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：97 年 08 月 01 日至 98 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

疫苗是近代醫學上的重要發明，過去人們對於傳染病，往往都無法醫治，但自從疫苗被發明之後，將許多傳染病亦都逐漸獲得控制。近年來，疫苗已經成為一種最符合經濟效益的疾病預防方法，疫苗的研發設計也廣泛受到各界重視。因此，藉由生物資訊的分析、使用最佳化方法發展電腦輔助疫苗設計系統對於生物學家設計疫苗上有很大的助益。在第一年執行期間，探討其可供分類之重要特徵與最佳分類器，並設計一套繼承式基因演算法，自動依據實驗資料篩選重要物理化學特性，研究細胞毒性T細胞及MHC class I抗原處理及表現路徑之相關致免疫性之重要特徵。因此，本計畫在第二年執行期間，預期藉由過去發展之最佳化方法來針對病毒感染機制及輔助性 T 細胞免疫反應之重要特徵做研究。實驗結果證明分析延伸性的物理化學特性用於預測 HIV-1 協同受器以及針對輔助性 T 細胞相關致免疫性路徑研究有良好的效果，初步成果已有論文發表。

關鍵詞：人類免疫缺陷病毒、協同受器、輔助性T細胞、主要組織相容性複合體第二型分子、物理化學特性、繼承式基因演算法、支持向量機器。

Abstract

The development of vaccines has been the most important achievement in preventive medicine. In the past years epidemic diseases were incurable to human beings. However, since the development of vaccines, many infectious diseases are gradually controlled. In recent years, the vaccine has become one of the most cost-effective method of disease prevention, and how to design a vaccine is widespread subject to public attention. For this reason, the design of vaccines for the biologists have a lot of help by the analysis of bioinformatics and develop computer-aided vaccine design systems using bio-inspired optimization methods. In the first year, this project assess different classifiers and features to solve classification problems and proposed an efficient inheritable genetic algorithm to aotumatically mine informative physicochemical properties according to experimental data for studying immunogenic pathway of cytotoxic T cell and predicting immunogenicity of MHC class I binding peptides. In the second year, we aims to design efficient methods to solve the classification problems of HIV-1 coreceptor usage and the prediction of immunogenicity induced by MHC class II binding peptides based on the achievements of the first-year project. The results show that the utilization and analysis of the mined informative physicochemical properties are effective in predicting HIV-1 coreceptor usage and immunogenicity. Relatived achievements were published in conference and journal papers.

Keywords: HIV-1, Coreceptor, Helper

T cell, Major histocompatibility complex class II molecules, Physicochemical properties, Inheritable genetic algorithm, Support vector machine.

二、研究目的

本計畫藉由發展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整合重要特徵因子探勘、分析、註記之應用系統來建構電腦補助疫苗設計系統。

在第一年的計畫中，對於物理化學特性做探討，並研究細胞毒性 T 細胞相關致免疫性之重要特徵，在此部分已有傑出的成果。因此，在第二年計劃中，將延伸性的分析物理化學特性用於預測 HIV-1 協同受器以及針對輔助性 T 細胞相關致免疫性路徑做進一步研究。

正確的預測HIV-1 協同受器及輔助性T細胞相關之致免疫性對於電腦輔助疫苗設計系統有極大的重要性。因此，本期中報告首先針對以下兩個研究方向，簡單說明研究目的及最佳化方法設計流程：

(1) 預測 HIV-1 協同受器

人類免疫缺陷病毒(Human Human Immunodeficiency Virus, HIV)為一種反轉錄病毒，會入侵宿主的免疫細胞使其失去作用，並整合入宿主細胞的基因組當中。近年來研究 HIV-1 治療及抑制病情的方法，在於控制 HIV-1 病毒的傳染途徑，進而幫助抑制病情。在有些訊號傳遞的機制中，除了主要受器之外，還會有協同受器來幫助接收訊息，並和主要受器一起完成整個機制。HIV-1 病毒的傳染途徑中，免疫細胞上的受器和協同受器都是不可或缺的，目前的研究，已有藥物可以阻隔 HIV-1 病毒和協同受器的連結，進而阻止病毒進入到免疫細胞中。因此，準確的預測出 HIV-1 病毒在進入免疫細胞時所使用的協同受器，對於治療 HIV-1 的病毒感染、發展疫苗設計上有很大的幫助。

