

DNA分子標誌資料 之遺傳變異的歸群分析

農試所作物組 呂秀英 魏夢麗 呂椿棠

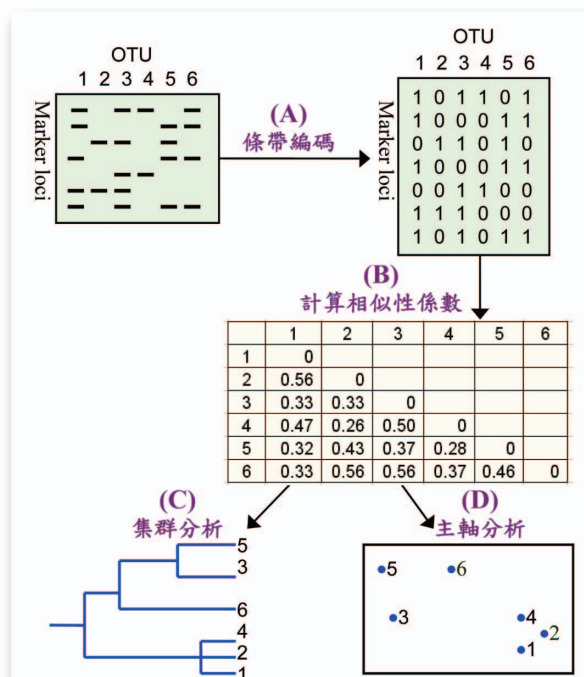
一、前言

遺傳標誌 (genetic marker) 是指可追蹤染色體、染色體某一片段或某個基因座 (gene locus) 在譜系中傳遞的任何一種遺傳特性，為表示遺傳多樣性 (或稱遺傳變異) 的有效手段。隨著遺傳學的不斷發展，遺傳標誌的種類和數量也持續增加，迄今主要分為形態標誌、細胞學標誌、生化標誌和分子標誌。其中前三種遺傳標誌都是以基因表現的結果 (外表型) 為基礎，是對基因的間接反映；而分子標誌是指能反映生物個體或種群間基因組中某種差異特徵的 DNA 片段或蛋白質，即以生物大分子的多型性 (polymorphism) 為基礎，直接反映基因組的遺傳變異。廣義的分子標誌包含可遺傳的且可檢測的 DNA 序列或蛋白質，其中蛋白質標誌包括種子貯藏蛋白質和同工酶。狹義的分子標誌只是指 DNA 分子標誌。分子標誌利用遺傳標誌與重要農藝性狀間之連鎖關係，已廣泛用於植物分子遺傳圖譜的建構、品種鑑別和品系純度的測定、重要農藝性狀基因定位與選殖、基因轉殖植物鑑定和分子標誌輔助選拔育種等方面。DNA 分子標誌的擴增產物在電泳凝膠上分離後，必須讀帶和統計分析所獲得的數據資料。DNA 電泳圖譜條帶的紀錄，經常依其條帶的有

和無兩種狀態來呈現紀錄，然後再依其出現狀態的相似程度進行歸群分析，資料處理流程如圖一所示。本文將對此統計分析流程及方法進行說明，以供參考。

二、DNA 分子標誌的類型和資料形式

DNA 分子標誌為近代生物技術發展下之產物，目前常用的分子標誌是由數種

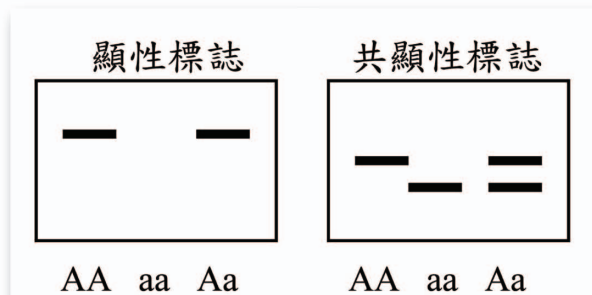


圖一、DNA 分子標誌資料處理的流程圖：(A) 電泳圖譜條帶紀錄以出現有無各以 0 和 1 編碼，(B) 估算兩兩操作分類單位 (OTU) 間的相似性係數，(C) 由相似性係數進行集群分析，(D) 由相似性係數進行主軸分析。

