國 立 交 通 大 學

生物資訊及系統生物研究所

碩 士 論 文

一個解決最小斷點全序化問題的貪婪演算法

A greedy algorithm for linearization of partially ordered genomes using breakpoint distance

研究生：黃晟宸
指導教授：盧錦隆 教授

中華民國　九十九　年　八　月

一個解決最小斷點全序化問題的貪婪演算法

A greedy algorithm for linearization of partially ordered genomes using breakpoint distance

研究生：黃晟宸　　Student：Chen-Cheng Huang
指導教授：盧錦隆 博士　　Advisor：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

August 2010

Hsinchu, Taiwan

中文摘要

基因或是標記在基因體上的全序關係對於大多數比較基因學的研究是非常重要的。然而除了少數的模式基因體之外，大多數的基因體尚未被完全的定序出來。目前的基因定位技術常常產生一些基因圖譜，裡頭有好幾個基因或標記會被放在同一個位置，或者會因為解析度不足的關係而遺失掉一些基因或標記，因此這些基因定位的技術只能提供基因在圖譜的偏序關係而不是全序關係。所以近年來，從一個偏序關係的基因體中推論出其全序關係的議題越來越受到關注。在本篇論文中，我們研究所謂的最小斷點全序化問題，其目的是要在一個偏序關係的基因體上找到一種可能的全序關係，使得該基因體跟另一個參考的全序關係基因體的斷點距離為最小。過去的研究已經證實了最小斷點全序化問題是 NP-hard，因此目前被提出的非指數時間演算法只是一些啟發式的或近似的演算法。在這個研究中，我們提出一個線性時間的貪婪演算法來解決一個最小斷點全序化問題的特例，即在給定的偏序關係基因圖譜上沒有基因或標記被遺失。

Abstract

The total order of the genes or markers on a genome is very important for most comparative genomics studies. However, except for a few model genomes, most genomes have not been completely sequenced yet. Current genetic mapping techniques often generate gene maps that have several genes or markers at the same position and/or are missing some other genes due to the lack of resolution in maps. They thus only suffice to produce partial orders, rather than total orders, in the gene maps. Therefore, there has been a growing interest recently in inferring the total order of genes or markers on a genome whose genes or markers are ordered partially. In this thesis, we study the so-called minimum breakpoint linearization (MBL) problem, which is to find the total order of a partially ordered genome that minimizes its breakpoint distance to a reference genome whose genes are already totally ordered (i.e., a completely sequenced genome). It was previously showed to be NP-hard

and therefore the non-exponential time algorithms proposed currently are just heuristic or approximate. In this study, we present a greedy algorithm in $O(n)$ running time for a special case of the MBL problem in which there are no missing genes in the given partially ordered gene maps.

Acknowledgement

感謝我的指導教授盧錦隆老師的耐心指導，讓我學到許多做研究應有的態度與方法，並且在我研究上遭遇到瓶頸時，不厭其煩地提供協助。也感謝時常鼓勵我的學姐、同學、學弟及朋友們。最後要感謝家人的栽培與支持，讓我得以無憂無慮的完成碩士學位。

哎呀！空白還這麼多，那我就再寫一點點好了。

「哎呀！不像話！」老大說了這句話後，轉身離開了實驗室。

昆澤科科的笑了兩聲。

「周末又去泡妞了喔？」吃素的忠翰猜測著老大抱怨的原因。

「有八卦～～！？」彥菱學姐發現事情並不單純。

昱全搶戲的眉頭一皺，若有所思的盯著螢幕。

這時候互亘發出了經典性的笑聲…

原來昱全跟互亘同一時間，從 msn 收到了張廠長的訊息。

突然！從一個深不見底的坑洞中，傳來一道聲音。

「嘖嘖…臨安捏麥賽喔 ~~~ 喔 ~~ 喔 ~ ！！」

揪竟～～八卦跟張廠長的訊息有什麼樣的糾葛，亦或是坑洞中的神秘聲音會帶來怎樣的線索？

讓我們在 Lu Lab 繼續看下去…

oops！不小心寫過頭了，結果空白變的更多…

致謝演員: 盧老大 彥菱 養葛葛 張廠長 忠翰 昆澤 芸蓁 昱全 互亘 (養葛葛比芸蓁難發現 :p)

Contents

List of figures

Chapter1.

