國 立 交 通 大 學

生物資訊及系統生物研究所

碩 士 論 文

大規模直向同源基因的偵測

Large-scale Orthology Detection

研 究 生：鐘仁駿
指導教授：林苕吟 博士
盧錦隆 博士

中華民國 一 百 零 一 年 二 月

大規模直向同源基因的偵測

Large-scale Orthology Detection

研 究 生：鐘仁駿　　Student：Jen-Chun Chung

指導教授： 林苕吟 博士　Advisor：Dr. Tiao-Yin Lin

盧錦隆 博士　　Dr. Chin Lung Lu

國 立 交 通 大 學

生 物 資 訊 及 系 統 生 物 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

February 2012, Hsinchu, Taiwan

中文摘要

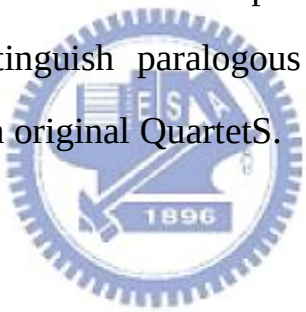
基因體快速定序技術的快速發展造成基因體序列資料的數量產生空前的成長。然而，以目前生物實驗的方法來找出基因功能的速率趕不上今日高通量定序技術的速率，使得許多已定序基因體內的基因功能仍未可得知。研究指出在不同物種內的直向同源基因應該有相同的功能。因此，找出直向同源基因在預測已定序基因體內的基因功能是有幫助的。最近有一個稱為 QuartetS 的方法被提出來執行大規模直向同源基因的偵測。QuartetS 的方法為先找到旁向同源基因，接著把那些非旁向同源基因的基因視為直向同源基因。為了找出在不同兩個物種內的兩個基因 *x* 和 *y* 是否為旁向同源基因，QuartetS 先建構由四個基因組成的基因樹，此基因樹為利用這兩個基因 *x* 和 *y* 以及第三個物種內的兩個旁向同源基因 z_1 和 z_2 所建構而成。QuartetS 利用一個近似方法在基因樹中找出樹根的位置。如果預測出的樹根位置在基因樹的內部枝幹上，那麼基因 *x* 和 *y* 就被視為旁向同源基因。否則，QuartetS 利用其他組旁向同源基因 z_1 和 z_2 並且重複上述的步驟。如果全部預先準備的旁向同源基因都不能夠用來證明基因 *x*

和 y 為旁向同源基因，那麼基因 x 和 y 就被視為直向同源基因。然而，QuartetS 的缺點有 2 個: (1)其假定物種演化的突變速率是固定的，(2)在基因樹中樹根的位置是利用近似方法估計而來的。在這份研究中，我們對 QuartetS 做了以下的改良: (1)物種的突變速率沒有假定為固定，(2)我們加入了相對於基因 x、y、z_1 和 z_2 為外群基因的的 5 個基因 o 來預測基因樹中樹根的位置。最後，實驗結果顯示從直向同源基因中區別出旁向同源基因的效能方面，我們改良的 QuartetS 方法確實是比原先 QuartetS 還要好的。

Abstract

The rapid development of genome sequencing technology has resulted in an unprecedented growth in the number of the genome sequence data. However, the rate of the current biological experimental methods to identify gene function can’t catch up with the rate of today's high-throughput sequencing technology, leading to that the functions of the genes in many sequenced genomes are still unknown. It has been reported that the orthologous genes in different species should have the same function. Hence, the identification of orthologous genes is helpful to the prediction of gene functions in the sequenced genomes. Recently, a method, called QuartetS, has been proposed to perform large-scale orthology detection. The approach of QuartetS is first to find the paralogous genes, and then consider those genes that are not paralogous as orthologous genes. To determine whether two genes, say *x* and *y*, from two different species are paralogous, QuartetS first constructed a quartet gene tree using these two genes and other two paralogous genes, say z_1 and z_2 from the third species. QuartetS used an method to approximately determine the location of the root in the quartet gene tree. If the predicted root is located in the inner edge of the quartet tree, then *x* and *y* are considered as paralogous genes. Otherwise, QuartetS used other pairs

of paralogous genes as z_1 and z_2 and repeated the above procedure. If all pre-prepared pairs of paralogous genes can't be used to prove that *x* and *y* are paralogous, then *x* and *y* are considered as a pair of the orthologous genes. However, the shortcomings of QuartetS are that the mutation rate in species evolution is assumed to be constant, and the location of the root in the quartet tree is estimated using an approximate method. In this study, we make the following modifications to improve QuartetS: (i) The mutation rate of species is not assumed to be constant. (ii) The location of the root in the quartet gene tree is predicted by adding the fifth gene *o* that is a outgroup gene with respect to genes *x*, *y*, z_1 and z_2. Finally, experimental results have shown that the performance of our improved QuartetS method to distinguish paralogous genes from orthologous genes is indeed better than original QuartetS.



Acknowledgement

時間過得很快，兩年的碩士生涯即將步入尾聲，伴隨著我的，將是一段未知的新的旅程。想起這兩年來的風風雨雨，點點滴滴依舊歷歷在目。能夠順利的完成學業，第一要先感謝家裡的人默默地給予我支持與鼓勵，能夠讓我去做自己想做的決定;其中我最感謝的是我女朋友，一路上陪著我從大學到研究所，在我最需要妳的時候，妳也永遠在我身旁，聽我抱怨也好，聊阿冏多麼欠打也好，如果說要打分數的話,相信妳一定是我的理想情人(100 分囉)!不過這不是情書，所以我們跳過這段好了。

特別要感謝的是研究室的成員，我要謝謝顆顆辛苦的 hold 住整個實驗室的行政工作，讓我得以專注地在我的研究上。感謝冏漢在研究上的給予我意見想法，此外每天一起進實驗室和回宿舍，每天一起嘴砲，也讓我這個愛講話的人有了宣洩的管道，之後不能在像以前一樣打你了，真是感到相當惋惜，不過希望你趕快畢業找到心愛的人來跟我炫耀吧!感謝學妹美齡貼心的送我生日早餐，當下真的相當感動，另外也常常討論折扣的優惠，讓我吃壽司省下不少錢。

也感謝瑋芸和熾隆，在我論文跑資料時 nice 的借我電腦來跑，才讓我的研究能順利完成。感謝大學部的學弟們，因為有你們我才得以繼續延續運動的習慣。感謝學校和慈濟基金會提供的獎助學金，在我需要幫忙時帶來一點溫暖，有機會我也會把你們的愛心延續下去的。最後要感謝我的指導教授盧錦隆老師，雖然您時常予以我們高規格要求，一開始的確很難適應，但後來發覺在過程中收穫良多，便不再那麼排斥，反而會開始去想想您所想要表達的意思;就如您所說的，要把一件事情做好是需要興趣加上能力，我會把這件事情牢記在心的!

