使用機器學習與特徵選擇預測明膠酶的受質切位

張浩禎　　朱彥煒
基因體暨生物資訊學研究所
ywchu@nchu.edu.tw

摘要

明膠酶(gelatinase)有 gelatinase A (MMP-2) 和 gelatinase B (MMP-9)兩型，具有裂解細胞外基質的活性。近來研究指出基質金屬蛋白酶家族在生理和病理機制上，有多種調控，如免疫反應、腫瘤發展和幹細胞分化等。MMP-2、MMP-9 會引發腫瘤轉移，其抑制藥物也進入臨床試驗，成功抑制腫瘤轉移但病人的存活率卻沒有提升，前人探究其原因為基質金屬蛋白酶家族同源性高，結構相似，使得抑制藥物專一性不高；以及對 MMP-2、MMP-9 受質的調控路徑尚未全面了解，以致產生副作用。因此，若能準確預測 gelatinase 之作用受質與位置，有益於其生理和病理作用機制之探討。本預測架構為整合二進制、物理化學屬性、蛋白質不穩定區段和溶劑可接觸性與二級結構等不同類別的資訊、加上我們首先採用的群間差異特徵，配合支持向量機的使用，建立第一層預測模型；接著將其輸出的預測機率做為第二層系統之特徵，同時比較多種機器學習方法來建構第二層的預測系統，MMP-2 及 MMP-9 受質的預測效能分別之馬修斯相關係數為 89.4%和 64.4%。並進一步利用物理化學屬性之特徵選擇，對 MMP-2 與 MMP-9 在活性中心的屬性進行分析，以提供藥物設計時參考。此預測系統建立完成後，有助於發現 MMP-2 與 MMP-9 新的可能受質，以推估其未發現的調控路徑。

關鍵詞：Gelatinase、MMP-2、MMP-9、支持向量機、最近鄰居法、群間差異。

1. 前言

蛋白質轉譯後修飾調控各種細胞代謝過程。常見的轉譯後修飾為可逆性。而蛋白水解為蛋白酶水解胺基酸與胺基酸之間連結的胜肽鍵；為不可逆性，因此在各種蛋白質體中，都有廣泛且重要的影響[1]。在人類基因體中，蛋白酶之基因約佔 2%，且其中約有 5-10% 已作為藥物標靶[2，3]。要了解特定蛋白酶在生物體內的功能與訊息調控路徑，需要辨別其受質，也就是裂解體研究；並利用受質切位的特異性作為藥物設計之參考。但蛋白質體的複雜動態性高，不同的組織、細胞、以及所處的發育或疾病時期，都有獨特的蛋白質組成和不同的轉譯後修飾情形，使得同一蛋白酶在每種蛋白質體中有不同的裂解體[1，4]。

基質金屬蛋白酶(matrix metalloproteinase)可降解細胞外基質與腫瘤發生有高相關性[5]。在二十年前就開始研究抑制藥物，並進入人類的臨床試驗。但試驗結果卻是失敗的，雖然可抑制腫瘤轉移但病人的生存率卻沒有上升，且引發肌肉疼痛或關節疾病等副作用。而後科學家探討其原因大約有兩點：一、MMPs 為 23 個結構相似且含鋅原子的胜肽內切酶，藥物無法專一性抑制特定 MMP，導致非目標 MMP 也被抑制，影響其他正常生理功能。二、近來研究指出除了細胞外基質，MMPs 更作用在細胞激素、細胞膜上受器和生長因子等，影響細胞生長、分化、移動等多樣的細胞機制。且由於任何一種 MMP 都未全面了解其受質所參與的調控路徑，無法全盤考量對其抑制後的影響範圍[6-8]。在 MMPs 之中，MMP-2 與 MM-9 受到較高的矚目。因為他們在多種的惡性腫瘤裡有高度表現，為癌症生物標記。MMP-2 和 MMP-9 因其結構和受質又稱明膠酶(gelatinase)，與其他 MMPs 結構不同在於催化區上有三個纖維鏈結蛋白域。MMP-2 與 MM-9 結構極為相似，差別在於催化區與血红素结合蛋白樣结構域之連結區段長短，而他們的受質有部分重疊，也各自催化不同受質，影響不同訊息路徑[9]。另外，相較於其他有特定胺基酸切位的蛋白酶，如 caspase 都切除在 aspartate 後的肽鍵[10]；MMP-9 與 MMP-2 無固定的切位胺基酸，且切位周邊的氨基酸特異性也不盡相同[11]，因此較困難預測其切位與受質。要瞭解 MMP-2 與 MM-9 的受質裂解體，可透過質譜分析，但不同時期且各種細胞之蛋白質體不盡相同，如利用實驗進行全面的受質辨認需花費長時間和高額金錢。因此，in silico 進行 MMP-2 和 MMP-9 受質切點預測可以進行大規模的受質裂解體的註解。在生物實驗設計上，可以減少且提供準確性高的候選者。