作者：呂研究員秀英
連絡電話：04-23302301-125

不同技術發展出來，各有其特性，例如以核酸隨機增殖多型性 (random amplified polymorphic DNA, RAPD) 及增殖片段長度多型性 (amplified fragment length polymorphism, AFLP) 方法對DNA品質要求不高，操作簡單易行，可檢測整個基因組，但由於一般表現為顯性遺傳，只能檢測顯性基因，不能區分顯性純合和雜合基因型；以限制片段長度多型性 (restriction fragment length polymorphism, RFLP) 及簡單重複序列 (simple sequence repeat, SSR) 方法在基因組中多型性較高，並且對位基因數目多，呈共顯性遺傳，可檢測共顯性基因，但必須依賴定序設計引物 (primer)，開發成本高。共顯性標誌 (co-dominant markers) 在每個膠槽都產生至少一個條帶 (遺傳對位基因座)，而顯性標誌 (dominant markers) 則預期會有空的對位基因座發生，示意如圖二。

在顯性標誌場合 (如RAPD 或 AFLP)，可以將電泳圖譜中條帶的有和無兩種狀態各以1和0來編碼登記，此屬於二元資料形式；而在共顯性標誌場合 (如SSR 或 RFLP)，則大多建議使用較多的標誌，以多對位基因的頻度 (frequency) 來登記，此為類別資料形式。SSR 和RFLP 也可直接用條帶的有和無來登記，不過若以這種二元性狀來建立資料，有可能會喪失某些共顯性標誌所特有的遺傳資訊。無論這些



圖二、顯性標誌和共顯性標誌條帶出現的特性差異示意圖。

條帶性狀是二元或類別資料，所組成的資料矩陣都是一個多變數資料形式 (如圖一(A) 範例)，即表中的行是作為變數的標誌位置 (marker loci)，而欄是操作分類單位 (operational taxonomic unit, OTU)，如品種、變種、個體等。

三、DNA 分子標誌的遺傳相似性係數

在DNA 分子標誌以條帶有無之二元資料呈現的場合，兩兩OTU間之分子標誌條帶之有無的編碼及條數統計，示意如圖三 (a= 共有帶數，b和c= 特異帶數，d= 共同沒有帶的數目)。

統計上OTU之間的相似程度，可用相似性係數、距離係數或相關係數來測量。農藝性狀符合計量值的資料形式，可直接計算兩OTU間的距離係數 (如歐氏距離) 或相關係數。但對於分子標誌的二元資料，必須就原始二元資料 (0,1) 先計算出每對OTU之間的相似性係數 (如圖一(B) 範例)。分子標誌二元資料的相似性係數，以Jaccard 係數和Nei & Li 係數最常被採用，其值都在 [0,1] 之間，但兩種方法的意義有很大的不同。以圖三為例說明，Jaccard 係數值 = $\frac{a}{a+b+c} = 3 / (3+2+3) = 0.375$ ，它不是看個別條帶，而是看膠片上的位置，其公式中的分母是能顯現相似程度 (即有發生條帶，無論是共有或特異) 的膠片位置總數，而分子是產生共有帶的膠片位置數。Nei & Li 係數 = $\frac{2a}{2a+b+c} = 2(3) / [2(3)+2+3] = 0.545$ ，它是看各OTU上的每一個條帶，即計算其出現的總帶數作為公式的分母，而分子是共有帶的條帶總數，換言之，Nei & Li 法對正向配對的共有帶 (a) 有更多的加權。

