物種演化距離計算之研究

The Study of Computing Evolutionary distance

孫光天	魏至軒	張定宗	劉紋君	徐禕佑
Koun-Tem Sun	Jr-Shiuan Wei	Ting-Tsung Chang	Wen-Chun Liu	Yi-Yu Hsu
國立臺南大學 數位學習科技系	國立臺南大學 數位學習科技系	國立成功大學 醫學院基礎醫學所	國立成功大學 醫學院基礎醫學所	國立臺南大學 數位學習科技系
ktsun@mail.nutn.edu.tw	*rick70002@yahoo.com.tw*	*ttchang@mail.ncku.edu.tw*	*ttchang@mail.ncku.edu.tw*	*ktsun@mail.nutn.edu.tw*

摘要

物種演化距離之計算，可幫助我們了解不同物種的親緣關係，在生物學上有極重要的地位，應用方向主要包括長期演化的親緣分析(phylogenetic analysis)和親緣樹(phylogenetic tree)的設計，以往在此類長期的演化研究的分析，常會使用不合於實際情形的強假設，使計算方法只適用於某些的情況，如穩定性、非穩定性、同質性與異質性，本研究就是要建立一個能適用於穩定性、非穩定性、同質性與異質性等不同情形皆適用的方法。我們將此演化距離計算公式，命名為兩階段法(Two-Phase method)，最後我們將兩階段法與JC、K2P、K3P、GTR、LogDet等方法做比較，發現本方法在不同情況下，皆有最高正確率，顯示本研究所提之技術，對於物種距離計算為一有效之工具。

關鍵字：演化率、演化公式、親緣分析、親緣樹、突變率

Abstract

We can understand the phylogeny of different species by calculating the evolutionary distance which is important to biology. The application of estimating evolutionary rates methods includes the design of phylogenetic analysis and phylogenetic tree. In the past, long-term evolutionary analyses were almost applied to unrealistic strong assumptions which only conformed to some situations, such as stationary, non-stationary, homogeneity, and heterogeneity. The point in our study is to derive a new method for estimating evolutionary rates which generally adapts to stationary, non-stationary, homogeneity, heterogeneity situations. We give a name to our evolutionary distance calculating method TP(Two-Phase method). Comparing the results with JC、K2P、K3P、GTR、LogDet, TP for estimating evolutionary rates has the remarkable accuracy than others, and the technology of the study is a sufficient tool for evolutionary distance calculating.

Keywords: evolutionary rate, evolutionary equation,phylogenetic analysis, phylogenetic tree, mutation rate

1、前言

以往在判斷物種之間的差異，我們只能透過外觀來觀察物種之間的相似，但由於科技的發達，物種的DNA序列一一被人類解碼出來，原本外觀相似的物種，如海豚和鯊魚，在DNA序列的微觀世界中，卻是完全不相同的，為了解物種之間的演化關係，各種計算演化機率模型一一被提出來。

而過去在物種演化距離計算的研究中，以三大方向為主，第一是採用指數分配作為演化距離推估的機率分配，以JC[1]、K2P[2]、K3P[2]、TN[6]等方法為主，其主要功能在於計算方便，對於早期資訊科技不發達，用這些方法亦可人工計算，但不適合用於序列資料屬於異質性(Heterogeneity)或非穩定性(non-stationary)的

情形，第二是採用General Time Reversible，利用對轉置機率矩陣的特徵值作運算的方式，也較能保留原始的轉置機率，以GTR[3]、GTR+Γ [7]等方法為主，解決了資料非穩定性時所產生的誤差，但其效能在序列資料屬於異質性時，效果不佳，第三個是採用對轉置機率矩陣作行列式的方式，是一個能完整保留原始的轉置機率，且較能適用於序列資料屬於異質性的方法，以LogDet[4]等方法為主，但在實驗結果發現同質性的情形下，LogDet的效果反而不佳，所以本研究的主旨就是要建立一個演化模型，在穩定性、非穩定性、同質性與異質性的情形下，均能精確

