人類蛋白質在細胞內位置的預測

徐筱姍[1]　游景盛[1,2]
[1]逢甲大學生醫資訊暨生醫工程工程學程
[2]逢甲大學資訊工程學系
yucs@fcu.edu.tw

摘要

細胞內蛋白質的功能與其所在子細胞位置的關係十分密切，蛋白質在細胞內若要正常運作必須要在正確的子細胞位置才能發揮其功能。因此，如果想要進一步了解未知功能的蛋白質，可以藉由確認蛋白質在細胞內的位置間接獲得重要線索。藉由預測蛋白質在細胞內的位置所使用的特徵，許多研究工作如人類發展藥物篩檢、疫苗設計與基因的註解具有醫藥發展的潛在應用。在這篇研究裡，我們利用 *n*-胜肽組成份特性以支持向量機器(Support Vector Machine, SVM)的演算法進行改進預測人類蛋白質在細胞內位置的準確性。我們將每條蛋白質的序列使用不同的特徵藉由支持向量機器分類器來進行第一階預測，再將其多樣性的結果由支持向量機器分類器進行第二階整合預測。在本研究使用的Hera資料集中，分屬於細胞內的九個位置的2233人類蛋白質序列，經十次交叉驗證可達整體預測準確性 71%，與前人的研究比較，同在 74%資料集涵蓋率的條件下，我們的方法可由 78%提升至 80%。另外前人研究中剩下的 26%無法預測的資料我們也可以達到 45%預測準確率。很明顯的我們的方法可以明確的提升預測準確性，我們預期這樣的技術將有助於未來對人類基因體學與蛋白質體學的相關研究。

關鍵詞：*n*-胜肽組成份、支持向量機器、子細胞位置。

Abstract

The biological function of a protein in a cell is often closely correlated with its subcellular localization. Hence, the information about where a protein localized often offers important clues toward knowing the function of an uncharacterized sequence. The protein subcellular localization can be used as an important feature to screen for drug candidates, vaccine design, and gene products annotation. Here, we applied the support vector machine algorithm to a benchmark dataset of human protein sequence based on multiple *n*-peptide composition properties. The first cascade in our approach is that we classify the protein sequence by different feature then use Support Vector Machine (SVM) algorithm to predict subcellular localization. In second cascade, we integrate the predicted results from the first step as the features to obtain the final prediction by SVM, too. We use the benchmark Hera dataset, which including 2233 human proteins separately in 9 subcellular localizations inside of cell. Our method improves an overall classification accuracy of 80% as estimated by using a 10-fold cross-validation test with coverage of 74% in previous work. For the rest 26% sequences, our method achieves an overall classification the accuracy of 45%. This research should provide an important tool in human genomics and proteomics studies.

Keywords: *n*-peptide composition、Support Vector Machine (SVM)、subcellular localization

1. 前言

生物學家長久以來對蛋白質如何運送和細胞生理調控的關係有著濃厚的興趣，在儲存大量蛋白質子細胞位置資訊的資料庫中，複雜的蛋白質網路關係難以令生物學家整理出一系統全貌。因為蛋白質的功能與蛋白質在細胞內位置有著非常密切的關係，藉由了解蛋白質在細胞內的位置仍有助於進一步了解蛋白質的相關功能。

近年來預測蛋白質子細胞位置的相關研究，生物資訊學者藉由各式不同的蛋白質序列特性以及演算法策略發展工具，如最近鄰算法(Nearest Neighbor approach) [1]、支持向量機器 [2, 3]、貝氏網路方法(Bayesian Network approach) [4]等機器學習的自動化分類的方法應用在預測蛋白質在細胞內位置的相關研究。其中，尋找有助於機械學習以進行分類的生物意義特徵十分重要。除了已知分泌蛋白與細胞核內的蛋白質先驅物(precursors)具有訊號序列能夠引導蛋白質進入正確的子細胞位置，細胞內許多蛋白質缺乏顯而易見的序列特徵。因此許多方法選擇以 N 端的序列區固定長度計算胺基酸的組成份 [5-7] 、合併使用序列的胺基酸組成份 [8, 9] 、計算雙胜肽(dipeptides)與間隔(gapped)雙胜肽的胺基酸的組成份[10]等特徵做為發展預測蛋白質在細胞內的位置的根據。

