基於胺基酸片段之統計與 SVM 預測線性抗原決定基

Linear Epitope Prediction Using Statistics of Amino Acid Segments and SVM

林雅琪、王信偉、吳偉國、白敦文*

國立台灣海洋大學 資訊工程學系

*twp@mail.ntou.edu.tw

摘要

B 細胞抗原決定基在設計開發疫苗和誘生抗體反應的研究上扮演著相當重要的角色。但以生物實驗直接辨識蛋白質分子中抗原決定基的位置是相當耗時且需要大量的實驗資源，所以透過資訊技術開發一個具高度正確預測率的工具，以加速並簡化實驗的進行是非常重要且具有挑戰性的。本論文提出一套可以有效改善原有只採用物理化學特性為基礎的線性抗原決定基預測系統，基於原系統物理化學性質之分析，再透過外加使用長度從 2 到 4 個已確認抗原特徵的胺基酸片段作為 SVM(Support Vector Machine)的特徵值，可結合原系統的預測優勢，明顯提升對線性抗原決定基預測的正確率。最後，本論文使用與 HIV 抗原相關的 10 條蛋白質序列進行預測分析，雖然實驗結果顯示新組合系統在平均敏感度降低 8.7%，但平均識別度提升 22%，整體的平均正確率改善 9.6%，平均陽性預測值也改善了 12.7%。若與知名的 BepiPred 系統比較，本系統的預測表現皆明顯優於該系統：平均敏感度高 3.2%、平均識別度高 15.5%、平均正確率高 8.8% 及平均陽性預測值高 12.9%。

關鍵字：線性抗原決定基、物理化學性質、胺基酸對、LEPD、SVM

ABSTRACT

B-cell epitopes play an important role for developing synthetic peptide vaccines and inducing antibody responses. Applying biological experiments for epitope identification is time consuming and demands a lot of experimental resources. Therefore, it is useful and challenging to develop a linear epitope prediction system with high precision rates based on the computational technologies. In this paper, a combinatorial method to improve physico-chemical property based linear epitope prediction systems is proposed. Here, a collected set of verified epitope and non-epitope segments ranging from 2 to 4 amino acids were trained and applied as the features of SVM (Support Vector Machine). With the combination of physico-chemical characteristics and SVM classifier, the performance can be effectively improved. Ten HIV related antigens were adopted for system evaluation, and the results showed that the proposed method decreased the average

sensitivity with 8.7% while the average specificity was increased by 22%. The average accuracy was increased by 9.6% and the average positive predictive value was increased by 12.7%. In comparison with the well-known BepiPred system, experimental results have shown that the proposed system outperforms BepiPred system in all respects, including a higher average sensitivity of 3.2%, specificity of 15.5%, accuracy of 8.8%, and positive predictive value of 12.9%.

Keyword：Linear Epitope、physico-chemical property、amino acid pair、LEPD、SVM

一、介紹

抗原決定基(epitope)是位於抗原表面的部分位置，它是可以引起免疫反應的部位。抗原決定基是抗體和抗原接合的關鍵，它可以被特定抗體辨識並結合。故抗原決定基的預測，可以作為生物學家在疫苗開發的前置分析，也因此它對疾病的預防與診斷極具重要性。抗原決定基的種類主要分為兩種：線性抗原決定基及組合式抗原決定基[1]，如圖 1 所示，中間藍色長鏈部分代表的是抗體的重鏈(heavy chain)；紅色短鏈則表示為抗體的輕鏈 (light chain)分子，綠色圓形與重鏈或輕鏈接觸的分子是代表抗原的位置，圖 1 左綠色的圓形為線性抗原決定基，它是由一段連續的胺基酸片段所組成，圖 1 右側綠色的圓形為組合式抗原決定基，它是由非連續的胺基酸片段在空間中摺疊後所組成。雖然估測約有 90%的抗原皆為組合式抗原決定基的形式存在[2]，但因為許多抗原結構仍然未被解析出來，因此在空間資訊仍無法取得的情形下，正確線性抗原決定基的預測更顯重要，因此本論文是針對線性抗原決定基的預測及分析進行討論。

