國 立 交 通 大 學

生物資訊研究所

碩 士 論 文

利用加權斷點距離建構原核生物的演化樹

Reconstructing Phylogenetic Trees of Prokaryotes Based on Weighted Breakpoint Distance

研究生：楊忠翰

指導教授：盧錦隆　博士

中華民國　九十八　年　六　月

利用加權斷點距離建構原核生物的演化樹

Reconstructing Phylogenetic Trees of Prokaryotes Based on Weighted Breakpoint Distance

研究生：楊忠翰　　Student：Chung-Han Yang

指導教授：盧錦隆　博士　　Advisor：Dr. Chin Lung Lu

國　立　交　通　大　學

生　物　資　訊　研　究　所

碩　士　論　文

A Thesis Submitted to Institute of Bioinformatics

College of Biological Science and Technology

National Chiao Tung University in partial Fulfillment of the Requirements

for the Degree of Master in

Biological Science and Technology

June 2009

Hsinchu, Taiwan, Republic of China

中華民國九十八年六月

利用加權斷點距離建構原核生物的演化樹

學生：楊忠翰　　指導教授：盧錦隆　博士

國立交通大學生物科技系生物資訊所碩士班

摘要

隨著DNA定序技術的發展，越來越多原核生物物種的完整基因體序列變得更加容易取得。因此，這給予我們一個機會得以藉由比較原核物種基因體之間的基因次序來推測出物種之間基因體規模的演化樹。在過去的研究中，一些利用基因次序的方法像是斷點距離可以用來建構出物種之間的演化關係。當一基因體其一組鄰近基因對的基因次序與在另一基因體上的直向同源基因對其基因次序不一致時，這被認為該鄰近基因對發生一次斷點，兩基因體之間的斷點總數量則為基因體之間的斷點距離。在這傳統的斷點距離中，假設所有在基因體上斷點的發生機率皆視為相同，然而已有文獻指出鄰近基因對可以被分為重組速率快或是重組速率慢的基因對。舉個例子來說，屬於同一個操作組的基因對會比屬於不同操作組的基因對更具有保留性。通常重組速率慢的基因對其彼此之

間的距離較近，反之重組速率快的基因對其彼此之間的距離較遠。根據以上所描述的特性，在這份研究中我們只考慮位於同股的鄰近基因對，並根據斷點是發生在重組速率快或是重組速率慢的基因對將斷點區分為長距離的斷點或是短距離的斷點這兩種類型。由於不同類型的斷點，其發生的機率不一樣，根據這樣的特性我們也定義出一個加權斷點距離並用此方法來衡量兩個原核生物基因體之間的演化距離。另外，我們發展出一個網站伺服器的工具稱之為wBPtree，其可利用原核生物整個基因體之間的重疊基因距離建構出原核生物的演化樹。除此之外，我們也利用一些蛋白細菌的基因體來測試wBPtree在建構演化樹的品質。相較於傳統的斷點距離所建構出的演化樹，我們wBPtree所建構出來的演化樹與參考樹(Eugeni Belda *et al.* 所屬研究團隊利用串接多個蛋白質序列所建構出來的演化樹)是相當一致的。這些結果已說明了我們的wBPtree可以做為一個有用的工具來建構出更準確與更穩定的原核生物基因體樹。

Reconstructing Phylogenetic Trees of Prokaryotes Based on Weighted Breakpoint Distance

Student: Chung-Han Yang Advisor: Dr. Chin Lung Lu

Institute of Bioinformatics
Department of Biological Science and Technology
National Chiao Tung University

ABSTRACT

As more and more complete genomes of prokaryotes are available, it provides us with an opportunity to reconstruct their genome trees based on a genome-scale phylogenetic inference by comparing gene orders between prokaryotic genomes. In the previous studies, some methods based on gene order, such as breakpoint distance, could be useful for reconstruction of the evolutionary relationships of species. It is considered that a breakpoint occurs when the gene order of an adjacent gene pair in a genome is different than that of its orthologous gene pair in another genome. The total number of breakpoints between two genomes is the breakpoint distance for these two genomes. In this original breakpoint distance, it is assumed that all the

breakpoints on a genome have the same probability to occur. However, it has been reported in the literature that adjacent gene pairs can be divided into two classes of fast- and slow-rearranging pairs. For example, a gene pair within an operon is more conservative than a gene pair whose genes are from different operons. Usually, the distance between the genes in a slow-rearranging pair is short and the distance between the genes in a fast-rearranging pair is long. Based on the property described above, we consider only about those adjacent gene pairs that are on the same strand in this study and further divide their breakpoints into two types that are short-distance breakpoints and long-distance breakpoints. Because the occurrence probabilities of short-distance breakpoints and long-distance breakpoints are different, we define a weighted breakpoint distance by assigning different weights to short- and long- distance breakpoints and use it to measure the evolutionary distance between two prokaryotic genomes. In addition, we have implemented a web-based tool, called wBPtree, for constructing the genome trees of prokaryotes based on weighted breakpoint distance between prokaryotic complete genomes. We have also tested our wBPtree on several Proteobacteria complete genomes to assess its quality of genome tree reconstruction. Compared with the phylogenetic trees produced by original breakpoint distance, the genome trees constructed by our

wBPtree are quite consistent with the reference trees that were reconstructed based on concatenation of multiple proteins. All these results have suggested that our wBPtree can serve as a useful tool for constructing more precise and robust genome trees for prokaryotic genomes.

Acknowledgement

光陰似箭，歲月如梭，經過兩年碩士班的生涯，終於得以畢業並邁向人生的下一個旅程，首先我要感謝親愛的父母以及兩位姐姐，由於家人默默的給予支持與鼓勵，並常常開導我，讓我能全心全力的做研究。

我要感謝智先在這兩年時間中給予我的幫助，時常與你討論進度讓我在研究方面事半功倍，我也絕不會忘記你所表現出的男子漢的態度以及運動的熱誠。感謝彥菱學姐兩年來的照顧，在研究有問題的時候都能很有耐心的為我解答，並且常常為大家策劃活動，讓實驗室總是充滿了溫馨。感謝演富學長以及禮瑋學長在碩一時帶給我的歡笑以及在研究方面的帶領與幫助。感謝明原學長在我寫程式以及研究上給予我的協助，讓我得以加速腳步進行研究，感謝慶恩，謝謝你時常與我一起熬夜挑戰如何逃離危險的環境，你那出眾的口才我也會時常銘記在心。感謝志偉這兩年來為大家管理實驗室的設備，時常給你添麻煩真是辛苦你了。感謝晟宸告訴我團隊合作的重要性，以及各種種族的特性。感謝芸蓁讓我重新體認跑跑車是多麼輕鬆溫馨但也緊張的休閒娛樂。感謝昆澤帶給大家對於蛋白結構方面的知識以及如何在緊張的生涯中放鬆自己。

我要特別感謝我的指導教授盧錦隆老師，由於您的督導，讓我得以突破許多研究上的難關，也讓我學習到正確的研究態度以及面對問題時該如分析並解決問題。

最後感謝實驗室所有同仁帶給我的歡樂時光以及吃不完的蛋糕之旅。

Contents

List of Figures

Chapter 1

Introduction

隨著定序技術的發展，越來越多物種 (Species)的全基因體 (Whole Genomes)被定序出來。生物學家得以利用這些龐大的資料來推論並建構出不同物種之間的演化關係。不同於利用多個物種之間共有的單一基因或是多個基因的核苷酸/胺基酸 (Nucleotide/Amino Acid)序列的比較並建構出多個物種之間的演化關係，基因體重組 (Genome Rearrangement)是藉由觀察基因體上基因的次序以及其方向性，相較於另一個物種基因體上直向同源基因 (Orthologous Genes)的次序以及方向性的差異來建構出不同物種之間的演化距離。前者只能觀察點突變 (Point Mutations)，例如核苷酸/胺基酸的取代 (Substitutions)、插入 (Insertions)、刪除 (Deletions)等；而基因體重組則是以基因 (Gene)為單位來觀察染色體上不同物種間的直向同源基因次序的改變。到目前為

止，利用基因體重組來分析全基因體上基因次序改變的研究越來越多，這些研究強力指出不管在原核生物 (Prokaryotes) 或是真核生物 (Eukaryotes)上，利用基因體重組來研究討論並推斷物種之間的演化史是非常具有可靠性的 [1,2]。

在生物學上，以下幾種事件會影響到基因在原核生物基因體上次序的改變。首先是倒位 (Inversion)跟易位 (Translocation)，當在比較物種關係很相近的基因體時，會偵測到許多倒位跟易位的事件 [3]. 倒位為染色體上某一段序列斷裂後，水平旋轉又黏合回去，使原本的基因次序發生改變 [4]。易位則是染色體某一段移到不同染色體的位置上 [5]。其次是刪除 (Deletion)，基因也許會因為單一的事件，或是在一個漸進的過程中被移除掉，這導致於在比較兩個物種的基因體時會產生缺口 (gap) [6,7]。第三個事件則是水平基因轉移 (Horizontal Gene Transfer, 簡稱為 HGT)，HGT 在原核生物的演化上扮演了一個非常重要的角色 [8,9,10,11]，HGT會插入一段外來基因到原本的染色體內，使原本的基因體多出一段基因。複製 (Duplication)則為在染色體複製時所發生的突變，會導致原本的基因體部分複製一段並黏合至基因體內。

