行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※　　DNA 序列中特殊調控訊號之組合型辨識方法　　※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：個別型計畫

計畫編號：NSC 90－2213－E－009 －169－

執行期間：90 年 8 月 1 日至 91 年 7 月 31 日

計畫主持人：國立交通大學資訊科學系 胡毓志

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：交通大學資訊科學系

中　華　民　國　91　年　9　月　18　日

行政院國家科學委員會專題研究計畫成果報告
國科會專題研究計畫成果報告撰寫格式說明
Preparation of NSC Project Reports

計畫編號：NSC 90-2213-E-009 -169-
執行期限：90 年 8 月 1 日至 91 年 7 月 31 日
主持人：胡毓志　交通大學資訊科學系
計畫參與人員：陳兆奕 徐英哲 邱炫超 李如珍 交通大學資訊科學系

一、中文摘要

生物學者已經確定基因的控制與調控主要是由位於基因周圍相當短的序列所決定的。這些序列在長度、位置、重複性、方向性，以及鹼基上各有不同。對分子生物學而言，尋找這些短的序列是個重要的問題，它有著其它重要的應用。儘管目前已存在許多不同的方式用來尋找這些訊號(signal，也就是短序列)，根據一些最新的研究顯示，這個問題仍有許多改善的空間。在 2000 年，Pevzner 與 Sze 兩人提出了特色構形偵測的挑戰問題。他們指出目前尋找特色構形的最新演算法也不能偵測出在挑戰問題裡的特色構形目標物。這篇論文，藉著使用反覆重新開始的設計，我們的新演算法可以正確地找出這些特色構形目標物。更者，再加上考慮某些轉錄因子會形成二聚體或是更複雜的結構，以及轉錄過程中可能涉及多個含有間隔的轉錄因子，我們將問題進一步延伸成尋找具有間隔變化的組合式訊息，使問題更有挑戰性。為了展示我們演算法的有效性，我們拿它來測試一系列新的挑戰問題，以及真實的基因家族，並與目前其它具有代表性的特色構形演算法做比較。

關鍵詞：基因調控、基因家族、調控訊號

Abstract

Biologists have determined that the control and regulation of gene expression is primarily determined by relatively short sequences in the region surrounding a gene. These sequences vary in length, position, redundancy, orientation, and bases. Finding these short sequences is a fundamental problem in molecular biology with important applications. Though there exist many different approaches to signal (i.e. short sequence) finding, some new study shows that this problem still leaves plenty of room for improvement. In 2000, Pevzner and Sze proposed the Challenge Problem of motif detection. They reported that most current motif finding algorithms are incapable of detecting the target motifs in their Challenge Problem. In this paper, we show that using an iterative-restart design, our new algorithm can correctly find the target motifs. Furthermore, taking into account the fact that some transcription factors form a dimer or even more complex structures, and transcription process can sometimes involve multiple factors with variable spacers in between, we extend the original problem to an even more challenging one by addressing the issue of combinatorial signals with gaps of variable lengths. To demonstrate the effectiveness of our algorithm, we tested it on a series of the new challenge problem as well as real regulons, and compared it with some current representative motif-finding algorithms.

Keywords: gene regulation, gene family, regulatory signal

二、緣由與目的

許多不同的基因體計劃已經產生了大量生物序列資料。然而，我們的生物知識進步的腳步跟不上資料產生的速度。這種不平衡的情況刺激了新方法與新器材的發展，以便於用在如新基因的註解(annotation)等方面上。其中一個最有發展潛力的新發明是 microarray gene chip，它可以同時直接測量基因體上每個基因 expression level 的改變[1][2]。根據 expression level 的改變，生物學家可以輕易隔離出共調控基因。這不僅提升了 gene expression 實驗的效率，也提供了以整個基因組巨觀的角度來觀察基因的行為。

由測量 gene expression 所隔離出來的共調控基因群聚只能顯示出哪些基因對同一個刺激因素有類似的反應。生物學家更想了解的是何種機制與這個反應有關係。細胞對刺激因素的反應是由轉錄因子的行為所控制。轉錄因子本身是一種特殊的蛋白質，它可以辨認特定的 DNA 序列。轉錄因子會連結到調控區，並與 RNA 聚　合(polymerase)起反應，藉此活化或抑制一組特定基因的表現。有了一個對特定刺激有相同反應的基因家族，我們想解答的問題是尋找他們的調控訊息(也稱為特色構形 motif 或模式 pattern)，也就是轉錄因子的結合區，這些結合區是控制基因所共有的區域。