由許多文獻已知預測不同物種的蛋白質子細胞位置的研究多具有很高的準確性。如 CELLO[2]預測格蘭氏陰性菌(Gram negative bacteria)準確度 88.9%； TargetP[1]預測植物蛋白質準確度 85.3%。而預測人類蛋白質的子細胞位置，PSLT[4]預測準確度可達 78%，但其中將近 30%的蛋白質無法進行預

測，我們希望應用支持向量機器自動化分類且從序列擷取特徵的方法進行預測，希望提升預測人類蛋白質在細胞內的位置的準確度。

在本研究中，我們計算序列中存在的 *n*-胜肽組成份之特徵並使用兩階段的支持向量機器(SVM)演算法預測人類蛋白質在細胞內的位置，除希望能提升人類蛋白質預測的準確性，同時釐清蛋白質的組成份上在不同細胞內的位置、在有不同的條件如學習資料集與預測對象的序列同源性在不同子細胞位置的意義與影響，以期應用於更多物種和更細緻的生物系統知識深入分析。

2. 材料方法

PSLT 方法在人類蛋白質在細胞內的位置的預測，單一位置預測最好的結果準確率雖高達 78%但資料庫的涵蓋率僅有 74%，因為其方法所使用的資料集僅有 74%可在 InterProt 資料庫[11]查得相對應的資訊。在本研究中我們使用相同的人類 Hera 資料集(www.mcb.mcgill.ca/~hera/PSLT)（如表 1）。

表 1. 人類 Hera 資料集。

子細胞位置(*l*)	數量
Cytosol	357
ER	340
Secreted	294
Golgi	96
Lysosome	91
Mitochondria	222
Nucleus	584
Peroxisome	44
Plasma membrane	205

我們使用 LIBSVM[12]預測人類蛋白質在細胞內的位置。首先第一階段以序列上胺基酸的組成份當作特徵，運用這些特徵訓練和測試，找出最適合的參數建立分類器。第二階段將第一階段的數個分類器預測結果為輸入值以確定預測蛋白質在細胞內的位置。我們並以十次交叉驗證(10-fold cross-validation test)程序評估預測的準確性。

我們將 n-peptide 特徵下述五類，分別是 20 種胺基酸組成份（C_n）、間隔胜肽組成份(D_g)（如圖一）、將序列等分 k 段並分別計算各段胺基酸的組成份(X_k)(如圖二)、將 20 種胺基酸依物理性質分類(如表二)後再計算分段胺基酸組成份(A_nX_k)（如圖三）、局部序列胺基酸組成份(N_iC)（如圖四）。