目前有多種蛋白酶受質切位的預測系統，其所使用的方法可分兩種[12]， 一、分數計算：GPS-CCD 利用 BLOSUM 62 進行胺基酸轉換，透過自行設計之演算法以預測 Calpain 受質切位。CaSPredictor 利用 BLOSUM 62 進行轉換並加入 PEST-like sequence 計算分數[13]。PoPS 讓使用者定義受質特異性的物理化學分數與權重來計算分數[14]。SitePrediction 利用每個位子的每種胺基酸發生的頻率和氨基酸替換矩陣計算分數以預測受質切位[15]。PoPS 和 SitePrediction 均可進行目前已發現的所有蛋白酶的受質切點預測，且提供二級結構

預測等額外資訊。二、機器學習： Pripper 使用 binary 編碼，分別利用支持向量機、隨機森林、J48 預測 caspase 受質切位[16]。另外一樣預測 caspase 受質切位的系統 Cascleave 除了 binary 編碼，更加入結構資訊和 Bi-profile Bayesian signature 編碼，以支持向量回歸來建立預測系統[17]。PCSS 讓使用者自行輸入訓練數據集並利用序列與結構等資訊編碼，利以支持向量機建立預測系統。也因為是使用者自行輸入訓練數據集，故可預測各種蛋白酶之受質[18]。

為了更精確地預測其受質切點以作為癌症標靶藥物設計之影響範圍參考，本研究建立 MMP-2 與 MMP-9 受質預測系統。此預測系統為兩層式，第一層建立利用二進制(binary)、物理化學屬性(physical-chemical property)、結構資訊、以及本實驗最新使用可以值展現數據庫中切位點(positive site)和非切位點(negative site)之胺基酸數量差異倍數(fold change)等特徵，並配合支持向量機建構四種特徵之模組。在第二層時，彙整每種特徵模組的預測信心指數並比較多種機器學習方法以建立模組。且第一層的每種特徵模組都測試七組不同的非切位點集，以得到適合且最能展現特徵編碼的特性之非切位點集。因此，整個預測系統可以學習更多的非切位點資訊。並針對其作為癌症標靶藥物失敗的原因，進一步利用物理化學屬性之特徵選擇，利用 MMP-2 與 MMP-9 在活性中心的屬性作為特徵，加上呈現資料庫中切點與非切點胺基酸組成差異倍數的資訊，提升預測系統準確度。

2. 材料與方法

2.1 資料收集

下載 MEROPS 9。5 版(西元 2011 年 7 月 1 日)[19]，篩選出人類的 MMP-2 與 MMP-9 受質，去除重複，並經過 Cd-hit 去除 70%相似的蛋白質，以整理出不冗餘的資料[20]。MMP-2 進行實驗的資料有 1269 筆受質切位在 630 個蛋白質中。MMP-9 的資料相對於 MMP-2 數量少，其受質切位數量為 269 筆，共在 42 個蛋白質中。沒有被註解為 MMP-2 與 MMP-9 作用的其他位點，均作為 negative site。因為 MMP-2 與 MMP-9 無特定胺基酸切位，使得 negative site 非常大量，分別為 330457 和 30558 筆。Training set 的好壞決定預測結果的準確度高地。因此本實驗隨機 7 次選出與 positive set 數量 1：1 的 negative sets(N1~N7)，各別與 positive set 形成 training set，共有 7 個 training set 並做不同編碼方式訓練，期望找到每一種編碼最好預測效能的 positive set 與 negative set 組合。其後，建立第二層預測時，將隨機挑選除了 N1~N7 以外的 negative set(N8)做為第二層之 training set。