2、方法

2.1、兩階段(Two-Phase)法

由於以往的方法都只能在特定的情形下，才會有效果，因此我們改進以往的方法，將其命名為Two-Phase method(二階段計算方法)。

第一階段，首先，第一階段我們將JC法作修訂，於機率矩陣的部份，我們將transversional的突變率放大，在此，由於學者Hoyle與Higgs[8]認為該權重值應定為2，所以我們將這個放大的權重(w_κ)定為2。

首先我們先將transversional的部份放大，如下矩陣：

$$\begin{pmatrix} - & p_{AT}\times w_\kappa & p_{AC}\times w_\kappa & p_{AG} \\ p_{TA}\times w_\kappa & - & p_{TC} & p_{TG}\times w_\kappa \\ p_{CA}\times w_\kappa & p_{CT} & - & p_{CG}\times w_\kappa \\ p_{GA} & p_{GT}\times w_\kappa & p_{GC}\times w_\kappa & - \end{pmatrix} \quad (1)$$

將乘過權種後的突變機率加總：

$$D_{Modify} = p_{AT}\times w_\kappa + p_{AC}\times w_\kappa + p_{AG} + ... + p_{GC}\times w_\kappa \quad (2)$$

，作為修正後的總突變機率，再來透過指數分配的特性，找出單位時間內的改變量，推導出TP(Two-Phase)法第一階段的公式。

$$K_{First\_Phase} = -\frac{3}{4}\ln\left(1-\frac{4}{3}D_{Modify}\right) \quad (3)$$

D_{Modify} ：修正後的總突變機率。

K_{First_Phase} ：TP法第一階段的每個位置平均突變次數，又稱突變率(mutation rate)。

第二階段，由於分析實驗資料的特性，發現在非穩定性的演化情形下，兩序列的核苷酸百分比之相關性會越高，兩序列間的距離就愈低，所在第二階段，我們希望利用兩序列的核苷酸A、T、C、G百分比的外積矩陣與真實突變機率矩陣的差異度，作為修正TP法在異質性的計算誤差。

利用序列x與序列y的A、T、C、G的百分比 $\Pi_x = \begin{bmatrix} \pi_A^x & \pi_T^x & \pi_C^x & \pi_G^x \end{bmatrix}$ 與 $\Pi_y = \begin{bmatrix} \pi_A^y & \pi_T^y & \pi_C^y & \pi_G^y \end{bmatrix}$ 計算出外積矩陣O：

$$O = \Pi_x \times \Pi_y = \begin{bmatrix} \pi_A^x\pi_A^y & \pi_A^x\pi_T^y & \pi_A^x\pi_C^y & \pi_A^x\pi_G^y \\ \pi_T^x\pi_A^y & \pi_T^x\pi_T^y & \pi_T^x\pi_C^y & \pi_T^x\pi_G^y \\ \pi_C^x\pi_A^y & \pi_C^x\pi_T^y & \pi_C^x\pi_C^y & \pi_C^x\pi_G^y \\ \pi_G^x\pi_A^y & \pi_G^x\pi_T^y & \pi_G^x\pi_C^y & \pi_G^x\pi_G^y \end{bmatrix} \quad (4)$$

$$r = \sum_{k\neq l\in A,T,C,G} P_{kl} \Big/ \sum_{i\neq j\in A,T,C,G} \pi_i^x\pi_j^y \quad (5)$$

，接下來將O乘上一個O與P的突變率比值r，再將真實的機率矩陣P與上述的乘上比值的外積矩相減，並取絕對值，將其矩陣內不包含對角線的值加總，作

$$\left[\left|P - O_{Modify}\right|\right] = \begin{bmatrix} - & |\pi_A^x\pi_T^y r - P_{AT}| & |\pi_A^x\pi_C^y r - P_{AC}| & |\pi_A^x\pi_G^y r - P_{AG}| \\ |\pi_T^x\pi_A^y r - P_{TA}| & - & |\pi_T^x\pi_C^y r - P_{TC}| & |\pi_T^x\pi_G^y r - P_{TG}| \\ |\pi_C^x\pi_A^y r - P_{CA}| & |\pi_C^x\pi_T^y r - P_{CT}| & - & |\pi_C^x\pi_G^y r - P_{CG}| \\ |\pi_G^x\pi_A^y r - P_{GA}| & |\pi_G^x\pi_T^y r - P_{GT}| & |\pi_G^x\pi_C^y r - P_{GC}| & - \end{bmatrix} \quad (6)$$

