

基因演算法應用於骨質疏鬆症之不同染色體基因多型性群之關聯性分析

楊正宏^a 何長軒^a 溫政浩^b 張學偉^b 莊麗月^c

^a高雄應用科技大學 電子工程系

^b高雄醫學大學 生物醫學暨環境生物學系

^c義守大學 化學工程系

^aInstitute of Electronic Engineering, National Kaohsiung University of Applied Sciences

^bFaculty of Biomedical Science and Environmental Biology, Kaohsiung Medical University

^cChemical Engineering, I-Shou University

摘要

對於研究骨質疏鬆症這種多基因遺傳因子疾病和單核苷酸基因多型性(Single nucleotide polymorphisms, SNPs)與遺傳變異之間的關係，連鎖不平衡(Linkage Disequilibrium, LD)為一重要的依據。然而根據LD的分析，只能夠探討同一染色體上多個特定SNP與遺傳表型的相關性。但影響骨質密度的基因，並不限於分布在同一染色體上。因此只用LD分析骨質疏鬆症反應路徑的相關基因之單核苷酸多型性，並不能完全正確地代表基因多型性與骨質疏鬆症的相關性。為了改善這個問題，我們利用基因演算法，從不同染色體中特定多個SNP與基因型的表現頻率中，找出特定SNP與其表現型之間高度相關性的類似haplotype組合。有別於LD的微觀，利用整體基因庫內宏觀的角度，此跨越染色體分析分散於基因庫內的多個SNP集合，對於研究遺傳表型與多個SNP的相互作用提供另一種新的方法。

關鍵字：單一核苷酸多型性、骨質疏鬆症、連鎖不平衡、基因演算法

1. 前言

近年來，人類基因圖譜的解碼，讓醫學界對於疾病的預測、預防以及治療有了全新的視野，科學家們希望能夠從眾多的基因中找出和疾病相關的基因來幫助診斷、預防甚至治療這類的疾病，其中基因多型性的研究近年來相當熱門，其研究重點除了與遺傳表現型(Phenotype)的相關性研究之外，也包括人種與族群不同的研究。在遺傳表現型方面，單一SNP與遺傳變異的相關性研究，目前已累積相當的研究數量，對於現

代SNP與基因多型性研究幫助相當大。藉由之前累積的大量資料與候選SNP(Candidate SNP)，對於現代的多發性因子疾病，如骨質疏鬆症(Osteoporosis)方面的研究提供了許多候選SNP資料。骨質疏鬆症為一種多發性因子的疾病，主要的影響因素除了環境因素，包括老化、停經、和BMI值的影響外，還有遺傳方面的影響。而現今探討骨質疏鬆症的相關SNP研究，主要著墨於骨質代謝的相關基因，如TNF- α 為影響骨質再吸收作用的主要因子之一[5][6]。另一方面，一些基因如vitamin D receptor (VDR), estrogen receptor (ESR), parathyroid hormone (PTH), glucocorticoid receptor, calcitonin receptor (CTR) (10-12), insulin-like growth factor-1, collagen 1- α -1 (COL1A1), interleukin-6, transforming growth factor-beta1 (TGF- β 1)和APOE (14,15)。對於評估SNP在遺傳學的作用，連鎖不平衡(Linkage disequilibrium, LD)為現今最常使用於分析多個SNP的出現，可以代表特定遺傳表現型的工具，如PLOD、TNFRSF11B [7][8]。另一方面，造成多發性因子的疾病其相關基因並不常位於同一染色體內，如上述提到造成骨質疏鬆症的相關基因，TNF- α (tumor necrosis factor- α)與Osteocalcin分別位於chromosome 6和chromosome 1 [6] [9]。鑒於LD只能分析同一染色體上的多個SNP的相關性，可能導致分析骨質疏鬆症相關基因(但位於不同的染色體上)而有所缺憾，因此需要一套系統可以分析多個不同染色體上各個SNP對於骨質疏鬆症的影響與相關性研究。針對上述要求，我們從文獻中找出與骨質疏鬆症相關的SNP，利用已知的實驗結果將其genotype排列組合，藉此找出幾種SNP的組合可以代表骨質疏鬆症發生的高危險群；但由於在多個SNP的組合之下，每個SNP各自又有三種



