行政院國家科學委員會專題研究計畫成果報告

辨識 PAI 的雛形系統之研究

A Study on the Prototype System for Identifying PAI

計畫編號：NSC 89-2218-E-126-006
執行期限：89 年 8 月 1 日至 90 年 7 月 31 日

主持人：林耀鈴 靜宜大學 資訊管理學系
共同主持人：蔡英德 靜宜大學 資訊管理學系
共同主持人：許芳榮 靜宜大學 會計學系
協同主持人：張晃猷 清華大學 生命科學所
計畫參與人員：郭嘉和 靜宜大學 資訊管理學系
E-mail: yllin@pu.du.tw

一、中文摘要

細菌感染仍是人類健康最主要的威脅之一。雖然抗生素治療仍被普遍使用，但是抗藥性菌株的廣泛流行，已明顯的降低抗生素的效用。最近由於細菌基因體學的發展，揭露了許多前所未知的細菌基因及相關資訊，其中有許多新基因值得作為未來藥物研發的對象。一般而言，與細菌致病能力有關的基因是最值得作進一步的探討，而許多具致病能力的細菌都帶有一長段的特殊 DNA 片段，稱作致病毒力基因島 (PAI)。PAI 的常見特徵包括 1)帶有數個致病相關基因； 2)G+C 的比例和該細菌其他基因體的區段有所不同； 3)兩端有特殊的插入序列，如 tRNA 基因序列或是同向的重複序列； 4)具有協助轉移或嵌入至細菌染色體上的酵素基因。目前為止，並無一方便迅速的方法在大量的基因片段的序列中偵測到 PAI。因此本計畫擬利用以上所述的 PAI 特徵，發展一軟體系統，用以迅速搜尋現有微生物基因庫中存在的 PAI

關鍵詞：PAI 、圖形化、G+C 比例、序列比對、生物特徵序列資料庫。

Abstract

Bacterial infections are still the major threat to human health. The recent progress on bacterial genome program provides many opportunities to identify novel targets for drug intervention. The pathogenic strain of a bacterium is different from its virulent relative in that it has acquired a large segment of DNA called *Pathogenicity Island (PAI*). PAIs normally carry genes encoding several virulence factors. The identification of a PAI in a bacterial genome can greatly facilitate the understanding of bacterial pathogenesis and development of antimicrobial we develop a software system that allows one to identify PAIs rapidly from thousands of contig in the bacterial genome database.

Keywords: PAI, Sequence Visulization, G+C ratio, Sequence Comparison, Bio-characteristics database.

二、緣由與目的

目前細菌性的傳染病仍是導致人類死亡的主要原因之一。已往都採用抗生素治療，但由於抗藥性菌株的廣泛流行，已明顯的降低抗生素的效用。最近由於細菌基因體學的發展，揭露了許多前所未知的細菌基因及相關資訊，其中有許多新基因值得作為未來藥物研發的對象。許多生物學家透過一系列微生物學和免疫學研究，想要發現新的抗生素。一般而言，與細菌致病能力有關的基因是最值得作進一步的探討。生物學家發現許多致病的細菌都有一段特殊的 DNA 片斷，稱作致病毒力基因島 (PAI)[13]。透過PAI 的研究，生物學家能夠發現這些致病細菌的許多重要特性，以便於能製造新的藥物。PAI 的常見特徵包括 1)帶有數個致病相關基因[12, 14, 17]；2)G+C 的比例和該細菌其他基因體的區段有所不同；3)兩端有特殊的插入序列，如 tRNA 基因序列或是同向的重複序列[4, 5, 8, 20, 21])具有協助轉移或嵌入至細菌染色體上的酵素基因。

目前PAI 的研究仍要透過繁鎖的實驗分析，並無一方便迅速的方法幫助生物學家在大量的基因片段序列中偵測到PAI 。因此本文利用以上所述的PAI 特徵，發展一軟體系統，用以迅速搜尋現有微生物基因庫中存在的PAI 。首先我們將建立與 PAI 特徵相關的資料庫，如插入序列資料庫。然後設計並發展快速搜尋PAI 的序列比對演算法。此外，這個系統將提供能縮短生物研究流程的實用軟

體工具，如DNA G+C的比例計算及蛋白質親水性比例計算等工具程式。此一軟體將先以已知具有PAI 的細菌基因庫如幽門曲狀桿菌作為測試對象，並進行修改。最後再針對包括肺炎克雷白氏菌等新完成的微生物基因庫進行比對。希望能及早發現這些微生物中基因體中所存有的PAI ，以供生物學家進一步之研究。相信此一軟體對於搜尋動物及植物病原細菌所帶有的PAI 也將會有很大的幫助。

