應用二進制粒子族群最佳化於操作組之預測

莊麗月
義守大學 化學工程系
chuang@isu.edu.tw

蔡瑞鴻
國立高雄應用科技大學 資訊工程系
1097308102@cc.kuas.edu.tw

楊正宏
稻江科技暨管理學院 網路系統學系
國立高雄應用科技大學 資訊工程系
chyang@cc.kuas.edu.tw

***摘要*—操作組(Operon)在藥物設計及蛋白質功能的研究上，透露許多具有價值的資訊，而相同操作組基因擁有相同或類似的生物功能，且會共同轉錄成 mRNA 序列，因此預測操作組成為掌握轉錄規則之關鍵。然而，經實驗方式檢測操作組有難度且耗時，為降低生物學家對於了解轉錄規則的困難度，本研究提出以二進制粒子族群最佳化(Binary Particle Swarm Optimization, BPSO)應用於基因體(Genome)操作組預測，藉由相鄰基因的距離及方向作為初始化依據，使得粒子族群於初始化後即可獲得優越的結果。本研究以大腸桿菌進行訓練，以該基因體相鄰基因之距離、代謝路徑及基因長度作為適應函數的評分依據，分別測試枯草桿菌、綠膿桿菌及葡萄球菌基因體的預測正確率(Accuracy)、敏感度(Sensitivity)及特異度(Specificity)。實驗結果顯示，本方法不僅減少預測所需的時間，亦能獲得較高預測正確率。**

***關鍵詞*—二進制粒子族群最佳化、操作組、預測正確率。**

一、前言

操作組(Operon)在藥物設計及蛋白質功能的研究上，揭露許多具有價值的資訊，例如葡萄球菌為社區及醫院傳染病的主要病原體[19]。操作組為轉錄基本單元，操作組中的基因會共同轉錄成單股 mRNA 序列，因此想要了解轉錄規則，預測操作組成為一大關鍵。操作組中包含一個或多個同向連續的基因，通常相同操作組的基因不是擁有相同的生物功能，即可能為功能互相影響。在原核生物中，細菌的基因體是由數千個基因所組成，人類第一個發現的操作組是位於大腸桿菌中的乳糖操作組，主要功能是將乳糖分解成葡萄糖及半乳糖，其中包含啟動子(Promoter)、終止子(Terminator) 、操縱基因(Operator) 及三個連續的結構基因(Structure gene)。結構基因依序為 *lac*Z、*lac*Y 及 *lac*A，皆由位於啟動子前端的調節基因(Regulator gene)*lac*I 控制基因表現。當乳糖不存在時，調節基因會製造抑制蛋白(repressor)與操縱基因結合，使 RNA 聚合酶無法與啟動子結合以抑制基因表現。若乳糖存在時，乳糖會與抑制蛋白結合使其無法活化，進而順利進行轉錄。然而，目前所認知的操作組不多，且生物學家以實驗檢測操作組是相當困難且耗時[8]。因此如何利用生物資訊的技術發展一個有效的預測方法，成為當前相當重要且待解決的議題。

近年來，許多學者已提出許多個能夠準確預測操作組的特性，目前常使用的特性主要分為下列 5 種類別[2]：相鄰基因的距離(Intergenic distance)、基因族群的保存(Conserved gene clusters) 、相關功能(Functional relations)、以基因體序列為基礎(Genome sequence based)及實驗證據(Experimental evidence)。在上述類別中，最具代表性的生物特性是在 Operon 邊界預測啟動子及終止子序列[8]，最簡單的預測特性則是觀察操作組中相鄰基因(Within operon pair, WO pair)距離，是否小於轉錄單元邊界的相鄰基因(Transcription unit border pair, TUB pair)距離，其不僅為最簡單的預測特性，且僅使用距離辨識 Operon 的方式，也能獲得不錯的結果[2]。

目前許多文獻中亦提出預測操作組的方法，包含利用支撐向量機(Support Vector Machine, SVM)[23]、隱藏馬爾克夫模型(Hidden Markov Model, HMM)[21] 、貝氏網路法(Bayesian network approach)[1]、模糊基因演算法(Fuzzy Genetic Algorithm, GA)[8]及基因演算法(Genetic

Algorithm, GA)[18]，皆有不錯的預測結果。其中 Jacob *et al.* [8]提出以模糊基因演算法為基礎，利用四個生物特性設計適應函數，使用的生物特性分別為：相鄰基因的距離、代謝路徑(Pathway)、保存於多個基因體中(Conservation across multiple genomes)及蛋白質功能的相似度(Similarity of protein function)。此方法特點在於依據上述特性設計一套評分方法，不需藉由複雜的數學公式計算，也能正確對每條染色體(Chromosome)進行評分，使基因演算法能夠依此為據，經由不斷迭代來提升預測正確率。另外，基因演算法[18]所使用的四個生物特性分別為：相鄰基因的距離、代謝路徑、相鄰類基因功能的聚簇 (Cluster of orthologous groups gene function, COG)和微陣列表示資料(Microarray expression data)，其中利用區域熵值最小化方法(Local-entropy-minimization method)計算相鄰基因的距離分數，若相鄰基因具有相同代謝路徑給予 1 分，否則不予給分。COG 及微陣列表示資料則是分別利用對數似然法(Log-likelihood)及皮爾森相關係數(Pearson correlation coefficient)進行評分，將以上 4 個生物特性的分數相加，作為相鄰基因的分數，藉此計算假定操作組(Putative operon)，最後再將所有假定操作組分數相加作為染色體分數，利用這套評分機制作為評估染色體優劣依據。上述文獻僅使用距離進行初始化[18]，卻忽略方向對於操作組預測的重要性，使得基因演算法無法於初始化時獲得較佳的母體染色體，更因設定較低的交配率及突變率，較難藉由迭代過程找尋染色體最佳解。