尋找特色構形的問題可以定義如下：給定一些以特定符號組成序列的樣本(例如由 A, G, C, T 符號組成的 DNA 序列)，在這些序列裡不同的位置上有未知隱含的模式(特色構形)，我們該如何找出這些未知的模式？根據特色構形的表示方式、重要性測量方式與尋找的策略的不同，這個問題發展出了許多不一樣的方法來解答[3-9]。雖然這些演算法實際上在許多不同的領域都被證明出很有效用，但新的報告指出好幾個具有代表性用以尋找特色構形

的演算法，仍然無法找出具有特殊形態的細微特色構形。這在尋找特色構形的挑戰問題中被提出來[10]。為了降低目前方法的做的限制，我們提出新的演算法 MERMAID，它採用矩陣來表示特色構形，它也可處理具有長度變化的間隔。這個方法延伸前人的做法，合併了數種特色構形重要性的測量方法，並改進重覆取樣的技術。我們以原來的挑戰問題與延伸的問題來證明新方法的效果，並與其它主要的特色構形尋找演算法做比較。為了驗證它在實際環境應用上的可行性，我們也用許多已知共同調控特色構形的酵母菌基因家族來測試 MERMAID。

分析基因體上的非編碼區以了解控制機制是個困難的問題。由於對示範基因相當密集的研究，某些調控，包括位置的影響、調控特色構形的多樣性、特色構形的方向性與不同特色構形的組合，雖然值得稱許，但仍不夠廣泛。儘管細微特色構形調控訊息的研究已經許多年了，但仍吸引了許多人的注意，因為它是研究基因體重要的步驟之一。逐漸增加的基因體基因活動知識，加上用來推論特色構形的演算法、聯結特色構形與活動的演算法，讓我們對未知的基因調控部份有更深的了解。

根據 Pevzner 與 Sze，儘管目前已存有許多各式各樣的演算法，這個問題仍然沒有被解決。他們發現許多常用的特色構形尋找演算法無法找到下面定義的挑戰問題：

$S = \{ s_1, \ldots, s_t \}$是有 t 條由 n 個字母組成的序列集合。每條序列包含(l, d)-signal，也就是一個長度為 l，不吻合個數(mismatch)為 d 的訊息。問題是如何找出正確的(l, d)-signal。

在他們的實驗中，他們在 20 條樣本序列裡隱藏了一個(15,4)-signal。為了驗證序列長度的影響，n 設為 100 至 1000。實驗結果顯示，當序列的長度漸增，MEME、CONSENSUS 與 Gibbs sampler 的效能劇烈下降。有兩個原因導致這失敗。第一個原因，演算法可能卡在 local optima。序列長度的增加產生更多的 local optima，因而使問題更加惡化。第二個原因，這些演算法期望在樣本中的訊息本身也會顯露自己。然而在挑戰問題中，樣本並沒有明確的訊息存在，只有 4 個不吻合的變形實例。Pevzner 與 Sze 提出 WINNOWER 和 SP-STAR 來解決挑戰問題，但 WINNOWER 的可應用性被複雜度所侷限，而 SP-STAR 的效能和其它演算法一樣隨著序列長度增加而大量下滑[10]。

三、設計考量

目前大多數的方法是基於 greedy 或是 stochastic hill-climbing 演算法，它們對比重矩陣序列上每個位置都有最佳化[3][4]。這些不但沒有效率，在含有細微訊息的長序列上因為可能有較多相似的的隨機模式同時存在於序列上，也會增加陷在 local optima 的機會。為了避免這個缺點，我們一開始允許每條長度為 l 的子字串都是可能的候選訊息。採納由[5]得來的想法，我們將這個特殊的子字串轉換成機率矩陣。這讓我們得到一組初始的機率矩陣，之後再做反覆的改進。我們使用初始的矩陣來標定那些吻合分數超過某個上限而可能是訊息的位置。最佳化的程序只檢查序列中這些可能的位置，而不是所有的位置。把注意力轉移到被視為可能是特色構形，而且與子字串相同或相似的模式上，我們可以在反覆改進的程序中大量減少搜尋的空間。