Contents

List of tables

List of figures

Chapter 1

Introduction

由於便宜且高通量基因體定序，使得在目前基因體序列資料的數量方面以空前的速度成長中。超過一千個原核生物的序列已經在公開資料庫可以取得，而且數以百計的細菌和古細菌的基因體定序計畫目前也正在進行中 [1]。在此同時，對於部分特定的物種，實驗研究正嘗試來解譯這些大量的基因體序列資料來找出它們的基因的功能、生物特性以及和人類疾病形成的原因是否有關連性。可惜的是以目前透過生物實驗研究來解譯這些基因功能的速率沒辦法趕上高通量定序技術所定序出的序列資料，而且這項差距在可預知的未來將會拉大，所以如何在兩者之間達成平衡就成為一個重要的議題。因為直向同源基因在不同的物種中通常科學家預期有著相同的分子功能，因此利用計算方法來尋找直向同源基因可以有效率的預測未知物種的基因功能，以彌平這樣的差距。直向同源基因的偵測在生物資訊的許多領域，

如比較生物學和分子演化學等扮演了很重要的角色。我們會對偵測直向同源基因有興趣是因為它們可能保留了一部分它們祖先的功能，其應用方法就是一開始找到已知和未知物種的直向同源基因，接著把在已知物種中解譯得到的基因功能資訊套用到未知的物種上，如此就可以大致上知道未知物種可能具備哪些功能。不過當快速的累積大量的基因體資料時，卻也為大規模直向同源基因的偵測方法帶來許多挑戰，因為要進行大規模直向同源基因的偵測，以目前的工具來說，是件相當耗時的工作。雖然如此，在直向同源基因偵測方法的改良部分則帶來更多機會和資源。目前有許多已發表的資料庫儲存了已解譯的序列資料，如 InParanoid [2]，OrthoMCL-DB [3]，COG-database [4]，Homogene [5]，eggNOG [6]，OMA Browser [7]和 Ensembl Compara [8]等等。在很多案例中，沒辦法正確的利用序列比對來找出直向同源基因，主要的困難點在於有旁向同源基因 (在同一基因體出現的同源基因)存在的緣故，使得要在同源基因中找出正確的直向同源基因變得相當困難。直向同源基因和旁向同源基因皆源自於相同的祖先基因但卻透過不一樣的演化路徑:直向同源基因源自於物種形成 (Speciation，也就是新物種產生的演化過程)，而旁向同源基因源自於複製 (Duplication，也就是基因複製的演化過程) [9]。

在之前的研究中，基因複製是新的基因功能出現的關鍵。當天擇壓力出現時，常壓制基因某些特別功能的突變，而複製則提供一個機會給被複製的基因 (i.e.旁向同源基因)來避免這樣的天擇壓力，並且經歷快速的突變後，最後導致新的基因功能的產生 [10，11]。因為基因複製不只是一個可以用來定義旁向同源基因的關鍵特徵，也是功能差異化的主要原因，所以在旁向同源基因的偵測中，它應該被考慮用在旁向同源基因中，把直向同源基因區別出來的關鍵演化證據。

其中一個用來推斷和利用演化證據來偵測直向同源基因的方法是透過一致性 (Reconciliation)，i.e.利用基因樹和物種樹的比較。它假設直向同源基因的基因樹應該和它們對應的物種樹一致，根據基因樹和它相對應的物種樹之間的拓樸差異可以用來推斷基因複製的出現和基因樹中旁向同源基因的存在。這些以建樹為主的方法發展出許多偵測直向同源基因的方法，包括 RAP [12]，RIO [13]和 Orthostrapper [14]。儘管如此，沒有一個上述以建樹為主的方法可以用在大規模直向同源基因的偵測，因為這些方法需要建立相當大的基因樹，而且物種樹的建立在實務應用上常常耗費太多的計算時間 [9，15]。基因樹的建立也存在些許誤差，雖然可以額外利用其他技術來改進它們的可靠性，但這部分也額外增加了計算時間上的成本 [16]。此外雖然以建樹為主的方法被公認為是比較準確的，但一份最近的研究 [17]指

出有另外一種方法不需要建立基因樹，卻能利用現在的計算能力來提供大規模的直向同源基因的偵測。這些研究方法是建立在所謂的Bidirectional Best Hit (BBH)方法上，i.e.如果位在基因體的基因和另外一個基因體內的基因有著最高的序列相似性(通常利用 E-value 或 BLAST 搜尋後的 Bit Score 來判斷)。然而這個 BBH 的方法會造成一些誤判 (False Positive)的預測結果，因為它們真正的直向同源基因在演化過程中被刪除了，所以一個旁向同源基因就被當成 BBH 內的基因。為了試著來減少這樣的誤判的預測結果，利用第三個基因體來做比較以尋找演化證據的方法被提出來了 [15，18-24]。

Yu et al. [24]在 2011 年提出了一個大規模偵測直向同源基因的方法，稱為 QuartetS。QuartetS 藉由基因複製事件的演化證據從直向同源基因中區分出旁向同源基因。這樣的演化證據是從兩個目標基因與第三個基因體上的兩個基因所組成的基因樹上所提供。QuartetS 基因樹的四個組成基因為 x、y、z_1 和 z_2，其中基因 x 和 y 為同源基因，另外資料庫裡的第三個基因體 Z 中的 z_1 和 z_2 則提供潛在複製事件的演化證據。

QuartetS 的主要想法就是去觀察基因 x 和 y 是不是源自於旁向同源基因 z_1 和 z_2 路徑上相同的複製事件，如果是那作者就推斷基因 x 和 y 為旁向同源基因。反之則認為這組旁向同源基因 z_1 和 z_2 無法證

明基因 *x* 和 *y* 的關係。QuartetS 的缺點為: (1)作者假設演化速率及時間一定，這樣的假設在實際應用到有些的序列資訊時將會有些誤差。(2)它們利用近似方法來推斷複製事件在基因樹上的位置，此法會產生比較多的誤判的預測結果。在本篇研究中，我們提出一個方法來改良 QuartetS:即我們利用外群基因體 *O* 上的基因 *o* 來找到演化上的證據，能夠同時提供大範圍的直向同源基因偵測。這個演化證據是藉由在一個由五個基因(兩個目標基因，兩個位於第三個基因體的基因，及一個位於外群基因體的基因)所建立的基因樹中找到一個推定的複製的位置來取得。除此之外，我們也不假定演化速率是固定的，而是根據序列的內容來做分析。我們的實驗結果顯示我們的方法比起 QuartetS 從直向同源基因中區分出較多的旁向同源基因，同時我們的方法也比 QuartetS 方法的 False Positive Rate 低。

Chapter 2

Preliminaries

在這個章節，我們將會介紹直向同源基因 (Orthologous Genes)及旁向同源基因 (Paralogous Genes)的基本觀念。另外我們也會在本章節中介紹兩種偵測直向同源基因的方法，分別是 Bidirectional Best Hit (BBH)以及 QuartetS。

2.1 Orthologous Gene and Paralogous Gene

同源的意思就是任兩個個體是由同一個共同的祖先所分化而來的關係。舉例來說，如果兩個基因它們有著相同的共同祖先，我們就稱它們為同源基因。同源基因可以被分為以下兩類:

1. 直向同源基因: 源自一個物種形成 (Speciation)，就是說不同物種內的基因是由最近的共同祖先的物種內的同一基因所分化而來。
2. 旁向同源基因: 源自一個物種複製 (Duplication)，就是說基因是由相同基因體內的單一基因所複製而來。

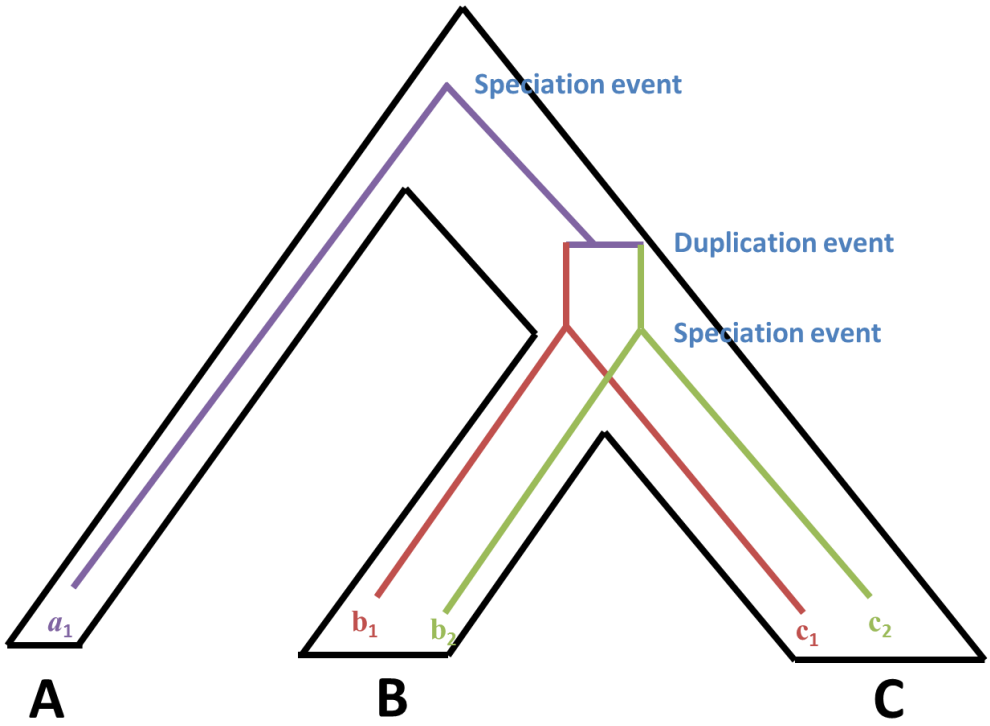


Figure 2.1: 可以用來區別直向同源基因和旁系同源基因的差異性，外圍樹型為物種樹，裡面樹型則為基因樹。(i)由基因樹可以知道基因 a_1 和 c_2 為直向同源基因，這是因為它們的共同祖先發生了物種形成事件。(ii)由基因樹可以知道基因 b_1 和 b_2 為旁向同源基因，這是因為它們的共同祖先發生了複製事件。