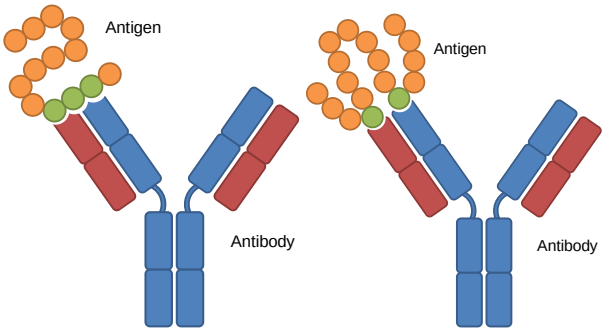

圖 1 線性抗原決定基(左)及組合式抗原決定基(右)

常見預測抗原決定基的方法是根據組成胺基酸的物理化學性質，例如結構的彈性[3]、表面可接觸性[4]、親水性[5]、二級結構[6]、抗原特性[7]等，現有許多線性抗原決定基預測系統，像是 PREDITOP[7]、PEOPLE[8]、BEPITOPE[9]、BcePred[10]都是以胺基酸的物理化學特性進行預測，不過 2005 年 Blythe and Flower[11]的研究結果顯示，僅以物理化學特性進行預測，最佳的結果也只能比隨機挑選好一點。因此許多研究嘗試以機器學習的方法改善預測準確度，例如 BepiPred[12]使用 Hidden Markov Model(HMM)結合 Levitt[13]和 Parker et al.[5]提出的兩種胺基酸物理化學性質進行預測。

根據 Chen[14]的研究結果顯示，某些胺基酸對(AAPs, Amino Acid Pairs)在抗原決定基出現的頻率遠大於在非抗原決定基出現的頻率，亦有某些特定的胺基酸對在非抗原決定基出現的頻率遠大於在抗原決定基出現的頻率，這個統計特性很值得參考使用，在其他的應用亦有許多成功的範例，例如以胺基酸對為主的統計特徵，成功地增加預測蛋白質二級結構內容的成功率[15-16]。

本論文採用長度為 2 到 4 的胺基酸片段作為統計特徵並結合 LEPD (Linear Epitope Prediction Database)[17]及 LIBSVM 函式庫 (A Library for Support Vector Machines)[18]進行線性抗原預測。LEPD 是一個目前具有全方位考量的線性抗原決定基預測系統，胺基酸的物理化學性質亦為其辨識抗原的主要考量，該系統採用一般的物理

化學性質包含二級結構特性、親水性、表面可接觸性、移動彈性及極性等。此外，過去研究亦發現，線性抗原決定基的位置並不僅位於整體抗原物理化學特性曲線的最高點區段，也有部分抗原決定基的位置是座落在局部區域的相對高點區段，因此 LEPD 系統可以同時對全域及區域的抗原片段進行分析預測，該理論是以數學型態學技術，擷取抗原特性具局部性相對高點的胺基酸片段，並視為抗原決定基之後選者，這個方法確實改善了其他使用物理化學性質預測方法的缺失。但由於 LEPD 預測的抗原決定基候選者相對過多，導致錯誤預測(false positive)偏高，為了可以有效提高其識別度(Specificity)，本實驗採用已知胺基酸片段的統計特徵和 SVM 的方法改善 LEPD 的預測結果。

二、研究方法

系統流程

本研究方法包含三個步驟，如圖 2 所示，第一階段是統計並計算已知抗原決定基資料集與非抗原決定基資料集中胺基酸片段內容所出現的次數，第二階段是藉由 1744 條訓練資料的胺基酸片段分數當特徵值進行訓練 SVM，並建立預測模型，最後步驟是將 LEPD 的預測結果，以上述的預測模型進行預測與再確認分析，下面章節將詳細介紹每一步驟。

2.1 步驟一

資料集

本論文採用 Bcipep 提供的 B 細胞抗原決定基資料集，該資料集包含了 1230 條長度由 4 到 57 個胺基酸片段且內容不重複的線性抗原決定基[19]，本論文需使用沒有重複的資料以避免重複資料影響到統計分析及預測結果，該資料集是用於統計長度從 2 到 4 的胺基酸片段的出現次數。