Introduction

在大多數的比較基因體學(comparative genomics)的研究中，基因(gene)或是標記(marker)在基因體(genome)上的全序關係(total order)是很重要的。但目前的基因定位技術所產生的基因圖譜(gene map)，時常發生解析度的不足(不同的基因或標記在基因體上對映到相同的位置)或是缺少基因上的資訊，導致目前還是有許多基因體尚未被完全的定序出來，使得那些基因圖譜只能提供相關的偏序關係(partial order)而不是全序關係。舉例來說，Gramene database包含了米、玉蜀黍、燕麥和其他穀類的許多偏序關係基因圖譜 [5]，即便將其中任一穀類的偏序關的基因圖譜結合在一起的話，也只能產生出一個偏序的關係，而不是全序的關係 [6] [7] [8]。

近年來，從偏序關係的基因體上推論出全序關係的議題越來越受到關注，使得原本基因體重組問題(genome rearrangement problem)延伸出許多偏序關係上最佳化的問題。例如偏序關係的反轉距離問

題(partial-order reversal distance problem) [4] [7] [8] 、最大共同區間全序化問題(maximum common interval linearization problem) [1]、最小斷點全序化問題(minimum breakpoint linearization problem) [1] [4]。

這些從基因體重組問題所延伸出來的偏序關係最佳化問題，基本的想法是在偏序關係的基因體上找到一種可能的全序關係，使得該基因體跟另一個參考的全序或偏序基因體的基因體重組距離(genome rearrangement distance) (例如，反轉距離或是斷點距離)為最小。由於一個基因體經由不同的基因定位技術，可能得到一個以上不同的偏序關係基因圖譜，因此上述的研究會將這不同的偏序關係基因圖譜結合在一起，並利用所謂的有向非循環圖(directed acyclic graph， 簡稱 DAG)來表示。

過去的研究已經證實了最小斷點全序化問題是 NP-hard [4]， 也提出了一些啟發式演算法(heuristic algorithm) [4]或是近似解演算法(approximation algorithm) [2]。在這個研究中，我們提出一個線性時間(linear time)的貪婪演算法(greedy algorithm)來解決一個最小斷點全序化問題的特例，即在給定的偏序關係基因圖譜上沒有基因或標記被遺失。

Chapter2.

Preliminaries

Definition 2.1. *一個基因圖譜是由有序的區塊(block)所組成的，每個區塊包含至少一個以上的基因，而基因所屬的區塊之間的次序，也代表著不同區塊裡的基因的次序，但同一個區塊內的基因無次序之分。*

Definition 2.2. *若一個區塊內只有一個基因，稱為平凡區塊(trivial block)。反之，稱為非平凡區塊(nontrivial block)。*

舉例來說，1 {2, 3} {4, 5} 6 代表各有兩個平凡跟非平凡區塊，且 1 在 2、3 之前，但是 2 跟 3 之間無次序之分。一個基因圖譜通常被有向非循環圖所表示，圖上的點代表基因，有向邊所連接的基因分別屬於相鄰的兩個區塊(Figure 2-1)，若兩個基因之間存在著有向邊或是有向路徑，則表示兩個基因之間有次序關係。當一個基因體存在兩個以上的基因圖譜時，通常可以透過遞移性(transitivity)去結合那些基因圖譜，也可看成在那些有向非循環圖裡合併那些有向邊(邊集合

化簡至最小)，進而得到更完整的偏序關係。然而在合併的過程中可能會有循環的產生，意指發生衝突的情況，例如在一個基因圖譜中，基因 a 在基因 b 之前，但在另一圖譜中基因 b 在基因 a 之前，在這種情況下勢必無法找到一種合理的全序關係。因此有些循環必須被破壞，例如重新排序或是從一些基因圖譜中刪除最小數量的基因，使得衝突的關係被破壞。在本篇論文中，我們假設在基因圖譜中並沒有衝突的情況發生。