本文提出一個有效的二進制粒子族群最佳化(Binary particle swarm Optimization, BPSO)方法，應用大腸桿菌(*Escherichia coli* K12-MG 1655)作為訓練資料，並以枯草桿菌(*Bacillus subtilis*)、綠膿桿菌(*Pseudomonas aeruginosa* PA01)及葡萄球菌(*Staphylococcus aureus*)作為測試資料。本研究在生物特性上的挑選，使用相鄰基因的距離、代謝路徑及基因長度(Gene length)作為計算適應函數值依據。經由實驗結果證明，使用距離、代謝路徑及基因距離進行操作組預測的結果，與其他使用 4 個特性的基因演算法進行比較，不僅減少預測所需的時間，亦能獲得較高預測正確率。

二、研究背景

(一) 問題定義

本研究將 OP 定義為陽性(Positive)，NOP 則定義為陰性(Negative)。圖 1 中箭頭代表基因，箭頭方向則為基因方向，其中灰色箭頭表示該操作組由一個基因組成，黑色箭頭表示該操作組由兩個以上基因組成，而白色箭頭則表示此基因尚未經實驗證實。由圖中可知，OP 是指位於操作組中的相鄰基因，而 NOP 必要條件與 OP 相同，即相鄰基因皆須位於相同方向，若操作組為單一基因組成，下一個基因為未知狀態，則將此對基因對稱為 NOP，但操作組由多個基因組成時，則會因操作組邊界尾端的不確定性[15]，無法將其稱為 NOP，而位於操作組第一個基因與前一個基因皆稱為 NOP。

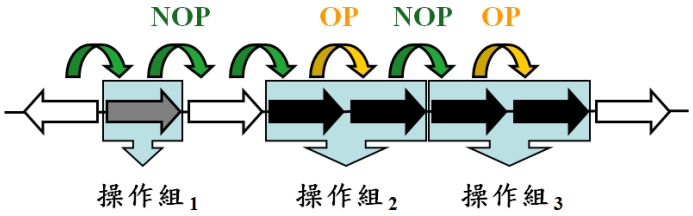


圖 1. 操作組對示意圖

(二) 生物特性

操作組中的基因具有許多相關生物特性，本研究所使用的特性為下列 3 項：

- **相鄰基因的距離：**此特性可預測具有完整基因體序列的操作組，於降解(Degradation)過程中保護 mRNA，因此位於相同操作組的基因具有距離較短之特性。在圖 2 中，$Gene_2$、$Gene_3$ 及 $Gene_4$ 位於相同操作組，因此三個基因之間的距離會比 $Gene_1$ 及 $Gene_2$ 或 $Gene_4$ 及 $Gene_5$ 之間的距離短，距離計算方式為相鄰基因的鹼基個數，會有重疊(Overlap)的情況發生。由圖 3 中我們可發現，OP 出現頻率最大的距離為-4[22]，而 NOP 距離分配頻率則是隨著距離增加，逐漸高於 OP 距離頻率，因此可藉由此特性辨別操作組。

圖 2. 操作組示意圖

- **代謝路徑**：相鄰基因若有相同的代謝路徑，則有可能為相同的操作組。
- **基因長度**：通常 TUB pair 會有較低的長度比例之自然對數值，計算方式為上游基因之長度除以下游基因之長度，接著再取自然對數，也就是說相鄰基因之長度會影響相鄰基因位於相同 Operon 的機率[4]。

(三) 正確率評估

陰陽性的評估方式如表 1 所示，若真實資料為陽性且評估正確，則稱為真陽性(True Positive, TP)，反之則稱為真陰性(True Negative, TN)。若真實資料為陽性中，卻評估為陰性，稱之為偽陰性(False Negative, FN)，反之則稱之為偽陽性(False Positive, FP)。藉由上述參數計算陽性預測率(Positive Prediction Rate)、陰性預測率(Negative Prediction Rate)、預測陽性能力的敏感度(Sensitivity, SN)、預測陰性能力的特異度(Specificity, SP)及評估整體預測能力的預測正確率(Accuracy, ACC)[4]，如表 2 所示。其中敏感度