2.2 建構預測系統

本研究利用機器學習方法建立預測系統，首先必須將蛋白質序列的資料型態轉換成機器學習所能讀取的特徵。本研究選用 binary、physical-chemical property、結構資訊、fold change 等特徵進行編碼，並利用 SVM 建構第一層的四種特徵模組，再將各種特徵模組預測結果的機率信心值彙整，經過第二層機器學習，以提升其準確率。第二層的預測模組將測試 5 種機器學習方法分別有 LibSVM、Naïve Bayes、Random Forest、IBK[21]以測試出準確度最高的方法，以建立兩層的預測系統。每種機器學習都以 10 倍交叉驗證進行預測效能計算。

2.3 第一層的特徵模組

第一層的特徵模組分別有二進制(binary)、物理化學屬性(physical-chemical property)、結構資訊、差異倍數(fold change)等四個。序列片段長(window size)的表示，被切的肽鍵往 N 端的第一個氨基酸為 P1，往 C 端的第一個胺基酸為 P1’。最長取到第二十個胺基酸 P20，P19，P18…P1P1’…P18’P19’P20’。

1. Binary：將測試 P20 到 P20’各種長度組合的胺基酸片段以向量的方式表示，故將 20 種胺基酸包含 Gap，以 21 個維度之向量做編碼，該胺基酸所占之維度設為 1。例如：
   Alanine(A) 000000000000000000001、
   Cysteine(C) 000000000000000000010。

2. Fold change：將 MMP-2 與 MMP-9 的 training set 中 positive site 和 negative site 的 P20-P20’片段輸入 Icelogo 軟體中[22]，依據下列公式，計算出每種胺基酸在每個位子上所出現的頻率並相除得到 fold Change 值。以顯示 positive site 和 negative site 每個胺基酸出現頻率的倍數差，以顯示不同 dataset 中 positive 與 negative 資料的差異性質。如利用 P+N1 建立模組，在進行 10-fold cross-validation 時，positive set 將分為 10 份，分別與總和 N2~N7 的 negative set 計算出 Fold Change 替換表。

$$\text{Frequency}+ = \frac{AA1}{all\ AA}\ \text{in positive set} \quad \text{(Eq. 1)}$$

$$\text{Frequency}- = \frac{AA1}{all\ AA}\ \text{in negative set} \quad \text{(Eq. 2)}$$

$$\text{Fold chang} = \frac{Frequency+}{Frequency-} \quad \text{(Eq. 3)}$$

3. 結構資訊： 將 dataset 中的蛋白質分送到 DISOPRED[23]和 NetSurfP[24]輸出的預測結果，取 P20-P20’片段進行編碼。編碼的資訊含有 DISOPRED 預測每個位子為 disorder 的機率和 NetSurfP 預測每個位子的 Relative Surface Accessibility、Absolute Surface Accessibility、Z-fit score、Probability for Alpha-Helix、Probability for Beta-strand、Probability for Coil。

4. Physical-chemical property：從 AAindex[25]資料庫下載胺基酸的物理化學性質，最初下載有 544 筆資料，去除空值與去除資料之間皮爾森相關係數大於 0.8 後，剩下 371 筆資料。加上由 Venkatarajan, M.S. 等人在 2001 年發表之論文，其將 237 個胺基酸物理化學性質，整合成 5 個向量之特徵[26]。而 gap 的數值為 20 個胺基酸的屬性數值之平均。因此，使用的物理化學特徵共有 376 個。取 P20-P20' 之切點與非切點片段用 376 筆物理化學性質編碼後，每一個位子將與 positive (+1) 和 negative(-1)進行皮爾森相關係數檢定，當相關係數|R|值大於 0.05 到 0.3 以及 P 值小於 0.001 時，此位子之相對特徵將被選出，其餘剔除。因此，非每一個位子都會進行編碼，且有進行編碼的位子的特徵數也不相同。

