生化反應路徑資料庫的內涵式查詢
Content-Based Retrieval in Biopathway Databases

王俊淵　劉志俊　曾文慶

中華大學資訊工程學系

ccliu@chu.edu.tw

摘要

後基因體時代，生物資訊研究的重心由基因體與蛋白體逐漸轉移到生化反應路徑。由於生化反應路徑的資料與日俱增，而且資料型態比基因與蛋白質來的複雜，因此生化反應路徑資料資料庫的查詢是一項充滿挑戰性的研究領域。目前主要的生化反應路徑資料庫所提供的查詢功能，皆是以參與反應的化合物或催化反應的酵素名稱或編號為查詢條件，欠缺反應路徑間相似性比較的能力。在本篇論文中，我們將多媒體資料的內涵式查詢的觀念擴充應用於生化反應路徑資料庫中，提出一種以矩陣為基礎的比對兩個生化反應路徑內容相似度的衡量方法，來讓使用者依樣本查詢(query by pathway example)生化反應路徑資料庫。可以在數量龐大的反應路徑中，有系統地找到功能相近的反應路徑，提供生物技術產業與製藥業研發的一項利器。

關鍵詞：生化反應路徑(biopathway)、生化反應路徑資料庫(pathway databases)、內涵式查詢(content-based retrieval) 、依樣本反應路徑查詢 (query by pathway example)

一、簡介與相關研究

生物遺傳學上的中心法則中指出，由去氧核醣核酸(DNA)轉錄為核醣核酸(RNA)，再由核醣核酸(RNA)轉譯為蛋白質(protein)，然後透過這些蛋白質之間的交互作用(protein-protein interaction)，形成生化反應路徑(biochemical pathway 或 biopathway) ，建構出生物體內各式各樣的生命現象。後基因體時代，生物科學家除了繼續對基因序列做更深入的解碼與蛋白質結構與功能研究之外，更進一步需要去了解各個蛋白質在生化反應路徑上扮演的角色，以及對反應路徑造成的影響。例如了解某種病毒在人體內致病的反應路徑後，找到關鍵性的反應，就能夠開發有效的抑制藥物，去治療此種病毒所造成的疾病。

在生化反應路徑的研究領域，除了利用實驗室的技術去探索與發現新的反應路徑之外，利用電腦來輔助進行生化反應路徑的研究，可以有效地加速研究的進展。但現有的反應路徑資料庫所提供的查詢功能，例如 KEGG [8][9]、WIT [15][18]與 BioCyc [10][12] ，皆是以參與反應的化合物或催化反應的酵素名稱或編號為查詢條件，而以靜態圖形或者網頁呈現查詢的結果，欠缺比較反應路徑間相似性的能力，未能充分發揮生化反應路徑資料庫的功用。

由於反應路徑的資料是一種網路結構，與基因序列一維字串的資料型態不同且較複雜，因此反應路徑的比對較基因序列的比對要困難許多。即使在同一物種的生化反應路徑中，每一個反應路徑都有其專司的功能，路徑與路徑之間，具有相互牽連的關係，路徑中出現的化合物或蛋白質，不單純出現在單一的路徑之中。不同物種的生化反應路徑的比對，雖然比對的複雜度更高，但能夠提供對未知或不全的反應路徑有價值的參考資料。

Küffner 等人提出一種反應路徑比對的方法[13] 。此方法由已知的反應路徑中，根據輸入的查詢條件，如反應物、產物、路徑長度、直徑與範圍等條件，找出符合條件的反應路徑。 Goesmann 等人[6]利用基因的註解及酵素的分類編號，來辨識生化反應路徑，並進一步去比較不同物種間的反應路徑，以了解物種之間反應路徑的差異性。Ogata 等人[14]利用一個圖形比較的演算法(heuristic graph comparison algorithm)，找出基因組(Genome)與生化反應路徑(Pathway)相對應的關係。再依照功能的分類，建立不同物種功能關係叢集表(FRECs)，由叢集表中可以找到反應路徑的差異。Dandekar 等人[4]利用多重序列比對的概念來比較多種微生物的反應路徑，並且以醣解作用做為例子，找出不同物種間近似的酵素。

*本論文研究為國科會補助之研究成果，計劃編號 NSC 91-2745-E-216-001

Sirava 等人[19]提出稱為 Pathfinder 的系統。此系統可將路徑中的每一個生化反應所使用的酵素，進行基因序列比對 BLAST [1]，之後將比對的結果，取最大的 p 值(p-value)，做為每一個反應的權重值，再將這些值取平均以後，做為不同物種的路徑相似度。

本文提出一種以矩陣為基礎的比對兩個生化反應路徑內容相似度的衡量方法。我們首先定義兩個生化反應途徑間相似度的計算方式，再進一步擴充運用在兩個生化反應網路間相似度的衡量，提供使用者依樣本查詢(query by pathway example)生化反應路徑資料庫的能力。

本論文的架構如下:第 2 節將對生化反應路徑的基本觀念作一介紹，並說明本文所提出的生化反應路徑資料庫系統整體架構，第 3 節說明現有生化反應路徑資料庫及其查詢功能，第 4 節說明依照內容比對兩個生化反應路徑相似度的衡量方法。第 5 章為本文結論。