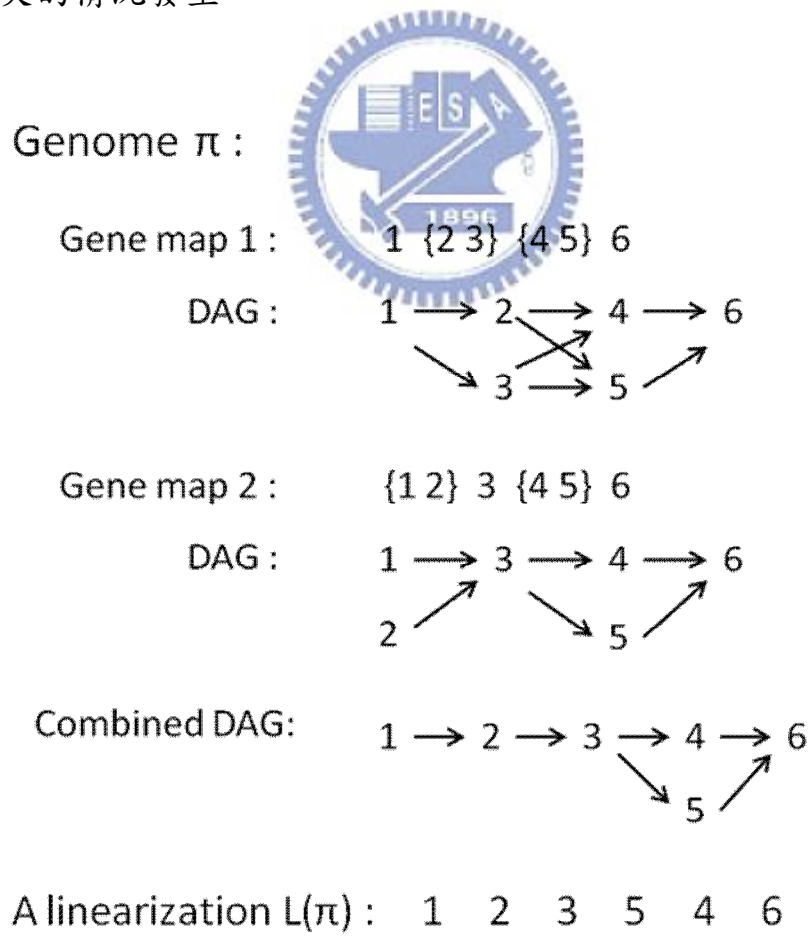

Figure 2-1. 兩個基因圖譜所組成的偏序關係基因體 π，L(π)是偏序關係基因體得一種可能的全序關係。

設π是一個基因體，用 m 個基因圖譜來表示。那些基因圖譜將會結合去建構出一個有向無循環圖，將表示為 DAG(π)，而這也表示著一個由基因集合{1, 2, … , n}所組成的偏序關係的基因體π。一個 DAG(π)的 L(π)表示 DAG(π)裡，由基因集合{1, 2, … , n}排列組合而成的一種可能的全序關係。通常一個 DAG(π)存在多種可能的全序化。設Γ為另一個基因體，由基因集合{1, 2, … , n}排列組合而成，且基因之間是全序關係。不失一般性的，我們假設Γ是恆等排列 (identity permutation) 1 2 3 4 … n。一個基因對(gene pair)在 L(π)裡相鄰，但不在Γ裡相鄰，則稱為 L(π)相對於Γ的斷點，而這些斷點的總合即是 L(π)跟Γ之間的斷點距離，用 $d_b(L(\pi), \Gamma)$表示其斷點距離。所謂的最小斷點全序化問題，如下

最小斷點全序化問題(minimum breakpoint linearization problem)

輸入: 一個由基因集合{1, 2, … , n}所組成的偏序關係為 P 的基因體 π。

輸出: 偏序關係 P 一種可能的全序化 L(π)，使得 $d_b(L(\pi), \Gamma)$為最小。

即是從一個偏序關係的基因體π中，找出一種可能的 L(π)，使得 $d_b(L(\pi), \Gamma)$為最小。

Chapter3.

Methods

Lemma 3.1. *若基因圖譜都沒有缺少的基因，則其結合後的有向無循環圖，可用有序的區塊表示其次序，也可視為一個基因圖譜。*

Proof. 對任一結合後的有向無循環圖，若其對應的基因圖譜數有m個，欲證其正確性。對基因圖譜數做數學歸納法，若基因圖譜數等於 1，成立。設基因圖譜數等於 m-1 時成立，欲證基因圖譜數為 m 時成立。由於基因圖譜數為 m-1 時成立，所以其 m-1 個基因圖譜會成為一個結合後的有向無循環圖，並對應到一個基因圖譜，此時 m-(m-1)所剩餘的基因圖譜數跟結合後的有向無循環圖所對應的基因圖譜數總合為 2，故需證明 m=2 時成立，用矛盾證法。欲證在結合後的有向無循環圖裡任兩個不同區塊的基因，不存在次序關係為錯。在結合後的有向無循環圖裡令兩個不同區塊的基因 x、y，則至少存在一個基因圖譜使得 x、y 屬於不同的區塊，在該基因圖譜裡 x、y 存在次序關係。由於結合後的有向無循環圖透過遞移性的方式結合