$$|\pi_A^x\pi_T^y r - P_{AT}| + |\pi_A^x\pi_C^y r - P_{AC}| + ... + |\pi_G^x\pi_C^y r - P_{GC}| = K_{Senond\_Phase} \quad (7)$$

為第二階段的演化量，本研究曾研究過單獨使用第二階段的演化量作為物種演化的依據，發現第二階段演化量的表現在異質性時，有很好的結果，所以可以補足第一階段異質性時不佳的表現。

$$K = K_{First\_Phase} \times w_{First\_Phase} + K_{Second\_Phase} \quad (8)$$

最後再將兩個階段的演化量分別乘上一個$0 < w < 1$的定值(本研究採用基因演算法去找出w值的最佳解)並相加即可。

2.2、RF-distance

由於我們是要計算不同機率模型所推估出來的樹和真實的物種演化樹的差異性，所以選用RF-distance做為評估演化樹優劣的方法，RF-distance主要是利用切割樹邊(vertex)，將樹的葉節點(leaf)分為不同群組的特性，來計算兩棵樹間差異的程度。

RF-distance[5]將兩棵樹邊(vertex)分為Good edge(好邊)以及Bad edge(壞邊)，其中Good edge代表切除兩棵樹邊(vertex)後會將兩棵樹分割出兩組一樣的群組。

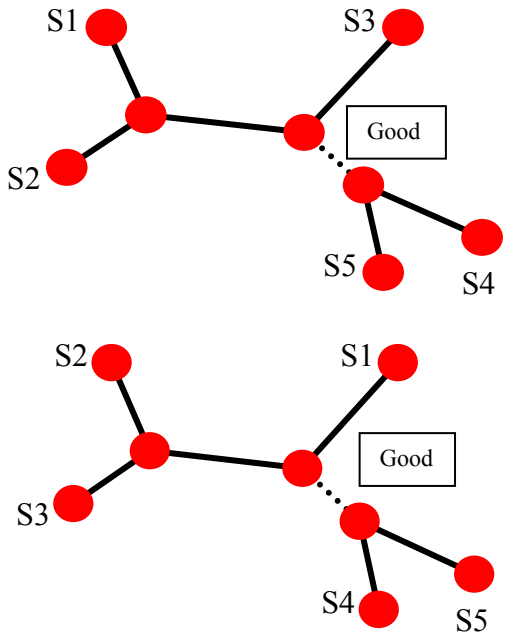

圖 1 Good edge示意圖

圖 1所示，虛線的內邊將兩顆樹各分為兩群子集合，左樹分為(S1,S2,S3)(S4,S5)兩個子集合，而右樹分為(S2,S3,S1)(S5,S4)兩個子集合，而兩顆樹的兩個子集合內的元素經過排序之後相同，所以這個內邊是一個Good edge，換句話說，這個內邊所切出來的分裂樹集合是相同的。

