

基於胺基酸組成份特徵整合支持向量機及遺傳演算法進行微生物蛋白質之子細胞位置預測與分析

林采妤 廖俊欽 游景盛
逢甲大學資訊工程學系
yuucs@fcu.edu.tw

摘要

蛋白質存在於細胞當中，而在細胞內不同位置上的蛋白質分別扮演著不同的角色。過去倚賴實驗方式來確認蛋白質細胞內的位置，既耗時又花費許多金錢，而生命科學領域中仍存在大量的基因與蛋白質亟需透過實驗了解其研究標的，因此近年來發展了許多預測蛋白質的子細胞位置(subcellular localization)工具，可大量的處理蛋白質的序列資料，加速細胞系統的基礎研究和相關的醫學製藥應用。在本論文中，我們利用 *n*-peptide 胺基酸序列組成份不同的特性作為特徵並透過機器學習的整合方法，結合支持向量機(Support Vector Machine; SVM)與遺傳演算法(Genetic Algorithm; GA)，在本研究中以三階層式的處理 – 第一層以遺傳演算法分別進行各支持向量機分類器所使用的特徵值篩選；第二層將合併第一層的所有分類器的結果進行篩選；第三層則合併前兩層結果，將第二層所篩選的分類器，以及其分類器在第一層所萃取的特徵值進行合併，並再次篩選，皆以馬修相關係數(Matthew's correlation coefficient; MCC)經五倍交叉驗證評估預測結果並與前人的方法進行比較，結果顯示：我們的方法在格蘭氏陽性菌的整體靈敏度(sensitivity)及準確率(accuracy)可以達到 94.7%及 98.1%，而古生菌的整體靈敏度更提升到了 98.9%。

關鍵詞：支持向量機、遺傳演算法、子細胞位置、*n*-peptide 胺基酸組成份

1 緒論

● 研究動機：

由於蛋白質可能存在一個或一個以上的細胞內位置，而這些存在的事實意味著一個細胞系統中複雜的調控網路關係，因此某些細胞生物學家可透過蛋白質於子細胞位置註解其功能和重要性，因此發展相關的預測分析工具成為近年來重要的課題。

近二十年來，蛋白質資料成長的速度已經越來越快，如果要再去倚賴實驗去分析蛋白質，想必會是相當耗時且損耗相當多的金錢，因此現在大多利用機器學習的方式去對蛋白質進行分析預測，透過預測出來的結果，可以有效應用在一些醫學議題上[1]，諸如藥物治療、標靶治療等，因此了解細胞內的蛋白質位置是相當重要的，而現在機器學習的方法也不少，我們希望與前人的方法相比較能有效提升預測準確

率(accuracy)、精確率(precision)等評估指標。

● 研究目的：

以目前知名的 PSORTb3.0[2]的工具方法為例，我們希望利用 *n*-peptide 胺基酸組成份為特徵(feature)的方式改進預測準確率，故本篇論文以相同的資料集，使用支持向量機演算法對細胞內的蛋白質位置進行預測，並結合遺傳演算法尋找最佳參數，並透過整合不同分類器的方法提升預測準確率。

2 文獻研究與回顧

由於基因體與蛋白質體學迅速的發展，資料量日漸龐大，僅僅單靠生化實驗方法無法追及大量定序資料註解的需求，因此學者們逐漸發展出透過機器學習的方式自動化預測蛋白質的子細胞位置[3]，從早期 1999 年 Nakai[4]等學者的研究，至 2003 年 Gardy 等人所發表的 PSORTb[5]為當時預測細菌細胞內蛋白質位置最為精確的一項工具，所涵蓋的生物包括有格蘭氏陰性菌(Gram-negative bacteria)、格蘭氏陽性菌(Gram-positive bacteria)，改版後的 PSORTb 2.0[6]更大幅提升預測的物種範圍和準確率；而同時間 PA2.5、CELLO、PSORTb2.0 等工具陸續提出改進預測的方法，如 Proteome Analyst version 2.5 (PA 2.5) [7] 根據同源序列的註解字辭以機器學習的方法預測蛋白質子細胞位置；SLP-Local[8]則是基於胺基酸群的區段組成份及距離去預測細胞內位置；Gneg-PLoc[9]則是使用了 *k*-nearest neighbor (KNN)演算法結合 GO terms 去預測細胞內位置；由 Yu 等人所發表的 CELLO[10]利用多種胺基酸組成份進行序列的特徵擷取並且結合支持向量機的機器學習方式，對蛋白質位置進行預測，其後[11]並進一步發展多個支持向量機的分類器整合，而其精確率、靈敏度、準確率每項都有達 85%以上。

近年由 Huang[12]所發表的研究中也有提到，透過虛擬胺基酸組成份(pseudo amino acid composition)結合類神經網路，對蛋白質位置預測可以有效的提升準確率。