在 HIV-1 的感染過程中，主要是使用 CCR5 (R5) 和 CXCR4 (X4)這兩種協同受器。病毒可分為使用單一種協同受器的 R5 和 X4，或兩種協同受器都可利用的 R5X4 三種類型。本研究目的，設計一個有效使用重要的物理化學特性的方法來預測三個類別的 HIV-1 協同受器。

(2) 預測 MHC class II 抗原決定位置

輔助性T細胞帶CD4 的細胞標記，當輔助性 T 細胞與抗原呈現細胞（antigen presenting cell, APC）上的MHC class II分子結合後會去辨認抗原，辨認出其專一性抗原後會受刺激而釋出白血球間激活素-1（interleukin-1，或稱介白素）的物質，這種物質會刺激輔助性T細胞釋出介白素-2，介白素-2 會刺激專一性的細胞毒性T細胞繁殖，且產生B細胞生長素的蛋白，這種蛋白會和介白素-1 共同作用使的B細胞大量繁殖產生專一性抗體。對於電腦輔助疫苗設計系統，準確的預測MHC class II epitopes將有極大的助益。

三、研究方法

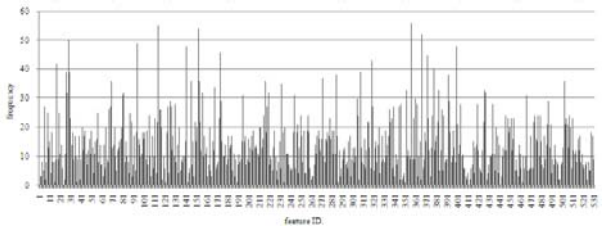
本研究計畫之核心為發展重要特徵挑選，從蛋白質序列中挑選重要特性來做分類、預測以及最佳化系統設計、估計蛋白質序列各位置的結合強度。本研究提出的方法 SVM-PCP 可以達到極高的 HIV-1 病毒在進入免疫細胞時所使用的協同受器的預測辨識率。

另外本研究設計使用智慧型基因演算法（intelligent genetic algorithm, IGA）加入一些外生性蛋白質片段的資訊來設計一個最佳化的特殊位置得分矩陣（position weight matrixes, PWMs）來預測 MHC class II 抗原決定位置。

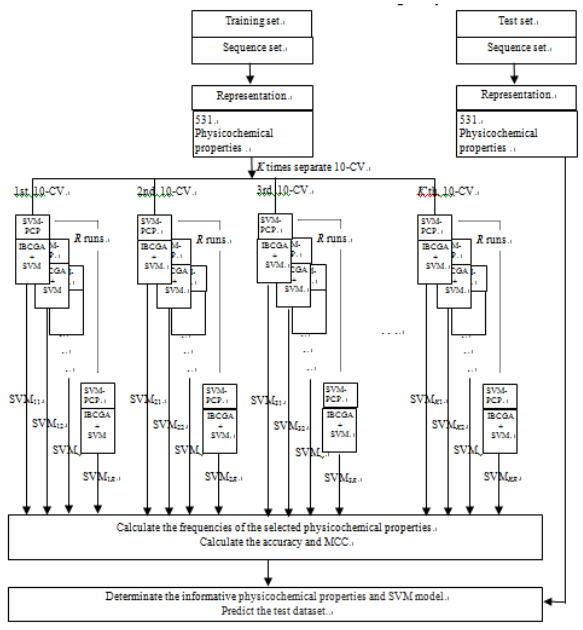
3.1 計算重要物理化學特性頻率預測系統 (SVM-PCP)

蛋白質之物理化學特性在生化反應上是最直覺的一個特徵且被廣泛的應用在生物資訊領域中。因此本研究採用 AAindex 資料庫[1]收集的五百三十一個物理化學特性，

使用吾人過去發展之繼承式基因演算法(IBCGA)[2]挑選出重要的物理化學特性及調整分類器參數來設計支持向量機器(Support Vector Machine, SVM)。其中將蛋白質序列資料分成 K=10 組獨立的資料集各別執行 R=20 次繼承式基因演算法，繼承式基因演算法在之前的研究中已經證明可有效處理最佳化問題，將 K*R=200 組挑選的物理化學特性計算出現的頻率分數後，以最高分的那組做為支持向量機器模型，此方法可以有效預測 HIV-1 協同受器。