所有基因圖譜的次序關係，則在結合後的有向無循環圖裡 x、y 也會屬於不同的區塊且存在次序關係，矛盾。故基因圖譜數為 m 時成立。

根據 Lemma 3.1 得知，若基因圖譜都沒有缺少的基因，合併後也可視為一個基因圖譜。後續為了方便起見，對於基因圖譜的討論都是基於沒有缺少的基因下去討論。

一個基因對在 L(π)裡相鄰，但不在 Γ 裡相鄰，則稱為 L(π)相對於 Γ 的斷點，所以當我們求最小的 $d_b(L(\pi), \Gamma)$時，問題等價於在 L(π)跟 Γ 裡，找出最多共同的相鄰基因對(common adjacent gene pairs)。

Definition 3. *可能的共同相鄰基因對(possible common adjacent gene pairs)的集合，是在 Γ 相鄰且在 DAG(π)裡相鄰或是沒有次序之分的基因對所組成的。*

最小斷點全序化問題，存在一種以上可能的全序化 L(π)，使得 $d_b(L(\pi), \Gamma)$為最小(即共同的相鄰基因對最多)。

Lemma 3.2. *對於最小斷點全序化問題，基因圖譜上若相鄰的兩個區塊，不同時為非平凡區塊，且區塊跟區塊之間存在可能共同的相鄰基因對。則此基因對必屬於某一組最佳解。*

Proof. 設基因 x 屬於第 i 個區塊，基因 x+1 屬於 i+1 個區塊。不失一般性令 i=1，第 i 個區塊為平凡區塊，第 i+1 個區塊為非平凡區塊，且令一偏序關係基因體 π_1，序列為 x […x+1…]…，(x, x+1)此基因對 $\notin$ 任一組最佳解且最佳解的基因對為 k 個。設另一偏序關係基因體 π_2 除了第一個跟第二個區塊為 x x+1 外，其餘都跟 π_1 相同，且最多的基因對為 j 個(Figure 3-1 a)。欲證 j = k，證 j < k 為矛盾。將 π_1 跟 π_2 的 x+1 都移到序列最後的區塊+1，則 π_1 跟 π_2 基因排列相同(Figure 3-1 b)，在這情況下，對 π_1 而言，最多的基因對為 k 或 k+1 或 k-1 個(例如，移動 x+1 破壞了原本(x+1, x+2)這對基因對)，對 π_2 而言，最多的基因對為 j-1 或 j-2 個。又 j < k，所以不管 k = j-1 或 k = j-2 或 k-1 = j-1 或 k-1 = j-2 或 k+1=j-1 或 k+1=j-2，矛盾。