從上述可見，利用機器學習方法針對有效的特徵以預測細胞內蛋白質的位置是相當重要的，而透過不同的方法結合，能更提升預測效果，進而將其研究結果應用在生技醫療上面，希望未來能普及的應用在醫學上。

3 材料與方法

- 建構資料集(dataset)

我們選用了與 PSORT3.0[2]這篇論文相同的序列資料集，經由篩選 Rey[6]等人收集自 PSORTdb 的微生物資料庫，針對原核生物並以格蘭氏陰性菌、格蘭氏陽性菌、古生菌為例。其中格蘭氏陰性菌的位置分為五個：細胞質(Cytoplasmic; C)、胞質膜(Cytoplasmic Membrane; CM)、周質空間(Periplasmic; P)、外膜(Outer Membrane; OM)、細胞外(Extracellular; EC)；而格蘭氏陽性菌及古生菌的位置則皆為四個：細胞質(Cytoplasmic; C)、胞質膜(Cytoplasmic Membrane; CM)、細胞壁(Cell Wall; CW)、細胞外(Extracellular; EC)其各個位置與蛋白質數量如下表 3.1：

表 3.1 建構用資料集

位置	格蘭氏陰性菌	格蘭氏陽性菌	古生菌
C	5025	1822	675
CM	1607	360	85
P	437	-	-
OM	541	-	-
CW	-	82	18
EC	419	312	27
Total	8029	2576	805

†位置縮寫：C=Cytoplasmic, CM=Cytoplasmic Membrane, P=Periplasmic, OM=Outer

- 獨立資料集(independent dataset)

為了測試我們的方法在其他的資料集是否能達到一樣的效能，因此我們亦採用了屬格蘭氏陰性菌的綠膿桿菌新定序株 (*Pseudomonas aeruginosa* PAO1)基因體資料作為獨立資料集 [13]，此資料集內部的資料包括著假設、未定屬性和過去不知道位置的蛋白質，而 PSORTb3.0 再從 PAO1 株取部分已確認各存於 Cytoplasmic、Periplasmic、Extracellular 的 171 個 Cytoplasmic 位置的蛋白質，因此我們將它當做獨立資料集來測試與比較。

- 支持向量機:

支持向量機[14]是一個透過使用者給定的特徵條件來將物件做分類的一種分類演算法，而我們利用蛋白質序列中的組成份做為特徵向量，來區分這個蛋白質是否屬於該位置。

本篇研究我們使用 LIBSVM[15]進行蛋白質子細胞位置的預測研究，以五倍的交叉驗證法(5-fold cross validation)對資料集做訓練以及測試的過程找出最佳化的參數，避免資料比數不同而使得可信度降低，而支持向量機當中 cost 值、gamma 值、和特徵選取參數等設定均影響預測分類的結果，因此我們結合遺傳演算法反覆學習

訓練。

- 遺傳演算法:

遺傳演算法是一個搜尋技術，應用於計算並且找最正確解或近似解去解決問題的一種演算法，而其主要理論為：複製(reproduction)、交配(crossover)、突變(mutation)，從生物的角度來看，父代會將基因遺傳給子代，而子代如果遺傳好的基因則這個演算法會繼續複製下去，不好的基因就會被淘汰掉。本篇論文採用 Lu 等人[16]的策略，針對 SVM 的 cost 和 gamma 參數與特徵組的不同特徵值進行特徵選取，其適應函數(fitness function)是傾向優先選擇 MCC 值高，且支持向量機的能夠產生出最佳的分類方式，而透過遺傳演算法還能有效的幫助研究把較能區分位置的特徵保留下來。

- 序列特徵:

由於蛋白質序列是由 20 種不同胺基酸排列而成的，因此我們透過不同方式的胺基酸組成份截取方式對於每一條序列做特性上的分類，並將其分類特性做為支持向量機的輸入特徵值。我們總共分為五類：分別是胺基酸組成份(C_n)、間隔胺基酸組成份(D_g)、依照本身物理特性來做分類再去計算胺基酸的組成份(A_nX_k)、依照局部特性整合以組成份表示(N_iC)、將序列分段分別計算胺基酸的組成份(X_k)。[10-11]

- 方法架構流程圖:

我們的方法總共透過三層的遺傳演算法進行蛋白質位置的預測，在第一層的遺傳演算法中，我們先將 PSORT3.0 序列資料集進行 28 種的 n -peptide 組成份的特徵轉換，分別視為各自獨立的支持向量機分類器並各執行第一次遺傳演算法以篩選特徵值；在第二層的遺傳演算法中，我們將上述 28 種特徵分類器預測結果合併後進行第二次的遺傳演算法重新特徵選擇取得最佳的分類器組合；在第三層的遺傳演算法中，我們會將第二層所得到的最佳分類器組合的原始序列特徵進行合併，之後再次使用遺傳演算法對這個合併後的特徵進行特徵值的篩選，然後我們就能得到最終的分類器，方法架構流程圖如圖 3.3。