二、反應路徑

2.1 反應路徑

生化反應(biochemical reactions)指的是在生物體內，為了維持生物個體機能正常運作，所進行的化學反應。生化反應與傳統的化學反應相似，其組成包含反應物(reactants)、催化劑(catalysis)、生成物(products)及一些輔助因子(cofactors)構成。生化反應與一般化學反應不同之處有二: 第一，在生化反應的過程中部分反應需要有能量的介入。例如: 利用 ATP 作為體內生化反的能量供應來源，當反應達成穩定時，ATP 的供給也會做調整，使得體內的狀態維持恆定。第二，為了使生物個體內的生化反應速率迅速進行，生化反應的另一個特徵，是它在反應的過程中，有許多的酵素(enzyme)參與反應。酵素是蛋白質的一種，每一種酵素具有高度的特異性(specificity)，只會與特定的反應物結合產生催化的作用。酵素可以加速反應進行的速率，但不會啟動新的生化反應。但少數酵素可在不同的生化反應中，扮演著不一樣的角色。依據實際反應需求，酵素其功能可藉由調節劑、抑制劑而有所改變。

圖 1 為醣解作用(glycolysis)的第一個反應。此反應的反應物為α-D-葡萄糖，生成物為α-D-葡萄糖-6 磷酸，反應的催化劑為己醣激酶(hexokinase, EC 2.7.1.1)，反應所需要的能量由 ATP 提供。

生化反應路徑(biopathways)是由一連串的生化反應所形成的網路。反應路徑進行的過程中，會經過許多的反應，產生一些中間產物(intermediates)，直到生成個體需要的生成物後，才會維持穩定。一個反應路徑進行時，生物體內其他地方也會有不同的反應路徑同時在進行。如果參與的反應物質有關連性，則這些生化反應路徑將會互相影響，使得生化反應路徑的複雜度大大提高。隨著生物技術的進步，產生了大量的生化反應路徑相關資訊。為了能夠運用電腦有效率地去輔助反應路徑的研究，以系統化的方式完整地紀錄這些反應路徑，是生物科技與生物資訊領域中一件刻不容緩的事情。

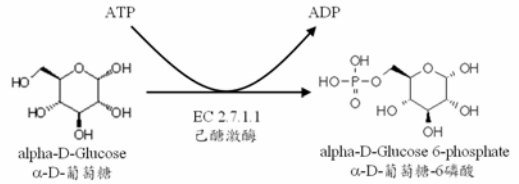


圖 1 葡萄糖分解反應

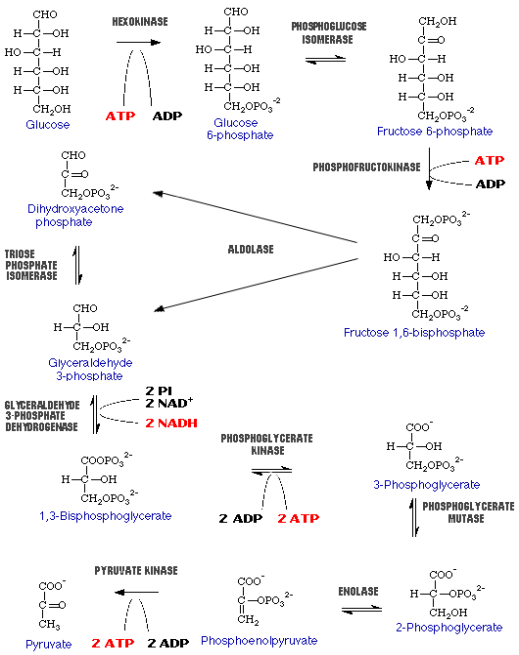


圖 2 醣解作用的生化反應路徑示意圖 (http://biotech.icmb.utexas.edu/glycolysis/pathway.html)

我們以一個最普遍的生化反應路徑為例來說明。醣解作用是最基本的能量代謝反應路徑，在整個代謝反應中，扮演產生能量產生的核心角色。由葡萄糖(Glucose)做為起始反應物，一直到丙酮酸(Pyruvate)的產生，此生化反應路徑需要經過 10 個主要的反應步驟，如圖 2 所示。每一個反應步驟中都有酵素的參與，在部分反應中有能量及其他的輔助因子參與反應，路徑的前半部（前五個步驟）需要消耗能量，後半部則會產生能量。由此例子可以看出生化反應路徑的資料與傳統文數字型態的資料，甚至與長字串為主的核酸序列或蛋白質序列，相較之下有著完全不同的資料型態。

