國立臺東大學資訊管理學系

碩士論文

Department of Information Science and Management systems

National Taitung University

Master Thesis

重建無親代資料樣本的親緣關係之研究

A Research of the Full Sibling Relationship Reconstruction without Parental Information

曾千秦

Chien-Chin Tseng

指導教授：陳彥宏 博士

Advisor: Yen-Hung Chen, Ph.D.

指導教授：林育珊 博士

Advisor: Yu-Shan Lin, Ph.D.

中 華 民 國 99 年 7 月

July, 2010

國立臺東大學
學位論文考試委員審定書

系所別：

本班 曾千秦 君

所提之論文 重建無親代資料樣本的親緣關係之研究

業經本委員會通過合於 碩士學位論文 條件

論文學位考試委員會：張耀中

（學位考試委員會主席）

林育珊

陳秀宏

（指導教授）

論文學位考試日期：99 年 07 月 27 日

國立臺東大學

附註：1. 本表一式二份經學位考試委員會簽後，正本送交系所辦公室及註冊組或進修部存查。

2. 本表為日夜學制通用，請依個人學制分送教務處或進修部辦理。

博碩士論文電子檔案上網授權書

（提供授權人裝訂於紙本論文書名頁之次頁用）

本授權書所授權之論文為授權人在 國立臺東大學 資訊管理學系碩士班 _____組 98 學年度第 二 學期取得 碩士 學位之論文。

論文題目：重建無親代資料樣本的親緣關係之研究

指導教授： 陳彥宏　林育珊　博士

茲同意將授權人擁有著作權之上列論文全文(含摘要)，非專屬、無償授權國家圖書館及本人畢業學校圖書館，不限地域、時間與次數，以微縮、光碟或其他各種數位化方式將上列論文重製，並得將數位化之上列論文及論文電子檔以上載網路方式，提供讀者基於個人非營利性質之線上檢索、閱覽、下載或列印。

- 讀者基非營利性質之線上檢索、閱覽、下載或列印上列論文，應依著作權法相關規定辦理。

授權人：曾千秦

簽　名： 曾千秦　　中華民國 99 年 07 月 27 日

博碩士論文授權書

本授權書所授權之論文爲本人在 國立臺東大學 資訊管理 系(所)

組 98 學年度第 2 學期取得 碩 士學位之論文

論文名稱：重建無親代資料樣本的親緣關係之研究

本人具有著作財產權之論文全文資料，授權予下列單位：

同意	不同意	單　位
☐	☑	國家圖書館
☑	☐	本人畢業學校圖書館
☐	☑	與本人畢業學校圖書館簽訂合作協議之資料庫業者

得不限地域、時間與次數以微縮、光碟或其他各種數位化方式重製後散布發行或上載網站，藉由網路傳輸，提供讀者基於個人非營利性質之線上檢索、閱覽、下載或列印。

☐同意 ☑不同意 本人畢業學校圖書館基於學術傳播之目的，在上述範圍內得再授權第三人進行資料重製。

本論文爲本人向經濟部智慧財產局申請專利(未申請者本條款請不予理會)的附件之一，申請文號爲：＿＿＿＿＿＿，請將全文資料延後半年再公開。

公開時程

立即公開	一年後公開	二年後公開	三年後公開
			✓

上述授權內容均無須訂立讓與及授權契約書。依本授權之發行權為非專屬性發行權利。依本授權所為之收錄、重製、發行及學術研發利用均為無償。上述同意與不同意之欄位若未勾選，本人同意視同授權。

指導教授姓名：陳玄宏 林育卅 (親筆簽名)

研究生簽名：曹千秦 (親筆正楷)

學　　號：g9701309 (務必填寫)

日　　期：中華民國 99 年 7 月 27 日

1.本授權書 (得自 http: //www.lib.nttu.edu.tw/theses/ 下載) 請以黑筆撰寫並影印裝訂於書名頁之次頁。

2.依據 91 學年度第一學期一次教務會議決議:研究生畢業論文「至少需授權學校圖書館數位化，並至遲於三年後上載網路供各界使用及校內瀏覽。」

授權書版本:2008/05/29

誌謝辭

回首往昔，去年這個時候是我歡送學長姐的日子，轉瞬間碩士班二年的時光已過，在這段期間有歡笑，有快樂也有挫折。陪我走過這段心歷路程，幫助我、協助我、鼓勵我的人不少，讓我可以順利完成學業。

在這段求學的日子哩，首先我要感謝的是：二年多來不辭辛苦指導我的陳彥宏教授及林育珊教授，他們不斷的在課業、生活上的教導支持以及諄諄教誨，同時訓練我獨立思考能力及論文的撰寫，才使論文得以完成。師恩浩蕩，在此謹致由衷感激。感謝口試委員張耀中教授在百忙之中抽空前來，並對本文詳細指導建議與校閱，讓我得以順利畢業。在學習期間承蒙研究所紹永學長、好姐妹宜澤學長、金門島島主維昇學長、永遠年輕的佳慧學姊與愛抽菸的阿宅學長的照顧與支持；再者感謝一起經歷碩班二年的愛喝酒龍富、精神導師聖熙、心思細膩文翔、吃吃吃延錄、單眼皮控弘侑（柚子）、娃娃音木須、好鄰居頂級食物級慧文（饅頭）與瘦身有成冠廷讓我在碩士生涯中參雜歡笑與淚水；還有羽哥每天的打氣、士綸的陪伴與幫忙、佳倩在台東相陪與加油、母后施芳與金玉滿堂眾人的支持，你們在我的這段人生中佔有極大的意義；還有我們的學弟妹，謝謝你們在我困難的時候幫忙我。在此要感謝的人實在是太多了，無法依依全部到初，僅以最虔誠感恩的心，謝謝大家的幫忙。

最後特要感謝的是我的父母和兩位姐姐與姐夫，在我研究所期間陪我度過喜怒哀樂，一路上給我的支持、鼓勵和關懷，由於他們在我背後無怨無悔的付出，使我無後顧之憂，順利的拿到碩士學位。

總之，畢業在即，網路歷歷在目，千言萬語難以道盡。內心一則以喜、一則以憂，喜者我終於可以畢業，憂者須再面對未知的挑戰；但不管如何，經過此段期間的焠鍊與蛻變，將使我更能面對挑戰，步向成功大道。

曾千秦　謹識

于　東大資管所　2009 年 07 月

摘要

旨在探討將一群未知親代血緣關係的物種，利用 4 對偶基因型資料找出親緣關係的分群演算法問題。分群演算法常被應用於廣泛的各學術研究中，在此提出的方法將可用於基因分群上。給定 n 個物種（elements），每個物種 i 有 l 個 locus，每個 locus 有兩個對偶基因以$<a_{ij},b_{ij}>$表示，$0<j\leq l$，a_{ij} and b_{ij} 為物種 i 在第 j locus 的 2 個對偶基因。在無法得知物種之間的親緣關係時，則可利用對偶基因的特性來重建親緣關係，本研究使用孟德爾遺傳定律作為規則，其中孟德爾提出 4-allele 就是物種的每個 locus 的所有對偶基因不超過 4 種來作為本文分群方法的主要條件，即 a full sibling group 必須要滿足 4 對偶基因（4-allele）的特性。因此完全親緣關係重建問題（full sibling reconstruction problem）定義為給定 n 個物種的集合 N，每個物種 i 有 l 個 locus，每個 locus 有 2 個對偶基因以$<a_{ij},b_{ij}>$，$0<j\leq l$，目的是要將 n 個物種進行分群且每個群必須是 a full sibling group 也就是要滿足 4-allele 特性（即 $\forall 1\leq j\leq l\ |\cup_{i\in S} a_{ij}\cup b_{ij}|\leq 4$）且群組的個數要最小。當 locus 的數目為 2（respectively, $O(n^3)$）時，這問題已被證明無法設計出 1.00014（respectively, 1.0065）倍多項式時間的近似演算法，除非 $RP=NP$。在本研究中，我們設計一個時間為 $O(n^2l)$的啟發式演算法來解決完全親緣關係重建問題並將應用生物上的資料進行模擬測試。

關鍵詞：分群、完全親緣關係重建問題、孟德爾遺傳法則、4 對偶基因條件、啟發式演算法

Abstract

Motivated by the reconstruction of the full sibling relationships from a single generation of individuals without parental information, many partitioning algorithms have been well studied. Given a set of elements N, a partition of N consists of dividing the set of elements into two or more disjoint groups (non-empty subsets of N) such that each element belongs to exactly one group. A full sibling group is defined a group of the diploid individual has at most k different alleles in each locus of diploid individuals, $k \in \{2,4\}$. The 4-allele condition is a necessary condition of Mendelian inheritance for individuals to be full siblings. Given a set of n diploid individuals of the same generation N, each individual i has l loci denoted by $<a_{ij},b_{ij}>$, $0<j\leq l$, in which a_{ij} and b_{ij} are the two alleles of the individual i at locus j. The full sibling reconstruction problem is to partition the given individuals with minimum number of groups such that each group S satisfy the 4-allele condition (i.e., $\forall 1\leq j\leq l$ $|\cup_{i \in S} a_{ij} \cup b_{ij}| \leq 4$). This problem was showed to be no 1.0065-approximation algorithm (respectively, 1.00014-approximation algorithm) unless $RP=NP$, when the number of loci is $O(n^3)$ (respectively, 2). In this thesis, we design an $O(n^2l)$-*time* heuristic algorithm to solve the full sibling reconstruction problem and perform the algorithm on some bioinformatics data.

Keywords: partition、The full sibling reconstruction problem、Mendelian inheritance laws、4-allele condition、heuristic algorithm

目 次

表 目 次

圖 目 次

第一章　緒論

本章節主要針對研究背景與動機做概要的介紹，闡明研究的問題與目的，然後描述本研究所探討的的演算法分群問題定義。本章共分為三節：第一節為研究背景與動機；第二節為研究目的；第三節為研究架構流程。

1.1 研究背景與動機

近年來，生物醫學資料的需求與重要性與日俱增，專家學者無不想盡速破解人類的基因密碼，目前生物學家已先行找出許多生物的基因資料片段，但面對龐大的基因資料，要如何判斷將具有親緣關係的物種基因資料放在一起，實為一件棘手的工程，為了讓生物學家或是醫藥研究相關的專家學者能直接獲取想要的基因資料，節省整理時間，便需要資訊相關人員先行在電腦上進行模擬測試。然而如何將一群未知親緣關係的物種分群就需要設計一些方法以及一些分群的條件限制，因此，非常多不同的分群方法紛紛被提出。

不同的限制條件與不同的演算法將會產生不同的分群結果，本研究的研究主要是探討 Berger-Wolf et al.（2005）這篇論文，所用的條件為孟德爾遺傳法則中的 4 對偶 (4-allele) 基因的特性將物種分群，群內的物種都有親緣（兄弟姐妹）關係稱為「a Full Sibling Group」。他們的方法是透過將此問題轉成最小集合涵蓋（Minimum Set Cover）方法（詳見第 2.4 節），把無親代資料的基因，重新分群找出各個具有親緣的體系。因此，本研究便朝向此方向探討一種分群演算法，其條件也是以孟德爾遺傳法則作為分群規則，將未分類的基因進行分群。基因分群問題從上述就得知是個面對龐大基因資料分類非常重要的方法，因此本研究主期望達到降低比對次數，使分群演算法錯誤率降低。

本研究將會常常使用到完全親緣關係（Full Sibling Relationship）與 4 對偶基因這兩名詞，因此先行在這做個定義介紹。以人類來說，是由男女雙方分別提供的基因染色體結合後，誕生所謂的下一代，基因型就會由上一代傳承到下一代，小孩拿到父母各一半的基因型，當他們擁有相同父母中所有特質時就稱為純親緣，亦稱為完全親緣關係（交通大學生物科技結構及細胞生物諮詢網，2000）；而所謂 4-對偶基因特性是從孟德爾遺傳定律可推導出在沒有親代資料樣本時，任兩人都可能是兄弟姐妹關係作為樣本規則。以某性徵為例，假設 A, B 兩人基因型完全不相同，A 控制此性徵的基因型為（1/2, 3/4），B 的基因型為（5/6, 7/8），他們仍有可能為同一對親代之子女。舉例來說若有一對父母的基因型，一個為（1/5, 3/7），另一位是（2/6, 4/8），他們就有可能生出像 A, B 兩基因型完全不同的子代。由此可知，在孟德爾遺傳定律的規則裡，任兩個基因型完全不同的兩人，仍有可能為同一對親代所生。