二、相關研究及現存系統

目前已有很多用來辨識分析生物序列的序列比對分析程式，如美國NCBI 的Blast [18]、日本DDBJ的FastA [7]、Smith-Waterman 、臺灣工研院的FLAG [10]等等。Blast [3, 4]從1990 年就一直發展至今，已經有很多系列版本被發展出來，可以運用在各種生物序列比對上。它是根據統計理論所設計的，雖然它也是一種heuristic 運算法，其計分方式卻與像dynamic programming 一樣，能模擬找出突變最少的狀況的過程。

FastA [16, 19]是在1985 年就已開發出來，至今亦經過多次改良。它是利用序列片段的相似性來推測可能有親緣關係的區域，再將少量找到的區域做嚴謹的Local Alignment 分析，可以節省許多計算的時間。Smith-Waterman[9] 利用Dynamic programming 的方法，能有系統地尋找同源的序列，可是它的速度太慢，必須用硬體加速才實用。FLAG 是臺灣工研院生醫中心最近發展的DNA 序列比對程式，特別是針對基因體與基因體之間的序列比對。

三、系統功能與描述

3.1 G+C 比例計算：

由於PAI 的特徵：G+C 的比例和該細菌其他基因體的區段有所不同，以及圖形化的考量，所以本系統具備G+C 比例計算的功能。只要在輸入DNA 序列檔案或直接輸入DNA 序列時，選擇DNA 型式，即可輸出G+C 比例圖形。在圖形底下的「輸入SIZE 」物件可以 控制G+C 比例，使之放大或縮小。

3.2 序列尋找

為了滿足生物學家的需要，以便他們可以做更多種類的分析、研究，本系統加入了序列尋找的功能。此功能提供多種條件選項讓使用者自己組合出符合需求的條件情況，條件選項可複選也可單選，會有不同的結果出現。若只選擇「(%)以上」和「(%)以下」，會尋找出符合條件並最長的一個序列。(如圖五) 若只選擇「bp 以上」和「bp 以下」，則不會找到任何序列。若複選這2 種條件選項，會尋找出符合條件的所有序列。(如圖六) 除了圖形結果輸出外，尋找到的序列文字資料也會同步輸出到「範圍序列」中。

3.3 序列比對

目前有很多的序列比對程式，而使用最廣的是FastA 與Blast 系列的程式。我們以這個兩程式加上最嚴謹的序列比對演算法(Local Alignment)，就它們的優缺點來考量：

FastA 系列: 利用序列片段的相似性來推測可能有親緣關係的區域，再將少量找到的區域做嚴謹的Local Alignment分析，可以節省許多計算的時間。靈敏、但是速度較Blast 慢。只能以核酸序列比對核酸資料庫，或以蛋白質序列比對蛋白質資料庫。TFastA 則以蛋白質序列比對核酸序列資料庫。

Blast 系列: 比對速度非常快，但在序列相似性較低時會有失誤，有可能會漏掉一些相關的序列。它是一組程式，可以根據輸入的序列與所選擇的資料庫種類而執行適當的程式。例如：BlastN 以DNA 序列去查詢DNA 資料庫、BlastP以蛋白質序列去查詢蛋白質資料庫、Blastx 以DNA 序列去查詢蛋白質資料庫、TBlastN 以蛋白質序列去查詢DNA資料庫。

Smith-Waterman系列: 是最嚴謹的序列比對，但處理速度太慢，而且會耗用很大的儲存空間，例如兩個 1Kb 的序列互相比對，至少要佔用 10_6 個位置存得分。在改善儲存空間方面，可以利用Linear space 的方法來降低對記憶體之需求量。可是在改善速度方面，雖然可用硬體加速來增加執行速度，但資料量一多起來，還是很花時間 [1]。

基於速度較快、功能較多的理由，本系統採用Blast程式來做序列比對。至於它的缺點：不夠嚴謹，本系統則另外增加Local Alignment演算法來輔助、補強。資料量很大時，建議使用Blast 來做序列比對，若要再做細微一點的比對時，便可使用Local Alignment 演算法來做較精確的局部比對。資料量比較小又不使用特殊功能時，即可直接利用Local Alignment 演算法來做序列比對。除了圖形結果輸出外，比對出來的序列文字資料也會同步輸出到「範圍序列」中。