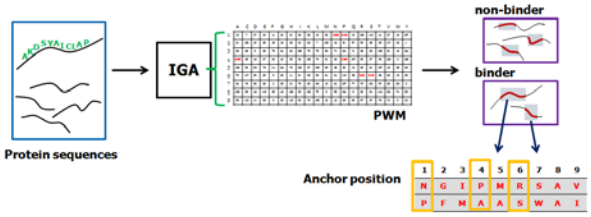
圖一、在兩百次實驗中，五百三十一種物理化學特性出現的頻率圖。



圖二、SVM-PCP 方法系統流程圖。SVM-PCP 會自動決定一組有資訊的物理化學特性以及支持向量機器模型，提供做為測試的資料使用。

3.2 最佳化的特殊位置得分矩陣預測系統

使用智慧型基因演算法(IGA)[3]設計一個最佳化的得分矩陣來計算每個位置的各個胺基酸結合強度，抗原決定位置的長度 M=9 且一共有二十種胺基酸，PWM 為 9*20 的矩陣，一共有 180 個參數做最佳化調整。在基因演算法編碼過程中，加入一些 SYFPEITHI 資料庫[4]所提供的外生性蛋白質片段資訊，統計結合胜肽之胺基酸出現頻率，來幫助找出的矩陣，不只具有準確的預測能力，還具備生物意義。



圖三、預測 MHC class II epitopes 系統流程圖。預測系統會輸出 binding peptides 和 binding sites。

四、結果與討論

以下針對預測 HIV-1 協同受器及與 MHC class II 分子結合之胜肽 epitope 位置的研究結果做個別描述。

4.1 預測 HIV-1 協同受器

只要含有 CD4 受體的細胞，都有可能成為 HIV 病毒的目標細胞。當 HIV 病毒要進入人體 T 細胞時，是由 HIV 病毒上的 gp120 膜蛋白的 V3 環狀區域和 T 細胞表面的 CD4 受體及協同受體作連結，使膜蛋白結構改變，進而讓病毒與 T 細胞作融合。V3 環狀區域長度大約三十五個胺基酸。由於 HIV 病毒突變的非常快，使得 V3 環狀區域的組成有很大的變化，造成預測上的困難。本研究包含下列三個方向做探討:

(1) 求出一組使得支持向量機器有效辨識的重要物理化學特性特徵集來預測 HIV-1 協同受器。表一顯示本研究提出之 SVM-PCP 有很好的預測結果。

表一、在固定 specificity 為 92.5%下，各種方法 sensitivity 和 AUC 的比較[5]。

Method	Sensitivity (%)	AUC
11/25rule	59.5	*
PSSM	71.9	0.90
[a]SVMbinary	76.4	0.91
SVM-PCP	88.9	0.94

(2) 建立一組新的資料集，將原有的 159 條序列擴增成 1225 條序列[6]，因此有較多的資料來做訓練產生支持向量機器模型。表二與表三分別顯示兩種不同數量的序列資料在訓練資料集與測試資料集中各種比較指標的數據結果。表四顯示綜合三種不同數量的序列資料比較彼此之間的辨識率，本研究提出的方法 SVM-PCP 可以達到極高的辨識率。

表二、比較Set157和Set139在訓練資料集中各種比較指標的統計結果。

	Statistics	Set157 [14] (best)	Set139 (best)	Set139(average)
R5	Accuracy (%)	75.00	99.00	98.33 ± 1.53
	MCC	0.58	0.84	0.82 ± 0.03
X4	Accuracy (%)	79.31	81.50	81.05 ± 4.56
	MCC	0.66	0.85	0.83 ± 0.04
R5X4	Accuracy (%)	40.00	72.73	69.16 ± 7.54
	MCC	0.18	0.75	0.72 ± 0.05
Overall Accuracy (%)		67.72	89.15	87.97 ± 1.60
Mean of Accuracies (%)		64.77	84.41	82.85 ± 2.35