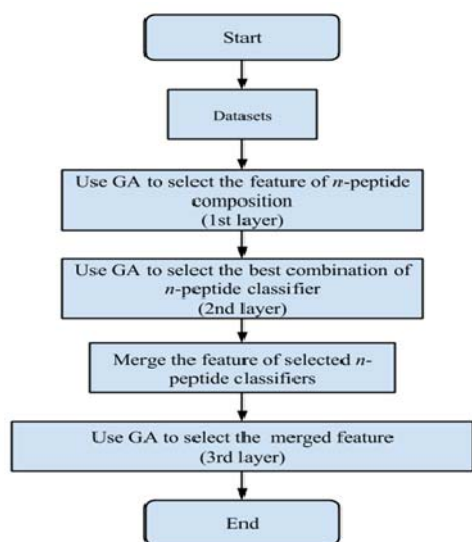


圖 3.3 方法架構流程圖

- 評估方法:
 1. Sensitivity(靈敏度): $TP/(TP+FN)$
 2. Specificity(專一性): $TN/(TN+FP)$
 3. Precision(精確率): $TP/(TP+FP)$
 4. Accuracy(準確率): $(TP+TN)/(TP+FP+TN+FN)$
 5. Matthew's correlation coefficient (MCC): $(TP*TN-FP*FN)/[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]^{1/2}$

我們採用 5 倍的交叉驗證(5-fold cross validation)，將資料集隨機分成五等份，將一份做為測試用的子資料集，剩餘四份合併做為訓練用的子資料集，共重複五次動作以完成五份子資料集各別的結果，確認每筆資料預測的真實性：True Positive(TP)、True Negative(TN)、False Positive(FP)、False Negative(FN)，並完成上述各項評估方法的計算。

4 研究結果與分析討論

- 使用單一特徵預測的結果:
 1. 格蘭氏陰性菌

其第一層遺傳演算法預測的最好結果如表 4.1，因此在這邊我們選用了四個分類器 $N_{13}C$ 、 X_3 、 X_6 、 X_7 ，再分別對四個位置進行預測。

由圖 4.1 當中我們發現到其實 X_7 的 MCC 值在這四類的分類器當中不是屬於最高的，不過他的精確率可以達到 97.4%，證明了他用少量的特徵可以達到跟其他特徵一樣的成功。

在圖 4.2 當中，我們可以看到特徵 $N_{13}C$ 在 Cytoplasmic Membrane 上的預測是有比較好的表現的，MCC 值達 0.91，準確率達到 97.3%。

表 4.1 格蘭氏陰性菌第一層遺傳演算法預測的結果

位置	分類器	MCC	選取特徵數
C	X_7	0.92	91
CM	$N_{13}C$	0.91	209
P	X_6	0.77	75
OM	X_3	0.89	41
EC	X_6	0.81	75

†位置縮寫：C=Cytoplasmic, CM=Cytoplasmic Membrane, P=Periplasmic, OM=Outer Membrane, EC=Extracellular.

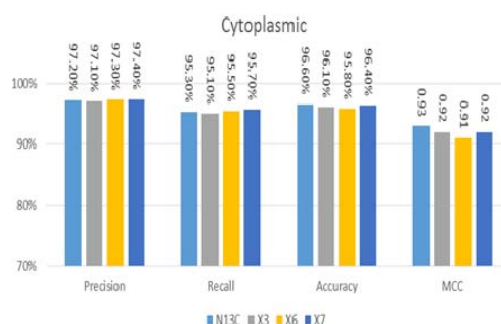


圖 4.1 以不同分類器預測格蘭氏陰性菌資料集 Cytoplasmic 位置的結果

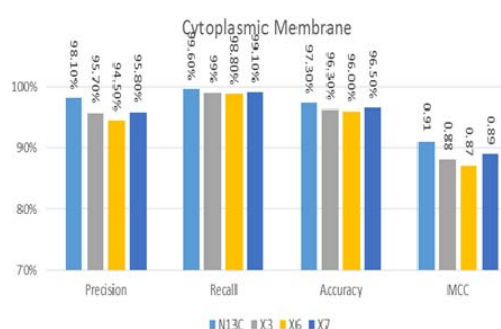


圖 4.2 以不同分類器預測格蘭氏陰性菌資料集 Cytoplasmic Membrane 位置的結果

在圖 4.3 當中，我們看到 X_6 在特徵當中的 MCC 值是最好的，有 0.77，而同時他在預測的準確率也達 97.80%，因此我們認為在 Periplasmic 這個位置來說透過 X_6 的特徵來預測是能有效提升準確率。