2.2 Bidirectional Best Hit (BBH)

Bidirectional Best Hit (BBH)方法不需要建立基因樹，卻能利用現在的計算能力來提供高通量的直向同源基因的偵測，這些研究方法是

建立在所謂的 Bidirectional Best Hit (BBH)方法上，i.e.如果位在基因體的基因和另外一個基因體內的基因有著最高的序列相似性(通常利用 E-value 或 BLAST 搜尋後的 Bit Score 來做判斷)。BBH 定義如 Figure 2.2 所示，兩個基因體 *X* 和 *Y* 分別有一組基因 x_1 和 y_1，當以基因 x_1 當作查詢基因，利用 BLAST 去對基因體 *Y* 內的所有基因做搜尋，所搜尋出與基因 x_1 最相似(Bit Score 最高)的基因為基因 y_1，反之我們也同樣的以基因 y_1 當作查詢基因，利用 BLAST 去對基因體 *X* 內的所有基因做搜尋，所搜尋出與基因 y_1 最相似的基因為基因 y_1，則基因 x_1 和 y_1 被推斷為直向同源基因。目前已經有文獻指出利用以上所介紹的 BBH 方法在偵測細菌基因組的直向同源基因有不錯的成效。

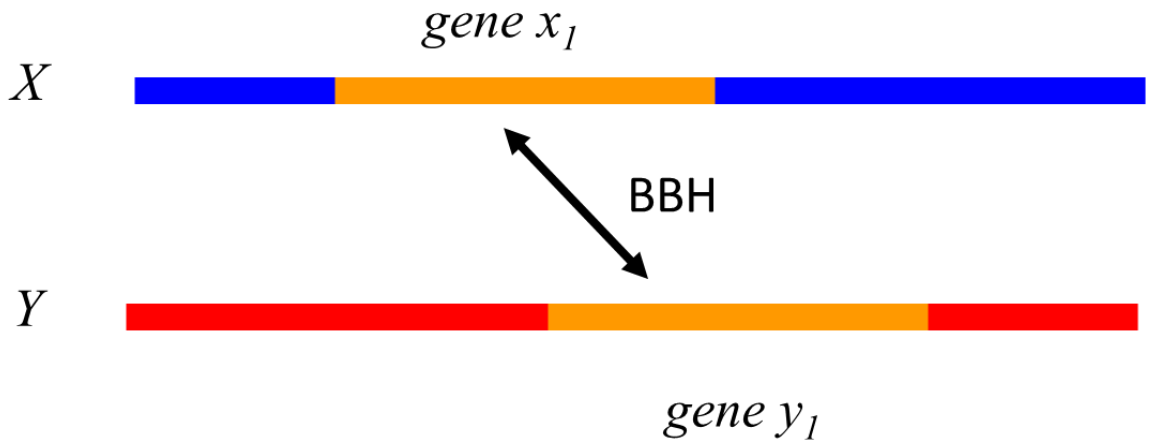


Figure 2.2: 分別位在基因體 X 和 Y 上的基因 x_1 和 y_1 彼此利用 BLAST 做搜尋後，得到兩者彼此序列相似性最高，故基因 x_1 和 y_1 為 BBH pair。

2.3 QuartetS

Yu et al. [24]在 2011 年提出了一個大規模偵測直向同源基因的方法，稱為 QuartetS。QuartetS 藉由基因複製事件的演化證據從直向同源基因中區分出旁向同源基因。這樣的演化證據是從兩個目標基因與第三個基因體上的兩個基因所組成的基因樹上所提供。QuartetS 基因樹的四個組成基因 x、y、z_1 和 z_2。如果分別位於基因體 X 和基因體 Y 內的基因 x 和 y 形成 BBH (Bidirectional Best Hit) Pair，則基因 x 和 y 為同源基因。另外資料庫裡的第三個基因體 Z 中的 z_1 和 z_2 則提供潛

在複製事件的演化證據。

QuartetS 的主要想法就是去觀察基因 *x* 和 *y* 是不是源自於旁向同源基因 z_1 和 z_2 路徑上相同的複製事件，如果是那作者就推斷基因 *x* 和 *y* 為旁向同源基因。反之則認為這組旁向同源基因 z_1 和 z_2 無法證明基因 *x* 和 *y* 的關係(Figure 2.3)。

