演化式計算輔助整合特徵模型預測蛋白質瓜胺酸化位置

陳啟瑋　　洪慧芝　　朱彥煒
國立中興大學基因體暨生物資訊學研究所
ywchu@nchu.edu.tw

摘要

蛋白質瓜胺酸化(Citrullination) 藉由胜肽精胺酸脫亞胺酶(Peptidylarginine Deiminase，PAD) 將受質蛋白上的精胺酸轉變成不帶電的瓜胺酸，瓜胺酸化與類風濕性關節炎、多發性硬化症和阿茲海默症等疾病有關。由於受限於目前的資料量，現今尚未有針對瓜胺酸化的相關工具提供使用。因此，本研究發展一套針對資料量不足情況下，使用機器學習建構預測模型之實驗方法，並用於發展蛋白質瓜胺酸化位置預測工具。從 8 種特徵；文獻收錄之催化規則、序列相似度、演化保留訊息、胺基酸理化和生化特性、二級結構、蛋白質不穩定結構及結構表面可接觸性，透過資料與編碼後產生解決出不同解決問題能力的預測模型，進而透過整合方式獲得進一步提升的預測能力。而在機器學習的過程中，要如何調整適當的學習權重以獲得較佳的預測模型，本研究以基因演算法並支持向量機作為適性函數，挑選出適合之特徵權重最為訓練。其過程中並與前人文獻中之胜肽精胺酸脫亞胺酶催化規則有所呼應，最終生物意義並打破以往挑選最佳預測模型方式，以模型學習過程及特徵重要性做為評估方式。最終模型準確可達 Sn：0.79、Sp：0.98、Acc：0.89 及 MCC：0.79。

關鍵詞：瓜胺酸化、胜肽精胺酸去亞胺酶、支持向量機、基因演算法、特徵選擇。

1. 前言

胜肽精胺酸脫亞胺酶 (Peptidylarginine Deiminase，PAD) 屬於蛋白質轉譯後修飾酵素之一，當 PADs 與鈣離結合使催化中心結構改變，進而得以和受質結合，將蛋白質上帶正電的精胺酸 (Arginine) 轉變成不帶電的瓜胺酸(Citrulline)，其過程稱為瓜胺酸化(Citrullination 或 Deimination)[1-4]。

PADs 有五種異構型(Isoforms)， PAD1-4 及 PAD6，並具有組織分布特異性，各自分布在特定的組織細胞中[5-8]。PAD1 位於皮膚表皮；PAD2 可見於各類器官，包括腦、骨骼肌、子宮、胰腺、涎腺、垂體腺、汗腺、脾、巨噬細胞和骨髓等。PAD3 表現於毛髮之毛囊細胞；PAD4 主要表現在淋巴細胞、內皮細胞、單核細胞和巨噬細胞，而 PAD6 主要表現在胚胎幹細胞與卵母細胞。其瓜胺酸化反應將受質蛋白質上，帶正電荷的精胺酸轉變成中性不帶電的瓜胺酸，可能造成蛋白質的構型改變，進而失去蛋白質原有之功能。蛋白質瓜胺酸化所牽涉的生理過程包括：上皮終端分化(Epithelial Terminal Differentiation)、神經生長、胚胎發育、基因表現之調節 (Gene Expression Regulation) 及細胞凋亡 (Apoptosis)。而病理表現則目前已知與類風濕性關節炎、多發性硬化症和阿茲海默症 (Alzheimer' s Disease) 等疾病有關[3, 5, 9-11]，其中 PAD4 是唯一具有細胞核定位訊號(Nuclear Localization Signal)，可使蛋白經由膜孔到細胞核中，主要表現於淋巴細胞、內皮細胞、單核細胞和巨噬細胞，此外，近幾年的文獻指出，PAD4 調節真核細胞中的基因的轉錄，並與 p53 在 DNA 修復都有密切的關係[12]。