如何衡量兩個生化反應路徑內容相似度是一項重要的研究主題。

2.2 酵素的分類

如前文所述，為了加快生物體內化學反應反應速率，都會有催化劑的參與，而生物體內最主要的催化劑則是酵素。由於有酵素的參與，使得反應速率大幅提高(最高可達 10^6 倍)，才能順利維持體內各項機能的運作。目前發現的酵素依照功能上來區分主要可以分成六大類別:氧化還原酶(Oxidoreductases)、轉移酶(Transferases)、水解酶(Hydrolases)、裂解酶(Lyases)、異構酶(Isomerases)、鍵結酶(Ligases)。國際生化命名委員會(IUBMB)為了統一酵素的分類，利用稱為 EC(Enzyme Classification number)來代表酵素的分類。EC 編號共有四組數字，第一個數字代表六大類別; 第二個數字代表次類別; 第三個數字代表亞次類別。在同一個分類中的兩個酵素，其催化的生化反應也會較為相似。以酒精去氫酶(alcohol dehydrogenase, EC 編號為 1.1.1.1)與高絲胺酸去氫酶(homoserine dehydrogenase, EC 編號為 1.1.1.3)為例，第一個 1 表示兩者的主類別為氧化還原酶，第二個 1 表示兩者的次類別為其供給者為-CHOH，第三個 1 表示兩者的亞次類別為其接受者為 NAD^+或 NADP，第四個數字表示序號，用以分別此兩個酵素。

2.3 生化反應路徑資料庫系統架構

我們的生化反應路徑資料庫系統整體架構，如圖3所示，包含反應路徑物件(pathway objects)、反應路徑模型(pathway modeling)、反應路徑資料庫(pathway database)、反應路徑模擬(pathway simulation)與反應路徑查詢(pathway retrieval)等五個部分。

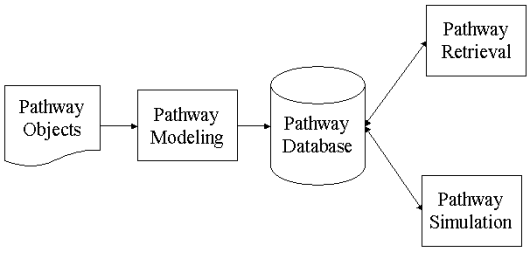


圖 3　生化反應路徑資料庫系統架構圖

一個反應路徑物件包含反應物、產物、酵素、輔助因子等參與生化反應的化合物，由這些化學物質共同組成生化反應路徑。在反應路徑模型的部分，我們提出可以利用Petri net圖形化工具來建立反應路徑的模型，並且以階層式結構呈現反應路徑圖形，達到視覺化的效果[16]。反應路徑資料庫儲存反應路徑的內容及每個反應路徑的模型，以提供生化反應路徑查詢資料之用。反應路徑模擬器係利用Petri net動態模擬的特性，讓建立的反應路徑模型在給予起始值與條件後，可以進行反應路徑的自動模擬。反應路徑查詢處理器可根據反應路徑的起始反應物、產物、催化劑(酵素)、及中間反應過程，進行兩個反應路徑間的相似性比對，找出兩反應路徑之間的相似度。

三、現有生化反應路徑資料庫

3.1 現有生化反應路徑資料庫

近幾年來，隨著生化反應路徑的角色日漸受到重視，許多生化反應路徑資料庫紛紛成立。一方面提供有需要的使用者，去查詢生化反應路徑的相關資訊，另一方面，可以作為各個研究單位研究成果相互交流資訊的媒介。這些與生化反應路徑相關的資料庫，散佈在歐洲、美洲以及亞洲。由此可知，全世界對於生化反應路徑的研究是相當的重視。以下是對目前主要的生化反應路徑資料庫的簡介。

3.1.1 Enzyme

Enzyme 是一個有關於酶的命名資料庫(http://www.expasy.ch/enzyme/)，由國際生化與分子生物聯合委員會(IUBMB)負責維護對於各種酶的命名資料。在 2002 年 8 月之時，其所包含酶的紀錄達到 3982 筆。近幾年來，Enzyme 資料庫提供許多公開的資料，作為其他代謝資料庫的連結與參考，並且允許使用者查詢與酶有關的生化反應。除此之外，Enzyme 資料庫與 SWISS-PROT 蛋白質序列資料庫，也會做資料同步更新與交互參考的動作。

3.1.2 KEGG

KEGG(Kyoto Encyclopedia of Genes and Genomes) 是日本京都大學化學研究所的研究小組，所開發的一個生物資料庫(http://www.genome.ad.jp/kegg/kegg2.html)。目前是日本 GenomeNet 伺服器上最核心的一個資料庫，主要提供對細胞或物種的基因資訊，做高層次的功能解釋。KEGG 由幾個主要的資料庫所構成，包含:PATHWAY 資料庫，用來記載生化反應的相關資訊;GENE 資料庫則是記載由基因序列計畫中所產生的基因與蛋白質資訊，LIGAND 資料庫則是記載有關於小分子化合物與細胞內生化反應的相關資訊。

3.1.3 WIT/MPW

WIT (What is There) 資料庫系統(http://wit.mcs.anl.gov/MPW/)，其設計的主要目的是用來 1.比較分析序列基因; 2.從染色體序列與 MPW 資料庫中，重新建構出代謝作用。在這個系統中，包含將近 40 個物種完整的與接近完整的基因資料。另外它們還有提供同源序列的資訊與記載基因產物在代謝反應路徑上的位置。透過 WIT 系統可以了解這些