評估兩個基因體之間的基因體重組距離可以藉由觀察基因對 (Gene Pairs)的方式來分析。藉著研究兩物種之間直向同源基因對的突變次數來得到物種彼此的斷點距離 (Breakpoint Distance) [12,13]。當在比較兩個基因體時，若是其中一個基因體中的一組鄰近基因對(Adjacent Gene Pairs)，其在另一個基因體的直向同源基因對的基因次序發生改變，則我們稱之為發生一次斷點 (Breakpoint)。統計兩個基因體之間所發生的斷點次數就是兩個基因體彼此的斷點距離。利用斷點距離則可得知兩個基因體彼此的演化距離。舉個例子來說，現在有兩個基因體G跟H，首先將G上的直向同源基因當作是參考的次序，並且將基因位在正股或是反股上給予標記，＋號代表是位於正股，－號則是位在反股上，之後跟H上直向同源基因的次序作比較。假設G上的次序為G=(-1,-2,3,4)，而H=(-4,-3,-2,1)則我們可以得知H所發生的斷點為(-3,-2)跟(-2,1)，而(-4,-3)由於是(3,4)的倒位，因此沒有產生斷點。同理可以得知G所發生的斷點次數與H相同，所以G跟H之間有兩個斷點。

然而傳統的斷點距離對於每組基因對發生斷點的機率皆視為相同，而忽略該基因對在演化史中扮演著重組速率快或是慢的角色從而影響到發生斷點的機率。對於鄰近基因對在演化過程中重組速率快慢的問題，

可以由Rocha *et al.* 所提出的模組 (Model)來討論與研究 [14]。此模組為探討基因次序保留性(Conservation)與演化時間彼此之間關連性，該模組將基因對區分為重組速率慢的基因對，也就是隨著時間演化比較不容易分開的基因對，以及重組速率快且隨著時間演化容易分開的基因對。此演化模組相較於以往所提出的模組更能反映出隨著演化時間長短基因次序保留性的改變。在生物體內，隨著基因對之間距離的不同，以及基因對是否受到同一個轉錄單元 (Transcription Unit)的調控都會影響到基因對的重組速率。於生物的實際例子可以用操作組 (Operon)來說明，操作組內的表現基因之間距離大部份都很短並位於同股上，而且會受到同一組調控因子所管理。同一個操作組內表現基因對的保留性也會比不在同一個操作組的保留性高 [15,16]

根據上述前人的研究，我們可以假設在同股鄰近基因對上，通常重組速率慢的基因對其彼此之間距離較近，而重組速率快的基因對其彼此之間距離較遠。根據這樣的假設，我們改進了傳統斷點的缺點，在這份研究中，我們只討論位於同股的鄰近基因對，並且將斷點區分為發生在重組速率慢的基因對以及發生在重組速率快的基因對，其分別定義為近距離斷點 (Short-distance Breakpoints) 以及遠距離斷點 (Long-distance

Breakpoints)由於兩種斷點在演化所扮演的重要性不一樣，在計算斷點距離上我們也會給予不同的加權比重。以上所計算出來兩物種之間的改良斷點距離我們稱之為加權斷點距離 (Weighted Breakpoint Distance)。

除此之外，我們也發展出一個網路伺服器工具 wBPtree (http://bioalgorithm.life.nctu.edu.tw/wBPtree)，可以利用原核生物物種全基因體的訊息來計算出物種兩兩之間的加權斷點距離，進而建構出原核物種之間的演化樹。為了測試我們所建構出的演化樹，我們利用一些原核生物全基因體的資料來建構出我們的演化樹，並跟利用傳統斷點所建構出的樹形之間作比較。這些原核生物的參考樹形為Eugeni Belda *et al* 藉由連結十個直系同源蛋白質序列所建構出來的樹形[1]。而後比較的結果也證實了對於原核生物來說，我們所提出的加權斷點距離比傳統的非加權斷點距離更能正確且穩定 (Robust)地來建構出原核生物之間的演化樹。

Chapter 2

Preliminaries

在這個章節，我們將會介紹直向同源基因 (Orthologous Genes)、水平基因轉移 (Horizontal Gene Transfer)和斷點 (Breakpoint)的基本概念。另外在本章節也會介紹兩種預測直向同源基因的方法：Bidirectional Best Hit (BBH)以及 INPARANOID。

2.1 Orthologous Genes and Horizontal Gene Transfer

基本上，當不同物種上的基因是由這些物種最後的共同祖先內的一個單一基因所分化而成，則這些基因彼此為直向同源基因。相對的，旁系同源基因 (Paralogous genes)為同一個基因體內的基因經由複製(Duplication)所產生的基因。通常在不同物種上的直向同源基因都會擁

有相同的功能。然而旁系同源基因通常會各自有不同的生物功能。

Figure 2.1 說明直向同源基因和旁系同源基因的演化過程，以及彼此之間關連性。(i) 最初基因 A 位於物種 w。(ii) 基因 A 經由複製作用，使同一個染色體內有兩個基因 A。(iii) 之後兩個基因 A 經由演化後，分化為兩個相關的基因 A1 以及 A2。因此這兩個基因彼此為旁系同源基因。(iv) 物種 x 和 y 分別由物種 w 演化而來，換句話說，經過物種形成事件(Speciation)後，w 為 x 以及 y 的最後共同祖先(Last Common Ancestor)。而基因 A1 的後裔則分別為 A1x 以及 A1y，而基因 A2 的後裔則分別為 A2x 以及 A2y。因此，基因 A1x 以及 A1y 互為直向同源基因。而 A2x 以及 A2y 彼此也是直向同源基因。

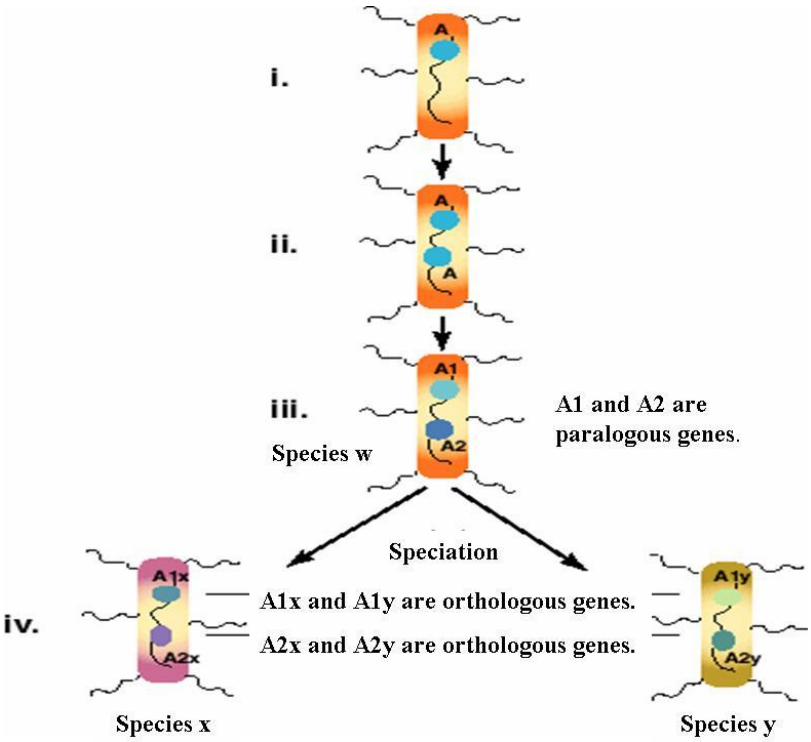

Figure 2.1: A1 和 A2 為旁系同源基因，而 A1x 和 A1y 為直向同源基因。

由於基因體內有許多基因序列，因此如何識別不同基因體之間的直向同源基因是一個重要的議題。可以簡單假設如果基因之間有直向同源的關係，則直向同源基因彼此之間的相似程度將遠高於與其它基因相比較的相似程度。在之後的章節中，我們將會介紹 Bidirectional Best Hit 以及 INPARANOID 兩種預測兩兩基因體彼此之間直向同源基因的方法。