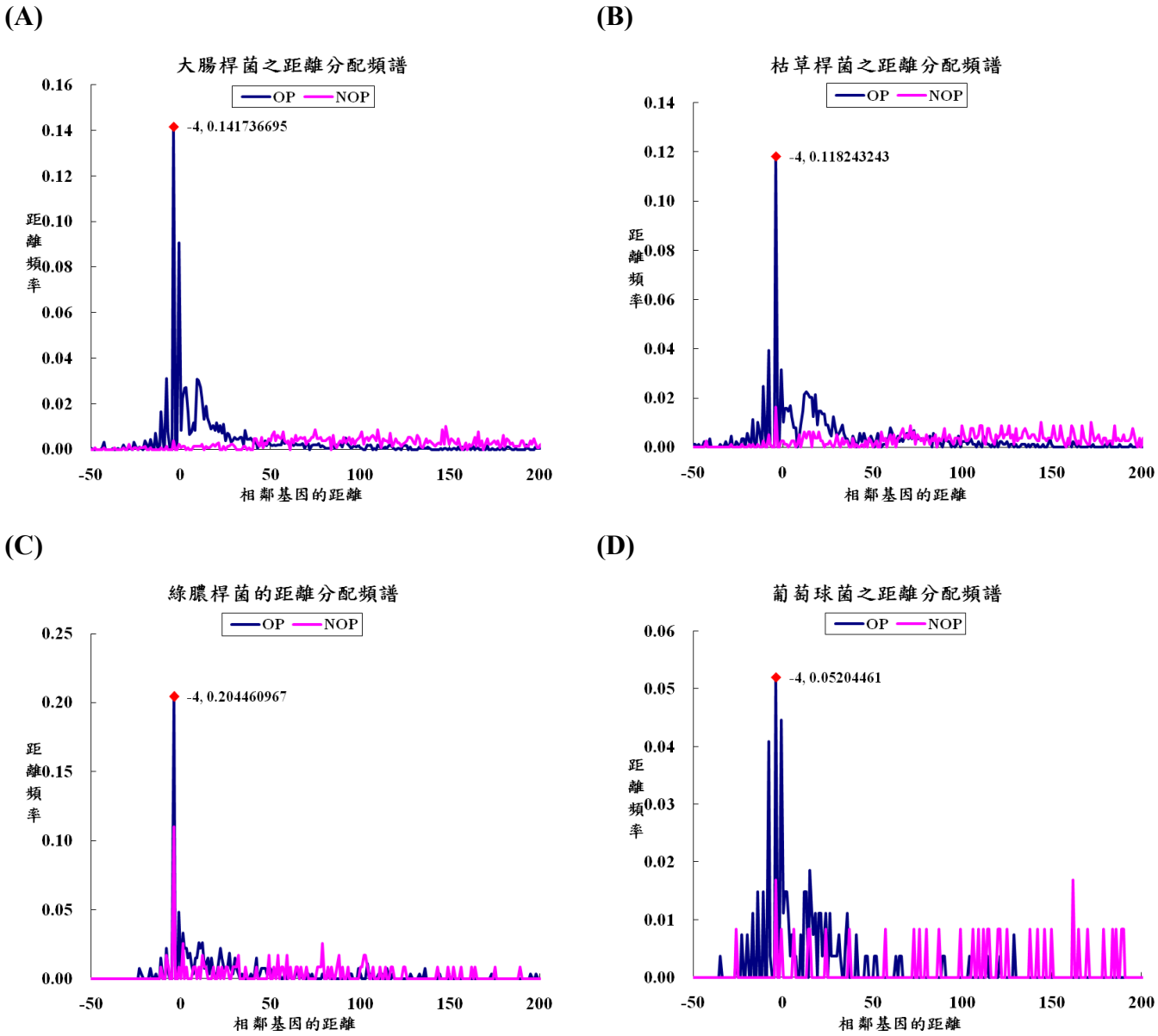


圖 3. 距離分配頻譜

表 1. 陰陽性評估表

真實資料 / 預測結果	陽性	陰性
陽性	真陽性(TP)	偽陽性(FP)
陰性	偽陰性(FN)	真陰性(TN)

表 2. 正確率評估公式表

評估值	評估公式
敏感度	SN=TP/(TP+FN)
特異度	SP=TN/(FP+TN)
陽性預測率	PPR=TP/(TP+FP)
陰性預測率	NPR=TN/(FN+TN)
預測正確率	ACC=(TP+TN)/(TP+FP+TN+FN)

與特異度存在相互拉扯(Trade-off)的關係，也就是說其中一者若提高，另一者必定會降低，因此需藉由接受器操作特性曲線(Receiver operating characteristic curve, ROC curve)表示所有敏感度及特異度的相互關係，將斜率為 1 的直線與 ROC 曲線相切，該切點則為兩者相加的最大值，當曲線下面積(Area under the curve, AUC)越大表示所得之預測正確率越高。

三、研究方法

(一) 資料來源

本研究大腸桿菌、枯草桿菌、綠膿桿菌及葡萄球菌基因體資料皆由 GenBank 資料庫(http://www.ncbi.nlm.nih.gov/)下載，分別擁有 4488、4225、5651 及 2845 個基因，其中記錄各個基因之定義、名稱、編號、起始位置、終止位置、方向及產物名稱。大腸桿菌及枯草桿菌的操作組實驗資料分別由 OperonDB (http://regulondb.ccg.unam.mx/) [13]及 DBTBS (http://dbtbs.hgc.jp/) [17]兩個最具代表性資料庫取得，*Pseudomonas aeruginosa* PA01 及 *Staphylococcus aureus* 基因體的操作組資料則由 ODB (http://odb.kuicr.kyoto-u.ac.jp/) [12]取得，其中記錄各操作組之名稱、基因個數、方向及基因名稱。代謝路徑及 COG 資料則分別由 KEGG (http://www.genome.ad.jp/kegg/pathway.html) 及 NCBI (http://www.ncbi.nlm.nih.gov/COG/)取得。其中大腸桿菌及枯草桿菌為目前已由實驗驗證之操作組[4]。因此本研究仿照文獻 GA 利用大腸桿菌基因體進行訓練[18]，並以枯草桿菌作為操作組之主要預測目標。此外，為了驗證 BPSO 較其他文獻優越[18][19]，本實驗亦加入綠膿桿菌及葡萄桿菌基因體進行預測。

(二) 二進制粒子族群最佳化

粒子族群最佳化 (Particle Swarm Optimization, PSO)為一種最佳化演算法[9]，最初概念源自於鳥類及魚類族群特性。在族群生活中，個體不僅會受到自身過去經驗及認知影響，亦受到該族群行為的影響。然而，PSO 實數編碼無法解決離散空間問題，因此 Kennedy 及 Eberhart 於 1997 年提出 BPSO，以克服 PSO 無法解決二進制編碼的問題[10]。BPSO 中的族群經由隨機方式初始化 N 個粒子後，接著每個粒子於 d 維空間移動搜尋，第 i 個粒子的位置及速度分別由 $x_i=(x_{i1},x_{i2},\ldots,x_{id})$ 及 $v_i=(v_{i1},v_{i2},\ldots,v_{id})$ 表示，且分別限制於 $[X_{\min},X_{\max}]^d$ 和 $[V_{\min},V_{\max}]^d$ 範圍中。經由迭代的更新及搜尋，每個粒子會將個體經驗記錄下來，而個體最佳適應函數值稱為個體最佳($pbest_i$)，族群中最佳 $pbest_i$ 則稱為群體最佳($gbest$)。BPSO 的更新公式如下：

$$v_{id}^{new}=w\times v_{id}^{old}+c_1\times r_1\times\left(pbest_{id}-x_{id}^{old}\right)+c_2\times r_2\times\left(gbest_{id}-x_{id}^{old}\right) \quad (1)$$

$$\text{if } v_{id}^{new}\notin(V_{\min},V_{\max})\text{ then } v_{id}^{new}=\max\left(\min\left(V_{\max},v_{id}^{new}\right),V_{\min}\right) \quad (2)$$

$$S\left(v_{id}^{new}\right)=\frac{1}{1+e^{-v_{id}^{new}}} \quad (3)$$

$$\text{if } \left(r_3<S\left(v_{id}^{new}\right)\right)\text{then } x_{id}^{new}=1\,\text{else}\,x_{id}^{new}=0 \quad (4)$$