然而，對於胜肽精胺酸脫亞胺酶，其受質之序列特定性以及蛋白質專一性，目前並不清楚，且其受質參與細胞內的調控路徑目前尚未全然釐清。若要進一步探究蛋白質瓜胺酸化所造成的結果，在生物上之意義或與疾病之相關性，尚須鑑定蛋白質瓜胺酸化，進而了解其生物作用與功能途徑[13]。因此有研究利用蛋白質陣列試圖尋找胜肽精胺酸脫亞胺酶之受質[14]。但人類蛋白質數量龐大且生物實驗費時費力，若能藉由資訊的計算提供最有可能之瓜胺酸化受質蛋白，可能將幫助實驗減少所需花費的金錢與時間。

但目前也無相關工具輔助實驗設計，為此，本研究主要在於建構瓜胺酸化預測工具，可望幫助生物學家尋找可能做為 PAD 受質之蛋白質，提供輔助實驗設計方向的建議資料，進一步做更深入的分析與討論。瓜胺酸化屬於蛋白質轉譯後修飾中的一種，而有收錄蛋白質轉譯後修飾相關資料的 Swiss-Prot 或 HPRD (Human Protein Reference Database) 等資料庫中，較多資料的為磷酸化、乙烯化、甲基化等，也因為這些數量較為豐富的後修飾非常適合利用數學、統計等領域並配合電腦的計算建構預測模型，許多這類後修飾的預測工具如雨後春筍般的被開發出來，對於資料量甚少的後修飾較為無人問津。而瓜胺酸化也同那些資料過於少量的後修飾一樣，有其生物上的重要性卻礙於目前現有資料而遲遲無法以生物資訊發展輔助工具，但這些資料量較不豐富的後修飾中存在著共同的因子，乃推測的資料多過實驗所證實之數量。然而，推測資料雖少部分可能是藉由實驗資料所做出的推測，其中包含著錯誤待實驗驗證之資料，但對於預測模型的好壞其必須包含容錯性，才具有將目前

資料以外的未知資料做出正確的預測。

基於此論點，本研究進而嘗試克服以資料量少無法建構預測模型之問題，並用以解決瓜胺酸化預測模型之建構，設計一套對於資料量不足的情況下所用之發展預測模型之方法，將所有的推測資料及實驗資料做為訓練集測試資料集，以 8 種特徵編碼產生不同性質之預測模型，進而將不同特徵編碼之模型所預測之結果利用整合方式建構第二層預測模型。並以基因演算法尋找其 8 種特徵其各自的重要性權重，進一步提升預測準確度其準確度達 0.80。

2. 材料與方法

2.1 蛋白質瓜胺酸化資料集

首先，必須先建立一個資料正確的瓜胺酸化資料集。處理的資料來源便是 UniProt 資料庫，UniProt 是現今最廣泛被利用的線上蛋白質資料庫之一，其 Protein knowledgebase (UniProtKB) 中包含兩個部分，分別為人工註解的 Swiss-Prot 與自動註解的 TrEMBL 資料集，而 Swiss-Prot 收錄的註解與文獻等資訊由專業的生物學家所提供，其準確性無庸置疑，因此本研究從中擷取與瓜胺酸化有關的蛋白質資料，共 84 個蛋白質 141 個瓜胺酸化位置，而 84 個蛋白中包含了 64 個組織蛋白(Histone)、10 個 MBP 蛋白(Myelin Basic Protein)、8 個 CXL(C-X-C Motif Chemokine) 和 2 個其他種類蛋白質。除了 UniProt 資料庫之外，並從 HIstome 人類組織蛋白轉譯後修飾資料庫中，獲得額外的組織蛋白瓜胺酸化位置共 6 筆。

使用一個完整、正確且資料量大的資料集，有利於使用機器學習方法做預測，然而，從 UniProt 中獲得了 141 個瓜胺酸位置，其中推測的位置有 109 筆，實際實驗之位置則有 32 筆，面臨到資料量過少，可能不足以反映瓜胺酸化問題之學習樣本。而推測之樣本數為實驗資料的 3 倍多，因此，本研究嘗試以推測之資料做為學習樣本，希望以此建構的預測模型，能預測出實驗之瓜胺酸位置。剔除推測與實驗衝突之瓜胺酸位置後，93 個推測之瓜胺酸位置資料集命為 S93 做為訓練預測模行之用，而從 UniProtKB 及 HIstome 收集實驗之 38 個瓜胺酸位置命為 S38 做測試之用。