水平基因轉移 (Horizontal Gene Transfer)為基因在不同的物種之間轉移，被認為是原核物種演化的主要因素 [21]。之前已有文獻討論水平

基因轉移會造成識別直向同源基因以及旁系同源基因的困難 [22]。如 Figure 2.2 的例子，基因 XA 以及基因 XB 被視為物種 A 以及物種 B 的同源基因, 事實上基因 XA 是由祖先直向演化而來，而基因 XB 是從物種 C 經由水平基因轉移而形成。在某些分析直向同源基因的方法上，如 Bidirectional Best Hit (會在之後作介紹)，基因 XA 以及 XB 會被認為有直向同源的關係。然而這兩個基因彼此並非是直向同源基因，因為 XA 及 XB 並非由物種 A 及物種 B 最後的共同祖先內的一個單一基因所分化而成。這些由水平基因轉移所造成的困擾在原核生物中是很常見的現象。

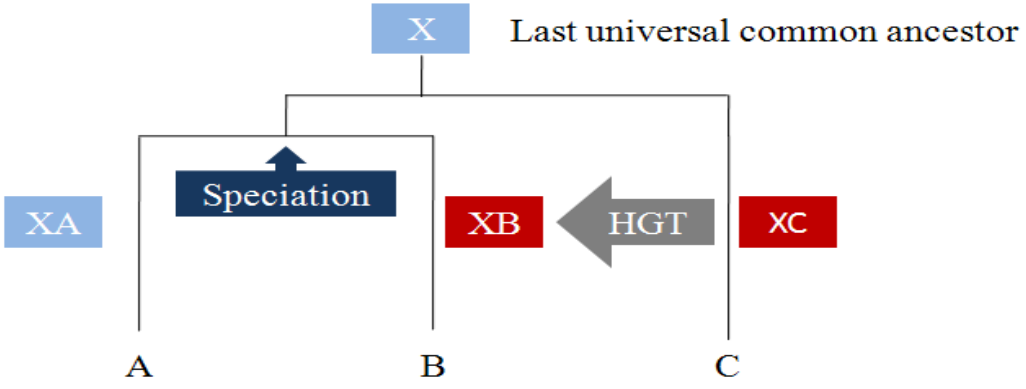

Figure 2.2: 水平基因轉移對尋找直向同源基因所造成的影響。物種 B 內的基因 XB 是由物種 C 內的基因 XC 水平基因轉移所形成。

2.2 Bidirectional Best Hit

Bidirectional Best Hit (BBH) 是一種預測兩個基因體彼此之間直向同源基因的方法。BBH 是利用 BLAST 程式去搜尋兩兩基因序列的相似性來預測兩兩基因序列之間是否擁有直向同源的關係。BBH 的定義如 Figure 2.3 所示，若當兩個基因體 G_i 跟 G_j 之間有一組基因 a 跟 b，當以基因 a 去比對 G_j 內所有的基因時，所搜尋出與基因 a 最相似的基因為基因 b，同理若基因 b 去比對 G_i 內所有基因時，所搜尋出與基因 b 最相似的基因為基因 a 時，則基因 a 與 b 則被視為直向同源基因。已有文獻證實 BBH 方法可以有效預測細菌基因體的直向同源基因[9]。

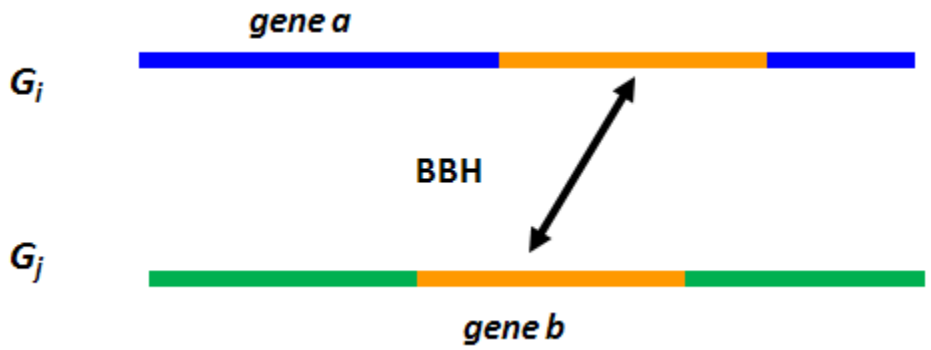

Figure 2.3: 基因 a 與基因 b 為直向同源基因。

2.3 INPARANOID

Remm *et al.* [23] 建構了一個用來尋找兩個物種間直向同源的工具，稱為 INPARANOID，而且這個工具也會考量物種內旁系同源 (Inparalogs) 的關係。根據 INPARANOID 的演算法可得知，INPARANOID 的方法可以視為是利用 BBH 方法的延伸。INPARANOID 的方法與步驟將會在下面的文章進行說明。

首先在給予兩條物種基因體後，INPARANOID 第一步會利用 BLAST 程式在所有兩兩基因序列之間進行序列比對。在經過事先已被定義好的限制條件篩選後，所有序列比對的相似分數都將在之後的步驟中被用來進行分析。

INPARANOID 在之後會藉由 BBH 的方法來尋找可能的直向同源關係並更進一步考量物種內旁系同源關係來預測出直向同源基因組。因此，一基因對於來自同一個物種的物種內同源基因也就是旁系同源基因的相似程度會比來自其他物種的基因更高。

接下來 INPARANOID 會根據以下介紹的分類演算法對所有預測的直向同源基因組進行分類。

(1) 若是直向同源組內經由 BBH 方法所找出的直向同源基因早已分類至相同的直向同源組則將兩個直向同源組合併。

(2) 若是直向同源組內主要的直向同源基因對於位於另一個基因體上的兩個直向同源基因都有相同的相似分數，則將兩個直向同源組合

併。

(3) 若是直向同源組內其中有直向同源基因早已屬於另一個擁有更高的相似程度直向同源基因組，則刪除一個新的直向基因組。

(4) 若是組中直向同源基因對的一個基因在其他直向同源基因組有更高的相似性則將兩個直向同源基因組合併。

(5) 其它所有屬於物種內旁系同源的重覆的組別根據他們對於直向同源基因的相似程度而分開。

最後，直向同源基因組的信任數值會藉由衡量每個直向同源基因組的可靠程度而評估出來。INPARANOID 是一個免費的軟體，可以到以下的網站下載使用 http://www.cbg.ki.se/inparanoid/

2.4 Breakpoint

我們藉由 Sankoff [13]於文獻中所採納的注釋來介紹斷點。首先 G 跟 H 為兩個字串(基因體)且利用正負符號(即＋或－)來標記基因的轉錄方向。則兩個字串為 $G = g_1 g_2 \ldots g_n$ 而 $H = h_1\ h_2 \ldots h_n$ 而 n 為直向同源基因的數量總數。在 G 中 g_i 位於 g_{i+1} 之前所以數列關係為 $1 \leqq i < n$。若在 G 中基因 a 與 b 相鄰且 a 位於 b 之前，但是在 H 中基因 a 與 b 並不相鄰或是相鄰但基因 a 並非位於 b 之前或是 $-b$ 並非在 $-a$ 之前，則於 G 中基因 a 與 b 產生一個斷點。在 G 中的斷點總數相等於在 H 中的斷點總數。

舉個例子來說，假設G = (-2, -1, -3, 4)而H = (3, 1, 2, 4)。必須注意的是在這邊為了可以方便觀察G與H之間的斷點數量，我們可以使用一個簡

單的手法。我們可以增加一個基因0放在G與H的起始端，並增加一個基因5放在G與H的結束位置。於是在經過簡單的處理後G = (0, -2, -1, -3, 4, 5)而H = (0, 3, 1, 2, 4, 5)。從定義我們可以得知G與H之間的斷點數量為2，因為在G中有兩個斷點分別為(0, -2)跟(-3, 4)。

Chapter 3

Methods

這在個章節中，我們將會介紹加權斷點距離並描述我們的演算法是如何藉由兩兩物種全基因體之間的加權斷點距離建構出物種之間的演化樹。

3.1 Weighted Breakpoint Distance

為了討論基因體重組的問題，我們用一個整數(integer)來表示兩個物種之間的一個直向同源基因，並利用正負符號(即＋或－)來表示其轉錄方向(如＋號代表了方向為 5’→3’，－號代表了方向為 3’←5’)。在這篇研究中，我們只著重在發生在於同股基因對的斷點，換句話說我們以發生於同股鄰近基因對的斷點為代表來討論物種之間的演化關係。

假設 $L=\{C_1,C_2,\ldots..,C_n\}$為兩個基因體 A 跟 B 之間的直向同源基因，我們根據 L 集合在 A 染色體以及 B 染色體上的位置分佈而給予不同的基因次序 $A=(a_1, a_2, \ldots, a_n)$而 $B=(b_1,b_2,\ldots,b_n)$，之後我們各別觀察是否有斷點發生於 A 跟 B 的同股鄰近基因對，並且依照基因對之間的距離將斷點定義為發生在重組速率比較慢，基因對比較不容易分開的近距離斷點，或是發生在重組速率比較快，基因對很快就會分開的遠距離斷點這兩種情形。上述用來區分鄰近基因對是否為重組速率快慢的距離可以由使用者依照生物學的知識來定義。當我們得到 A 跟 B 基因體上各別的斷點次數後，A 跟 B 之間的加權斷點距離可以由以下公式來推導出來。

$$Dis(A,B)=W_s\times\frac{(X_a^s+X_b^s)}{2\times\min(S_a,S_b)}+W_l\times\frac{\left(X_a^l+X_b^l\right)}{2\times\min(L_a,L_b)}$$

