行政院國家科學委員會專題研究計畫　成果報告

選擇最佳演化樹策略及演化樹合併演算法之研究

計畫類別：個別型計畫
計畫編號：NSC94-2213-E-216-031-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：中華大學生物資訊學系

計畫主持人：吳哲賢

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 95 年 10 月 19 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：94-2213-E-216-031-
執行期限：94 年 8 月 1 日至 95 年 7 月 31 日

主持人：吳哲賢　　中華大學生物資訊學系
E-mail: jswu@chu.edu.tw

一、中文摘要

建構演化樹是分析物種間演化過程，最基本及重要的工具。現今的演化樹建構工具相當豐富，但是不同工具得到的演化樹也不盡相同。本計劃首先選出十個當前大家常用的演化樹建構工具，及十個不同的物種群。利用演化樹 RF 距離演算法，評估出 PHYLIP 及 ClustalW 為最佳的兩個演化樹建構工具。接著設計演化樹合併演算法，合併上述兩個工具所得到的演化樹，並証明所得到的新演化樹，評估 RF(Robinson-Foulds)距離為最佳。最後利用實驗的結果，驗證我們的演化樹建構工具，為當今評估 RF 距離之最佳演化樹建構工具。

關鍵詞：演化樹、RF 距離

Abstract

Construction of phylogenetic tree is the most basic and important tool to analyze the evaluated process among objects. There are many construction tools of phylogenetic trees, each tool does not construct the same phylogenetic tree. We first select 10 popular construction tools of phylogenetic tree and 10 groups of objects, measure their RF (Robinson-Foulds) distances, and decide the first two optimal tools are PHYLIP and ClustalW. And then we derive algorithms to combine the above two phylogenetic trees, and prove the combined tree is the optimal of measuring RF distance. The experimental results finally show that our tool is the optimal.

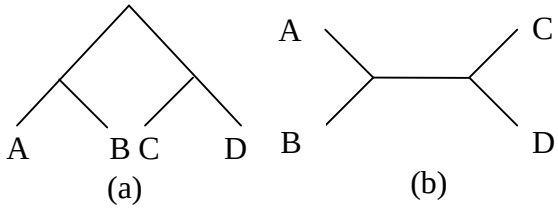
Keywords: Phylogenetic Trees, RF distance

二、簡介

生物之間的演化關係是現今生物學家最關切探討的問題。所謂演化，是指生物在變異、遺傳與自然選擇作用下的演變發展，物種淘汰和物種產生的過程。

生物的演變歷史，如同一棵樹[1-3]。演化樹的建構可以協助了解演化過程及其歷史[4]。研究生物的演化史的方法很多，我們可以經由物種所遺留下來的遺骸或是化石，從中獲取其 DNA 或是蛋白質的序列來建構演化樹。觀察演化樹中物種的相對地位以及位置，可以充分的了解它們的演化過程及親疏遠近。

演化樹的表示方法有兩種，分別是有根樹和無根樹。表示物種或是基因在演化時間上的先後順序可以用有根樹表示，會涉及到物種是不是祖先的問題。而無根樹主要是用來表示物種與物種之間的親疏遠近的距離，不會考慮到祖先的問題。圖一為有根樹(a)與無根樹(b)的表示方式。



圖一 四物種 ABCD 之(a)有根樹與(b)無根樹

建構最佳演化樹是 NP-hard 的問題[5-7]。而目前分析演化樹的方法有很多種，依據輸入資料方式的種類，有特徵法、距離法與序列法。用特徵演化樹(Character Tree)的方法[8]，是將物種之間的特徵值做比較的方法。例如物種是否有翅膀、是否用鰓或是肺呼吸或用腳行走等。這些外表結構的差異

性就是特徵比對所在。

距離演化樹(Distance Tree)的方法，利用了物種與物種之間的距離而建構出的演化樹[9]。利用基因的序列計算出兩兩物種之間的距離，並且用一個 N 乘 N 的矩陣儲存距離，N 代表物種個數。序列演化樹(Sequence Tree)[10]，是由多個物種之基因序列經由比對後，考慮 DNA 序列中核甘酸之排列情況。