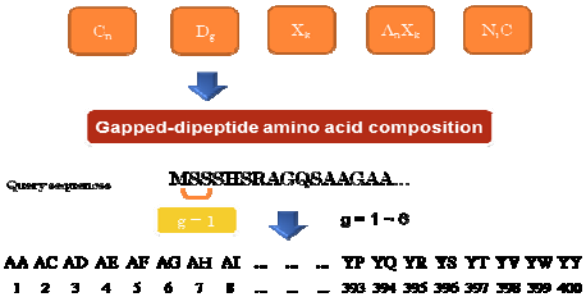

圖 1. 間隔雙胜肽組成份計算。

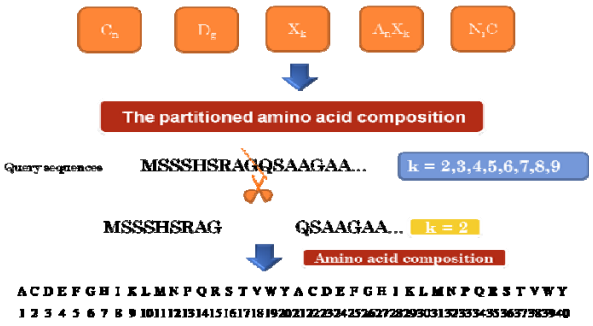

圖 2. 序列等分 k 段胺基酸組成份計算。

表 2. 胺基酸物理化學特性分類表。

分類代碼	分類依形成	胺基酸依種類
E	Acidic 酸性	DE
	Basic 鹼性	HKR
	Aromatic 芳香基團	FWY
	Amide 醯胺團	NQ
	Small hydroxyl 微羥基	ST
	Sulfur-containing 含硫氨基酸	CM
	Aliphatic 1 不飽和脂肪酸	AGP
	Aliphatic 2 飽和脂肪酸	ILV
F	Acidic 酸性	DE
	Basic 鹼性	HKR
	Polar 極性	CGNQSTY
	Nonpolar 非極性	AFILMPVW
P	Lowpolarity 低極性	LIFWCMVY
	Neutralpolarity 中極性	PATGS
	Highpolarity 高極性	HQRKNED
S	Acidic 酸性	DE
	Basic 鹼性	HKR
	Aromatic 芳香基團	FWY
	Amide 醯胺	NQ
	Small hydroxyl 微羥基	ST
	Sulfur-containing 含硫氨基酸	CM
	Aliphatic 脂肪酸	AGPILV
V	Small 小	GASCTPD
	Medium 中	NVEQIL
	Large 大	MHKFRYW
Z	Lowpolarizability 低極化	GASDT
	Mediumpolarizability 中極化	CPNVEQIL
	Highpolarizability 高極化	KMHFRYW
H	Polar 極性	RKEDQN
	Neutral 中性	GASTPHY
	Hydrophobic 疏水性	CVLIMFW

圖 3. 序列等分 k 段之胺基酸物化特性組成份。

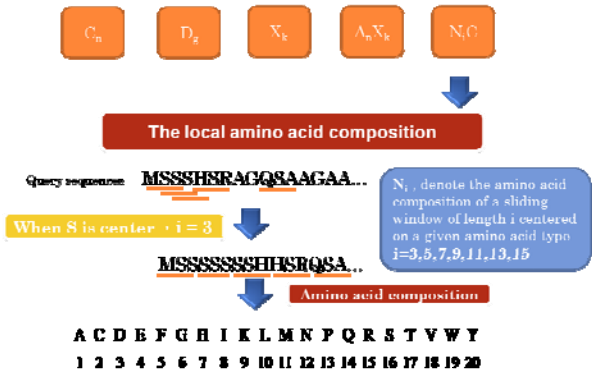

圖 4. 局部序列胺基酸組成份計算。

3. 預測表現評估

在本論文中，我們使用準確率進行預測的表現評估，並將結果和 PSLT 結果列於表 3:

P_ℓ 表示每個位置預測的準確率，ℓ 是細胞內的位置，L_ℓ 在 ℓ 的位置預測正確的蛋白質的數量， L_ℓ 在 ℓ 位置蛋白質的數量，如公式(1)。

$$P_\ell = \frac{L_\ell}{N} \times 100 \ \% \qquad (1)$$

P 表示整體的預測準確率，將每個位置的準確度累加除以九個位置，如公式(2)。

$$P = \frac{\sum_\ell P_\ell}{\ell} \times 100 \ \% \qquad (2)$$

4. 結果與討論

在本論文結果和 PSLT 結果列於表 3:

我們根據支持向量機器預測所得九個子細胞位置預測的機率值最大者視為其預測信賴值進行排序，前 74%視為 74%涵蓋率與 PSLT 進行比較，其中預測準確率達 80%，整體提升了 2%，九個位置分別：Cytosol（69%）、ER（76%）、Secreted（90%）、Golgi（29%）、Lysosome（67%）、Mitochondria（83%）、Nuclear（92%）、Peroxisome（14%）、Plasma membrane（80%）。超過半數的子細胞位置預測準確率皆有提升，Cytosol 提升 4%、ER 提升 7%、Secreted 提升 1%、Lysosome 提升 7%、Mitochondria 提升 6%。

表 3. 根據 Hera 人類蛋白質資料集，比較本論文預測蛋白質子細胞位置的預測結果與 PSLT 方法。

子細胞位置 (l)	PSLT Coverage = 74%	Our method Coverage = 74%	Our method Coverage = 26%
Cytosol	65(83)	69(93)	42(77)
ER	69(82)	76(86)	53(76)
Secreted	89(93)	90(93)	43(51)
Golgi	60(64)	29(54)	32(55)
Lysosome	60(71)	67(72)	23(55)
Mitochondria	67(71)	83(87)	43(63)
Nuclear	93(97)	92(96)	60(80)
Peroxisome	43(64)	14(36)	6(20)
Plasma Membrane	89(91)	80(86)	40(62)
Accuracy	78(86)	80(89)	45(68)

