生物技術國家型科技計畫補助成果，計畫編號 92農科-3.1.1-牧-U1(3) 以及 93 農科-4.2.3-牧-U1，NSC-94-2213-E-216-033。

參考文獻：

[1] Ernst Kretschmann, Wolfgang Fleischmann and Rolf Apweiler, "Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT," *Bioinformatics, 17, 920- 926, 2001.*

[2] Wolfgang Fleischmann,Steffen Moller,Alain Gateau and Rolf Apweiler, "A novel method for automatic functional annotation of proteins," *Bioinformatics, 15, 228-23, 1999.*

[3] Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler1, Marie-Claude Blatter,Anne Estreicher, Elisabeth Gasteiger, Maria J. Martin1, Karine Micho, Claire O'Donovan1, Isabelle Phan, Sandrine Pilbout and Michel Schneider, "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003," *Bioinformatics, 17, 920-926, 2001.*

[4] Lynette Hirschman, Jong C. Park , Junichi Tsujii, Limsoon Wong, Cathy H. Wu, "Accomplishments and challenges in literature data mining for biology," *Bioinformatics, 18, 1553-156, 2002.*

[5] Joshua M. Temkin and Mark R. Gilder , " Extraction of protein interaction information from unstructured text using a context-free grammar," *Bioinformatics, 19, 2046-205 , 2003.*

[6] Nikolai Daraselia., Anton Yuryev, Sergei Egorov,Svetalana Novichkova, Alexander Nikitin and Ilya Mazo , "Extracting human protein interactions from MEDLINE using a full-sentence parser," *Bioinformatics, 20, 604-611, 2004.*

[7] See-Kiong Ng, Marie Wong, "Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts", *Pacific Symposium on Biocomputing (PSB), 1999.*

[8] Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, Andrey Rzhetsky, "a natural-language processing system for the extraction of molecular pathways from journal articles," *Bioinformatics, 17, S74-S82, 2001.*

[9] E. Segal ., H. Wang, D. Koller, "Discovering molecular pathways from protein interaction and gene expression data," *Bioinformatics, 19, i264-i272, 2003.*

[10] B.J. Stapley, G. Benoit , "Biobibliometrics: Information Retrieval And Visualization From Co-Occurrences of Gene Names In Medline Abstracts", *Pacific Symposium on Biocomputing(PSB)*, 2000.

[11] Stefan Schuster, David A. Fell, and Thomas Dandekar , "A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks", *Nature Biotechnology* Vol 18, March, 2000.

[12] Stefan Schuster, David A. Fell, and Thomas Dandekar , "METATOOL: for studying metabolic networks", *Bioinformatics Vol 15 No.3 1999 Pages 251-257.*

[13] Schuster, S., Dandekar, T. & Fell, D., "Detection of elementary flux modes in biochemicalnetworks: a promising tool for pathway analysis and metabolic engineering", *TIBTECH, 17, 53-60, 1999.*

[14] S.Schuster, C.Hilgetag, J.H.Woods, D.A.Fell, "Reaction routes in biochemical reaction systems:Algebraic properties,validated calculation procedure and example from nucleotide metabolism.", *Journal of Mathematical Biology Issue: Volume 45, Number 2, 153 – 181, August 2002.*

[15] Schuster, S. and Hilgetag, C. Stapley, G. Benoit , "On elementary flux modes in biochemical reactionsystems at steady state.", *J. Biol. Syst. 2, 165-182, 1994.*

[16] Schuster, S., Hilgetag, C., Woods, J. H. and Fell, D. A. , "Elementary modes of functioningin biochemical networks. In: Computation in Cellular and Molecular Biological Systems", *World Scientific,* Singapore P 151-165, 1996.

[17] Hartwell , "A robust view of biochemical pathways", *Nature 387,* 855-857, 1997.

[18] Nathan D. Price, Jason A. Papin, and Bernhard Palsson, "Determination of Redundancy and Systems Properties of the Metabolic Network of Helicobacter pylori Using Genome-Scale Extreme Pathway Analysis", *Genome Research*, 12:760-769, 2002.