應用增強型粒子族群最佳化演算法於單核苷酸多型性標籤選取問題

楊正宏
國立高雄應用科技大學
電子工程系
chyang@cc.kuas.edu.tw

侯又壬
國立高雄應用科技大學
電子工程系
1096305143@cc.kuas.edu.tw

何長軒
國立高雄應用科技大學
電子工程系
1094320138@cc.kuas.edu.tw

莊麗月
義守大學
化學工程系
chuang@isu.edu.tw

摘要

單核苷酸多型性是由基因中單一個核苷酸的改變所造成的，且被視為人體基因多樣性的主要原因之一。過去都是利用 SNP 來進行研究基因變異與疾病之間的關係，若透過 haplotype 來研究基因變異與疾病間的關係，不但效率高，且節省成本。為了能夠以少量資料找到蘊含豐富資訊量的 TagSNPs，我們透過改良式的 PSO，來挑選最好的單核苷酸多型性標籤。本研究在 PSO 裡加上了粒子與相互(mutually)之間的資訊交換，以達到更好的搜尋效果。本實驗測試資料集來自 HapMap 並且以重建 Haplotype 的正確率和 GEVALT 所求得的解作比較，結果顯示本實驗提出的改良式 PSO 確實能夠改善預測正確率，達到挑選出較好的單核苷酸多型性標籤的目的。

關鍵詞：單核苷酸多型性、haplotype、Tag SNP、PSO

ABSTRACT

Single Nucleotide Polymorphism is caused by single change of one Nucleotide in the gene, and is a major impact on variety of the human gene. In the past, SNP is used to process genetic mutation of research and relation between disease. We try to find out the relation between genetic mutation and disease through haplotype. It's not only high in efficiency, but also save the cost. In order to be able to use a small amount of data and contains rich information of TagSNPs, we improved PSO to select better tagSNPs. In this paper we propose a new condition to original PSO, it is the exchange of mutually information of particles, called Mbest, to achieve better search results. The experimental data sets we used is download from HapMap, and compared the accuracy of reconstruct Haplotype and the result of using GEVALT. The results showed that the improved PSO this study proposed can indeed improve the accuracy of divination. We improved that Enhance PSO will select better tagSNPs.

Keyword : Single Nucleotide Polymorphism, haplotype、Tag SNP、PSO.

1.前言

單核苷酸多型性 (Single Nucleotide Polymorphism, SNP)為基因組的一個特異和定位點出現兩個或多個的核苷酸可能性；SNP 的產生可能會造成蛋白質表現的改變，所以 SNP 也成為影響人類在特徵上差異的關鍵，例如有些人可能特別容易或特別不容易患上某些疾病，或對於治療藥物的反應性有所差異等現象。SNP 的發生頻率非常高，且每個人的 DNA 上所發生的 SNP 皆不同，故 SNP 常被當作一種基因標記（genetic marker）。大約每一千個鹼基對就會有單核苷酸變異的現象發生，由於單核苷酸多型性序列資料的變異有限，一般進行疾病基因之多型性變化分析時會將整個染色體的 SNPs 進行基因定型(Genotyping)的動作。然而 SNP 資料太過龐大，因此在進行 genotyping 的過程中必然會花費大量的金錢和時間；為了解決上述情況，我們從原始龐大的 SNP 資料中挑選出有少量但具有足夠資訊可表示原始資訊的 SNPs，以達到降低 genotyping 所需成本的目的，這種挑選 SNP 的過程稱為 Tag SNP Selection。

在人類整個基因組中大約有數百萬個 SNP，然而這些相鄰的 SNP 位點會有整群遺

傳給下一代的特性。這些位於同一染色體上某一區域的一組具有相關性的 SNP，就稱之為單體型(haplotype)。它代表一個特定區域上的 SNP 組合，在臨床醫學應用上，可根據較少的疾病或病人檢體中，找到更為有用的變異組合，若能了解導致這些疾病的基因變異的話，人類便能夠對這類的疾病做出適當的治療和預防。單就藥物反應的觀點來看，Haplotype 確實是比 SNP 更具有統計上的意義。

一般而言，genotyping 挑選出來的 SNP 都是從單體型(haplotype)中挑選的，故又稱為 haplotype tagging SNPs (htSNPs)。為了挑選出最佳的 htSNPs，在過去出現了許多的方法，不斷被更新改進。其中有一種較被廣泛使用的方法是利用人類基因體中區塊狀結構的特性，也就是將人類基因體視為許多互斥區塊的集合，而且每一個區塊內包含一個很小的 common haplotype 集合(其集合可定義整個族群中該區塊至少 80%的變異)，即區塊的變異度(haplotype diversity)都相當有限[2][13]。然而在使用上述的方式之下會產生一個問題，就是每一個區塊內的變異度相當的低，使得同一區塊內所有 SNP 中會有許多重覆且多餘的資訊，因此必須從每個區塊中僅挑選出足以分辨所有 common haplotype 的 SNP，藉此達到特定的比率[4][15]。

