國立交通大學

生物資訊及系統生物研究所

碩士論文

利用 V3 環狀序列的物化特性預測 HIV-1 病毒類型

Prediction of HIV-1 tropisms by using physicochemical properties of V3 loop sequences

研 究 生：許凱迪

指導教授：何信瑩　教授

中 華 民 國 九 十 八 年 七 月

利用 V3 環狀序列的物化特性預測 HIV-1 病毒類型

Prediction of HIV-1 tropisms by using physicochemical properties of V3 loop sequences

研 究 生：許凱迪　　Student：Kai-Ti Hsu

指導教授：何信瑩　　Advisor：Shinn-Ying Ho

國 立 交 通 大 學

生物資訊研究所

碩 士 論 文

A Thesis Submitted to Institute of Bioinformatics and
Systems Biology Department of Biological Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of Master in
Bioinformatics

July 2009

Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 八 年 七 月

利用 V3 環狀序列的物化特性預測 HIV-1 病毒類型

學生：許凱迪　　指導教授：何信瑩

國立交通大學生物資訊研究所碩士班

摘　　要

預測人類免疫缺失病毒(Human Immunodeficiency Virus, HIV)在進入人體的細胞時，是利用哪一種的協同受體(coreceptor)，是近年研究人類免疫缺失病毒科學家的目標。因為協同受體是人類免疫缺失病毒在進入細胞時，所必須要有的輔助受體，如果能準確的預測協同受體的使用種類，則可以利用藥物有效的阻隔人類免疫缺失病毒進入到細胞內的途徑，並藉此減緩病情的擴散。

根據人類缺失免疫病毒所連結協同受體的不同，可將病毒分成三類，分別是(1)只利用 CCR5 協同受體進入目標細胞的 R5 型病毒、(2)只利用 CXCR4 協同受體進入目標細胞的 X4 型病毒以及(3)可同時利用 CCR5 和 CXCR4 兩種協同受體，進入目標細胞的 R5X4 型病毒。

在之前的研究中指出， gp120 蛋白質上的 V3 環狀序列，是辨別人類缺失免疫病毒連接哪一種協同受體的重要序列。因此，我們的目的是利用分析 V3 環狀序列的物化特性，來預測其所屬的人類免疫缺失病毒是屬於哪一種類型的病毒。並且找出一組可適用於不同資料及分類器的物化特性組，以助於往後相關的預測改良及人類免疫缺失病毒的研究。

這個問題的困難點在於(1)V3 環狀序列的不定長度和(2)V3 環狀序列的多變性。我們為了克服這兩個困難，採用了 AAindex 編碼。AAindex 可以將序列的物化特性，利用序列本身的胺基酸組成來取得平均值，由於是考慮整條序列的組成，因此對於上述兩個問題都能有效的處理，並有效提升了預測的準確度。

此外我們還利用遺傳式雙目標基因演算法來挑選分類器所使用的物化特性組，讓分類器能使用最少的物化特性就能得到最好的預測能力。挑出來的物化特性組，也藉由進一步的分析，來探討在其他資料和分類器中，物化特性組的分類能力。最後我們搜尋了相關的文獻，以找出物化特性所含的生物意義

Prediction of HIV-1 tropisms by using physicochemical properties of V3 loop sequences

Student：Kai-Ti Hsu　　Advisor：Dr. Shinn-Ying Ho

Institute of Bioinformatics and Systems Biology

National Chiao Tung University

ABSTRACT

Bioinformatics methods for predicting the T cell coreceptor usage from the array of membrane protein of HIV-1 are investigated. Because the coreceptor is necessary for entrances of HIV-1, if we confirm the coreceptor usage, the pathway of HIV could be blocked.

According to the usage of coreceptor, HIV-1could be classified to three tropisms, R5, X4 and R5/X4. The V3 loop sequences were used as input data since the critical properties of sequences that can decide the HIV-1 tropisms. In this study, we aim to propose an effective prediction method for dealing with the three-class prediction problem of R5, X4 and R5/X4 HIV-1 tropisms and find out a feature set of informative physicochemical properties which can be utilized at different data and classifier.

The difficulties of problem are different length and highly variable composition of V3 loop sequences. We deal with the difficulties by physicochemical properties of AAindex because it counts the physicochemical properties of whole sequences, not the specific sequences site. Otherwise, the IBCGA was used to select the physicochemical properties and SVM parameters that make the highest accuracy with least number of properties. Finally, we analyzed the characteristic of selected physicochemical properties to confirm our discovery and have new finding.

誌　　謝

首先，我要特別感謝我的指導教授－何信瑩老師，在我研究陷入瓶頸時，能很快的引領我找到問題的所在，讓我能逐步完成研究以及學習如何思考分析研究方向。在何老師身上不僅學到許多作研究的方法和秘訣，也學到了做人處事的態度應變，這些都要感謝老師的不吝指導。另外還要感謝實驗室學長姐在平時能常常和我討論，在研究過程中不至於有太多停頓，並且幫忙處理機器及程式上的問題。也要感謝學弟們可以幫忙一起搜尋相關資料，並且也給了我許多鼓勵。
最後要謝謝我的家人，因為有他們支持及幫忙，使我沒有經濟及家庭的壓力，在求學階段中可以專心於學業上的研究。

目 錄

表目錄

圖目錄

符號說明

γ ：SVM 核心參數

C ：SVM 花費參數

h ：資料分類類型個數

n ：全部的物化特性個數

r ：選擇的物化特性個數

f_t ：評估函數值，為結合了 t 組實驗的評估值

S_{jk} ：主效果值

F_t ：表示 j 參數在 t 個實驗中 k 水準情況下的值

r_{start} ：開始搜索物化特性的個數

r_{end} ：結束搜索物化特性的個數

S_r ：第 r 個物化特性的分數

$F(P_i)$ ：第 i 個物化特性被選擇的次數

m ：物化特性組中所含的特性個數

一、緒論

1.1 研究背景與動機

人類免疫缺失病毒(Human Immunodeficiency Virus, HIV)為一種反轉錄病毒，會入侵宿主的 T 細胞使其失去作用，並整合入宿主細胞的基因組當中。對於人類免疫缺失病毒的治療，以往都注重在控制病毒中的反轉錄酶的作用，但是成效不彰。因此近年來研究治療及抑制病情的方法，在於控制人類免疫缺失病毒的傳染途徑，進而幫助抑制病情。

在病毒的傳染途徑中，免疫細胞上的受體(receptor)和協同受體(coreceptor)都是不可或缺的。在先前的研究中，病毒表面的 gp120 蛋白質上，和有 V3 環狀序列，此序列是判斷連接何種協同受體的主要依據[1]。如果可以藉由 V3 環狀序列，來準確預測出病毒所使用的協同受體，便可以利用藥物來阻止病毒進入細胞，進而抑制人類免疫缺失病毒的細胞感染及擴散，對於治療人類免疫缺失病毒的感染有很大的幫助[2]。

1.2 研究目的

本研究主要有兩個方向，首先是找出新的方法，可以更準確的預測出人類免疫缺失病毒在進入免疫細胞時，所使用的協同受體。我們利用智慧型雙目標演算法和 SVM 分類器為基礎，提出一個新的預測方法 SVM-PCP，並比較之前的研究所使用的方法，進而證明本論文提出的 SVM-PCP 能夠更準確的預測 HIV 病毒所利用的協同受體的種類。另外，我們希望能夠挑選出一組物化特性，並可以是用在不同資料庫及分類器下。並找出相關研究中所提到，有關於那些特徵所含有的生物意義，使其可供其他生物學家利用。