目前的演化樹建構工具種類豐富，功能越來越強大。當輸入同一組的資料情況下，不同的演化樹建構工具中，所分析出來的結果，會產生不同的演化樹[11-16]。如何去評估演化樹的好壞，並且從中挑選出最佳的演化樹，我們將介紹 RF 距離(Robinson-Foulds distance)[17]方法。RF 距離可以比較出演化樹之間的差異程度，所以我們將利用這個方法分析演化樹。

接著介紹演化樹合併演算法[18]。在給定一組相同物種情況下，兩種不同的演化樹建構工具，所形成的兩棵演化樹。我們將這兩棵演化樹合併成一棵，並且利用 RF 距離的比較方法分析這三棵演化樹，然後證明合併後的演化樹，為三棵演化樹中的相對最佳化樹。

最後我們提出最佳演化樹建構工具。包含了利用 RF 距離評估最佳演化樹、最佳演化樹建構工具評估演算法、演化樹建構工具之現況分析，與利用 RF 距離評估之最佳演化樹建構工具。

三、演化樹 RF 距離與合併演算法

[RF 距離之計算方法]

演化樹的比較方法有很多種，我們主要探討利用 RF(Robinson-Foulds distance)距離來比較演化樹之間的差異程度。RF 距離的方法，主要是利用演化樹結構裡面的內部邊，切割演化樹的內部邊，產生不同集合之特性，利用這些集合計算出 RF 距離。其演算法如下：

1. 輸入 Tree1 與 Tree2//其外部節點皆為 N。
2. 對 Tree1 與 Tree2 拿掉每個內部的邊，形成兩個集合//內部邊個數為 N-3。
3. 利用 RF 距離公式及所形成之集合計算 RF 距離。

RF 距離公式如下:

$$RF\%(T_1,T_2)=\frac{|split(T_1)|+|split(T_2)|-2|split(T_1)\cap split(T_2)|}{2(N-3)}$$

參數定義如下:

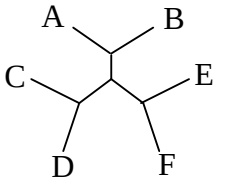
N:物種(外部節點)的個數。

|split(tree)|:樹的分裂樹集合個數。

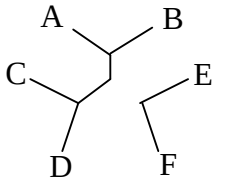
T1:Tree1、T2:Tree2

$0 \le RF \le 1$

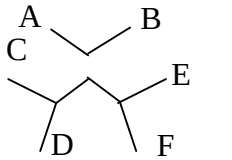
所謂樹的分裂樹(split tree)[10]，即切割樹內部的邊所得到兩個群組(cluster)的集合。計算樹的分裂樹之集合個數，即為樹的內部邊的個數。圖二為六個物種之演化樹，其分裂樹集合個數為圖二(a)、(b)與(c)三種情況。



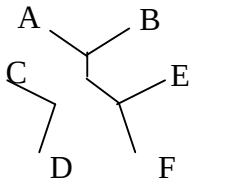
圖二 六個物種之演化樹



圖二(a) 切割第一條內部邊的集合 (EF ABCD)



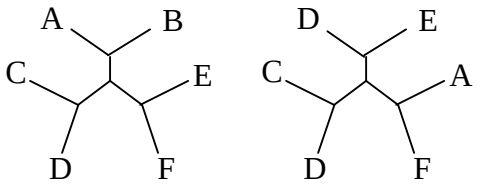
圖二(b) 切割第二條內部邊的集合 (AB CDEF)



圖二(c) 切割第三條內部邊集合 (CD ABEF)

公式中|split(T1)∩split(T2)|是指 tree1 與 tree2 分裂樹集合相同的個數。圖三為兩棵不同演化樹，都有三組分裂樹集合，有一組相同分裂樹集合(AB CDEF)。

Tree1: Tree2:



圖三 兩棵不同演化樹 Tree1 與 Tree2

Tree1 的三組分裂樹集合:

(AB CDEF) (CD ABEF) (EF ABCD)

Tree2 的三組分裂樹集合:

(AB CDEF) (CF ABDE) (DE ABCF)

共同分裂樹集合(共一組):