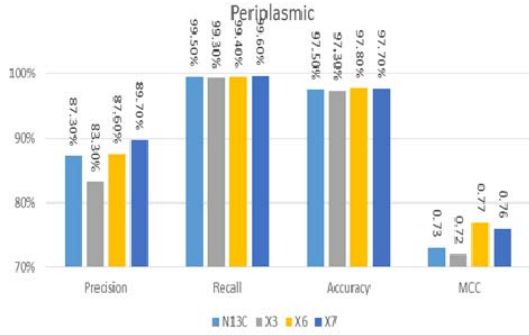


圖 4.3 以不同分類器預測格蘭氏陰性菌資料集 Periplasmic 位置的結果

在圖 4.4 當中，我們可以看到使用 X_3 及 X_6 來當做特徵時，他們 MCC 都是 0.89，且預測 accuracy 都達 98.7%，可見得使用 X_n 做為特徵在預測 Outer Membrane 是有一定的可信度。

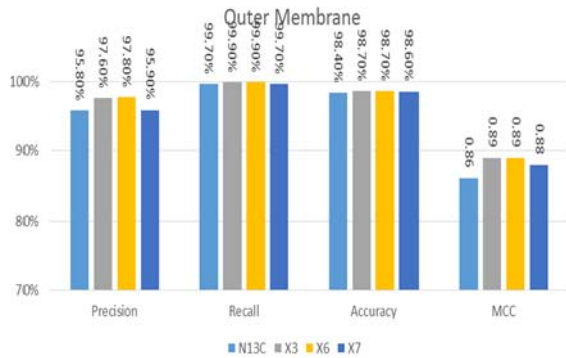


圖 4.4 以不同分類器預測格蘭氏陰性菌資料集 Outer Membrane 位置結果

在圖 4.5 當中，我們發現到 X_6 及 X_7 的 MCC 值都是高於 0.80 的，而他們的準確率也有達到 98%，因此我們認為在 Extracellular 位置上透過 X_n 特徵的分類是有效的。

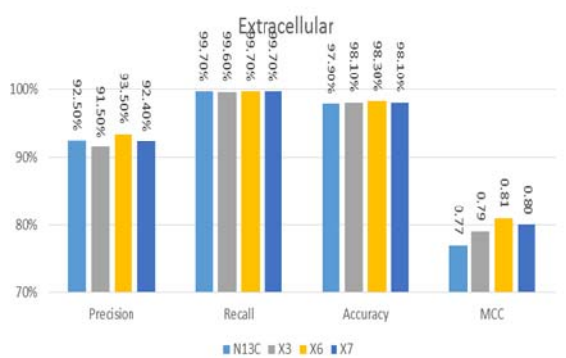


圖 4.5 以不同分類器預測格蘭氏陰性菌資料集 Extracellular 位置結果

2. 格蘭氏陽性菌

在格蘭氏陽性菌方面，其第一層遺傳演算法預測的結果如表 4.2，因為 $N_{15}C$ 在 Cytoplasmic 及 Cytoplasmic Membrane 的 MCC 值最高，因此在這邊我們選用了三個分類器 D 、 $N_{15}C$ 、 X_6 再分別對四個位置進行預測。

表 4.2 格蘭氏陽性菌資料集第一層遺傳演算法預測的結果

位置 [†]	分類器	MCC	選取特徵數
C	$N_{15}C$	0.93	186
CM	$N_{15}C$	0.91	151
CW	D	0.90	49
EC	X_6	0.85	54

[†]位置縮寫：C=Cytoplasmic, CM=Cytoplasmic Membrane, CW=Cell Wall, EC=Extracellular.

在圖 4.6 與圖 4.7 當中，我們可以看到 $N_{15}C$ 的 MCC 值各分別為 0.93 與 0.91，因此我們認為 $N_{15}C$ 在 Cytoplasmic 與 Cytoplasmic Membrane 位置上的預測是有一定的可信度的。