1.3 論文架構

本論文分為六個章節，其結構如下：

第一章、緒論：包含研究動機與背景、研究目的及論文架構。

第二章、HIV-1 機制：簡介人類免疫缺失病毒的感染途徑、所使用的序列、協同受體和相關的生物背景。

第三章、現有方法：介紹 SVM-PCP 所使用的 SVM 分類器和智慧型雙目標演算法，以及之前用來預測人類免疫病毒種類的方法。

第四章、研究方法：定義研究問題，並對所使用的方法及資料庫作介紹和比較。

第五章、研究結果：比較 SVM-PCP 和之前實驗的結果，並評估挑出來的特性組用在其他資料及分類器上的結果，以及討論生物特性的相關意義。

第六章、結論與未來展望。

二、HIV-1 機制

2.1 免疫細胞感染路徑

人類免疫缺失病毒主要以含有 CD4 受體的細胞為目標細胞。當人類免疫缺失病毒要進入目標細胞時，會經由以下步驟：

(1) 人類免疫缺失病毒的外層膜蛋白 Gp120 會先和目標細胞上的 CD4 受體作結合，使得 gp120 中的 V3 環狀序列暴露出來。

(2) V3 環狀序列會利用中央區域和尖端區域兩部份，去和協同受體作連結，並且將 gp120 膜蛋白固定在細胞表面。

(3) Gp120 蛋白固定在細胞表面後，會產生結構變化，讓跨膜糖蛋白 gp41 直接和細胞表面接觸，並和細胞膜融合，產生運輸通道。

(4) 人類免疫缺失病毒的核心 RNA 會藉由 gp41 產生的通道進入目標細胞。此時目標細胞已被感染。

圖 1 為感染路徑的前半部份，為 gp120 和 CD4 連結後，使得 V3 環狀序列露出，並和協同受體連結的過程。

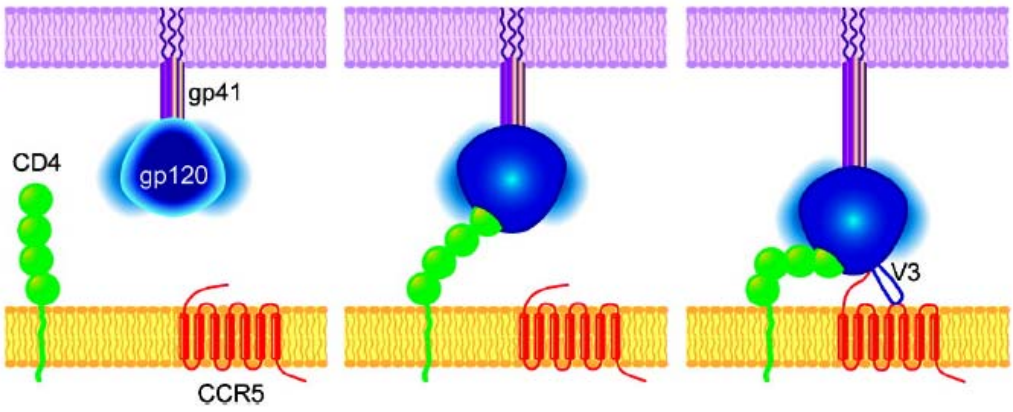

圖 1 gp120 連結目標細胞表面受體過程圖

資料來源：從 Brower 的研究論文中取得[3]

說明：圖 1 為 gp120 中的 V3 環狀區域如何和細胞表面的協同受體作結合的過程圖。其中 gp120 會先和 CD4 結合，之後 V3 環狀區域才會暴露在 gp120 外側，並和協同受體結合。

2.2 GP120 蛋白

Gp120 是人類免疫缺失病毒表面的糖蛋白之一，是位於病毒表面最外層的蛋白質，會和 gp41 共同幫助人類免疫缺失病毒進入目標細胞。gp120 糖蛋白中含有多個多變性環狀序列(variable loop)，而 V3 環狀序列為其中一個多變型環狀序列，同時 V3 環狀序列也是 gp120 蛋白中會直接和協同受體作連結的結構。

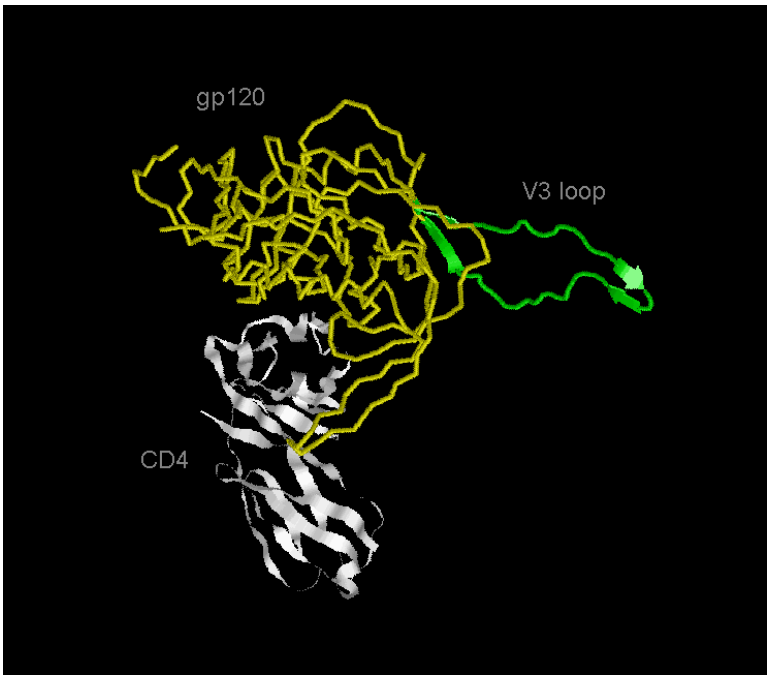

圖 2 V3 環狀序列、gp120 及 CD4 結構圖

資料來源：結構資料是從蛋白質資料庫(protein data base, PDB)中取得，PDB 編號為 2B4C，取得結構資料後利用 Rasmol 軟體繪圖而成。

說明：圖 2 為 gp120 蛋白質、CD4 受體和 V3 環狀序列的結構及位置圖。圖中僅呈現 gp120、CD4 及 V3 環狀序列的骨架結構，並利用不同顏色來區分。圖中綠色部分是露出在蛋白質外的 V3 環狀序列，黃色部分為 gp120 的蛋白質結構，而白色的部分是和 gp120 連結的 CD4 受體。

2.3 V3 環狀序列

V3 環狀序列(V3 loop sequences)為一段大約長 30~38 個胺基酸的環狀結構序列，為人類免疫缺失病毒和協同受體結合的序列。因為人類免疫缺失病毒的快速突變，使得 V3 環狀序列的胺基酸組成會產生較大的變異，而 V3 環狀序列的高變異性，也增加了預測協同受體的難度。在預測的過程中，我們要克服的問題就是不同的 V3 環狀序列中所含有的不同長度和高變異性。