基因與蛋白質的功能。MPW 是建構於 WIT 系統之下的一個專司處理代謝生化反應路徑的資料庫。MPW 存放的資料，包含主要與次要的代謝反應、膜的傳輸作用、訊號傳遞的反應路徑、細胞內的傳輸、轉錄與轉譯等，作為重建代謝反應路徑基礎。

3.1.4 BioCyc/MetaCyc

BioCyc 是一個專門收集生化反應路徑與基因組相關的知識庫(http://biocyc.org/)。資料庫以物種作為分類，不僅建立個別物種的生化反應路徑與基因組資料庫，更包含詳細的文獻資料。其中 MetaCyc 是一個專門整理代謝生化反應路徑的資料庫，主要的功用是作為 1.基因組在生化反應路徑上的分析; 2.代謝工程上的運用; 3.生物化學上的教材等三個方面。Metacyc 上的資料，超過 150 個物種的生化反應路徑。其中最完整的部分，則為大腸桿菌(*E.coil*)的生化反應路徑，與 EcoCyc 資料庫緊密結合，目前的資料持續在增加中。

3.2 現有生化反應路徑資料庫的查詢功能

目前的生化反應路徑資料庫所提供的查詢功能，仍然以關鍵字為條件的文字查詢為主。此外，也有一些生化反應路徑資料庫對生化反應路徑進行分類，提供使用者生化反應路徑階層式目錄瀏覽的功能。以下將針對兩種現有資料庫查詢功能做進一步說明。

- 關鍵字查詢: 所謂的關鍵字查詢，是以生化反應路徑資料中關鍵文字作為查詢的條件。關鍵字一般包含:生化反應路徑的名稱、反應物、中間反應的產物、最後的生成物、輔酶與酶等。這種查詢方式是目前生化反應路徑資料庫使用最多也是最基本的方式。在表 1 中整理現有各主要生化反應路徑資料庫關鍵字查詢功能。
- 階層式分類目錄瀏覽: 不同於文字查詢方式，有些生化反應路徑資料庫，將生化反應路徑依照名稱與功能，建立階層式的分類目錄，以供使用者瀏覽。這種方式的優點，有利於使用者透過瀏覽分類目錄，選擇需要的生化反應路徑。

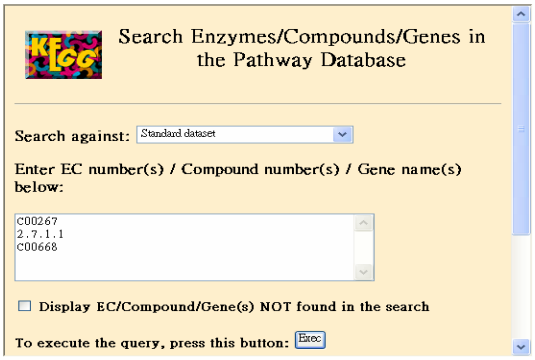


圖 4 KEGG 資料庫範例查詢

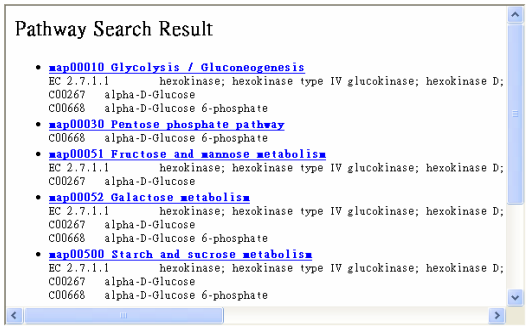


圖 5 KEGG 資料庫範例查詢結果(部分)

圖 4 為 KEGG 資料庫所提供之關鍵字查詢介面。使用者可以鍵入酵素的 EC 編號、化合物的代號或基因的編號作為查詢的條件。例如鍵入 C00267(α-D-葡萄糖)、2.7.1.1(己醣激酶)與 C00668(α-D-葡萄糖-6 磷酸) 作為查詢的條件。查詢結果如圖 5 所示，符合一個以上查詢條件生化反應路徑會逐筆列出。

表 1. 生化反應路徑查詢功能整理

資料庫名稱	查詢功能說明
ENZYME	主要以輸入酶的編號(EC 編號)、輔助因子與小分子化合物(chemical compounds)，作為查詢的條件。
EMP/MPW	除了酶的編號(EC 編號)外，還包括反應的中間產物、生成物與生化反應的分類等。
UM-BBD	內容以微生物為主，查詢方式與前兩者不同的地方，是它們有針對不同的微生物，去做生化反應路徑的查詢。除了文字查詢外，還有提供階層式的目錄查詢。 http://umbbd.ahc.umn.edu/
KEGG	查詢方式包含關鍵字的輸入與階層式的目錄查詢。其中 KEGG 的查詢結果，將生化反應路徑，以圖形的方式呈現，並提供主要生物資料庫交互參考超連接。
BioCyc	查詢方式包含關鍵字的查詢、階層式的目錄查詢與對基因組序列進行 BLAST 序列比對等。