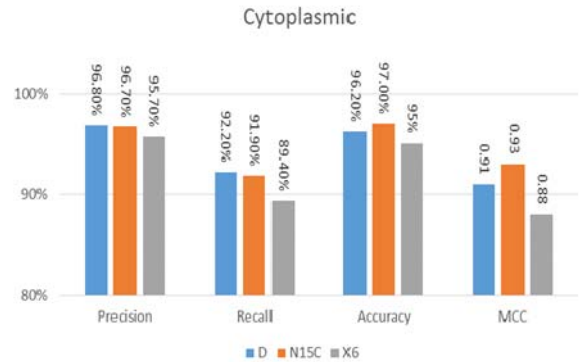


圖 4.6 以不同分類器預測格蘭氏陽性菌資料集 Cytoplasmic 位置結果

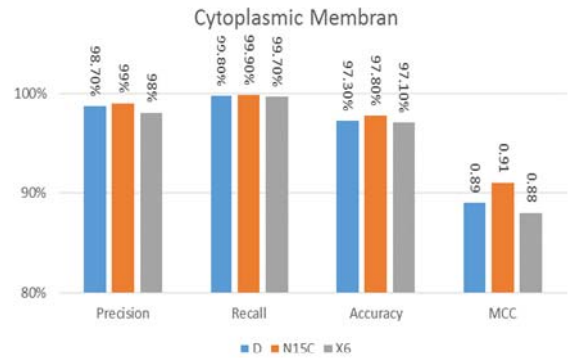


圖 4.7 以不同分類器預測格蘭氏陽性菌資料集 Cytoplasmic Membrane 位置的結果

在圖 4.8 當中，我們可以明顯的看到 D 特徵分類器的準確率達 99.4%與 $N_{15}C$ 不相上下，但是在比較 MCC 值就顯得有相當大的落差了， D 高於 $N_{15}C$ 有 0.06，因此我們認為 Cell Wall 透過 D 特徵來分類是有效的。

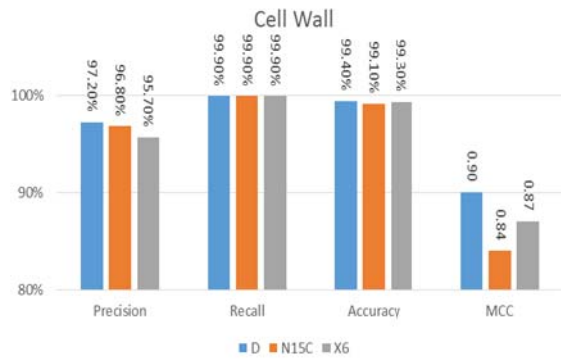


圖 4.8 以不同分類器預測格蘭氏陽性菌資料集 Cell Wall 位置結果

在圖 4.9 當中，我們可以看到 $N_{15}C$ 及 X_6 在 Extracellular 上的表現是較佳的，但是因為 X_6 可以透過較少的特徵就能達到與 $N_{15}C$ 分類器一樣的功效，因此我們認為 X_6 在 Extracellular 位置上能較有效的預測出蛋白質位置。

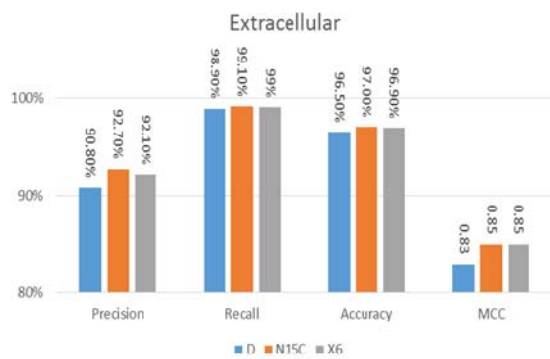


圖 4.9 以不同分類器預測格蘭氏陽性菌資料集 Extracellular 位置結果

3. 古生菌

在格蘭氏陽性菌方面，其第一層遺傳演算法預測的結果如表 4.3，因此在這邊我們選用了四個分類器 C 、 D 、 D_{j1} 、 D_{j3} 再分別對四個位置進行預測。

表 4.3 古生菌資料集第一層遺傳演算法預測的結果

位置 [†]	分類器	MCC	選取特徵數
C	D_{j1}	0.95	83
CM	D	0.93	58
CW	C	1.00	5
EC	D_{j3}	1.00	40

[†]位置縮寫：C=Cytoplasmic, CM=Cytoplasmic Membrane, CW=Cell Wall, EC=Extracellular.

在圖 4.10 當中，我們可以看到 D_{j1} 的 MCC 高達 0.95，準確率也有 98.6%，因此我們認為透過 D_{j1} 做為特徵來預測 Cytoplasmic 是有效的，雖然 D 的 MCC 值也達 0.94，但是我們還是覺得 D_{j1} 是較好的分類特徵。

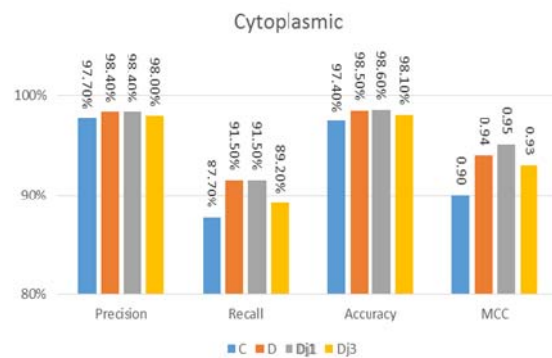


圖 4.10 以不同分類器預測古生菌資料集 Cytoplasmic 位置結果

在圖 4.11 當中， D 與 D_{j1} 特徵在預測時的 MCC 值都是 0.93，但是在特徵數比較上， D 特徵能透過較少的特徵就能達到相同的 MCC 值，因此我們認為在 Cytoplasmic Membrane 位置上， D 特徵是比較有意義的。

