以生物資訊最佳化方法發展電腦輔助疫苗設計系統

Developing Computer-aided Vaccine Design Systems Using Bio-inspired Optimization Methods

計畫編號：NSC 96-2628-E-009-141-MY3
執行期限：96 年 08 月 01 日至 97 年 07 月 31 日
主持人　：何信瑩　國立交通大學生物科技學系（所）
syho@mail.nctu.edu.tw

一、中文摘要

免疫資訊學中的一個重要的目標即是發展快速有效的演算法來預測胜肽致免疫性並建立致免疫性路徑模型。過去研究包含預測蛋白質體對受質蛋白之切點、與抗原轉送蛋白結合之胜肽及主要組織相容性複合物結合之抗原決定部位。本計畫擬建立一個更完整的預測平台來幫助設計疫苗，而首先提出擴大研究兩個重要問題:預測與泛素結合之殘基來篩選容易被泛素所標定之蛋白質與預測致免疫性來進一步篩選能引起免疫反應之胜肽。與泛素結合之蛋白質能被蛋白質體所分解，實為進入致免疫性路徑之重要步驟，因此本計畫探討其可供分類之重要特徵與最佳分類器，同時提出一套繼承式基因演算法來自動依據實驗資料篩選重要物化特性。同時對於預測胜肽之致免疫性，本計畫發現使用繼承式基因演算法篩選之重要物化特性能提供更準確之預測辨識率。而相關之網路預測系統也被實作供生物學家使用。

關鍵詞：預測胜肽致免疫性、預測與泛素結合之殘基、直交實驗設計、繼承式基因演算法、支持向量機器

Abstract

To develop an efficient and effective algorithm to predict peptide immunogenicity and model immunogenic pathway is one goal of immunoinformatics. In the past, related studies focus on prediction of cleavage site of proteasome, peptides binding to transporter associated with antigen processing (TAP) and peptides binding to major histocompatibility complex (MHC). In order to develop a comprehansive prediction system to acclerate the process of vaccines design, this project firstly proposed and studied two important problems of prediction of ubiquitylation sites and prediction of immunogenicity of MHC class I binding peptides. The ubiquitylated protein will be degraded by proteasome. It is an important step for protein to enter the immunogenic pathway and induce immune response. This project assess different classifiers and features to classify ubiquitylation sites and proposed an efficient inheritable genetic algorithm to aotumatically mine informative physicochemical properties according to experimental data. For predicting immunogenicity of MHC class I binding peptides, the informative physicochemical properties mined by inheritable genetic algorithm performs well, compared with other methods. The corresponding web servers are free available for experimental biologist.

Keywords: Prediction of peptide immunogenicity; Prediction of ubiquitylation sites; Orthogonal experimental design; Inheritable genetic algorithm; Support vector machine;

二、研究目的

本計畫主要目的是發展一套輔助免疫學家設計疫苗之電腦系統。本計畫之核心為發展數個能適用於探勘各種生物免疫現象之重要特徵因子的高性能大量參數最佳化之生物資訊演算法，並結合生物免疫知識庫發展成一整合型重要特徵因子探勘、分析、註記的應用系統。

首先，本計劃將針對物理化學特性做探討，結合支持向量機與本實驗室發展之繼承式雙目標基因演算法來建構繼承式物理化學特徵探勘系統。並利用此系統來研究細胞毒性 T 細胞相關致免疫性之重要特徵。最後將依據探勘之結果設計網路伺服器供使用者查詢。

對於設計一套電腦輔助疫苗設計系統而言，正確預測細胞毒性 T 細胞相關之致免疫性將有極大的助益。而在過去，學者一般認為胜肽之致免疫性與胜肽及 MHC class I 結合之親和力有關。然而有幾篇研究卻指出胜肽與 MHC class I 結合之親和力和致免疫性並無明顯相關性。因此本計畫擬大規模探討細胞毒性 T 細胞相關致免疫性與胜肽及 MHC class I 親和力之關係。並希望藉由已知實驗資料來篩遠對於致免疫性相關之物理化學特徵。

同時，正確預測與泛素結合之蛋白質將有助於辨識容易進入致免疫性路徑之蛋白來協助疫苗設計。有鑑於此，本期中報告首先針對以下兩個具有傑出成果的問題為例，簡單說明進行最佳化方法設計的流程：