在以上的公式中，X_a^s跟X_b^s分別表示在 A 與 B 基因體上的直向同源基因中，斷點發生於重組速率慢的鄰近基因對的近距離斷點。X_a^l跟X_b^l分別表示在 A 跟 B 直向同源基因中，斷點發生於重組速率快的鄰近基因對

的遠距離斷點。S_a跟S_b分別表示為A與B基因體上重組速率慢的直系同源鄰近基因對。L_a跟L_b分別表示為A與B基因體上重組速率快的直系同源鄰近基因對。$Dis(A,B)$表示為基因體A跟基因體B之間的加權斷點距離，而W_s跟W_l則代表了使用者認為在演化距離慢以及演化距離快的鄰近基因對所形成得斷點時所給予的加權比重。

3.2 Algorithm

Figure 3.1 為我們演算法的流程圖，這個演算法是藉由加權斷點距離來建構出原核生物的基因體演化樹。

第一步為使用者輸入多個物種的編號 (accession number)，我們的工具會根據編號從 NCBI (National Centre for Biotechnology Information) 上抓取完整的基因體序列。第二步使用者可以選擇是否要排除分析為推定的(putative)、 假設的(hypothetical)或是未知的(unknown)的基因以及水平基因轉移 (horizontal gene transfer)的基因。這些被註解為推定與假設的基因資訊是由 NCBI 對完整基因體序列中每個蛋白質編碼序列(coding sequence, 簡稱 CDS)的分析所得來的。然而這些基因尚未用生物實驗的

方法證實其真正的基因功能，因此若是要嚴謹的分析物種，使用者可以選擇是否要濾掉這些基因。水平轉移基因為不同物種的外來基因轉移至基因體內，這在原核生物中是很常發生的現象 [9]，可能會導致分析物種演化時產生錯誤的推論，因此 wBPtree 藉由 HGT-DB [9] 資料庫內關於水平基因轉移的資訊提供了讓使用者可以選擇是否要排除這些分析為水平基因轉移的基因的選項。HGT-DB 資料庫提供了大量原核生物全基因體序列中所可能為水平基因轉移的基因名單。

接下來我們會利用 BlASTP 的程式來搜尋出兩個基因體中所有可能的直向同源基因。我們所使用的方法為 Bidirectional Best Hit (BBH)。BBH 定義為當基因 *a* 跟 *b* 各別為兩個基因體 *A* 跟 *B* 的基因時，利用基因 *a* 去搜尋基因體 *B* 中跟 *a* 最相似的基因時找到 *b*。而利用基因 *b* 去搜尋基因體 *A* 中跟 *b* 最相似的基因時也找到 *a*。則我們可以推測出 *a* 跟 *b* 彼此是直向同源基因。先前有研究探討說在細菌的基因體中利用 BBH 的方法來找出預測的直向同源基因體非常準確[17]。另外我們也提供了另外一個選擇，利用 INPARANOID [9] 這個工具來搜尋兩個基因體間可能的直向同源基因。之前有文獻認為在已有的五個自動偵測直向同源基因的方法中，INPARANOID 是最好的[18]。

之後每個基因體中的鄰近基因對，我們會根據使用者所定義的距離來判斷該鄰近基因對是重組速率慢的鄰近基因對，或是重組速度快的鄰近基因對。

接下來對於任何兩個基因體，我們會將兩基因體彼此之間的斷點根據斷點發生於演化速度慢或是發生於演化速度快的鄰近基因對而分類近距離斷點或是遠距離斷點並算出兩個基因體之間的加權斷點距離。最後我們會根據 UPGMA、NJ 以及 FM 三種建構演化樹的方法以及每個基因體之間的加權斷點距離所建出的矩陣來建構出根據使用者所輸入的原核生物基因體資料的基因體演化樹。

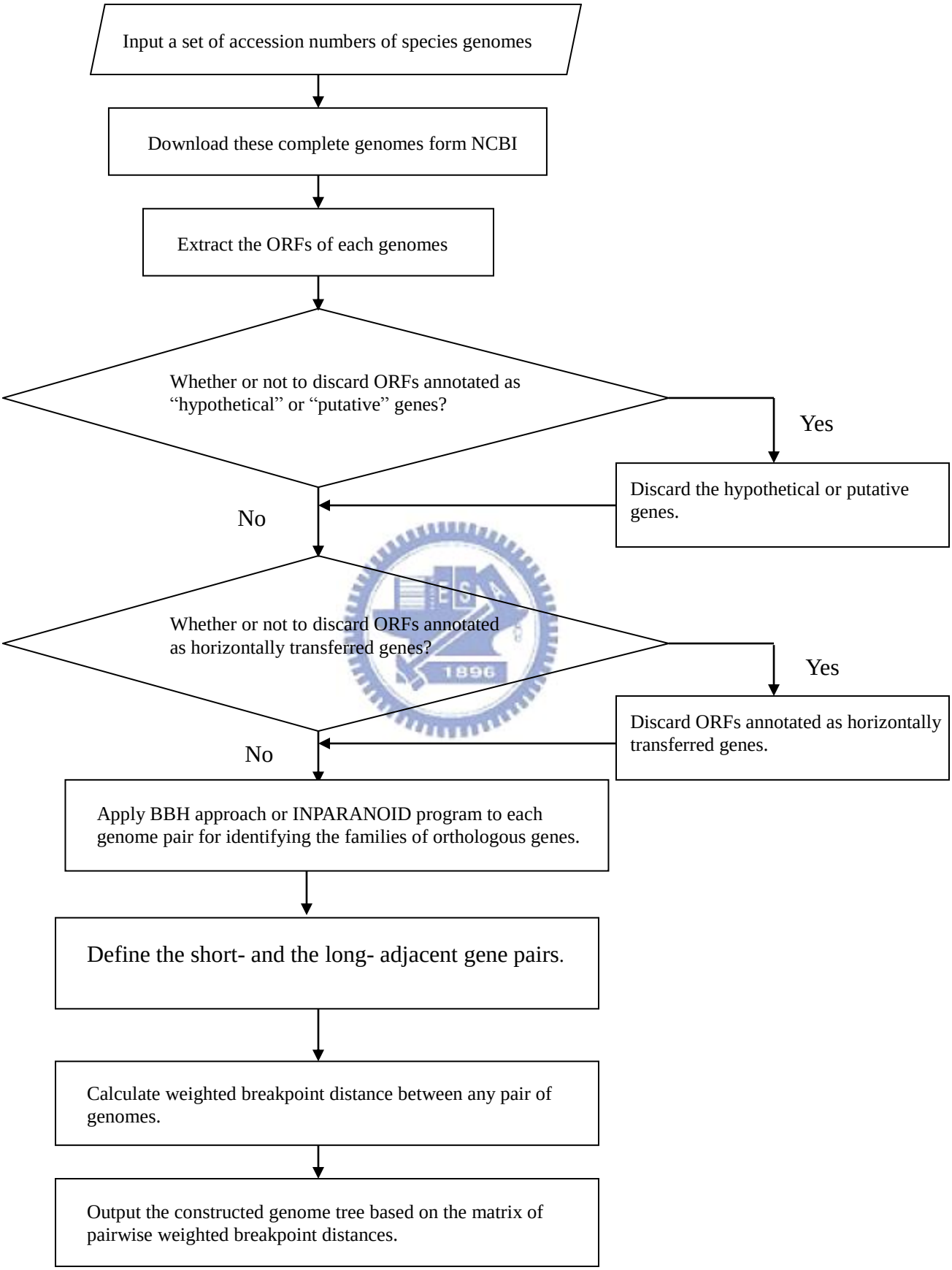

Figure 3.1: 演算法的流程圖.