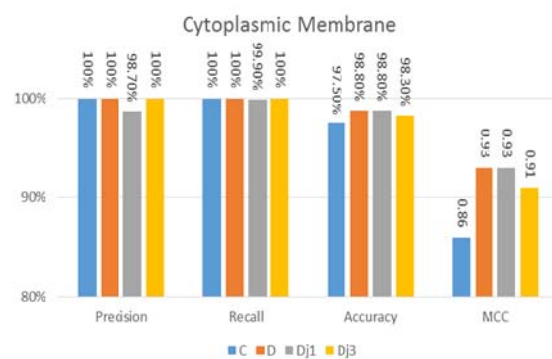


圖 4.11 以不同分類器預測古生菌資料集 Cytoplasmic Membrane 位置結果

在圖 4.12 當中，因為 Cell Wall 是細胞內位置較明顯的，因此在決定特徵上，我們就直接選用特徵數較少的 C 特徵當做是最好的特徵分類器。

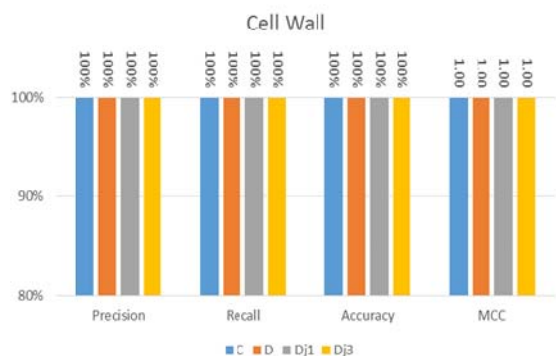


圖 4.12 古生菌資料集使用不同分類器測 Cell Wall 結果

在圖 4.13 當中，我們值皆可以看到 D_β 在 MCC 的預測上面是 1.00 的，所以我們將他當做是 Extracellular 位置上面一個較有意義的特徵。

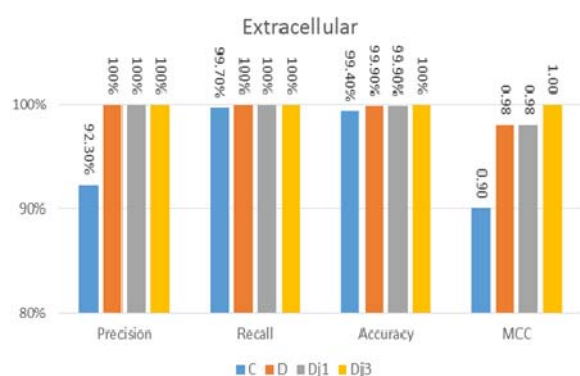


圖 4.13 古生菌資料集使用不同分類器測 Extracellular 結果

● 合併特徵所預測的結果

因為我們的特徵分類器有 28 個，從單一特徵分類器的預測結果來看，能預測出位置的特徵都不同，我們將特徵分類器做合併再進行預測。

第二層將合併第一層的所有分類器的結果進行篩選，第三層則整併前兩層結果，將第二層所篩選的分類器，以及其分類器在第一層所萃取的特徵值進行合併，再次篩選，得到最後的合併特徵分類器進行預測。

1. 格蘭氏陰性菌

在格蘭氏陰性菌方面，詳見表 4.4，在位置預測方面其實是有優有劣的，不過我們在 Outer Membrane 及 Extracellular 這兩個位置上 MCC 值有明顯的比 PSORTb3.0 方法明顯要高的很多。

表 4.4 格蘭氏陰性菌資料集與其他方法的比較結果

	位置	精確率	靈敏度	準確率	MCC
我們的方法	C	97.6%	98%	97.2%	0.94
	CM	96.8%	88.7%	97.2%	0.91
	P	90.2%	69.8%	97.9%	0.78
	OM	97.9%	86%	98.9%	0.91
	EC	93.8%	71.8%	98.3%	0.81
	Overall	97.1%	93.6%	97.8%	-

	位置	精確率	靈敏度	準確率	MCC
PSORTb v.30.	C	98.1%	98.4%	97.5%	0.94
	CM	96.4%	91.5%	97.6%	0.92
	P	91.3%	72.4%	99.0%	0.81
	OM	94.3%	75.3%	99.2%	0.84
	EC	92.0%	61.3%	98.3%	0.74
	Overall	97.3%	94.1%	98.3%	-

†位置縮寫：C=Cytoplasmic, CM=Cytoplasmic Membrane, P=Periplasmic, OM=Outer Membrane, EC=Extracellular.