訓練與測試的資料集則是採用 Chen *et al.*所提供長度皆為 20 的 872 條抗原決定基(epitope)和 872 條非抗原決定基(non-epitope)，這 872 條抗原決定基原始來源同樣為上述的 Bcipep 資料庫，由於機器學習適合使用固定長度的訓練資料，但 Bcipep 提供的抗原決定基資料集的長度不等，故 Chen 研究團隊將 Bcipep 資料集中長度大於 20 的抗原決定基片段，截去兩側的胺基酸，僅保留片段中間長度為 20 的抗原決定基，長度不足 20 的片段則往兩側延伸直到長度達到 20 為止，例如長度為 8 的抗原決定基，就必須在兩側各增加 6 個鄰近的胺基酸片段，鄰近的 6 個胺基酸來源就是從原始蛋白質序列中取得，經由截長補短及移除重複片段的過程後，最後得到 872 條抗原決定基。另外 872 條非抗原決定基則是從 Swiss-Prot[20]資料庫中任意選出長度為 20 但不與抗原決定基重複的胺基酸片段。

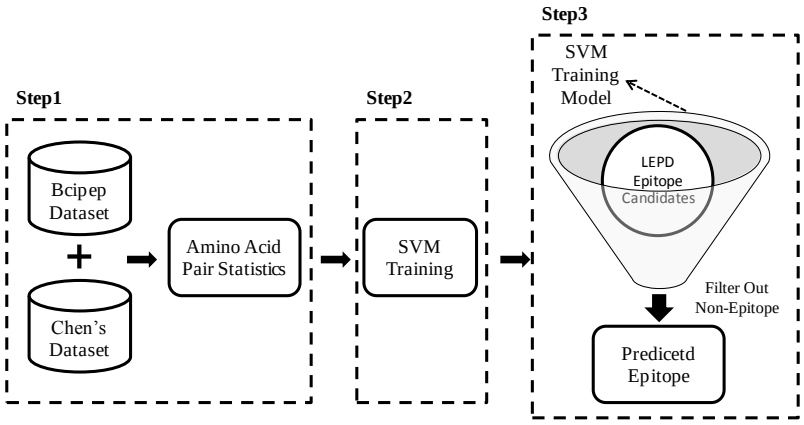

圖 2 系統流程圖

胺基酸片段統計分析

本實驗參考 Chen將長度為2的胺基酸對作為特徵預測線性抗原決定基，並延伸採用長度為 3 或 4 個胺基酸片段作為特徵值之分析，下面我們詳細介紹這三個特徵，表 1 定義了 9 個胺基酸片段統計所需變數。

表 2 胺基酸片段統計所使用的變數代號

	變數代號	說明
(1)	f^{+}_{AAP}	抗原決定基資料集中胺基酸對出現的機率
(2)	f^{-}_{AAP}	非抗原決定基資料集中胺基酸對出現的機率
(3)	R_{AAP}	每個胺基酸對的分數
(4)	N^{+}_{3AAP}	抗原決定基資料集中胺基酸片段長度為3所出現的次數
(5)	$Total_{3AAP}$	抗原決定基資料集中所有胺基酸片段長度為3出現的總次數
(6)	R_{3AAP}	胺基酸片段長度為3的分數
(7)	N^{+}_{4AAP}	抗原決定基資料集中胺基酸片段長度為4所出現的次數
(8)	$Total_{4AAP}$	抗原決定基資料集中所有胺基酸片段長度為4出現的總次數
(9)	R_{4AAP}	胺基酸片段長度為4的分數

表 3 列舉少數長度為 2 的胺基酸對中，抗原決定基與非抗原決定基胺基酸對的組成比較範例

AAP	抗原決定基(%)	非抗原決定(%)	比例
CW	0.043	0.006	7.062
MW	0.104	0.018	5.717
MN	0.018	0.097	0.189
HF	0.043	0.187	0.228

長度為 2 的胺基酸對

首先統計 400 個所有可能胺基酸對的組合分別在抗原決定基資料集及非抗原決定基資料集裡出現的機率。

表 3 列出一些範例說明胺基酸對在抗原決定基資料集及非抗原決定基資料集裡的出現機率的明顯差異性，其中"比例"所代表的意思是胺基酸對在抗原決定基資料集的出現機率除以同一胺基酸對在非抗原決定基資料集的出現機率，由表中可以看出某些胺基酸對(AAPs)在抗原決定基出現的機率遠大於在非抗原決定基出現的機率，亦有某些胺基酸對在非抗原決定基出現的機率遠大於在抗原決定基出現的機率。大部分來說，如果某個較常在抗原決定基資料集裡出現的胺基酸對，這個胺基酸是抗原決定基的機率就比較高。為了能使用這個性質，我們將每個胺基酸對是抗原決定基的可能性表示為一個分數，以方便後續之計算。分數(R_{AAP})的算法第一步驟如方程式(1)所示，將抗原決定基資料集中胺基酸對出現的機率(f^{+}_{AAP})除以非抗原決定基資料集裡胺基酸對出現的機率(f^{-}_{AAP})，再取其 log 值。第二步驟如方程式(2)所示，將第一步驟的計算結果正規化，使分數介於 0 到 1 之間。