2.4 第二層整合模組

每種編碼的最好預測效能之 training set 來建立預測模組。重新隨機選取不包含第一層之 7 組 negative set 且與 positive site 數量 1:1 的 negative site 形成 negative set 8 (N8)：送給由四種特徵用 SVM 建立的第一層特徵模組。SVM 其輸出的預測結果有 positive site 和 negative site 的機率，此機率作為第二層之特徵。第二層分別 B、PCP、DS 特徵模組預測結果以及彙整 B、PCP、DS、FG 特徵模組預測結果作為特徵輸入給機器學習建立預測系統。將測試 5 種機器學習方法： LibSVM、Multilayer Perceptron、Naïve Bayes、Random Forest、IBK，找出最好的第二層預測方法。

2.5 數據集相似度分析

Positive set 和 7 個 negative ste(N1~N7)將分別利用 BLOSUM62 分數矩陣比對兩個胺基酸的相似度。將每一個位子的每一個胺基酸與相同位子的其他胺基酸的比對分數平均。舉例：共有 100 筆資料，第一個位子的第一個胺基酸將與其他 99 個胺基酸利用 BLOSUM62 矩陣轉換分數並加總後平均，即為第一個位子的相似度分數。將有 P20~P20'的平均分數。

2.6 模組預測能力評估

當預測結果為切點且實際亦為切點則為 True Positive(TP)，若預測結果為切點但實際卻非切點為 False Positive(FP)，而預測結果為非切點但實際卻為切點則為 False Negative(FN)，預測結果為非切點且實際亦非切點為 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切點的辨識能力，其公式如下：

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

Sensitivity(Sn)，顯示預測結果對於切點的辨識能力，其公式如下：

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

3. 結果

3.1 第一層預測效能

Binary model (B)-Binary 編碼呈現蛋白質的一級序列訊息。由序列片段由切位往 N 端和 C 端各 20 個胺基酸進行準確度測試。MMP-2 與 MMP-9 當片段從 P3 開始時，MCC 都有明顯上升。而 MMP-2 的片段長度從 N 端 P3 到 P15 以其 C 端 P3'到 P20 都有 MCC 都有 0.5 以上，其 MCC 最好的收斂約在 P3 到 P13'之間。MMP-9 的片段長度從 N 端 P3 到 P15 以其 C 端 P2'到 P20'有 MCC 都有 0.4 以上，其 MCC 最好的收斂約在 P3 到 P9'之間 (圖 1)。收斂區域，呈現了 MMP-2 與 MMP-9 切位附近的氨基酸特異性之決定長度。MMP-2 每組 training set 的 Sn 都有 0.9 以上，Sp 都在 0.85 以上，ACC 接近 0.9。而預測效果最好的 training set 為第 7 組(P4-P5')，其 MCC 為 0.808。MMP-2 最好與最差的 MCC 相距 0.034(表 1)。而 MMP-9 每組的 Sn 都有 0.7 以上，Sp 都在 0.75 以上，ACC 約為 0.75。預測效果最好的 training set 為第 7 組(P3-P5')，其 MCC 為 0.570。MMP-9 最好與最差的 MCC 相距 0.108 (表 2)。

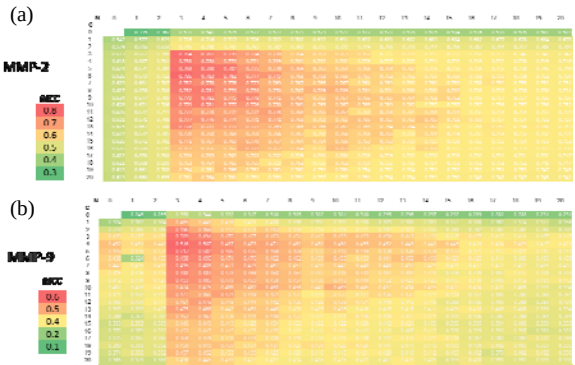


圖 1 MCC of vary window size by binary coding。 X,Y 軸分別表示往 N 端和 C 端延伸的氨基酸長度。(a) Heat map of MMP-2。(b) Heat map of MMP-9。

表 1
MMP-2 的第一層特徵模組內之 training sets 最高與最低 MCC 差異。

MMP-2 first layer models	Highest	Lowest	Difference
Binary	0.808 (7)	0.774 (2)	0.034
Fold change	0.796 (3)	0.687 (2)	0.109
Structure	0.649(7)	0.609 (2)	0.04
Physical-chemical property (|R| > 0.05)	0.814 (1)	0.778 (6)	0.036