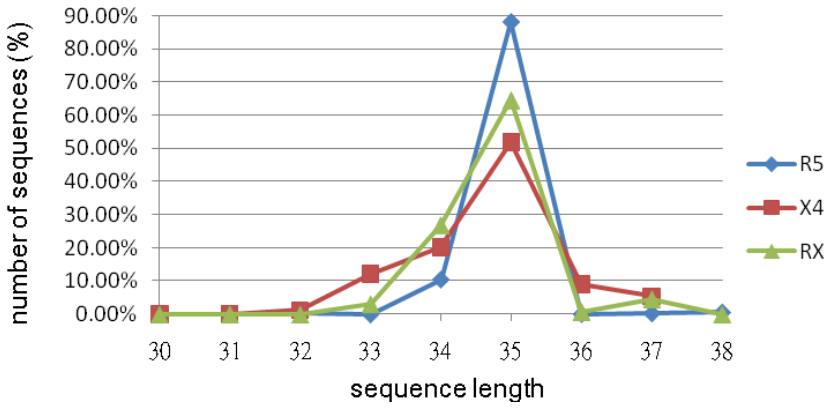

圖 3 V3 環狀序列長度分佈圖

資料來源：由 Data1225 的資料依照序列長度比例作圖而成

說明：圖 3 是用我們從資料庫得到的序列(Data1225)，統計每一類型病毒的序列長度分佈所作的分布圖。雖然序列長度還是以 35 個胺基酸為主，但是有約 20%的序列含有不同個數的胺基酸，並且在 X4 和 R5X4 兩類的病毒序列中，序列長度不為 35 個胺基酸的比例，遠比 R5 型病毒要來的多。下表為 V3 環狀序列長度統計的詳細資料。

表 1 序列長度分佈表

	R5		X4		RX		Total	
長度	個數	(百分比)	個數	(百分比)	個數	(百分比)	個數	(百分比)
30	1	(0.11%)	0	(0.00%)	0	(0.00%)	1	(0.08%)
31	0	(0.00%)	0	(0.00%)	0	(0.00%)	0	(0.00%)
32	2	(0.21%)	2	(1.22%)	0	(0.00%)	4	(0.33%)
33	0	(0.00%)	20	(12.20%)	4	(3.08%)	24	(1.96%)
34	98	(10.53%)	33	(20.12%)	35	(26.92%)	166	(13.55%)
35	820	(88.08%)	85	(51.83%)	84	(64.62%)	989	(80.73%)
36	1	(0.11%)	15	(9.15%)	1	(0.77%)	17	(1.39%)
37	3	(0.32%)	9	(5.49%)	6	(4.62%)	18	(1.47%)
38	6	(0.64%)	0	(0.00%)	0	(0.00%)	6	(0.49%)

資料來源：由 Data1225 的資料依照序列長度比例作表而成

說明：表 1 是 Data1225 資料中 V3 序列長度的個數及百分比分佈表，由此表可以更明顯的看出 X4 和 R5X4 兩型的病毒，他們的序列長度比 R5 型的病毒變化性更大。

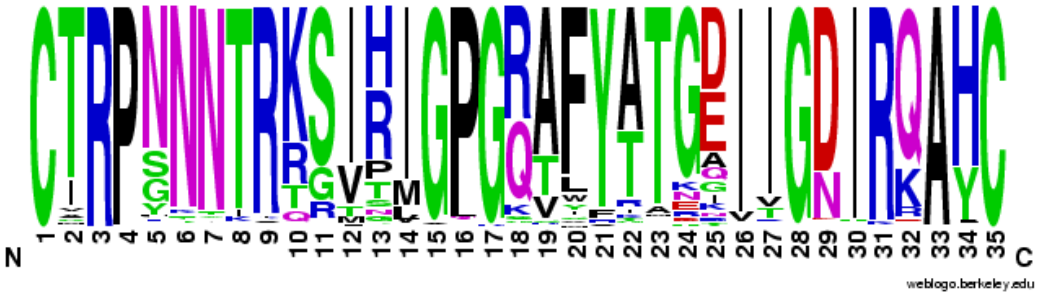

圖 4 V3 環狀序列胺基酸分布圖

資料來源:由 Data1225 作多重序列分析,並和先前研究提供的序列 HXB2 [1]作比較，以取得 35 個胺基酸長度的序列。

說明:圖 4 是利用 Data1225 中的全部序列作多重序列比對(multiple sequence alignment)的結果，並利用先前研究所提到的標準序列 HXB2 [1]來作比對，將沒有對應到序列上的位置刪除，藉此得到只含有 35 個胺基酸長度的序列組。再利用 WebLogo [4]所提供的軟體，對序列資料作胺基酸組成分析圖。

由圖可發現，經過序列比對以及刪除胺基酸較少出現的位置過後，序列資料的變異性還是非常的大，尤其是靠近 V3 環狀結構的尖端(13~20)和中央環狀部分(5~11, 22~30)，都是變異性較大的區域，這些區域也是影響 V3 環狀序列和協同受體連結的主要區域 [5]。

2.4 協同受體

在本篇論文中，主要研究目標為人類免疫缺失病毒的感染過程中所使用的協同受體。協同受體為人類免疫缺失病毒在進入目標細胞時，除了主要的 CD4 受體外，會幫助病毒連結在細胞外的蛋白質受體。人類免疫缺失病毒主要是使用 CCR5 和 CXCR4 這兩種協同受體，而根據所利用的協同受體可三種類型，分別是僅可以利用 CCR5 的 R5 型病毒或利用 CXCR4 的 X4 型病毒，以及兩種協同受體都可以利用的 R5X4 型病毒。我們的研究便是在預測病毒是這三類中的哪一種類型。

CCR5 和 CXCR4 都是趨化素受體(chemokine receptor)的一種 [6]，而趨化素受體為一種 G 蛋白受體(G protein receptor)。G 蛋白受體的主要結構為穿膜的蛋白質受體，其結構是穿過細胞膜，並且可以分別和細胞膜外及膜內的物質作連結，而 V3 環狀序列是和受體的膜外部份作連結。G 蛋白受體的膜外結構含有三個膜外環狀結構(extracellular loop)和 N 端結構(N-terminal)，這些膜外結構的組成和性質，是影響人類免疫缺失病毒連結的特性之一。

人類免疫缺失病毒所使用的兩種協同受體，都利用胱氨酸(cystine)產生兩組雙硫鍵，分別會將協同受體的 N 端(N-terminal)和第一膜外環(first extracellular loop) 以及第二和第三膜外環(second and third extracellular loop)連結，形成兩組會和 V3 環狀序列接觸的區域。這兩組區域分別會和 V3 環狀序列以及尖端兩部分作連結[5]。

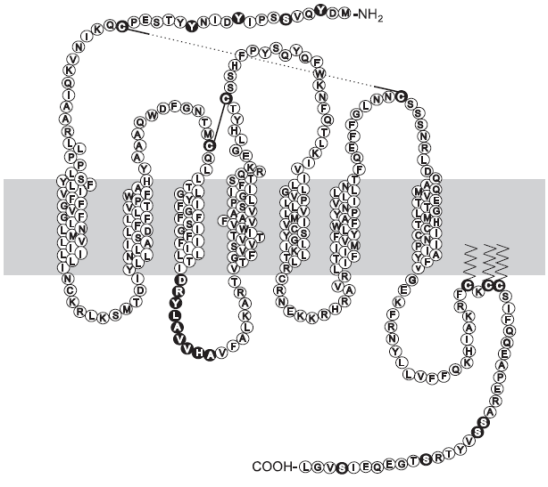

圖 5 CCR5 結構圖

資料來源：由 Oppermann 等人的研究中節錄的圖片[6]