表三、比較Set157和Set139在測試資料集中各種比較指標的統計結果。

	Statistics	Set157 (best)	Set139 (best)	Set139(average)
R5	Accuracy (%)	78.57 (11/14) a	91.67 (11/12)	90.21 ± 8.90
	MCC	0.47	0.71	0.64 ± 0.14
X4	Accuracy (%)	70.00 (7/10)	80.00 (8/10)	68.55 ± 15.90
	MCC	0.47	0.76	0.63 ± 0.14
R5X4	Accuracy (%)	50.00 (3/6)	60.00 (3/5)	20.20 ± 17.13
	MCC	0.38	0.61	0.10 ± 0.21
Overall Accuracy (%)		70.00	81.48	69.22 ± 7.48
Mean of Accuracies (%)		66.19	77.22	59.65 ± 7.95

表四、比較Set139、Set1225及Set3fold，三種序列資料的辨識率及統計結果。

	Statistics	Set139 (average)	Set1225 (average)	Set3fold (average)
R5	Accuracy (%)	90.2083 ± 8.8979	97.4382 ± 1.2016	97.4895 ± 1.0069
	Correct number(n)	11 ± 1	175 ± 3	302 ± 3
	Total number(n)	12	178	310
X4	Accuracy (%)	68.5500 ± 15.8953	79.0968 ± 5.5944	73.5502 ± 5.9010
	Correct number(n)	7 ± 3	25 ± 2	41 ± 3
	Total number(n)	10	31	55
R5X4	Accuracy (%)	20.2000 ± 17.1297	37.4133 ± 8.1919	34.4504 ± 5.8295
	Correct number(n)	1 ± 1	9 ± 2	15 ± 3
	Total number(n)	5	25	43
Overall Accuracy (%)		69.2222 ± 7.4759	88.5954 ± 1.3985	87.6185 ± 1.0703
Correct number(n)		19 ± 2	209 ± 4	358 ± 4
Total number(n)		27	234	408

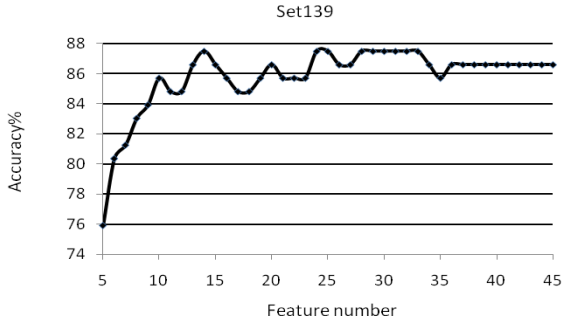
(3) 找出重要意義的物理化學特性有效設計支持向量機器，從具有鑑別度的物理化學特性更進一步去了解 HIV-1 協同受器的結構特性。

圖四、圖五與圖六顯示出在不同資料集中，使用 IBCGA 找出不同的特徵數的辨識率曲線圖，由圖中可以明顯看出辨識率最高的位置在特徵數約十四、十五的地方。因此，表五列出在 Set3fold 資料集中，使用 IBCGA 依序找出具有最高辨識率的十四個重要物理化學特性，並做描述說明。

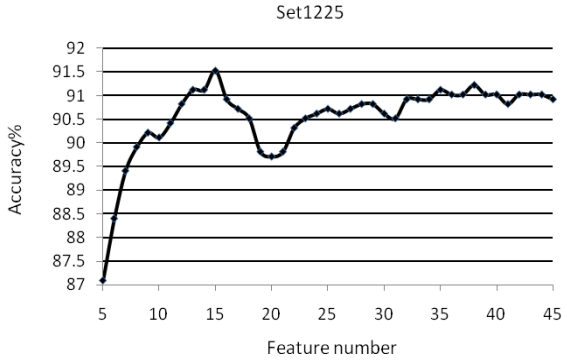
依據文獻對已知物化特性的了解分析發現，本研究找出的前 5 名重要特徵都應證是對 HIV-1 協同受器的結構特性有密切關係。

