

人類群體基因組單體型推論問題之研究

楊惠娥

成功大學資訊管理研究所

隨著人類基因體計劃的完成，我們對人類DNA 的結構及序列已有初步認識，但仍無法確知其功能。DNA的功能可能是疾病研究的關鍵，為了確定其功能，科學家們會針對人類每條染色體上稱為基因組單體型(haplotype) 的基因鹼基組合來做進一步的分析。

由於直接取得基因組單體型資料將耗費許多成本及時間，科學家們通常會使用在一對染色體上稱為基因型(genotype) 的基因鹼基混合敘述資料，來代替基因組單體型資料以進行分析。然而，基因型資料所含的資訊並不足以正確地推論出每條染色體上基因的鹼基組合，因此我們必須求解人類群體基因組單體型推論(Population Haplotype Inference, PHI) 問題，以從一群人的基因型資料來推論出其相對應的基因組單體型資料。在眾多的PHI 問題相關文獻中，近年來以應用最大簡約原則(Pure Parsimony) 的PHI 問題(PHI problem based on pure parsimony criterion, HIPP) 之相關研究最受矚目。HIPP 問題旨在使用最少的不同基因組單體型，來解讀一個給定的基因型矩陣。

本論文特別針對HIPP 問題加以探討，在探討並整理文獻中各類求解PHI 問題之方法後，我們依照基因型資料間的相互包容或排斥的數學關係，提出兩個結合數學規劃與生物意義的啟發式演算法。其中，第一個演算法利用基因型資料間的容斥關係，將可行解空間大幅縮小，以整合的基因型資料來求解一個整數規劃問題；而第二個演算法賦予那些可解讀較多基因型的基因組單體型更高的權重，並使用貪婪(greedy) 的方式來選取權重較大的基因組單體型以求解HIPP 問題。

本論文並執行大規模的程式測試，來分析各類解法在求解HIPP 問題時之效率及效果雙方面的表現。最後，我們提出一個可用來處理大規模(large scale)HIPP 問題的技巧，將該問題切成較小規模的子問題，求解各子問題之後再將其解組合之；另外，我們亦針對文獻中一個極有效率的啟發式演算法加以改良，使之可以有效地求解出任一HIPP 問題的所有最佳解。

關鍵字：基因組單體型推論；基因型；最大簡約；啟發式演算法；整數規劃