而在查詢結果的呈現方式，主要可以分為文字模式與圖形模式兩大類。文字模式是將生化反應路徑以文字一一的呈現出來，包含:參與反應的基質、輔助因子、小分子化合物、酶及生成物等。有些生化反應資料庫，還提供一些參考連結到相關的網站，讓使用者能夠更深入了解每個查詢結果的細節。圖形模式是將生

化反應路徑視覺化表示，讓使用者能夠從圖上，清楚了解每個反應之間的關係。一般來說，一個完整的生化反應路徑，通常相當的複雜，為了讓使用者能夠清楚了解反應的細節，又要避免過多資訊增加查詢使用上的負擔，在設計上，盡可能以圖形提供各種複雜程度的查詢結果。

四、生化反應路徑資料的內涵式查詢

4.1 反應途徑的相似型比對

如前所述，一個生化反應路徑是由一連串的生化反應所形成的網路。為了避免混淆，我們將一連串的生化反應所形成的網路稱為***反應網路***(***reaction networks***)，而在反應網路中的一條路徑，也就是線性的一連串生化反應稱為***反應途徑***(***reaction paths***)。

定義 1　一個長度為 L 的反應途徑 Path = { R, P, π, L}。其中 R 為起始反應物的集合; P 為最終產物的集合; π 為中間反應的集合。

反應途徑的資料結構為一條串列(linked list)。其中起始反應物與最終產物表示反應途徑的起始點與終止點，中間反應集合表示由起始反應物到最終產物，需要經過的反應所形成的集合。在比較兩個反應途徑的相似度時，我們可以將反應途徑相似度，分成兩個部份來計算: 第一部份計算反應路徑的起始反應物/產物的相似度，第二部份針對反應的集合，找出兩個反應集合之間的相似度。最後將兩個部份的相似度依權重加總，便可得到完整的反應路徑間的相似度。計算流程如圖 6 所示。

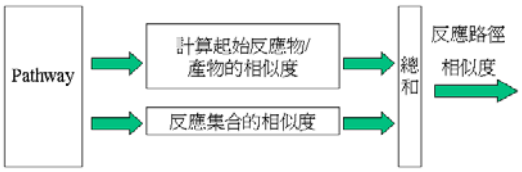


圖 6　反應途徑相似度計算流程

在一個反應途徑中，起始反應物與最終產物都是化合物。不同化合物之間的相似度計算非常複雜，我們目前對兩化合物的相似度設定採取最簡單的方式: 完全相同的化合物，相似度為 1; 不相同的化合物，相似度為 0。

在兩個生化反應的相似度計算部份，由於每一個酵素大多對應至一個生化反應，我們可以利用每個酵素的 EC 值做為生化反應相似度判斷的標準。EC 值四個碼由左到右將酵素的功能，逐層進行分類，所以編號由左到右越相近的酵素，表示對應的反應也越近似。所以我們可以設定如表 2 的兩個酵素(反應) 相似度評分表。例如 EC 2.7.1.1 與 EC 2.7.1.2，都是作用於六碳糖上的己醣激酶，皆可催化圖 1 的反應，故兩者間的相似度可設為 0.8。

表 2 比較兩個生化反應相似度評分表

EC number 四碼相同的個數	相似度
四碼相同	1.0
前三碼相同	0.8
前兩碼相同	0.6
第一碼相同	0.4
都不同	0

兩個中間反應集合的相似度，可利用反應與反應兩兩比對的方式，找出最大的相似度，然後加總當作反應集合的相似分數。

對於兩條反應途徑$Path_1 = \{ R_1, P_1, \pi_1, L_1\}$與$Path_2 = \{ R_2, P_2, \pi_2, L_2\}$，兩者間的反應途徑相似度的計算公式如公式(1)，其中w_1、w_2與w_3，分別表示起始反應物相似度$Reactant(R_1, R_2)$、最終產物相似度$Product(P_1, P_2)$與中間反應集合相似度$Path(\pi_1, \pi_2)$，在計算整個反應途徑相似度中，所佔的權值。其值可依照實際情形做調整。在本論文中，w_1、w_2與w_3的設定分別為0.2、0.2與0.6。

$$Similarity(Path_1, Path_2) = w_1 \times Reactant(R_1, R_2) + w_2 \times Product(P_1, P_2) + w_3 \times Path(\pi_1, \pi_2)/L_1 \quad (1)$$