公式(1)中，w 為 BPSO 之慣性權重，c_1 及 c_2 分別為 $pbest_i$ 及 $gbest$ 的學習因子，r_1 及 r_2 為 0 至 1 之間的亂數，v_{id}^{old} 及 v_{id}^{new} 分別為更新前及更新後的速度。公式(2)則用來限制更新後的粒子速度，使粒子速度能介於 $V_{\max}$ 及 $V_{\min}$ 之間。接著將更新後的速度代入公式(3)，會得到一個 0.0025 至 0.9975 之間的數值。最後根據公式(4)來判斷粒子

維度，其中 r_3 為 0 至 1 之間的亂數，若更新後的速度較快，相對 $S\left(v_{id}^{new}\right)$ 較大，則粒子為 1 的機率較高，但因限制速度的關係，使得 BPSO 最低也有 0.0025 的機會可以轉變狀態。BPSO 的詳細步驟分別介紹於下列小節。

2.1 編碼

本研究以二進制編碼方式建構粒子族群，當編碼為 1 時，表示此基因與下一個基因為 OP；編碼為 0，則表示此基因與下一個基因為 NOP，亦可視為該操作組的最後一個基因。如圖 4 所示，若產生的二進制粒子為 110010，表示 $Gene_1$、$Gene_2$ 及 $Gene_3$ 屬於第一個操作組，$Gene_4$ 自成為一個操作組，而 $Gene_5$ 及 $Gene_6$ 則屬於最後一個操作組，其中編碼為 0 的 $Gene_3$、$Gene_4$ 及 $Gene_6$，則為操作組的結尾基因。

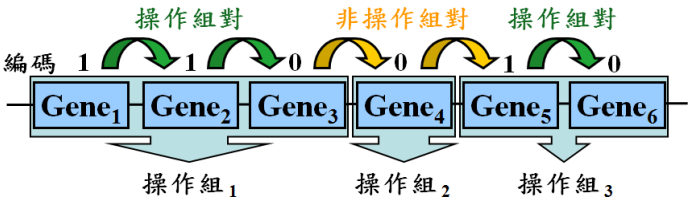


圖 4. 編碼示意圖

2.2 初始化

在本研究中，我們利用相鄰基因的距離及基因方向產生 20 個二進制粒子，每個粒子藉由隨機產生 0 到 600 的亂數進行初始化。如圖 5 所示，假設取亂數 75 作為判斷門檻值，$Gene_1$ 及 $Gene_2$ 的距離為 70，且於基因體中為同方向的情況，則將 $Gene_1$ 編碼為 1；而當 $Gene_2$ 及 $Gene_3$ 的距離為 80，且大於判斷門檻值，則將 $Gene_2$ 編碼為 0。若相鄰基因的距離小於亂數所設的門檻值，但方向相反，仍將其編碼為 0。計算距離的方式如公式 5 所示[16]，其中 $Gene_1$_finish 為上游基因的鹼基終結位置，$Gene_2$_start 則為下游基因的鹼基起始位置。

$$distance=Gene_2\_start - (Gene_1\_finish+1) \qquad (5)$$

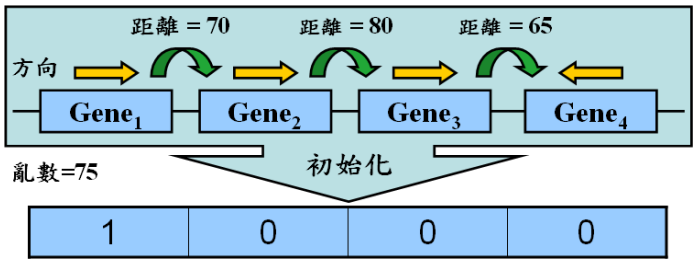


圖 5. 初始化示意圖

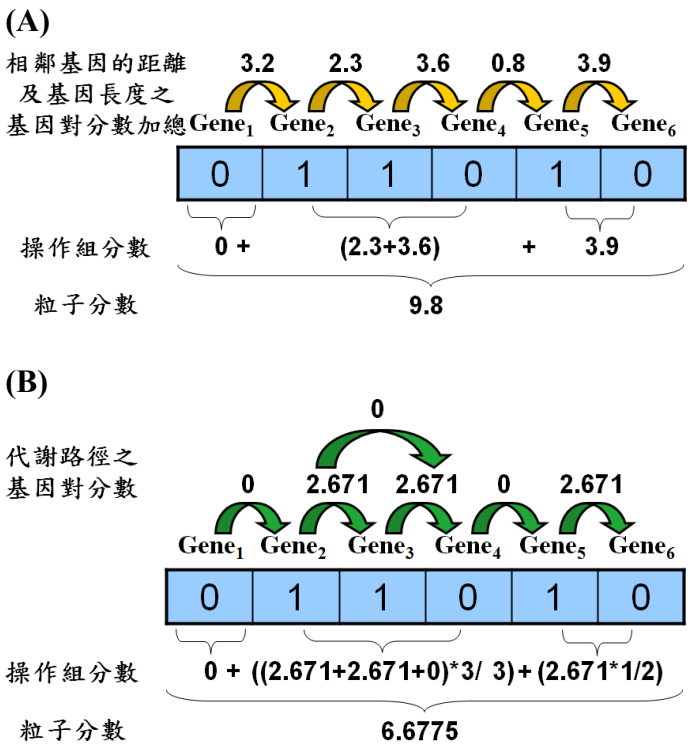
2.3 適應函數

由文獻得知，對於操作組的預測問題，相鄰基因的距離特性具有相當影響力 [1, 7, 19, 20, 22]，因此本研究以此特性為主，並加入其它文獻所使用的相鄰基因的距離、代謝路徑及基因長度特性進行實驗，特性的適應函數評估方式皆以

