行政院國家科學委員會專題研究計畫　成果報告

利用基因規劃預測核糖核酸二級結構

計畫類別：個別型計畫
計畫編號：NSC91-2213-E-009-099-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立交通大學資訊科學學系

計畫主持人：胡毓志

計畫參與人員：徐英哲，陳兆奕，王美華，林婉嫻

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 92 年 9 月 17 日

行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※

※

※

※　　　　利用基因規劃預測核糖核酸二級結構

※

※

※

※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：個別型計畫

計畫編號：NSC 91－2213－E－009 －099－

執行期間：91 年 8 月 1 日至　92 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　92　年　9　月　12　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 91-2213-E-009 -099-

執行期限：91 年 8 月 1 日至 92 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 林婉嫻 王美華　交通大學資訊科學系

一、中文摘要

對於一群俱有同源關係或功能相似的 RNA 序列，共同組態可能代表 RNA 序列與調控蛋白質的結合區。與 DNA 不同的是，RNA motif 在結構的保留遠比在序列的保留程度來得高，透過對結構 motif 的瞭解，我們可以更進一步認識調控活動。目前已經有許多相關的預測系統及工具，但大多數集中在單一 RNA 的結構預測，而非針對一群 RNA，近來，雖有少數針對 RNA 群集的結構預測，但它們僅僅能找到較簡單的結構。有鍵於此，我們提出以基因規劃為基礎的 RNA motif 預測系統，它能夠辨識較 stem-loop 更複雜的結構。我們以三種不同形式的 RNA 資料驗證其效能。

關鍵詞: 基因調控、基因家族、調控訊號

Abstract

Given a set of homologous or functionally related RNA sequences, the consensus motifs may represent the binding sites of RNA regulatory proteins. Unlike DNA motifs, RNA motifs are more conserved in structures than in sequences. Knowing the structural motifs can help us gain a deeper insight of the regulation activities. There have been various studies of RNA secondary structure prediction, but most of them are not focused on finding motifs from sets of functionally related sequences. Although recent research shows some new approaches to RNA motif finding, they are limited to find relatively simple structures, e.g. stem-loops. Here, we propose a novel genetic programming approach to RNA secondary structure prediction. It is capable of finding more complex structures than stem-loops. To demonstrate the performance of our new approach as well as to keep the consistency

of our comparative study, we first tested it on the same data sets previously used to verify the current prediction systems. Besides, to show the flexibility of our new approach, we also tested it on a data set that contains pseudoknot motifs that most current systems cannot identify. A web-based user interface of the prediction system is set up at http://bioinfo.cis.nctu.edu.tw/service/gprm/.

Keywords: RNA motif，基因規劃

二、緣由與目的

就如同蛋白質的二級結構一樣，核醣核酸二級結構在傳統上被視為是構成三度空間結構的中間步驟。另外，因為核醣核酸序列的演化是受結構控制，所以這些我們感興趣的核醣核酸，其實是在由鹼基配對交互作用所形成的二級結構或是三級結構上具有一致的特徵，而不是在序列上擁有相似的片段，因此當我們在研究核醣核酸的特性時，結構是比序列更為重要的議題。

此外，核醣核酸在生物體內所扮演的角色相當重要，儲存在去氧核醣核酸(DNA)序列上的遺傳訊息，都必須藉由核醣核酸的傳遞才能夠真正發揮功能，產生出需要的蛋白質，如果少了這種聯繫的物質，整個生物體的活動將會停擺。在訊息傳遞的過程中，核醣核酸的二級結構發揮了類似模板(template)的功能，因為它所形成的特定結構只能與特定的物質結合，就好比鑰匙和鎖孔的關係一樣，如此一來就可以避免在複雜的生物機制中出現錯誤。

