分別以連鎖不平衡及拉氏鬆弛法選取代表性單核苷酸多型性之研究

馬家宜

成功大學工業與資訊管理學系碩士班

在DNA序列可能發生的眾多差異性當中，單核苷酸多型性(Single Nucleotide Polymorphism，SNP) 是最常發生的一種遺傳變異，由多個SNP鹼基所組成之序列稱為基因組單體形(Haplotype)，此序列的改變對疾病的發生及人類特徵的顯現有重大關聯，可被應用於辨識不同疾病及其他相關之醫學研究上。由於目前已發現的SNP資料量龐大，為了節省SNP資料庫所需的高成本花費，許多研究建議以被稱為tagSNP的SNP序列資料部份集合來代表原本全部的SNP序列。

tagSNP可依其應用目的而有不同的定義，本研究首先針對文獻中最常被使用之tagSNP定義，將其選取問題(Selection Problem) 轉換成一個具有多重最佳解的0,1二元整數規劃問題，以選出可辨識出所有的Haplotype序列樣式之最小SNP部分集合。由於過去研究多著重於改善求解方法之效率，並未評估所求得之最佳解與其它最佳解間之資訊差異，因此本研究提出一個以圖型理論為基礎的啟發式演算法，先求解出所有的最佳解，再採用連鎖不平衡(Linkage Disequilibrium，LD) 觀念以計算已被選取之tagSNP與尚未被選出之其它SNP間的相互關連性，作為該最佳解所包含Haplotype資訊量多寡的評估指標，並選取其中最多資訊者為最終最佳解。此外，我們亦提出一個可同時考慮極小化tagSNP個數與極大化LD值之雙目標數學規劃模式以求解類似問題。

在求解大規模的tagSNP選取問題上，本研究提出一個以拉氏鬆弛法為基礎的啟發式演算法LRH，採用次梯度(subgradient) 法更新拉氏乘數(Lagrangian multiplier)，使之逐漸逼近最佳解。我們亦在求解過程中加入貪婪演算法的觀念，藉由固定部份SNP欄位以逐漸縮減問題規模，改善求解速度及求解品質。此外，我們亦提出一個結合LRH與最佳化軟體CPLEX兩者優點的二階段求解方法；數值測試結果顯示該二階段解法的確可以在更短的時間內選取出品質更佳之tagSNP解。最後，本研究提出一個整數規劃模式以在具有容量限制的生物晶片上選取較可靠的tagSNP解，並呈現容量限制下限與辨識之可靠性間的關係圖以供後續研究參考。

***關鍵字*：**連鎖不平衡；標記型單核苷酸多型性；基因組單體型；演算法；拉氏鬆弛法

論文指導教授: *王逸琳* ilinwang@mail.ncku.edu.tw
http://ilin.iim.ncku.edu.tw
***馬家宜* 於 2006 年就讀成大工資所碩士班，於 2008 年獲得碩士學位**