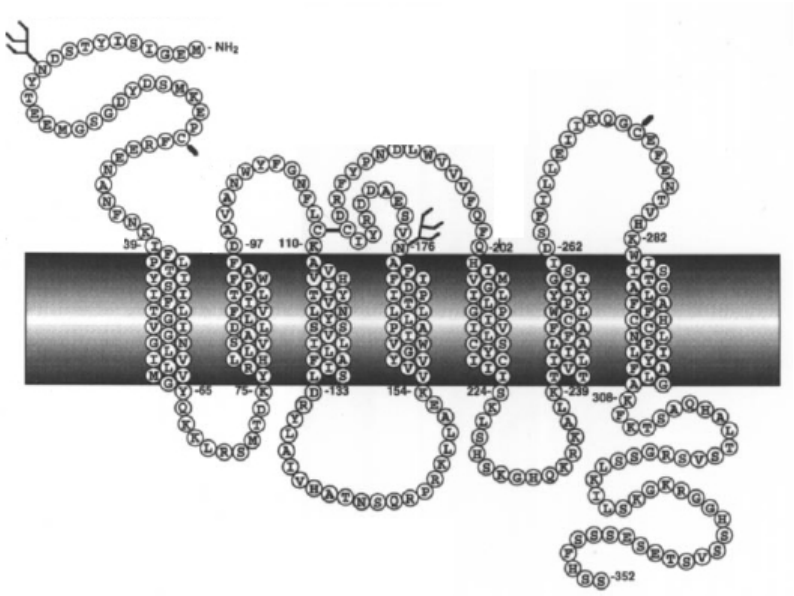

圖 6 CXCR4 結構圖

資料來源：Berson 等人的研究論文中得到的[7]。

說明：圖 5 和圖 6 中，協同受體的大結構都差不多，為含有七個穿膜螺旋的 G 蛋白結構。而其中主要影響人類免疫缺失病毒連結的，是細胞外 N 端部分和外環部分的序列組成。

三、現有方法

3.1 SVM 分類器

SVM(Support vector machine)分類器為一種處理分類問題的學習模型。SVM 再處理分類問題時，會尋找一個超平面(hyperplane)來將不同類的資料作區分，並且此超平面和資料群會有最大的距離空間。而為了使不同類型的資料能夠更有效的被分開，SVM 使用了幾種不同的核心方程式(kernel function)，來讓資料解搜尋空間，從低維度提升到高維度，讓尋找出來的超平面可以更有效的分隔不同類型的資料。在本研究中所利用的核心方程式為放射基底方程式(radial basis function)，是最常使用在 SVM 中的一種核心方程式，其公式如下：

$$K(x_i, x_j) = \exp(-\gamma \, x_i - x_j), \gamma > 0 \qquad (1)$$

在公式中，核心參數(kernel parameter) γ 被用來決定資料要如何去轉換到高維度的解搜尋空間中，而花費參數(cost parameter) C 是用來對 SVM 模型結果中的錯誤作出懲罰的參數。而 γ 和 C 在建造 SVM 模型中，都會被拿來調整，直到得到最好的表現量為止。在多類型的分類問題中，SVM 分類器會採取一對一的方式，將問題轉化成多個分兩類的問題並加以解決。當資料還有 h 種類別時，SVM 分類器會建造出 $h(h\text{-}1)/2$ 個子分類器，每一個子分類器都會將資料轉換成只有兩類的情況，接著個別處理分兩類的問題。當全部的子分類器得到結果之後，會利用投票的方式來得到最後的預測結果。本研究是使用了三類型的資料(h=3)，而使用的 SVM 分類器是整合 2.81 版的 LIBSVM [8]套件所得。

3.2 物化特性

物化特性為分析生化反應時，最直觀式的特性，因此我們把它用來應用在預測方法中。Amino acid indices (AAindex) database 中搜集了許多已經發表的胺基酸物化特性，在最新的版本中(AAindex 9.0版 [9])，總共含有544個物化特性，我們把其中有胺基酸特性值為NA的特性刪除，以讓每個特性都含有20個相對應的胺基酸特性值，篩選後還有531個物化特性 [10, 11]。因此在本研究中，我們會從531種物化特性中來挑選要使用在分類器中的特性，並由那些挑選出來的特性來預測HIV病毒的協同受體。

胺基酸序列中每個物化特性的值，是經由下面步驟來得到的：

(1) 從AAindex資料庫中，取得序列中物化特性對應到每個胺基酸的值，此時會有胺基酸序列長度 × 531(sequence length × 531)個值。

(2) 把序列中對應所選擇物化特性的值拿來做平均，每條序列在這步驟會得到531個值

(3) 將步驟(2)所得到的值作一般化，使得值的範圍在[-1, 1]之間，給SVM分類器使用。

3.3 繼承式雙目標基因演算法

為了要從 531 個物化特性中，挑出數量最少的物化特性，並得到最好的預測結果，我們使用繼承式雙目標基因演算法(Inheritable bi-objective genetic algorithm, IBCGA)[12]作為參數最佳化的工具。IBCGA 的原理為智慧型基因演算法(Intelligent genetic algorithm, IGA)[13]，並加入了繼承的機制，使結果更快得到收斂。IGA 結合了基因演算法與直交實驗兩種方法，除了可以快速的收斂之外，也會有較高的精準度。

IGA 是採用分裂並個別解決的策略(divide-and-conquer)與在交配(crossover)的過程中會使用直交表(orthogonal arrays)來挑選好的參數，使得 IGA 能克服大量參數和染色體過長的問題。在本研究中，IGA 可以有效的在解空間中去尋找並發現 C(n,r)的最佳解，n 為全部的物化特性個數，而 r 為挑選出來的物化特性個數。並且可從 C(n,r)的最佳解，藉由 IBCGA 的繼承機制，在 C(n,r±1)的解空間中去找尋最佳解[12]。所以，IBCGA 在限定 r 的範圍時，能夠有效率的獲得一組高準確度的物化特性參數。在本研究中，r 的範圍為 5~45 個物化特性。

適應性函數(fitness function)是用來評斷 IBCGA 的結果好壞的指標，我們所使用的適應性函數為十折交叉驗證(10-fold cross validation)所得到的全部準確度(overall accuracy)。

3.3.1 染色體編碼

IBCGA 中利用了 IGA 的原理，因此也會用到基因演算法中的染色體編碼來操作。這些基因演算法用的染色體被稱作 GA 染色體(GA-chromosome)，而組成 GA 染色體中的基因，則被稱為 GA 基因(GA-gene)。IBCGA 中的染色體，是包含了二維化選擇的物化特性基因和 SVM 分類器中的參數基因兩部份的 GA 基因所組成的 GA 染色體。而本研究中的每條 GA 染色體中，包含了 531 個物化特性基因和兩個 4 位元(4-bit)的 GA 基因來調整 SVM 中的 C 和 γ 參數。而我們利用 b_i 來選擇物化特性是否被使用，當 b_i=0 時，表示第 i 個物化特性沒有被使用在 SVM 分類器中，當 b_i=1 時，則表示第 i 個物化特性被包含在 SVM 分類器的物化特性組中。而 C 和 γ 參數則是由 16 個值下去作挑選，為$\{2^{-7}, 2^{-6}\ldots, 2^{8}\}$中所包含的值。

3.3.2 直交實驗設計

利用直交表來設計直交實驗和分析參數，為同時分析多個參數影響的有效方法 [11]。全因子參數實驗(complete factorial experiment)會去測量每個參數結合產

生的所有可能性，但是這樣會使得實驗過為龐大，因此我們使用直交表來作部分因子參數實驗(fractional factorial experiment)。直交實驗設計便是設計有效率的實驗參數，使得在部分因子參數實驗中可以獲得最好的參數組合。