舉例來說，圖7以反應途徑Q做為查詢反應途徑樣本與三個反應途徑P_1、P_2、P_3做比較。分別可以得到相似度為

$$Similarity(Q, P_1) = 0.2*(1) + 0.2*(1) + 0.6*\frac{0.8}{1} = 0.88$$

$$Similarity(Q, P_2) = 0.2*(1) + 0.2*(1) + 0.6*\frac{0}{1} = 0.4$$

$$Similarity(Q, P_3) = 0.2*(0) + 0.2*(1) + 0.6*\frac{0}{1} = 0.2$$

所以相似度比較結果為Q與P_1比較相近，P_2次之，P_3最不相似。

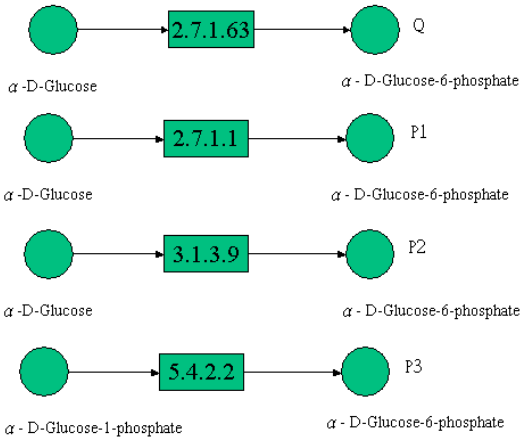


圖 7 反應途徑相似度比較範例

4.2 反應網路的矩陣表示法

前面我們提出的反應途徑比對方式，能夠針對直線的反應途徑比較相似度，而一個生化反應路徑，不單單只有直線的結構，而是一個複雜的生化反應網路地圖。我們可以用有向圖 G (V, E)來表示，其中 G 中的節點集合 V 代表反應網路中所有的反應，同樣地以我們酶的 EC 編號做為對應反應的代表，G 中的邊線集合 E 代表與兩個相鄰反應之間的連續關係。以圖 8 為例，此反應網路對應之有向圖 G (V, E) = ({5.1.3.9, 2.7.1.11, 4.1.2.13, 1.2.1.12, 2.7.2.3, 3.6.1.7, 3.1.3.11}, {(5.1.3.9, 2.7.1.11), (2.7.1.11, 4.1.2.13), (4.1.2.13, 1.2.1.12), (1.2.1.12, 2.7.2.3), (1.2.1.12, 3.6.1.7), (3.1.3.11, 5.1.3.9}) 。

在考慮兩反應網路間的相似性時，由於研究中的反應網路的資料往往是殘缺不全的，我們想要知道的是兩個反應網路之間有哪些反應途徑是一樣的或近似的，這種資訊可以給研究者一些研究方向與實驗設計的指引。因此，我們可以加總兩個反應網路之間的所有對應反應途徑的相似度來求反應網路間的相似度。為了能夠有系統進行兩個反應網路中所有反應途徑的配對比較，我們定義了反應網路的矩陣表示法，稱為 ***反應矩陣*** (***reaction matrixes***)。

定義 2　給定一反應網路 G (V, E) ，其對應之反應矩陣 M_L 為一個 N x N 的矩陣，其中 N 為反應網路 G 中出現的反應個數，L 表示此反應矩陣所記錄所有反應途徑的長度。矩陣中每一個元素 p_{ij} $(1 \le i \le N, 1 \le j \le N)$表示由起始反應 R_i 到終止反應 R_j 的反應途徑。$|M_k|$表示矩陣中值非零的元素個數，也就是反應網路中長度為 L 的反應途徑個數。

圖 9 為圖 8 中的反應網路所對應之反應矩陣，為一個 7×7 矩陣(M_1) ，記錄所有長度為 1 的反應途徑。矩陣的列表示反應途徑的起始反應，矩陣的行表示反應途徑的終止反應。元素 $p_{12} = 0$ 表示反應網路中沒有由 1.2.1.12 起始而以 2.7.1.11 終止的反應途徑; $p_{13} = 1$ 表示反應網路中有由 1.2.1.12 起始而以 2.7.2.3 終止的反應途徑。每一列中元素的總和，表示該反應連接其他反應的個數，例如矩陣的第 1 列元素的總和為 2，表示有 2 個反應途徑以 1.2.1.12 為起始反應，而其終止反應分別為 2.7.2.3 與 3.6.1.7。

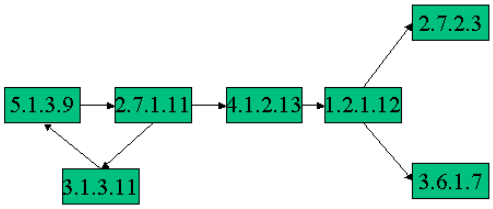


圖 8 反應網路之有向圖表示法

起點＼終點	1 .2 .1 .12	2 .7 .1 .11	2 .7 .2 .3	3 .1 .3 .11	3 .6 .1 .7	4 .1 .2 .13	5 .1 .3 .9
1.2.1.12	–	0	1	0	1	0	0
2.7.1.11	0	–	0	1	0	1	0
2.7.2.3	0	0	–	0	0	0	0
3.1.3.11	0	0	0	–	0	0	1
3.6.1.7	0	0	0	0	–	0	0
4.1.2.13	1	0	0	0	0	–	0
5.1.3.9	0	1	0	0	0	0	–

圖 9 反應網路之矩陣表示法

使用矩陣表示法來表示反應網路的一項優點是我們可以經由類似矩陣乘法的運算，由反應網路中的兩個反應矩陣 M_X 與 M_Y 推導出 M_{X+Y}。

定義 3　假設一個反應個數為 N 的反應網路 G (V, E) 的兩個反應矩陣 M_X 與 M_Y，則

$$M_{X+Y} = M_X \otimes M_Y \qquad (2)$$

其中 ⊗ 運算涵義為 $\begin{cases} \gamma_{ik} = 0, \ i = k \\ \gamma_{ik} = \alpha_{ij}\beta_{jk}, \ i \neq k \end{cases}$