Berger-Wolf et al.（2005）提出完全親緣關係重建問題（Full Sibling Reconstruction Problem）。給定 n 個物種的集合 N，每個物種 i 有 l 個 Locus，每個 locus 有兩個對偶基因以$<a_{ij},b_{ij}>$，$0<j\leq l$，完全親緣關係重建問題的目的是要將 n 個物種進行分群且每個群必須是 a Full Sibling Group 也就是要滿足 4 對偶基因特性（即 $\forall 1\leq j\leq l \mid \cup_{i \in S} a_{ij} \cup b_{ij} \mid \leq 4$）且群組的個數要最小。他們的方法假設每個物種不需要親代（父母）的樣本的資料即可分群。

本研究正式定義此問題如下：

問題：完全親緣關係重建問題（Full Sibling Reconstruction Problem）

輸入：n 個物種的集合 N，每個物種 i 有 l 個 Locus，每個 Locus 有兩個對偶基因 $<a_{ij},b_{ij}>$，$0<j\leq l$

輸出：n 個物種分成 1 個以上互斥的群

目標：分群個數最少並且滿足 4 對偶基因特性（即 $\forall 1\leq j\leq l \mid \cup_{i \in S} a_{ij} \cup b_{ij} \mid \leq 4$）

底下本研究舉例說明此問題如下：給定 8 個已知基因的物種，每個物種內含有 2 個 Locus，每個 Locus 含有 2 個對偶基因如表 1 所示，本研究將輸入的物種依照 4 對偶基因規則（符合孟德爾遺傳定律之親緣關係的物種分到同一個群集），如將物種（2，3，4，5，6）分成一群，Locus 1 內有 3 種不同的 Alleles = {149, 155, 177}，Locus 2 內有 4 種不同的 Alleles = {245, 253, 267, 277}，則此群組滿足 4 對偶特性，此外將物種（1，7，8）分為一群，locus 1 內有 4 種不同的 Alleles = {149, 151, 167, 173}，Locus 2 內有 3 種不同的 Alleles = {243, 251, 255}同樣也滿足 4 對偶基因特性。因此，此範例將會把這 8 個物種分成兩群（2，3，4，5，6）與（1，7，8），如表 2 所示。

表 1　樣本資料範例針對完全親緣關係重建問題

Animal	**Locus 1　Locus 2** **allelel1/allele2**
1	149/167　243/255
2	149/155　245/267
3	149/177　245/253
4	155/155　253/253
5	149/155　245/267
6	149/155　245/277
7	149/151　251/255
8	149/173　255/255

表 2　最佳的分群結果針對完全親緣關係重建問題

Sibling Groups:
2，3，4，5，6
1，7，8

1.2 研究目的

本研究主要探討在沒有親代資訊下，利用Berger-Wolf et al.（2007）中提及的規則以模擬產生無親代生物基因資料，將這些資料重新找出彼此完全親緣關係的問題。其方法是延伸於使用最小集合涵蓋方法，將它應用於4對偶基因型的分群上，並有學者Berger-Wolf 等人於2005年提出，將完全親緣關係重建問題所用的分群方法。然而所使用的方法在顯示結果上較為緩慢，需進行集合樣本與其他基因型一一比對至全部基因型比對結束，再另起一集合樣本與全部基因型進行比對，並重複此動作，並重複此動作，循環至所有基因型組成的所有可能的集合分群完畢才結束，即不計算個體間的相似度，直接把符合遺傳規則的基因型置於同一群組，依照此方法比對完全部基因型的分群；此方法較直觀也較容易理解，壞處便是分群速度較慢。對於上述分群績效的問題，本研究為使其錯誤率降低求得之分群結果，讓分群速度較佳，經由上述動機，本研究目的為朝向設計一種分群演算法來幫助基因型分群，並觀察其錯誤率與時間複雜度，以模擬生物的基因資料進行分群，藉著分析後的結果來說明本研究方法的執行情況。

1.3 研究流程

本研究回顧與生物資訊相關的背景知識及相關文獻探討後，吸取先前研究者的經驗，並嘗試以相同的孟德爾遺傳法則為基礎，以發展出一套啟發式演算法，並透過測試與驗證，以達到改善其錯誤率與執行效率。本文將分為以下五大部分，完整說明及研究流程，如圖 1 所示。

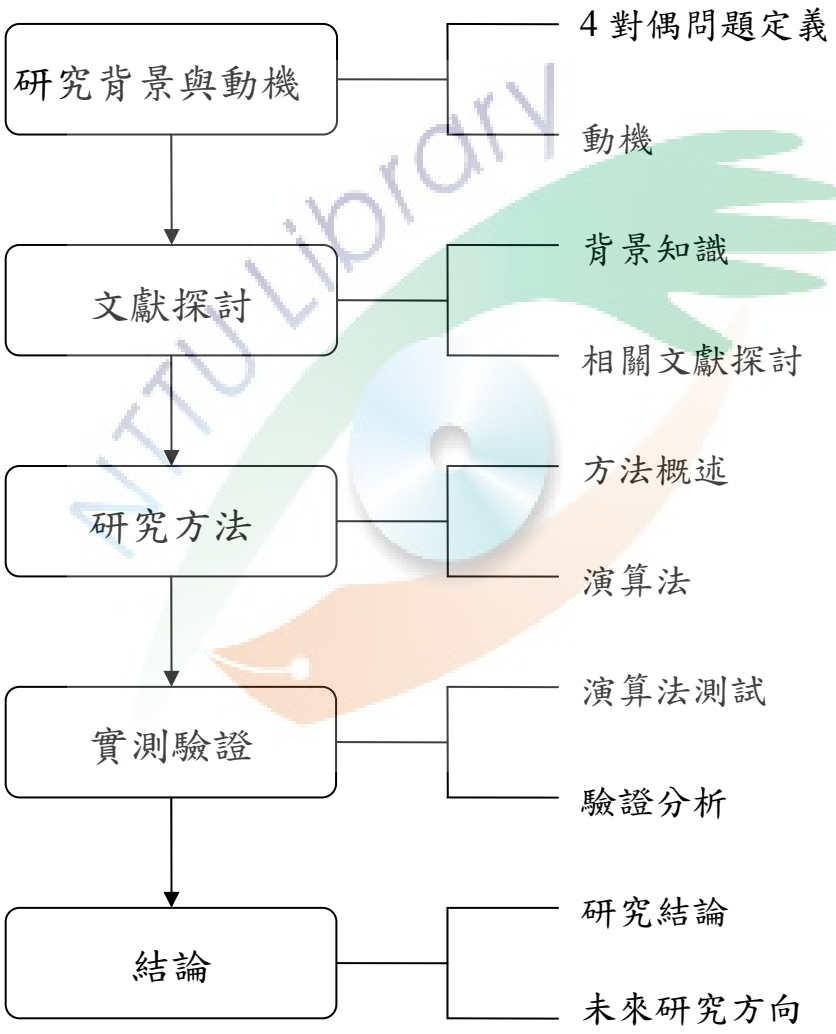

圖 1　研究流程

根據上述流程圖，可將本文之研究內容描述如下：

1、研究背景與動機

本文主要目的在於將未知親代血緣關係的4對偶基因型資料做親緣關係的分群，利用基因分群找出具有親緣關係的演算法。探討4對偶基因的定義及其動機為重建無親代資料的親緣關係，所以在此將提出一個新的分群演算法改善分群錯誤率，再與其他分群方法比較其優缺點。

2、文獻回顧

本研究主要文獻分別為：（1）詞義定義，提供不了解生物詞義的閱讀者作一整理描述，並稍作舉例；（2）針對本研究相關4對偶基因重建兄弟關係的問題，先前生物研究學者提出的一系列分群演算法應用在生物醫學上的論文，並了解各學者提出的方法，以及改善內容與對此課題的分析方法；（3）介紹有關最小集合涵蓋的問題及其應用的領域；（4）4對偶基因的最小集合涵蓋問題，回顧之前學者Berger-Wolf et al.（2005）對於4對偶基因的最小集合涵蓋所使用的方法及其步驟。透過文獻回顧的四個部份來協助了解此研究領域之現況與發展內容，並收集Berger-Wolf et al.（2005）及Chaovalitwongse et al.（2007）所提出之孟德爾法則的約束下，以4對偶基因為規則的演算法應用方式及特色，以助於本研究進行無親代資料的分群演算法。

3、研究方法

依據文獻回顧所得到的資料，參考各學者所運用的演算方法及步驟，確立本研究將採用之演算方法及演算流程，解決分群過程效率問題，並以這方法與最小集合涵蓋方法作為比較，確定此方法能讓分群過程中改善錯誤率。

4、實測驗證

此章節本研究會展示整個實驗數據的模擬結果，藉此觀察本研究的方法執行錯誤率與正確性。

5、結論與未來研究方向

根據前述所驗證之分析及比較結果做出結論，並說明在建立演算法的過程中所遇到之問題，最後提出未來仍須改進的方向。

第二章　文獻回顧

在第二章探討其相關研究中，首先本研究在2.1節將先對本文中常出現的生物資訊相關名詞做一個概略的定義整理。進而，在2.2節中將完全親緣重建的相關問題及一些分群演算法的一系列論文做一些概要性說明。2.3節將介紹有關最小集合涵蓋的問題及其應用的領域。2.4節將針對學者Berger-Wolf 等人於2005年所提出的完全親緣關係重建問題，並回顧其提出之透過轉換到最小集合涵蓋問題來解決此問題的方法及其步驟。

2.1 背景知識

在進入生物資訊領域時，通常會有些專有名詞無法讓大眾能夠在閱讀同時了解或者無法確定自我認知是否正確，所以需要一些淺顯易懂的解釋來方便大家進一步認識生物資訊，畢竟生物資訊與人類解開演化進步的秘密是息息相關。以下為描述常用在生物資訊上的基本名詞（交通大學生物科技結構及細胞生物諮詢網，2000；吳英嬌、吳惠生，1986；中華民國鑑識科學學會，2009）。

醫學領域通常與本研究所要提到的生物有關，不過醫學多半是關心在人體構造、生理疾病等，而在生物領域上則是生命體內部的構造組合，例如細胞（Cell）、病毒（Virus）、基因（Gene）等生命體內可能有關的組合或致命的微小分子，而本研究將會提及所有生物體所擁有的物質，生物便是由基因所構成，簡而言之基因即為一段可傳遞遺傳訊息的序列（Sequence）。它不僅可予以研究專家許多遺傳演化的資訊，使研究專家得以抽絲剝繭般地解開難懂的基因密碼，也可以找出基因突變的遺傳因子。而生命體都是由上一代的兩種基因遺傳所組成的，每個

人的基因都不會完全相同，就算是直系血親也會因為父母親的兩種基因結合而只出現相似的情況，這就是所謂的基因型（Genotype）（交通大學生物科技結構及細胞生物諮詢網，2000）。基因位於染色體上，每條染色體上擁有許多不同的性狀的基因，可用來控制不同的特徵，例如身高、膚色、長相、性別等。每個控制性狀的基因，在染色體上都各自有屬於自己的位置，如同座位般，稱為基因座（Locus），換句話說就是基因的地址。其複數型為Loci（交通大學生物科技結構及細胞生物諮詢網，2000）。

以人類來說，父母分別提供的基因染色體結合後，產生子女，基因型就會由父母遺傳到子女，所以父母與小孩才會具有親緣關係，而小孩則是拿到父母各一半的基因型。若父母生了兩個小孩，則這兩個小孩便是兄弟姐妹（Siblings），當他們擁有相同父母中所有特質時就稱為完全親緣；如果是同父異母或同母異父的則稱為半親緣。在本研究中提及親緣時，本研究只考慮所謂的完全親緣關係（交通大學生物科技結構及細胞生物諮詢網，2000）。