Chapter 4

Implementation

根據在上個章節所描述的演算法，我們架設了一個網路伺服器工具名為 wBPtree (short for Genome **Tree** Based on **W**eight **B**reak**P**oint Distance) (http://bioalgorithm.life.nctu.edu.tw/wBPtree/)。wBPtree 所使用的核心程式是利用程式語言 C 以及 Perl 所寫成，而網頁介面則是藉由 PHP 所架設而成。此網路工具可以提供使用者在線分析物種之間的關係(Figure 4.1)。

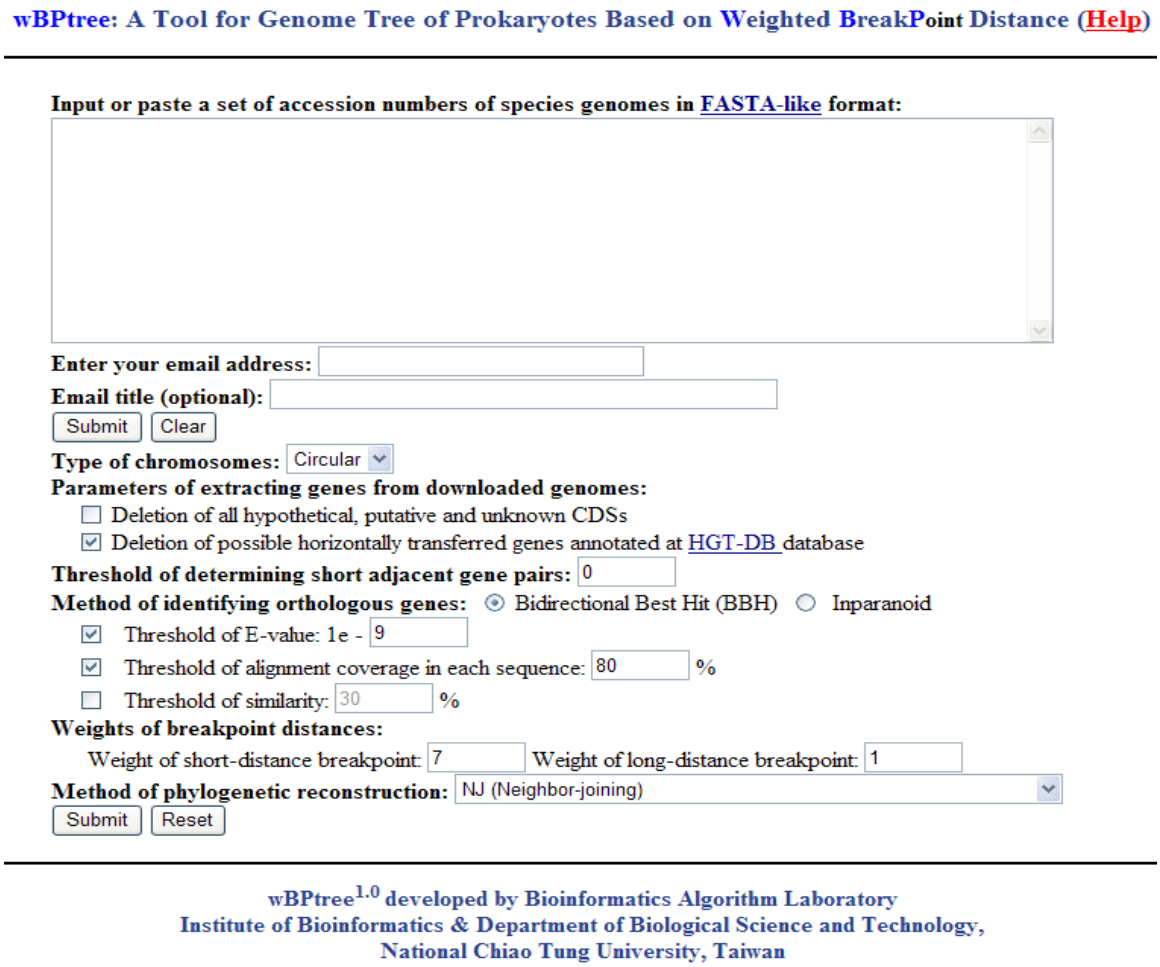

Figure 4.1: wBPtree 的網頁介面。

4.1 Input of wBPtree

1. 首先需要使用 FASTA-Like 的格式將原核生物物種於 NCNI 上的編號輸入或是複製貼上。以下是三個 γ-proteobacteria 利用 FASTA-Like 格式輸入的範例。之後 wBPtree 會自動從 NCBI 下載所有輸入的原核生物全基因體資訊。

>Ba
NC_002528
>Ec
NC_000913
>Hi
NC_000907

2. 當使用者所送出的工作完成後，系統會藉由使用者填入的電子郵件地址送出郵件告知已可觀看分析結果。若是有需要使用者也可填入電子郵件主旨可供收到郵件時得以識別。

3. 如果使用者只需要使用預設的參數作分析，則只需要點擊"Submit"按鈕，或是也可以根據之後的參數介紹調整參數。

4. 選擇所要分析的原核生物物種染色體類型為環狀或是線狀。

5. 選擇是否要刪除在 NCBI 基因體檔案內被註解為推定的、 假設的或是未知的的基因。

6. 選擇是否要刪除基因體中被 HGT-DB 資料庫分析為由水平基因轉移所形成的基因。

7. 使用者可以依照生物知識決定基因體中鄰近基因對之間的距離限制，小於限制則為重組速率慢的鄰近基因對，而超過這個限制的則為重組速度快的鄰近基因對。

8. 使用可以選擇要用哪種預測值向同源基因的方法，我們提供了 Bidirectional Best Hit (BBH)跟 INPARANOID 兩種方法。我們也提供

讓使用者也能更進一步調整更嚴謹的限制條件去偵測直向同源基因的選項，包括了 E-value 的限制、序列必須參與多少比例於比對以及基因序列的相似程度。

9. 使用者可以決定近距離斷點的斷點距離以及遠距離斷點的斷點距離其加權比重的調整。

10. 點擊"Submit"按鈕即可進行分析。

4.2 Output of wBPtree

wBPtree首先會於輸出頁面中顯示使用者輸入的原核生物全基因體資料以及使用者設定的分析參數。之後會根據物種之間的加權斷點距離而顯示物種的加權斷點距離矩陣，如Figure 4.2。