(1) 預測與 MHC class I 結合之胜肽的致免疫性

對於設計一套電腦輔助疫苗設計系統，正確預測與 MHC class I 結合之胜肽的致免疫性將有極大的助益。在過去，學者一般認為胜肽之致免疫性與胜肽及 MHC class I 結合之親和力有關。然而有幾篇研究卻指出胜肽與 MHC class I 結合之親和力和致免疫性並無明顯相關性。本研究提出以繼承式基因演算法 (Inheritable Bi-objective genetic algorithm, IBCGA)來同時篩選重要物化特性並設計分類器的方法來預測致免疫性。POPI 預測系統比使用一班特徵篩選法更能夠更準確的預測致免疫性，並且比序列排比方法與以結合親和力作預測之方法準確，將可利用於設計疫苗。本研究更進一步證明與 MHC class I 之結合親和力與胜肽致免疫性並無明顯相關性。

(2) 以生物資訊方法辨識 Ubiquitylation 之殘基

Ubiquitylation 是一種蛋白質轉錄後修飾作用，能夠將泛素結合在蛋白質上使其被蛋白質體所分解。其在抗原處理路徑中扮演重要角色。因此，準確預測 Ubiquitylation 之殘基將有助於篩選容易進入抗原表現路徑之蛋白質，進而幫助疫苗設計。本研究先針對三種主要可由序列取得之特徵與三種分類器進行分析研究，並提出一套重要特徵探勘系統來加強辨識率。最後將之應用到辨識可能的 Ubiquitylation 殘基，並利用 C5.0 建構之決策樹與精簡的分類規則提供可解讀知識與交互驗證。

三、研究方法

本研究計畫之核心即為發展重要特性探勘系統，並搭配不同之可從蛋白質序列取得之特性來分類及預測蛋白質之致免疫性及辨識 Ubiquitylation 之殘基。

3.1 重要物理化學特性探勘系統

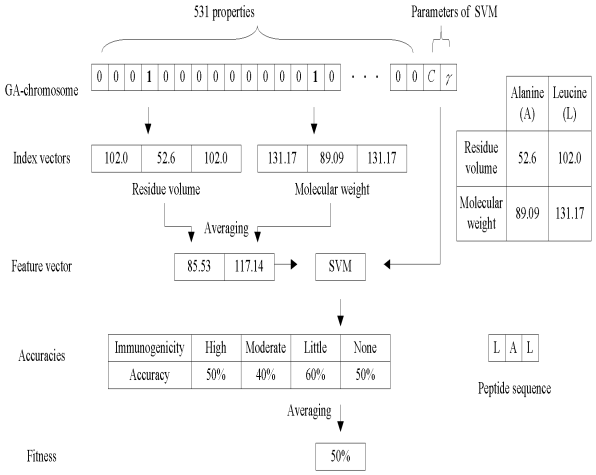
蛋白質之基本物理化學特性決定其功能與特性。因此本研究從 AAindex 資料庫[1]收集了五百三十一個物理化學特性(排除十三個有資料缺損的特性不使用)，每個特性皆包含對應二十個胺基酸的特性數值。由於從 n 個特性中找出少數重要之 m 個特性存在有巨大的搜尋空間 $C(n,m) = n!/(m!(n-m)!)$。即為 0/1 組合最佳化問題。

過去之研究多以專家知識或以簡單的排序篩選法找重要的特徵。然而此類方法會因有限之知識與忽略特性間之交互作用而無法做有效之篩選。本研究即提出以吾人過去發展之繼承式雙目標基因演算法來自動篩選重要特性與調整分類器參數來設計最佳化之支持向量機器(Support Vector Machine, SVM)。

繼承式基因演算法在之前的研究中已經被證明可以有效率的處理 0/1 組合最佳化問題[2]。其利用了以直交實驗設計[3]為

基礎的智慧型交配運算與繼承式演算法來快速找到最佳解。本研究使用之基因演算法染色體編碼及解碼如圖一所式。

繼承式基因演算法已證明能在預測蛋白質細胞核內位置[4]、預測致免疫性[5]與預測 Ubiquitylation 之殘基[6]等問題上獲得良好之效果，相關研究成果將於第四節說明。