對偶基因（Allele）是控制性狀的一組基因中的其中一個，不同的對偶基因在遺傳特徵上所屬將會不同，如頭髮顏色、血型、膚色或身高等等。細分的話可分成顯性與隱性兩種。顯性對偶基因會比隱性的對偶基因更為明顯。例如控制豌豆莖高矮基因的實驗，呈現出高莖的性狀基因為T，導致矮的為t，T和t雖然都是同種基因，但它們卻所屬不同的對偶基因（交通大學生物科技結構及細胞生物諮詢網，2000）。

基因能控制繁衍的物種形式，所以古人說：「種瓜得瓜，種豆得豆」的意思為不論任何物種都能繁殖出跟自己同種的下一代，而控制繁衍出與自己同種的下一代必須靠基因來控制，就如同上述之父母雙方各傳承部分基因給小孩，而父母就代表著親代，小孩代表子代。根據英漢生物化學辭典（吳英嬌、吳惠生，1986）

中的解釋之，即指「親代是一種細胞它能透過細胞分裂產生出與它相關的細胞，是指一種DNA分子或染色體，透過複製形成與它相關的分子或染色體」一種細胞它能通過細胞分裂產生出與它相關的細胞，亦為一種DNA分子或染色體，通過複製形成與它相關的分子或染色體，即所謂的”親代”。中華民國鑑識科學學會（2009）的網站上定義如何判定親緣關係，則是利用了生物科技對生物樣品分析其遺傳標記以研判其親緣關係。無親代資料，即為只有子代的基因資料，父母親的基因資料可能因遺失等因素而無法得知。所以本研究主要探討在無親代的資料下，利用大量的基因樣本分門別類地找出具有血緣（親緣）關係的兄弟姐妹的組合精確度問題，目的為降低分群組數。

2.2 一些親緣關係重建問題的分群演算法

分群是一種十分廣泛被應用到各個領域的問題，也是一個非常需要深入研究的問題。分群定義為把一堆未曾分類過的資料，經由分群演算法將資料分門別類。常見的分群方式有兩種，一種為 Partition-Based（分割式）分群；一種為階層式分群。給定一集合 N 內包含 n 個元素，分割式分群指的是將集合中的資料分成兩個或更多個互斥的群（Cluster）（N 的非空子集），也就是每個元素必須屬於某一群集（Han and Kamber, 2006; Tan, Steinbach and Kumar, 2006）。階層式的分群方式則是將已給定的集合建構成一巢狀結構的群集。

在生物資訊與作業研究的領域上，有相當多的分群演算法已被提出來重建兄弟關係的群組。Almudevar and Field （1999）提出一個分群演算法利用 Minimal Sibling Groups Under Likelihood 方法，透過 DNA Markers 將有親緣的放在同一群裡面，求最少可以分成的群數，他們所採用的資料是有關漁業養殖的資料。Konovalov, Manning and Henshaw （2004）則利用 java 程式實做一個新的分群方法稱為 KinGroup，評估所有可能的分割情況後，重新建構物種演化關係。Beyer and May （2003）所提出的圖形理論的分割法（Partition Population using Likelihood Graphs）求物種的親緣分群。Wang（2004）提出的 Simulated Annealing（模擬退火法）重建親緣關係並實作成軟體套件 COLONY。Fernández and Toro（2006）也提出 Blind Search Algorithm 類似 Simulated Annealing 找出家譜中高關聯性的共同祖先矩陣。

Berger-Wolf et al.（2005）及 Chaovalitwongse et al.（2007）提出了以孟德爾遺傳法則為基礎，在一群物種中重建親緣關係稱為親緣關係重建問題（Full Sibling Reconstruction Problem）其分群限制為滿足 4 對偶特性。他們將此問題轉

換到最小集合涵蓋（Minimum Set Cover, MSC）問題後利用 Greedy 的方法直接把符合遺傳規則的分在同一群（見 2.4 節），可在無親代資料下重建整個親緣體系。其缺點是分群速率不好（考慮性徵越多越慢）。而他們採取使用 MSC 的原因是因為當要能涵蓋全部個體，也就是把相似度最高的放在一起，使分群最少。所謂的 4 對偶基因則是一個群組內的物種每個 Locus 的所有對偶基因不超過 4 種，這是一種常見於辨別是否為親緣關係的方法。Sheikh et al.（2007）與 Berger-Wolf et al.（2007）另外提出一種分群的方法，原理則是利用微衛星體資料（Microsatellite Data）統計和啟發式技術（Heuristic Techniques）來重建親緣關係，並且必須仰賴於事先瞭解的各種基因特徵。上述這兩種方法是專為大量基因樣本和小家族群式的資料而設計，野生群種並不屬於這兩種。所以他們在此也提出使用 2 對偶基因的最小集合涵蓋技術這是利用孟德爾遺傳定律演化來重建親緣關係群組。給定 n 個物種的集合 N，每個物種 i 有 l 個 Locus，每個 Locus 有兩個對偶基因以$<a_{ij},b_{ij}>$，$0<j\leq 1$，滿足 2 對偶基的限制使指一個群組內的每個 locus 的每一個對偶基因它的不同個數最多是 2（即 $\forall 1\leq j\leq 1 \quad |\cup_{i\in S} a_{ij}|\leq 2$ and $|\cup_{i\in S} b_{ij}|\leq 2$）。很顯然的 2 對偶基的限制遠比 4 對偶基因限制來的嚴格。Sheikh, et al.（2007）與 Berger-Wolf et al.（2007）提出滿足 2 對偶基因限制的完全親緣關係重建問題並將它轉換到最小集合涵蓋問題同時使用模擬和實際的物種（小蘿蔔、鮭魚及小蝦）來進行比較和驗證其方法。Sheikh et al.（2008）提出一些一致性 consensus 的方法來重建親緣關係，利用基因數據來分析親緣關係在許多生物學上是非常重要的，其中包括在許多保護生物學與遺傳學上，並使用了 Strict Consensus、Voting Consensus、Majority Consensus 等三種不同的一致性技術來解決親緣關係重建問題。Sheikh et al.（2009）認為用微衛星體（Microsatellites）資料重建完全親緣關係是一個很好的研究，但他們發現半親緣關係重建較少人進行研究，進而在理論上為其關鍵點。因而提出不同構想的半親緣重建問題，並證明此問題為 NP-Hard，並為此問題設計其啟發式演算法，利用生物和模擬的資料集進行分析實驗，然後與先前的軟體 COLONY （Wang, 2004）進行比較。Ashley,

et al.[b]（2009）設計了一個名為 KINALYZER 軟體利用顯性標記基因座資料重建無父母資訊的完全親緣群組，如微衛星體。並採用新的演算法來重建親緣關係，此問題的基本假設仍然是以孟德爾遺傳法則為基礎並利用最小 2 對偶基因集合涵蓋法找到最少的親緣群組。這是一個「Greedy Consensus」的方法，可以重建親緣群組的基因座子集合並找出他們的部份一致性。Ashley et al.（2009）他們證明了親緣關係重建問題法設計出多項式時間的 1.0065 倍（當 Locus 的數目為 $O(n^3)$）及 1.00014 倍（當 Locus 的數目為 *2*）的近似演算法除非 *RP=NP*，*n* 為物種的個數。之後 Ashley et al.（2010）的論文 Survey（收集）了目前的全親緣重建的微衛星遺傳標記方法。接著介紹以孟德爾法則和其延伸出即使是以錯誤數據為基礎的親緣關係重建新組合問題並提出一些相關演算法。本研究將利用表 3 作一相關研究的統整。

表 3　相關親緣關係的分群法研究整理

作者	發表時間	內容概述
Almudevar and Field	1999 年	提出一個分群演算法利用 Minimal Sibling Groups Under Likelihood 的方法，將有親緣的放在同一群裡面，求最少可以分成的群數，採用的是有關漁業養殖的資料。
Beyer and May	2003 年	提出圖形理論的分割法（Partition Population using Likelihood Graphs）求物種的親緣分群，且使用單基因座共顯性標記 Single-Locus co-dominant marker。
Konovalov, Manning and Henshaw	2004 年	利用 java 程式實做一個新的分群方法稱為 KinGroup 用來重新建構物種演化關係。
Wang	2004 年	Simulated Annealing（模擬退火法）重建親緣關係並實作成軟體套件 COLONY。
Berger-Wolf et al.	2005 年	提出親緣關係重建問題（Full Sibling Reconstruction Problem），利用最小集合涵蓋法（Minimum Set Cover, MSC）與 4 對偶限制，在無親代資料下可重建整個親緣體系。
Fernández and Toro	2006 年	提出 Blind Search Algorithm 類似 Simulated Annealing 找出家譜中高關聯性的共同祖先矩陣（matrix）。
Berger-Wolf et al.	2007 年	提出滿足 2 對偶基因限制的完全親緣關係重建問題並將它轉換到最小集合涵蓋問題，並利用 2 對偶基因的最小集合涵蓋的技術進行分群，同時使用模擬和實際的物種（小蘿蔔、鮭魚及小蝦）來進行比較和驗證其方法。
Chaovalitwongse et al.	2007 年	利用最小集合涵蓋的演算法解決親緣關係重建問題（Berger-Wolf et al. 2005 的期刊版本）。
Sheikh et al.	2007 年	為了親緣關係重建提出滿足 2-對偶基因限制的完全親緣關係重建之組合最佳化問題，並將此問

作者	發表時間	內容概述
		題轉換到最小集合涵蓋問題且利用 2 對偶基因的最小集合涵蓋的技術進行分群。（Berger-Wolf et al. 2007 的會議版本）
Sheikh et al.	2008 年	研究了利用 Consensus 的方法來重建親緣關係，並且討論了不同的一致性（Consensus）方法。
Ashley et al.[a]	2009 年	證明了親緣關係重建問題（Full Sibling Reconstruction Problem）無法設計出多項式時間的 1.0065 倍(當 locus 的數目為 $O(n^3)$)及 1.00014 倍（當 Locus 的數目為 *2*）的近似演算法，除非 *RP=NP*，*n* 為物種的個數。
Ashley et al.[b]	2009 年	提出KINALYZER軟體利用顯性標記基因座資料重建無父母資訊的完全親緣群組，利用 2 對偶基因的最小集合涵蓋法找到最少的親緣群組。
Sheikh et al.	2009 年	進行半親緣關係重建研究，利用生物和模擬的資料集，他們提出實驗結果，設計半親緣的最小集合涵蓋問題的演算法並與先前的軟體 COLONY 進行比較。
Ashley et al.	2010 年	Survey 目前的全親緣重建的微衛星遺傳標記方法。接著介紹以孟德爾法則和其延伸出即使是以錯誤數據為基礎的親緣關係重建的新組合問題並提出一些相關演算法。

(本研究整理)

2.3 最小集合涵蓋（Minimum Set Cover）

最小集合涵蓋（Minimum Set Cover, MSC）問題為一傳統且重要的最佳化組合問題。給定一集合 U 內有 n 個元素$\{u_1,u_2,\ldots,u_n\}$，另有 k 個集合$\{S_1,S_2,\ldots,S_k\}$，每個集合包含部份 U 內的元素。最小集合涵蓋問題是在 k 個集合內找出最少的集合且可以包含所有 U 內的元素。此問題有被廣泛的應用在許多領域上，例如生物資訊（Doi and Imai, 1997, 1999; Duitama et al., 2009; Huang et al., 2005; Jabado et al., 2006; Zhao et al., 2005）與作業研究上（Beasley, 1990; Caprara et al., 1999, 2000; Drezner and Hamacher, 2002; Farahani and Hekmatfar, 2009; Lan et al., 2007）。

在生物資訊方面， Berger-Wolf et al.（2005）提出了以孟德爾遺傳法則為基礎的物種分群問題即是轉換到利用最小集合涵蓋問題上之後透過 MSC 的 Greedy 的演算法解決此問題。引子對的選取問題定義為給定 k 個引子（Sets），n 個目標序列（Elements），要找出最少的引子可以釣出（Hybridize）所有 n 個目標基因。這個問題也是可被 Model 到集合涵蓋問題並透過一些集合涵蓋的演算法來解決並應用在疾病的偵測上（Doi and Imai, 1997, 1999; Duitama et al., 2009; Huang et al., 2005; Jabado et al., 2006）。Zhao et al.（2005）定義了最小化 siRNA 涵蓋問題。siRNA 為 RNA 內的一子字串可以干擾 RNA 來抑制基因的表現（RNAi）。給定 k 個 siRNA（小干擾 RNA），α個目標基因（Target Genes）序列與β個非目標基因（Off-Target Genes）序列的集合。某個 siRNA 是某個目標基因（或非目標基因）序列的子字串，則此 siRNA 涵蓋（Cover）這個目標基因（或非目標基因）。此問題是要找出最少的 siRNA（Sets）可以涵蓋（Cover）所有的目標基因（Elements）但是不能涵蓋任何一個非目標基因。也就是說找出的這些 siRNA 可抑制所有的目標基因。在不考慮非目標基因的部份，此問題同樣可轉換至最小集合涵蓋問題