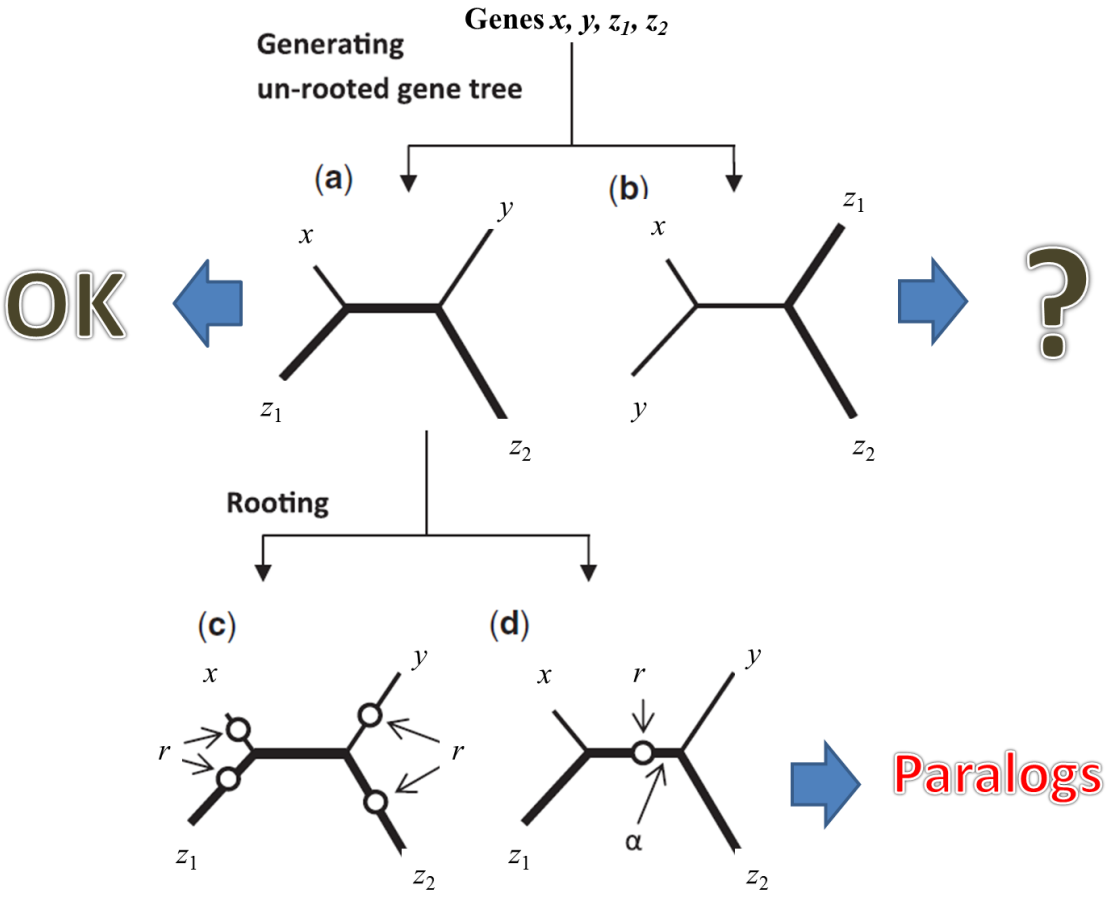


Figure 2.3: 為基因 *x*、*y*、z_1 和 z_2 所建立的無樹根的基因樹。以 a、b、c 和 d 圖解釋 QuartetS 的作法。

觀察 Figure 2.3 得到這四張圖為基因 x、y、z_1 和 z_2 所建立的無樹根的基因樹。其中旁向同源基因 z_1 和 z_2 之間的路徑(較粗部分)指出可能發生複製的地方，先看 b 圖，基因 x 和 y 的演化路徑和基因 z_1 和 z_2 的演化路徑無重疊關係，故我們無法從 b 圖得知基因 x 和 y 的關係。以 a 圖而言如果基因 x 和 y 的演化路徑和旁向同源基因 z_1 和 z_2 有重疊關係，代表我們可以在重疊的路徑上找到來證明基因 x 和 y 關係的演化證據，則我們可以拿來進行下一步找樹根的動作。接著先看 d 圖，如果基因 x 和 y 的最近共同祖先(樹根)落在內部枝幹 (Inner Branch)上，則我們推斷基因 x 和 y 為旁向同源基因。反之，如果基因 x 和 y 的最近共同祖先落在外部枝幹 (Outer Branch)上，則作者認為這一組基因 z_1 和 z_2 並無法用來推斷基因 x 和 y 的關係，所以就換掉這組 z_1 和 z_2，再從第三個基因體 Z 中取出下一組 z_1 和 z_2，重新執行建樹的步驟。從這邊我們知道樹根將決定基因 x 和 y 的關係，因此尋找樹根的位置將是這個方法的關鍵點。作者假設四個基因 x、y、z_1、z_2 以相同的突變速率和演化時間從它們最近的共同祖先 (即樹根 r)分化，並且利用一個 α 值來判斷樹根 r 的可靠性。如 Figure 2.4 所示，α 值為樹根 r 到最近的內部節點的距離(框起來的部分)。

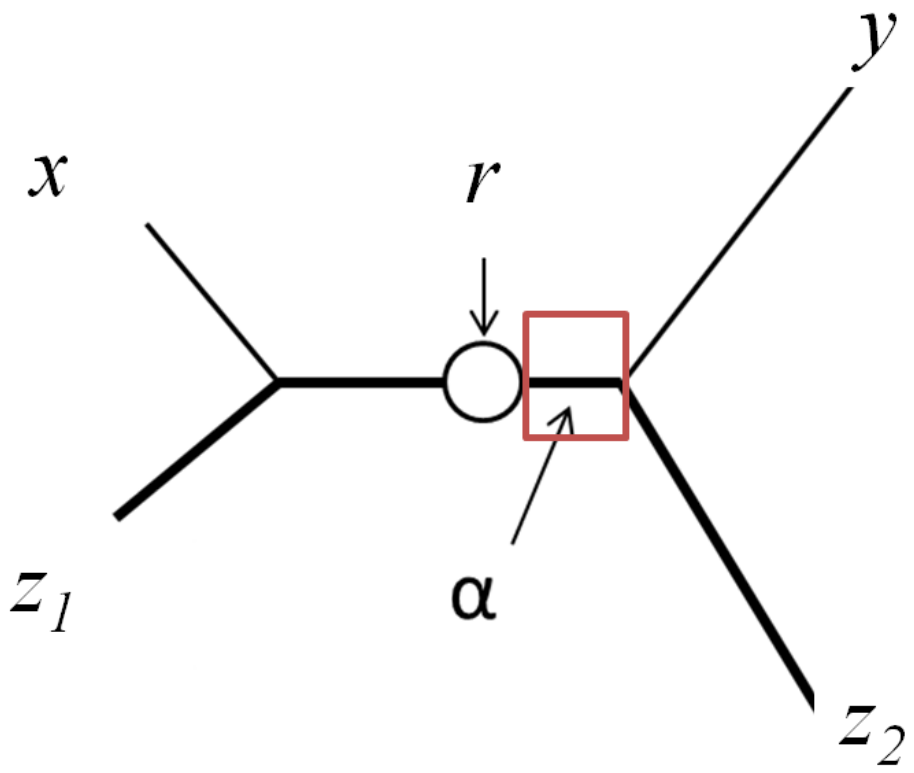


Figure 2.4: α 值在基因樹中的位置。

作者利用了以下方程式來測量 α 的近似值。

$$\alpha = \frac{1}{2}\left[\min\left(S_{x,z1}, S_{y,z2}\right) - \frac{1}{4}\left(S_{y,z1} + S_{x,z2} + S_{x,y} + S_{z1,z2}\right)\right]$$

其中 $S_{i,j}$ 為利用 BLAST 的 Bit Scores 來表示基因 i 和 j 的相似性。

如果 α 大於預先定義的門檻值(Ω，預設為 20)，則作者推斷基因 *x* 和 *y* 為旁向同源基因。隨著 Ω 值越大，會得到比較少的旁向同源基因以及比較多的直向同源基因，反之亦然。

Chapter 3

Methods

在這個章節中，我們提出改良 QuartetS 的方法，並且說明如何利用 EC Number 來進行效能評估。

3.1 The Idea of Our Method

QuartetS 的缺點為作者假設演化速率及時間一定，但這樣的假設在實際應用到序列資訊時將會有些誤差，且它們利用近似方法來計算樹根的位置，會產生比較多的誤判的預測結果。我們的方法改良自 QuartetS，修正部分如下所述: (1)我們不假設基因演化速率是一致的，因為現實生活中物種演化速率未必都一致，有些物種有不一致的現象，有些物種則一致，(2)我們利用基因 x、y、z_1、z_2 和 o 建立基因樹以找出樹根的位置，取代 QuartetS 使用近似方法來做分析，增加樹幹位置的可靠性。我們的方法找出藉由第三基因體 Z 內的旁向同源基因 z_1 和 z_2 及外群基因體 O 內的基因 o 所建構的基因樹中的演化證據來找

出分別位在基因體 *X* 和基因體 *Y* 的基因 *x* 和 *y* 的同源關係。如果內群基因的最近共同祖先(即樹根)位在內部枝幹上，如 Figure 3.1 所示，則基因 *x* 和 *y* 可能源自於旁向同源基因 z_1 和 z_2 的基因複製事件，則我們推斷基因 *x* 和 *y* 為旁向同源基因。反之，如果我們利用外群基因 *o* 在內群的外部枝幹找到內群基因的最近共同祖先，這樣的資訊沒辦法證明基因 *x* 和 *y* 源自於旁向同源基因 z_1 和 z_2 的基因複製事件，所以不能拿來證明基因 *x* 和 *y* 的同源關係，如 Figure 3.2。