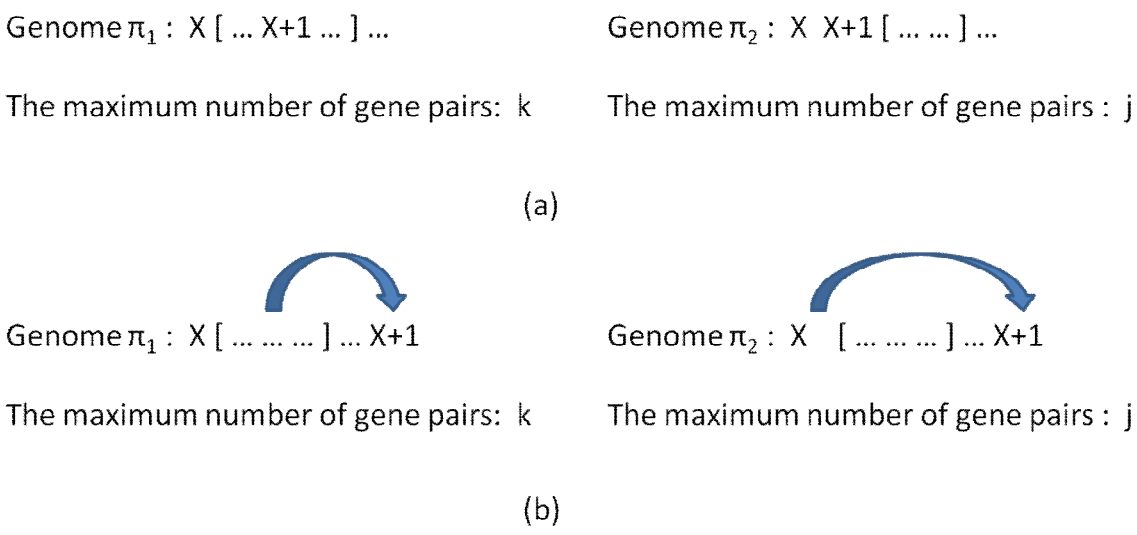

Figure 3-1. Lemma 3.2.說明圖

根據 Lemma 3.2 得知，基因圖譜上若相鄰的兩個區塊，不同時為非平凡區塊，且區塊跟區塊之間存在可能共同的相鄰基因對，則可選取該基因對，得到可能的最佳解。最小斷點全序化問題，由於 Lemma 3.1，可簡化為從單一基因圖譜上，找到一組可能的全序化 $L(\pi)$，使得 $d_b(L(\pi), \Gamma)$為最小。又根據 Lemma 3.2，最小斷點全序化問題裡的基因圖譜(由平凡跟非平凡的區塊所組成的)將可以簡化為只需考慮非平凡的區塊所組成的部分。

Definition 4. *若存在連續的可能的共同相鄰基因對且個數為極大，則稱之鏈(chain)。例如 i·i+1，i+1·i+2，...，i+k·(i+k)+1。*

Lemma 3.3. *若一鏈包含三個以上的區塊(例如區塊為 i、i+1、i+2)，且中間的區塊為非平凡區塊包含一個以上不屬於此鏈的基因，則在最佳解中至少有一個基因對屬於此鏈，不能被選取(至少出現一個斷點)。*

Proof. 令一鏈包含三個以上的區塊且中間的區塊為非平凡區塊包含一個以上不屬於此鏈的基因 x，若在這個非平凡區塊上同時選取了此鏈上的基因對，則 x 不論放哪都會造成矛盾。

根據 Lemma 3.3 可以得知，在相同區塊內的共同相鄰基因對盡

可能的越多越好。

對於非平凡區塊跟非平凡區塊之間，可能的共同相鄰基因對可視為一個點，而每個點都屬於一個階層，由左往右遞增，且每個點都屬於某個集合，若兩個點包含同一個基因，且兩個點所屬的階層相差一，則此兩個點為同一集合(Figure 3-2, 3-3)。同集合的點不能連續在相鄰的階層被選取，且一個階層只能選取一個點，而對於同一個階層，可能會出現多個點可以選擇，這時我們的演算法會根據優先權的高低(Figure 3-3)，來決定被選取的點。有向邊只能從階層 i 連往階層 j，對所有的 i、j 而言，1 <= i < j 且有向邊代表著所選的基因對的順序性。非平凡區塊跟非平凡區塊之間選擇基因對的問題將轉成圖型問題(directed k-partite graph)。