核醣核酸的二級結構是生物體之所以能正常運作不可或缺的重要因素，過去做核醣核酸二級結構預測(RNA secondary structure prediction)的方法有很多，包括用動態程式規劃(dynamic programming)的方法尋找化學上能量最穩定的結構，或是以排比(alignment)的方式，利用一條已知二級結構核醣核酸序列上的資訊，去預測另外一條結構未知的相關(related)核醣核酸序列，以及用基因演算法(genetic algorithm)的方式尋找二級結構和摺疊路徑(folding pathway)等[3,4,5]。但是上述這些方法都只針對單一條核醣核酸序列提供唯一的最佳二級結構預測結果，或是包含多個次佳的結果。

另外還有一個研究的主題乃是針對某家族的核醣核酸序列，尋找他們共同的二級結構，在這個部分所使用的方法包括用多重序列排比(multiple sequence alignment)的方法找尋共變(covariance)區域，例如 SLASH[6]，CLUSTALW[7]等，或是運用能量計算搭配基因演算法或是其他資料探勘(data mining)的方法來尋找[8, 9]。

本研究的方向是屬於第二部分的主題，也就是希望能從一個家族的核醣核酸序列中，找出其共同的結構，因為這些共同的結構在生物演化上可能是有意義的，他們可能控制一種重要的生物機能，所以在經過長時間的演化之後，這些結構仍然保留至今。

本研究所採用的方式是基因規劃(genetic programming)，並且直接以共同二級結構作為我們演化的目標，這樣的做法到目前為止還沒有人嘗試過。大部分的研究學者多把能量或是序列排比當成輔助工具，之前雖然有人用基因演算法的方式來預測核醣核酸的二級結構，但是他們的方法中，最重要的適應函數(fitness function)仍然是以能量為基準，運用基因演算法的特性來提供一些在能量上屬於較佳結構的解。況且截至目前為止，要以能量為基準去預測包含擬節(pseudo-knot)

的結構，仍然沒任何成效很好的估計方法。因此，本研究嘗試不以能量為適應函數，也不從序列排比的結果來預測結構，而是直接針對序列本身，並以較直觀的評分方式來當成我們的適應函數，希望藉此提供另一種研究的方向。

三、設計考量

如果針對一條陌生的核醣核酸序列要預測它的二級結構，我們可以找到很多可能的結構，而且這個數目隨著序列長度的增加成指數成長。舉例來說：一條包含 200 個鹼基的核醣核酸序列就有超過10^{50}個可能的鹼基對結構。而我們需要做的事情是從錯誤的結構中找出生物學上正確的結構，因此需要一個可以給正確結構最高分數的函數以及可以計算所有可能結構之分數的一個演算法。

本研究設計的目的，是希望能從一組相關的核醣核酸序列中找出他們共同的二級結構，如果我們先把每一條序列的二級結構預測出來，再找他們共同結構的話，這樣的搜尋範圍不僅龐大而且有許多變異，所以本研究直接以共同二級結構為搜尋的目標，這樣做雖然大幅縮小了搜尋範圍，但是對於一組我們不熟悉的序列資料而言，我們要的結果仍然藏身於10^{10}甚至更多的可能答案之中。

關於核醣核酸二級結構預測，本研究根據下面兩個假設：

假設一：同一個家族的核醣核酸序列有共同的二級結構。

同一個家族的核醣核酸序列之所以會被分類在一起，就是因為他們具有類似的表現型，而就目前生物學家的了解，控制核醣核酸表現結果的主要部分就是他們所形成的二級結構。本研究假設在一組被歸類為同一家族的相關核醣核酸序列中，存在某些共同的結構，而這些結構就是決定此一家族核醣核酸共同表現型的原因。

假設二：我們所要尋找的共同結構不會任意出現在隨機產生的序列中。

本研究所要尋找的共同結構應該具有演化上的意義。核醣核酸在演化的過程當中，序列的內容可能會經過多次突變，但是其重要結構仍然被保留下來，所以這些結構在演化的過程中必定扮演很重要的角色。因此我們假設這樣的重要結構應該不是偶然形成，也就是說在我們隨機產生的核醣核酸序列中不應該會經常出現。