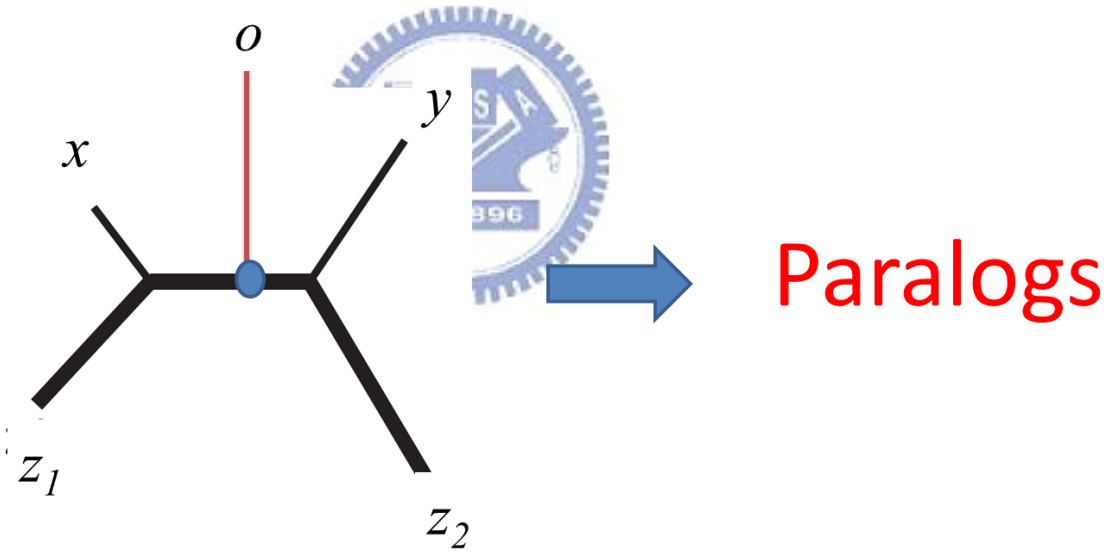


Figure 3.1: 內群基因的最近共同祖先位在內部枝幹上。

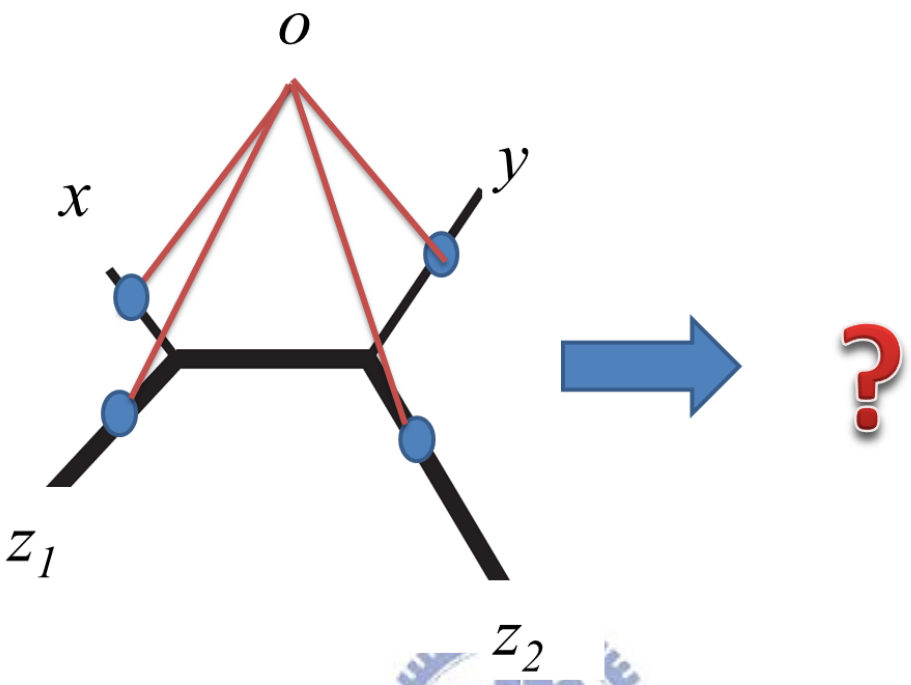


Figure 3.2: 內群基因的最近共同祖先位在外部枝幹上。

3.2 Algorithm

我們的方法流程如 Figure 3.3 所示，執行步驟如下:

Step 1

輸入一連串包括外群基因體的編號(Accession Numbers)。

Step 2

從 NCBI 的 FTP 下載這些基因體的資料。

Step 3

利用 BLAST 來對任兩個基因體做全部對全部 (All Against All)的搜尋。

Step 3-a-i

這階段就是從兩個 BLAST 檔案中擷取出 BBH Pair。

Step 3-a-ii

輸入基因體 *X* 和 *Y* 的編號以輸出 *X* 和 *Y* 的 BBH Pair。

Step 3-a-iii

分別從基因體 *X* 和 *Y* 的 BBH Pair 取出基因 *x* 和 *y*。

Step 3-b-i

接著挑選出資料庫中扣除基因體 *X* 和 *Y* 之外的第三個基因體 *Z*，並利用 BLAST 檔案來挑選出旁向同源基因 z_1 和 z_2。

旁向同源基因 z_1 和 z_2 詳細挑選方式為:

A. 先以基因 *x* 為查詢基因，再到基因體 *Z* 中利用 BLAST 做尋找，如果有找到基因且其 Bit Score 高的則當作 z_1 的候選基因;接著以基因 *y* 為查詢基因，再到基因體 *Z* 中利用 BLAST 做尋找，如果找到且 Bit Score 高的則當作 z_1 的候選基因。

B. 基因 z_2 則利用基因 z_1 當作查詢基因，再到基因體 *Z* 中利用 BLAST 做尋找，如果找到且 Bit Score 高的則當作 z_2 的候選基因。這樣可以確保基因 z_1 和 z_2 之間存在著同源的關係。

Step 3-b-ii

從外群基因體 *O* 中挑選出一個外群基因 *o*。這邊解釋一下外群基因 *o* 的詳細挑選方式:

A. 我們的程式會先以基因 *x* 為查詢基因，再到外群基因體 *O* 中利用 BLAST 做尋找，如果找到且 Bit Score 為最低，則當作 *o* 的候選基因，外群基因體預設有四組，所以一共會找出四個候選外群基因 *o*。

B. 接著從這四個候選外群基因 *o* 中挑選出一個基因 *o*，利用全區比對工具 (Global Alignment Tool) Needle 來對外群基因 *o* 與內群基因的關係做比較。如果這個外群基因 *o* 對內群基因的分數最高者比起內群基因彼此之間的分數最低者還要低，則這個外群基因 *o* 證實可以用來當作外群基因使用，反之則這個外群基

因 *o* 不可以用來當作外群基因使用，程式就換到四個候選外群基因的下一個基因 *o*，繼續以上的流程;而如果這四個候選外群基因 *o* 都證實不能用來當作外群基因，則程式隨機挑選四個外群基因體的一個基因 *o*，來繼續執行上述所提的全區比對，直到找到合適的外群基因 *o* 為止。

Step 4

我們可以利用基因 *x*、*y*、z_1、z_2和 *o* 透過多重序列比對(Multiple Sequence Alignment)來建構基因樹，這邊我們用的工具為 Clustalw 以進行多重序列比對，另外用 Protdist 和 Neighbor 等工具來建構基因樹。

Step 5

如果我們利用外群基因體在內群的外部樹幹上找到最近的共同祖先，這樣的資訊並不夠可以拿來做證明基因 *x* 和 *y* 的同源關係;反之如果在內群的內部枝幹上找到共同祖先，則我們可以推斷基因 *x* 和 *y* 為旁向同源基因。

Step 6

如果搜尋資料庫都不能有適當的旁向同源基因 z_1，z_2和外群基因 *o* 來證明基因 *x* 和 *y* 為旁向同源基因，則我們推斷基因 *x* 和 *y* 為直向同源基因。