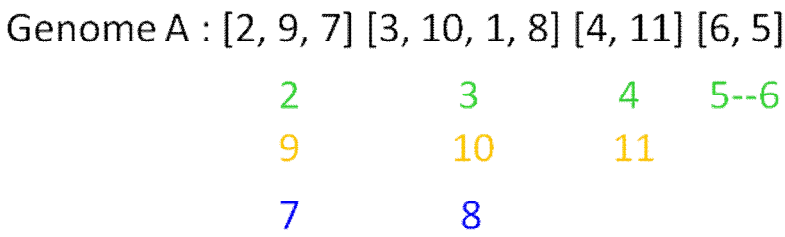

Figure 3-2. 圖中 5 跟 6 已經確定了彼此之間的順序(同區塊內基因對越多越好)。

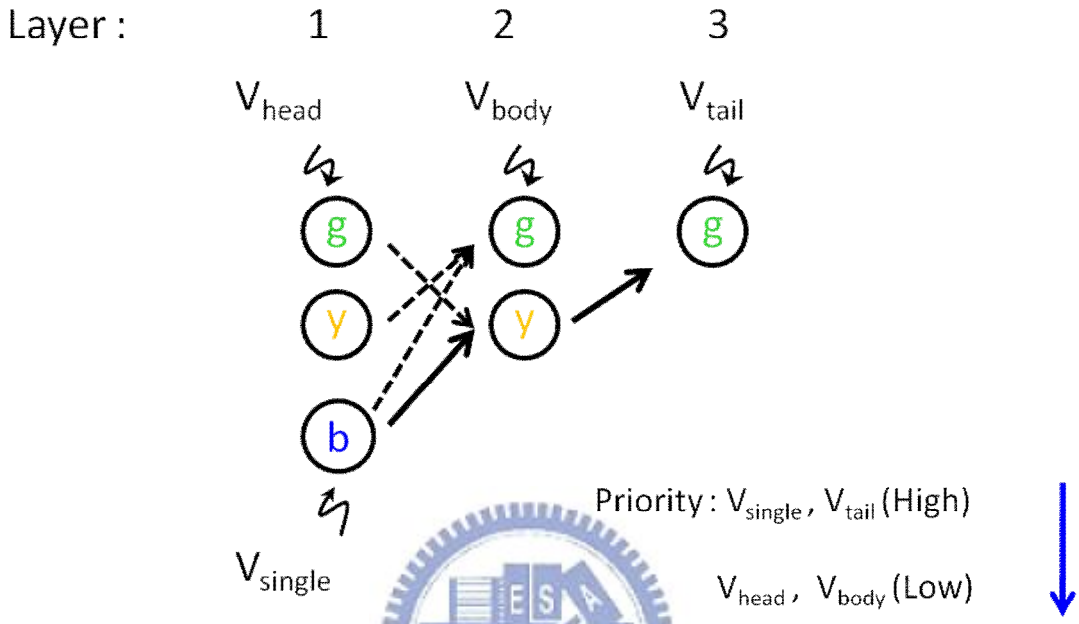

Figure 3-3. 有向 K 分圖

所以在經由 Lemma 3.1、Lemma 3.2、Lemma 3.3 我們的演算法設計如下。

演算法

輸入: 一個偏序關係為 P 的基因體 π。

輸出: 相對於 P 的一種可能的全序化，該全序化跟恆等排列之間的斷點距離為最小。

開始

在非平凡區塊裡，選取所有可能的共同相鄰基因對;

由 第一個區塊開始 至 最後一個區塊

若在區塊之間合法的可能的共同相鄰基因對的個數>=2

若高優先權基因對個數>=1，

任選一個高優先權基因對;

其餘的，執行替代路徑;

其餘的，直接選取可能的共同相鄰基因對;

將所有選取的基因對依其選取的關係排列;

剩餘未決定的基因就依其輸入順序決定其相對位置;

結束

演算法 替代路徑

輸入: 兩個以上優先權相同的基因對。

輸出: 一個替代路徑。

開始

對每個優先權相同的基因對

往右(階層遞增)找尋相對應的鏈;

個別找出一個 階層為第一大與第二大的鏈;

由 階層第二大的鏈的 tail 開始

邊往左(階層變小)交替出現於階層前二大的鏈之間;

直到 邊連接到輸入的那些基因對

結束

對於我們演算法的時間複雜度，由於在所有非平凡區塊裡，選取所有可能的共同相鄰基因對(若存在基因 i 跟 i+1 屬於同一個區塊，則選取該對基因對)，從 n 個基因的基因圖譜上由左到右檢視一遍 $O(n)$。在區塊跟區塊之間選取可能的共同相鄰基因對的時間花費有兩種情況，若不同時為非平凡區塊，則直接選取該對基因對。若

同時為非平凡區塊，且存在多個可能的共同相鄰基因對，則以優先權高低來判斷選取基因對 $O(1)$或執行替代路徑 $O(n)$。所以我們的演算法總合的時間複雜度為 $O(n)$。

Lemma 3.4. *我們的演算法在非平凡區塊所組成的序列中所找到的為最佳解.*

Proof. 欲證最佳解在非平凡區塊所組成的序列中所找到的基因對的個數大於我們的演算法所找到的基因對的個數為矛盾。

若最佳解的基因對的個數大於我們的演算法所找到的基因對的個數，則至少存在一個階層 i，最佳解有選到點，我們的演算法沒得選。不失一般性，討論階層最小的那層。對於那層最佳解有選到的點，有四種情況，該點可能是 single，head，body，tail。對於 single 跟 head 而言，由於是該集合裡階層最小的點，故不可能發生因為階層 i-1 選了同集合的點而導致階層 i 的點不能選的情況。所以若只存在一個點在階層 i， 若為 single 或 head 則我們的演算法必然會選。對於 body 跟 tail 而言，有可能發生因為階層 i-1 選了同集合的點而導致階層 i 的點不能選的情況。若發生上述情況，設階層 i-1 所選的點為 g_{i-1} 且屬於 Green 集合，而導致階層 i 不能選 g_i，則又可分為兩種情況。情況一，g_{i-1} 屬於一個替代路徑(alternative path)且開始於階