(AB CDEF)

帶入 RF 距離公式如下:

$$RF\%(T_1,T_2)=\frac{|split(T_1)|+|split(T_2)|-2|split(T_1)\cap split(T_2)|}{2(N-3)}$$

$$=\frac{3+3-2\cdot 1}{2(6-3)}=\frac{4}{6}=0.67$$

RF 值分析：

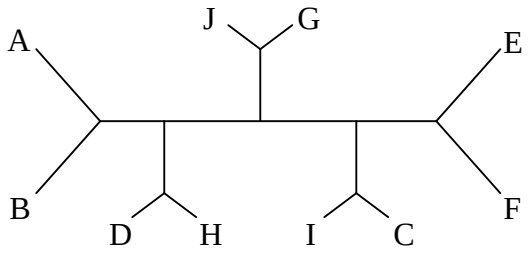
1. 兩棵樹間的 RF 值，從公式中分析得知，介於 0 與 1 之間。
2. 如果 RF 值越小，表示此兩棵數越相似。
3. 當兩棵樹間的 RF 值為 0，表示有相同分裂樹集合，也就是說兩棵樹完全相同。

[演化樹合併演算法]

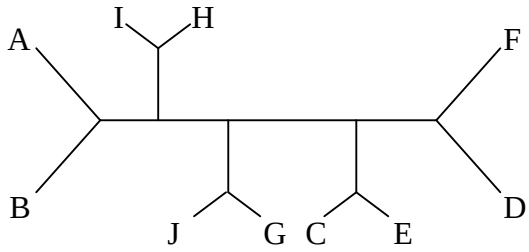
在給定一組相同物種情況下，輸入到兩種不同的演化樹建構工具，形成的兩棵演化樹。將這兩棵演化樹作合併的動作，並且利用 RF 距離的比較方法分析這三棵演化樹，然後證明合併後的演化樹，為三棵演化樹中的相對最佳演化樹。其演算法如下:

1. 求出兩棵演化樹的相同分裂樹集合。
2. 依照相同的分裂樹集合，對樹的外部節點切割。
3. 針對不相同的分裂樹集合中選擇適當集合再作切割。
4. 最後依照切割出來的結果，建構出新的合成演化樹。

舉例說明，Tree1 與 Tree2 為 10 個物種 A 到 J 之不同演化樹，圖四為 Tree1(a)與 Tree2(b)樹形結構。



圖四(a) Tree1 樹形



圖四(b) Tree2 樹形

Tree1 的七組分裂樹集合:

(AB DHJGICEF) (DH ABJGICEF)
(JG ABDHICEF) (IC ABDHJGEF)
(EF ABDHJGIC) (ABDH JGICEF)
(ICEF ABDHJG)

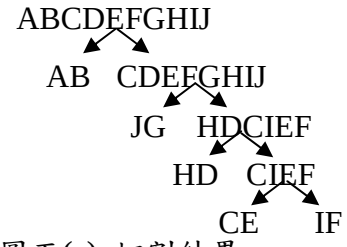
Tree2 的七組分裂樹集合:

(AB IHJGCEFD) (IH ABJGCEFD)
(JG ABIHCEFD) (CE ABIHJGFD)
(FD ABIHJGCE) (ABIH JGCEFD)
(CEFD ABIHJG)

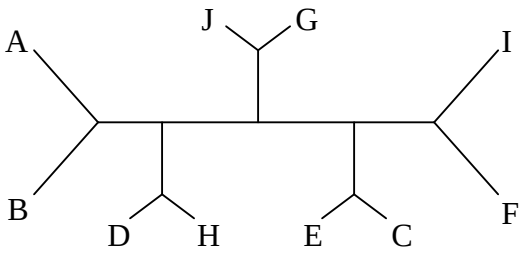
相同的分裂樹集合(共 2 組):

Tree1(AB DHJGICEF)與
Tree2(AB IHJGCEFD)
Tree1(JG ABDHICEF)與
Tree2(JG ABIHCEFD)

利用演化樹合併演算法，依照兩棵樹相同分裂樹集合，對樹的外部節點做切割。外部節點切割後，對剩下的 HDCIEF 做檢查，最後切割成 HD CE IF 情況。圖五(a)為外部節點切割情況，圖五(b)為合併後的新演化樹 Tree3。