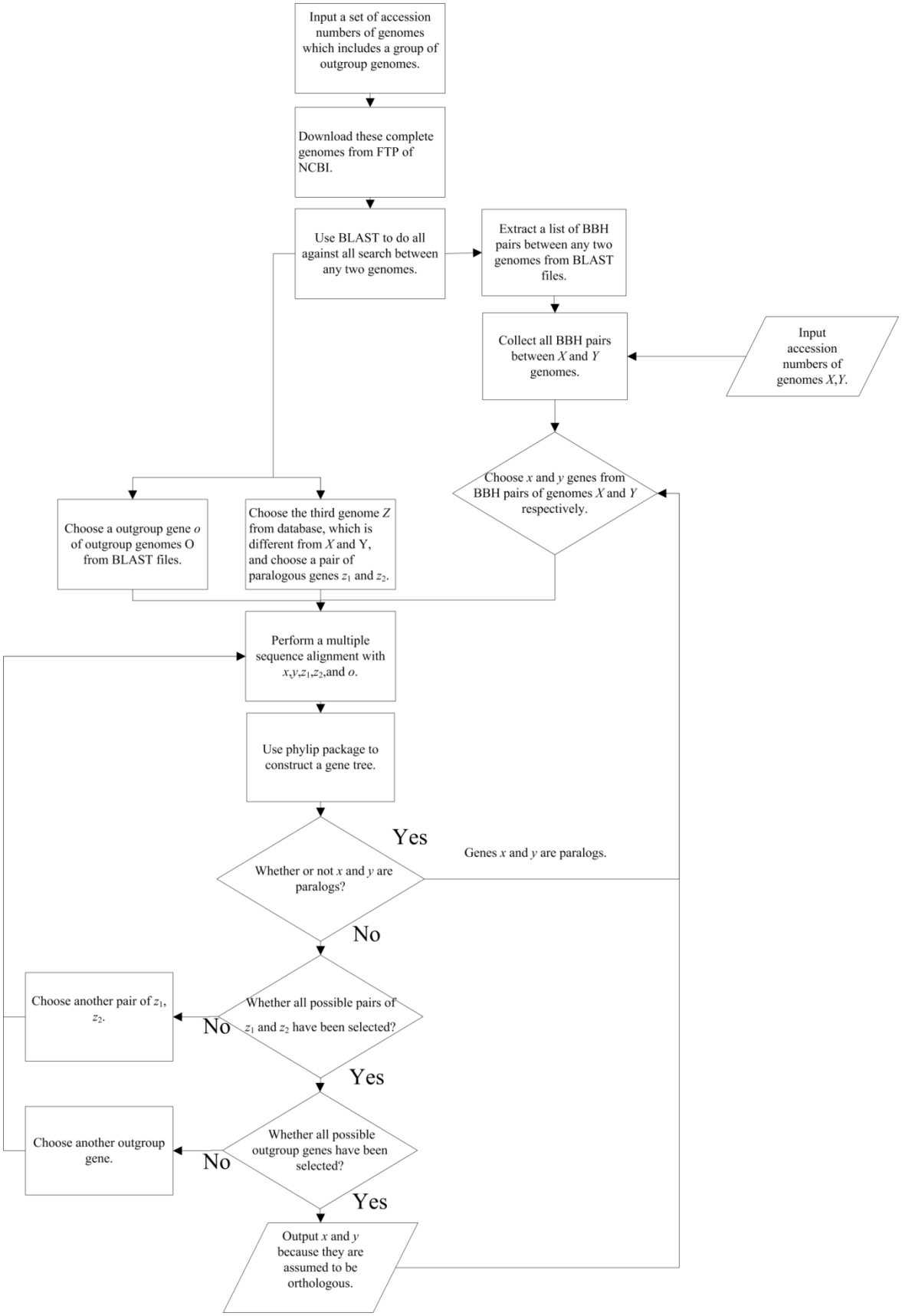


Figure 3.3: 我們的方法的流程圖。

3.3 BBH Method and EC Number

我們 BBH 比對範圍至少要大於序列長度的 50%，且一對一(Pairwise)比對結果所利用的 Bit Score 至少要大於 50。另外我們使用 EC Number 來表示基因的功能，這是由瑞士生物資訊研究所利用酵素的命名來建立 EC Number 的資料庫 [26]，它們根據蛋白質的酵素功能來訂出 EC Number。我們拿它來當參考對象。我們也把有超過一個 EC Number 以上的蛋白質移除。我們使用 EC Number 的理由為既然直向同源基因有相同的功能，大體上我們預期直向同源基因有相同的 EC Number。所以說如果直向同源基因有相同的 EC Number，我們就認為此項結果為 True Positive，否則就是 False Positive。

Chapter 4

Experimental Results

在本章節中，我們會以實驗的方式將我們的方法在進行大規模偵測直向同源基因時所預測出的結果與 QuartetS 做比較，並且說明我們的方法的優缺點。另外我們的方法和 QuartetS 所使用的 BBH Pair 會有些不同，這是因為 BBH Pair 數量會因為資料庫的大小而有所差異。

4.1 7 γ-Proteobacteria Genomes and 4 Outgroup Genomes

Table 4.1 則列出了我們進行實驗所使用的七組 γ-Proteobacterial 基因體。Table 4.2 為我們進行實驗使用的四組外群基因體。

Abbreviation	Species Name	Accession Number
eco (大腸桿菌)	Escherichia coli K12	NC_000913
vch (腸炎弧菌)	Vibrio cholerae O1 biovar eltor str. N16961	NC_002505
pae (綠膿桿菌)	Pseudomonas aeruginosa PAO1	NC_002516
hdu (杜克嗜血杆菌)	Haemophilus ducreyi 35000HP	NC_002940
stm (鼠傷寒沙門氏菌)	Salmonella typhimurium LT2	NC_003197
wgl (初級內共生菌)	Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	NC_004344
BBp (蚜蟲內共生菌)	Buchnera aphidicola str. Bp (Baizongia pistaciae)	NC_004545

Table 4.1: 七組 γ -Proteobacterial 基因體。

Abbreviation	Species Name	Accession Number
xfa (葉緣焦枯菌)	Xylella fastidiosa 9a5c	NC_002488
xcc (十字花科黑腐病菌)	Xanthomonas campestris pv. campestris str.	NC_003902
xac (地毯草黃單胞菌)	Xanthomonas axonopodis pv. citri str. 306	NC_003919
xft (葉緣焦枯菌)	Xylella fastidiosa Temecula1	NC_004556

Table 4.2: 四組外群基因體。

4.2 Experimental Results

我們方法的實驗結果如 Table 4.3 所示，右箭號所框起來部分為我們的方法預測的結果，左箭號框起來部分則為 QuartetS 的結果。我們以 Table 4.4 結果來看，第一欄為找出的旁向同源基因的數量，第二欄為找出的直向同源基因的數量。第三欄為我們的方法找出的旁向同源基因與 QuartetS 的方法找出的直向同源基因的交集。第四欄為我們的方法找出的直向同源基因與 QuartetS 的方法找出的旁向同源基因的交集。由上表第三欄可以看出 QuartetS 找出的直向同源基因有 416 個基因被我們的方法推斷為旁向同源基因，且由第一欄我們的方法找出來的旁向同源基因比 QuartetS 多，其他六組實驗結果也有同樣的現象，這代表說我們的方法比較能夠在直向同源基因中把更多的旁向同源基因給區別出來。

QuartetS　　Our method

Experimental pairs	Paralogous genes	Orthologous genes	A∩B	C∩D
NC_000913(eco)-NC_002516(pae)	79-477	1577-1179	416	18
NC_000913(eco)-NC_004545(BBp)	2-74	471-399	73	1
NC_002505(vch)-NC_004545(BBp)	12-61	414-365	55	6
NC_002516(pae)-NC_003197(stm)	77-442	1475-1110	378	13
NC_002516(pae)-NC_004545(BBp)	18-91	397-324	82	9
NC_002940(hdu)-NC_004545(BBp)	10-52	366-324	44	2
NC_004344(wgl)-NC_004545(BBp)	5-29	331-307	24	0

A: paralogous genes of our method
B: orthologous genes of QuartetS
C: orthologous genes of our method
D: paralogous genes of QuartetS

Table 4.3: 這是我們方法的實驗結果。

Experimental pairs	Paralogous genes	Orthologous genes	A∩B	C∩D
NC_000913(eco)-NC_002516(pae)	79-477	1577-1179	416	18

A: paralogous genes of our method
B: orthologous genes of QuartetS
C: orthologous genes of our method
D: paralogous genes of QuartetS