$$R_{AAP} = \log\left(\frac{f^{+}_{AAP}}{f^{-}_{AAP}}\right) \quad (1)$$

$$nR_{AAP} = \frac{R_{AAP} - \min}{\max - \min} \quad (2)$$

變數 max 和 min 代表在正規化之前 R_{AAP} 的最大值與最小值，而正規化的目的則是為了避免某一個極大或極小的特徵值影響之後的分類器學習。

因為訓練資料集的每條抗原決定基及非抗

原決定基長度皆為 20，故每一條序列都存在 19 個胺基酸對，我們將每條序列中胺基酸對所對應的出現頻率分數累加之後再除以 19，即為這條序列的分數。舉例來說，PISPIETVPV 這條長度為 10 的序列，就可以分成 9 個胺基酸對，分別是：PI、IS、SP、PI、IE、ET、TV、VP、PV，此序列的胺基酸對分數就等於 9 個胺基酸對的分數累加之後再除以 9。依照上述算法，最後可以得到每條訓練資料的胺基酸對分數，我們以此分數當作 SVM 的第一個特徵值。

長度為 3 和 4 的胺基酸片段

首先統計分析抗原決定基資料集，$20^3 = 8000$ 個長度為 3 的胺基酸片段的出現次數與 $20^4 = 160000$ 個長度為 4 的胺基酸片段的出現次數。由於非抗原決定基資料集裡長度為 3 與長度為 4 的胺基酸片段有些出現次數為 0，故計算長度為 3 與長度為 4 的胺基酸片段分數與計算長度為 2 的胺基酸對分數的方法有些不同。長度為 3 的胺基酸片段分數（R_{3AAP}）計算方法如方程式(3)、(4)所示，直接將抗原決定基資料集長度為 3 的胺基酸片段出現的次數（N^{+}_{3AAP}）除以總出現次數（$Total_{3AAP}$）後再正規化，使分數介於 0 到 1 之間。長度為 4 的胺基酸片段分數也是同樣的算法，如方程式(5)、(6)所示。

$$R_{3AAP} = \frac{N^{+}_{3AAP}}{Total_{3AAP}} \quad (3)$$

$$nR_{3AAP} = \frac{R_{3AAP} - \min_3}{\max_3 - \min_3} \quad (4)$$

$$R_{4AAP} = \frac{N^{+}_{4AAP}}{Total_{4AAP}} \quad (5)$$

$$nR_{4AAP} = \frac{R_{4AAP} - \min_4}{\max_4 - \min_4} \quad (6)$$

2.2 步驟二

SVM

決定上述與出現頻率相關的三個特徵後，輸入所有訓練資料集及相對的特徵值到 LIBSVM，本實驗使用分五群交叉驗證(Fivefold Cross-Validation)的方式，測試資料平均分成五等份，挑選四份作為訓練資料集以建立分類模型，剩下的一份作為驗證用資料集，每一等份都經由設定為驗證用的資料集後，最後計算五次個別分類準確率的平均，即完成分類驗證的動作。再將欲預測的蛋白質序列至輸入至 LEPD 系統，並將 LEPD 預測出的線性抗原決定基候選者交由 SVM 進行分類，最後移除 SVM 視為非線性抗原決定基的候選者，以改善 LEPD 預測的線性抗原決定基過多的問題，實際降低錯誤預測的數量。