在資料的處理上，將蛋白質上有註解精胺酸 (Arginine) 發生瓜胺酸化之蛋白質挑出做為 Positive data 瓜胺酸化位置，而蛋白質序列上其餘沒註解的精胺酸視為 Negative data 非瓜胺酸化位置。S38 的 Negative data 為 376 筆，S93 的 Negative data 為 768 筆。然而，以機器學習建構預測模型的過程中， Negative data 的使用將會影響模型的好壞，Positive 與 Negative 若數量差距過大往往造成學習產生偏頗的現象，因此本研究以 Positive 與 Negative 數量 1 比 1 的比例隨機挑選 Negative。

2.2 胜肽精胺酸去亞胺酶催化規則

使用機器學習方法建構預測模型需要一個完整、正確且資料量大的資料集，才可能做出較為準確的預測，而瓜胺酸化目前在蛋白質後修飾資料庫中，收錄的筆數甚少。因此，本研究從前人研究所發表的文獻中，收集可能之瓜胺酸化催化特性規則，希望以此規則輔助可能尚未足以代表瓜胺酸化問題空間的資料集，進行預測模型的建構。

由 Tarcsa *et al.*[15]與 György *et al.*[5]獲得 PADs 可能對於蛋白質一級結構和二級結構上的催化規則。 從 Kyouhei *et al.*[4]獲得 PAD4 偏好受質組織蛋白(Histone) 上精胺酸位於 N 端或 C 端。此外，於 Stensland *et al.*[16]獲得組織蛋白與聚角蛋白微絲蛋白(Filaggrin) 之瓜氨酸化位置周圍胺基酸組成可能偏好特性。

為了測試收集之規則，是否與目前已知的瓜胺酸化資料有所衝突，進一步以規則比對 S93 及 S38 資料集內的瓜胺酸與非瓜胺酸化位置，其中二級結構以 PSIPRED 預測工具取得結構資訊。

2.3 特徵模型

本研究主要使用機器學習做為預測方法，首先必須將資料抽象化，從中提取出做為學習之特徵，若所選擇之特徵較符合問題空間，可能獲得較佳之預測結果。因此，特徵的好壞影響預測其準確度。若要使電腦能夠自動對資料提取出特徵，作為分類之用所牽涉之技術較為複雜。為此，本研究以 PADs 受質喜好特性規則(Rule)、胺基酸向量化(Binary)、演化保留訊息(PSSM)、胺基酸理化和生化特性(AAIndex)、二級結構(Secondary Structure)、蛋白質結構穩定性(Disorder) 及蛋白質結構可接觸性與二級結構(Surface Accessibility and Secondary Structure) 做為特徵，分別依照不同特徵以 LIBSVM 實現支持向量機建構預測模型，進而評估與探討其資料在不同編碼下之特性。

2.3.1 規則編碼

藉由文獻收集之 PADs 受質可能喜好規則，包含一級結構 5 條規則、二級結構 4 條規則及精胺酸周圍四個位置喜好與非喜好之胺基酸 8 條規則，共 17 條規則，依照其規則與序列及結構匹配度編碼為：匹配 1、不匹配 0。其中精胺酸接近 N 端或 C 端評估方式為：該精胺酸位置除以蛋白質序列總長，若該值小於等於 0.3 或大於等於 0.7 將被視為接近 N 或 C 端，其編碼為 1，反之則為 0。一條序列片段將測試 17 個規則，並編碼為長度 17bit 的特徵值。但其規則參考之精胺酸周圍資訊甚少，一級結構僅一個胺基酸，或上下游 2 個胺基酸，二級結構僅考慮精胺酸結構，可能使機器學習過程中遺漏掉重要的訊息，因此，也進一步測試其他使用精胺酸周圍訊息之編碼。

2.3.2 胺基酸向量編碼

機器學習常將一筆資料以向量的方式表示，故將 20 種胺基酸包含 Gap，以 20 個維度之向量做編碼，該胺基酸所占之維度設為 1，例如：A 為[1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0] 、 Y 為[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1]，所有編碼為表 3.5 所示。因此，一個胺基酸以 20 個維度之向量所表示，共有序列片段胺的基酸數量乘上 20 個維度之向量，以胺基酸序列相似性做為機器學習之特徵。