挑選 htSNPs 的另外一種方法是透過計算 SNP 之間關聯性的高低 (如：Linkage Disequilibrium ; LD)，這種方法主要目的是讓所有 SNP 都能夠和被挑選出來的 htSNPs 集合中至少有一個 SNP 具有某種程度的關聯性(如：r^2 = 0.8%) [1][6]。透過這類的方法，htSNPs 集合即可能間接把和該疾病有關的 SNP 推測出來。直到最近幾年，開始利用從所有 SNP 中被挑選出來的 htSNPs 來預測剩餘未被挑選的 SNPs，並且以預測的平均錯誤率降至最小為目的來挑選最佳的 htSNPs [5][14][16]。其中，Halldorsson 所提出的方法，不同於一般 block-base 方式，而是屬於 block-free 的方法。一般 block-base 在挑選 htSNPs 時會限制只能夠在同一個區塊內選取，區塊和區塊間在挑選 htSNPs 時不會互相影響；然而 block-free 則能夠在整個基因體中挑選 htSNPs，而不會侷限在同一個區塊中。

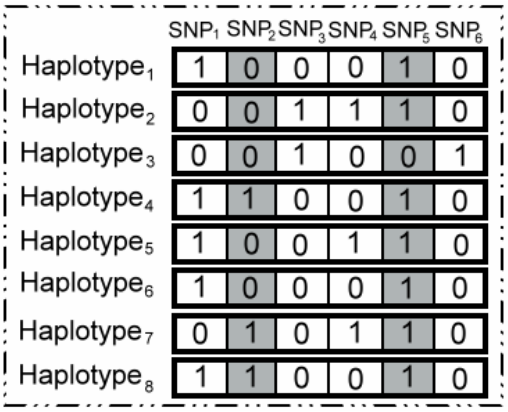
在本研究中，我們從所有和疾病相關聯的 SNPs 中挑選出部份的 SNPs 子集合，並且使這個子集合所蘊含的資訊量與原本 SNPs 集合所包含的資訊量之間的誤差達到最小。本研究以改良式的粒子族群最佳化演算法來挑選出 htSNPs，並且仿效 Halldorsson 等人的方法利用從所有 SNP 中被挑選出來的 htSNPs 來預測剩餘未被挑選的 SNP，有效的降低其預測的平均錯誤率，挑選出最佳的 htSNPs。再根據文獻(Oh 2004)[12]中所提出的區域搜尋方法，將原先的設定做簡化及改良，以較簡單的方式直接增加或刪除 tag SNP 的數量來節省運算時間，並維持原先區域搜尋的效果，對染色體做最後的修正。

2.相關研究探討

2.1 重建 Haplotype 的方法

目前已有許多重建 Tag SNP 的方式被提出，例如以正確率來考量的重建方式，透過某些特定數量的 Tag SNP 選取，用來重建整個 haplotype，並且將重建後的錯誤率降至最低。一般 genotyping 所需花費的時間非常久，加上資料集非常龐大，是一種效率較差的方式。為提升選取的效率，只利用將被挑選的 Tag SNP 來進行 genotyping，剩下沒被挑選的 SNP 則是直接由 tag SNP 的 allele 來進行預測，如此將可達到高效率和高正確率的目的。

如圖 1 所示，設有 8 個 haplotype，每個 haplotype 分別有 6 個 SNP，利用挑選 SNP_2、SNP_5 作為 tag SNP。圖 2 為進行測試的 haplotype，每個 haplotype 中的 SNP 以 0 和 1 來區分；皆以 test $haplotype_1$ 為重建的例子，由於 test $haplotype_1$ 的 SNP_2 和 SNP_5 分別為 0 和 1，因此我們從圖 1 的 8 個 haplotype 中挑選出 SNP_2、SNP_5 均分別為 0 和 1 的 haplotype (挑選結果為 $haplotype_1$、$haplotype_2$、$haplotype_5$、$haplotype_6$ 符合)，如圖 3。接著將此 4 個 haplotype 中 4 個未挑選的 SNP (SNP_1、SNP_3、SNP_4、SNP_6) 進行投票。以 SNP_1 為例，四個 haplotype 的 SNP_1 中有三個為 1，一個為 0，如此經過投票，我們預測 test $haplotype_1$ 的 SNP_1 為 1；同理，針對 test $haplotype_1$ 的 SNP_3 進行投票，結果僅有一個 haplotype 的 SNP_3 為 1，剩餘的都為 0，因此 test haplotype 的 SNP_3 預測結果為 0。如此方法可將剩餘的 haplotype 之 SNP 全部預測出來，之後再根據預測的統計結果計算出預測的正確率，以評估該 tag SNP 的好壞。