效能計算

敏感度(Sensitivity)、識別度(Specificity)、正確率(Accuracy)和陽性預測值(Positive Predictive Value)這四個指標常用來評估預測方法的效能。在式(7)中的敏感度代表正確預測抗原決定基的機率；式(8)中的識別度代表正確預測非抗原決定基的機率；式(9)中的正確率代表正確預測抗原決定基及非抗原決定基的機率；式(10)中的陽性預測值代表預測為抗原決定基的正確機率。這些參數的方程式定義如下：

$$Sensitivity = \frac{TP}{TP + FN} \quad (7)$$

$$Specificity = \frac{TN}{TN + FP} \quad (8)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (9)$$

$$PPV = \frac{TP}{TP + FP} \quad (10)$$

其中 TP、FP、TN、FN 分別代表 true positive、

false positive、true negative 和 false negative 的縮寫。

三、結果

本實驗採用 HIV(Human Immunodeficiency Virus)[21]抗原決定基資料集進行測試，該資料集包含 10 條蛋白質序列。如表 4 所示，第一行為蛋白質名稱，第二行至第五行分別為 TP、FP、TN 及 FN 的數值，第六行至第九行分別為預測的敏感度、識別度、正確率及陽性預測值，表格中白底部分的列資料顯示為原有 LEPD 系統的預測結果，灰底部分的列資料則為本論文所提出的整合系統經過濾 LEPD 候選者之後的結果，最後一列的資料顯示同樣序列資料經 BepiPred 預測的表現。

實驗結果顯示，LEPD 系統的線性抗原決定基的預測表現在經由本系統進一步過濾之後，雖然預測結果的平均敏感度降低了 8.7%，但平均識別度則改善了 22%，整體的平均正確率也改善了 9.6%，平均陽性預測值更是增加了 12.7%。若採用同一 HIV 資料集並選擇知名的 BepiPred 預測系統進行預測及比較，其結果顯示由 BepiPred 預測的平均敏感度為 48.8%、平均識別度為 61.2%、平均正確率為 56.3%及平均陽性預測值為 60.9%，本系統整體表現皆優於 BepiPred 系統，包括平均敏感度高 3.2%、平均識別度高 15.5%、平均正確率高 8.8%及平均陽性預測值高 12.9%。圖 3 至圖 6 分別為本系統過濾前及過濾後的敏感度、識別度、正確率、及陽性預測值的比較圖，x 軸代表的是表 4 中蛋白質序列的編號，藍色方塊的線條代表 LEPD 的預測結果，紅色三角形的線條代表經由胺基酸片段特徵及 SVM 過濾後的結果，圖中可看出在大多數的序列中，本系統能有效地提升識別度、正確率及陽性預測值，並且大部分有效地維持原有敏感度的表現，但其中第三條與第四條蛋白質的敏感度降低特別多，分別降低了 0.27 與 0.25，這是因為這兩條序列的某些胺基酸片段在 Bcipep 抗原決定基資料集中出現的機率較低，並且較常出現在 Chen 所提供的非抗原決定基資料集的緣故，未來隨著經過實驗驗證的線性抗原決定基資料集的增加，本系統將可以收錄並訓練更多的線性抗原決定基片段，該統計資訊一定可以使預測的結果更增完善。

表 4 HIV 資料集預測結果

Protein Name	TP	FP	FN	TN	Sensitivity	Specificity	Accuracy	PPV
1. Integrase	65	94	52	77	0.555556	0.450292	0.493056	0.408805
	65	36	52	135	0.555556	0.789474	0.694444	0.643564
2. NEF	95	13	84	14	0.530726	0.518519	0.529126	0.879630
	92	8	87	19	0.513967	0.703704	0.538835	0.920000
3. Protease	16	34	2	47	0.888889	0.580247	0.636364	0.320000
	11	6	7	75	0.611111	0.925926	0.868687	0.647059
4. RT	54	243	32	231	0.627907	0.487342	0.508929	0.181818
	32	28	54	446	0.372093	0.940928	0.853571	0.533333
5. REV	44	21	29	22	0.602740	0.511628	0.568965	0.676923
	42	13	31	30	0.575342	0.697674	0.620690	0.763636

6.	**TAT**	53	8	29	11	0.646341	0.578947	0.633663	0.868852
		43	8	39	11	0.524390	0.578947	0.534653	0.843137
7.	**gp160**	288	159	256	153	0.529412	0.490385	0.515187	0.644295
		269	106	275	206	0.494485	0.660256	0.554907	0.717333
8.	**p17**	46	4	43	39	0.516854	0.906977	0.643939	0.920000
		46	4	43	39	0.516854	0.906977	0.643939	0.920000
9.	**p24**	101	18	92	20	0.523316	0.526316	0.52381	0.848740
		100	9	93	29	0.518135	0.763158	0.558442	0.917431
10.	**p2p7p1p6**	30	53	16	38	0.652174	0.417582	0.49635	0.361446
		24	27	22	64	0.521739	0.703297	0.642336	0.470588
Average						0.607391	0.546824	0.554939	0.611051
						0.520367	0.767034	0.651050	0.737608
BepiPred						0.488419	0.611719	0.563224	0.609086