直交表中的每一行(row)代表用來組合的參數數量，而每一列(column)代表每次組合參數時的改變值。參數的主效果(main effect)為考慮到其他參數不影響的情況下，單一參數選擇對結果的影響。我們利用製造直交表，來預估每個參數的主效果，以選擇出對實驗影響最大的參數。二因素直交表為一個$L_M(2^{M-1})$的表格，包含了M行和M-1列，如表2中$M=8$，為分析7個參數的直交表，而兩水準則是決定在實驗中是否選擇參數。f_t為評估函數值，為結合了t組實驗的評估值。我們定義在k水準下每個參數j的主效果值為S_{jk}，其中j=1, . . . ,M-1和k=1, 2，並得到公式：

$$S_{jk}=\sum f_t \times F_t,\ t=1,\ldots,M \qquad (2)$$

F_t=1表示j參數在t個實驗中k水準情況下的值。當$S_{j1}>S_{j2}$時，表示j參數在水準一的情況下會比水準二的情況下有較多的貢獻。利用主效果的值我們可以決定每個參數在何種情況下會有較多的貢獻，並藉此選擇最好的參數組合 [11, 13]。

表2是隨機挑出七個物化特性所得到的直交表範例。計算MED=$|S_{j1}-S_{j2}|$來判斷影響最大的參數，在本範例中影響最大的參數為第五個參數，其MED值為8.08。

表 2 主效果分析範例

	Factors							score
	1	2	3	4	5	6	7	
1	1	1	1	1	1	1	1	86.9033
2	1	1	1	2	2	2	2	82.4969
3	1	2	2	1	1	2	2	81.2729
4	1	2	2	2	2	1	1	82.4969
5	2	1	2	1	2	1	2	83.8433
6	2	1	2	2	1	2	1	85.5569
7	2	2	1	1	2	2	1	82.7417
8	2	2	1	2	1	1	2	85.9241
Sj1	333.17	338.80	338.07	334.76	339.66	339.17	337.70	
Sj2	338.07	332.44	333.17	336.47	331.58	332.07	333.54	
MED	4.8960	6.3648	4.8960	1.7136	8.0783	7.0991	4.1616	
Rank	4	3	4	7	1	2	6	
Better Level	2	1	1	2	1	1	1	

資料來源：是用SVM中用來訓練的資料，經由隨機挑選的物化特性(30, 43,70,95,142,205,281)作主效果分析後所得到的結果。

3.3.3 IBCGA 流程

使用適應性函數 f(X)執行 IBCGA 時，我們在每一次的執行過程中可以得到一組參數解 X_r，其中 r 是從之前定義的範圍中搜尋，r 的值為 $r=r_{start}, r_{start}+1, \ldots, r_{end}$。當定義好 r_{start} 和 r_{end} 的範圍值後，IBCGA 會依照以下步驟執行：

(1) 初始化(Initiation)：隨機產生一組起始的群組(population)，其中包含了 N_{pop} 條染色體 (individuals)。在染色體中的 n 個 GA 基因，包含了 r 個 1 值和 n-r 個 0 值的參數，其中 $r = r_{start}$。

(2) 評估(Evaluation)：利用 f(X)來評估每條染色體的值。

(3) 選擇 (Selection)：利用傳統 GA 的比較方法，隨機比較兩條染色體，並將有較好結果的染色體放到交配區(mating pool)中。

(4) 交配(Crossover)：在交配區中選擇 $p_c \times N_{pop}$ 個親代染色體，並將直交實驗運用到交配過程中，來輔助親代染色體的交配。其中 p_c 為交配發生率(crossover probability)。

(5) 突變(Mutation)：隨機選擇 $p_m \times N_{pop}$ 個子代染色體來進行交換式突變(swap mutation)以產生新的子代染色體，為 p_m 突變發生率。在突變的過程中，為了避免最好的結果被消滅掉，因此我們會先把最好的子代染色體挑出來，使其不參與突變步驟。

(6) 結束測試(Termination test)：如果 X_r 參數解滿足停止條件(stopping condition)，則輸出 X_r 中最好的染色體參數。沒有達到停止條件時，則回到步驟(2)。在本研究中，我們所使用的停止條件是執行了 40 個世代。

(7) 參數遺傳(Inheritance)：當 $r < r_{end}$ 時，在每條染色體中隨機選擇一個 GA 基因，使基因的值從 0 突變成 1，以增加 r 的數量成 $r+1$ 並回到步驟(2)。當不滿足 $r < r_{end}$ 時，停止 IBCGA。

3.4 HIV 相關研究

利用 V3 環狀序列來預測人類免疫缺失病毒的研究，大約是從 1995 年左右開始，就有利用電荷規則(charge rule)來作簡單的預測 [14]。這時候的分析主要是將人類免疫缺失病毒分成兩種類型，利用 CCR5 作為協同受體和利用 CXCR4 作為協同受體兩種，而可以同時利用兩種協同受體的病毒則分在使用 CXCR4 這類型當中。一直到了 2008 年才有第一篇研究，是直接將人類免疫缺失病毒分成三類型來作預測 [15]。以下介紹幾種較多人使用及效果較好的方法。

3.4.1 方法介紹

1. 電荷規則(charge rule)：電荷規則又稱為 11/25 規則(11/25 rule)，主要是因

為這個預測方法是看第 11 和第 25 個位置的胺基酸，是由哪種胺基酸組成，並且再參考 V3 環狀序列的整體帶電量，以推斷出人類免疫缺失病毒是哪一型的病毒。當第 11 個和第 25 個位置的胺基酸為帶正電的胺基酸，並且整體帶電量大於+5，便將序列所屬的病毒歸類在 X4 型病毒中[14]。如果是偏負電的 V3 環狀序列，則會被歸類成 R5 型病毒。

2. 位置加權矩陣(Position Specific Score Matrix, PSSM)：2003 年 Jensen 提出了位置加權矩陣的方法[1]。位置加權矩陣是將 V3 環狀序列長度固定後，把每個位置的不同胺基酸都給予一個統計上的值，並將胺基酸和位置的對應表作成一個 20*35 的矩陣。之後個別統計出 R5 和 X4 型病毒的胺基酸組成所接近的胺基酸對應值，並計算出標準值，以作為分類的依據。當有新的序列進來時，每個胺基酸都會有一個對應的分數，而序列便會取得整體的分數，再用此分數來判別是哪一型的病毒[1]。

3. SVM(Support vector machines)：在 2007 年時，Sing 利用了 SVM 來和其他方法作比較，並表示 SVM 是用來分類中最好的分類器[2]。這裡的 SVM 所搭配的分類依據，是將每個位置的胺基酸用二維編碼的方式，將 20 個胺基酸分別編成 20 個不同的二維組成，如 A=00000000000000000001，所以 20 個胺基酸就有 20 組不同的組合。利用這種編碼方式將每個位置的胺基酸編碼，最後會得到 20*35 個特性值，並用這些特性值作為 SVM 分類的依據[2]。

4. SVM+結構特性：2007 年還有一篇研究，是結合了結構特性來預測人類免疫病毒的類型[16]。這篇所使用的結構特性，是將胺基酸依照特性來分群，在將各群中胺基酸的距離來作為結構特性。用來分群的特性有氫鍵提供者(hydrogen-bond donor)、氫鍵接收者(hydrogen-bond acceptor)、離子鍵提供者和接收者(ambivalent donor/acceptor)、含脂肪族胺基酸(aliphatic)以及含苯環胺基酸(aromatic ring)六大群。每個胺基酸可以被包含在兩個以上不同群中，本篇的結構特性是看不同群的胺基酸之間的距離，來作為分類的特性[16]。