Tag SNPs：SNP_2、SNP_5
Tagged SNPs：SNP_1、SNP_3、SNP_4、SNP_6

圖 1 Haplotype Dataset

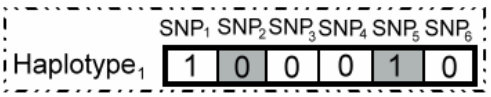


圖 2 進行測試的 haplotype

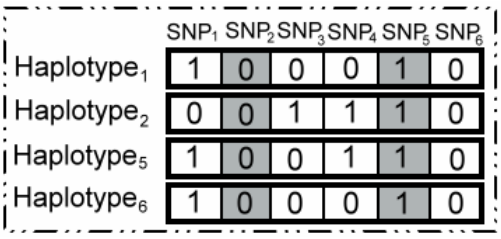


圖 3 挑選後的 Haplotype

2.2 STAMPA

STAMPA 是由 Eran Halperin 等人所提出的 tag SNP 選取方法[3]，是一種新的 haplotype 預測的方法，並且結合了動態規劃來選取適當數量的 tag SNP。目的是以能夠找到最大 haplotype 預測正確率的 tag SNP 為出發點所設計，但是在測試的過程中我們發現 haplotype 平均預測正確率並不高，因此我們利用 STAMPA 挑選 tag SNP 的方法並加入粒子最佳化演算法來選取更好的 tag SNP 組合，以改善 haplotype 預測平均正確率。

2.3 GEVALT

GEVALT 是由 Haploview 和 Gerbil 還有 STAMPA 所構成的程式，程式可從網站 http://acgt.cs.tau.ac.il/gevalt/下載。將 GEVALT 裡 STAMPA 所計算之正確率數值與本研究提出的方法 EBPSO 對於所求得的正確率做比較，顯示出改善的結果。

3.研究方法

3.1 Haploview

為了確保資料集內所有 SNP 的品質都在一定的水準之上，每個測試資料在透過 Gerbil 進行 haplotype phasing 之前，利用 Haploview [8]來進行篩選，並對每個 SNP 進行品質控制，如此解決了因為採樣的不足或是因為資料遺失使該 SNP 變的沒有意義的情形。

3.2 二進制粒子族群最佳化(Enhanced Binary Particle Swarm Optimization)

以下簡述本研究所提方法，Kennedy 和 Eberhart 於 1997 年提出了離散二進制粒子族群最佳化(Binary Particle Swarm Optimization, BPSO)，使得粒子族群最佳化更容易解決組合最佳化的問題。

假設族群大小為 m 並且在 d 維的搜尋空間中搜尋最佳解，則粒子的位置和速度表示形式如下：

$$X_{id} = \{X_{i1},\ X_{i2},\ ...,\ X_{id}\}$$

$$V_{id} = \{V_{i1},\ V_{i2},\ ...,\ V_{id}\}, i = 1,\ 2,\ ...,\ m$$

X_{id} 為粒子位置的集合，而 V_{id} 為速度的集合，在每一次迭代中，粒子透過下列(1)、(2)兩個公式來更新速度及位置：

$$V_{id}^{t+1} = W \cdot V_{id}^{t} + C_1 \cdot rand \cdot (pBest_{id} - X_{id}) + C_2 \cdot rand \cdot (gBest - X_{id}) \quad (1)$$

$$S(V_{id}^{t+1}) = \frac{1}{1 + e^{-v_{id}^{t+1}}} \quad (2)$$

$$if\ (rand < S(v_{id}^{t+1}))$$
$$then\ \ X_{id}^{t+1} = 1\ \ ;\ \ else\ \ X_{id}^{t+1} = 0 \quad (3)$$

其中公式(2)、(3)的 $S(v_{id}^{t+1})$ 為 Sigmoid 函數，此函數能夠將輸入的值限制於 0 和 1 之間，因此，透過此函數讓粒子更新過後的速度限制在[0, 1]的範圍。

本研究提出的一種改良式 BPSO，與原先 BPSO 不同的是加上了相互之間的資訊交換，其速度更新公式如下公式(4)：

$$V_{id}^{t+1} = W \cdot V_{id}^{t} + c_1 \cdot rand \cdot (pBest_{id} - X_{id}) + c_2 \cdot rand \cdot (gBest - X_{id}) + c_3 \cdot rand \cdot (mBest - X_{id}) \quad (4)$$

簡單來說，在過去使用的 BPSO 只應用到了各個粒子的 pBest 和群體的 gBest 來影響每個粒子下一次移動的速度產生，其結果有可能落入區域最佳解，或者是無法確實及有效的找尋最佳解，而產生一個搜尋上的盲點；在公式(4)可發現一個式子 $c_3 \cdot rand \cdot (mBest - x_{id})$，就是將當次迭代之中，將各個粒子更新後的位置多作一次比較，找出最好的解作為 mBest 使每個粒子加上了一個影響的條件，如此不但能避免落入區域最佳解，還能增加其搜尋到最佳解的機率。此想法源自於基因演算法(Genetic Algorithm, GA)的交配模式，可以發現其主要跳脫區域最佳解的機制在於交配和突變兩個過程，所使用到的資訊也是目前染色體所擁有的資訊，然而原始的 PSO 並無交配和突變的動作，利用這個觀念，如果仿照 GA 模式，增加目前迭代中每個粒子移動的資訊給其他粒子參考，便能與 GA 有相似的最佳化求解優勢，因此便考慮到 PSO 在做最佳化的過程當中，加上了目前每個粒子相互之間有用的資訊做為條件，去影響每個粒子移動的速度，彌補了原始 PSO 中沒有交配的程序，如此將更有效的找尋到最佳解。而 mBest 的值有兩種可能的情形，一種即本身的值就等於全域最佳解(gBest)，另一種是比 gBest 稍微差的結果，雖然後者適應值較 gBest 差，但在整體上並非必需排除的，由研究中可以得知它也會間接的影響粒子往最佳解的方向移動，這也就是主要將 mBest 加入速度更新公式的原因，將其相互之間的資訊再做交換，避免落入區域最佳解的情形，在搜尋最佳解的過程上扮演著提供重要且有用資訊的角色。