圖一、繼承式基因演算法之個體適應性衡量示意圖。圖中說明如何從基因演算法之染色體作解碼並套用支持向量機器來衡量其適應性。

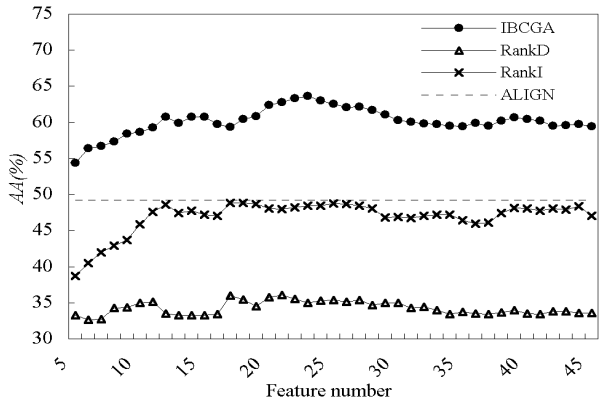
四、結果與討論

以下針對預測與 MHC class I 結合之胜肽之致免疫性與辨識 Ubiquitylation 殘基之結果分別描述之。

4.1 預測與 MHC class I 結合之胜肽的致免疫性

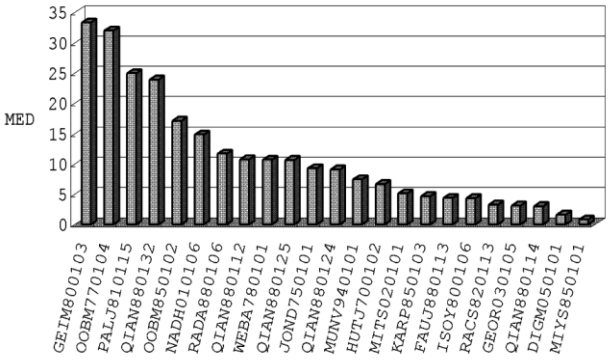
為了研究與 MHC class I 結合之胜肽的致免疫性，本研究從 MHCPEP 資料庫[7]中取出與人類 MHC class I 結合之胜肽序列與致免疫性資料建立了一個資料集 PEPMHCI。

由於實驗資料之致免疫性分為四個等級(無、小、中、大)，並且資料量不平均，因此本研究採用四類辨識率當作目標函數。由圖二實驗結果顯示繼承式基因演算法能有效地根據實驗資料找出重要之物化特性。不論使用多少個重要物化特性皆比使用預設支持向量機參數之排序篩選法 RankD 及使用 IBCGA 篩選出最佳參數之排序篩選法 RankI 選出之物化特性的辨識率高出許多。



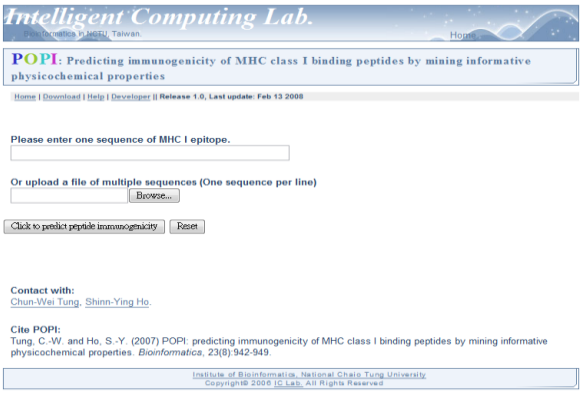
圖二、IBCGA 與其他方法篩選結果之比較。

由於篩選出之重要特性之重要性排名將能提供重要資訊給生物學家，因此吾人採用了直交實驗設計中用來評估個別特徵的主效果指標(main effect difference, MED)來分析篩選出來的物化特性。圖三即為以主效果排名後的結果，主效果越高的特性其對於系統的貢獻越高。為提供免疫學家方便設計疫苗之工具，吾人並實作 POPI 網路伺服器供免費使用(圖四)。

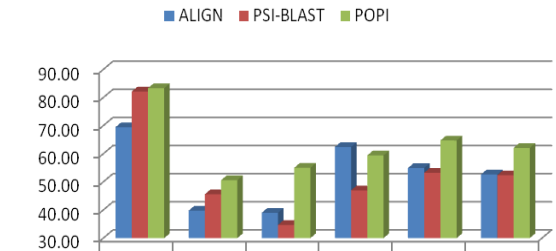


圖三、 以主效果排序篩選出之 23 個重要物化特性。

由於序列相似性通常能提供蛋白質功能之預測，常用之序列比對方法 ALIGN 及 PSI-BLAST 之預測辨識率被實作來與 POPI 做比較。由圖五可知 POPI 提供比序列比對方法準確約 10%之預測辨識率(64.72%)。此外，過去研究經常將胜肽與 MHC 之結合親和力作為其致免疫性之指標，因此吾人實作 AFFIPRE 方法來分析其預測效果。