表 2
MMP-9 的第一層特徵模組內之 training sets 最高與最低 MCC 差異。

MMP-9 first layer models	Highest	Lowest	Difference
Binary	0.570 (7)	0.462 (5)	0.108
Fold change	0.386 (2)	0.305 (4)	0.081
Structure	0.400 (2)	0.309 (1)	0.091
Physical-chemical property (|R| > 0.2)	0.613 (7)	0.446 (5)	0.167

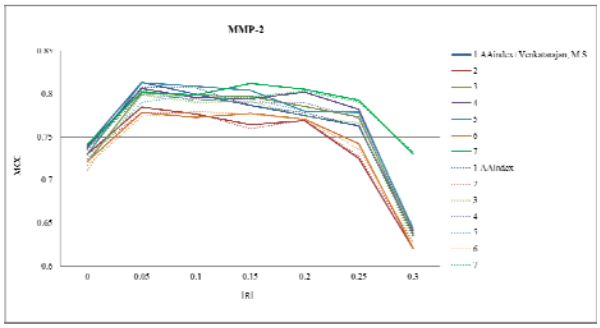
Fold change model (FG)- MMP-2 與 MMP-9 以 icelogo 輸出的 fold change 值進行 P20-P20'共 40 個胺基酸編碼，以展現每個位子的 positive 與 negative 的每一種胺基酸比例差異倍數值。MMP-2 每組的 Sn、Sp 都在 0.83 以上，ACC 在 0.85 左右，預測效果最好的 training set 為第 3 組，其 MCC 為 0.796。MMP-2 在 fold change 編碼上最好與最差的 MCC 相距 0.109(表 1)。MMP-9 每組的 Sn 約為 0.6，Sp 差距較大，ACC 約為 0.65。預測效果最好的 training set 為第 2 組，其 MCC 為 0.386。最好與最差的 MCC 相距 0.081(表 2)。

Structure model (S)-利用結構預測系統輸出的結果作為特徵，MMP-2 每組的 Sn、Sp 都接近 0.8 以上，ACC 約 0.8。預測效果最好的 training set 為第 7 組，其 MCC 為 0.649。MMP-2 最好與最差的 MCC 相距 0.04(表 1)。MMP-9 每組的 Sn、Sp 都在 0.65 以上，ACC 約在 0.65。預測效果最好的 training set 為第 2 組，其 MCC 為 0.4。MMP-9 最好與最差的 MCC 相距 0.091(表 2)。

Physical-chemical property model (PCP)- MMP-2 與 MMP-9 的物理化學屬性編碼，先以 P20-P20'片段以相關係數檢定進行特徵選擇，最後呈現不連續性片段的編碼，以凸顯特定位子的重要性。經過相關係數 0 到 0.3 的門檻選擇，不同門檻編碼的位子和使用的特徵不盡相同。MMP-2 的資料在|R|>0.05 時，所選擇出的特徵建構的模組之 MCC 約 0.8。隨著|R|上升，MCC 逐漸下降到約 0.6(圖 2(a))。MMP-2 的第 1 組且當|R|>0.05 時，有最高 MCC 為 0.814。在同樣的|R| > 0.05 下，最差的 MCC 為第 6 組的 0.778，最好與最差的 MCC 相距 0.036(表 1)。MMP-**9** 在 physical-chemical property 編碼，以不同相關係數篩選特徵，其隨著|R|上升，每組的 MCC 趨勢不盡相同(圖 2(b))。MMP-9 之第 7 組 training set 隨著|R|上升到 0.2 時，有最好的 MCC 為 0.613。在同樣的|R| > 0.2 下，最差的 MCC 為第 5 組的 0.446，最好與最差的 MCC 相距 0.167(表 2)。

加入 Venkatarajan, M.S. 等人在 2001 年發表之五維胺基酸特性，確實可以提升準確度。隨著相關係數上升，被選擇出來的特徵數下降，在 AAindex 佔大量特徵下，五維胺基酸特性約佔整體特徵的 1%。探討 physical-chemical property 編碼中最好的模組，MMP-2 在|R|>0.1 時，五維特徵可使 MCC 上升 0.01(圖 2(a))。MMP-9 在|R|>0.1 到 0.2，五維特徵都有使 MCC 上升。尤其在 R=0.2 時，上升了 0.04(圖 2(b))。