然而，當目標訊息十分的細微，例如(15, 4)-signal，若只考慮我們選擇出來可能的訊息位置，這種偏好(bias)可能是有害處的。上面的偏好是基於樣本中的目標訊息有足夠的規律性，因此最終可以最佳化的方式來推導出正確目標訊息。不幸的是，如果訊息本身的規律性無法與其它相似的隨機模式有所區分的話，這個樂觀的假設並不成立。因此，誤把隨機模式當成是真的訊息的機會增高。這個演算法會誤導致變形的模式，而非真正的訊息。

處理細微的訊息時，由於相似隨機模式的影響，stochastic 最佳化並不保證能找出正確的目標訊息。然而，收斂的模式必定與目標訊息很接近，因為隨機模式也必定與目標訊息有些許的相似；否則的話，他們也不會被選擇參與最佳化的程序。假設目標訊息如預期一般是樣本中最保留的模式，而我們使用一個訊息實例做為最佳化的起點。不論最後的模式收斂為何，至少它會比當做起點的子字串更接近目標訊息，即使它跟目標訊息並不完全相同。既然收斂的模式接近目標訊息，一個改進的方式即是再使用這個模式當做起點，重新執行最佳化的程序。我們可以用改進過的模式當起點，反覆地執行最佳化程序，直到不再有改進為止。我們期望這個反覆開始的策略可以成功找出挑戰問題中類似(l, d)-signal 的細微訊息。

Pevzner 與 Sze 提出了 SP-STAR 的延伸來處理有間隔的訊息，但他們只處理了固定間隔的問題。然而在實際的領域，特色構形含有長度不一的間隔，而且起始轉錄需要兩個或多個鄰近的轉錄因子同時連結。因此 Pevzner 與 Sze 提出的挑戰問題很自然地可延伸成尋找組合式的(l, d)-signal。組合式的(l, d)-signal 是由多個(l, d)-signal 所組成，而兩個訊息之間的間隔長度介於特定範圍之內。例如，一個組合訊息(l, d)- X(m, n)-(l, d)-signal 是由兩個(l, d)-signal 所組成，而間隔的長度介於 m 至 n 個鹼基對之間。注意，不同元件(component)的訊息長度與突變數目可能不同。

通常有兩種方法來尋找組合式訊息。第一種方法有兩個步驟。首先第一個步驟是尋找可能的訊息元件。第二個步驟是將訊息元件結合成組合式訊息，並驗證他們的重要性。這個方法只有在元件本身有足夠的重要性，因此在第一步驟可被隔離出來，以便於第二步驟的組合性測式時，才能顯示出它的效果。當訊息元件只有在組合時才能顯示出重要性，前述的方法會低估元件間的互動性而無法找出組合。為了避免這個限制，一個替代的方法是直接尋找組合式訊息。根據上述的設計考量，我們發展出 MERMAID(Matrix-based Enumeration and Ranking of Motifs with gAps by an Iterative-restart

Design)來處理細微的組合式訊息。MERMAID 列舉在給定範圍內所有可能不同間隔長度的子字串組合。它為每個訊息元件建立一個機率矩陣，然後使用反覆開始的程序對矩陣做最佳化。假設間隔的範圍相當小，則 MERMAID 的時間複雜度並不會劇烈增加。

四、結果與討論

這篇論文的其中一個目標，是證明可應用簡單的重覆開始策略，我們的特色構形偵測演算法可以找到細微的訊息，例如(15, 4)-signal。根據定義，我們重現了挑戰問題，並用它來比較我們與其它人的演算法。我們用八個樣本資料來測式 MERMAID，如同 Pevzner 與 Sze 所為，每個樣本包含 20 條 i.i.d.的 1000 個鹼基對序列。設 K 是樣本中已知的訊息位置集合，而 P 是預測出來的位置集合。效能係數定義為 $|K \cap P| / |K \cup P|$。更者，為了顯示出 MERMAID 由於重覆開始策略，與結合數個客觀函數的最佳化程序合力的結果，可幫助找出細微的訊息，我們在樣本中隨機的位置上置入由 MEME 所找出來，與目標訊息的 mismatch 值最小的特色構形。然後我們重新執行 MEME。重複上述過程，並檢查光以重覆開始策略是否能改進 MEME 的效能。我們測試 MEME 的原因是 MERMAID 採用了相同的特色構形列舉方式。因為 MEME 徹底測試了樣本中的每個子字串，我們置入的子字串會在下個回合使用到。我們只置入最接近實際訊息(也就是有最小的 mismatch)的特色構形，以確保樣本的基本分布幾乎沒有改變。雖然我們沒有重寫 MEME 程式，這個約略的模擬仍然可以反映出它的效能。結果指出 MERMAID 表現得比 CONSENSUS、Gibbs sampler 與 MEME(有或沒有重覆開始)要好很多。對於有間隔長度變化的特色構形，我們先測試 20 條長度 1000 鹼基對序列的 (6,1)-X(1,g)-(6,1)-signal，g 值變化由 3 到 9。實驗的結果顯示 MERMAID 的效能相當穩定，直到間隔長度為 9。我們也用同樣的資料測試 CONSENSUS、Gibbs sampler、MEME 與 oligonucleotide analysis。以上每個演算法的效能係數在各種長度都接近零。