圖四、 POPI 網路預測系統截圖。



	None	Little	Moderate	High	OA	AA
ALIGN	69.44	39.76	39.00	62.38	54.91	52.64
PSI-BLAST	82.14	45.59	34.67	46.99	53.23	52.35
POPI	83.33	50.60	55.00	59.41	64.72	62.09

圖五、 POPI 與序對比對方法之比較。

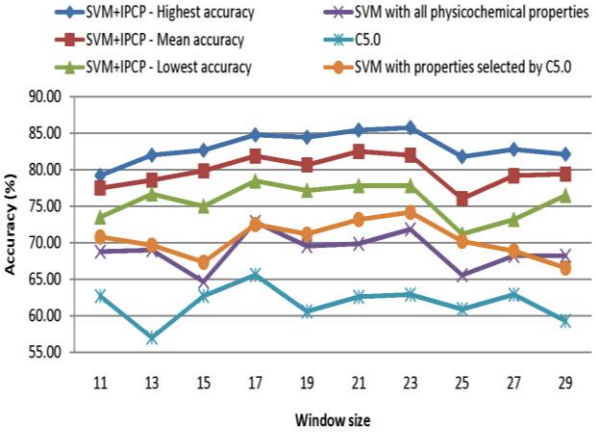
表一顯示 POPI 提供比 AFFIPRE 準確約 20%之辨識率(60.63%)。此結果與過去之研究相符[8, 9]。

表一、 POPI 與 AFFIPRE 方法之比較。

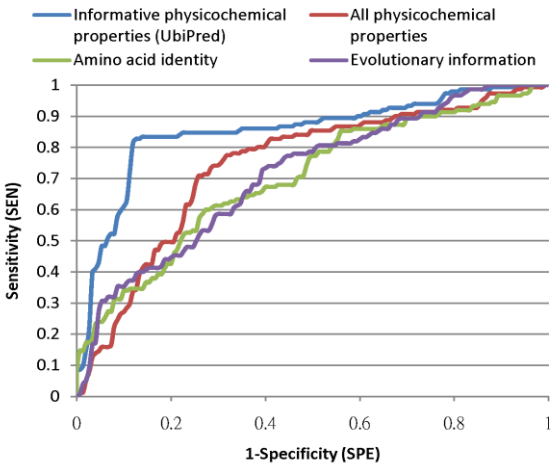
Immunogenicity class	AFFIPRE		POPI	
	ACC (%)	*MCC*	*ACC (%)*	*MCC*
None & Little	35.63	0.17	80.46	0.39
Moderate	32.26	0.01	25.81	0.23
High	52.38	0.15	45.24	0.27
OA	39.38		60.63	
AA	40.09		50.50	

4.2 以生物資訊方法辨識 Ubiquitylation 之殘基

本研究首先從 UbiProt 資料庫[10]中萃取 Ubiquitylation 殘基之資料與其對應之蛋白質序列進而建立一個資料集 UBIDATA。對於預測 Ubiquitylation 殘基，環境資訊將能提供更多資訊，因此本研究採用移動視窗(sliding window)方法來應用環境資訊。



圖六、 繼承式基因演算法篩選之物化特性(IPCP)與 C5.0 之比較。



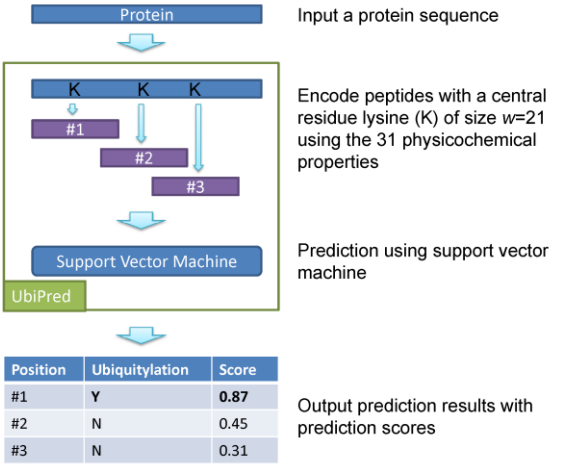
圖七、各種方法之 ROC 圖比較。

在衡量了三種分類器(貝氏分類器、最近鄰居分類器與支持向量機)與三種特徵(二元編碼、演化資訊與全部物化特性)在不同移動視窗大小下對於 Ubiquitylation 殘基之分類效果後，由實驗結果可知，對於三種特徵，使用支持向量機之效果最好。而採用全部物化特性之預測辨識率(84.44%)比採用二元編碼(65.67%)與演化資訊(66.33%)之預測辨識率好。