3.4 生物特徵序列資料庫

本系統提供生物特徵序列資料庫，讓Blast 、Local Alignment 做資料庫序列比對時使用。採用動態的方式來建構資料庫，使用者除了可以利用工具列編輯資料庫外，還可以用Excel 依照格式編輯，並存成CSV 檔案格式，需要使用時再輸入即可。因此，使用者可以依據需要，編輯好幾組資料庫表單，等於是使用者可以自己幫資料庫分門別類，增加比對時的多樣性、機動力。在對資料庫做比對時，會將

結果依據資料庫表單上的顏色設定分別塗色，以幫助使用者容易分辨。

3.5 使用者設定

由於本系統是採用圖形化呈現方式，所以顏色就變得相對重要。考慮到不同的使用者可能對不同的顏色比較敏感，在某些顏色組合下，能看得比較清楚、容易辨識。所以，本系統提供顏色設定的功能，讓使用者可以自己控制二組圖形、背景、序列、標線的顏色。其中「序列」指的是20 個胺基酸，每一個胺基酸有其代表的顏色，並提供「儲存」的功能，以便以後還可以使用同樣的顏色設定。

3.6 便利的小功能

為了讓使用者能夠更方便、迅速辨識PAI ，本系統含有很多便利的小功能來幫助使用者。以下便是其中的一些：

局部放大
若資料量太大，使用者想看清楚局部區域的圖形時，便可利用此功能，把滑鼠拖曳選擇的部份，依與整個畫面的比例大小放大在畫面上。

局部顯示
若資料量太大，使用者想看清楚局部區域的序列資料時，便可利用此功能，把滑鼠拖曳選擇的部份，將其代表的序列資料顯示在「範圍序列」裡。另外，在序列尋找、比對完之後，結果所顯示的區域上方都會一個指示的倒三角形。以滑鼠右鍵點選這些標記，會將此標記代表的序列及相關資訊顯示在「範圍序列」中。

對照比較
本系統可同時提供二組比例大小不一樣的圖形，以便使用者可以對照比較。只要在「輸入SIZE 」設定即可，字型的顏色是代表圖形的
顏色。

對照序列
另外，為了讓使用者清楚分辨胺基酸或鹼基，本系統設定只要出現在畫面上序列長度少於50 時，就會在每一個胺基酸或鹼基的正上方位置秀出代號並塗色以便辨識。

水平捲軸
若已利用局部放大的功能，將局部區域放大到整個畫面上，便可使用水平捲軸的功能，藉移動捲軸來觀看其它區域的圖形。另外，也可利用滑鼠右鍵按住圖形不放，然後再水平移動，一樣會有跟水平捲軸相同的功能。

標線
按滑鼠左鍵點在圖形上，則會標示此位置的一條直線，並且在上方會顯示此位置的資訊及胺基酸或鹼基代號。除了出現上述的標線外，還會將此位置的數值及本身加左右各30個的胺基酸或鹼基顯示在「目標序列」中。

儲存序列資料
為了方便使用者記錄序列資料，本系統提供儲存序列資料的功能，可以將顯示在範圍序列中的序列資料存成TXT 純文字檔。

3.7 系統應用測試

我們藉由以下的設備來對幾組DNA 序列作Blast 比對：Windows 98 SE 作業系統、Celeron1000 CPU 、256RAM 記憶體。輸入Pseudomonas aeruginosa PA01 (6.2MB) 的繪圖時間為2.0秒。比對的對象分別為PAGI-1 (65KB,)、PAI_O157 (56KB)、SPI3 (21KB)、PAI 6 (7KB, partial)，E-value 設定為1e-299 ，處理時間分別為6.7 秒、5.2 秒、4.8 秒、4.6 秒。輸入Escherichia coli O157:H7 (5.6MB) 的繪圖時間為1.6 秒。比對的對象分別為PAGI - 1 、PAI _O157 、SPI3 、PAI 6 ，E-value 設定為3e-11 ，處理時間分別為4.7 秒、5.9 秒、4.2 秒、4.2 秒。輸入Pyrococcus abyssi (1.7MB) 的繪圖時間為0.7 秒。比對的對象分別為PAGI - 1 、PAI _O157 、SPI3 、PAI 6 ，E-value 設定為3e-11 ，處理時間分別為1.9 秒、2.0 秒、1.8 秒、1.7 秒。若比對出來的結果愈多，則處理的時間會相對地增加。