上。Zhao et al.（2005）設計了一個 Exact Branch-and-Bound Algorithm 及 Probabilistic Greedy Algorithm 解決此問題並模擬到兩個基因家族（FREP Genes from Biomphalaria Glabrata and the Olfactory Genes from Caenorhabditiselegans）。

在作業研究方面，在此領域通常是將集合涵蓋問題應用至遞送（Delivery）和路由（Route）問題上。最有名的問題為空服員排班（Airline Crew Scheduling）問題。此問題假設本研究有 *n* 個航線（Elements）及 *k* 個空服員，每個空服員被要求要服務某些航線的行程（Sets），為找出最少空服員的行程可以服務所有的航線。Caprara et al.（1999）將此問題轉到最小集合涵蓋問題上用來解決 Italian Railway Company, Ferrovie Dello Stato SPA 的排班，並提出 Lagrangian-Based Heuristic Algorithm。Caprara et al.（2000）之後提出一篇 Survey 文獻來比較多個正確（Exact）及啟發式（Heuristic）的演算法。最近 Lan et al.（2007）提出一簡單的啟發式演算法來解最小集合涵蓋問題並用來測試 OR-Library （Beasley, 1990）。另外，一些作業研究上探討的設備配置問題（Facility Location Problem）也可以被轉換到最小集合涵蓋問題，並有很多演算法用來設計解決這些問題（Drezner and Hamacher, 2002; Farahani and Hekmatfar, 2009）。

然而，最小集合涵蓋問題已知為 NP-Complete （Garey and Johnson, 1979）。因此，目前為找到多項式時間的正確演算法（Exact Algorithm）是不可能的，除非 P=NP。因此，許多的近似演算法（Approximation Algorithms）（Cormen et al., 2001; Goldsmidt, 1993; Johnson, 1974; Lan et al., 2007; Lovász, 1975; Slavík, 1997）、隨機演算法（Randomized Algorithms）（Haouari, 2002; Vazirani, 2001）、啟發式演算法（Heuristic Algorithms） （Beasley, 1990; Caprara et al. 1999, 2000; Ceria, 1998; Lan, 2007）與適當指數時間的近似演算法（Moderately Exponential Approximation Algorithms）（Bourgeois et al. 2009; Cygan et al. 2009） 有被大量的提出。另外也有非常多的指數時間的正確演算法被提出（Balas and Carrera,

1996; Beasley and Ornsten, 1992; Björklund, 2009; Caprara et al., 2000; Fomin et al., 2005; Mannino and Sassano, 1995; Rooij and Bodlaender, 2008），上述方法多半是利用 Branch-and-Bound 或 Branch-and-Cut （Balas and Carrera, 1996; Beasley and Ornsten, 1992; Fomin et al., 2005; Mannino and Sassano, 1995; Rooij and Bodlaender, 2008）。在近似演算法方面，Greedy 的演算法可找到近似率為 H(n)（Harmonic number）（Johnson, 1974; Vazirani, 2001），實際的 Tight 為 $\ln n - \ln \ln n + \ln 2 - 1$，也是目前最小集合涵蓋問題最好的近似演算法（Slavík, 1997）。也被證明近似率不會低於$(1-\varepsilon) \ln n$，$\varepsilon > 0$，除非 $NP \subset DTIME(n^{\log \log n})$ （Feige, 1998）。

2.4 完全親緣關係重建問題轉換到最小集合涵蓋問題

完全親緣關係重建問題（Full Sibling Reconstruction Problem）是由 Berger-Wolf 等人於 2005 年提出，基於孟德爾遺傳法則作為理論基礎，重建完全親緣關係。而每個親緣群組的集合一定要滿足 4 對偶基因特性，所謂 4 對偶基因特性是從孟德爾遺傳定律可推導出在沒有親代資料樣本時，任兩人都可能是兄弟姐妹關係作為樣本規則。假設 A、B 兩人的基因型完全不相同，但因為父母基因本身就不可能相同，所以還是有可能生出像 A 與 B 兩個基因型完全不同的子代。由此可知，在孟德爾遺傳定律的規則裡，任兩個基因型完全不同的兩人，仍有可能為同一對親代所生。

經由上述的理由，Berger-Wolf et al.（2005）提出完全親緣關係重建問題，將 n 個物種分群使其達到滿足孟德爾遺傳法則中 4 對偶基因之限制且分群的個數要最少，每個物種 i 有 l 個 Locus，每個 Locus 有兩個對偶基因 $<a_{ij},b_{ij}>$，$0<j\leq l$，他們將此問題轉到最小集合涵蓋問題上。他們的作法概念是先找出所有 $O(n^2)$ 集合，每個集合會滿足 4 對偶基因特性（每個集合內有 j 個 Locus，每個 Locus 最多 4 種對偶基因）。集合的產生可以針對每一對（兩個）物種每個 Locus 最多 4 種對偶基因，把這些 Locus 的對偶基因（最多 4 個）視做一個集合。之後某物種隸屬某個集合是指如果此物種的每個 Locus 內的兩個 Alleles 出現此集合內的話。目標是要找出最少的集合可以 Cover 所有的物種。如此即可轉換到最小集合涵蓋問題上，之後再透過最小集合涵蓋問題的的演算法將物種進行分群，所得之結果。這裡的最小集合涵蓋演算法是透過轉換到整數規劃來解。

本研究描述此演算法如下：

1. 建立一個合理的集合 $S_1, S_2,.....,S_m$，每個集合滿足 4 對偶基因的限制。作法為任兩個物種即可產生一個集合。
2. 每一個物種 x 放入到集合 S_i 只有在此物種的每個基因座的對偶基因符合 S_i 內相對基因座的對偶基因時。
3. 針對步驟二中所產生的集合，利用整數規劃的最小集合涵蓋演算法找出最小集合涵蓋，而這些集合即是此問題的結果。

整數規劃的最小集合涵蓋演算法是假設有 n 個元素，m 個集合，s_j 為第 j 個集合，每個集合變成一個變數 x_i，當第 i 個集合被挑選時 x_i 設為 1 否則為 0。底下我們列出最小集合涵蓋問題轉到整數規劃的公式：

$$\text{Min} \sum_{i=1}^{m} x_i$$

$$AX \geq 1$$

$$x_i \in \{0,1\}$$

A 為一個 n*m 的元素-集合矩陣 $a_{ij} = \begin{cases} 1 & if \quad i \in s_j \\ 0 & otherwise \end{cases}$

X 為一個 m*1 的矩陣 $\begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_m \end{bmatrix}$

以下將舉例說明此演算法。假設本研究有 A、B、C、D 四個物種，每個物種有兩個 Locus，每個 Locus 有兩個對偶基因，如表 4 所示。步驟一本研究將可產生如表 5 所示之六個集合，分別為 $Set_{A,B}$、$Set_{A,C}$ 、$Set_{A,D}$ 、$Set_{B,C}$ 、$Set_{B,D}$ 、

$Set_{C,D}$，每個集合的每個 locus 都滿足 4 對偶基因的限制。依照步驟二將物種 A、B、C、D 分別放入此六個集合中。例如集合 $Set_{A,B}$ 內有 A，B，C 三個物種。執行完步驟二後，每個集合的內容將為 $Set_{A,B}$ ={A,B,C}，$Set_{A,C}$ ={A,C}，$Set_{A,D}$ ={A,D}，$Set_{B,C}$ ={B,C}，$Set_{B,D}$ ={B,D}，$Set_{C,D}$ ={C,D}。將 $Set_{A,B}$ 設為 x_1，$Set_{A,C}$ 設為 x_2，$Set_{A,D}$ 設為 x_3，$Set_{B,C}$ 設為 x_4，$Set_{B,D}$ 設為 x_5，$Set_{C,D}$ 設為 x_6，之後利用整數規劃的最小集合涵蓋演算法法，(見表 6)，將可產生兩個集合為 $Set_{A,B}$，$Set_{B,D}$。這兩個集合將為此演算法之分群結果。

表 4　樣本資料範例針對最小 4 對偶基因集合涵蓋的演算法

	Locus 1	**Locus2**
物種 A	1/1	2/3
物種 B	1/4	5/6
物種 C	1/4	2/6
物種 D	7/8	9/6

表 5　物種的集合

	Locus1	**Locus2**
$\text{Set}_{A,B}$	{1,4}	{2,3,5,6}
$\text{Set}_{A,C}$	{1,4}	{2,3,6}
$\text{Set}_{A,D}$	{1,7,8}	{2,3,6,9}
$\text{Set}_{B,C}$	{1,4}	{2,5,6}
$\text{Set}_{B,D}$	{1.4,7,8}	{5,6,9}
$\text{Set}_{C,D}$	{1,4,7,8}	{2,6,9}

表 6　整數規劃的最小集合涵蓋演算法

$$\text{Min} \sum_{i=1}^{6} x_i$$

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} \geq \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$x_i \in \{0, 1\}$$

第三章 研究方法

本章中，3.1 節概述本研究將預計解決的完全親緣關係重建問題，以及所使用的分群方法概述。隨後本研究提出的演算法將在 3.2 節作說明。

3.1 方法概述

Berger-Wolf et al.（2005）演算法中，是將一筆未分類的基因資料中挑一個做為被比對的樣本，之後依序資料內的基因放入樣本比對是否符合 4 對偶基因。而本研究是從未分類的基因資料內，隨機挑出兩筆資料放在同一個群集。原因是應用孟德爾遺傳法則其中的一個特性推論出任兩人都可能是兄弟姐妹關係。由此兩筆資料之基因型，可建立符合此親緣群集的規則。然後再依序將樣本放入此群集進行比對，若此樣本之基因型符合此親緣群集的規則，則可與樣本放在同一個群集，若不符合則留在未分類的基因資料中，反覆進行此動作直至分群完畢。

步驟如下：

1. 首先將未分類的物種基因資料放在未分類的群組中。
2. 從未分類的物種隨機挑兩物種，建立一開始的群組比對樣本。
3. 對於每筆基因資料，若可以加入這新群組（也就是需要滿足 4 對偶基因的限制），便可加入此群組，否則便留在未分類群組中。
4. 重複第二項與第三項步驟，直到所有物種分類完畢。