層 i-1 之前。情況二，其餘的情況。

若為情況一，由於知道g_{i-1}不是tail(因為g_i存在)，且替代路徑最後一個點必為tail，所以必存在一個不是g_i的點在階層i可被我們的演算法所選(Figure 3-4)。若為情況二，則討論在階層i-1是否存在其他不是g_{i-1}的點被最佳解所選。若不存在，則在此階層i-1我們的演算法將會比最佳解多選一個點，即g_{i-1}，又因為階層i是最佳解有選到點，而我們沒有選到點的最小階層，所以從階層1到階層i我們的演算法所找到的基因對的個數會等於最佳解的基因對的個數。若存在其他不是g_{i-1}的點被最佳解所選，不失一般性設該點屬於Purple集合，且該點有四種可能，single，head，body，tail。若為single則我們的演算法在階層i-1時會選擇優先權高的Purple的single點而不是Green的點，則與"因為選了g_{i-1}而導致g_i不能選取"之前的假設矛盾。若為head或body則表示必存在tail且該tail屬於階層j，j >= i，則在階層i勢必存在不是g_i的點可以被選取，則與"在階層i最佳解有選取基因對，而我們的演算法沒得選"之前的假設矛盾。若p_{i-1}為tail，則我們演算法會選擇優先權高的Purple點，而不是g_{i-1} 除非我們的演算法因為在階層i-2選了Purple點，導致階層i-1無法在選Purple點，則用i-1代入之前的情況判斷。情況一，p_{i-2}屬於一個替代路徑且開始於階層i-2之前。情況二，其餘的情況(Figure 3-5)。由於情況判斷時，情況一必為矛盾。情況二

時，又有四種假設，其中只有X_k(X集合，k為當下情況判斷所輸入的階層)為tail時不會矛盾，但若情況判斷式一直重複到階層1前都沒矛盾時， 則在階層1時不存在點為tail，故”至少存在一個階層 i，最佳解有選到點，我們的演算法沒得選”這個假設矛盾，得證。