5. 演化式類神經網路(Eevolved neural networks)：之前的方法都是將人類免疫缺失病毒分成兩類來做預測，2008 年 Lamers 的研究是第一篇將病毒分成三種類別來作預測的研究[15]。Lamers 的研究是利用胺基酸的物化特性來作為編碼，並使用演化式類神經網路當作分類器，來處理病毒的建模及預測。Lamers 使用了 9 種的胺基酸特性，在固定 V3 環狀序列長度為 35 個胺基酸的情況下，可以得到 9*35 個特性值，此外他還加入了序列整體的帶電量(total charge)和等電子點(isoelectric point)，所以分類器會在 9*35+2 個特性中作挑選[15]。

表 3 為上述方法的比較，列出主要使用的分類器和分類用的特性為哪些，並作簡單的介紹。

表 3 人類免疫缺失病毒預測方法比較

Method	Year	Tropisms	Classifier	Properties	Other
Charge rule [14]	1995	Two	Charge value	11/25 site charge and net charge	Earliest method of prediction
PSSM [1]	2003	Two	Statistical cutoff value	PSSM	Effective improved the prediction accuracy
SVM [2] with binary coding	2007	Two	SVM	Binary coding sequences	Best method before 2007 (compare with ANN, PSSM, decision tree, charge rule, SVM)
SVM[16] with structure properties	2007	Two	SVM	Structure properties	Feature set contain sequences and structure properties
ENN [15]	2008	Three	ENN	Amino acid properties	First using the three classes of tropism to predict

資料來源：整理[15]、[1]、[2]、[14]和[16]論文中資料

說明：表 3 是論文比較表，主要由方法和所使用的特性作比較，並依年代排序。

3.4.2 方法分析

在預測的過程中，主要會有兩個問題：V3 環狀序列的不定長度和高變異性。在上面所提到的方法中，都無法克服序列的不定長度。因為分類所用的特性，都是使用固定位置及長度的編碼，所以必須要有相同長度的序列。因此，多重序列比對(multiple sequences alignment)在之前的方法中是必要的，在序列比對後，會因為長度不同的關係，產生部分胺基酸被刪除以及空白序列(gap)的產生，這些都會影響到預測的結果。

序列的高變異性會造成每個位置上有更多變化的胺基酸組成，之前的分類特性都是根據位置來作編碼，高變異性會讓依靠特定位置來作預測的方法，預測能力降低。因此在這種高變異性的環境下，以每個單一位置的特性作分類的方法較

不適合，以整條序列作為參考的方法可以考慮到整體的變異程度，是適合應用在此問題的方法。

我們的方法為了克服以上兩個問題，採用了 AAindex 作為我們的分類特性。AAindex 會將物化特性在整條序列中所得到的值平均，使得每個特性在序列中只有一個輸出的結果，這樣編碼的好處是可以考慮到整條序列的變異程度，並且可以處理不同長度的序列。所以我們的方法是不需要作序列比對便可將病毒分類，也可以顧到序列整體的特性。

除了編碼的方式較適用於這個問題外，我們也可以處理將病毒分成三類型的問題，並得到不錯的結果。我們採用 IBCGA 來挑選分類所使用的物化特性組，因為 IBCGA 可以有效的處理大量參數的問題，所以儘管是三種類型的病毒，也可以找出適合用來分類的物化特性。

而我們所使用的物化特性，都是可以從生物的角度來切入作相關分析，相較於之前的方法所採用的其他特性，能讓我們的結果再從其他研究來作驗證及解釋，並且可以讓生物學家應用到分析人類免疫缺失病毒中。所以在最後我們討論了所挑選出來的物化特性，看能不能從中找到一些新的貢獻。

四、研究方法

4.1 題目定義

本篇研究的主題是利用人類免疫缺失病毒上 V3 環狀序列的胺基酸序列，來預測其所屬的人類免疫缺失病毒是 R5、X4 和 R5X4 這三型中的哪一種。並討論所挑選出的物化特性組，其中所含的生物意義和相關研究。

我們是利用 AAindex 來轉換 V3 環狀序列，使其變成物化特性的數值，接著將物化特性值和要選擇的核心參數(kernel parameter)編碼成 IBCGA 的染色體，再利用 IBCGA 來選擇使用的物化特性和核心參數值。IBCGA 選出的物化特性和核心參數值，會在 SVM 分類器中利用十折交叉驗證(10-fold cross valuation)來訓練資料，並製造分類模型。最後我們挑選出最好的物化特性組，並探討物化特性組的分類能力以及那些物化特性的生物意義。

4.2 研究資料

主要的實驗資料是從Los Alamos National Laboratory HIV Sequence Database 而來，此資料庫為目前在人類免疫缺失病毒的研究上常被使用的資料庫。我們以含有V3 環狀區域為篩選條件，分別尋找利用CCR5、CXCR4以及CCR5/CXCR4三種類型的人類免疫缺失病毒的序列。依此找到的序列數量為2,940條，包含2,235條利用CCR5作協同受體的序列、383條利用CXCR4的序列以及322條利用CCR5/CXCR4的序列。此外還有另外兩組的研究資料，是從之前的研究[2, 15]取得的。這兩組實驗資料便用來比較我們的分類能力和之前的研究成果。

4.2.1 資料篩選

從資料庫搜集的序列中，許多序列的胺基酸組成是相同的，但是卻有不同的accession number。為了避免大量相同的序列會產生過度吻合(overfitting)，讓結果偏向特定的序列群，我們採取以下兩步驟來使得資料能更一般化：

1. 當相同胺基酸組成的序列使用兩種或兩種以上不同的協同受體，我們稱那些序列為衝突的序列(conflict)，並把它們全部移除。
2. 為了不被大量重複的序列所影響，因此我們把重覆的序列刪除，這些重複的序列為多餘的序列(duplicate)，每一組重覆的序列中，我們只留下一條作為代表。

經由以上兩步驟的資料篩選後，我們得到 1,225 條序列，包含 931 條使用 CCR5 作協同受體的序列、164 條使用 CXCR4 以及 130 條使用 CCR5/CXCR4 兩種協同受體的序列。

表 4 原始資料及篩選過後的序列數量

	R5	X4	R5X4	Total
Original Data	2235	383	322	2940
Duplicate and conflict	1304	220	191	1715
Handled Data	931	163	131	1225

資料來源：由 Los Alamos National Laboratory HIV Sequence Database 得到序列資料，並去除掉重複的序列和有衝突的序列後，剩下的序列為用在實驗中的序列。

4.2.2 資料分類

論文中有三組實驗資料。第一組資料為 Data2class，是由 Richard 等人[2]篩選的實驗資料，這組資料是將利用 CCR5/CXCR4 的病毒歸類於 CXCR4 類別裡面，因此這組資料只分為兩類，分別為使用 CCR5 以及 CXCR4 兩種，我們使用這組資料和之前的工具作比較。第一組資料包含了 423 條利用 CCR5 及 84 條利用 CXCR4 作協同受體的序列。

第二組資料為 Data139，是根據 Lamers 等人[15]所提供的 accession number 來搜尋，所找到的一組資料，其中包含了 139 條序列。Lamers[15]這篇論文是第一篇將病毒的協同受體同時分成三類並作預測的研究，我們利用這組資料來比較分成三類時的分類器好壞。