3.2 演算法

演算法如下：

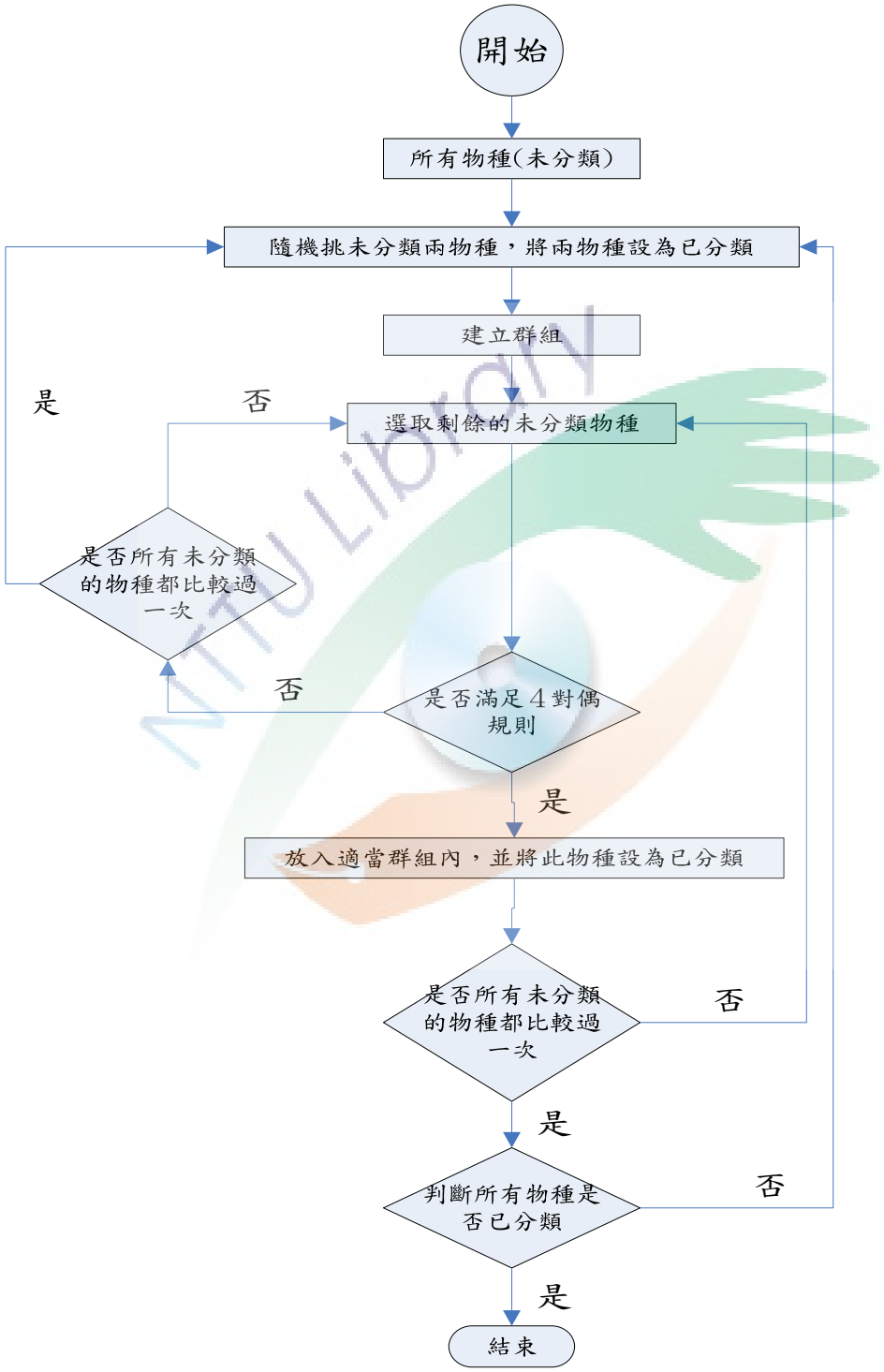

圖 2　演算法流程圖

根據上述演算法流程圖，可將本文之演算法敘述如下：

步驟一：將未分類的物種放在未分類的群組中，即是所有物種。

步驟二：在所有物種中，隨機挑出兩個物種。

步驟三：將挑出的兩物種建立為一個群組。

步驟四：選取兩物種以外的物種進入此群組進行比對。

步驟五：判斷加入的物種是否滿足 4-對偶規則。

若滿足 4-對偶規則，則跳入步驟六。

否則跳回步驟四，將未比對過的物種依序加入群組比對；若無剩餘未比對之物種，則跳入步驟八。

步驟六：將滿足 4-對偶規則的物種放入群組。

步驟七：判斷是否將所有物種比對完畢。

若比對完畢，則跳入步驟八。

否則跳回步驟五，將尚未比對過的物種放入是否滿足 4-對偶規則中。

步驟八：判斷是否每個物種皆有分到群組。

若每個物種皆有群組，則結束。

否則跳回步驟二，將剩餘未分到群組的物種隨機挑出兩物種建立群組。

本研究所提出的演算法是屬於啟發式的演算法。近年來觀察自然界的生物行為，獲得概念來設計求解啟發式演算法的精確性問題已成為熱門的研究領域，而這類型的演算法概念不僅是模仿自然界行為，更是藉由觀察自然界所獲得的靈感，一般大家所聽過的例子有：基因演算法是源自於生物演化所提出、Ant 演算法是透過蟻群的覓食行為、粒子演算法為效法鳥類覓食行為、模擬退火演算法是藉由觀察固體加熱後控制降溫使分子重新排列等...。由於這些演算法非常的具有彈性，可用於求解許多不同問題，作法簡單、求解效率高，因此目前已成為數量方法中最為熱門的方法之一。這類型的演算法他們雖然可以應用在非常廣泛的問

題上，也能求出不錯的解，但是沒辦法確定它最壞的情況；它通常也可在合理的時間內解出答案，但也沒辦法保證它是否每次都能以同樣的速度求解。

本研究演算法的設計構想是來自於 Berger-Wolf et al.（2005）的最小集合涵蓋法。透過了解最小集合涵蓋法後，可發現逐步作基因比對的效果很緩慢，循序比對比完後才能重啟新的一個樣本繼續與其他基因型進行的比對。從這個方法可以很直覺的想到，若事先取兩個成一群當樣本，再與其他基因進行比對，其速度上的增加本研究將透過實驗來進行驗證。任兩個物種事先成一群作樣本的目的在於加快分群比對的速度。

如下舉例說明本研究的演算法步驟。首先給定一群物種 $N=\{1, 2, 3, 4, 5, 6, 7, 8\}$，每一個物種的基因具有兩個基因座，每個基因座含有兩個對偶基因，如表 6 所示。開始對 N 進行分群，隨機抽取兩物種 $S=\{2, 3\}$得到一分群，將其他未分類物種放入此進行比對（需滿足 4 對偶基因的限制），若無法足特性，便回到 N。如表 7 所示為此種分群也是最佳的分群結果。

1. 八個物種 $N=\{1, 2, 3, 4, 5, 6, 7, 8\}$。
2. 基於孟德爾遺傳法則，隨機將任兩樣本當作同一個親緣群集 $S_1=\{2, 3\}$。
3. 把其餘樣本 $N=\{1, 4, 5, 6, 7, 8\}$一個一個放進親緣群集 S_1 進行比對，能夠滿足 4 對偶基因的限制則留在親緣群集 S_1，而最佳分群情況則是 $S_1=\{2, 3, 4, 5, 6\}$，能以最少的群數，容納最多的樣本。
4. 而未分類的樣本 N 中尚有$\{1, 7, 8\}$，將任兩樣本作為親緣群集 $S_2=\{7, 8\}$，將剩下的樣本 1 放入親緣群集 S_2比對，滿足 4 對偶基因的限制則留在親緣群集 S_2，輸出產成兩個分群 $S_1=\{2, 3, 4, 5, 6\}$、$S_2=\{1, 7, 8\}$。

表 7　針對本研究演算法之樣本資料範例

Animal	**Locus 1　Locus 2** **allelel1/allele2**
1	149/167　243/255
2	149/155　245/267
3	149/177　245/253
4	155/155　253/253
5	149/155　245/267
6	149/155　245/277
7	149/151　251/255
8	149/173　255/255

表 8　本研究演算法之分群結果

Sibling Groups:
2，3，4，5，6
1，7，8

3.3 演算法的時間複雜度分析

因為 Ashley et al. （2009）他們證明了完全親緣關係重建問題法設計出多項式時間的 1.0065 倍（當 Locus 的數目為 $O(n^3)$）及 1.00014 倍（當 Locus 的數目為 *2*）的近似演算法除非 *RP=NP*，*n* 為物種的個數。所以目前無法在多項式的時間複雜度下找到此兩問題的最佳解（optimal solution）。我們的啟發是演算法是可以在 $O(n^2l)$ time 找到一組合理的解（feasible solution），*l* 為每個物種有 *l* 個 locus。底下我們分析第二節的演算法的時間複雜度。

演算法如下:

步驟一：將未分類的物種基因資料放在未分類的群組中。

步驟二：在所有物種中，隨機挑出兩個物種。

步驟三：將挑出的兩物種建立為一個群組。

步驟四：選取兩物種以外的物種進入此群組進行比對。

步驟五：判斷加入的物種是否滿足 4 對偶規則。

若滿足 4 對偶規則，則跳入步驟六。

否則跳回步驟四，將未比對過的物種依序加入群組比對；若無剩餘未比對之物種，則跳入步驟八。

步驟六：將滿足 4 對偶規則的物種放入群組。

步驟七：判斷是否將所有物種比對完畢。

若比對完畢，則跳入步驟八。

否則跳回步驟五，將尚未比對過的物種放入是否滿足 4 對偶規則中。

步驟八：判斷是否每個物種皆有分到群組。

若每個物種皆有群組，則結束。

否則跳回步驟二，將剩餘未分到群組的物種隨機挑出兩物種建立群組。

在此演算法中，第一步驟，一開始 n 個物種都沒分類所以只要 O(1) time。在第二步驟跟步驟三，隨機挑兩個只要在常數時間即可完成。判斷加入的物種是否滿足 4 對偶特性需要花 $O(l)$的時間，因為我們有 l 個 locus，所以步驟四到七需要花 $O(nl)$的時間，因為最多可能會有 $O(n)$個群組。因為步驟八需要將還沒分類的物種都要分好類，所以我們的方法將可在 $O(n^2l)$ time 完成。

第四章 實測驗證

本章中，針對本研究所設計的演算法與未知親緣關係之下，利用分群法重建問題求得較低的分群錯誤率來進行實際的程式模擬，因為完全親緣關係重建問題（full sibling reconstruction problem）在目前為止，還未有較佳的分群速率被提出，所以採用 Berger-Wolf et al.（2005）提出的最小集合涵蓋演算法來與本研究的演算法進行比較。因為親緣重建問題是被轉換到使用最小集合涵蓋法來分群，所以針對完全親緣重建問題，實作本研究提出之演算法的程式。此演算法的時間複雜度為 $O(n^2l)$ time。雖然實作的結果顯示，使用的方法分群精確性較差，然而本研究的方法其分群時間卻有比最佳的最小集合涵蓋演算法快的多。此程式是執行在相同的電腦上。CPU 型態：Pentium(R) Dual-Core CPU T4200 2.00GHz，作業系統：Microsoft Windows XP。

以 n 個物種為一個未分類的集合 N。對於每個物種 i 有 l 個 Locus，每個 Locus 有兩個對偶基因 $<a_{ij},b_{ij}>$，而 n 個物種可分成 1 個以上互斥的群。使用的資料是從 Berger-Wolf et al.（2005）的論文中所提出的與 4 對偶基因參數範圍來進行資料模擬。在接下來的第一節中，利用上述論文所提及的模擬資料進行實測本研究的演算法，並觀察分群後的錯誤率。錯誤率我們是透過觀察這些物種的最佳分群結果(透過 Berger-Wolf et al.（2005）的整數規劃最小涵蓋演算法)與我們的演算法出來的分群的結果之間的距離，距離我們是透過 Gusfeld（2002）的分群距離(partition-distance problem)來計算的，第一節我們簡單的敘述此分群距離的算法。第二節我們對我們的演算法進行程式實驗模擬，算出完全親緣關係重建問題的分群的結果，並觀察錯誤率。

4.1 partition-distance 問題

針對 *n* 個元素，給定兩個不同的分群結果 P_1 有 *i* 個群$\{C_{1,1},C_{1,2},\ldots,C_{1,i}\}$及 P_2 有 *j* 個群$\{C_{2,1},C_{2,2},\ldots,C_{2,j}\}$及，我們說 P_1 及 P_2 是一*致(identical)*定義為在 P_1 中的每個群都會在 P_2 中對應到一個裡面元素完全相同的群(反之亦然)。Almudevar and Field （1999）定義的分群距離函數 d_A：$d_A(P_1,P_2)$-> R^+ ，為移除最少的元素使得移除後兩個分割會變成一致的，移除的個數則為 P_1，P_2 間的距離 。舉例如下：給定 9 個元素 {1, 2, 3, 4, 5, 6, 7, 8, 9 }，對這些元素進行兩種分群演算法，假設得到 P_1 分成 3 群為{(1,2), (4,5,6,7), (3,8,9)}。P_2 也分成 3 群為{(1,2,4,5), (8,9), (3,6,7)}。我們移除元素 {3, 4, 5}使得 P_1 和 P_2 變成一致，因此 $d_A(P_1,P_2)$ 是 3。Gusfield（2002）提出一個 $O((i+j)^3)$的演算法，轉換計算兩個分群的距離到最大分配（maximum weighted assignment）的問題上。我們的研究中，將使用此分群距離來當做我們的錯誤率。