	BAP	BBp	BSg	bfl	ecc	eco	ecs	ece	hdu	hin	pae	ppu	pst	pmu	sfl	sfx	son	stm	stt	sty	vch	vpa	vvu	wgl	xac	xcc	xfa	xft	ype	ypk
BAP	556	449	495	351	494	510	512	508	375	397	368	379	393	412	494	495	412	502	504	504	435	441	435	351	348	342	310	334	497	488
BBp	0.071	496	424	338	439	455	454	450	346	380	323	335	343	385	438	444	376	447	447	445	394	399	389	331	319	311	266	288	446	439
BSg	0.000	0.017	525	333	465	484	482	480	366	390	359	376	388	395	472	475	395	477	484	483	411	422	416	335	336	331	303	310	477	471
bfl	0.580	0.559	0.834	575	502	519	520	515	392	414	380	404	408	437	508	509	423	515	515	514	441	441	422	380	364	353	334	348	512	501
ecc	1.138	0.502	1.172	0.316	4755	3255	3310	3290	948	1114	1557	1535	1532	1277	2894	2842	1432	2911	2768	2768	1358	1416	1342	513	1157	1144	745	761	2149	2133
eco	1.069	0.452	1.130	0.238	0.151	3966	3439	3408	981	1152	1619	1591	1572	1317	2988	2920	1480	2971	2825	2831	1406	1450	1378	539	1225	1210	775	790	2186	2156
ecs	1.012	0.419	1.127	0.203	0.185	0.103	4778	4130	979	1142	1611	1608	1578	1310	2991	2925	1474	2983	2826	2835	1391	1453	1374	535	1209	1191	780	794	2214	2186
ece	1.002	0.411	1.074	0.204	0.152	0.097	0.080	4731	972	1134	1587	1584	1565	1304	3011	2949	1464	2960	2799	2808	1375	1441	1366	530	1197	1180	768	781	2185	2165
hdu	3.005	1.672	3.438	1.767	2.755	2.728	2.648	2.634	1593	979	711	699	709	1057	953	951	839	970	976	975	844	863	826	401	629	608	514	530	956	936
hin	2.322	1.321	2.692	1.166	2.693	2.489	2.471	2.440	2.771	1586	841	837	841	1239	1108	1106	933	1135	1131	1126	935	963	923	403	698	687	554	573	1108	1091
pae	2.746	1.367	2.985	1.765	2.648	2.746	2.754	2.667	3.447	3.591	5227	2981	2682	947	1445	1429	1477	1604	1545	1548	1183	1251	1148	397	1528	1512	853	870	1475	1454
ppu	2.016	1.223	1.919	1.445	2.648	2.602	2.561	2.446	3.670	3.128	0.994	4891	2785	935	1430	1417	1442	1557	1498	1500	1171	1220	1125	414	1452	1432	824	840	1404	1386
pst	2.295	1.243	2.248	1.418	2.521	2.585	2.657	2.553	3.507	3.219	0.893	0.697	4857	932	1415	1404	1367	1528	1465	1464	1153	1212	1137	431	1501	1496	860	868	1441	1420
pmu	2.029	1.048	2.214	0.981	2.073	1.988	2.018	1.911	2.316	1.026	3.047	3.181	2.646	1900	1259	1259	1058	1301	1291	1288	1048	1079	1025	416	757	756	579	605	1264	1245
sfl	1.039	0.435	1.204	0.284	0.216	0.136	0.132	0.118	2.611	2.494	2.671	2.607	2.454	1.871	4016	3236	1379	2655	2551	2555	1320	1377	1319	522	1115	1095	727	750	2022	1994
sfx	1.019	0.428	1.163	0.229	0.199	0.135	0.126	0.120	2.640	2.411	2.609	2.500	2.416	1.840	0.112	3827	1363	2624	2512	2521	1315	1364	1304	515	1110	1084	727	750	1997	1974
son	1.676	0.864	1.917	0.847	1.743	1.754	1.673	1.693	2.751	2.613	2.279	2.286	1.979	2.315	1.698	1.660	4035	1445	1408	1403	1444	1502	1422	449	1146	1114	741	769	1360	1335
stm	0.897	0.460	0.866	0.209	0.226	0.226	0.272	0.252	2.645	2.374	2.589	2.451	2.470	1.846	0.218	0.230	1.770	4019	3356	3366	1395	1432	1349	530	1191	1188	769	785	2157	2132
stt	0.840	0.433	0.850	0.212	0.205	0.240	0.205	0.197	2.656	2.421	2.669	2.488	2.547	1.891	0.241	0.255	1.667	0.061	3806	3694	1371	1409	1331	529	1167	1165	766	782	2099	2070
sty	0.888	0.436	0.888	0.212	0.200	0.247	0.208	0.193	2.590	2.333	2.604	2.467	2.496	1.859	0.253	0.274	1.670	0.070	0.013	3868	1367	1407	1327	531	1163	1163	764	780	2098	2071
vch	1.363	0.807	1.549	0.707	1.203	1.208	1.158	1.151	2.607	2.127	1.997	2.014	1.909	1.655	1.145	1.108	1.248	1.272	1.283	1.243	2511	1872	1797	446	912	894	655	674	1347	1327
vpa	1.323	0.678	1.586	0.685	1.171	1.296	1.212	1.247	2.637	2.105	1.963	2.161	1.757	1.831	1.155	1.095	1.130	1.209	1.180	1.189	0.167	2818	2015	450	950	945	671	698	1389	1371
vvu	1.529	0.784	2.001	0.726	1.203	1.276	1.235	1.279	2.683	2.073	2.219	1.994	2.013	1.842	1.234	1.193	1.141	1.332	1.332	1.313	0.226	0.071	2601	438	886	867	636	668	1301	1278
wgl	1.118	0.975	1.191	0.701	1.231	1.157	1.058	1.078	3.288	2.723	3.685	2.763	2.995	2.091	1.323	1.239	1.956	1.105	1.143	1.102	1.536	1.533	1.669	599	367	354	342	351	532	521
xac	2.470	1.352	3.292	1.564	2.923	2.880	2.869	2.946	3.640	3.140	2.832	2.708	2.587	3.053	3.038	2.999	2.619	2.904	2.987	2.983	2.551	2.350	2.294	2.933	4055	3187	1279	1306	1118	1102
xcc	2.741	1.488	3.042	1.602	3.070	3.147	3.065	3.179	3.808	3.369	2.975	2.969	2.803	3.343	3.349	3.225	2.692	3.131	3.235	3.228	2.769	2.605	2.638	3.263	0.090	3977	1271	1302	1084	1072
xfa	3.311	2.038	4.045	1.798	2.667	2.823	2.756	2.818	3.444	3.387	2.449	2.223	2.360	2.903	2.734	2.710	2.270	2.637	2.725	2.737	2.503	2.325	2.338	3.835	0.487	0.482	2335	1527	746	727
xft	2.572	1.700	3.665	1.460	2.386	2.503	2.478	2.507	3.105	2.966	2.381	2.138	2.252	2.437	2.464	2.431	2.106	2.360	2.537	2.547	2.453	2.156	2.227	3.630	0.470	0.471	0.185	1811	762	747
ype	1.106	0.544	1.274	0.288	0.721	0.718	0.656	0.697	2.594	2.399	2.405	2.401	2.155	1.869	0.594	0.533	1.576	0.644	0.616	0.592	1.233	0.977	1.062	1.068	2.575	2.715	2.503	2.214	3585	3103
ypk	1.255	0.575	1.433	0.275	0.782	0.740	0.687	0.729	2.744	2.537	2.367	2.377	2.283	1.978	0.635	0.565	1.661	0.665	0.647	0.633	1.252	1.043	1.080	1.164	2.283	2.468	2.364	2.169	0.026	3809

Figure 4.2: 30 γ-Proteobacteria 的加權斷點距離矩陣。

對角線的數字表示每個物種內的基因總數，或是經過排除推定的、假設的或是未知的的基因以及水平基因轉移事件後的基因總數，使用者

點擊數字上的超連結後後可以觀看各基因的詳細資料。

在上三角的矩陣中，每個數字代表兩兩基因體之間的直向同源基因數量，使用者在點擊數字上的超連結後可以連結到兩兩基因體彼此直向同源基因的詳細資料。除了兩兩基因體之間直向同源基因的關係外也包括各基因體內重組速率慢或是重組速率快的直向同源基因對的數量以及詳細資訊。

在下三角的矩陣中，每個數字代表兩兩基因體之間的加權斷點距離，使用者可於點擊超連結後得到兩兩基因體的基因次序以及加權斷點距離的詳細資訊。

最後wBPtree會根據兩兩基因體的加權斷點距離而畫出演化樹。我們提供了UPGMA、NJ以及FM三種方法所建構出的演化樹的資訊。

Chapter 5

Experiments

在本章節中，我們將會藉由實驗將加權斷點距離與傳統斷點距離所建構的物種關係與參考樹作比較並比較加權斷點距離與傳統斷點距離的優缺點。

5.1 30 γ-Proteobacteria complete genomes

在實驗中，我們選擇了 30 條 Gamma-Proteobacteria 來當作測試的資料。

裡面包含了*Buchnera aphidicola* str. APS (我們將名稱縮短為 Bap, NC_002528)，*Buchnera aphidicola* str. Bp (BBp, NC_004545)、*Buchnera aphidicola* str. Sg (BSg, NC_004061)、*Blochmannia floridanus* (bfl,

NC_005061)、*Escherichia coli* CFT073 (ecc, NC_004431)、*Escherichia coli* K12 (eco, NC_000913)、*Escherichia coli* 0157-H7 (ecs, NC_002695)、*Escherichia coli* 0157:H7 EDL933 (ece, NC_002655)、*Haemophilus ducreyi* 35000HP (hdu, NC_002940)、*Haemophilus ducreyi* Rd KW20 (hin, NC_000907)、*Pseudomonas aeruginosa* PAO1 (pae, NC_002516)、*Pseudomonas putida* KT2440 (ppu, NC_002947)、*Pseudomonas syringae* pv. tomato str. DC3000 (pst, NC_004578)、*Pasteurella multocida* Pm70 (pmu, NC_002663)、*Shigella flexneri* 2a str. 301 (sfl, NC_004337)、*Shigella flexneri* 2a str. 2457T (sfx, NC_004741)、*Shewanella oneidensis* MR-1 (son, NC_004347)、*Salmonella typhimurium* LT2 (stm, NC_003197)、*Salmonella enterica* subsp. enterica serovar Typhi Ty2 (stt, NC_004631)、*Salmonella enterica* subsp. enterica serovar Typhi str. CT18 (sty, NC_003198)、*Vibrio cholerae* O1 biovar eltor str. N16961 (vch, NC_002505)、*Vibrio parahaemolyticus* RIMD 2210633 (vpa, NC_004603)、*Vibrio vulnificus* CMCP6 (vvu, NC_004459)、*Wigglesworthia glossinidia* endosymbiont of *Glossina brevipalpis* (wgl, NC_004344)、*Xanthomonas axonopodis* pv. citri str. 306 (xac, NC_003919)、*Xanthomonas campestris* pv. campestris str. ATCC 33913 (xcc, NC_003902)、*Xylella fastidiosa* 9a5c (xfa,

NC_002488)、*Xylella fastidiosa* Temecula1 (xft, NC_004556)、*Yersinia pestis* CO92 (ype, NC_003143) 以及 *Yersinia pestis* KIM (ypk, NC_004088)。我們所使用的參考樹為連結十個直系同源蛋白質序列所建構出來的樹形(Figure 5.1)[1]。並將用我們的工具所建構出來的樹形(Figure5.2)與用傳統非加權的斷點距離所建構出來的樹形(Figure 5.3)做個比較。

Figure 5.1為一個好的參考樹形，同為內共生菌 (endosymbiotic)的物種(bfl, wgl, BBp, BAp, BSg)在演化樹上自己形成單一菌門，其他物種群集在演化的分析上也與之前的研究一致[19,20]。不可避免地，在網路基因體資料庫內，總會包含了註解不當的基因。在嚴謹的條件下，我們也許應該排除這些CDSs在資料庫內被註解為推定的、假設的或是未知的基因，然而在物種*W. brevipalpisa* (wgl)中大部分的CDSs都被註解為推定的、假設的或是未知的基因。因此若是排除這些CDSs會導致*W. brevipalpisa* (wgl)與其他物種之間找不到直向同源基因，取而代之的，在實驗中，我們移除了那些在HGT-DB資料庫[9]內被註解為水平基因轉移的基因，並且在搜尋直向同源基因時設定了更為嚴謹的條件，比如說在使用BBH的方法來搜尋直向同源基因時，我們將E-value限定為最少要