圖五(a) 切割結果



圖五(b) 新演化樹 Tree3

接著，將 Tree1、Tree2 與 Tree3 作 RF 距離分析，得知 Tree3 之 TotalRF 為 0.8571，為三棵樹中最小者，故為相對最佳演化樹。其結果如表一。

表一 Tree1、Tree2 與 Tree3 的 RF 分析表

	Tree1	Tree2	Tree3	TotalRF
Tree1	0	0.7143	0.2857	1
Tree2	0.7143	0	0.5714	1.2857
Tree3	0.2857	0.5714	0	0.8571

四、最佳演化樹建構工具

本計劃首先選出十個當前大家常用的演化樹建構工具，及十個不同的物種群。利用演化樹 RF 距離演算法，評估出 PHYLIP 及 ClustalW 為最佳的兩個演化樹建構工具。接著設計演化樹合併演算法，合併上述兩個工具所得到的演化樹，並証明所得到的新演化樹，評估 RF(Robinson-Foulds)距離為最佳。最後利用實驗的結果，驗證我們的演化樹建構工具，為當今評估 RF 距離之最佳演化樹建構工具。

[利用 RF 距離評估最佳演化樹演算法]

首先，我們先介紹利用 RF 距離評估最佳演化樹演算法，其演算法的步驟如下:

1. 計算兩兩演化樹之 RF 距離。
2. 加總每棵演化樹與其他演化樹 RF 距離。
3. 加總距離最少者即為最佳演化樹。

[標準化 RF 距離總合值]

接著，提出標準化 RF 距離總合值，其演算法步驟如下:

1. 將所有演化樹之 RF 距離總合再加總。
2. 個別演化樹之 RF 距離總合除以上述加總值即為標準值。
3. 所有演化樹之標準值總合為 1。

舉例說明，給定四棵相同物種的不同演化樹，Tree1，Tree2，Tree3，及 Tree4。兩兩演化樹 RF 值的計算結果如下表二:

表二: Tree1 到 Tree4 的 RF 分析表

	Tree1	Tree2	Tree3	Tree4
Tree1	0	0.6	0.5	0.4
Tree2	0.6	0	0.2	0.8
Tree3	0.5	0.2	0	0.7
Tree4	0.4	0.8	0.7	0

計算每棵樹的 RF 總值之後，會得到 Tree1 = 1.5，Tree2 = 1.6，Tree3 = 1.4，Tree4 = 1.9，因為 Tree3 的 1.4(0.5 + 0.2 + 0.7)為最小，即是四棵演化樹中的相對最佳演化樹。表三為相對最佳演化樹分析結果。

表三 相對最佳演化樹分析

	Tree1	Tree2	Tree3	Tree4	TotalRF
Tree1	0	0.6	0.5	0.4	1.5
Tree2	0.6	0	0.2	0.8	1.6
Tree3	0.5	0.2	0	0.7	1.4
Tree4	0.4	0.8	0.7	0	1.9

RF 總值之加總為 1.5 + 1.6 + 1.4 + 1.9 = 6.4，把 Tree1 到 Tree4 的個別 RF 總值除以加總值 6.4，標準值: Tree1'= 0.23，Tree2' = 0.25，Tree3' = 0.22，Tree4'= 0.30。所以 Tree3' 之標準值 0.22 為最小，且標準值總合為 1(0.23+0.25+0.22+0.30)。表四為標準化分析結果。

表四 標準化分析

	Tree 1	Tree 2	Tree 3	Tree4	Total RF	標準值
Tree 1	0	0.6	0.5	0.4	1.5	0.23
Tree 2	0.6	0	0.2	0.8	1.6	0.25
Tree 3	0.5	0.2	0	0.7	1.4	0.22
Tree 4	0.4	0.8	0.7	0	1.9	0.30

[最佳演化樹建構工具評估演算法]

在這一節裡，我們接著提出最佳演化樹建構工具評估演算法，其演算法步驟如下:

1. 給予 M 種不同演化樹建構工具以及 N 組不同物種群。
2. 計算個別演化樹建構工具及個別物種群之標準值。
3. 加總每個演化樹建構工具之標準值。
4. 再將每個工具之加總值標準化。
5. 標準化加總值最少者即為最佳演化樹建構工具。

舉例說明，延續表四的例子，假設給予 4 種不同演化樹建構工具以及 3 組不同物種群。第一組 Data1 標準值分別為 0.23、0.25、0.22 以及 0.30，再針對 Data2 至 Data3，算出所有標準值。將每個工具之所有標準值加總起來，再除以 3 即為加總值之標準化。最後，加總值表準化最少者 Tool3 即為最佳演化樹建構工具。表五為加總值標準化後的結果。

表五 加總值標準化

	Data1	Data2	Data3	標準值之加總	加總值標準化
Tool1	0.23	0.24	0.29	0.76	0.2533
Tool2	0.25	0.23	0.25	0.73	0.2433
Tool3	0.22	0.27	0.22	0.71	0.2366
Tool4	0.30	0.26	0.24	0.8	0.2666

[RF 距離評估之最佳演化樹建構工具]

最後提出利用 RF 距離評估之最佳演化樹建構工具，其演算法步驟如下:
1. 將前兩名(Phylip 與 ClustalW)之演化樹利用演化樹合併演算法作合併。
2. 分析新演化樹，其加總值標準化後的值為最少，即為最佳演化樹建構工具。

五、實驗結果

我們將所挑選出的 10 組常用之不同的演化樹建構工具與 10 組不同的 DNA 之物種群，這 10 組測試的物種，都是目前常用工具所公開使用的測試資料。接著評估演化樹工具之現況分析，並且挑選出前兩名演化樹建構工具。

常用的 10 組演化樹建構工具如下:
1. MEGA
2. Spectrum
3. Phylip
4. ClustalW
5. PTP
6. START2
7. Splittree
8. Swaap
9. T-rex
10. Wet

所挑選的 10 組測試物種群如下：
1.AspA(大腸桿菌，物種數為 6 筆)
2.Algae(海藻，物種數為 8 筆)
3.Bacillus cereus(芽胞桿菌，物種 18 筆)
4.Drosophila(果蠅，物種數為 10 筆)
5.Gophers(包 含 talpoides 、 bottae 、 underwoodi，物種數為 14 筆)
6.Hepatitis C virus (C 型肝炎，物種數 8 筆)
7.HIV(愛滋病，物種數為 10 筆)
8.Klebsiella pneumoniae(肺炎，物種 16 筆)
9.Primates(靈長類，物種數為 12 筆)
10.Streptococcus agalactiae(鏈球菌，物種數為 20 筆)

分析果可以發現，Phylip 值為 0.075518 排名第一，ClustalW 為 0.082033 排名第二。表六為 10 種工具分析結果。

表六 10 種工具分析結果

	標準值之加總	加總值標準化
MEGA	0.997987	0.099799
Spectrum	1.039335	0.103934
Phylip	0.755179	0.075518
ClustalW	0.820334	0.082033
PTP	1.208570	0.120857
START2	0.833011	0.083301
Splittree	1.205733	0.120573
T-rex	1.385283	0.138528
Wet	0.825933	0.082593
swaap	0.928634	0.092863

接著利用 RF 距離評估之最佳演化樹建構工具，將表六的分析結果中的前兩名演化樹建構工具(Phylip 與 ClustalW)取出。接著利用演化樹合併演算法作合併動作，將合併出的新演化樹作 RF 距離分析結果。表七為 Phylip、ClustalW 與新演化樹之 RF 分析結果。

表七 Phylip、ClustalW 與新演化樹之 RF 分析結果

	新演化樹	Phylip	ClustalW	RF 加總
新演化樹	0	0.73333	0.80000	1.53333
Phylip	0.73333	0	0.93333	1.66666
ClustalW	0.80000	0.93333	0	1.73333

最後將新演化樹與 10 種軟體作評估演化樹工具之現況分析，並且依照分析結果作排名。表八為新演化樹與 10 種軟體之分析與排名結果。發現新演化樹為全部最佳演化樹建構工具。