表 3. 相鄰基因的距離之評分表

間距	分數	間距	分數	間距	分數
[-∞, -99]	-0.82457	[30, 39]	0.568643	[170, 179]	-1.83357
[-100, -91]	0	[40, 49]	-0.67375	[180, 189]	-1.98772
[-90, -81]	1.478014	[50, 59]	-0.52852	[190, 199]	-1.51772
[-80, -71]	0	[60, 69]	-0.43437	[200, 209]	-2.35497
[-70, -61]	-0.31375	[70, 79]	-0.6435	[210, 219]	-1.98772
[-60, -51]	0	[80, 89]	-0.6322	[220, 229]	-3.4918
[-50, -41]	0.533552	[90, 99]	-0.55887	[230, 239]	-2.23556
[-40, -31]	-0.22673	[100, 109]	-1.48787	[240, 249]	-2.25966
[-30, -21]	0.379401	[110, 119]	-1.15683	[250, 259]	-2.79865
[-20, -11]	2.019145	[120, 129]	-1.43768	[260, 269]	0
[-10, -1]	2.22656	[130, 139]	-1.84221	[270, 279]	-3.33417
[0, 9]	2.2105	[140, 149]	-2.66512	[280, 289]	-2.1329
[10, 19]	2.340637	[150, 159]	-1.80384	[290, 299]	-2.83947
[20, 29]	1.564274	[160, 169]	-1.78965	[300, ∞]	-2.96611

註：利用對數似然法，針對每段距離間距進行評分。



註: 子圖(A)為相鄰基因的距離及基因長度的評分方法，子圖(A)為代謝路徑的評分方法。

圖 6. 適應函數示意圖

$$\mathrm{LL}_{\mathrm{Property}}\left(gene_i, gene_j\right) = \ln\left(\frac{N_{WO}(\mathrm{Property})/TN_{WO}}{N_{TUB}(\mathrm{Property})/TN_{TUB}}\right) \quad (6)$$

$$\mathrm{LL}_{glr}\left(gene_i, gene_j\right) = \ln\left(\frac{length_i}{length_j}\right), j = i+1 \quad (7)$$

對數似然法計算，評估方式如公式(6)，其中 N_{WO}(Property)及 N_{TUB}(Property)分別代表 WO pair 及 TUB pair 具有相同特性的基因對個數，而 TN_{WO} 及 TN_{TUB} 則分別代表 WO pair 及 TUB pair 的基因對個數。距離的適應函數如表 3 所示，若相鄰基因的距離介於該間距中，則以該間距所得的分數作為評分依據。例如 OP 的距離為 5，該距離介於 0 至 9 之間，所得的分數則為 2.2105。根據公式計算，若兩基因具有相同代謝路徑則給予 2.671，否則不予計算其適應函數。公式(7)則是代入相鄰基因的長度，來評估該粒子適應值，分子為上游基因長度，分母為下游基因長度。以圖 6 為例說明詳細的評分程序，粒子編碼為 011010，以對數似然法評估各個相鄰基因對的分數後，計算各操作組中基因對的平均分數。由於本研究不將單一基因的操作組列入評分，故第一個操作組的分數為 0。如圖 6 所示，相鄰基因的距離及基因長度特性皆僅評估與下游基因的關係，代謝路徑特性則評估假定操作組內的關係，分別計算粒子分數後，最後將(A)及(B)粒子分數加總，即為該粒子之適應值。