註:表格內的數值為本論文預測最高分結果的準確率，括弧內為前兩高分結果的準確率評估結果。

我們將資料集另外 26%的預測結果視為 PSLT 方法無法進行預測之資料集部份，平均準確性達 45%，九個子細胞位置預測的結果：Cytosol（42%）、ER（53%）、Secreted（43%）、Golgi（32%）、Lysosome（23%）、Mitochondria（43%）、Nuclear（60%）、Peroxisome（6%）、Plasma membrane（40%）。

本論文整體預測準確性可達 71%，同樣在九個子細胞位置中，預測第一高分為預測結果的平均準確性各別為：Cytosol（60%）、ER（70%）、Secreted（81%）、Golgi（30%）、Lysosome（52%）、Mitochondria（74%）、Nuclear（85%）、Peroxisome（11%）、Plasma membrane（70%）。另外若以前兩高分者為預測結果計算正確的準確性結果中，當涵蓋率 74%時最高準確率為 86%，預測九個位置的結果：cytosol（83%）、ER（82%）、Secreted（93%）、Golgi（64%）、Lysosome（71%）、Mitochondria（71%）、Nuclear（97%）、Peroxisome（64%）、Plasma membrane（91%）。

與前人預測方法比較下，當資料集的預測涵蓋率為 74%時，我們的方法的準確率可提升 2%。而對於 26%資料在原方法無法預測的情況下，依然可預測且準確性可到 45%，應可視為有用的輔助資料。

本論文使用序列比對軟體 ALIGN[13]針對個別蛋白質序列取得與對應的學習資料中可獲得的最大同源序列相似度值，並與所預測的子細胞位置計算所得機率值分析預測的結果，探討不同子細胞位置的預測信賴度與學習資料的關係。

如圖 5 中藍色的長條圖表示不同預測機率值區間的預測準確率，如預測的機率值介於 0.9~1 之區間的所有蛋白質序列，其預測準確率為 92.4%，機率值介於 0.8~0.9 之區間預測準確率為 91.3%，機率值介於 0.7~0.8 之區間預測準確率為 79.4%，機率值介於 0.6~0.7 之區間預測準確率為 65.5%，機率值介於 0.5~0.6 之區間預測準確率為 52.6%，機率值介於 0.4~0.5 之區間預測準確率為 48.4%，機

率值介於 0.3~0.4 之區間預測準確率為 44%，機率值介於 0.2~0.3 之區間預測準確率為 32.2%，預測的機率值介於 0.1~0.2 之區間並無序列。隨著預測機率值越低，預測準確性也越低。

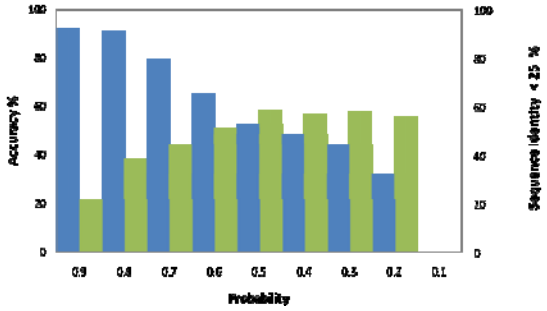

圖 5. 預測準確性與序列相似度的關係圖。X 軸為機械學習預測九個位置的機率值區間。藍色的長條圖表示該區間的預測準確率，綠色長條圖表示在區間中序列相似度小於 25%的序列在區間裡所有的序列的比例。

我們同時針對資料集中序列與序列間同源的屬性討論是否可能影響預測蛋白質在細胞內的位置的結果。根據序列比對軟體 ALIGN 計算序列與序列之間的相似度，從中選出序列相似度最高的序列做為序列與序列間同源性的依據。若序列相似度小於 25%，視為兩序列間無同源關係，我們計算在區間中序列相似度小於 25%的序列在區間裡所有的序列的比例，如圖 5 綠色的長條圖所示。當機率值在 0.5 左右會有交點，表示同源序列在同區間的資料中占 50%時，預測準確率達 66%，機率值在 0.6 時，當同源序列在同區間的資料中占 60%時，預測準確率達 50%以上。當資料集中同源序列比例越高，預測的準確性也越高，整體來看同源序列比例約占 40%，但機率值小於 0.5 時，預測準確率則急遽下降，當機率值大於 0.5、同源序列比例占 50%以上，預測的準確率可達到 60%以上的準確性。