2. 格蘭氏陽性菌

在格蘭氏陽性菌方面，詳見表 4.5，我們的方法在 MCC 的比較上就比 PSORTb3.0 較佳，如果單看一個位置的話，在 Cell Wall 上的 MCC 比原本的 0.80 高了 0.12，來到了 0.92。

表 4.5 格蘭氏陽性菌資料集與其他方法的比較結果

	位置	精確率	靈敏度	準確率	MCC
我們的方法	C	97.5%	98.7%	97.3%	0.93
	CM	99.4%	86.1%	98%	0.91
	CW	94.9%	90.2%	99.5%	0.92
	EC	95.9%	82.7%	97.5%	0.88
	Overall	97.5%	94.7%	98.1%	-

	位置	精確率	靈敏度	準確率	MCC
PSORTb v.30.	C	99.0%	96.9%	96.8%	0.91
	CM	96.0%	86.2%	97.7%	0.90
	CW	88.9%	72.7%	99.6%	0.80
	EC	92.9%	67.6%	97.3%	0.78
	Overall	98.2%	93.1%	97.9%	-

†位置縮寫：C=Cytoplasmic, CM=Cytoplasmic Membrane, CW=Cell Wall, EC=Extracellular.

3. 古生菌

在古生菌方面，在整體的精確率比前人的方法高了 1.9%，來到 99.1%；整體的靈敏度比前人的方法高了 5.5%來到 98.9%；整體準確率比前人方法高了 1.8%來到 99.5%，在 Cell Wall 及 Extracellular 的 MCC 值都達到 1.00，詳見表 4.6。

表 4.6 古生菌資料集與其他方法的比較結果

	位置	精確率	靈敏度	準確率	MCC
我們的方法	C	99%	100%	99.1%	0.97
	CM	100%	89.4%	98.9%	0.94
	CW	100%	100%	100%	1.00
	EC	100%	100%	100%	1.00
	Overall	99.1%	98.9%	99.5%	-
PSORTb v.30.	C	98.1%	96.9%	95.8%	0.85
	CM	88.2%	79.8%	96.6%	0.82
	CW	100%	83.3%	99.6%	0.91
	EC	100%	54.2%	98.6%	0.73
	Overall	97.2%	93.4%	97.7%	-

† 位置縮寫：C=Cytoplasmic, CM=Cytoplasmic Membrane, CW=Cell Wall, EC=Extracellular.

● 對 PAO1 進行預測

爲了比較我們分類器的準確率，我們使用額外的獨立資料集進行預測，而爲了減少測試和訓練資料集的誤差，因此我們選用了格蘭氏陰性菌有機體綠膿桿菌 PAO1，而其中總共包含 171 個 Cytoplasmic 位置的蛋白質，接著用我們的方法與前人的方法對資料集進行預測，預測結果我們的靈敏度只有 81.3% 比 PSORTb3.0 的靈敏度值要來的低了 10%。

而會造成預測結果較差的原因，是因爲 PSORTb3.0 的預測合併了 BLAST 等許多工具的使用，會有較大的序列背景資料庫可以查詢，因此比起未經學習的機器學習類工具的準確率更高，這樣的比較方法略爲不公平，但是在精確率的上我們都接近 96%，因此我們預測出來的結果精確率還是相當高的。

爲了補足我們方法的缺點，因此我們再以 CELLO 這個預測工具的答案作爲輔助成爲整合的方法，因爲我們的分類方式是每個位置都是一個分類器，而分類器的預測答案只會“是這個位置”或“不是這個位置”，因此當我們的位置分類器都預測不到的時候才採用 CELLO 的答案，因爲 CELLO 一定會給一個答案，因此對於無法預測的序列我們就透過 CELLO 工具來補足我們的缺點，預測結果如表 4.7。

表 4.7 對格蘭氏陰性菌 PAO1 資料集進行預測的結果

方法	靈敏度	精確率
我們的方法	81.3%	95.9%
PSORTb3.0	91.8%	96.3%
整合的方法	95.3%	95.3%

● 將整合的分類器對格蘭氏陰性菌資料集進行檢驗

將整合過後的分類器對原始的格蘭氏陰性菌資料集進行預測，發現會有六個蛋白質的細胞位置無法預測，如表 4.8。

而在其餘的位置則預測到有 3 個蛋白質帶有多個位置的可能，如表 4.9，其中 GI number 爲

19944 被預測出是 Cytoplasmic 與 Periplasmic；GI number 爲 81364502 被預測出來是 Cytoplasmic 與 Cytoplasmic Membrane；GI number 爲 47606230 被預測出來是 Cytoplasmic 與 Extracellular。

表 4.8 整合方法在格蘭氏陰性菌 PAO1 無法預測的蛋白質

NCBI GI number	資料庫註解
129616	Major pilu subunit operon regulatory protein PapB
729182	Competence protein S.
730255	PA-I galactophilic lectin
55584093	Secretion monitor
76364240	Cold shock protein CspA
15597937	50S ribosomal protein L20