四、計畫成果與討論

這個系統將幫助生物學家快速地辨識、分析PAI ，並且預期發現新的PAI 。利用獲知的資訊製造新的藥物，以抑止細菌感染所造的傳染病。更進一步，希望此套系統可運用於動植物的DNA 分析上，以促進生物學的發展。由於本系統擁有眾多功能，除了可以辨識、分析PAI 外，更可以運用在其它的生物資料分析上。今後，也仍會繼續增加其它能幫助生物學家做分析工作的功能，以期本系統能真正達到協助生物學家研究的目的。

五、參考文獻

[1] 序列資料庫搜尋程式
http://binfo.ym.edu.tw/post/topics/srch_prg.htm

[2] 淺談人類基因體計畫
http://life.nthu.edu.tw/~stass/animalfarm/5th/One-5.htm

[3] S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, Basic local alignment search tools. *J. Mol. Biol* **21**5: 403-410, 1990.

[4] S.F. Altschul, T.L. Madden, Alejandro A.Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and D.J. Lipman, *Gapped BLAST and PSI-BLAS*T: a new generation of protein database search programs, Nucleic Acids Res. 25:3389-3402, 1997.

[5] C. Buchrieser, R. Brosch, S. Bach, A. Guiyoule and E. Carniel, The high pathogenicity island of *Yersinia tuberculosis* can be inserted into any of the three chromosomal *asn* tRNA genes. *Mol. Microbiol.,* **3**0: 965-978, 1998.

[6] T.H. Cormen, C.E. Leiserson, R.L. Rivest, Introduction to Algorithms : Dynamic Programming, p301-328.

[7] DNA Data Bank of Japan http://www.ddbj.nig.ac.jp/

[8] U. Dobrindt, P.S. Cohen, M. Utley, I. Mühldorfer and J. Hacker, The *leu*X-encoded tRNA5 Leu but not the pathogenicity islands I and II influence the survival of the uropathogenic *Escherichia coli* strain 536 CD-1 mouse bladder mucus in the stationary phase. *FEMS Microbiol. Lett.,* **16**2: 135-141, 1998.

[9] Richard Durbin, S.R. Eddy, Anders Krogh, Graeme Mitchison, Biological sequence analysis : Pairwise alignment, p12-45.

[10] Fast Local Alignment for Gigabases http://flag.itri.org.tw/

[11] Gusfield , *Algorithms on Strings, Trees, and Sequence*s, Cambridge University Press, 1997.

[12] J. Hacker, L. Bender, M. Ott, J. Wingender, B. Lund, T. Marre and W. Goebel, Deletions of chromosomal regions coding for fimbriae and hemolysins occur *in vivo* and *in vitro* in various extraintestinal *Escherichia coli* isolates. *Micro. Pathog.,* 8: 213-225, 1990.

[13] J. Hacker, G. Blum-Oehler, I. Mühldorfer and H. Tschäpe, Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol. Microbiol.,* **2**3: 1089-1097, 1997.

[14] J. Hacker, S. Knapp and W. Goebel, Spontaneous deletions and flanking regions of the chromosomal inherited hemolysin determinant of *Escherichia coli* O6 strain. *J. Bacteriol.,* **15**4: 1145-1152, 1993.

[15] Human Genome Project http://www.ornl.gov/hgmis/

[16] D.J. Lipman, W.R. Person: *Rapid and sensitive protein similarity searche*s, Science, 227, 1435-1441, 1985.

[17] D. Low, V. David, D. Lark, G. Schoolnik and S. Falkow, Gene clusters governing the production of hemolysin and mannose-resistant hemagglutination are closely linked in *Escherichia coli* serotype O4 and O6 isolates from urinary track infections. *Infect. Immun.,* **4**3: 353-358, 1984.

[18] National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov/

[19] W.R. Pearson, D.J. Lipman: *Imported tools for biological sequence comparis*on, Proc. Natl. Acad. Sci. USA, 85, 2444-2448, 1988.

[20] A. Ritter, D. Gally, P.B. Olsen, U. Dobrindt., A. Friedrich, P. Klemm and J. Hacker, The Pai-associated leuX specific tRNA5 Leu affects type 1 fimbriation in pathogenic *Escherichia coli* by control of FimB recombinase expression. *Mol. Microbiol.,* **2**5:871-882, 1997.

[21] H. Schmidt, J. Scheet, C. Janetzki-Mittermann, M. Datz and H. Karch, An *ileX* tRNA gene is located close to the Shiga toxin II operon in enterohemorrhagic *Escherichia coli* O157 and non-O157 strains. *FEMS Microbiol. Lett.,* **14**9: 39-44, 1997.