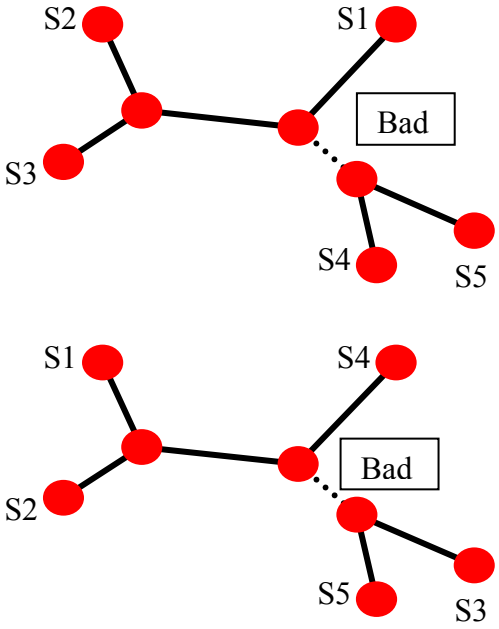

圖 2 Bad edge示意圖

而Bad edge則代表切除邊後會將兩棵樹分割出兩個不一樣的群組，如圖 2所示，虛線的內邊將兩顆樹各分為兩群子集合，左樹分為(S1,S2,S4)(S3,S5)兩個子集合，而右樹分為(S2,S3,S1)(S5,S4)兩個子集合，而兩顆樹的兩個子集合內的元素經過排序之後不相同，所以這個內邊是一個Bad edge，換句話說，這個內邊所切出來的分裂樹集合是不相同的。

最後將上述所統計的數據，透過RF-distance的演算法，即可求得樹與樹之間的差異度。

RF-Distance演算法(Algorithm)：

(1) 讀入2 個*Tree*。

(2) For (i=1 ; i<=n-3 ; i++)

//n為葉點個數 ;n-3為內部邊的總數

Do (分別對$Tree_1$ 與$Tree_2$ 拿掉第i個內部邊，各別形成兩個子集合)。

(3) 對兩顆樹的兩個子集合分別做排序。

(4) 比對$Tree_1$ 與$Tree_2$ 集合是否相等，計算出兩棵樹相同的分裂樹集合個數。

(5) 帶入下列公式求出RF 值

$$RF\%(Tree_1, Tree_2) = \frac{|split(Tree_1)| + |split(Tree_2)| - 2|split(Tree_1) \cap split(Tree_2)|}{2(N-3)} \quad (9)$$

N所指的是物種的個數，而$split(\ )$指該樹的分裂樹集合數，$|split(Tree_1) \cap split(Tree_2)|$指兩分裂樹集合相同個數。

在評估演化距離計算時，我們就選用這個方法來評估其優劣。

3、實驗與測試

在計算演化所使用的資料，根據Lockhart[4]等學者所做的研究，利用行光合作用的物種(photosynthetic)(8種)、動物(7種)和蜜峰(6種)三類的樣本去判斷距離公式的好壞，本研究去NCBI資料庫下載這三類共20種(其中動物的蠑螈(salamander)序列不完整，不予以討論)物種序列，作為計算突變率估計的資料來源。

演化距離計算的流程一共分為四個步驟，如圖 3所示：

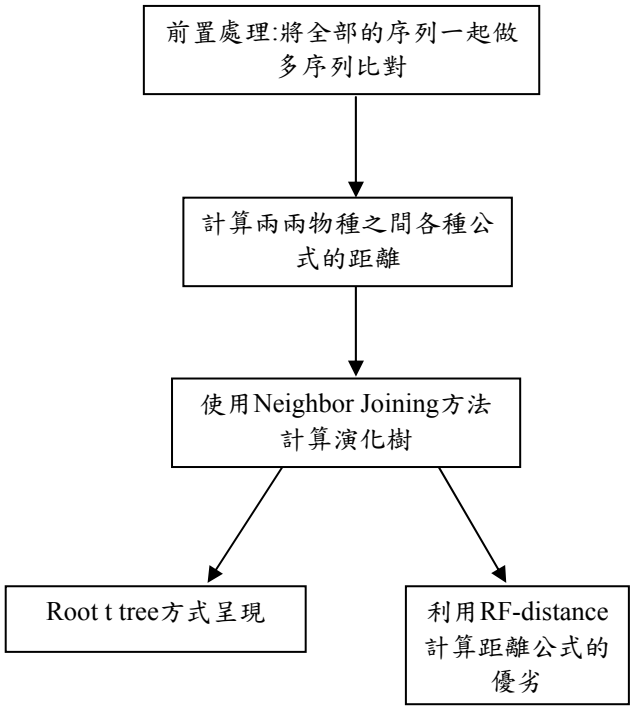

圖 3 流程圖

第一步：資料前置處理，將所需序列由NCBI資料庫下載，透過Multiple Sequence Alignment將序列做序列比對的處理，並將比對後產生gap("-")的位置從所有的序列中去除。