下圖 4 顯示，有 A、B、C 三個粒子，X 為最佳解的位置，經過第一次移動後，產生了 gBest 於粒子 A 的位置，而 mBest 即為所有粒子當次迭代過後較佳的粒子，也就是 C_1 的位置；以 B 粒子來說，經過一次移動到了 B_1 的位置，對於下一次的移動，若是一般的 PSO 情形，B_1 會受 gBest 的影響移動到 B_2，但此時並不為最佳解，若 B_1 受到 gBest 和 mBest 的影響，則會移動到 B_2'的位置，成為新的 gBest。藉此說明了當 mBest 和 gBest 不為相同粒子的情況下，會使粒子跳脫 gBest 的影響往更好的方向移動，又因為 mBest 恆小於或等於 gBest，所以此方法能降低粒子落入區域最佳解的機率。

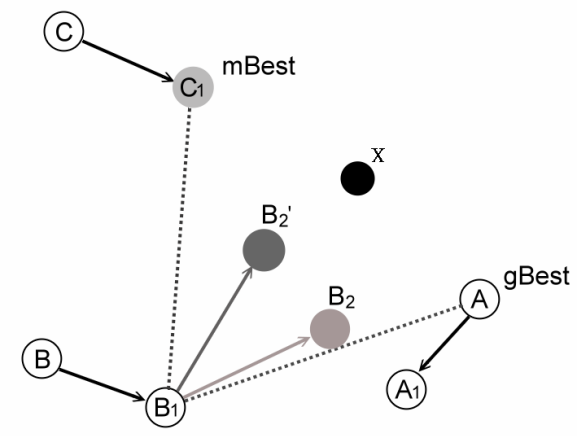


圖 4 EBPSO 粒子移動圖

單核苷酸多型性標籤選取問題的目的在於從大量的 SNP 資訊集合中，挑選出最有利用性、資訊量最高的 SNP，即挑選 tag SNP 的過程。在機器學習的領域當中特徵選取問題是屬於最佳化問題的一種，因此可以使用最佳化演算法來解決單核苷酸多型性標籤選取問題，本研究便透過粒子最佳化來改善文獻(Eran Halpern 2005)[3]中利用動態規劃法挑選 SNP 時造成平均預測正確率較差的情形。

3.2 粒子初始化

為符合 PSO 問題形式，必須先設計出粒子的模式，以二進制的編碼方式來設計 PSO 的粒子，其概念如下：

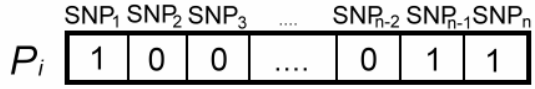


圖 5 粒子的設計圖

如上圖 5 所示，每個粒子皆以二進制的方式來設計 $P_i = \{p_{i1}, p_{i2}, \ldots, p_{in}\}$ 且 $p_{in} = \{0, 1\}$，其中 i 表示粒子的編號，n 則表示 SNP 數量；若假設 $p_{in} = 1$ 則表示第 i 個粒子的第 n 個 SNP 被挑選為 tag SNP，反之亦然。舉例來說，假設一粒子的內容如 $P_i = \{1, 0, 1, 0, 0, 1, 0\}$ 即粒子中被選為 tag SNP 的 SNP 為 $\{SNP_1, SNP_3, SNP_6\}$，剩餘的 SNP 則是將要被預測的 SNP，即 tagged SNP。