表五、十四個挑選的物理化學特性列表。

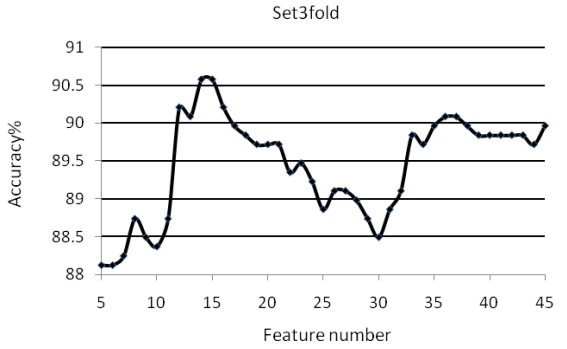
Feature ID.	AAindex identity	Description
30	CHAM830107	A parameter of charge transfer capability (Charton-Charton, 1983)
43	CHOP780206	Normalized frequency of N-terminal non helical region (Chou-Fasman, 1978b)
70	EISD860102	Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)
95	FINA910104	Helix termination parameter at posision j+1 (Finkelstein et al., 1991)
142	KARP850101	Flexibility parameter for no rigid neighbors (Karplus-Schulz, 1985)
205	NAKH920104	AA composition of EXT2 of single-spanning proteins (Nakashima-Nishikawa, 1992)
281	QIAN880124	Weights for beta-sheet at the window position of 4 (Qian-Sejnowski, 1988)
320	RADA880107	Energy transfer from out to in(95%buried) (Radzicka-Wolfenden, 1988)
335	RICJ880114	Relative preference value at C1 (Richardson-Richardson, 1988)
360	SNEP660102	Principal component II (Sneath, 1966)
386	WERD780103	Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga, 1978)
392	WOLS870103	Principal property value z3 (Wold et al., 1987)
475	TSAJ990102	Volumes not including the crystallographic waters using the ProtOr (Tsai et al., 1999)
479	WILM950102	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)



圖四、Set139 資料集不同特徵數的辨識率曲線圖。



圖五、Set1225 資料集不同特徵數的辨識率曲線圖。



圖六、Set3fold 資料集不同特徵數的辨識率曲線圖。

4.2 預測 MHC class II 抗原決定位置

為了研究與 MHC class II 分子結合之胜肽抗原決定位置，本研究使用 IEDB 資料庫[7]中取出與老鼠和人類 MHC class II 分子結合之胜肽序列建立資料集。

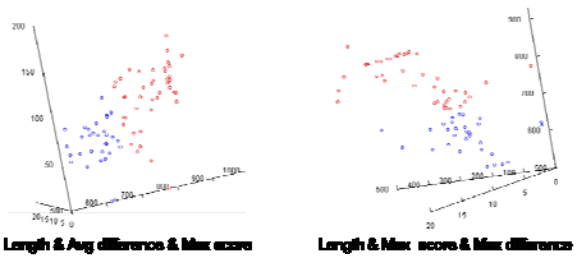
在初步結果中，與 MOEA[8]做比較，使用較小的 population size = 50 可以達到和多

目標方法(population size = 500)有相似的結果。

表六、本研究方法和 MOEA 方法之比較。

Type	Allele	AUC	
		MOEA	t3-All
Mouse	I-Ab	0.919	0.880
HLA	DRB1-0101	0.651	0.663
	DRB1-0404	0.786	0.787
	DRB1-0701	0.735	0.745
	DRB1-1302	0.820	0.832
	DRB4-0101	0.759	0.792
	DRB5-0101	0.660	0.709

在預測 MHC class II 抗原決定位置，一般常使用特殊位置結合強度，找尋胺基酸序列中分數最高的片段做為預測的抗原決定位置。本研究針對一些相關的特徵做分析，發現在三維空間中，新增加的特徵有效區分結合胜肽之類別。圖七顯示在 IAb 資料集中，除了最高分數外，使用胺基酸序列長度及序列中使用移動視窗(sliding window)方法取得每個片段分數之平均差值與最大差值做三維空間的資料分佈圖。



圖七、三維空間的資料分佈圖。

五、學術論文成果

本計畫兩年之執行期間，成果相當豐碩，其反映在已發表之相關學術論文的質量上。針對本計畫之核心技術，大量參數研發最佳化之演算法中，繼承式基因演算法(Inheritable Bi-objective genetic algorithm, IBCGA)之應用有顯著成果。如應用繼承式基因演算法挑選 GO terms、預測蛋白質毒性。