第三組資料為 Data1225，是使用前面提到從資料庫搜集和篩選的全部資料，包含了分成三類的 1,225 條序列。

表 5 三組資料的序列數比較

	R5	R5X4	X4	Total
Data2class	423	-	84	507
Data139	72	27	40	139
Data1225	931	130	164	1225

資料來源：為三組不同的資料數比較，分別是從之前研究[2, 15]和資料庫得來的資料。上表是每種類型病毒在資料中的詳細序列數量。

4.3 使用方法

利用 IBCGA 來選擇所使用的物化特性和 SVM 分類器的參數，可以幫助改善預測人類免疫缺失病毒所使用的協同受體的能力，而分析選出的物化特性組，不僅可以增加分類預測的可信程度，也可以幫助更多的了解人類免疫缺失病毒和協同受體連結的機制。因為 IBCGA 所得到的物化特性組不是固定的唯一解，所以我們在得到物化特性組和預測模型後，還必須利用其他方法來確認哪一組的物

化特性為最有效的。

4.3.1 SVM-PCP

SVM-PCP 為我們用來預測人類免疫缺失病毒使用的協同受體的方法，其中包含了利用 IBCGA 來選擇物化特性組和使用 SVM 分類器作機器學習的訓練。SVM-PCP 會自動決定一組物化特性組並建立 SVM 的訓練模型(SVM-model)。以下是 SVM-PCP 的執行步驟。

首先我們準備了 K 組的獨立資料來當作機器學習中訓練的資料。接著，每一組的資料都會執行 R 次的獨立訓練，所以我們會得到 $K\times R$ 組的訓練資料和物化特性組。在本研究中，我們採用 K=10 和 R=20，因此經過 IBCGA 的選擇後，總共得到 200 組的物化特性組，而每一組都含有 m 個物化特性。這些物化特性組會經過 SVM 分類器，最後得到 SVM 模型以用來做測試。

4.3.2 挑選物化特性組

之後要評估得到的物化特性組的分類能力，因此我們計算了選擇物化特性的頻率 $F()$，$F()$會計算訓練得到的全部物化特性組中，每一個物化特性被選擇的次數。統計完選擇次數後，會在計算每一個物化特性組的特性分數 S_r 其中 r = 1, …, $K\times R$，的 S_r 計算公式為：

$$S_r = \left(\sum_{i=1}^{m} F(P_i)\right)/\mathrm{m} \qquad (3)$$

其中 $F(P_i)$為第 i 個物化特性被選擇的次數，m 為特性組中所含的物化特性個數。最後我們選擇了有最大 S_r 值的物化特性組作為我們的分類依據。

五、研究結果

5.1 SVM-PCP 效能評估

我們利用不同資料組來和之前的研究成果比較，先評估 SVM-PCP 在病毒分成兩類及三類這兩種不同情況下的分類能力。之後再把挑選出來的物化特性組，用來評估在不同資料及分類器下的分類能力。

5.1.1 兩類型分析

因為在2007年之前的研究，主要是將人類免疫缺失病毒分成兩大類，分別是R5及X4兩類型，而R5X4這一型的病毒，通常會被歸類在X4這一類當中[2]。因此我們先利用Data2class這一組資料來評估當人類免疫缺失病毒只分成兩類時，我們所使用的IBCGA效能如何，並和之前的預測方法作比較。

在這組實驗當中，我們將資料分成五比一，其中五份的部分拿來作訓練，剩下的一份為測試資料。而我們所比較的方法有11/25 rule[14, 17]、PSSM[1]以及利用二進位編碼的SVM分類器[2]。

表 6 兩類型病毒預測比較

Method	Sensitivity (%)	AUC
11/25 rule[2]	59.5	*
PSSM[2]	71.9	0.90
SVM_{binary}[2]	76.4	0.91
SVM-PCP	88.9	0.94

資料來源：由Sing[2]得到的序列預測結果並和我們的方法作比較

說明：表6為比較結果，比較的標準是先固定預測的特異性(specificity)在92.5%，然後比較實驗的敏感性(sensitivity)，以及由訓練資料畫出來的ROC曲線所包含的面積AUC(area under the ROC curve)[2]。

由表中可以看到，當固定住預測的特異性結果時，利用IBCGA來挑選分類器所用的物化特性組，會比之前的方法好很多。除此之外，我們利用了分出去的測試資料來評估預測能力的好壞，所得到的準確度(accuracy)也有90.22%。

5.1.2 三種類分析

在2008年，Lamers將人類免疫缺失病毒分成三類來研究，並發展預測工具[15]。為了得到精準的結果，我們也採用將病毒分成三類的方法，也就是分成R5、X4和R5X4來作分析預測，並評估所使用方法的好壞。

我們使用從Lamers提供的accession number所找出來的Data139來作分三類的效能評估及比較。我們先將Data139利用Lamers所使用的比例來分訓練和測試資料，並將訓練及測試的結果來作比較。表7和表8分別為訓練和測試結果。因為Lamers在論文中是採用訓練中最好的一組來呈現[15]，所以我們也是用最好的結果來作比較。

表 7 比較 Lamers 訓練結果

	R5	X4	R5X4	Overall	Mean
Lamers[15]	75.00%	79.31%	40.00%	67.72%	64.77%
SVM-PCP	99.00%	81.50%	72.73%	89.15%	84.41%

資料來源：Lamers[15]的訓練(training)資料和我們的訓練結果作比較

表 8 比較 Lamers 測試結果

	R5	X4	R5X4	Overall	Mean
Lamers[15]	78.57%	70.00%	50.00%	70.00%	66.19%
SVM-PCP	91.67%	80.00%	60.00%	81.48%	77.22%

資料來源：Lamers[15]的測試(test)資料和我們的測試結果作比較

說明：從表7和表8可以看出，不管是訓練還是後面的測試，我們的方法SVM-PCP再分三類的資料時，都會比之前的有較好的準確率。

在Lamers的論文中只有預測的序列數和準確度，表中的MCC(Matthew's correlation coefficient)是我們利用公式以及他所給的資料計算得到。

5.2 統計分析

由 5.1 的效能評估可看到，利用 IBCGA 挑選物化特性組的 SVM-PCP 方法，會比之前研究所使用的預測方法有更好的效能。因此我們利用新得到的資料 Data1225 配合 SVM-PCP，來作預測模型的選擇以及物化特性的分析。我們將 Data1225 隨機分成三份，其中兩份作訓練，剩下一份作測試，並利用 SVM-PCP 來得到訓練模型。最後從 SVM-PCP 所產生的訓練模型中，挑選出物化特性分數最高的那組來作分析。

5.2.1 選出物化特性組

我們計算每一個物化特性在 SVM-PCP 訓練得到的模型中，所出現的個數，物化特性出現的個數即為分數。接著計算每一個模型中，所挑選出來的物化特性組的物化特性分數合，並由分數和作為評估訓練模型的好壞。表 9 是將所挑選的物化特性組和全部模型的準確率比較。

表 9 選擇模型與平均值比較

	R5	X4	R5X4	Training accuracy	Number of feature	Feature score	Test accuracy
PCP-model	98.55%	75.22%	52.87%	90.58%	14	29.43	90.44%
Average of models	96.11%	81.09%	59.97%	91.11%	35.2	20.74	87.51%