尋找生物上有意義的特色構形，其困難在於以下幾點變化(1)特色構形每個位置上的鹼基，(2)特色構形在序列上的位置，(3)序列上特色構形出現的多樣性。除此之外，許多生物上有重要意義的特色構形很短，而且只有與其它特色構形組合時才有重要性，因此更難尋找[11]。儘管許多蛋白質與 DNA 的結合區域只與有限數目的鄰近核甘酸建立連繫，有不少轉錄因子會與兩個或以上由非保留區所分隔的短保留序列連結。

不少用來從功能性相關的生物序列上辨認共同的特色構形方法被提出來。每個方法用來尋找特色構形的模型不同。我們回顧這些方法，並與我們的做比較。選擇這些方法是基於他們發展成熟，且可由網路上免費取得，還有不同想法，如表五所示。雖然這些方法只能尋找連續的核甘酸，有些演算法採用這些方法並發展成可偵測有間隔的特色構形[7][11][12]。MERMAID 不同於目前只能偵測固定間隔特色構形的演算法，它能尋找組合式變化間隔的細微訊息。實驗顯示 MERMAID 不僅在人造領域上可偵測由鄰近元件組成的組合式訊息，也可成功辨認在實際 regulon 中有間隔的已知特色構形。MERMAID 萃取單一變化間隔的特色構形能力可進一步推廣到辨認特色構形的組合。

五、參考文獻

[1] DeRisi, J., Iyer, V. and Brown, P., "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale", *Science*, Vol 278, (1997) 680-696

[2] Wodicak, L., Dong, H., Mittmann, M., Ho, M. and Lockhart, D., "Genome-wide Expression Monitoring in Saccharomyces cerevisiae", *Nature Biotechnology*, Vol 15, (1997) 1359-1367.

[3] Hertz, G., Hartzell III, G. and Stormo, G., "Identification of Consensus Patterns in Unaligned DNA Sequences Known to be Functionally Related", *Computer Applications in Biosciences*, Vol 6, No 2, (1990) 81-92

[4] Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A. and Wootton, J., "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignments", *Science*, Vol 262, (1993) 208-214.

[5] Bailey, T. and Elkan, C., "Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization", *Machine Learning*, 21, (1995) 51-80

[6] Hu, Y., Sandmeyer, S. and Kibler, D., "Detecting Motifs from Sequences", in Proceedings of the 16th International Conference on Machine Learning, (1999) 181-190

[7] Li, M., Ma, B. and Wang, L. "Finding Similar Regions in Mary Strings", in Proceedings of the 31st ACM Annual Symposium on Theory of Computing, (1999), 473-482.

[8] Gelfand, M., Koonin, E. and Mironov, A., "Prediction of Transcription Regulatory Sites in Archaea by a Comparative Genomic Approach", *Nucleic Acids Research*, Vol 28(3), (2000) 695-705.

[9] van Helden, J., Andre, B, and Collado-Vides, J., "Extracting Regulatory Sites from the Upstream Region of Yeast Genes by Computational Analysis of Oligonucleotide Frequencies", *Journal of Molecular Biology, 281*, (1998) 827-842.

[10] Pevzner, P. and Sze, S. "Combinatorial Approaches to Finding Subtle Signals in DNA Sequences", in Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, (2000).

[11] van Helden, J., Rios, A. F. and Collado-Vides, J., "Discovering Regulatory Elements in Non-coding Sequences by Analysis of Spaced Dyads", *Nucleic Acids Research*, Vol 28, (2000) 1808-1818.

[12] Rocke, E. and Tompa, M. "An Algorithm for Finding Novel Gapped Motifs in DNA Sequences", in *RECOMM*-98, (1998) pp. 228-233.