2.4 演算法流程

本研究方法之流程步驟依序如下：

步驟一：設定粒子族群最佳化的相關參數，分別為族群大小 N、迭代次數 G。

步驟二：建構 BPSO 粒子族群，產生 20 個亂數，其值皆介於 0 至 600 之間，根據相鄰基因距離及方向限制進行初始化。

步驟三：計算每個粒子的適應函數值。

步驟四：若第 i 個粒子之適應函數值較 $pbest_i$ 佳，則進行取代 $pbest_i$。

步驟五：若某一粒子之 $pbest_i$ 較 $gbest$ 佳，則以 $pbest_i$ 取代 $gbest$。。

步驟六：根據更新公式(1)-(4)對每一個粒子的位置及速度進行更新。

步驟七：判斷是否達到所設定之迭代次數，若符合則終止程式，否則回步驟三。

四、結果與討論

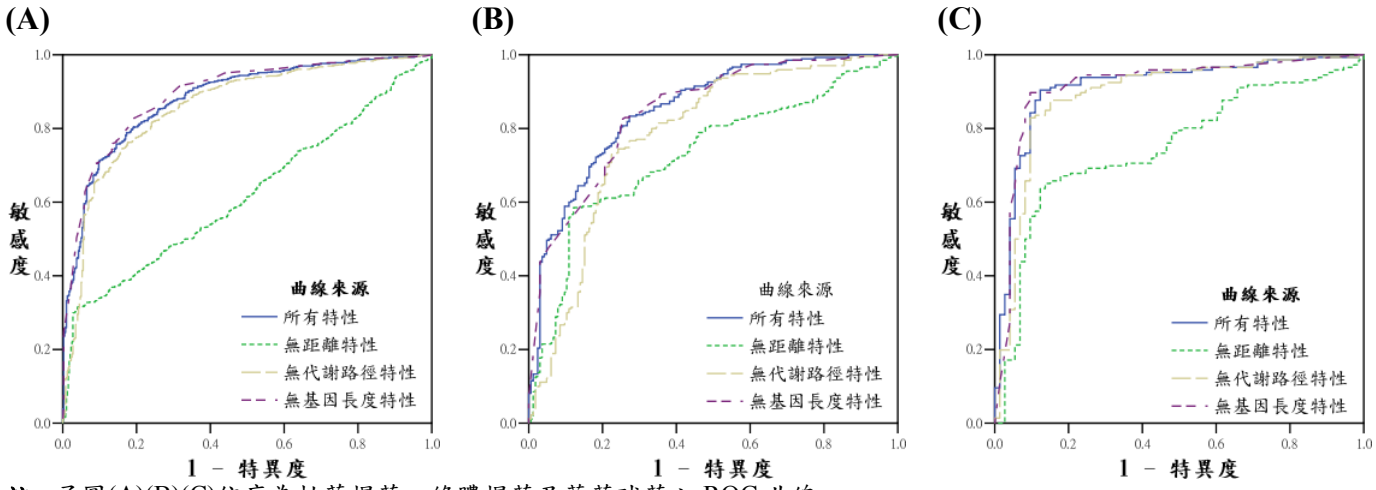
(一) 參數設定

本研究於粒子族群最佳化的參數設定如下，粒子族群個數 $N = 20$，迭代次數 G = 100，慣性權重值 $w = 1$，學習因子 $c_1 = c_2 = 2$，r_1、r_2 及 r_3 皆為 0 至 1 之間的亂數，速度限制的 $V_{\max}$ 及 $V_{\min}$ 分別為 6 和-6[10]。由於文獻之初始化亂數門檻值設定為 300_{bps}[18]及 600_{bps}[8]，本實驗為與文獻預測結果進行比較，並突顯亂數門檻值對於初始化的改善，因此參照文獻設定初始化亂數的範圍。

(二) 結果

由圖 7(A)(B)(C)可分別看到枯草桿菌、綠膿桿菌及葡萄球菌使用所有特性及刪去某一項特性的

ROC 曲線，三個特性分別為基因間之距離、代謝路徑及基因長度。我們從圖中可發現，刪除距離特性的曲線下面積明顯減少，由此可知，刪除距離特性大幅降低了操作組的預測正確率，藉此可得知相鄰基因間之距離對於操作組預測的重要性。另外，亦可從圖中得知代謝路徑對於操作組的重要程度，雖然刪除該特性後，曲線下面積不如上述特性明顯減少，但仍可由 ROC 曲線得知代謝路徑特性對於操作組預測之貢獻。然而，比較所有特性曲線與無基因長度特性曲線的面積，雖然差距不大，但該特性與其他特性結合進行預測，仍對辨識操作組具有相當程度幫助。在本研究中，經由 BPSO 迭代過程，找尋適應函數值最佳的粒子，再與實驗證實的操作組資料進行比對，分別記錄 TP、FN、TN 及 FP，作為計算預測正確率、敏感度及特異度的依據。此外，本研究結果亦與已發表文獻進行比較，包括基因演算法[18]、模糊基因演算法[8]、支撐向量機[23]及其它預測方法[3-6, 11, 12, 14, 19, 20]，以突顯本研究方法之效能。在表 4 中，我們列出文獻於預測枯草桿菌操作組所使用之特性，由表中可觀察到大部份研究方法皆使用四個以上之特性進行預測，較本研究所使用的特性多。由表 5 得知，



註: 子圖(A)(B)(C)依序為枯草桿菌、綠膿桿菌及葡萄球菌之 ROC 曲線。

圖 7. ROC 曲線

表 4. 預測枯草桿菌操作組之特性使用列表

研究方法	使用特性
BPSO	ID, pathway, and gene length ratio
GA [18]	ID, pathway, COG, and microarray
FGA [8]	ID, pathway, homologous genes, and protein functions
SVM [23]	ID, pathway, homologous genes, and phylogenetic profile
Using both genome-specific and general genomic information [4]	ID, homologous genes, phylogenetic distance, motif, GO, and gene length ratio
DVDA [5]	Homologous genes
FGENESB (http://www.softberry.com)	ID, GOC, promoter, and terminator
ODB [12]	ID, pathway, microarray, and GOC
OFS [20]	ID, common gene annotation, and GOC
OPERON [6]	Gene cluster conservation
JPOP [3]	ID, COG, and phylogenetic profile
VIMSS [14]	ID, comparative features, COG and CAI
UNIPOP [11]	Homologous genes

註: 使用特性欄位中的 ID, GOC, GO 及 CAI 依序表示 intergenic distance, gene order conservation, gene ontology, codon adaptation index;

表 5. 研究方法之正確率比較表

基因體	研究方法	正確率(%)	敏感度(%)	特異度(%)
枯草桿菌	BPSO (**初始門檻** = 600_{bps})	**92.1**	**93.0**	89.9
	BPSO (**初始門檻** = 300_{bps})	90.5	88.7	94.5
	GA [18]	88.3	87.3	89.7
	FGA [8]	88.2	N/A	N/A
	SVM [23]	88.9	90.0	86.0
	Using both genome-specific and general genomic information [4]	90.2	N/A	N/A
	DVDA [5]	48.5	31.9	93.2
	FGENESB (http://www.softberry.com)	77.1	72.1	90.4
	ODB [12]	63.2	49.9	**99.2**
	OFS [20]	68.3	76.5	43.9
	OPERON [6]	62.9	53.1	89.2
	JPOP [3]	74.6	72.0	90.0
	VIMSS [14]	78.0	76.4	87.1
	UNIPOP [11]	79.2	78.2	82.1
綠膿桿菌	BPSO (**初始門檻** = 600_{bps})	**93.3**	**93.0**	93.9
	BPSO (**初始門檻** = 300_{bps})	91.0	88.5	**95.1**
	GA [18]	81.3	87.0	76.3
葡萄球菌	BPSO (**初始門檻** = 600_{bps})	**95.9**	**95.9**	**95.8**
	BPSO (**初始門檻** = 300_{bps})	93.6	92.4	95.8
	Genome-wide operon prediction in *Staphylococcus aureus* [19]	92.0	N/A	N/A