2.3.3 演化保留訊息編碼

此方法利用蛋白質演化訊息作為特徵，蛋白質具有功能的區域其演化過程中較不易產生突變，得以將功能性保留下來，在特定的胺基酸片段下，瓜胺酸化之精胺酸周圍之評分矩陣，通常具有胺基酸保留性的特性，因此可能適合用來做為分類之特徵。本研究利用 PSI-BLAST 將蛋白序列對 NCBI nr (Nonredundant) protein dataset v2010_07_23 做比對，以指令 “blastpgp -d nr -i Preotein.txt -j 3 -h 0.001 -o Preotein.out -Q Preotein.pss”獲得每個胺基酸在該位置之演化保留分數，其輸出之結果如圖 2.7，因此，一個胺基酸有 20 個突變機率分數，編碼方式共有序列片段的胺基酸數量乘上 20 個維度之向量。

2.3.4 胺基酸理化與生化特性編碼

AAindex (Amino Acid Index Database) 收錄五百種以上藉由實驗所定義的胺基酸理化和生化特性，而要在眾多的特性中選擇適合用以解決問題的胺基酸特性較為不易。而有研究 William *et al.*將其胺基酸性質依照相似性簡化其繁多的特性，而 Mathura *et al.*從文獻中整理並歸納 5 種特性。William 簡化成 Polarity、Secondary structure、Molecular size or volume、Codon diversity 及 Electrostatic charge 5 種，Mathura *et al.* 歸納為 5 種特性；Hydrophobicity、Side chain length、α-helix propensity 、 Number of codons 及 β-strand propensity，每個胺基酸擁有一個數值來代表其理化及生化特性，例如 Side chain length 較短的 Glycine 值越大，Side chain length 較長的胺基酸其值越小。進一步將兩篇研究將眾多的胺基酸特性化中簡為 5 種特性的結果，藉由皮爾森相關係數計算其兩組簡化結果之間的相似性，如表 1。僅只有 Hydrophobicity 與 Polarity 有 0.7 以上較高的相似度，因此將兩篇研究所歸納之胺基酸特性結果視為兩種特徵做編碼之用，一個胺基酸將有 5 個值所表示，共有序列片段的胺基酸數量乘上 5 個維度之向量。

2.3.5 二級結構編碼

從文獻中發現，PADs 的受質喜好與結構有相關，因此若以蛋白質結構做學習之特徵可能有助於預測，但並非所有蛋白質被解結構，而較大的蛋白質其結晶結構也並非完全，導致以資訊方法解決問題將有所限制，為此，本研究以二級結構預測工具 PSIPRED 獲得蛋白質序列其可能的結構資訊。PSIPRED 提供三種二級結構的預測分別為 Helix、Strand 及 Coil，以及對該位置所做的預測信心分數 (Confidence Score)，。進而將 Helix 編碼為 100、Strand 編碼為 010、Coil 編碼為 001 及 Gap 編碼為 000，包含預測信心分數一個胺基酸以 4 個值表示，共有序列片段的胺基酸數量乘上 4 個維度之向量。

表 1 Mathura 和 William 兩編碼相關係數

	hydrophobicity	side chain length	α-helix propensity	number of codons	β-strand propensity
Polarity	0.80	-0.55	0.13	-0.13	0.03
Secondary structure	0.39	0.60	0.66	-0.08	-0.13
Molecular size or volume	-0.34	-0.15	0.13	-0.11	-0.03
Codon diversity	0.17	0.49	-0.56	-0.57	-0.09
Electrostatic charge	-0.11	-0.13	0.14	-0.19	-0.24

2.3.6 蛋白質結構穩定性編碼

根據 Kyouhei *et al.*及 György *et al.*均認為不穩定結構 (Disorder) 對於 PADs 存在重要性，因此也加入該結構做為學習之特徵，可能將有助於預測。為此，本研究以不穩定結構預測工具 DISOPRED 獲得蛋白質序列其可能的結構資訊。DISOPRED 提供兩種結構的預測，分別為穩定 (Ordered) 與不穩定 (Disorder)，以及該位置不穩定可能性分數 (Disorder Probability)，若該值大於門檻值將預測為不穩定。進而將一個胺基酸以 2 個值表示，穩定編碼為 0、不穩定編碼為 1 及可能性分數，共有序列片段的胺基酸數量乘上 2 個維度之向量。