進一步篩選重要物化特性將能提供

生物學家相關知識並改善辨識率，因此本研究使用決策樹演算法 C5.0 與繼承式基因演算法分別篩選兩組物化特性。由圖六之結果可知，在任何移動式窗大小下，繼承式基因演算法皆有較好的結果。

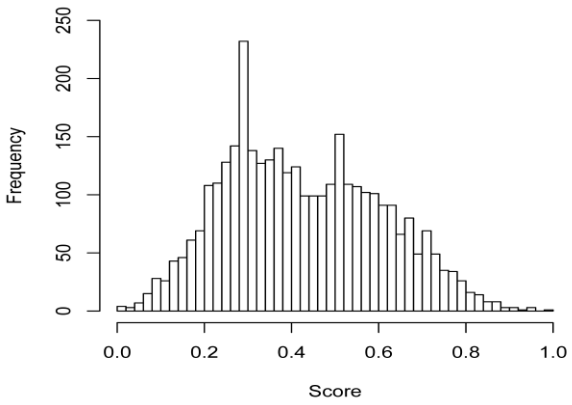
圖七展示了使用支持向量機與四種特徵之 ROC(receiver operating characteristic)圖。由圖中可知使用繼承式基因演算法篩選之重要物化特性能有效預測 Ubiquitylation 殘基。為提供生物學家使用，本研究採用重要物化特性與支持向量機發展了一套 UbiPred 預測系統。其系統流程如圖八所示，並可於 http://iclab.life.nctu.edu.tw/ubipred 使用。



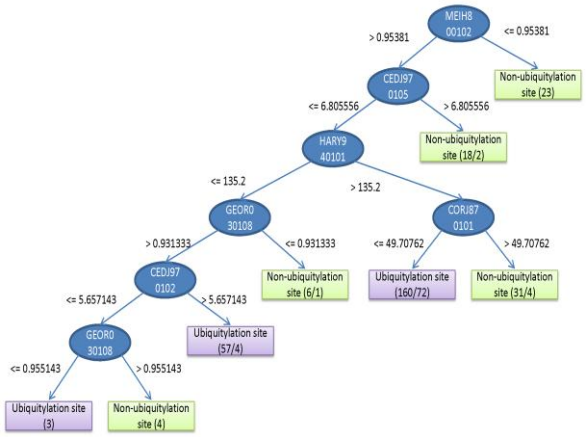
圖八、UbiPred 之系統流程圖。

由於尚有許多 Ubiquitylation 殘基尚未被發現。因此本研究將 UbiPred 應用在從 UbiProt 取出之蛋白質上來辨識可能之 Ubiquitylation 殘基。其結果如圖八所示，分數大於 0.5 之殘基即有可能是 Ubiquitylation 殘基(如圖九所示)。在使用了較嚴格的分數(>0.85)後，一共 23 個有可能的 Ubiquitylation 殘基被 UbiPred 所辨識。

為了提供更直覺之人類可解讀知識並交叉驗證辨識出之 Ubiquitylation 殘基，本研究採用 C5.0 來建置決策樹(圖十)並將其應用來驗證這 23 個殘基。結果發現，此 23 個殘基在套用 C5.0 建置之決策樹後皆被歸類為 Ubiquitylation 殘基。生物學家可進一步實驗分析其 Ubiquitylation 之可能性。



圖九、3424 個受測殘基之分數。



圖十、C5.0 建置之決策樹

五、學術論文成果

在等待本計畫審核期間，對於預測與 MHC class I 結合的胜肽之致免疫性研究成果已有重大突破，並已發表在國際著名的 Bioinformatics 期刊上。相關之論文如下:

C.-W. Tung and **S.-Y. Ho***, "POPI:Predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties," *Bioinformatics*, vol. 23, issue 8, pp. 942-949, 2007. (SCI)

本計畫在第一年之執行期間，成果相當豐碩，其反映在已發表之相關學術論文的質量上。其中預測蛋白質細胞內位置之研究將對於篩選膜蛋白之表面抗原有相當助益。相關之學術論文如下:

W.-L. Huang, C.-W. Tung, S.-W. Ho, S.-F. Hwang and **S.-Y. Ho***, "ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization," *BMC Bioinformatics*, vol. 9, pp. 80, 2008. (SCI)

對於辨識 Ubiquitylation 殘基來幫助預測易進入致免疫性路徑之胜肽之研究成果已投稿 BMC Bioinformatics，並正在修訂中。相關之學術論文如下:

C.-W. Tung and **S.-Y. Ho***, "Computational identification of ubiquitylation sites from protein sequences," *BMC Bioinformatics*, under revision. (SCI)

而預測與 MHC class II 結合的胜肽之致免疫性成果亦發表在國際研討會中。相關之研討會論文如下:

C.-W. Tung and **S.-Y. Ho***, 2007, "Mining physicochemical properties for predicting immunogenicity of MHC class II binding peptides." *18th International Conference on Genome Informatics*. Biopolis, Singapore.