第二步：計算距離矩陣，透過文獻所提及的方法和本研究所提出的Two-Phase方法，去計算每個方法的距離矩陣。

第三步：建立演化樹，透過上述方法所求得的距離矩陣，利用Neighbor-Joining[9]法求出該方法的演化樹。

第四步：計算建立出來的演化樹和真實演化樹的差距，利用RF-distance[5]計算兩兩演化樹的距離。

4、結果與討論

在實驗的部份，我們利用行光合作用的物種(photosynthetic)(8種)、動物(7種)和蜜峰(6種)三類的樣本去判斷距離公式的好壞，我們將序列透過刪除位點相同核苷酸的部份，建立每個物種A、C、T、G百分比不相同的情形，又稱之為物種間的異質性。

而實驗的部份會包括六的部份，對應到表 1中的(1)~(6)：光合作用物種使用序列全長(1)、光合作用物種使用序列全長去除相同的部份(2)、動物使用序列全長(3)、動物使用序列全長去除相同的部份(4)、蜜蜂使用序列全長(5)、蜜蜂使用序列全長去除相同的部份(6)，(1)(3)(5)的物種關係為同質性，(2)(4)(6)的物種關係為異質性，(1)(2)(3)(4)的物種關係為穩定性，(5)(6)的物種關係為非穩性。

表 1 各距離計算方法之RF-distance錯誤百分比表

	(1)	(2)	(3)	(4)	(5)	(6)
TP	0.0%	0.0%	33.0%	33.0%	0.0%	0.0%
JC	0.0%	0.0%	67.0%	33.0%	20.0%	0.0%
TN	0.0%	0.0%	67.0%	67.0%	0.0%	20.0%
K2P	33.0%	33.0%	67.0%	67.0%	20.0%	20.0%
K3P	0.0%	0.0%	67.0%	33.0%	20.0%	20.0%
LogDet	0.0%	0.0%	33.0%	67.0%	0.0%	20.0%
ModifyLogDet	0.0%	0.0%	67.0%	67.0%	20.0%	20.0%
paralinear	0.0%	0.0%	67.0%	67.0%	0.0%	20.0%
Ta/Ts	0.0%	0.0%	67.0%	67.0%	20.0%	0.0%

GTR	0.0%	0.0%	67.0%	33.0%	20.0%	20.0%
GTR+G	0.0%	0.0%	67.0%	67.0%	40.0%	20.0%

樹狀結構的建立是用neighbor-joining的方法建立，並採用Tree-View做為樹狀輸出的應用程式，由表1得知，LogDet方法在比率異質性或非穩定性(non-stationary)演化的情形下，效果較佳，反之JC法在比率同質性或穩定性(stationary)演化的情形下，效果則較佳。

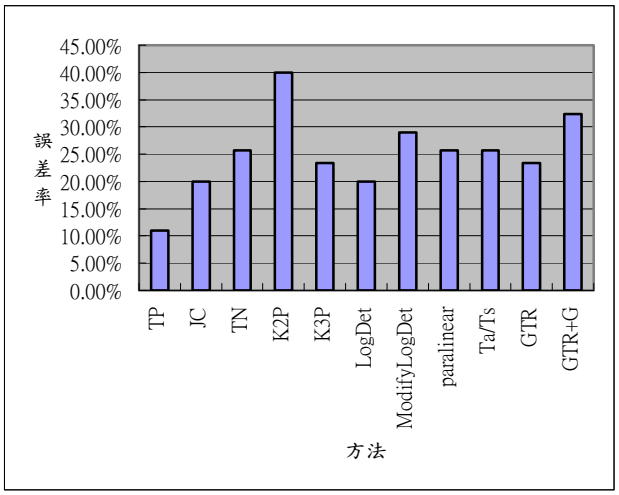

圖 4各物種距離計算方法誤差百分率圖

表 2各距離計算方法之特性表

方向	方法	特性
運用指數分配的累積機率函數	JC,K2P,K3P,TN等	於物種間的演化為穩定性與同質性時較佳
(普遍時間逆轉法) General Time Reversible	GTR,GTR+Gamma等	於物種間的演化為同質性時較佳，於非、穩定性時皆佳
採用對轉置機率矩陣作行列式的方式	LogDet(LD), ModifyLD,paralinear等	於物種間的演化為非穩定性或異質性時較佳