2.3.7 蛋白質結構可接觸性與二級結構編碼

胺基酸胜肽透過折疊使蛋白質產生有功能的構型，其中胺基酸位於蛋白質內或蛋白質外對於研究蛋白質之間的交互作用以及三級結構有非常大的重要性。而酵素與受質結合進行催化與結構表面可接觸性可能也具有影響，因此，從 NetSurfP ver. 1.1 獲得蛋白質結構表面可接觸性預測資訊，其提供資訊。其預測結果提供二級結構可能性分數、Relative Surface Accessibility、Absolute Surface Accessibility 以及胺基酸可能為 Buried 或 Exposed 共 7 個資訊，進而將一個胺基酸編碼為 Buried；10、Exposed；01、Gap；00 及 5~10 欄位之值，共有序列片段的胺基酸數量乘上 8 個維度之向量表示。

2.4 第二層預測模型

本研究以 PADs 受質喜好特性規則(Rule)、胺基酸向量化(Binary)、演化保留訊息(PSSM)、胺基

酸理化和生化特性(AAIndex)、二級結構(Secondary Structure)、蛋白質結構穩定性(Disorder) 及蛋白質結構可接觸性與二級結構(Surface Accessibility and Secondary Structure) 做為特徵，分別依照不同特徵建構預測模型。然而使用不同特徵所建構之預測模型，具有其不同的預測之特性及效能，為整合不同特性之預測模型，進而將 8 個特徵之預測結果透過支持向量機建構第二層預測模型，以進一步提升預測準確度。

2.5 演化式計算搜尋特徵模型權重

使用不同特徵所建構出的預測模型，其具有解決不同問題之能力，再加上使用之訓練資料與其中負向資料的差異，使得不同編碼與訓練資料產生更為複雜且較難選擇適合解決問題之模型。且並非所有特徵都適合用來解決瓜胺酸化預測的問題，因此勢必將調整 8 種特徵模型其權重，以改善第二層預測模型之學習效果。

為解決該問題，本研究利用基因演算法(Genetic Algorithm) 尋找每個特徵之重要性權重分數。其基因演算法參數設置為：Crossover rate 0.7、Mutationrate rate 0.05、Population 200、Chromosome length 40、Generations 20。

自然界中的染色體隱含著遺傳等訊息，而在人工的基因演算法染色體隱藏著解決問題的答案，因此以 40 個 0 與 1 組成 40Bit 2 進制長度的染色體，8 個 Bit 表示 1 個權重，再透過轉換成 10 進制除以 8Bit 最大值 31 做為權重值。該權重乘上特徵模型 l 預測之信心分數，做為特徵模型選擇所需建構模型之輸入，其模型以測試資料之準確度 (MCC) 做為該染色體之適應分數 (Fitness Value)，準確度越高表示適應分數越大，使得在演化過程中得以被保留並可能成為最終獲得之解答。

2.6 預測模型效能評估

當預測結果為瓜胺酸化位置且實際亦為瓜胺酸化則為 True Positive (TP)，若預測結果為瓜胺酸化位置但實際為非瓜胺酸化為 False Positive (FP)，而預測結果為非瓜胺酸化位置但實際卻為瓜胺酸化則為 False Negative (FN)，預測結果為非瓜胺酸化位置且實際亦非瓜安酸化為 True Negative (TN)。將利用下列公式評估預測系統的準確性。Matthews Correlation Coefficient (MCC) 顯示 Positive 和 Negative 的相關度，其值最大為 1 表預測完全正確，最小為-1 表示預測完全與正確結果相反 ，其公式如下：