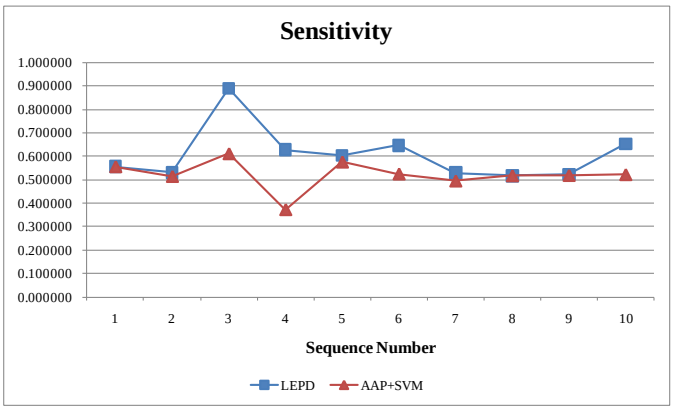

圖 3 敏感度結果比較圖

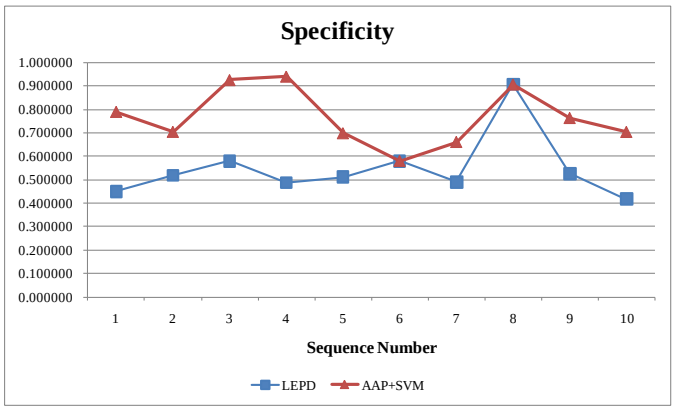

圖 4 識別度結果比較圖

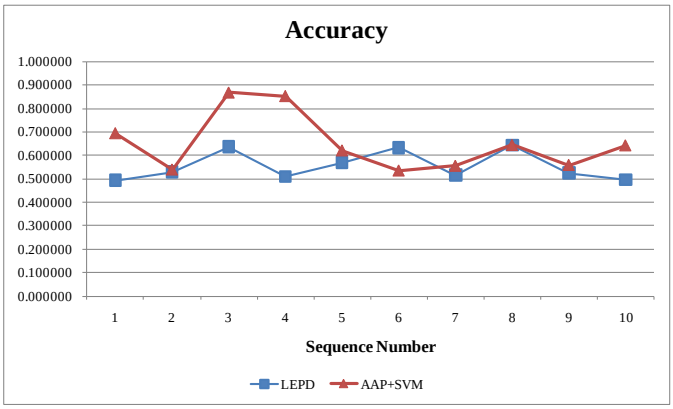

圖 5 正確率結果比較圖

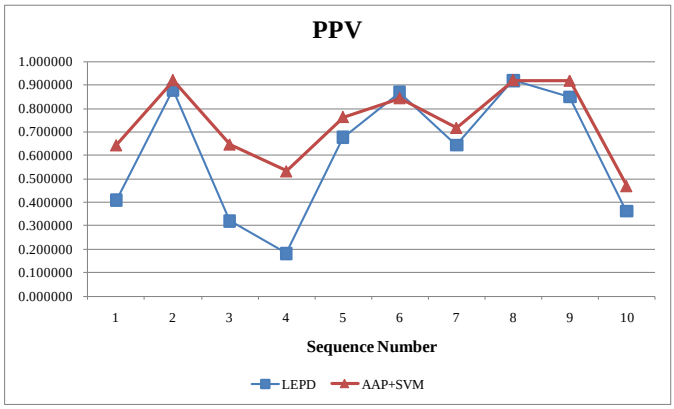

圖 6 陽性預測值結果比較圖

四、結論

本實驗採用長度從 2 到 4 個胺基酸的片段作為 SVM 分類特徵，結合以胺基酸物理化學性質為主的 LEPD 預測系統進行預測線性抗原決定基的改善，與先前方法對照之下，結合 SVM 技術的預測結果在識別度、正確率及陽性預測值皆有提升。當與 BepiPred 系統比較時，本論文提出整合系統的預測結果也較佳。線性抗原決定位置的預測，對設計開發疫苗及疾病診斷等生物研究與醫學相關應用是極具價值的。若要進一步提升本系統的預測準確度，將挑選更精確的特徵值及陸續增加抗原決定基資料集，進而減少假警報的錯誤率，以達到更佳的預測結果。

誌謝

本論文完成承蒙國科會之計畫經費贊助，計畫編號為 NSC98-2627-B-019-003 及 NSC 98-2221-E-019 -031 -MY2，僅此誌謝。

[1] Barlow, D.J., Edwards, M.S., and Thornton, J.M., *Continuous and discontinuous protein antigenic determinants.* Nature, 1986. **322**(6081): p. 747-8.

[2] Walter, G., *Production and use of antibodies against synthetic peptides.* J Immunol Methods, 1986. **88**(2): p. 149-61.

[3] Vihinen, M., Torkkila, E., and Riikonen, P., *Accuracy of protein flexibility predictions.* Proteins, 1994. **19**(2): p. 141-9.

[4] Emini, E.A., et al., *Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide.* J Virol, 1985. **55**(3): p. 836-9.

[5] Parker, J.M., Guo, D., and Hodges, R.S., *New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites.* Biochemistry, 1986. **25**(19): p. 5425-32.

[6] Debelle, L., et al., *Predictions of the secondary structure and antigenicity of human and bovine tropoelastins.* Eur Biophys J, 1992. **21**(5): p. 321-9.

[7] Kolaskar, A.S. and Tongaonkar, P.C., *A semi-empirical method for prediction of antigenic determinants on protein antigens.* FEBS Lett, 1990. **276**(1-2): p. 172-4.

[8] Alix, A.J., *Predictive estimation of protein linear epitopes by using the program PEOPLE.* Vaccine, 1999. **18**(3-4): p. 311-4.