Table 4.4: 大腸桿菌對綠膿桿菌的實驗結果。

另外我們觀察大腸桿菌(NC_000913，eco)對綠膿桿菌(NC_002516，pae)的實驗結果(Table 4.5)。QuartetS 的 EC Numbers 有 379 組，True

Postives 有 372 組，False Positives 有 7 組;我們的方法的 EC Numbers 有 302 組，True Postives 有 298 組，False Positives 有 4 組。我們針對 QuartetS 的 7 組 False Positives 去做觀察，發現 QuartetS 的 7 組 False Positives 中,其中有 3 組結果被我們的方法推斷為旁向同源基因(Table 4.6 框起來處)。

NC_000913(eco)-NC_002516(pae)

	QuartetS	Our method
EC Numbers	379	302
True Postives	372	298
False Positives	7	4

Table 4.5: 使用 EC Number 來表示功能後，大腸桿菌對綠膿桿菌的實驗結果。

NC_000913(eco)	NC_002516(pae)
NP_415027.1(multifunctional acyl-CoA thioesterase I and protease I and lysophospholipase L1	NP_251546.1(tesA gene product)
NP_416140.1(adenosine deaminase)	NP_248838.1(unnamed protein product)
NP_417432.1(periplasmic L-asparaginase II)	NP_250028.1(ansB gene product)
NP_417819.1(aminodeoxychorismate synthase, subunit II)	NP_249340.1(trpG gene product)
NP_417585.2(propionate kinase/acetate kinase C, anaerobic)	NP_249527.1(ackA gene product)
NP_417606.1(tagatose 6-phosphate aldolase 1, kbaY subunit)	NP_249246.1(fda gene product)
NP_418619.1(L-ribulose 5-phosphate 4-epimerase)	NP_250374.1(unnamed protein product) 54

Table 4.6: 我們把 QuartetS 的 7 組 False Positives 找出來，發現其中有 3 組結果被我們的方法推斷為旁向同源基因。

接著我們觀察綠膿桿菌(NC_002516，pae)對鼠傷寒沙門氏菌(NC_003197，stm)的實驗結果(Table 4.7)。QuartetS 的 EC Numbers 有 334 組， True Postives 有 328 組，False Positives 有 6 組;我們的方法的 EC Numbers 有 276 組，True Postives 有 275 組，False Positives 有 1 組。我們針對 QuartetS 的 6 組 False Positives 去做觀察，發現 QuartetS 的 6 組 False Positives 中，其中有 3 組結果被我們的方法推斷為旁向同源基因(Table 4.8 框起來處)，其他 2 組 BBH 為我們所沒有的。

NC_002516(pae)-NC_003197(stm)

	QuartetS	Our method
EC Numbers	334	276
True Postives	328	275
False Positives	6	1

Table 4.7: 使用 EC Number 來表示功能後，綠膿桿菌對鼠傷寒沙門氏菌的實驗結果。

NC_002516(pae)	NC_003197(stm)
NP_248838.1(unnamed protein product)	NP_460426.1(adenosine deaminase)
NP_249246.1(fda gene product)	NP_462166.1(tagatose-bisphosphate aldolase)
NP_249340.1(trpG gene product	NP_462372.1(para-aminobenzoate synthase component II)
NP_251594.1(cobI gene product)	NP_460969.1(cobalt-precorrin-2 C(20)-methyltransferase)
NP_249527.1(ackA gene product)	NP_462156.1(propionate/acetate kinase)
NP_250374.1(unnamed protein product)	NP_463249.1(L-ribulose-5-phosphate 4-epimerase)

Table 4.8: 我們把 QuartetS 的六組 False Positives 找出來，發現其中有三組結果被我們的方法推斷為旁向同源基因，其中兩組 BBH Pair 為我們所沒有的，所以沒有辦法做比較。

最後我們觀察大腸桿菌(NC_000913，eco)對蚜蟲內共生菌(NC_004545，BBp)的實驗結果(Table 4.9)。QuartetS 的 EC Numbers 有 230 組， True Postives 有 229 組， False Positives 有 1 組;我們的方法的 EC Numbers 有 183 組，True Postives 有 183 組，False Positives 有 0 組。

我們針對 QuartetS 的 1 組 False Positives 去做觀察，發現 QuartetS 的 1 組 False Positives 中，這 1 組結果被我們的方法推斷為旁向同源基因(Table 4.10)。

NC_000913(eco)-NC_004545(BBp)

	QuartetS	Our method
EC Numbers	230	183
True Postives	229	183
False Positives	1	0

Table 4.9: 使用 EC Number 來表示功能後，大腸桿菌對蚜蟲內共生菌的實驗結果。

NC_000913(eco)	NC_004545(BBp)
NP_415711.2 (lytic murein endotransglycosylase E)	NP_777867.1 (membrane-bound lytic murein transglycosylase E)

Table 4.10: 我們把 QuartetS 的一組 False Positives 找出來，發現其中有一組結果被我們的方法推斷為旁向同源基因。

我們統計 QuartetS 對我們的方法的實驗結果(Table 4.11)。QuartetS 的 False Positives 有 12 組，全部推斷結果有 1467 組， False Positives Rate 為 0.8%;我們的方法的 False Positives 有 5 組，全部推斷結果有 1219 組，False Positives Rate 有 0.4%。從 Figure 4.1 中應該可以看出我們的方法準確度較 QuartetS 高。

False positive rate (FPR)

	QuartetS	Our method
False Positives	12	5
All Evaluated Predictions	1467	1219
False Positive Rate	0.8%	0.4%

Table 4.11: 我們的方法與 QuartetS 的方法的 FPR 比較。

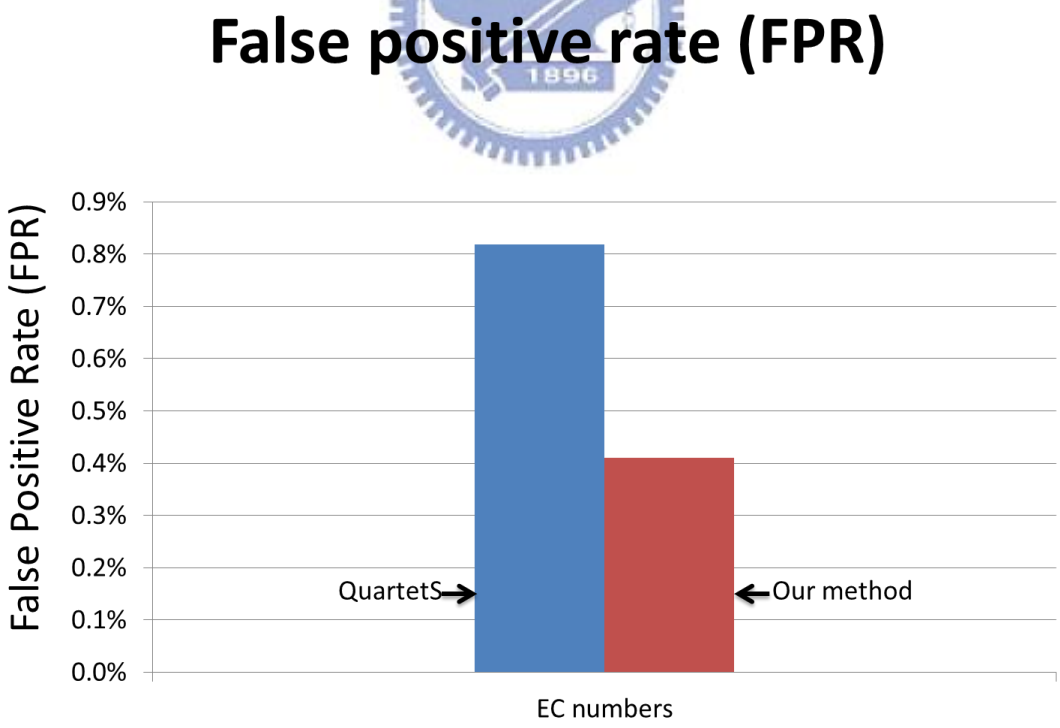


Figure 4.1: 我們計算 False Positive Rate 發現我們的方法準確度比 QuartetS 高。

4.3 Execution Time Requirement in Our Program

我們在進行實驗時，發現我們的方法執行時間上比起現行的方法或應用程式來的長。因此我們另外針對我們的方法的每個部分的執行時間做統計(Table 4.12)。從表格中可以看出，進行多重序列比對時用的 Clustalw 和建基因樹時用的 Phylip 的套件佔了幾乎全體的 93.66% (Figure 4.2)。因此針對執行時間這部分，如果能將這部分的缺點彌補起來，則整體執行時間將大幅加快。