10^{-9}，而每個CDS最少要有80%的序列參與在比對(alignment)內。另外我們參考了*E.coli*中操作組內表現基因彼此之間的距離[15,16]，將鄰近基因對彼此之間距離為70bps以內的鄰近基因對視為重組速率較慢，比較不容易分開的基因對，大於70bps的則為重組速率快，很容易就拆開的基因對。在計算公式上，對於所有參與這次實驗的變形菌(Proteobacteria)我們在W_s跟W_l給予了比率為7:1的加權。最後我們使用了NJ方法建構實驗的演化樹。

根據我們的工具所建構出的演化樹(Figure5.2)跟參考樹可以說是幾乎一致的。雖然*Blochmannia floridanus* (bfl)並沒有跟*W. brevipalpisa* (wgl)成為一個群集，但鄰近於內共生菌(endosymbiotic)的物種。與之相比，在傳統斷點所產生的樹形中 (Figure 5.3) *Blochmannia floridanus* (bfl)卻是跟*Yersinia*成為群集。另外在我們所建構出的樹形中，*E.coli*與*Shigella*各自形成自己的群集並且彼此之間都很靠近，然而在參考樹以及傳統的breakpoint樹形中，*E.coli*與*Shigella*成為混雜在一起的群集。最後則是有關*Haemophilus*物種群集位置的討論。先前的研究指出*Haemophilu*由於總體基因重組的速率是其他人的兩倍[1]，因此在傳統斷點所產生的樹形跟參考樹相比*Haemophilus*比*Vibrio*更遠離內共生菌 (endosymbiotic)的