註: 數據中的 N/A 表示文獻中無該資料。

初始門檻值為 600_{bps} 情況下，枯草桿菌、綠膿桿菌及葡萄球菌等基因體之操作組預測正確率分別為 92.1%、93.3%及 95.9%，其結果不僅優於初始門檻值為 600_{bps} 實驗結果，亦優於所有比較之文獻。於敏感度及特異度方面，本研究除枯草桿菌之特異度較 ODB 低，其餘皆優於比較文獻。因此，本研究方法對於基因體操作組之預測具有相當程度之貢獻。

(三) 討論

在所有比較的文獻中，GA 及 BPSO 皆為最佳化演算法，但文獻 GA 卻有特性重疊之缺點，因為該研究方法同時使用代謝路徑及 COG 兩種"功能相關"類別之特性，導致特異度提高且敏感度下降，使得操作組的預測個數減少[8]。不僅如此，由表 5 亦可以發現初始門檻值設定為 300_{bps} 時，會使敏感度及特異度無法達到良好平衡，而無法有效提升預測正確率。因此我們將初始門檻值調整為 600_{bps}，使敏感度及特異度之間差距縮小，相對提升操作組之預測正確率。然而，本研究優於文獻的原因，不僅為上述所提及之優點，因此本文將預測正確率提升之因素，分為下列幾項說明：(i)BPSO 的優勢；(ii)初始化的改善；(iii)以統計基礎所設計之適應函數；(iv)生物特性的選擇。

i. BPSO 的優勢

操作組預測問題中，有許多研究方法僅根據相鄰基因之特性辨別 WO pair 及 TUB pair，卻忽略了附近基因的生物特性，因而降低操作組的預測正確率。為克服此缺點，本研究使用 BPSO 評估鄰近基因的生物特性，另外，為使 BPSO 能有效搜尋最佳解，本方法設定慣性權重為 1，並限制粒子更新速度，以提升 BPSO 的預測效能。若粒子速度接近於 0，狀態轉變的機率增加，同時表示 BPSO 正在進行全域搜尋，若速度越靠近 6 或-6，狀態轉變的機率則會降低，則表示 BPSO 正在進行區域搜尋。總之，BPSO 不僅能在解空間中進行全域探索，亦可於區域空間中進行區域探勘，因而提升搜尋到最佳解的機率。

ii. 初始化的改善

初始化步驟對於操作組預測問題相當重要，若能於初始化即獲得較好的粒子族群，再經由迭代過程的更新，即可有效提升操作組之預測

正確率。本研究根據基因方向及相鄰基因之距離進行初始化，由表 5 得知，當距離門檻值為 300_{bps} 時，可獲得較低之敏感度及較高特異度，若將距離門檻值設定為 600_{bps}，則會提高敏感度且降低特異度，同時也提升預測正確率。換句話說，越低的距離門檻值會提高預測結果之特異度，故必須尋找一個較佳的門檻值，使得敏感度及特異度達到良好平衡，以獲得較佳之預測正確率。

iii. 以統計基礎所設計之適應函數

目前研究方法中，仍無法設計出一個與預測正確率成比例的適應函數，因為相鄰基因即使擁有相同生物特性，也不一定位於相同的操作組中，因此如何有效改善適應函數，為操作組預測之重要課題。本研究以對數似然法計算各粒子的適應函數值，藉由該方法以統計基礎建構的特性，以提升適應函數值與預測正確率之間的比例正確性。實驗結果證明，本研究所使用的適應函數確實能幫助我們搜尋到較佳的粒子，並有效提升預測正確率。

iv. 生物特性的選擇

大量的大腸桿菌實驗驗證資料可經由 RegulonDB 資料庫下載，但其他基因體之相關資料卻不如大腸桿菌豐富，為可廣泛應用本研究方法於操作組預測，本研究以基因體普遍擁有的特性進行預測。理論上，使用越多的特性進行預測，可獲得較佳的預測正確率，由表 4 可知，三個使用頻率較高的特性分別為相鄰基因間之距離、代謝路徑及同源基因(homologous gene)特性，然而文獻 DVDA 僅使用同源基因進行預測[5]，其實驗結果之預測正確率及敏感度僅分別獲得 48.5%及 31.9%[11]。另外，經由文獻[2]評估，以 DVDA 研究方法預測大腸桿菌及枯草桿菌之實驗結果，其敏感度及特異度皆小於 50%，且預測正確率皆低於 20%，經觀察上述結果後，本研究決定不以該特性作為預測依據。基因距離特性於最近幾年提出，且經文獻[4]證明該特性之操作組預測能力。由於相鄰基因的距離已由許多文獻證明其預測能力，因此本研究以相鄰基因之距離為根基，並加入上述兩個特性進行操作組之預測。

由於整體方法的改善，使本研究之預測正確率皆優於相關文獻，其中只有枯草桿菌的特異度低於 ODB，主要原因在於 ODB 研究方法僅著重於 NOP 預測，而忽略預測 OP 的重要性，因此使敏感度及特異度無法達到平衡，預測正確率也只獲得了 63.2%。相較於該方法，本研究的敏感度及特異度皆相差於 40%以下，所以能夠得到較佳的預測正確率。