(a)



(b)

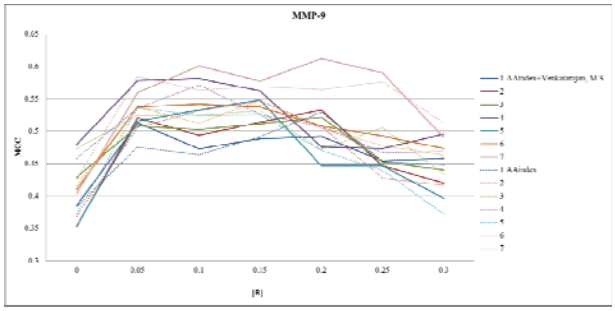


圖 2 Physical-chemical property models 相關係數之特徵選擇後 MCC 趨勢。實線為 AAindex 加上五維特徵。虛線為只有 AAindex 之特徵。(a)MMP-2 之 7 個模組，Y 軸從 MCC 為 0.6 開始。(b) MMP-9 之 7 個模組， Y 軸從 MCC 為 0.3 開始。

3.2 第二層預測效能

整合每種特徵最好的預測效果組別，MMP-2 之 binary 特徵預測效能最好是第 7 組，fold change 為第 3 組， structure 為第 7 組，physical-chemical property 為第 1 組(表 1)。MMP-9 之 binary 特徵預測效能最好是第 7 組，fold change 為第 2 組，structure 為第 2 組，physical-chemical property 為第 7 組(表 2)。第 2 層的建立首先比較彙整 3 個特徵模組(B、PCP、S)以及加入數據庫的差異倍數資訊的 4 個(B、

PCP、S、FG)特徵模組的預測效果並測試 4 種機器學習方法: LibSVM、Naïve Bayes、Random Forest、IBK。MMP-2 第二層的預測結果 3 或 4 個特徵模組，MCC 都在 0.8 以上，最好的為彙整 4 個特徵模組並利用 LibSVM 建構的預測系統，其 MCC 為 0.894。測試有無 FG 模組的輸出預測機率的特徵，可以發現加入 FG 模組的預測準確度 MCC 平均可以提升 5.52%(表 3)。MMP-9 第二層的預測效能結果，最好的 MCC 為彙整 4 個特徵模組並利用 IBK (K=23) 機器學習法建構的預測系統，其 MCC 為 0.644。加入 FG 模組的預測機率，預測準確度 MCC 平均可以提升 10.62%(表 4)。最後的完整預測系統架構，是由第一層每種不同特徵之最好 MCC 之模組，和彙整四個特徵模組的第二層最好 MCC 之機器學習方法所建成。

表 3 MMP-2 第二層利用不同機器學習比較有無 FG 模組的 MCC

MMP-2 second layer	Naïve Bayes	Random Forest	IBK	LibSVM
3 models	0.82	0.821	0.829	0.834
4 models	0.847	0.888	0.893	0.894
Increase	2.7%	6.7%	6.4%	6%

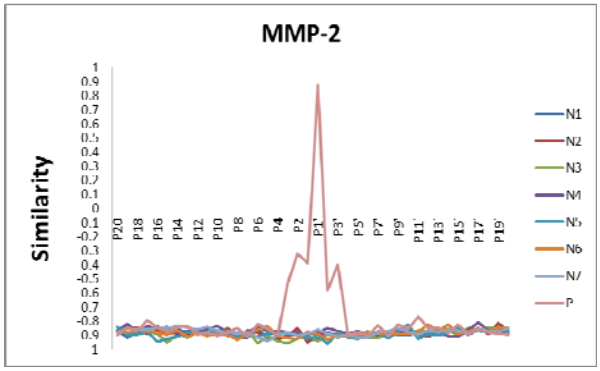
表 4 MMP-9 第二層利用不同機器學習比較有無 FG 模組的 MCC

MMP-9 second layer	Naïve Bayes	Random Forest	IBK	LibSVM
3 models	0.522	0.447	0.503	0.512
4 models	0.602	0.6	0.644	0.625
Increase	8%	15.3%	14.1%	11.3%