使用 GEVALT 內的 Grebil 工具，執行

genotype 轉 haplotype 的動作，接下來是用 STAMPA 產生一組解加入 PSO 初始流程的粒子當中。接著進行初始化粒子，一般粒子的初始化方式都是以隨機的方來產生，但是每個資料集的 SNP 數量皆不同，且所有挑選的 tag SNP 數量都是以相同的機率來產生的話那效率會變的很差，為了減少修正粒子所需要的時間，在初始化粒子時將挑選該 SNP 為 tag SNP 的機率以 *t*/*n* 來計算，*t* 表示所要挑選的 tag SNP 數目，*n* 則表示 SNP 的數目，如此一來一旦想要挑選的 tag SNP 數目越少，每個 SNP 會被挑選為 tag SNP 的機率便會隨之降低；反之，若想要挑選的 tag SNP 數目越多，每個 SNP 會被挑選為 tag SNP 的機率越高，如此便能夠有效的降低修正 tag SNP 數量所花費的時間，也挑選了適量的 tag SNP。為了確保經過 PSO 演算過程後的每個粒子所代表的 tag SNP 數量維持在一定的個數，我們新增了一個修正的步驟，也就是在 PSO 迭代過程中，同時再進行**區域搜尋(local search)**的動作。過程中會根據粒子本身所對應的 tag SNP 數目與目前所要挑選的 tag SNP 數目做比較，若目前挑選的 tag SNP 數量大於所要挑選的 tag SNP 數目，則進行刪除的步驟，反之則進行增加 tag SNP 的步驟。在修正的兩個程序(新增、刪除)中，若要進行刪除一個 tag 的動作，則會依序將 tag SNP 中所有為 1 的基因輪流設為 0，並且計算該粒子的適應值，如圖 7 所示，粒子中若有四個 1，則會產生四種不同可能染色體且每個粒子都有自己的適應值，最後挑選出適應值最高的粒子作為修正的結果；同樣的，若是新增一個 tag 則會以粒子中為 0 的位置來依序更改為 1，如圖 8 所示，將原來的粒子中四個為 0 的基因輪流設為 1 並且計算各自的適應值，最後再挑選出最佳的粒子作為修正後的結果。從上面敘述和圖中很容易發現移除和新增的過程，一次只能對一個 tag SNP 做新增或刪除的動作，因此在修正的步驟當中，我們不僅要決定該粒子要使用新增或刪除步驟來進行修正，同時也需要決定進行修正的次數以達到正確修正的目的。

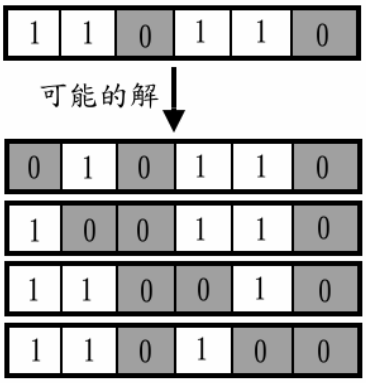


圖 7 刪除程序

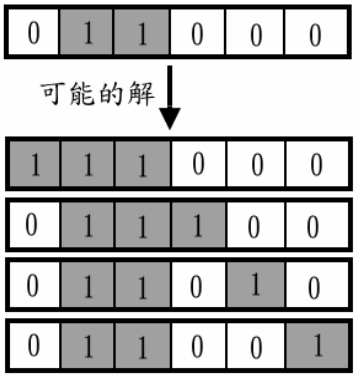


圖 8 新增程序

將修正後的 haplotype 傳入 GEVALT，使用 STAMPA 計算各粒子的適應值(也就是正確率)。此時即可依照適應值分別求出 pBest、gBest 和 mBest 的數值，將 pBest、gBest、mBest 和其它粒子資訊代入速度更新公式，更新每個粒子的位置及資訊，接著進行迭代流程，直到符合終止條件才結束。

4.結果與討論

4.1 測試資料集

本研究所使用的測試資料集是從多篇文獻或是 HapMap 網站中所取得的 genotype 資料集，包含了 LPL、STEAP 及取自(S.B. Gabriel 2002)[15]中的 PopulationD 取用 10 個資料集(D1、D8、D9、D10、D28、D29、D32、D41、D42、D49)進行測試，因此於本研究中總共使用了 12 個測試資料。

4.2 參數設定

為了證明 EBPSO 的求解能力，使用

STAMPA 加上 EBPSO 進行預測，又因必須公平的與其他文獻作比較，所以在本研究中 GEVALT 的各項參數，我們採用與該文獻作者所相同的初始設定，距離限制參數=30 (Eran Halpern 2005)[3]，即所有挑選出來的 tag SNP 當中選取任兩個 tag SNP，其距離皆不超過 30，而 EBPSO 部份粒子數目為 10、迭代次數為 50。

4.3 STAMPA、EBPSO 和 STAMPA-EBPSO 研究結果與比較

表 1 是採用 STAMPA 與 EBPSO 兩種方法實驗，並以 tagSNP 數為 2 的數據比較結果，可發現 EBPSO 方法就已經能在初期有提升的效果。表二是針對正確率達到 100%所需的 tagSNP 個數比較，發現 STAMPA-EBPSO 可產生優於 STAMPA 或者 EBPSO 的效果。

表 1. STAMPA 和 EBPSO 在 2 個 tagSNP 執行下的預測正確率比較

資料集	tagSNP 數為 2	
	STAMPA	EBPSO
PopulationD -D1	61.24	65.33
PopulationD -D8	75.66	78.78
PopulationD -D9	68.48	73.94
PopulationD -D10	79.74	84.96
PopulationD -D28	72.91	76.41
PopulationD -D29	75.81	77.47
PopulationD -D32	73.85	74.31
PopulationD -D41	78.27	79.10
PopulationD -D42	69.13	70.83
PopulationD -D49	78.46	83.50
LPL	82.93	84.53
STEAP	84.00	90.67
Averge	75.04	78.32