五、結論

本研究提出以二進制粒子族群最佳化應用於基因體操作組之預測，在初始化的部份，除利用距離的特性，亦增加方向限制，使得演算法於迭代步驟前即可獲得較佳的粒子族群。適應函數則是以大腸桿菌作為訓練資料，並根據其基因間之距離、代謝路徑及基因長度特性作為評分依據，接著再藉由位置及速度之更新公式，於每次迭代對粒子族群進行更新，以獲得較佳的預測正確率。實驗結果顯示，本研究方法不僅提升枯草桿菌、綠膿桿菌及葡萄球菌操作組的預測正確率，也因使用較少特性作為評分準則的關係，大幅降低預測所需時間，並達到最佳的成本效益。未來研究方向將會增加不同生物特性進行判斷，或利用其他演算法進行預測，以達到更佳預測結果。

六、參考文獻

[1] J. Bockhorst, M. Craven, D. Page, J. Shavlik and J. Glasner, “A Bayesian network approach to operon prediction”, Bioinformatics, Vol. 19, pp.1227-1235, 2003.

[2] R.W.W. Brouwer, O.P. kuipers and S.A.F.T. van Hijum, “The relative value of operon predictions”, Bioinformatics, pp.1-9, 2008.

[3] X. Chen, Z. Su, P. Dam, B. Palenik, Y. Xu and T. Jiang, “Operon prediction by comparative genomics : an application to the *Synechococcus* sp. WH8102 genome”, Nucleic Acids Research, Vol. 32, no. 7, pp.2147-2157, 2004.

[4] P. Dam, V. Olman, K. Harris, Z. Su and Y. Xu, “Operon prediction using both genome-specific and general genomic information”, Nucleic Acids Research, Vol. 35, no. 1, pp.288-298, 2007.

[5] M.T. Edwards, S.C.G. Rison, N.G. Stoker and L. Wernisch, “A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context”, Nucleic Acids Research, Vol. 33, no. 10, pp.3253-3262, 2005.

[6] M.D. Ermolaeva, O. White and S.L. Salzberg, “Prediction of operons in microbial genomes”, Nucleic Acids Research, Vol. 29, no. 5, pp.1216-1221, 2001.

[7] M.H. Gabriel and C.V. Julio, “A powerful non-homology method for the prediction of operons in prokaryotes”, Bioinformatics, Vol. 18, pp.S329-S336, 2002.

[8] E. Jacob, R. Sasikumar and K.N.R. Nair, “A fuzzy guided genetic algorithm for operon prediction”, Bioinformatics, Vol. 21, no. 8, pp.1403-1407, 2004.

[9] J. Kennedy, R.C. Eberhart, “Particle Swarm Optimization”, Proc. of IEEE international Conference on Neural Networks (ICCN), Vol. 4, pp.1942-1948, Perth, Australia, 1995.

[10] J. Kennedy and R.C. Eberhart, “A discrete binary version of the particle swarm algorithm”, System, Man, and Cybernetics, Computational Cybernetics and Simulation, IEEE International Conference, Vol. 5, pp.4104-4108, 1997.

[11] G. Li, D. Che and Y. Xu, “A universal operon prediction for prokaryotic genomes”, Journal of Bioinformatics and Computational Biology, Vol. 7, no. 1, pp.19-38, 2009.

[12] S. Okuda, T. Katayama, S. Kawashima, S. Goto and M. Kanehisa, “ODB: a database of operons accumulating known operons across multiple genomes”, Nucleic Acids Research, Vol. 34, pp.D358-D362, 2006.

[13] M. Pertea, K. Ayanbule, M. Smedinghoff and S. L. Salzberg, “OperonDB: a comprehensive database of predicted operons in microbial genomes”, Nucleic Acids Research, pp.1-4, 2008.

[14] M.N. Price, K.H. Huang, E.J. Alm and A.P. Arkin, “A novel method for accurate operon predictions in all sequenced prokaryotes”, Nucleic Acids Research, Vol. 33, no. 3, pp.880-892, 2005.

[15] C. Sabatti, L. Rohlin, M.K. Oh and J.C. Liao, “Co-expression pattern from DNA microarray experiments as a tool for operon prediction”, Nucleic Acids Research, Vol. 30, no. 13, pp.2886-2893, 2002.

[16] H. Salgado, M.H. Gabriel, T.F. Smith and C.V. Julio, “Operons in Escherichia coli: Genomic analyses and predictions”, Proceedings of the National Academy of Sciences of the United States of America, Vol. 97, no. 12, pp.6652-6657, 2000.

[17] N. Sierro, Y. Makita, M. de Hoon and K. Nakai, “DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information”, Nucleic Acids Research, Vol. 36, pp.D93-D96, 2008.

[18] S. Wang, Y. Wang, W. Du, F. Sun, X. Wang, C. Zhou and Y. Liang, “A multi-approaches-guided genetic algorithm with application to operon prediction”, Artificial Intelligence in Medicine, Vol. 41, pp.151-159, 2007.

[19] L. Wang, J.D. Trawick, R. Yamamoto and C. Zamudio, “Genome-wide operon prediction in *Staphylococcus aureus*”, Nucleic Acids Research, Vol. 32, no. 12, pp.3689-3702, 2004.

[20] B.P. Westover, J.D. Buhler, J.L. Sonnenburg and J. I. Gordon, “Operon prediction without a training set”, Bioinformatics, Vol. 21, no. 7, pp.880-888, 2004.

[21] T. Yada, M. Nakao, Y. Totoki and K. Nakai, “Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models”, Bioinformatics, Vol. 15, no. 12, pp.987-993, 1999.

[22] Y. Yan and J. Moult, “Decection of Operons”, Wiley InterScience, Vol. 64, pp.615-628, 2006.

[23] G.Q. Zhang, Z.W. Cao, Q.M. Luo, Y.D. Cai and Y.X. Li, “Operon prediction based on SVM”, Comput Biol Chem, Vol. 30, pp.233-240, 2006.