4. 討論

MMP-2 在四種編碼中，最高與最低的 MCC 相差約在 0.03 左右，與 MMP-9 的編碼相差有 0.1 之多。探究其原因為 MMP-2 與 MMP-9 資料相差五倍。有 1269 positive sites 的 MMP-2 在每種編碼上都有比 MMP-9 的 MCC 高出 0.2 以上，因為 MMP-2 資料量多較易得到好的 negative site 建構預測系統。MMP-9 資料量少，在隨機選取時相較於 MMP-2 較易選到集中性質的 negative site 導致預測效能偏頗，影響預測系統的正確度。至於，每種編碼的最高與最低 MCC 的組別都不盡相同，是因為隨機選取的 negative site 的組成不同，對於各編碼的屬性有不同的影響。以 BLOSUM62 計算 positive 與 negative sites 胺基酸相似率，可以看出每組 negative set 相對於 positive 的趨勢線都不盡相同(圖 3)。各個 negative set 與 positive set 相似度的差異，導致預測系統訓練時的不同，因此準確度會依 negative set 變化。有趣的是，MMP-2 在 P4 到 P4'片段(圖 6(a)) 以及 MMP-9 在 P13、P3、P1-P3'、P12'、P15'、P18' 等位置(圖 6(b))，positive 與 negative sites 胺基酸相似曲線明顯不同。這些位置剛好對應到 physical-chemical property 編碼經過相關係數特徵選擇後留下的有編碼位置。表示使用相關係數作為特徵選擇，不僅 MCC 會上升且也符合序列相似度的關係。

(a)



(b)

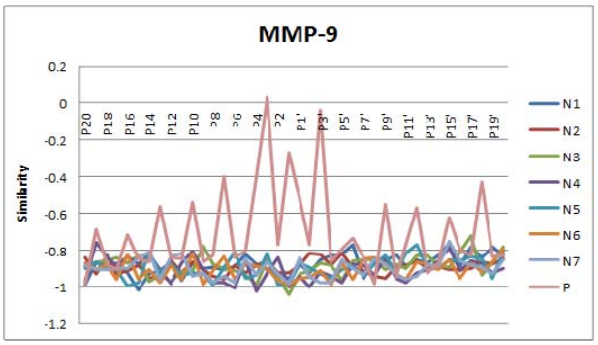


圖 3 切點與非切點之胺基酸序列的相似比率。 P 為切點資料庫。N1 為第一組之非切點資料庫，以此類推。(a)為 MMP-2 的資料庫相似度趨勢圖。(b)為 MMP-9 的資料庫相似度趨勢圖。

本實驗流程設計，讓機器學習更多樣的 negative set，讓每種編碼擁有最適合的 training set。並以第二層彙整不同意義之編碼使之帶有更多 negative 的訊息。 以解決 negative site 過於龐大之問題。可作為之後相關研究之預測系統架構之參考。本實驗最先使用 Icelogo 所提供的 fold change 數值，以顯示 training set 裡 positive 和 negative 的倍數差異。這個特徵在第二層彙整時，佔很大的重要性，他可以顯著地幫助預測較能的提升。此編碼亦可解決 negative data 過多的問題，如能選擇出適當的非切點群以代表全部的非切點資料，即可得到良好的 fold change 值，以展現全面性的差異。

相較於現有的預測系統，本系統更加友善且準確。本系統為現在唯一預測 MMP-2 與 MMP-9 受質切點的專職預測系統。系統的輸入，將設計可以一次輸入大量蛋白質序列，以實現 proteome 的預測功用。MMP-2 與 MMP-9 和腫瘤轉移有密切關係。為全面了解其受質所參與的調控路徑，本研究分別建立 MMP-2 與 MMP-9 受質切位的預測系統，以預測新的可能受質，之後將推估其他調控路徑，作為在製作抑制子時的全盤考慮與規劃，以避免副作用。

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。

參考文獻

[1] Doucet A, Butler GS, 1. Doucet A, Butler GS, Rodríguez D, Prudova A, Overall CM: Metadegradomics. Molecular & Cellular Proteomics 2008, 7(10):1925-1951.

[2] Puente XS, Sánchez LM, Overall CM, López-Otín C: Human and mouse proteases: a comparative genomic approach. Nature Reviews Genetics 2003, 4(7):544-558.