4.2 實測案例

本節中，實測案例的分群演算法資料是取自於 Berger-Wolf et al.（2005）中提及的規則所得到的資料，建構的資料是依照此論文規則所模擬產生的，使用模擬資料進行分群演算法時，針對此模擬資料進行隨機抽取兩筆作為一開始的分群樣本，其餘資料依序放入此群組作比對。Berger-Wolf et al.（2005）設定有 F 對成年，每對會產生至少一個小孩(juveniles)，然後對這些小孩來進行分群。假設有 F 個女性物種與 M 個男性物種，每個親代生物種都有 *l* 個 locus，每個 locus 最多有 a 個 alleles。接下來亂數產生 jF 物種（小孩），進行分類。對於每個物種的每個 locus 的 alleles 是從男性物種亂數取一個 alleles，女性物種處亂數取一個 alleles，總共會取 *2l* 個 alleles。每對親代生物種最多產生 o 對小孩（此 o 個小孩都是從相同的親代生物種亂數產生的），底下我們列出參數的設定：

- 生物種的女性(母) F=10，男性(公)M=10，十對親代生物種。
- Locus 的個數 *l*=2; 4; 6; 10。
- 每個 locus 最多幾個不同的 alleles a= 2; 5; 10; 20。
- 小孩（分類的物種，juveniles）的個數參數 j＝10;20;50;100。
- 每對親代生物種最多產生小孩個數（offspring）o=2;5;10;30;50。

本研究將實作我們所提出的啟發式演算法。這個程式是透過 JAVA 程式碼所完成的，針對完全親緣關係重建的 4 對偶基因分群問題，並計算錯誤率。本研究可以從此程式的執行結果清楚的得知執行的速度與分群結果。在程式的部份，使用 Eclipse 開發工具來撰寫程式碼，從外部將模擬的基因資料，如下圖 3 所示之文字檔（數據 1.txt）匯入程式部分，進行資料分群。程式編輯器 Eclipse 介面，下如圖 4 所示分為三個部份：(1) 左邊部分為建立了 Java 專案後，會出現整個套件資料夾。(2) 右邊部分為程式碼撰寫地方。(3) 中間主控台顯示的是本研究

執行程式碼後，顯示出分群的結果。本研究提出之演算法會依照外部資料的多寡進行隨機抽取比對樣本，進行分群，只要將所要匯入的資料文字檔放進套件資料夾後，按下執行，即可在主控台看到程式執行結果。本研究使用了兩種演算法（1）針對完全親緣關係重建的 4 對偶基因問題的啟發式演算法（2）完全親緣關係重建問題的最佳分群演算法，此演算法是採用 Berger-Wolf et al.（2005）提出將此問轉到最小集合涵蓋問題上，並透過整數規劃完成，整數規劃部份我們是透過 Matlab 7 來實作。之後將這兩個演算法分群後的結果進行比較，並計算錯誤率（錯誤率是最佳的分群結果與我們的演算法所產生的結果之間的分群距離除以所有的物種數，分群距離計算可參考 4.1 節）。底下我們利用圖 5 到圖 10 顯示我們的實驗結果。

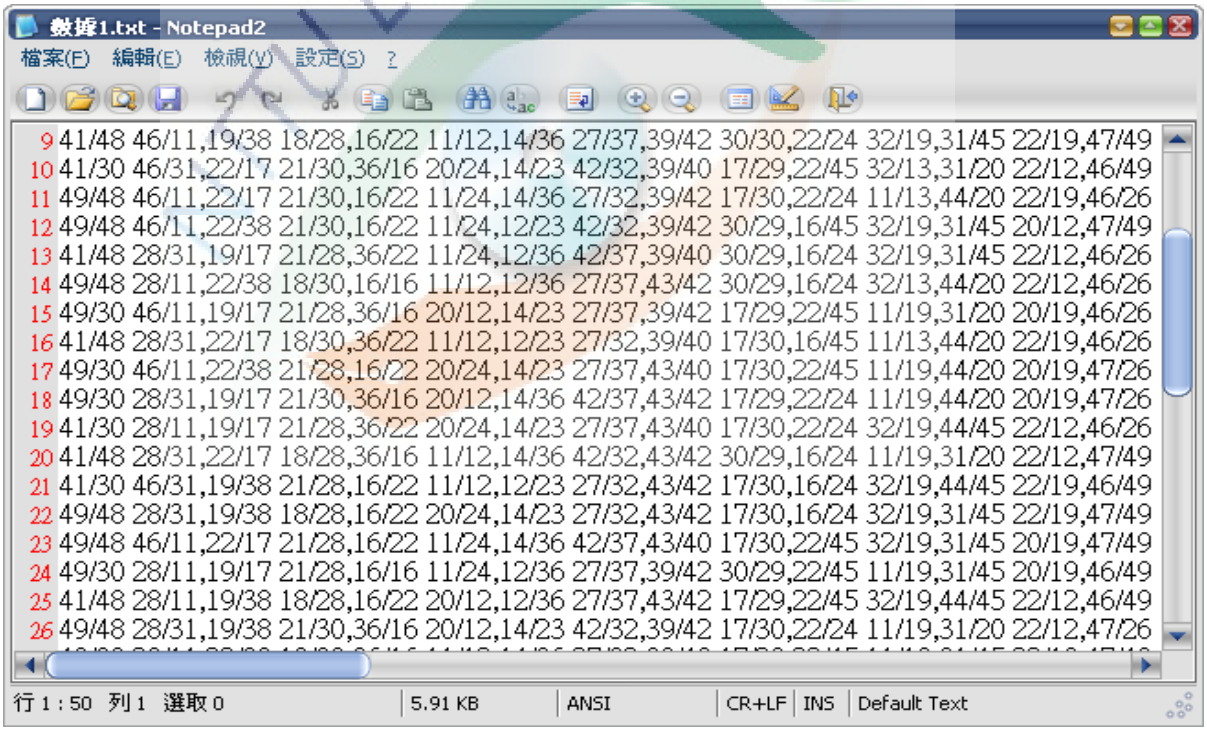

圖 3　匯入的文字檔

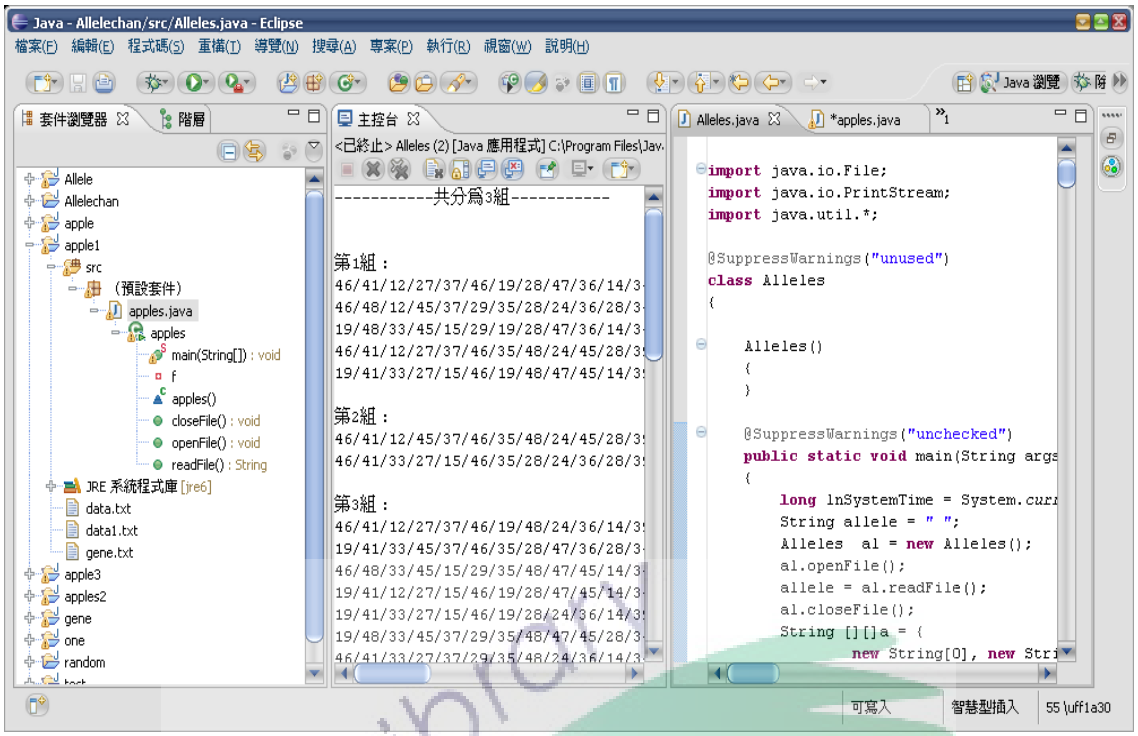

圖 4 執行後結果畫面

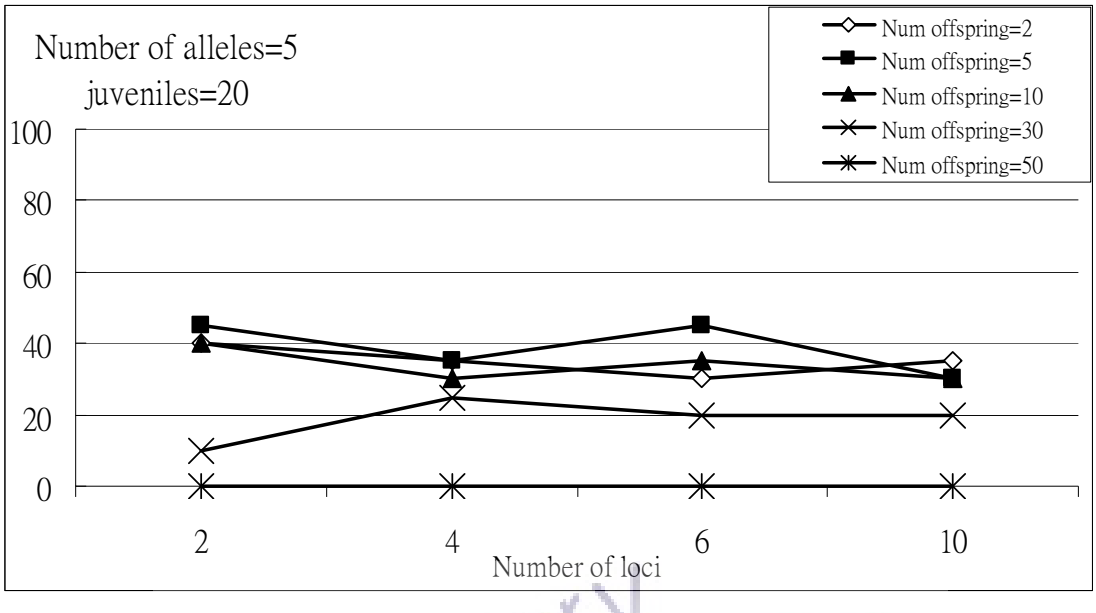

圖 5　基因座的個數對錯誤率的影響

圖 5 中列出的是我們的演算法與 Berger-Wolf et al.（2005）的整數規劃最小集合涵蓋演算法之錯誤率的比較。橫軸部份為基因座的個數，縱軸部分為我們的錯誤率。舉例而言，物種數目為 20，不同的對偶基因最多為 5 的時候，當基因座的個數為 2，offspring 為 10 的情況下，兩個演算法的分割距離為 8，所以錯誤率為 8 / 20 = 40%。由本張圖的情形觀察，當物種個數與對偶基因數為固定時，基因座的個數增加，其實並沒有很顯著影響到錯誤率的上升與下降。因此基因座的個數增減在本研究實驗結果之下，對錯誤率影響是不大的。