不同的 genotype，在這樣複雜的組合之下所產生的可能組合數相當的大，要從這麼多的可能組合下找出高低骨密度表現頻率的差異最高的組合要耗費不少時間，因此我們利用一種容易應用於各種最佳化問題演算法-基因演算法從不同染色體上特定的 SNP 及其 genotype 中找出高度相關性的類似 haplotype 組合(因為 haplotype 適用於同一染色體，本方法非侷限於同一染色體，所以稱為類似 haplotype)。經過測試，基因演算法確實能在短時間內找到最具代表性的組合，而不必將所有的可能組合都一一計算比較。

2. 基因演算法

基因演算法(Genetic Algorithm, GA)又稱遺傳算法，是藉由模擬生物自然演化的過程來獲得最佳解[2][3]。在基因演算法中將許多染色體(Chromosome)集成一個族群，在經過每一次迭代過程中透過選擇(Selection)、交配(Crossover)、突變(Mutation)等機制，保留適應值(Fitness Value)較高的染色體以獲得更好的解。本研究藉由基因演算法能從廣大的解空間中迅速獲得最佳解的特性，評估來自各種不同染色體的 SNP 及其相對應的 genotype 在不同組合之下組成的類似 haplotype 組合對骨質疏鬆症發生比例的影響可能性。

在基因演算法中由於族群裡每一個染色體皆視為一組解，故我們依照所選取 SNP 及其相對應的 genotype 來設計染色體。染色體中的每一個基因(Gene)皆代表著每一次所選取的 SNP 即其相對應 genotype 的情形，如 (S_i, G_i) , $i=1, 2, \dots, n$ ，其中 S_i 表示所選取的 SNP， G_i 表示 S_i 當中被選取的 genotype 其中 $n = \text{number of SNP}$ ，再依照所要選取的 SNP 數量來產生染色體。

SNP 編碼的部分由於不可重複選取，所以在編碼上以不重複為原則，而 genotype 選取的部分由於是依附在各自的 SNP 之下所以沒有重複的限制。染色體整體的長度也會隨著選取 SNP 的數量不同而有所變動。

在目標函數設計部份，本研究將資料中所有的樣本根據骨密度(Bone mineral density, BMD) T score 的高低分為高骨密度和低骨密度兩群，將每個染色體分別代入高骨密度群和低骨密度群中計算出滿足該染色體所代表解的樣本數，透過此方式將自高骨密度群和低骨密

度群所得到的樣本數代入適應值函數(Fitness function)中計算其適應值(Fitness value)，其計算公式如下：

$$\frac{|H_BMD - L_BMD|}{ALL_H_BMD + ALL_L_BMD} * 100\%$$

H_BMD :高骨密度中符合選取條件的樣本數。

L_BMD :低骨密度中符合選取條件的樣本數。

ALL_H_BMD :高骨密度總樣本數。

ALL_L_BMD :低骨密度總樣本數。

族群初始化的部份利用隨機的方式來產生初始族群(Initial Population)。初始族群中每個染色體針對兩個不同的部分其基因隨機產生的範圍也有所不同，本研究資料中共有 11 個來自不同染色體的 SNP，而每個 SNP 皆有三種不同的 genotype，故產生的範圍 $S_i = [1, \dots, 11]$ 、 $G_i = [1, \dots, 3]$, $i=1, \dots, n$ 。

在染色體選擇的部份，本研究是採用明確取樣(Deterministic sampling)的方法，即在不考慮選取機率高低的情況之下直接從族群中挑選出足夠的最佳染色體。此方法最大的好處就是可以避免某個適應值特別高的染色體佔滿整個交配池，造成提早收斂(Convergence)陷入局部最佳化的困境。

交配和突變的部份則是分別採用雙點交配(Two-point Crossover)的方式和隨機選取一個基因來突變。然而在經過交配和突變過後產生的子代 $S_1, S_2, \dots, S_n$ 的部分必須避免重複的情況發生，一旦發生上述情形便以隨機的方式將重複的 SNP 突變來解決。

本研究中所使用參數設定族群大小為 500、交配率為 0.7、突變率為 1.0、世代數則設為 50。

3. 結果與討論

本研究實作基因演算法所使用骨質疏鬆症相關的 SNP 資料方面，總共有 11 個已証實的 SNP 與其相關實驗資料，即基因定性(Qualification)的結果。將此 11 個 SNP 分別表示為 SNP-1~SNP-11 如 Table 1。表中 Chr 和 Genotype 分別表示該 SNP 所在的染色體及其三種不同的基因型，也包含了 SNP1~11 進行 PCR-RFLP (polymerase chain reaction-restriction length polymorphism)基因定性所需的引子(primer)與限制酶的資訊，也提供聚合酶鏈鎖反應(PCR)的 annealing 溫