$\gamma_{ik} \in M_{X+Y}, \alpha_{ij} \in M_X, and \ \beta_{jk} \in M_Y$

以上公式表示若由反應 i 到反應 j 有一個長度為 X 的反應途徑，且由反應 j 到反應 k 有一個長度為 Y 的反應途徑，則存在由反應 i 到反應 k 的一個長度為 $X+Y$ 之反應途徑。

對於一個反應網路，我們可由其有向圖得到對應長度為 1 之反應矩陣 M_1，再由公式(2)求得 $M_2, M_3, \ldots, M_L$。其中 L 為反應網路中最長反應途徑的長度。由於 ⊗ 運算的特性，長度大於 L 的反應矩陣皆為零矩陣，即使反應網路中有迴路亦然。

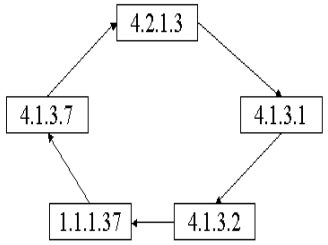
4.3 反應網路的相似型比對

前文中我們定義了反應網路的兩條反應途徑間的相似度比對方法。但線性反應途徑只是反應網路中的一種特例，事實上一個典型的反應網路不像線性反應途徑這麼單純，而是可能包含多種結構，如: 直線的反應途徑、分支路徑與迴路(循環) 路徑的反應網路。並且多個小的反應網路可再組成一個大的反應網路。由於反應矩陣能夠完整地記錄一個反應網路中的所有反應途徑，我們可以藉由比較兩個反應矩陣之間的相似度來求得兩個反應網路的相似度。

給定兩個要比較的反應矩陣為 M 與 N，M 為一個 P x P 的反應矩陣，α_{ij} $(1 \le i \le P, 1 \le j \le P)$ 表示 M 矩陣中的每一個非零元素; N 為一個 Q x Q 的反應矩陣，β_{kl} $(1 \le k \le P, 1 \le l \le P)$表示 N 矩陣中的每一個非零元素。則兩反應矩陣的相似度可依照以下公式進行計算:

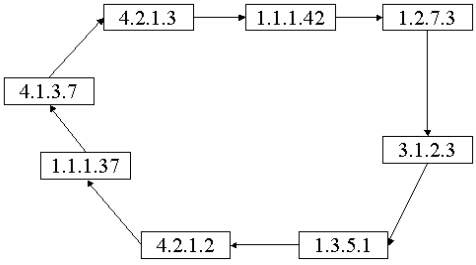
$$Similarity(M,N)=\frac{\sum_{\forall \alpha_{ij}\in M,\alpha_{ij}\neq 0} \underset{\forall \beta_{kl}\in N,\beta_{kl}\neq 0}{Max}(Similarity(\alpha_{ij},\beta_{kl}))}{|M|} \quad (3)$$

舉例說明如下。我們以乙醛酸循環(Glyoxylate cycle) 為查詢樣本，對檸檬酸循環(TCA cycle)做相似度比對。乙醛酸循環與乙醛酸循環反應路徑矩陣如圖 10 所示; 檸檬酸循環與檸檬酸循環反應路徑矩陣如圖 11 所示。乙醛酸循環長度為 1 之反應矩陣之非零元素共有 5 個，與檸檬酸循環中最近似反應途徑的相似度分別為 0.5、0.6、0.7、1.0 與 1.0，將這些個別路徑的相似度，加總取平均，得到最後路徑的相似度為 0.76，表示將乙醛酸循環對應到檸檬酸循環中，會得到 0.76 的相似度。



起點＼終點	4	4	4	1	4
	.2	.1	.1	.1	.1
	.1	.3	.3	.1	.3
	.3	.1	.2	.37	.7
4.2.1.3	–	1	0	0	0
4.1.3.1	0	–	1	0	0
4.1.3.2	0	0	–	1	0
1.1.1.37	0	0	0	–	1
4.1.3.7	1	0	0	0	–

圖 10　乙醛酸循環與乙醛酸循環反應路徑矩陣



	4.2.1.3	1.1.1.42	1.2.7.3	3.1.2.3	1.3.5.1	4.2.1.2	1.1.1.37	4.1.3.7
4.2.1.3	–	1	0	0	0	0	0	0
1.1.1.42	0	–	1	0	0	0	0	0
1.2.7.3	0	0	–	1	0	0	0	0
3.1.2.3	0	0	0	–	1	0	0	0
1.3.5.1	0	0	0	0	–	1	0	0
4.2.1.2	0	0	0	0	0	–	1	0
1.1.1.37	0	0	0	0	0	0	–	1
4.1.3.7	1	0	0	0	0	0	0	–