[9] Odorico, M. and Pellequer, J.L., *BEPITOPE: predicting the location of continuous epitopes and patterns in proteins.* J Mol Recognit, 2003. **16**(1): p. 20-2.

[10] Saha, S.a.R., G.P.S., *BcePred: Prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties.* Lecture Notes in Computer Science, 2004. **3239**: p. 197-204.

[11] Blythe, M.J. and Flower, D.R., *Benchmarking B cell epitope prediction: underperformance of existing methods.* Protein Sci, 2005. **14**(1): p. 246-8.

[12] Larsen, J.E., Lund, O., and Nielsen, M., *Improved method for predicting linear B-cell epitopes.* Immunome Res, 2006. **2**: p. 2.

[13] Levitt, M., *Conformational preferences of amino acids in globular proteins.*

Biochemistry, 1978. **17**(20): p. 4277-85.

[14] Chen, J., et al., *Prediction of linear B-cell epitopes using amino acid pair antigenicity scale*. Amino Acids, 2007. **33**(3): p. 423-8.

[15] Chou, K.C., *Using pair-coupled amino acid composition to predict protein secondary structure content.* J Protein Chem, 1999. **18**(4): p. 473-80.

[16] Liu, W. and Chou, K.C., *Prediction of protein secondary structure content.* Protein Eng, 1999. **12**(12): p. 1041-50.

[17] Chang, H.T., Liu, C.H., and Pai, T.W., *Estimation and extraction of B-cell linear epitopes predicted by mathematical morphology approaches.* J Mol Recognit, 2008. **21**(6): p. 431-41.

[18] Chang, C.C. and Lin, C.J., *LIBSVM: a library for support vector machines.* 2001.

[19] Saha, S., Bhasin, M., and Raghava, G.P., *Bcipep: a database of B-cell epitopes.* BMC Genomics, 2005. **6**(1): p. 79.

[20] Bairoch, A. and Apweiler, R., *The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.* Nucleic Acids Res, 2000. **28**(1): p. 45-8.

[21] Kober, B., et al., HIV Immunology and HIV/SIV Vaccine Databases 2003, 2003.