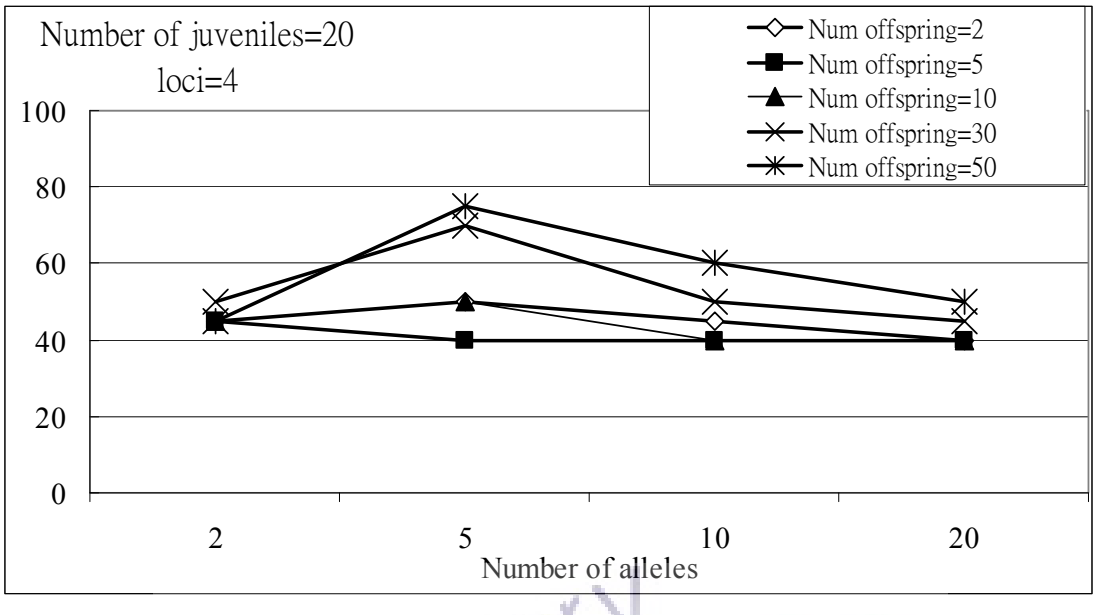

圖 6　對偶基因的個數對錯誤率的影響

圖 6 是我們的演算法與 Berger-Wolf et al.（2005）的整數規劃最小集合涵蓋演算法之錯誤率的比較。舉例而言，當我們固定 locus 個數為 4 與物種個數為 20 時，不同的對偶基因個數最多為 2，offspring 為 5 的情況下，兩個演算法的分割距離為 9，所以錯誤率為 9 / 20 = 45%。在觀察本研究實驗數據時，我們發現同一對父母生的小孩與對偶基因並不會很明顯的影響到我們的錯誤率高低，所以對偶基因個數的多寡，在本研究執行的結果下，對於錯誤率也不會有顯著影響。

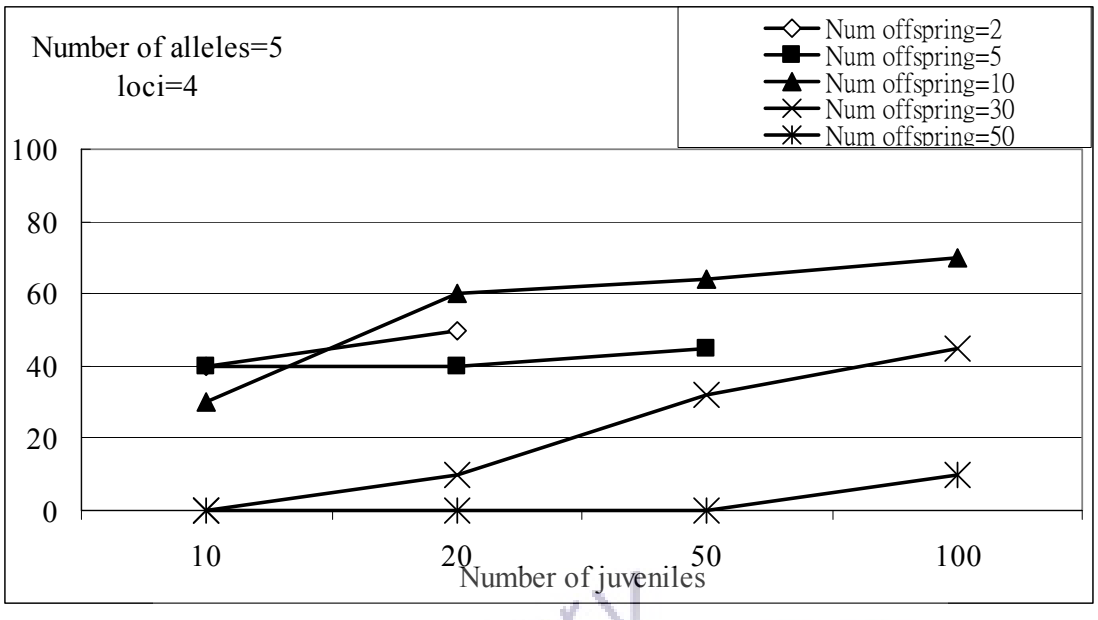

圖 7　物種個數對錯誤率的影響

圖 7 是我們的演算法與 Berger-Wolf et al.（2005）的整數規劃最小集合涵蓋演算法之錯誤率的比較。舉例而言，當我們固定 locus 個數為 4 與不同的對偶基因個數最多為 5 時，產生物種個數為 20，offspring 為 10 的情況下，兩個演算法的分割距離為 12，所以錯誤率為 12 / 20 = 60%。在觀察本研究實驗數據時，我們發現當物種數目越多，錯誤率也會跟著提高。根據本研究的參數設定，當同一對親代生物種所生的小孩個數最多為 2 時，因為十對親代生物種加起來小孩個數只有 20 個，因此只能測量到當物種個數為 20 的時候；小孩個數為 5 時，也如上述所說，十對親代生物種加起來小孩個數最多只有 50 個，因此也無法達到測量 100 的物種個數。

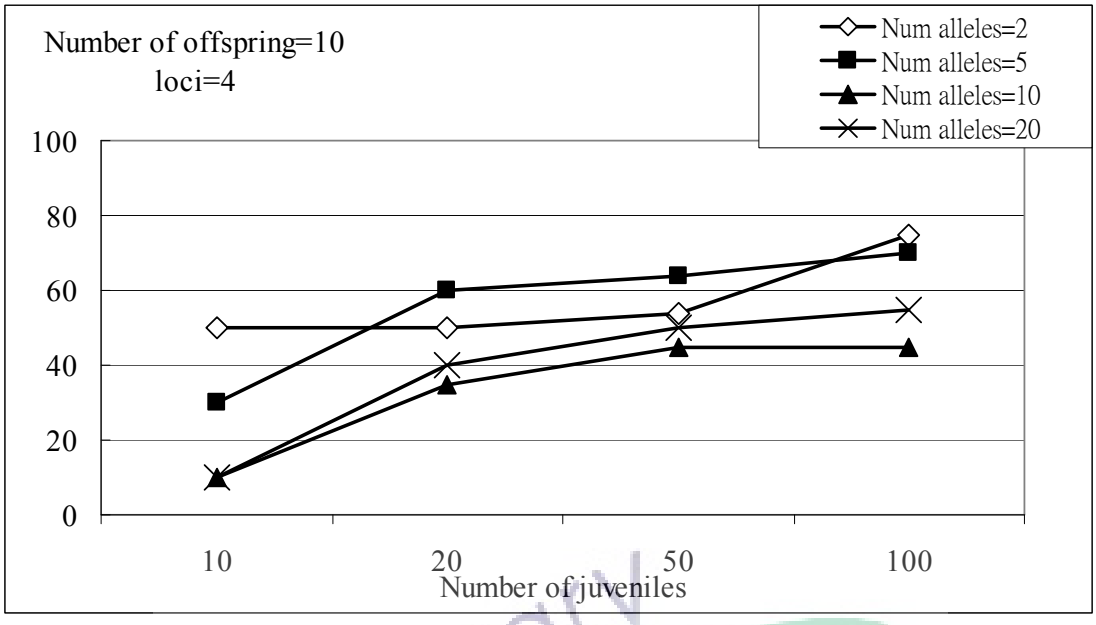

圖 8　物種個數對錯誤率的影響

圖 8 也是針對物種個數對錯誤率的影響，但變數改為不同的對偶基因的個數。舉例而言，當我們固定 locus 個數為 4 與 offspring 為 10 時，產生物種個數為 50，不同的對偶基因個數最多為 2 的情況下，兩個演算法的分割距離為 22，所以錯誤率為 22 / 50 = 54%。在觀察本研究實驗數據時，我們發現錯誤率會隨著物種的增加而向上攀升。

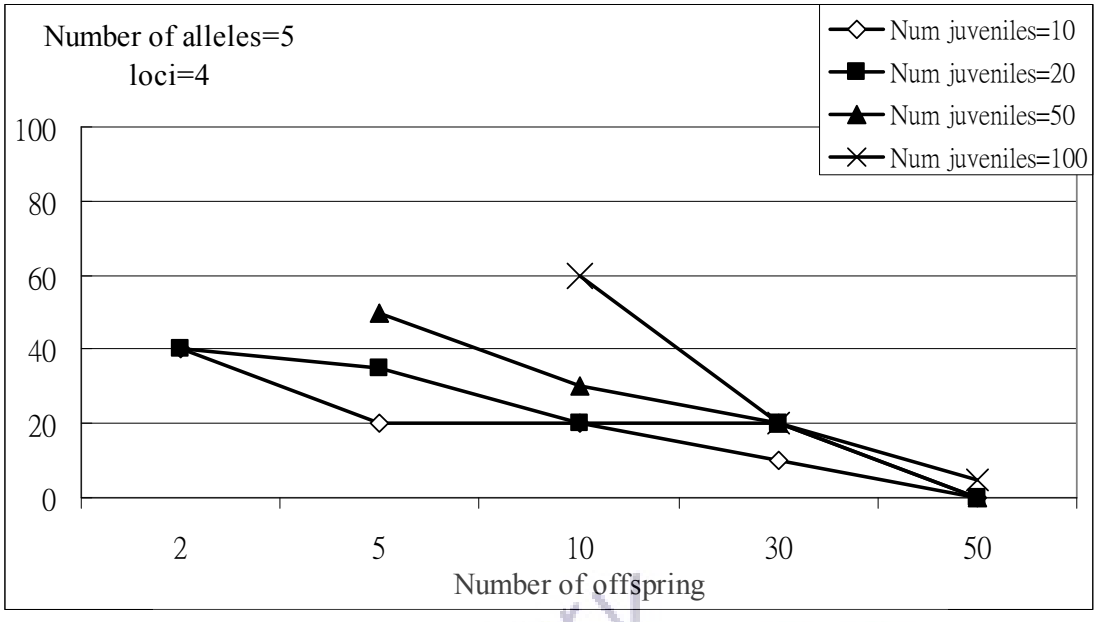

圖 9　子代個數對錯誤率的影響

圖 9 為我們的演算法與 Berger-Wolf et al.（2005）的整數規劃最小集合涵蓋演算法之錯誤率的比較。舉例而言，當我們固定 locus 個數為 4 與不同的對偶基因個數最多為 5 時，offspring 為 10，物種個數為 10 的情況下，兩個演算法的分割距離為 2，所以錯誤率為 2 / 10 = 20%。在觀察本研究實驗數據時，我們可以發現當相同子代增加時，我們的錯誤率就會隨著減少。根據本研究的參數設定，當測量物種個數為 100 時，同一對親代生物種的子代個數最多為 2 與 5 時，沒有辦法達到 100 個物種個數，所以無法量測；當物種個數為 50 時，子代個數最多為 2 時也無法量測。