圖 11　檸檬酸循環與檸檬酸循環反應路徑矩陣

事實上，這兩個反應路徑都由檸檬酸做為反應的起點。兩個反應的最大不同點，在於檸檬酸循環會將六碳結構的化合物，分解成兩個二氧化碳與一個四碳結構的化合物，並且產生能量。檸檬酸循環大多出現在動物體內。而乙醛酸循環中，乙醛酸為一個二碳的化合物，在反應路徑的中間會與乙醯輔酶A發生作用，變成四碳結構的化合物，並且回到檸檬酸循環的後半段反應。乙醛酸循環大多存在於植物與微生物中。

給定兩個要比較的反應網路為 G_1 與 G_2，假設其所有非零矩陣之反應矩陣分別為 M_1, M_2, …, M_{L1} 與 N_1, N_2, …, N_{L2}，則兩反應網路的相似度為

$$Similarity(G_1,G_2)=\frac{\sum_{i=1}^{L_1}\sum_{j=1}^{L_2} Similarity(M_i,N_j)}{L_1L_2} \quad (4)$$

五、結論

在本篇論文中，我們提出一種以矩陣為基礎的比對兩個生化反應路徑內容相似度的衡量方法，來讓使用者依樣本查詢生化反應路徑資料庫。我們首先定義兩個生化反應途徑間相似度的計算方式，再進一步擴充運用在兩個生化反應網路間相似度的衡量，此方法可以在數量龐大的反應路徑資料庫中，有系統地找到功能相近的反應路徑。在未來工作方面，我們計畫整合網路上現有的反應路徑資料庫，依照本文的查詢處理方法建構一完整的反應路徑資料庫。此外，我們將研究如何以有效率的演算法來加速反應路徑的相似性比對。最後為了能夠讓反應路徑查詢結果，具有資料交換的功能，可以利用 XML 語言來描述反應路徑查詢結果。

六、參考文獻

[1] Stephen F. Altschul et al., “Basic Local Alignment Search Tool. ” J. Mol. Biol., Vol. 215, pp. 403-410, 1990.

[2] Moritz Y. Becker and Isabel Rojas , “A graph layout algorithm for drawing metabolic pathways,” Bioinformatics, Vol. 17, pp.461-467, 2001.

[3] Cornish-Bowden, A. and Hofmeyr, J.H , “METAMODEL: A program for modeling and control analysis of metabolic pathways on the IBM pc and compatibles,” Comput. Applic. Biosci., Vol. 7, pp. 89-93, 1991.

[4] Dandekar T., Schuster S., Snel B. et al.,” Pathway alignment: application to the

comparative analysis of glycolytic enzymes," Biochem. J., 1: pp.15-24., 1999.

[5] Garfinkel D. ,"Computer modeling of metabolic pathways," Trends Biochem. Sci., Vol. 6, pp.69-71, 1981.

[6] Alexander Goesmann et al., "PathFinder: reconstruction and dynamic visualization of metabolic pathways", Bioinformatics, 18, pp.124-129, 2002

[7] R. Hofestädt and S. Thelen, "Quantitative modeling of biochemical network," In Silico Biology, Vol. 1, pp.39-53, 1998.

[8] Minoru Kanehisa, et al, "The KEGG databases at GenomeNet," Nucl. Acids. Res. , Vol. 30, pp.42-46, 2002.

[9] Minoru Kanehisa and Susumu Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," Nucl. Acids. Res. , Vol. 28, pp.27-30, 2000.

[10] Peter D. Karp, et al, "The MetaCyc Database", Nucl. Acids. Res. ,30, pp.59-61, 2002.

[11] Karp, P.D. " Pathway Databases: A Case Study in Computational Symbolic Theories," Science, Vol. 293, pp. 2040-2044, 2001.

[12] Peter D. Karp, et al, "The EcoCyc and MetaCyc databases," Nucl. Acids. Res. , Vol. 28, pp.56-59, 2000.

[13] Robert Küffner, Ralf Zimmer, and Thomas Lengauer, "Pathway analysis in metabolic databases via differential metabolic display (DMD)," Bioinformatics, 16, pp.825-836, 2000.

[14] Hiroyuki Ogata, et al, "A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters," Nucl. Acids. Res. ,28, 4021-4028, 2000.

[15] Ross Overbeek, et al, "WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction," Nucl. Acids. Res. , Vol. 28, pp.123-125, 2000.

[16] James L. Peterson, "Petri Nets," ACM Computing Surveys, Vol. 9, Issue 3, Sep. 1977.

[17] Christophe H. Schilling and Bernhard O. Palsson, "The underlying pathway structure of biochemical reaction networks," Proc. Natl. Acad. Sci. USA, Vol. 95, pp.4193-4198, 1998.

[18] E. Selkov, et al, "MPW: the Metabolic Pathways Database," Nucl. Acids. Res. , Vol. 26, pp.43-45, 1998.

[19] M. Sirava, et al, "BioMiner—modeling, analyzing, and visualizing biochemical pathways and networks," Bioinformatics, Vol 18, pp.s219-s230, 2002.

[20] Takako Takai-Igarashi and Tsuguchika Kaminuma, " A Pathway Finding System for the Cell Signaling Networks Database," In Silico Biology, Vol. 1, pp.129-146, 1999.