在此物種演化距離計算的研究中，經歸察其推導特性，大約可分三大方向，第一是採用指數分配作為演化距離推估的機率分配，以JC、K2P、K3P、TN等方法為主，其主要功能在於計算方便，對於早期資訊科技不發達，用這些方法亦可人工計算，但不適合用於序列資料屬於異質性(Heterogeneity)或非穩定性(non-stationary)的情形，第二是採用普遍時間逆轉方法(GTR,General Time Reversible)，利用對轉置機率矩陣的特徵值作運算的方式，也較能保留原始的轉置機率，以GTR)、GTR+Γ等方法為主，解決了資料非穩定性時所產生的誤差，但其效能在序列資料屬於異質性時，效果不佳，第三個是採用對轉置機率矩陣作行列式的方式，是一個能完整保留原始的轉置機率，且較能適用於序列資料屬於異質性的方法，以LogDet等方法為主，但在實驗結果發現同質性的情形下，LogDet的效果反而不佳，主要的受到同結構與不同結構轉換之間的差異，所以本研究提出了一個同質性、異質性、穩定性與非穩定性的情形下，都能使用的方法-TP(Two Phase)方法

本研究的第一個部份，將修正後的物種演化距離計算公式-兩階段法(Two-Phase)與JC、K2P、K3P、GTR、LogDet等方法做比較，使用行光合作用的物種(photosynthetic)(8種)、動物(7種)和蜜峰(6種)三類的樣本去建構的不同計算方法的演化樹再與真實物種演化樹作比較，利用RF-distance去計算其差異，結果證明本研究所提之修正公式，在真實物種序列NCBI資料庫之分析，不論於穩定性、非穩定性、同質性、異質性的演化，結果都優於其它方法，可成為生物資訊在物種親緣分析上一有用之技術。

參考文獻

[1] T. H. Jukes and C. R. Cantor, "Evolution of protein molecules," *Mammalian Protein Metabolism*, vol. 3, pp. 21–132, 1969.

[2] M. Kimura, "Estimation of evolutionary distances between homologous nucleotide sequences," *Proc. Natl. Acad. Sci.*, vol. 78, pp. 454-458, 1981.

[3] C. Lanave, G. Preparata, C. Sacone, and G. Serio, "A new method for calculating evolutionary substitution rates," *Journal of Molecular Evolution*, vol. 20, pp. 86-93, 1984.

[4] P. J. Lockhart, M. A. Steel, M. D. Hendy, and D. Penny, "Recovering Evolutionary Trees under a More Realistic Model of Sequence Evolution," *Mol. Biol. Evol.*, vol. 11, pp. 605-612, 1994.

[5] D. Robinson and L. Foulds, *Mathematical*

Biosciences, vol. 55, 1981.

[6] F. Tajima and M. Nei, "Estimation of Evolutionary Distance between Nucleotide Sequences," *Mol. Bid. Evol.*, vol. 1, pp. 269-285, 1984.

[7] P. J. Waddell and M. A. Steel, "General Time-Reversible Distances with Unequal Rates across Sites: Mixing G and Inverse Gaussian Distributions with Invariant Sites," *MOLECULAR PHYLOGENETICS AND EVOLUTION*, vol. 8, pp. 398-414, 1997.

[8] D. Hoyle and P. Higgs, "Factors Affecting the Errors in the Estimation of Evolutionary Distances Between Sequences " *Mol. Biol. Evol.*, vol. 20, pp. 1-9, 2003.

[9] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Mol. Biol. Evol*, vol. 4, pp. 406-425, 1987.