另一方面，我們根據不同子細胞位置詳細探討預測準確性與序列相似度的關係。由圖 6，大部分支持向量機器預測的機率值越高，同源的序列數量也越多。在 ER 子細胞位置發現即使同源序列很少但預測的準確性還是可以維持在 50%以上，在 Nuclear 同源序列很多預測準確性很高。從趨勢線的交叉口的大小，Secreted 子細胞位置即使同源序列很多，沒辦法可以有很高的準確性，這個現象可能是因為這個位置的蛋白質性質呈多樣性，在 Peroxisome 子細胞位置發現現象完全相反。我們發現在不同的位置即使有同樣的條件下，會有不同的預測結果。

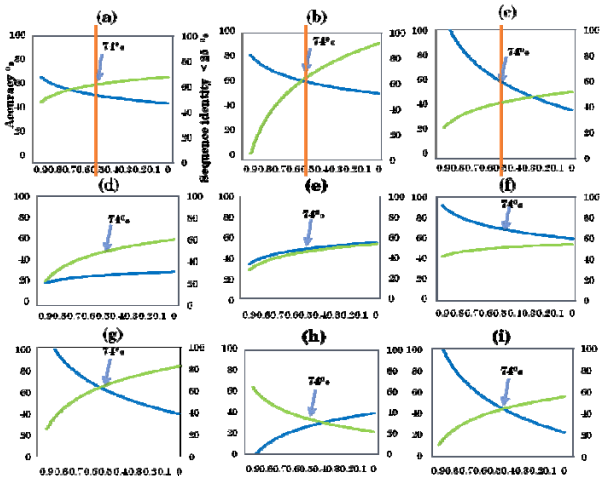

圖 6. 說明預測準確性與序列相似度的關係。(a)Cytosol 、(b)ER 、(c)Secreted 、(d)Golgi 、(e)Lysosome 、(f)Mitochondria 、(g)Nuclear 、(h)Peroxisome、(i)Plasmamembrane。藍色線代表預測準確性趨勢線，綠色線代表序列相似度小於 25%的序列在區間裡所有的序列的比例的趨勢線。

參考文獻

[1] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence," *J Mol Biol*, vol. 300, pp. 1005-16, Jul 21 2000.

[2] 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.

[3] 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.

[4] M. S. Scott, D. Y. Thomas, and M. T. Hallett, "Predicting subcellular localization via protein motif co-occurrence," *Genome Res*, vol. 14, pp. 1957-66, Oct 2004.

[5] O. Emanuelsson, H. Nielsen, and G. von Heijne, "ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites," *Protein Sci*, vol. 8, pp. 978-84, May 1999.

[6] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne, "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites," *Protein Eng*, vol. 10, pp. 1-6, Jan 1997.

[7] M. Reczko and A. Hatzigerrorgiou, "Prediction of the subcellular localization of eukaryotic proteins using sequence signals and composition," *Proteomics*, vol. 4, pp. 1591-6, Jun 2004.

[8] A. Hoglund, P. Donnes, T. Blum, H. W. Adolph, and O. Kohlbacher, "MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition," *Bioinformatics*, vol. 22, pp. 1158-65, May 15 2006.

[9] K. Nakai and M. Kanehisa, "A knowledge base for predicting protein localization sites in eukaryotic cells," *Genomics*, vol. 14, pp. 897-911, Dec 1992.

[10] K. J. Park and M. Kanehisa, "Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs," *Bioinformatics*, vol. 19, pp. 1656-63, Sep 1 2003.

[11] R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D.

Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. Sigrist, and E. M. Zdobnov, "The InterPro database, an integrated documentation resource for protein families, domains and functional sites," *Nucleic Acids Res*, vol. 29, pp. 37–40, 2001.

[12] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines.," 2001, p. Software available from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[13] E. W. Myers and W. Miller, "Optimal alignments in linear space," *Comput Appl Biosci*, vol. 4, pp. 11-7, Mar 1988.