表 2. STAMPA、EBPSO 和 STAMPA-EBPSO 方法達到正確率 100%所需 tagSNP 個數比較

資料集	tag SNP 個數		
	STAMPA	EBPSO	**STAMPA-EBPSO**
PopulationD -D1	40	40	**40**
PopulationD -D8	41	41	**41**
PopulationD -D9	48	48	**48**
PopulationD -D10	35	35	**35**
PopulationD -D28	64	64	**64**
PopulationD -D29	50	50	**50**
PopulationD -D32	49	50	**49**
PopulationD -D41	76	**75**	**75**
PopulationD -D42	76	76	**76**
PopulationD -D49	13	13	**13**
LPL	47	48	**47**
STEAP	15	15	**15**

針對下圖 9 可發現，EBPSO 在初期的效果很好，但在中後段平均上卻劣於 STAMPA，因此才將 EBPSO 和 STAMPA 做結合，取兩方法的優點改善正確率，在 4.4 節會做更詳盡的分析。

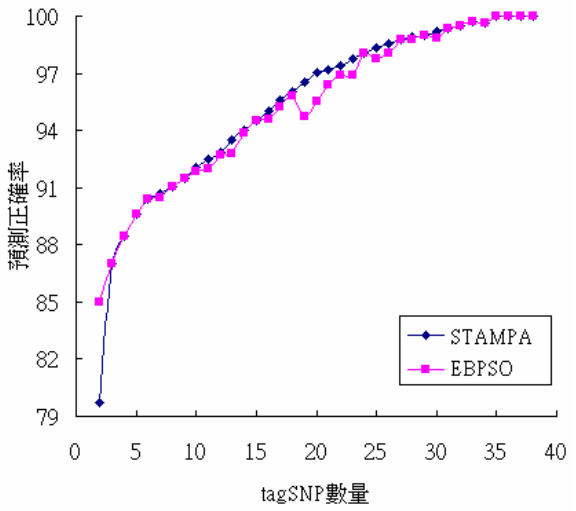


圖 9 STAMPA 和 EBPSO 於 PopulationD-D10 的結果比較圖

4.4 討論

本研究將原始 BPSO 的速度更新公式裡，加入粒子相互之間的資訊交換，結果顯示增強了 PSO 的效能，並將其應用於 tagSNP 選取及預測正確率的問題上，將 STAMPA 與本研究的 EBPSO 做結合，成為新的預測方法；參考文獻(Eran Halpern 2005)[3]的方式並將其演算法進行改善與比較，文獻中所使用的單核苷酸標籤選取方法是先設計了一套新的重建

haplotype 方法，並結合 STAMPA 的動態規劃手法進行單核苷酸標籤選取。為了改良該文獻所提出的動態規劃單核苷酸標籤選取方法，我們以延用了原作者的動態規劃法並且透過特徵選取的觀念，並加入改良式的 EBPSO 最佳化演算法，再與 STAMPA 所執行後的結果相比較，獲得更高的平均預測正確率。舉圖例來說，最直觀的從 X 軸和 Y 軸的關係可以發現，tagSNP 數量越多，則預測出的正確率就越高，接著比較 STAMPA 和本研究提出的 EBPSO 方法結合產生的結果；在最少 tag SNP 的情況之下所呈現的預測正確率，由於 STAMPA 挑選 tag SNP 時至少要挑選兩個 tag SNP 才能夠進行預測，因此我們就以最低限度的兩個 tag SNP 作為初始，又可由表 1 結果顯示，在兩個 tag SNP 的條件下，只使用 EBPSO 方法，平均下來就可以提升 3.28%，雖然整體正確率曲線差距不大，但是使用 EBPSO 的演算法所呈現的結果，就已經能找到比較好的正確率，證明 EBPSO 可以在少量 tagSNP 的情況下提升正確率。由表 2 可以發現，某些資料集在使用 STAMPA、EBPSO 和 STAMPA-EBPSO 各自達到 100%正確率的 tagSNP 數量並不相等，例如在 PopulationD-D41 資料集的結果可發現，新的 EBPSO 方法即可以利用 75 個 tagSNP 就達到 100%的正確率，若只是用 STAMPA 方法則需 76 個 tagSNP 才可達成，可得知 EBPSO 在達到最高正確率速度方面也有改善。