顯然在所有對位基因都產生條帶的場合，Nei & Li法將共有帶視為兩個條帶的作法，能對正向配對 (a) 和沒有配對 (b,

c)之間的差異性區分開來，但Jaccard法是將兩種特異帶算成是兩個不同膠帶位置，卻將共有帶只視為一個位置而已，以致往往低估了相似性。因此，對於每個膠槽都產生至少一個條帶(遺傳對位基因座)之

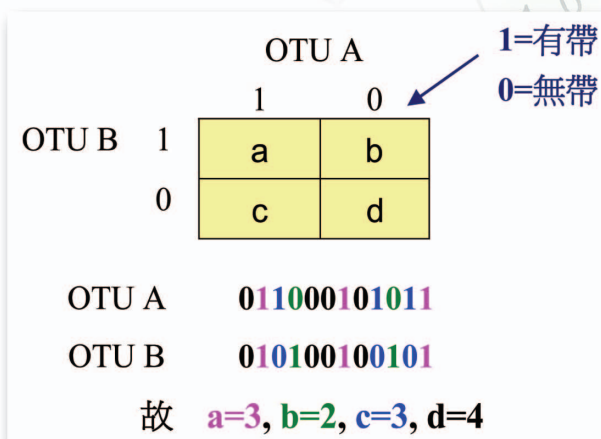
場合，即很少會產生沒有任何條帶(“空”的對位基因座)時，如SSR和RFLP，建議採用Nei & Li係數最理想；倘若預期會有空的對位基因座發生，如RAPD和AFLP，則可以使用Jaccard係數。以此類推，倘若有針對同一試驗材料進行SSR和AFLP所得之圖譜條帶紀錄，欲合併資料一起進行統計分析時，則應當使用Nei & Li係數。

至於其他相似性係數，如Gower係數與Jaccard係數的公式是相同的，而Sorensen係數和Dice係數與Nei & Li係數的公式也都相同。

四、DNA分子標誌的多變數分析

求得兩兩OTU間的相似性係數所組成的關聯矩陣(association matrix)後，就可以對OTU進行群聚類別的關係探討。農業和生物科技之多變數資料進行關聯分析所得的結果，常以集群樹狀圖(cluster dendrogram)和序列分布圖(ordination dendrogram)來表示，如圖一(C)、(D)範例。在實際應用上通常會同時依據兩者群聚結果，以相輔相成來決定群數。

多變數分析中的集群分析(cluster



圖三、兩兩操作分類單位(OTU)間分子標誌條帶之有無的編碼及條數統計示意圖(a=共有帶數，b和c=特異帶數，d=共同沒有帶的數目)。

analysis)所使用的距離矩陣法又可再分為近鄰法、遠鄰法、Ward最小變方集群法、均連法等，不同之抉擇往往造成不同的分群結果。均連法(average linkage)又稱為非加權配對算術平均法

(unweighted pair group method using arithmetic average, UPGMA)，是DNA分子標誌資料進行集群分析最常使用的方法。集群分析整個分析結果是以樹狀圖呈現，凡樣本間的相似性愈高時，則愈早相連，相連後可形成數小群，最後再依據群間相似性高低，逐步相連直到最後成為單一群體為止。至於要區分為幾群，通常以樹狀圖的分枝狀況來決定，但可在樹狀圖任何層級上被切分。

而DNA分子標誌之各OTU在二維或三維空間上排列的相互關係，最常使用多變數分析中的主軸分析(PCoA, principal coordinate analysis，或稱主座標分析，即度量多維尺度法—MDS)，而非主成分分析(PCA, principal component analysis)。這是因為PCA只限於資料形式為測量的數量變數，而PCoA為PCA的一般化版本，能執行各種度量距離測量，因此可適用於各種多變數資料形式，即使是存在與否的二元資料場合。利用PCoA所得的序列分布圖來呈現OTU間的親緣關係，即由各點(OTU)落在圖上之距離遠近，可用來輔助上述集群分析之樹狀圖的分群結果判斷，倘發現點間有群團現象表示可能是同一群。