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}} \quad \text{(Eq. 4)}$$

Accuracy (Acc)，顯示整體預測結果的預測正確率，其公式如下：

$$Acc = \frac{TP+TN}{TP+FP+TN+FN} \quad \text{(Eq. 5)}$$

Specificity (Sp)，顯示預測結果對於非瓜胺酸化位置的辨識能力，其公式如下：

$$Sp = \frac{TN}{TN+FP} \quad \text{(Eq. 6)}$$

Sensitivity(Sn)，顯示預測結果對於瓜胺酸化位置的辨識能力，其公式如下：

$$Sn = \frac{TP}{TP+FN} \quad \text{(Eq. 7)}$$

3. 結果與討論

3.1 特徵模型預測準確度

將 8 種特徵編碼分別建構 8 個預測模型，以及將 8 種特徵混合建構一個預測模型，共 9 種方式建構預測模型，以三組交叉驗證方式評估準確度，其結果如表 2。並以 S38 測試其預測準確度，結果如表 3。由表 2 顯示，透過交叉驗證評估發現所有特徵模型，其預測準確度略顯較高。其原因在於透過推測資料整理之訓練集資料間較為相似，而本研究嘗試以推測之資料做為學習樣本，希望以此建構的預測模型，能預測出實驗之瓜胺酸位置。因此由實驗資料整理之測試集，其準確度優劣才是本研究著重的焦點。

然而，不同特徵編碼其可解決之問題也不盡相同，從表 3 測試結果可發現蛋白質不穩定結構(Disorder) 與蛋白質結構可接觸性與二級結構(NetSurfP) 對於大部分瓜胺酸化位置有較高的預測準確度，這呼應了胜肽精胺酸脫亞胺酶在選擇受質催化規則中的特性。而其餘特徵編碼可能善於預測非瓜胺酸化位置。在使用所有 8 種豐富且能解決不同問題空間的特徵編碼模型中(All)，未能獲得較好的預測準確度，正是因為其中存在著干擾學習及不必要之特徵，在以往的研究中，將會以此預測模型進行特徵選擇，使模型預測準確度提高，但在此瓜胺酸化的預測中，資料數量尚未達使用特徵選擇的條件，斷然以特徵選擇提升準確度可能將造成過度擬合(Overfitting)，使得模型的容錯性降低進而無法預測目前尚未確定的瓜胺酸化位置。因此，進而以整合 8 個特徵模型之預測結果，建構第二層預測模型(Meta)。然而，其準確度不及使用所有編碼(All)，其原因可能在於整合的過程中，機器學習對於 8 個特徵模型其重要性尚未掌握，因此勢必將 8 種特徵模型其學習權重再做修正。

表 2 訓練集三組交叉驗證之預測準確度

Training S93	Sn	Sp	Acc	MCC
AAIndex 2005	0.92	0.98	0.95	0.91
AAIndex 2001	0.98	0.98	0.98	0.96
Binary	0.98	1.00	0.99	0.98
DISOPRED	0.92	0.86	0.89	0.79
NetSurfP	0.98	0.95	0.96	0.93
PSIPRED	0.91	0.94	0.93	0.85

PSSM	0.97	0.93	0.95	0.90
Rule	0.87	0.86	0.87	0.74
All	0.99	0.97	0.98	0.96

3.2 演化式搜尋特徵模型權重之預測準確度

不同特徵模型其擅長預測資料不盡相同，反映出訓練資料與特徵編碼產生之差異性。使用 8 種不同特徵建構 8 個預測模型，以其預測結果之信心分數做為建構第二層模型之輸入，準確度 Sn 達 0.71、Sp 0.97、Acc 0.85 及 MCC 0.72。透過基因演算法調整其特徵模型適合之學習權重後，準確度 Sn 為 0.79，相較於未調整前權重提升 0.08。而 MCC 也從 0.72 提升至 0.79，上升 0.07。且使用基因演算法調整權重所建構之第二層預測模型，準確度均高於單一特徵模型準確度(MCC)。

表 3 測試集預測準確度

Test S38	Sn	Sp	Acc	MCC
AAIndex 2005	0.68	0.94	0.82	0.65
AAIndex 2001	0.63	0.95	0.80	0.62
Binary	0.71	0.98	0.85	0.72
DISOPRED	0.82	0.71	0.76	0.53
NetSurfP	0.92	0.86	0.89	0.78
PSIPRED	0.74	0.88	0.81	0.63
PSSM	0.66	0.90	0.79	0.59
Rule	0.68	0.94	0.82	0.66
All	0.79	0.95	0.87	0.76
Meta	0.71	0.97	0.85	0.72
GA	0.79	0.98	0.89	0.79