下圖 10 和圖 12 分別為 PopulationD-D41 和 PopulationD-D42 以原始 STAMPA 和 STAMPA+EBPSO 方法所呈現的結果，將少量 tagSNP 的線段放大，並由圖 11 和圖 13 可以發現在 5 個 tagSNP5 之前有明顯的改善。另外兩張圖(圖 14、圖 15 分別是使用 STEAP、PopulationD-D49 兩個資料集做實驗，可以直觀的看出，不論是用 STAMPA 或者 STAMPA-EBPSO 來執行，結果一樣顯示在少量 tagSNP 時預測值同樣會大於原始 STAMPA，並且在最差的情況也只是和 STAMPA 的結果一樣，因此整體的預測結果定不會比 STAMPA 差。圖 16 為 PopulationD-D42 使用 BPSO 方法的結果，發現只使用 BPSO 方法呈現的線段，則會不如 STAMPA 或 STAMPA-EBPSO 的線段那樣平滑，是因為 PSO 在移動過程中，漸漸修正位置的情形，如果在某次迭代粒子位置變差，在下次移動中將會引導粒子往更好的位置移動，如此線段才會顯示抖動不平滑，整體平均上就因抖動情形而相對較差，所以才需要 STAMPA 與 BPSO 的結合，又由於 EBPSO 對於找尋最佳解的效果又高於 BPSO，不但能提高正確率，又能使曲線如 STAMPA 般平滑，因此運用兩方法的優點作綜合性提升正確率。觀察整個 tagSNP 數量與預測正確率的關聯曲線，可以發現如果以原始 STAMPA-EBPSO 方法去預測正確率，相較於 STAMPA 方法的優勢在於 tagSNP 數量較少的時候，會有優異的正確率表現，證明能夠有效的以少量的 tagSNP 去預測其它 SNP。隨著 tagSNP 各數增加，結果也會有些微的改善，雖然不如少量 tagSNP 的改善幅度，但最差的結果也只是和 STAMPA 一樣，再加上 tagSNP 數量增加的預測結果，需要預測的 SNP 數量相對減少，因此整體上結果是有改善的。