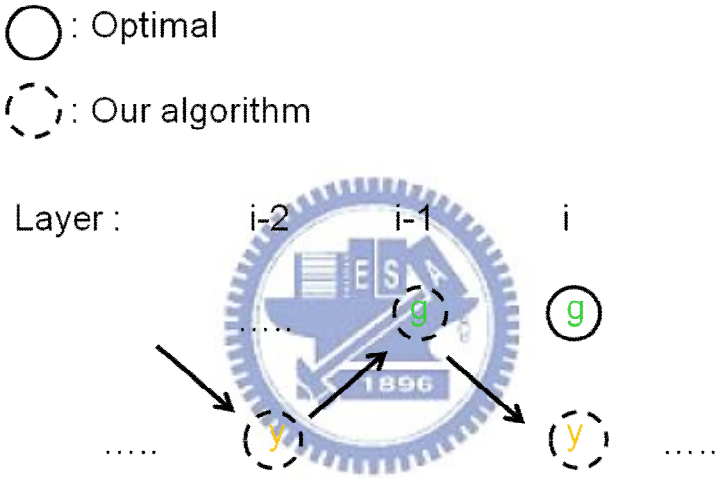

Figure 3-4

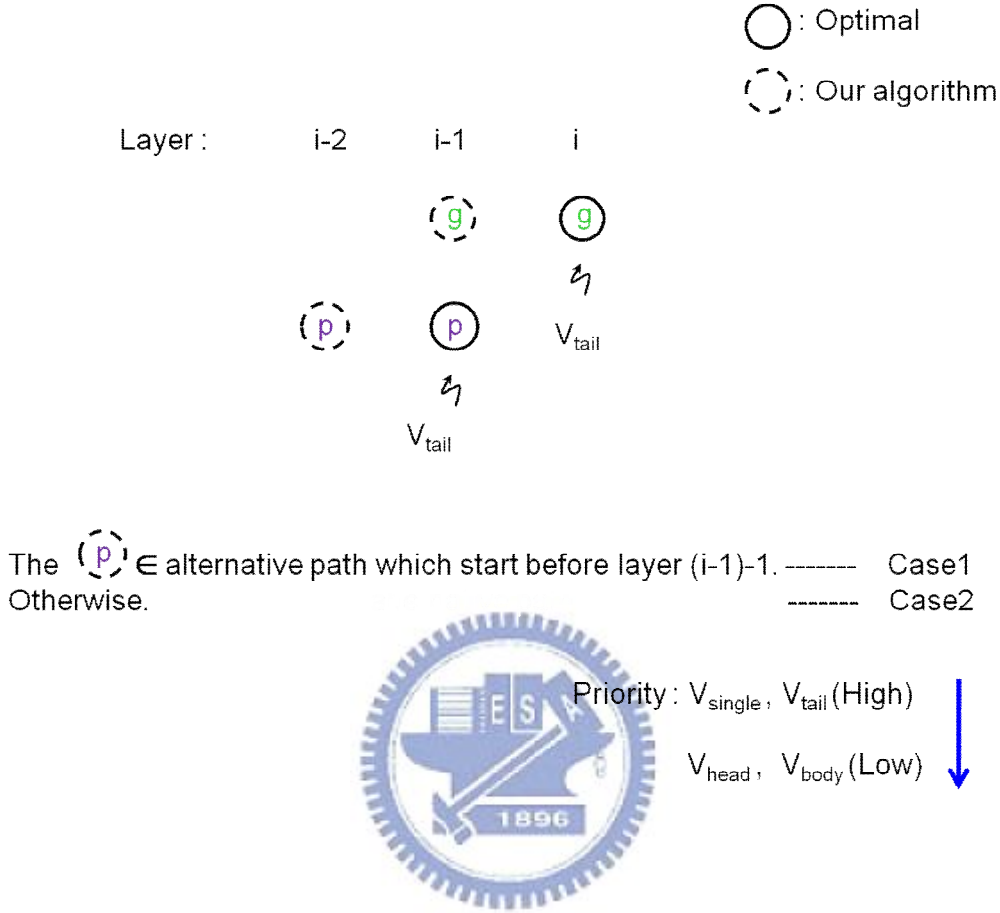
: Optimal
: Our algorithm
Layer :
i-2
i-1
i
g
g
p
p
V_{tail}
V_{tail}
The p ∈ alternative path which start before layer (i-1)-1. ------- Case1
Otherwise. ------- Case2
Priority : V_{single} , V_{tail} (High)
V_{head} , V_{body} (Low)

Figure 3-5

Chapter4.

Conclusions

在本篇論文中，對於一個特殊的最小斷點距離問題，即提供偏序關係的基因圖譜都沒有缺少基因或標記的話，我們提出了一個線性時間的貪婪演算法來解決這個問題，同時也證明了這個演算法可以得到最佳化的解。若利用其它不同的基因重組距離，如 SCJ (single cut or join) [3]，來取代斷點距離時，那麼本論文中所探討的特殊最小基因重組距離問題是不是也能在多項式時間(polynomial time)內被解決將是一個未來值得被研究的問題。

References

1. G. Blin, E. Blais, D. Hermelin, P. Guillon, M. Blanchette, and N. El-Mabrouk (2007) Gene Maps Linearization Using Genomic Rearrangement Distances. *J. Computational Biology*, **14**, 394-407.

2. X. Chen and Y. Cui (2009) An Approximation Algorithm for the Minimum Breakpoint Linearization Problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **6**, 401-409.

3. P Feijão and J Meidanis (2009) SCJ: a variant of breakpoint distance for which sorting, genome median and genome halving problems are easy. *Lect. Notes Bioinformatics*, **5724**, 85-96.

4. Z. Fu and T. Jiang (2007) Computing the Breakpoint Distance between Partially Ordered Genomes. *J. Bioinformatics and Computational Biology*, **5**, 1087-1101.

5. D. Ware, P. Jaiswal, J. Ni, X. Pan, K. Chang, K. Clark, L. Teytelman, S. Schmidt, W. Zhao, S. Cartinhour, S. McCouch, and L. Stein (2002) Gramene: A Resource for Comparative Grass Genomics. *Nucleic Acids Research*, **30**, 103-105.

6. I.V. Yap, D. Schneider, J. Kleinberg, D. Matthews, S. Cartinhour,

and S.R. McCouch (2003) A Graph-Theoretic Approach to Comparing and Integrating Genetic, Physical and Sequence-Based Maps. *Genetics*, **165**, 2235-2247.

7. C. Zheng, A. Lenert, and D. Sankoff (2005) Reversal Distance for Partially Ordered Genomes. *Bioinformatics*, **21**, i502-i508.

8. C. Zheng and D. Sankoff (2006) Genome rearrangements with partially ordered chromosomes. *Journal of Combinatorial Optimization*, **11**, 133-144.