表 4.9 整合方法在格蘭氏陰性菌 PAO1 預測多個位置的蛋白質

NCBI GI number	資料庫註解
119944	Ferredoxin-2
81364502	NADH-quinone oxidoreductase subunit 15
47606230	Yop proteins translocation protein X

5 結論

透過支持向量機結合遺傳演算法基於蛋白質胺基酸組成份的方法可以有效的預測出格蘭氏陰性菌、格蘭氏陽性菌、古生菌細胞內蛋白質的位置，透過與 PSORTb3.0 方法的比較之下，格蘭氏陰性菌的整體精確率及靈敏度及準確率輸的分數都不超過 0.5 個百分點；在格蘭氏陽性菌的整體靈敏度及準確率從原本的 93.1% 提升到 94.7% 及 97.9% 提升到 98.1%；古生菌的靈敏度更從 93.4% 提升到了 98.9%，將近五個百分點。

爲了改善第一層遺傳演算法的結果，我們使用三層的遺傳演算法，從表 4.1 與表 4.4 的結果來看，格蘭氏陰性菌的 Cytoplasmic 與 Outer Membrane 的 MCC 分別從 0.92 提升到 0.94 與 0.89 提升到 0.91；從表 4.2 與表 4.5 的結果來看，格蘭氏陽性菌 Extracellular 的 MCC 從原本的 0.85 提升到了 0.88；從表 4.3 與表 4.6 的結果來看，古生菌的 Cytoplasmic 的 MCC 從原本 0.95 提升到了 0.98，從結果來看，使用三層的整合機器學習演算法架構確實能有助於提升預測結果。

參考文獻

- [1] M. C. Hung and W. Link, "Protein localization in disease and therapy," *J Cell Sci*, vol. 124, pp. 3381-92, Oct 15 2011.
- [2] N. Y. Yu, J. R. Wagner, M. R. Laird, G. Melli, S. Rey, R. Lo, *et al.*, "PSORTb 3.0: improved

protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes," *Bioinformatics*, vol. 26, pp. 1608-15, Jul 1 2010.

[3] S. Hua and Z. Sun, "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics*, vol. 17, pp. 721-8, Aug 2001.

[4] K. Nakai and P. Horton, "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization," *Trends Biochem Sci*, vol. 24, pp. 34-6, Jan 1999.

[5] J. L. Gardy, C. Spencer, K. Wang, M. Ester, G. E. Tusnady, I. Simon, *et al.*, "PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria," *Nucleic Acids Res*, vol. 31, pp. 3613-7, Jul 1 2003.

[6] S. Rey, M. Acab, J. L. Gardy, M. R. Laird, K. deFays, C. Lambert, *et al.*, "PSORTdb: a protein subcellular localization database for bacteria," *Nucleic Acids Res*, vol. 33, pp. D164-8, Jan 1 2005.

[7] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. S. Wishart, B. Poulin, *et al.*, "Predicting subcellular localization of proteins using machine-learned classifiers," *Bioinformatics*, vol. 20, pp. 547-56, Mar 1 2004.

[8] S. Matsuda, J. P. Vert, H. Saigo, N. Ueda, H. Toh, and T. Akutsu, "A novel representation of protein sequences for prediction of subcellular location using support vector machines," *Protein Sci*, vol. 14, pp. 2804-13, Nov 2005.

[9] K. C. Chou and H. B. Shen, "Large-scale predictions of gram-negative bacterial protein subcellular locations," *J Proteome Res*, vol. 5, pp. 3420-8, Dec 2006.

[10] C. S. Yu, C. J. Lin, and J. K. Hwang, "Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions," *Protein Sci*, vol. 13, pp. 1402-6, May 2004.

[11] C. S. Yu, Y. C. Chen, C. H. Lu, and J. K. Hwang, "Prediction of protein subcellular localization," *Proteins*, vol. 64, pp. 643-51, Aug 15 2006.

[12] C. Huang and J. Yuan, "Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites," *Biosystems*, vol. 113, pp. 50-7, Jul 2013.

[13] C. K. Stover, X. Q. Pham, A. L. Erwin, S. D.

Mizoguchi, P. Warrener, M. J. Hickey, *et al.*, "Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen," *Nature*, vol. 406, pp. 959-64, Aug 31 2000.

[14] V. N. Vapnik, *The nature of statistical learning theory*. New York: Springer, 1995.

[15] C. C. Chang and C. J. Lin, "LIBSVM: A Library for Support Vector Machines," *Acm Transactions on Intelligent Systems and Technology*, vol. 2, 2011.

[16] C. H. Lu, Y. C. Chen, C. S. Yu, and J. K. Hwang, "Predicting disulfide connectivity patterns," *Proteins*, vol. 67, pp. 262-70, May 1 2007.