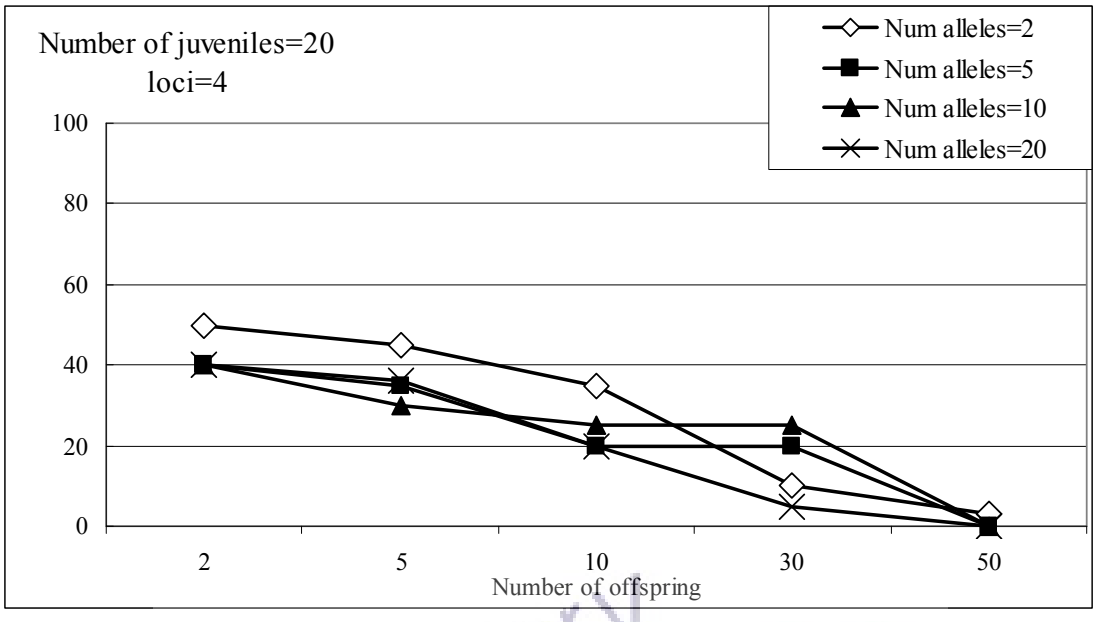

圖 10　子代個數對錯誤率的影響

圖 10 也是針對子代個數對錯誤率的影響，但變數改為不同的對偶基因的個數。舉例而言，當我們固定 locus 個數為 4 與物種個數為 20 時，offspring 為 30，不同的對偶基因個數最多為 10 的情況下，兩個演算法的分割距離為 5，所以錯誤率為 5/ 20 = 25%。在觀察本研究實驗數據，我們可以發現相同子代個數增加時，錯誤率就會隨之降低。

本研究中我們並不針對兩個演算法時間進行比較，因為 Berger-Wolf et al.（2005）的演算法是透過整數規劃來分群，而整數規劃一般的執行時間呈指數成長，且為 NP-Complete，我們的演算法時間複雜度為 $O(n^2l)$，所以在執行上會明顯的快很多，而在這六項實驗中，我們的演算法在執行時間上平均執行一次約一秒的時間。

第五章　結論與未來研究方向

本研究針對完全親緣關係重建問題，成功的設計了一個啟發式的分群演算法，在此演算法下，並不需要親代資料，其時間複雜度為 $O(n^2l)$，l 為每個物種基因座的個數。之後產生一些生物資料進行模擬測試，並與此問題的最佳分群演算法進行比較，比較結果可知，我們的演算法錯誤率都在 0%至 75%之間，而且在執行時間上快很多。

本研究提出的演算法在錯誤率部分仍有改善空間。在未來的研究方向除了我們可設計較好的演算法改善錯誤率之外，在我們的演算法部分，我們建議可在於未分群的物種要加入某個群組時，可透過一些較快速的搜尋演算法來達成，例如:binary search 或是透過一些雜湊函數，來改善其錯誤率與執行速度。此外，是否能證明完全親緣關係重建問題是否為 NP-Complete 或是我們的演算法產生的分群距離最佳的分群結果是否有一個倍數的保證也是未來可以計畫去研究的議題。

參考文獻

中文部份

中華民國鑑識科學學會（2009）。親子 DNA 鑑定實驗室認證技術規範草案，http://docs.google.com/viewer?a=v&q=cache:yvVWa4pKK6wJ:tafs.cid.cpu.edu.tw/downloap/relationship%2520test.pdf+定義+親緣&hl=zh-TW&gl=tw&sig=AHIEtbS1Jn9ohBQg9_LzbKh76NyNNPesAw&pli=1。

交通大學生物科技結構及細胞生物諮詢網（2000）。http://140.113.239.106/sections.php?op=listarticles&secid=7。

吳英嬌、吳惠生（1986）。英漢生物化學大辭典，台北：名山。

英文部分

Almudevar, A. and Field, C. (1999). Estimation of single generation sibling relationships based on DNA markers. *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 136-165.

Ashley[a], M. V., Berger-Wolf, T. Y., Berman, P., Chaovalitwongse, W., DasGupta, B., and Kao, M. Y. (2009). On Approximating four covering and packing problems. *Journal of Computer and System Sciences*, 75, 287-302.

Ashley, M. V., Berger-Wolf, T. Y., Caballero, I. C., Chaovalitwongse, W., DasGupta, B., and Sheikh, S. I. (2010). Full sibling reconstruction in wild populations from microsatellite genetic markers. *Computational Biology: New Research*, 231-258.

Ashley[b], M. V., Caballero, I. C., Chaovalitwongse, W., DasGupta, B., Govindan, P., Sheikh, S. I., and Berger-Wolf, T. Y. (2009). KINALYZER, A computer program for reconstructing sibling groups. *Molecular Ecology Resources*, 9, 1127-1131.

Balas, E. and Carrera, M. (1996). A dynamic subgradient-based branch-and-bound procedure for set covering. *Operations Research*, 44, 875-890.

Beasley, J. E. (1990). A Lagrangian heuristic for set covering problems. *Naval Research Logistics*, 37, 151-164.

Beasley, J. E. and Ornsten, K. J. (1992). Enhancing an algorithm for set covering problems. *European Journal of Operational Research*, 58, 293-300.

Beyer, J. and May, B. (2003). A graph-theoretic approach to the partition of individuals into full-sib families. *Molecular Ecology*, 12, 2243-2250.

Berger-Wolf, T. Y., DasGupta, B., Chaovalitwongse, W., and Ashley, M. V. (2005). Combinatorial reconstruction of sibling relationships. *Proceedings of the 6th International Symposium on Computational Biology and Genome Informatics*, 1252-1255.

Berger-Wolf, T. Y., Sheikh, S. I., DasGupta, B., Ashley, M. V., Caballero, I. C., Chaovalitwongse, W., and Putrevu, S. L. (2007). Reconstructing sibling relationships in wild populations. *Bioinformatics*, 23, i49-i56.

Björklund, A., Husfeldt, T., and Koivisto, M. (2009). Set partitioning via inclusion-exclusion. *SIAM Journal Computing*, 39, 546-563.

Bourgeois, N., Escoffier, B., and Paschos, V. T. (2009). Efficient approximation of min set cover by moderately exponential algorithms. *Theoretical Computer Science*, 410, 2184-2195.

Caprara, A., Fischetti, M., and Toth, P. (1999). A heuristic method for the set covering problem. *Operations Research*, 47, 730-743.

Caprara, A., Fischetti, M., and Toth, P. (2000). Algorithms for the set covering problem. *Annals of Operations Research*, 98, 353-371.

Ceria, S., Nobili, S. P., and Sassano, A. (1998). A Lagrangian-based heuristic for large-scale set covering problems. *Mathematical Programming*, 81, 215-228.

Chaovalitwongse, W., Berger-Wolf, T. Y., DasGupta, B., and Ashley, M. V. (2007). Set covering approach for reconstruction of sibling relationships. *Optimization Methods and Software*, 22 (1), 11-24.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithm* (2nd edition). Cambridge: MIT Press.

Cygan, M., Kowalik, Ł., and Wykurz, M. (2009). Exponential-time approximation of weighted set cover. *Information Processing Letters*, 109, 957-961.

Doi, K. and Imai, H. (1997). Greedy algorithms for finding a small set of primers satisfying cover and length resolution conditions in PCR experiments. *Genome Informatics*, 8, 43-52.

Doi, K. and Imai, H. (1999). A greedy algorithm for minimizing the number of primers in multiple PCR experiments. *Genome Informatics*, 10, 73-82.

Drezner, Z. and Hamacher, H. W. (2002). *Facility location: applications and theory*. Series, Springer.

Duitama, J., Kumar, D. M., Hemphill, E., Khan, M., Măndoiu, I. I., and Nelson, C. E. (2009). PrimerHunter**:** a primer design tool for PCR-based virus subtype identification. *Nucleic Acids Research*, 37, 2483-2492.

Farahani R. Z. and Hekmatfar, M. (2009). *Facility Location: Concepts, Models, Algorithms and Case Studies*, Springer Verlag.

Feige, U. (1998). A threshold of ln for approximating set cover. *Journal ACM*, 45, 634-652.

Fernández, J. and Toro, M. A. (2006). A new method to estimate relatedness from molecular markers. *Molecular ecology*, 15, 1657-1667.

Fomin, F. V., Grandoni, F., and Kratsch, D. (2005). *Measure and conquer: domination - a case study. Lecture Notes in Computer Science (LNCS),* Springer-Verlag, 3580, 191–203.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. San Francisco: W.H. Freeman and Company.

Goldsmidt, O., Hochbaum, D. S., and Yu, G. (1993). A modified greedy heuristic for the set covering problem with improved worst case bound. *Information Processing Letters*, 48, 305-310.

Gusfield, D. (2002). Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters*, 82, 159–164.

Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques*. (2nd ed.). San Francisco: Academic Press.

Haouari, M. and Chaouachi, J. S. (2002). A probabilistic greedy search algorithm for combinatorial optimization with application to the set covering problem. *Journal of the Operational Research Society*, 53, 792-799.

Huang, Y. C., Chang, C. F., Chan, C. H., Yeh, T. J., Chang, Y. C., Chen, C. C., and Kao, C. Y. (2005). Integrated minimum-set primers and unique probe design algorithms for differential detection on symptom-related pathogens. *Bioinformatics*, 21, 4330-4337.

Jabado, O. J., Palacios, G., Kapoor, V., Hui, J., Renwick, N., Zhai, J., Briese, T., and Lipkin, W. I. (2006). Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. *Nucleic Acids Research*, 34, 6605-6611.

Johnson, D. S. (1974). Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences*, 9, 256-278.

Konovalov, D. A., Manning, C., and Henshaw, M. T. (2004). KinGroup: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Molecular Ecology*, 4, 779-782.

Lan, G., DePuy, G. W., and Whitehouse, G. E. (2007). Discrete optimization an effective and simple heuristic for the set covering problem. *European Journal of Operational Research*, 176, 1387-1403.

Lovász, L. (1975). On the ratio of optimal integral and fractional covers. *Discrete Mathematics*, 13, 383-390.

Mannino, C. and Sassano, A. (1995). Solving hard set covering problems. *Operations Research Letters*, 18, 1-5.

Rooij, J. M. M. and Bodlaender, H. L. (2008). Design by measure and conquer, a faster exact algorithm for dominating. *Proceeding International Symposium on Theoretical Aspects of Computer Science*, 657-668.

Sheikh, S. I., Berger-Wolf, T. Y., Chaovalitwongse, W., DasGupta, B., and Ashley, M. V. (2007). Reconstructing sibling relationships from microsatellite data. *European Conf. on Computational Biology (ECCB)*.

Sheikh, S. I., Berger-Wolf, T. Y., Khokhar, A., and DasGupta, B. (2008). Consensus methods for reconstruction of sibling relationships from genetic data. *Association for the Advancement of Artificial Intelligence(AAAI)*, 97-102.

Sheikh, S. I., Berger-Wolf, T. Y., Khokhar, A., Caballero, I. C., Ashley, M. V., Chaovalitwongse, W., and DasGupta, B. (2009). Combinatorial reconstruction of half-sibling groups. *8th International Conference on Computational Systems Bioinformatics (CSB)*, 59-67.

Slavík, P. (1997). A tight analysis of the greedy algorithm for set cover. *Journal of Algorithms*, 25, 237-254.

Tan, P. N., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*. Addison-Wesley.

Vazirani, V. V. (2001). *Approximation Algorithms*. Berlin: Springer-Verlag.

Wang, J. (2004). Sibship reconstruction from genetic data with typing errors. *Genetics*, 166, 1963-1979.

Zhao, W., Fanning, M. L., and Lane, T. (2005). Efficient RNAi-based gene family knockdown via set cover optimization. *Artificial Intelligence in Medicine*, 35, 61-73.