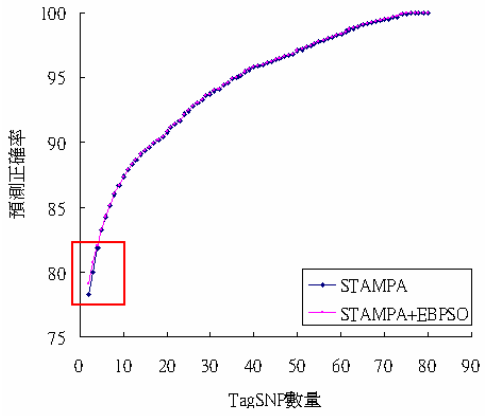


圖 10 STAMPA 和 STAMPA-EBPSO 用於 PopulationD-D41 的結果比較圖

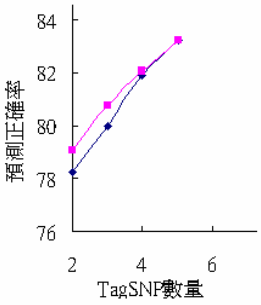


圖 11 PopulationD-D41 在 5 個 tagSNP 情況下的放大結果比較圖

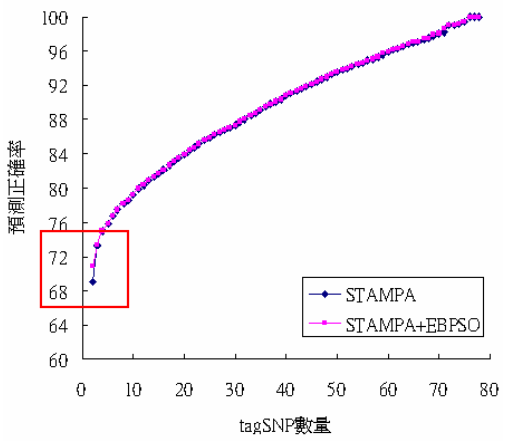


圖 12 STAMPA 和 STAMPA-EBPSO 用於 PopulationD-D42 的結果比較圖

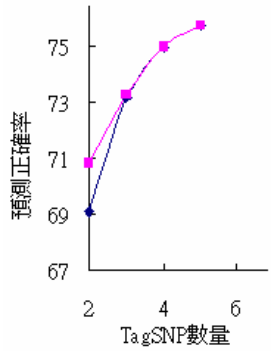


圖 13 PopulationD-D42 在 5 個 tagSNP 情況下的放大結果比較圖

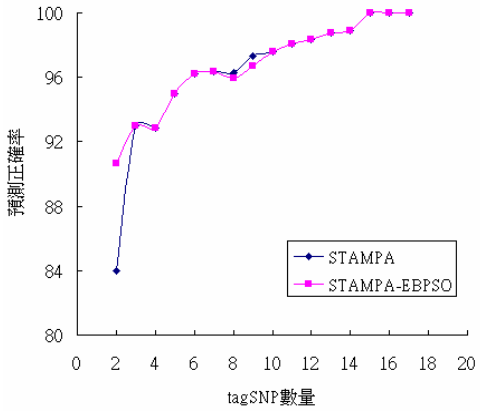


圖 14 STAMPA 和 STAMPA-EBPSO 用於 STEAP 的結果比較圖

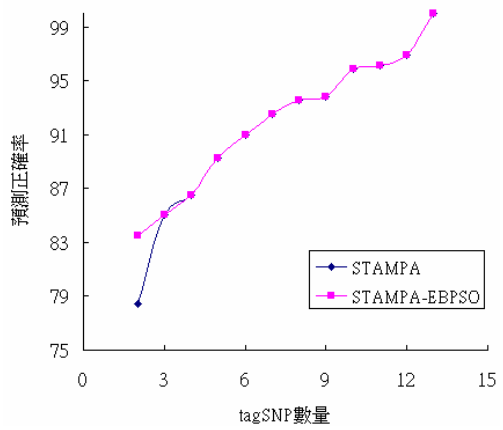


圖 15 STAMPA 和 STAMPA-EBPSO 用於 PopulationD-D 49 的結果比較圖

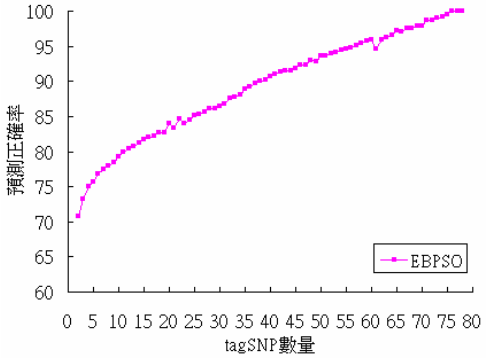


圖 16 EBPSO 用於 PopulationD-D42 的結果圖

5.結論

在本文中，我們提出了一個改良 BPSO 的演算法，為 EBPSO，以 mBest 加入速度更新的條件，降低落入區域最佳解的情形，並將此方法加以應用於單核苷酸多型性標籤選取問題；原始問題是以 STAMPA 方法去預測正確率，優點在於對多數 tagSNP 的預測效果較為穩定，但初期預測效果並不理想，在本實驗中將新的 EBPSO 方法於同樣條件下進行預測，結果顯示，EBPSO 方法在挑選極少數量的 tagSNP 情況下平均預測效果有顯著的提升，並且能在某些資料集中以較少的 tagSNP 達成 100%的正確率，因此結合 STAMPA 與 EBPSO 方法，綜合兩方法在各方面的優勢，有效的提升整體預測正確率。

參考文獻

[1] Carlson, C. S., M. A. Eberle, M.J. Rieder, Q. Yi, L. Kruglyak, Nickerson DA “Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage

disequilibrium," ***Am J Human Genet,*** 74: 106-120, 2004.

[2] Daly, M. J., J. D. Rioux, S. F. Schafiner, T. J. Hudson, and E. S. Lander "High-resolution haplotype structure in the human genome," ***Nature Genetics,*** 29(2): 229-232, 2001.

[3] Halperin, E., G.. Kimmel and R. Shamir, "Tag SNP selection in genotype data for maximizing SNP prediction accuracy," ***Bioinformatics,*** 21: 195-203, 2005.

[4] Johnson, G.. C., L. Esposito, B. J. Barratt, et al. "Haplotype tagging for the identification of common disease genes," ***Nat Genet.,*** 29(2):233- 7, 2001.

[5] Halldorsson, BV., V. Bafna, R. Lippert, R. S. Schwartz, F. M. D. L. Vega, A. G.. Clark, and S. Istrail, "Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies," ***Genome Research,*** 14: 1633-1640, 2004.

[6] Chapman, J.M., J.D. Cooper, J.A. Todd, D.G.. Clayton, "Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power," ***Human Heredity,*** 56: 18-31, 2003.

[7] Kennedy, J., and R.C. Eberhart, "A discrete binary version of the particle swarm algorithm," ***IEEE***: 4104-4108, 1997.

[8] Barrett JC., B. Fry, J. Maller, MJ. Daly, "Haploview: analysis and visualization of LD and haplotype maps," ***Bioinformatics,*** 21(2): 263-265, 2005.

[9] Zhang, K., T. Chen, M. Waterman, and F. Sun, "A set of dynamic programming algorithms for haplotype block partitioning and tag SNP selection via haplotype data or genotype data," ***In Proceedings of Discrete Mathematics and Theoretical Computer Science Workshop on SNP,*** Rutgers University Busch Campus, Piscataway,NJ, USA, pp. 1–26, 2003.

[10] Burkett, KM, M. Ghadessi, B. McNeney, J. Graham, D. Daley, "A comparison of five methods for selecting tagging single nucleotide polymorphisms," ***BMC Genetics,*** 6(1): S71, 2005.

[11] Stephens, M., N.J. Smith, P. Donnelly, "A new statistical method for haplotype reconstruction from population data," ***America journal of Human Genetics,*** 68(4): 978-989, 2001.

[12] Oh, I.-S., J.-S. Lee, and B.-R. Moon, "Hybrid Genetic Algorithms for Feature Selection," ***In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence,*** 26(11): 1424-1437, 2004.

[13] Patil, N., A. J. Berno, D. A. Hinds, et al., "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21," ***Science,*** 294: 1719-1723, 2001.

[14] Sebastiani, P., R. Lazarus, ST. Weiss, LM. Kunkel, IS. Kohane, MF. Ramoni, "Minimal haplotype tagging. Proceedings of the National," ***Academy of Sciences,*** 100: 9900-9905, 2003.

[15] Gabriel, S. B., S. F. Schaffner, H. Nguyen, et al. "The structure of haplotype blocks in the human genome," ***Science,*** 296: 2225-2229, 2002.

[16] Bafna, V., B. V. Halldorsson, R. Schwartz, A. G. Clark, S. Istrail, "Haplotypes and informative SNP selection algorithms:don;t block out information," ***Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology,*** (RECOMB 03): 19-27, 2003.