為探討基因演算法找出之權重結果並分析，圖 1 為基因演算法尋找之準確度可達 0.79 的 78 條最佳權重組，及尚未調整前之權重。在未調整前(Meta)使用之權重為 AAIndex 2005：9.71、AAIndex 2001：12.3、Binary：17.32、DISOPRED：1.49、NetSurfP：9.09、PSIPRED：4.03、PSSM：5.24、Rule：1.65，以 Binary 序列相似性視為較重要之特徵，其次為胺基酸理化及生化特性。而藉由基因演算法尋找學習權重，平均 78 條後的權重為 AAIndex 2005：0、AAIndex 2001：0、Binary：0.39、DISOPRED：0.35、NetSurfP：0.69、PSIPRED：0.56、PSSM：0.50、Rule：0.49，以蛋白質結構可接觸性與二級結構較為重要，其次為演化保留性及規則。而在未調整前支持向量機認為有較高重要性之 AAIndex 在基因演算法之下被認為是完全無重要之特徵，這可能是由於 AAIndex 包含許多不同的胺基酸特性，在尚未從其中特性做挑選的前提下，混和多種特性可能因此導致對於解決瓜胺酸化問題便的較無幫助。而基因演算法將其權重調整為零可提升準確度，這也驗證移除會干擾學習之特徵可獲得較好的準確度。

由圖 1 分析 78 條最佳權重組合，有 0.92 預測瓜胺酸化準確度最高的 NetSurfP (藍色)，以及 0.98 最佳非瓜胺酸化預測準確度的 Binary (綠色)，以尋找權重的過程中可以發現，當 Binary (綠色)有較高權重值時 NetSurfP (藍色)權重較低，而 NetSurfP (藍色)權重較高時，則 Binary (綠色)權重較低，還無法使兩者做緊密的結合，這可能是由於這兩種編碼受限於目前訓練資料貧乏之缺陷所導致，或僅演化 20 世代尚未找尋到最佳學習權重。而在第 28 染色體之權重組以 Rule 及 DISOPRED 最高，這也說明文獻中之規則也可扮演機器學習上重要的學習項目。

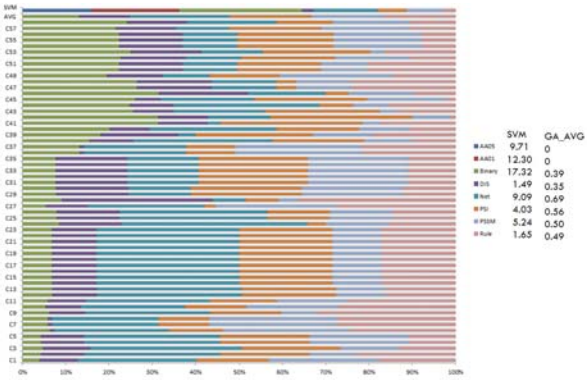

圖 1 最佳權重組

4. 結論

PADs 與許多疾病有關，如類風濕性關節炎，多發性硬化症，牛皮癬，阿茲罕默症，也關係到細胞抗癌機制作用的過程，但其詳細的調控機制仍未非常明朗。而現今尚未有瓜胺酸化相關的生物資訊工具可供使用，因此，本研究主要在於建構瓜胺酸化預測工具，進而有辦法對未知是否會受該酵素作用的蛋白質序列作預測，幫助生物化學與分子生物研究者在設計蛋白質與研究。

本研究以機器學習方法透過不同特徵編碼與訓練資料產生出不同的解決問題空間之能力，透過整合其 8 種結果建構第二層預測模型，並利用基因演算法調整適當的學習權重，其過程中也驗證學習特徵與文獻規則具有相呼應。最終提升預測準確度並使準確度優於單一特徵模型，模型準確達 Sn：0.79、Sp：0.98、Acc：0.89 及 MCC：0.79。

誌謝

本論文承蒙行政院國家科學委員會之計畫經費贊助，計畫編號為 NSC 101-2221-E-005-085 與 NSC 102-2221-E-005-081，僅此誌謝。

5. 參考文獻

[1] J. E. Jones, C. P. Causey, B. Knuckley, J. L. Slack-Noyes, and P. R. Thompson, “Protein arginine deiminase 4 (PAD4): Current understanding and future therapeutic potential,” Current opinion in drug discovery & development, vol. 12, p. 616, 2009.