表八 新演化樹與 10 種軟體分析排名結果

	標準值之加總	加總值標準化
1. 新演化樹	0.64438	0.06444
2. Phylip	0.70072	0.07007
3. ClustalW	0.74436	0.07444
4. Wet	0.76138	0.07614
5. START2	0.79761	0.07976
6. Swap	0.88450	0.08845
7. MEGA	0.93953	0.09395
8. Spectrum	0.98732	0.09873
9. PTP	1.11991	0.11199
10. Splittree	1.12711	0.11271
11. T-rex	1.29317	0.12932

六、結論

在這篇論文中首先利用 RF 距離演算法，評估出 PHYLIP 及 ClustalW 為目前最佳的兩個演化樹建構工具。接著設計演算法合併上述兩個工具所得到的演化樹，証明所得到的新演化樹 RF 距離為最佳。最後利用實驗，驗證我們的演化樹建構工具，為當今評估 RF 距離之最佳演化樹建構工具。

七、參考文獻

[1] T. Jiang, E. Lawler, and L. Wang, “Aligning sequences via an evolutionary tree: complexity and approximation”, In Proceedings of the Twenty-Sixth Annual ACM Symposium on the Theory of Computing ACM, pp. 760-769,1994.

[2] S. Kannan, and T. Warnow, “Inferring evolutionary history from DNA sequences”, SIAM Journal on Computing, pp. 713-737, 1994.

[3] L. Klotz, and R. Blanken, “A practical method for calculating evolutionary trees from sequence data”, Journal of Theoretical Biology, pp. 261-272, 1981.

[4] W. Day, “Computational complexity of inferring phylogenies by dissimilarity matrices”, Bulletin of Mathematical Biology, Vol. 49, No. 4, pp. 461-467, 1987.

[5] W. Day, and D. Sankoff, “Computational complexity of inferring phylogenies by compatibility”, Systematic Zoology, vol. 35, No. 2, pp. 224-229, 1986.

[6] D. Robinson, and L. Foulds, “Mathematical Biosciences”, vol. 53, pp. 131–147, 1981.

[7] D. Sankoff, “Minimal mutation trees of sequences”, SIAM Journal on Applied Mathematics, vol. 28, pp. 35-42, 1975.

[8] R. Agarwala, and D. Fernandez-Baca, “A polynomial- time algorithm for the perfect phylogeny problem when the number of character states is fixed”, SIAM Journal on Computing, pp. 1216-1224, 1994.

[9] J. Hein, “An optimal algorithm to reconstruct trees from additive distance data”, Bulletin of Mathematical Biology, pp. 597-603, 1989.

[10] S. Altschul, and D. Lipman, “Trees, stars, and multiple biological sequence alignment”, SIAM Journal of Applied Mathematics, vol. 49, pp. 197-209, 1989.

[11] L. Cavalli-Sforza, and A. Edwards, “Phylogenetic analysis: models and estimation procedures”, Evolution, vol. 32, pp. 233 –257, 1967.

[12] L. Joseph, Thorley, and D. Roderic, “RadCon：Phylogenetic tree comparison and consensus”, Bioinformatics, vol. 16, pp. 486-487, 2000.

[13] C. Korostensky, and G. H. Gonnet, “Using traveling salesman problem algorithms for evolutionary tree construction”, Bioinformatics, vol. 16, pp. 619-627, 2000.

[14] A. Meade, D. Corne, M. Pagel, and R. Sibly, “Using Evolutionary Algorithms to Estimate Transition Rates of Discrete Characteristics in Phylogenetic Trees”, Congress of Evolutionary Computation, pp. 1170-1176, IEEE 2001.

[15] K. Ming-Yang, “Tree Contractions and Evolutionary Trees”, SIAM Journal on Computing, pp. 1592-1616, December 1998.

[16] D. Sankoff, “Minimal mutation trees of sequences”, SIAM Journal on Applied Mathematics, vol. 28, pp. 35-42, 1975.

[17] D. Robinson, and L. Foulds, “Mathematical Biosciences”, vol. 53, pp. 131–147, 1981.

[18] 朱廷翰，利用 RF 距離評估演化樹及合併演化樹演算法，中華大學資訊工程研究所，2005.