其中對於電腦補助疫苗設計之相關研究已有不錯成果，如預測人類免疫缺陷病毒(HIV)的協同受器 CCR5 (R5)、CXCR4 (X4)和兩種協同受器都可利用的 R5X4、辨識 Ubiquitylation 殘基來幫助預測易進入致免疫性路徑之胜肽、預測蛋白質細胞內位置、預測與 MHC class I 結合的胜肽之致免疫性、預測與 MHC class II 結合的胜肽之致免疫性。

相關之學術期刊論文如下:

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)

C.-W. Tung and **S.-Y. Ho***, "Computational identification of ubiquitylation sites from protein sequences," *BMC Bioinformatics*, 9, 310, 2008.(SCI) (Highly accessed)

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)

W.-L. Huang, C.-W. Tung, H.-L. Huang and **S.-Y. Ho***, "Predicting protein subnuclear localization using GO-amino-acid composition features," *BioSystems*, under minor revision, 2009. (SCI)

相關之學術研討會論文如下:

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

W.-L. Huang, C.-W. Tung, S.-W. Ho and **S.-Y. Ho***, "ProLoc-rGO: Using rule-based

knowledge with Gene Ontology terms for prediction of protein subnuclear localization," *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. (CIBCB).* Sun Valley, Idaho, USA, 201-206, 2008.

C.-T. Tsai, W.-L. Huang, S.-J. Ho, L.-S. Shu and **S.-Y. Ho***, " Virulent-GO: Prediction of virulent proteins inbacterial pathogens utilizing Gene Ontology terms, " *International Conference on Bioinformatics and Bioengineering (ICBB)*, May 27-29, Tokyo, Japan, 2009.

K.-T. Hsu, H.-L. Huang, C.-W. Tung, Y.-H. Chen, and **S.-Y. Ho***, " Analysis of physicochemical properties on prediction of R5, X4 and R5X4 HIV-1 coreceptor usage," *International Conference on Bioinformatics and Bioengineering (ICBB)*, May 27-29, Tokyo, Japan, 2009.

六、計畫成果自評

吾人所提之研究計畫在第二年執行期間，研究進度順利，研究成果亦已投稿相關學術論文，尚有部分研究成果正在進行整理中，已達預期目標。

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 bi-objective 0/1 combinatorial optimization problems and its applications," *IEEE Trans. Syst. Man Cybern. Part B-Cybern.*, vol. 34, pp. 609-620, 2004a.

[3] S.-Y. Ho, and L.-S. Shu, "Intelligent evolutionary algorithms for large parameter optimization problems," *IEEE Transactions on Evolutionary Computation* 8(6): 522-541, 2004.

[4] H.-G. Rammensee, J. Bachmann, N.-P.-N. Emmerich, O.-A. Bachor and S. Stevanović, "SYFPEITHI: database for MHC ligands and peptide motifs," *Immunogenetics*, 50: 213–219, 1999.

[5] T. Sing, A.-J Low, N. Beerenwinkel, O. Sander, P. Cheung, F.-S Domingues, J. Büch, M. Dämer, R. Kaiser, T. Lengauer and P.-R. Harrigan "Predicting HIV coreceptor usage on the basis of genetic and clinical covariates," *Antiviral therapy*, vol. 12, pp. 1097 – 1106, 2007.

[6] Los Alamos National Laboratory HIV Sequence Database, http://www.hiv.lanl.gov/.

[7] B. Peters, J. Sidney, P. Bourne, H.-H. Bui, S. Buus, G. Doh, W. Fleri, M. Kronenberg, R. Kubo, O. Lund, D. Nemazee, J.-V. Ponomarenko, M. Sathiamurthy, S. Schoenberger, S. Stewart, P. Surko, S. Way, S. Wilson, A. Sette*, "The Immune Epitope Database and Analysis Resource: From Vision to Blueprint," *PLoS Biology*, vol. 3, Issue 3, e91, 2005.

[8] Rajapakse, M., B. Schmidt, et al. (2007). "Predicting peptides binding to MHC class II molecules using multi-objective evolutionary algorithms." *BMC Bioinformatics* 8.