[3] Overall CM, Blobel CP: In search of partners: linking extracellular proteases to substrates. Nature Reviews Molecular Cell Biology 2007, 8(3):245-257.

[4] López-Otín C, Overall CM: Protease degradomics: a new challenge for proteomics. Nature Reviews Molecular Cell Biology 2002, 3(7):509-519.

[5] Kessenbrock K, Plaks V, Werb Z: Matrix metalloproteinases: regulators of the tumor microenvironment. Cell 2010, 141(1):52-67.

[6] Coussens LM, Fingleton B, Matrisian LM: Matrix metalloproteinase inhibitors and cancer—trials and tribulations. Science 2002, 295(5564):2387-2392.

[7] Turk B: Targeting proteases: successes, failures and future prospects. Nature reviews Drug discovery 2006, 5(9):785-799.

[8] Drag M, Salvesen GS: Emerging principles in protease-based drug discovery. Nature reviews Drug discovery 2010, 9(9):690-701.

[9] Bauvois B: New facets of matrix metalloproteinases MMP-2 and MMP-9 as cell surface transducers: Outside-in signaling and relationship to tumor progression. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer 2011.

[10] Timmer JC, Zhu W, Pop C, Regan T, Snipas SJ, Eroshkin AM, Riedl SJ, Salvesen GS: Structural and kinetic determinants of protease substrates. Nature structural & molecular biology 2009, 16(10):1101-1108.

[11] Prudova A, Auf Dem Keller U, Butler GS, Overall CM: Multiplex N-terminome analysis of MMP-2 and MMP-9 substrate degradomes by iTRAQ-TAILS quantitative proteomics. Molecular & Cellular Proteomics 2010, 9(5):894-911.

[12] SONG J, TAN H, BOYD SE, SHEN H, MAHMOOD K, WEBB GI, AKUTSU T, WHISSTOCK JC, PIKE RN: Bioinformatic approaches for predicting substrates of proteases. Journal of bioinformatics and computational biology 2011, 9(1):149.

[13] Liu Z, Cao J, Gao X, Ma Q, Ren J, Xue Y: GPS-CCD: a novel computational program for the prediction of calpain cleavage sites. PLoS One 2011, 6(4):e19001.

[14] Boyd SE, de la Banda MG, Pike RN, Whisstock JC, Rudy GB: PoPS: a computational tool for modeling and predicting protease specificity. In: 2004. IEEE: 372-381.

[15] Verspurten J, Gevaert K, Declercq W, Vandenabeele P: SitePredicting the cleavage of proteinase substrates. Trends in biochemical sciences 2009, 34(7):319-323.

[16] Piippo M, Lietzén N, Nevalainen OS, Salmi J, Nyman TA: Pripper: prediction of caspase cleavage sites from whole proteomes. BMC Bioinformatics 2010, 11(1):320.

[17] Song J, Tan H, Shen H, Mahmood K, Boyd SE, Webb GI, Akutsu T, Whisstock JC: Cascleave: towards more accurate prediction of caspase substrate cleavage sites. Bioinformatics 2010, 26(6):752-760.

[18] Barkan DT, Hostetter DR, Mahrus S, Pieper U, Wells JA, Craik CS, Sali A: Prediction of protease substrates using sequence and structure features. Bioinformatics 2010, 26(14):1714-1722.

[19] Rawlings ND, Barrett AJ, Bateman A: MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic acids research 2012, 40(D1):D343-D350.

[20] Li W, Godzik A: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006, 22(13):1658-1659.

[21] Witten IH, Frank E: Data Mining: Practical Machine Learing Tools and and Techniques. 2005.

[22] Colaert N, Helsens K, Martens L, Vandekerckhove J, Gevaert K: Improved visualization of protein consensus sequences by iceLogo. Nature methods 2009, 6(11):786-787.

[23] Ward J, Sodhi J, McGuffin L, Buxton B, Jones D: Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. Journal of molecular biology 2004, 337(3):635-645.

[24] Bent P, Thomas P, Pernille A, Morten N, Claus L: A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Structural Biology 2009, 9.

[25] Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M: AAindex: amino acid index database, progress report 2008. Nucleic acids research 2008, 36(suppl 1):D202-D205.

[26] Venkatarajan MS, Braun W: New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties. Journal of Molecular Modeling 2001, 7(12):445-453.