另外，還有一點和過去研究很不一樣的地方，那就是本研究將尋找共同核醣核酸二級結構的目標定義為一個藉由監督式學習(supervised learning)來獲得最佳答案的問題，這也就是為什麼本研究會假設目標共同二級結構不會任意出現在隨機產生序列上的原因，而且根據這項假設，我們會產生一組負面背景資料(negative set)當成監督式學習中的錯誤範例。綜合上述這些特性，我們必須從龐大的搜尋空間中找到最佳的，再加上考慮到本研究所定義的結構描述語言之彈性，我們選擇了基因規劃當作模型[1, 2]。

沒有選擇基因演算法的原因是，本研究直接將共同結構描述於母體中的個體，這樣的好處就是在本模型演化結束之後，母體就包含我們要的答案，也就是這組序列資料的共同二級結構，接下來只需要依據這個答案的描述在每一條序列上標出結構出現的位置就大功告成了，如此一來省略了基因演算法中編碼和解碼的繁瑣工作，況且我們還可以看到整個共同二級結構演化的過程，而這個過程也可能透露出一些有用的資訊可供研究。

四、結果與討論

從實驗結果我們歸納出下面幾點特性：

1. 在絕大多數的情形下，微調(refinement)確實可以幫助我們獲得較好的預測結構。
2. 突變率在本模型中對結果的影響比起交換率要來的大。
3. 個體族群的大小在本研究的測試資料中不具有決定性的影響。
4. 對於序列個數較少資料，增加負面背景序列的倍數確實能幫助找到較好的答案。

從上面的結果看來，微調不僅可以幫助我們提昇最後答案的正確率，而且也縮小每一次實驗結果間的差距。在我們測試的資料中除了最後一組 miR 的結果，經過微調後相關係數下降的可能不算小，會出現這樣的情況主要是因為我們微調的部分是在二級結構上面，而且傾向選擇條件較嚴格的描述來避免過多錯誤的正預測，但是這組序列在莖幹上面較為複雜，很有可能因為較嚴格的二級結構而流失了一部份原本已經取到的鹼基對，不過還好差距不算太大，所以我們仍然認為微調是一個很好的運算子。

從 IRE like、soil-borne mosaic virus 和 Phe-tRNA 的實驗結果中可以看到，當我們在固定交換率的情形下，隨著突變率的增加，最後預測結果的相關係數有提昇之現象，至於當固定突變率而改變交換率的實驗中，我們看不出交換率的改變和相關係數有類似的關係，這很有可能失因為本研究在設計突變運算子的時候，將它設計的比較彈性，除了改變結構範圍的內容之外，結構裡莖幹的相對位置也可以改變，相較之下，交換運算子只能選擇和自己相同結構的個體來交換，這可能是交換運算子對整個結果的貢獻似乎不大的原因。

另外一個同樣看不出明顯關聯的是相關係數和個體族群大小間的關係，但是我們不認為增加母體中個體的總數對演化的結果沒有幫助。他們之間關聯不明顯的原因很可能是因為這些我們測試的資料都沒有包含太多的莖幹所致，最大的結構是 Phe-tRNA 這組，擁有四個莖幹，所有可能的結構相對位置有 105 種，所以將個體總數設定成 500 或是 1000 的話或許還可以有不錯的答案，但是當莖幹個數等於五時，單單結構的相對位置就有 945 種，在這種情況下，1000 的個體總數都不夠了，這也是一個將來必須處理的問題。

至於增加負面背景序列總數與相關係數之間的關係，以 soil-borne mosaic virus 這組資料的實驗結果最為明顯，隨著負面背景序列倍數的增加，相關係數的差距從 60%變成 80%，增加了 20 個百分點，其他例如 archaea 16S rRNA、Phe-tRNA 和 miR 的資料中也可以看出這樣的趨勢，至於 IRE like 和 HIVRT 這兩組較簡單的資料，其實幾乎每一次都可以找到全區的最佳解(global optimal)，所以增加負面背景序列的總數不會再有什麼特別的影響。

本研究嘗試利用基因規劃的模型去尋找一組相關核醣核酸序列中共同的二級結構，與過去方法不同的地方是，本研究以序列本身當作材料，直接以共同的二級結構為搜尋最終目標，透過我們定義的結構及莖幹描述語言來幫助共同二級結構預測。