針對大量參數研發最佳化之演算法為本計畫之核心技術。計畫執行期間對於最佳化演算法之研發亦非常卓越，包括多目標模擬退火演算法 (Multi-Objective Simulated Annealing, MOSA)、直交粒子群最佳化演算法 (Orthogonal Particle Swarm Optimization, OPSO)。相關之學術論文如下:

S.-Y. Ho, H.-S. Lin, W.-H. Liauh and S.-J. Ho, "OPSO: Orthogonal Particle Swarm Optimization and Its Application to Task Assignment Problems," *IEEE Trans. Systems, Man, and Cybernetics -Part A, Systems and Humans*, 38 (2), pp. 288-298, MAR 2008. (SCI, EI)

M.-H. Hung, S.-J. Ho, L.-S. Shu, S.-F. Hwang, and **S.-Y. Ho***, "A Novel Multiobjective Simulated Annealing Algorithm for Designing Robust PID Controllers," *IEEE Trans. Systems, Man, and Cybernetics-Part A, Systems and Humans*, 38 (2), pp. 319-330, MAR 2008. (SCI, EI)

而智慧型基因演算法 (Intelligent Genetic Algorithm, IGA)應用在重建基因調控網路亦有顯著成果。相關之研討會論文如下:

Y.-H. Chen and **S.-Y. Ho***, 2007, "GRNet: An efficient and robust evolutionary method for reconstructing gene regulatory networks," *18th International Conference on Genome Informatics*, Biopolis, Singapore.

六、計畫成果自評

吾人所提之研究計畫在第一執行期間，研究進度順利，研究成果亦已投稿相關學術論文，已達原計畫研究目標，成果豐碩。本報告提及之 POPI 及 UbiPred 預測系統可於下列網址中免費使用：http://iclab.life.nctu.edu.tw/POPI 及 http://iclab.life.nct.edu.tw/ubipred。

參考文獻

[1] S. Kawashima, and M. Kanehisa, "AAindex: amino acid index database," *Nucleic Acids Res*, vol. 28, no. 1, pp. 374, Jan 1, 2000.

[2] S. Y. Ho, J. H. Chen, and M. H. Huang, "Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications," *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics*, vol. 34, no. 1, pp. 609-620, Feb, 2004.

[3] A. Dey, Orthogonal *fractional* factorial *designs*, New York: Wiley, 1985.

[4] W. L. Huang, C. W. Tung, H. L. Huang *et al.*, "ProLoc: Prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features," *Biosystems*, vol. 90, no. 2, pp. 573-81, Sep-Oct, 2007.

[5] C. W. Tung, and S. Y. Ho, "POPI:

predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties," *Bioinformatics*, vol. 23, no. 8, pp. 942-9, Apr 15, 2007.

[6] C. W. Tung, and S. Y. Ho, "Computational identification of ubiquitylation sites from protein sequences," *BMC Bioinformatics*, under revision, 2008.

[7] V. Brusic, G. Rudy, and L. C. Harrison, "MHCPEP, a database of MHC-binding peptides: update 1997," *Nucleic Acids Res*, vol. 26, no. 1, pp. 368-71, Jan 1, 1998.

[8] A. Sette, A. Vitiello, B. Reherman *et al.*, "The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes," *J Immunol*, vol. 153, no. 12, pp. 5586-92, Dec 15, 1994.

[9] J. Ochoa-Garay, D. M. McKinney, H. H. Kochounian *et al.*, "The ability of peptides to induce cytotoxic T cells in vitro does not strongly correlate with their affinity for the H-2Ld molecule: implications for vaccine design and immunotherapy," *Mol Immunol*, vol. 34, no. 3, pp. 273-81, Feb, 1997.

[10] A. L. Chernorudskiy, A. Garcia, E. V. Eremin *et al.*, "UbiProt: a database of ubiquitylated proteins," *BMC Bioinformatics*, vol. 8, pp. 126, 2007.