93.66%

	Needle	Clustalw	Protdist	Neighbor	Others	Total
Sum (ms)	1339	186037	67993	27418	17504	300484
Average (ms)	16.13	252.08	92.13	37.15		
%	0.45	61.91	22.63	9.12	5.83	

Table 4.12: 我們的方法所使用的工具其時間成本比較表。

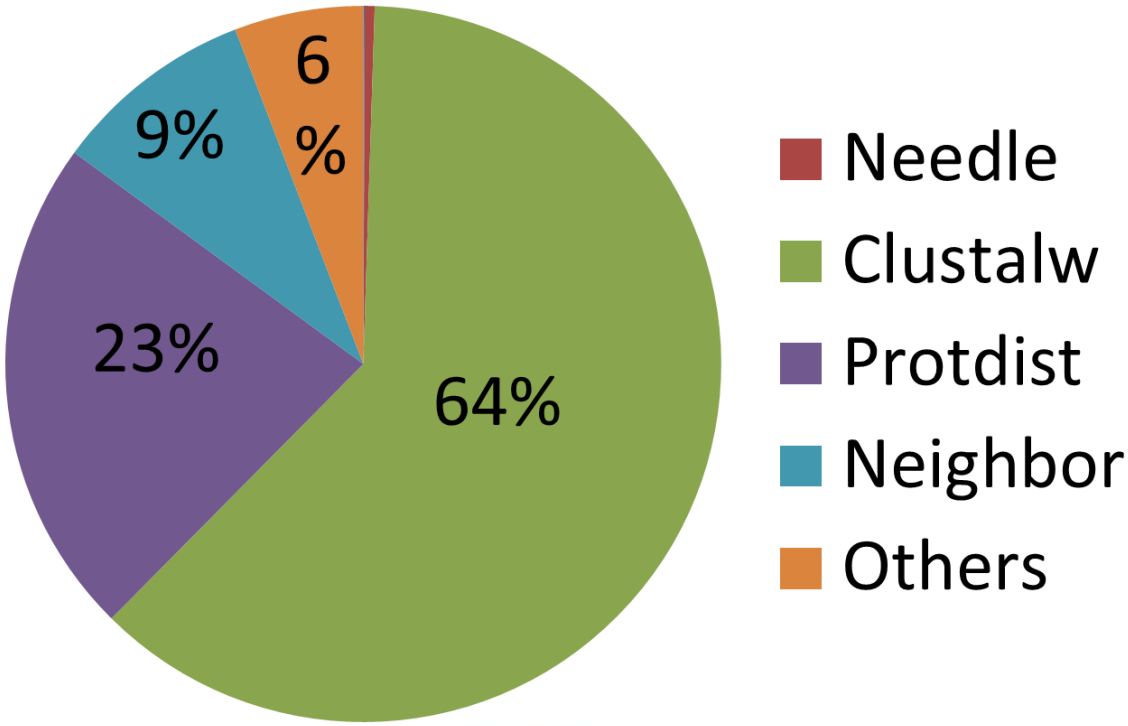


Figure 4.2: 我們的方法中，每個程式執行時間所佔比例的圓餅圖。

Chapter 5

Conclusion

在本篇論文中，我們針對大規模偵測直向同源基因的問題進行研究。QuartetS 藉由兩個目標基因與第三個基因體上的兩個基因所組成的基因樹上所提供的基因複製事件的演化證據來從直向同源基因中區分出旁向同源基因。在這份研究中，我們去改良 QuartetS 的方法，我們除了不假定演化速率是固定的之外，我們也藉由原有的四個基因並且加入的五個為外群基因所形成的基因樹來直接推論出可能的複製事件的位置。我們的實驗結果顯示，因為我們的方法比較能夠在直向同源基因中把更多的旁向同源基因給區別出來，所以我們改良的方法在成效方面確實是比原來的 QuartetS 還要來的好的。

相較於原來的 QuartetS，目前我們改良方法的執行時間還是高出很多，未來我們的重點將著重在考慮其他更可靠的尋找外群基因的方法，以減少我們方法所需要的執行時間，並增進我們的方法的正確性。

References

1. Liolios,K., Chen,I.M., Mavromatis,K., Tavernarakis,N., Hugenholtz,P., Markowitz,V.M. and Kyrpides,N.C. 2010 The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*., 38, D346–D354.

2. Berglund AC, Sjölund E, Ostlund G, Sonnhammer EL 2008: InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Research*, , 36 Database: D263-266.

3. Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS 2006: OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, , 34 Database: D363-368.

4. Tatusov RL, Galperin MY, Natale DA, Koonin EV 2000: The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28:33-36.

5. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ,

Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E 2008: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, , 36 Database: D13-21.

6. Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P 008: eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research*, , 36 Database: D250-254.

7. Schneider A, Dessimoz C, Gonnet GH 2007: OMA Browser-exploring orthologous relations across 352 complete genomes. *Bioinformatics*, 23(16):2180-2182.

8. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E 2007: Ensembl 2007. *Nucleic Acids Research*,, 35 Database: D610-617.

9. Koonin,E.V. 2005 Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, 39, 309–338.

10. Ohta,T. (2003) Evolution by gene duplication revisited: differentiation of regulatory elements versus proteins. *Genetica*, 118, 209–216.

11. Serres,M.H., Kerr,A.R., McCormack,T.J. and Riley,M. 2009 Evolution by leaps: gene duplication in bacteria. *Biology Direct*, 4, 46.

12. Dufayard,J.F., Duret,L., Penel,S., Gouy,M., Rechenmann,F. and Perriere,G. 2005 Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, 21, 2596–2603.

13. Zmasek,C.M. and Eddy,S.R. 2002 RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3, 14.

14. Hollich,V., Storm,C.E. and Sonnhammer,E.L. 2002 OrthoGUI: graphical presentation of Orthostrapper results. *Bioinformatics*, 18, 1272–1273.

15. 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.

16. Salter,L.A. and Pearl,D.K. 2001 Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. *Systematic Biology*, 50, 7–17.

17. Altenhoff,A.M. and Dessimoz,C. 2009 Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, 5, e1000262.

18. Li,L., Stoeckert,C.J. Jr and Roos,D.S. 2003 OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*earch, 13, 2178–2189.

19. Dessimoz,C., Cannarozzi,G., Gil,M., Margadant,D., Roth,A., Schneider,A. and Gonnet,G.H. 2005 OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: Introduction and first achievements. *Compare Genomics*, 3678, 61–72.

20. Alexeyenko,A., Tamas,I., Liu,G. and Sonnhammer,E.L. 2006)Automatic clustering of orthologs and in paralogs shared by multiple proteomes. *Bioinformatics*, 22, e9–e15.

21. Dessimoz,C., Boeckmann,B., Roth,A.C. and Gonnet,G.H. 2006 Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Research*., 34, 3309–3316.

22. Fulton,D.L., Li,Y.Y., Laird,M.R., Horsman,B.G., Roche,F.M. and Brinkman,F.S. 2006 Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics*, 7, 270.

23. Roth,A.C., Gonnet,G.H. and Dessimoz,C. 2008 Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*, 9, 518.

24. Yu,C., Zavaljevski,N., Desai,V. and Reifman,J. 2011 QuartetS: a fast and accurate algorithm for large-scale orthology detection. *Nucleic Acids Research*., 39, e88.

25. C. Dessimoz et al. 2000 Bairoch, A.: The ENZYME database in 2000. *Nucleic Acids Research* 28 304–305