本研究沒有使用複雜的評分方式，只根據兩個合理且容易瞭解的假設來設計適應函數，利用這樣的方法即可以達到不錯的預測結果，又加上為了預測二級結構而提出可以描述任何二級結構的語言以及兩種不同的莖幹描述方法，希望可以讓未來的研究者在這方面的研究議題上多了一種可供選擇的工具。在二級結構預測方面有幾點值得注意的，第一，本研究的預測是沒有任何結構上的限制，過去的研究中常因為描述語言以及能量評估上的侷限，使得必須將重要的擬結結構屏除在外，但是本研究所用的描述語言相當彈性，而且只根據序列本身的資訊當成適應函數評分的標準，所以不會有這樣的問題。

第二，在與過去研究的實驗比較中，IRE like 和 archaea 16s rRNA 這兩組資

料我們都有較好的表現，而 HIVRT 這組資料我們也可以達到 99%的正確率。他們的研究使用了序列排比和能量計算的方式，而本研究採用完全不同於這兩者的方法也得到幾乎相同的結果，這顯示用基因規劃來解二級結構預測的問題確實是可行的。

本研究提出兩種不同的莖幹描述方法，因為它們各有優劣，所以在目前仍然不能捨棄任何一種，然而在未來我們希望能夠整合這兩種描述方法，擷取它們的優點並且消除它們的缺點。如果要達到這個目標，我們認為多一些化學方面的背景知識可能是有益的。

專家的背景知識一直以來都是解決問題的一項利器，對所要解決的問題有更深入的了解，一般來說就越能夠達到事半功倍的效果。以本研究為例，除了了解到核醣核酸在二級結構上要具有較多的特性，而且在核醣核酸的莖幹配對中常會出現非標準華特生-克力克的配對之外，若能在加入其它化學上面的知識或許可以得到更好的結果，或許我們可以在候選莖幹的產生過程中就過濾掉不可能的莖幹，如此一來不僅所需要的搜尋時間會變短，而且因為可能的莖幹個數減少而更有機會找到正確答案。

五、參考文獻

1. Banzhaf, W., Nordin, P., Keller, R.E., Francone, F.D. (1998). *Genetic Programming: An Introduction On the Automatic Evolution of Computer Programs and Its Application.*
2. Koza, J.R., *Genetic Programming: on the programming of computers by means of natural selection.* MIT Press. (1992).
3. Batenburg, F.H.D. van, Gultyaev, A.P. and Pleij, C.W.A. (1995). An APL-programmed Genetic Algorithm for the Prediction of RNA Secondary Structure. *J. Theor. Biol.* 174, 269-280.
4. Chen, J.-H., Le, S.-Y. and Maizel, J.V. (2000). Prediction of common secondary structures of RNAs: a genetic algorithm approach. *Nucleic Acids Res.,* Vol. 28, No. 4, 991-999.
5. Shapiro, B.A., Bengali, D., Kasprzak, W. and Potts, M.J. (2001). RNA Folding Pathway Functional Intermediates: Their Prediction and Analysis. *J. Mol. Biol.* 312, 27-44.
6. Gorodkin, J., Stricklin, S.L. and Stormo, G..D. (2001). Discovering common stem-loop motifs in unaligned RAN sequences. *Nucleic Acids Res.,* Vol. 29, No. 10, 2135-2144.
7. Thompson, J.D., Higgins, D.G.. and Gibson, T.J. (1994). CLUSTALW：improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.,* 22, 4673-4680.
8. Bouthinon, D. and Soldano, H. (1999). A new method to predict the consensus secondary structure of a set of unaligned RNA sequences. *Bioinformatics,* 15 10, 785-798.
9. Gultyaev, A.P., Batenburg, F. H. D. van and Pleij, C.W. A. (1995). The Computer Simulation of RNA Folding Pathways Using a Genetic Algorithm. *J. Mol. Biol.* 250, 37-51.