[2] Z. Ke, Y. Zhou, P. Hu, S. Wang, D. Xie, and Y. Zhang, “Active Site Cysteine Is Protonated in the PAD4 Michaelis Complex: Evidence from Born− Oppenheimer Ab Initio QM/MM Molecular Dynamics Simulations,” The Journal of Physical Chemistry B, vol. 113, pp. 12750-12758, 2009.

[3] T. A. Cafaro, S. Santo, L. A. Robles, N. Crim, J. A. Urrets-Zavalia, and H. M. Serra, “Peptidylarginine deiminase type 2 is over expressed in the glaucomatous

optic nerve," Molecular vision, vol. 16, p. 1654, 2010.

[4] K. Arita, T. Shimizu, H. Hashimoto, Y. Hidaka, M. Yamada, and M. Sato, "Structural basis for histone N-terminal recognition by human peptidylarginine deiminase 4," Proceedings of the National Academy of Sciences, vol. 103, pp. 5291-5296, 2006.

[5] B. György, E. Tóth, E. Tarcsa, A. Falus, and E. I. Buzás, "Citrullination: a posttranslational modification in health and disease," The international journal of biochemistry & cell biology, vol. 38, pp. 1662-1677, 2006.

[6] C. Foulquier, M. Sebbag, C. Clavel, S. Chapuy- Regaud, R. Al Badine, M. C. Méchin, et al., "Peptidyl arginine deiminase type 2 (PAD- 2) and PAD- 4 but not PAD-1, PAD- 3, and PAD- 6 are expressed in rheumatoid arthritis synovium in close association with tissue inflammation," Arthritis & Rheumatism, vol. 56, pp. 3541-3553, 2007.

[7] E. R. Vossenaar, A. J. Zendman, W. J. van Venrooij, and G. J. Pruijn, "PAD, a growing family of citrullinating enzymes: genes, features and involvement in disease," Bioessays, vol. 25, pp. 1106-1118, 2003.

[8] C. Anzilotti, F. Pratesi, C. Tommasi, and P. Migliorini, "Peptidylarginine deiminase 4 and citrullination in health and disease," Autoimmunity reviews, vol. 9, pp. 158-160, 2010.

[9] M. Moscarello, L. Pritzker, F. Mastronardi, and D. Wood, "Peptidylarginine deiminase: a candidate factor in demyelinating disease," Journal of neurochemistry, vol. 81, pp. 335-343, 2002.

[10] Z. Baka, P. Barta, G. Losonczy, T. Krenács, J. Pápay, E. Szarka, et al., "Specific expression of PAD4 and citrullinated proteins in lung cancer is not associated with anti-CCP antibody production," International immunology, vol. 23, pp. 405-414, 2011.

[11] Z. Baka, B. György, P. Géher, E. I. Buzás, A. Falus, and G. Nagy, "Citrullination under physiological and pathological conditions," Joint Bone Spine, vol. 79, pp. 431-436, 2012.

[12] C. Tanikawa, K. Ueda, H. Nakagawa, N. Yoshida, Y. Nakamura, and K. Matsuda, "Regulation of protein Citrullination through p53/PADI4 network in DNA damage response," Cancer research, vol. 69, pp. 8761-8769, 2009.

[13] S. K. Bhattacharya, "Retinal deimination in aging and disease," IUBMB life, vol. 61, pp. 504-509, 2009.

[14] Q. Guo, M. T. Bedford, and W. Fast, "Discovery of peptidylarginine deiminase-4 substrates by protein array: antagonistic citrullination and methylation of human ribosomal protein S2," Molecular bioSystems, vol. 7, pp. 2286-2295, 2011.

[15] E. Tarcsa, L. N. Marekov, G. Mei, G. Melino, S.-C. Lee, and P. M. Steinert, "Protein unfolding by peptidylarginine deiminase Substrate specificity and structural relationships of the natural substrates trichohyalin and filaggrin," Journal of Biological Chemistry, vol. 271, pp. 30709-30716, 1996.

[16] M. E. Stensland, S. Pollmann, Ø. Molberg, L. M. Sollid, and B. Fleckenstein, "Primary sequence, together with other factors, influence peptide deimination by peptidylarginine deiminase-4," Biological chemistry, vol. 390, pp. 99-107, 2009.