物種群，然而由於我們的工具所使用的加權斷點距離更加重視重組速率慢的鄰近基因對，因此在我們的演化樹上，*Haemophilus*物種位置與參考樹是一致的。

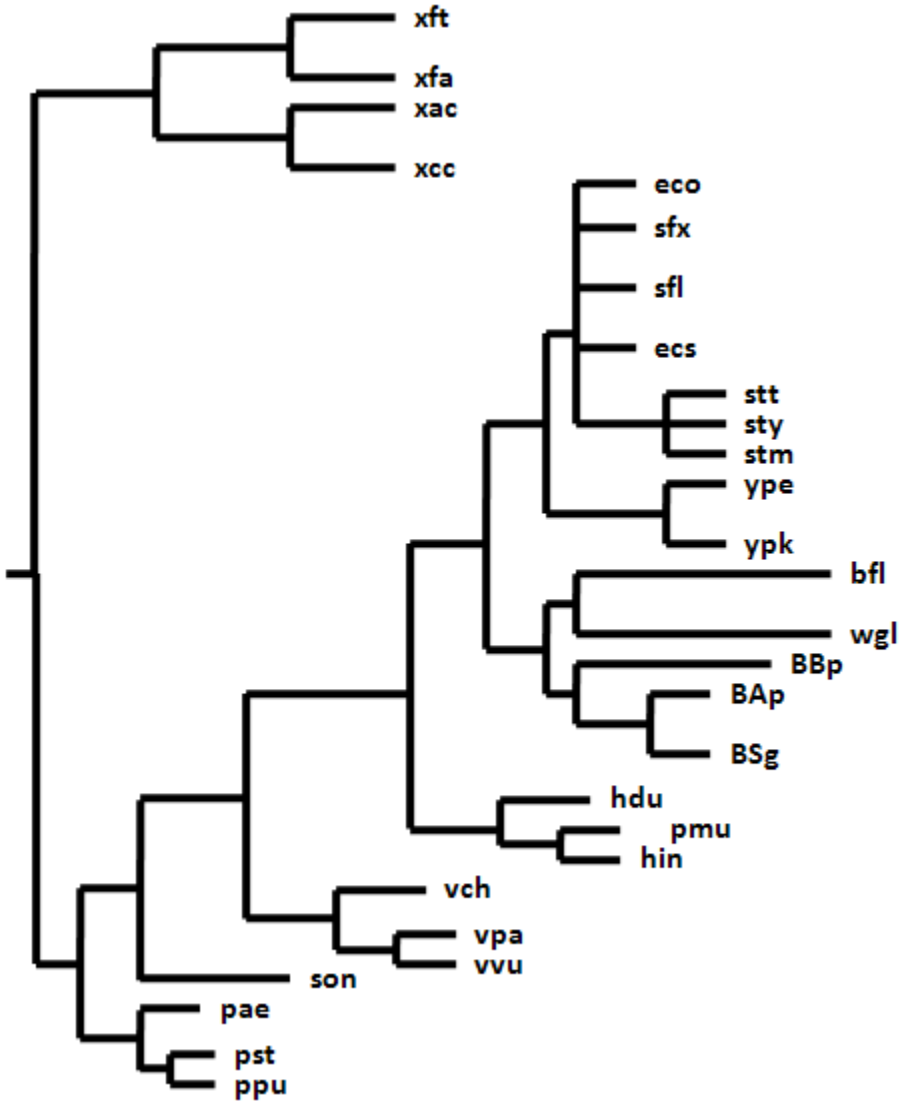

Figure 5.1: 連結十個直系同源蛋白質序列所建構出來的樹形。

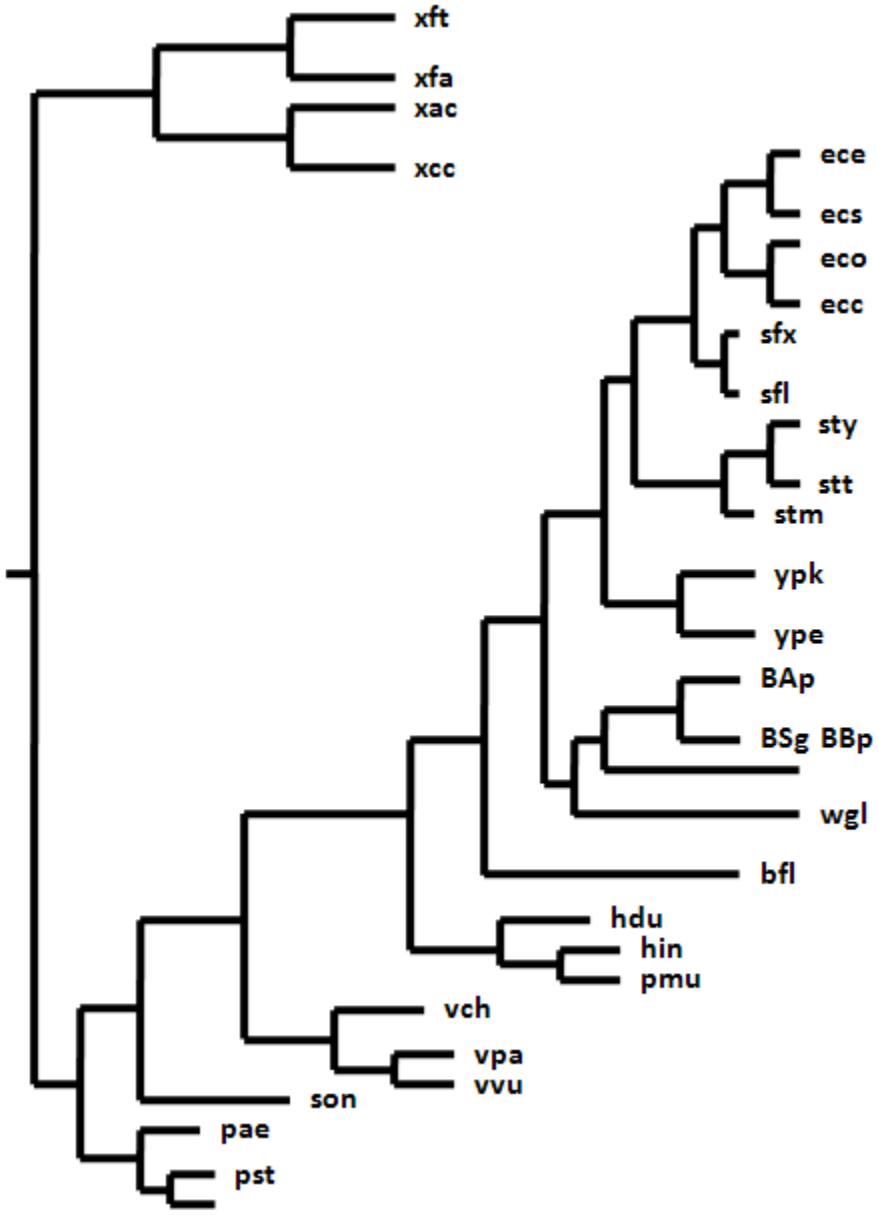

Figure 5.2: 利用加權斷點距離所建構的樹形。

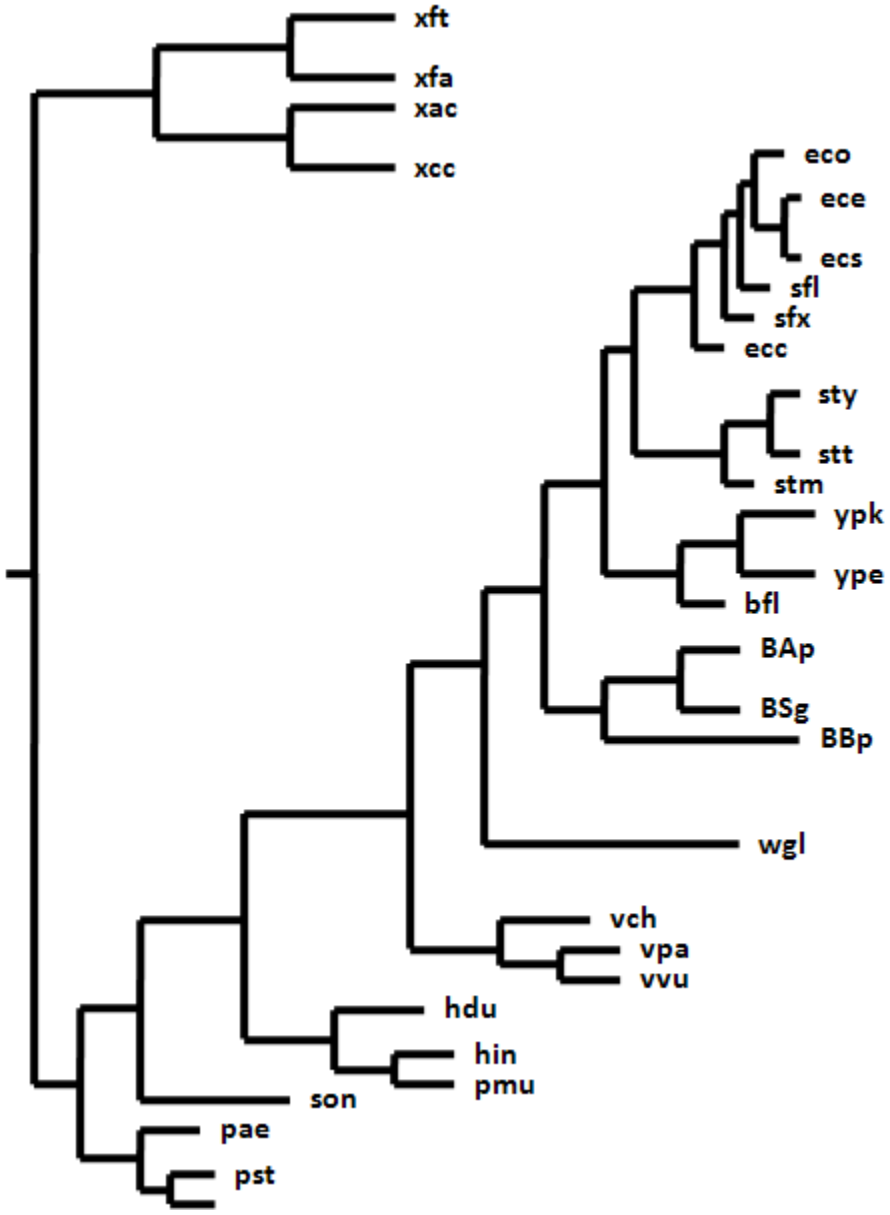

Figure 5.3: 傳統的斷點距離所建構出來的樹形。

Chapter 6

Conclusion

在這篇論文中我們藉由近距離斷點以及遠距離斷點所構成的加權斷點距離來討論並建構起原核生物物種的演化樹。並根據這項研究我們也建構了一個網路伺服器工具，稱為wBPtree以供使用者在線分析。

在計算傳統斷點距離時，由於所有基因對的重組機率都是隨機的，也因此無法區別出那些在演化上重組機率比較低的基因對跟比較高的基因對的改變。為了解決這樣的問題，我們提出一個加權的斷點觀念，並且將我們的討論專注在同股的鄰近基因對上，以區別出斷點是發生在重組機率低或是重組機率高的鄰近基因對。兩種斷點由於發生的機率不一樣當然也就給予不同的加權來進行分析與重新建構出物種間的演化樹，而我們的實驗也證實藉由加權斷點距離所建構出的樹形比傳統的斷

點距離方法更接近參考樹形，這證明了加權斷點距離比傳統的方法更準確也更有生物意義。我們相信加權斷點距離對於原核生物物種之間的演化關係上可以提供更深入的分析與探討。

References

1. Belda, E., Moya, A. and Silva, F. J. 2005 Genome rearrangement distances and gene order phylogeny in gamma-Proteobacteria. *Molecular Biology and Evolution*. **22:**1456-1467.
2. Moret, B.M. and Warnow, T. 2005 Advances in phylogeny reconstruction from gene order and content data. *Methods in Enzymology.* **395:**673-700.
3. Hughes, D. 2000. Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. *Genome Biology.* **22:**520-532.
4. Eisen, J. A., Heidelberg, J. F., White, O. and Salzberg, S. L. 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biology*. **1:** 0011.1-0011.9.
5. Tillier, E. R. and Collins, R. A. 2000. Genome rearrangement by replication-directed translocation. *Nature Genetics*. **26:**195-197.
6. Andersson, J. O. and Andersson, S. G. 2001. Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. *Molecular Biology and Evolution.* **18:**829-839.
7. Silva, F. J., Latorre, A. and Moya, A. 2001. Genome size reduction

through multiple events of gene disintegration in *Buchnera* APS. *Trends in Genetics*. **17:**615-618.

8. Koonin, E. V. and Galperin, M. Y. 1997. Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Current Opinion in Genetics & Development.* **7:**757-763.
9. Garcia-Vallve, S., Guzman, E., Montero, M. A. and Romeu, A. 2003. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Research.* **31:**187-189.
10. Ochman, H., Lawrence, J. G. and Groisman, E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature.* **405:**299-304.
11. Boucher, Y., Douady, C. J., Papke, R. T., Walsh, D. A., Boudreau,M. E., Nesbo, C. L., Case, R. J. and Doolittle, W. F. 2003. Lateral gene transfer and the origins of prokaryotic groups. *Annual Review of Geneticst.* **37:**283-328.
12. Nadeau, J. H. and Taylor, B. A. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences of the United States of Americz.* **81:**814-818.
13. Sankoff,D. 1999 Genome rearrangement with gene families. *Bioinformatics*. **15:**909-917.

14. Rocha E. P. 2006. Inference and analysis of the relative stability of bacterial chromosomes. *Molecular Biology and Evolution*. **23:**513-22.
15. Moreno-Hagelsieb, G., Treviño, V., Pérez-Rueda, E., Smith, T. F. and Collado-Vides, J. 2001. Transcription unit conservation in the three domains of life: a perspective from Escherichia coli. *Trends in Genetics*. **17:**175-7.
16. Okuda, S., Kawashima, S., Kobayashi, K., Ogasawara, N., Kanehisa, M. and Goto, S. 2007. Characterization of relationships between transcriptional units and operon structures in Bacillus subtilis and Escherichia coli. *BMC Genomics*. **8:**48.
17. Tatusov, R. L., Koonin, E. V. and Lipman, D. J. 1997. A genomic perspective on protein families. *Science*. **278:**631-637.
18. Hulsen, T., Huynen, M. A., de Vlieg, J. and Groenen, P. M. 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biology*. **7:**4.
19. Sauer, C., Stackebrandt, E., Gadau, J., Holldobler, B. and Gross, R. 2000. Systematic relationships and cospeciation of bacterial endosymbionts and their carpenter ant host species: proposalof the new taxon *Candidatus Blochmannia* gen. nov. *International journal of Systematic and Evolutionary Microbiology*. **50:**1877-1886.

20. Gil, R., Silva, F. J., Zientz, E., Delmotte, F., Gonzalez-Candelas, F., Latorre, A., Rausell, C., Kamerbeek, J., Gadau, J., Holldobler, B., van Ham, R. C., Gross, R., and Moya, A. 2003. The genome sequence of Blochmannia floridanus: comparative analysis of reduced genomes. *Proceedings of the National Academy of Sciences of the United States of Americz*. **100:**9388-9393.

21. Koonin, E.V., Makarova, K.S. and Aravind, L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Review of Microbiology*. 55:709-742.

22. Koonin, E.V. 2005. Orthologs, paralogs, and evolutionary Genomics. *Annual Review of Genetics*. **39**: 309-338.

23. Remm, M., Storm, C.E. and Sonnhammer, E.L. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology.* **314:**1041-1052.