資料來源：用 Data1225 作 SVM-PCP 得到的資料

說明：PCP-model 是我們用物化特性的分數，挑選出來的一組模型。我們那她來和 200 組模型的平均值作比較，看利用特性的分數挑出來的模型，是否比其他模型好。而從表可以發現，儘管訓練的資料稍微較差一點，但是測試的結果是讓人滿意的。

利用物化特性來評估訓練模型的好壞，可以避免訓練過程的過度吻合(overfitting)，也可以挑選出分類能力較好的物化特性組。並且測試使用物化特性挑出來的訓練模型，也有不錯的準確率。表 10 為我們所挑選出來的物化特性組，其中包含了 14 個物化特性。物化特性是根據主效果分析(MED analysis)的排名作為排序。

表 10 物化特性組中的 14 個特性

Feature ID	Ranking	AAindex identity	Description
43	1	CHOP780206	Normalized frequency of N-terminal non helical region (Chou-Fasman, 1978b)
70	2	EISD860102	Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)
475	2	TSAJ990102	Volumes not including the crystallographic waters using the ProtOr (Tsai et al., 1999)
205	4	NAKH920104	AA composition of EXT2 of single-spanning proteins (Nakashima-Nishikawa, 1992)
281	4	QIAN880124	Weights for beta-sheet at the window position of 4 (Qian-Sejnowski, 1988)
320	4	RADA880107	Energy transfer from out to in(95%buried) (Radzicka-Wolfenden, 1988)
335	7	RICJ880114	Relative preference value at C1 (Richardson-Richardson, 1988)
386	8	WERD780103	Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga, 1978)
479	9	WILM950102	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)

30	10	CHAM830107	A parameter of charge transfer capability (Charton-Charton, 1983)
392	11	WOLS870103	Principal property value z3 (Wold et al., 1987)
142	12	KARP850101	Flexibility parameter for no rigid neighbors (Karplus-Schulz, 1985)
360	12	SNEP660102	Principal component II (Sneath, 1966)
95	14	FINA910104	Helix termination parameter at posision j+1 (Finkelstein et al., 1991)

資料來源：利用物化特性分數所挑選出來的物化特性組，為分數最高的訓練模型所使用的物化特性組，其中包含了 14 個特性。

說明：14 個物化特性，包含 AAindex 的 ID 和解釋，是用主效果分析結果來作為排序的依據。

5.2.2 主效果分析

評估單一特性對預測協同受體的影響是在往後的相關發展時，很重要的一環，我們利用了主效果分析(MED analysis)來評估單一特性的影響力。主效果分析是利用直交表來看在不考慮其他物化特性影響的情形下，單一物化特性對實驗結果的影響程度。當主效果分析的值越大時，表示此特性的影響越大。圖7為選擇的物化特性組作主效果分析後的分數直條圖，表11為詳細的物化特性資料和分數。

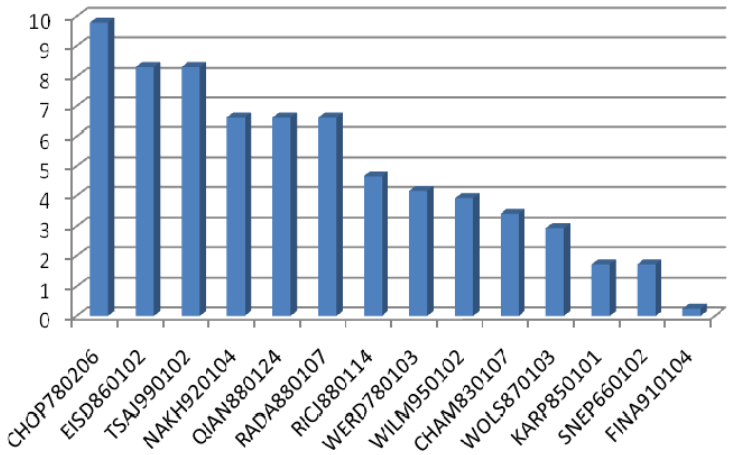

圖 7 主效果分析長條圖

資料來源：將挑選出的14個物化特性作主效果分析，將結果畫出來的長條圖。

說明：圖中的縱軸(Y)為主效果分析的分數，橫軸(X)為所代表的物化特性Identity。分數一樣時以AAindex的ID作為排序標準。因為這些特性已經是從531個特性中挑出來的，所以影響力都相當的大，導致比較其中單一特性的影響力時，各特性之間的差別性不大。

表 11 主效果分析中各物化特性分數

AAindex ID	Score	Feature descriptions
CHOP780206	9.791916	Normalized frequency of N-terminal non helical region
EISD860102	8.323135	Atom-based hydrophobic moment
TSAJ990102	8.323135	Volumes not including the crystallographic waters using the ProtOr
NAKH920104	6.609543	AA composition of EXT2 of single-spanning proteins
QIAN880124	6.609543	Weights for beta-sheet at the window position of 4
RADA880107	6.609543	Energy transfer from out to in(95%buried)
RICJ880114	4.651138	Relative preference value at C1
WERD780103	4.161575	Free energy change of alpha(Ri) to alpha(Rh)
WILM950102	3.916763	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O
CHAM830107	3.42717	A parameter of charge transfer capability
WOLS870103	2.937576	Principal property value z3
KARP850101	1.713577	Flexibility parameter for no rigid neighbors
SNEP660102	1.713577	Principal component II
FINA910104	0.244797	Helix termination parameter at posision j+1

資料來源：將挑選出的14個物化特性作主效果分析，將結果作成的表格。

說明：主效果分析結果的詳細資料，包含了物化特性的敘述和identity，以及主效果分析的值。其中值越大表示對這組資料而言，這個特性的影響力越大。

由以上圖表可以看到，在我們的物化特性組中影響最大的編號CHOP780206的”Normalized frequency of N-terminal non helical region” [18]，此特性主效果分析的值為9.792。而影響力最小的是編號FINA910104的”Helix termination parameter at posision j+1”[19]，主效果分析的分數為0.245。因為這些特性是已經篩選過的，所以整體而言每個特性的影響力都很大，導致比較單一特性時每個特性間的差別會比較小。

5.2.3 評估物化特性組

為了測試利用物化特性分數所挑選出來的物化特性組的分類能力，我們將這組特性放在其他資料和分類器中，來測試是否一樣有好的預測準確度。在不同資料方面，我們使用了 Data2class 和 Data139 來作比較。而在不同分類器的部分，我們利用 ANN 和 C4.5 兩種分類器來作評估。比較的過程一樣是將資料分成訓練和測試資料，並且使用由 5.2.1 所挑選出來的物化特性組來作訓練的分類模型並測試。

表 12 不同資料庫比較

	R5	X4	R5X4	Overall
Data1225	98.82%	58.82%	-	92.16%
Data139	100.00%	60.00%	20.00%	70.37%

資料來源：利用 14 個挑選出的物化特性，將 data2class 和 data139 作分類訓練及預測的結果。

說明：為了確認這 14 個物化特性的分類能力，我們將這些特性應用在其他資料上面，以比較在資料不同的情況下，是否也有好的預測能力。在 data139 的結果不是那麼的好，主要是因為資料數太少，導致這組資料不足以代表全體，所以會有較差的結果。

由表 12 可以看到，將這 14 個物化特性用在 Data2class 時，也可以有不錯的辨識度，但在 Data139 的部分，因為用來作訓練及測試的序列數過少，會造成資料的取樣性不足，使得結果稍為不理想。因此當資料數夠大時，利用我們所挑選出來的 14 個物化特性，可以有效的去預測人類免疫缺失病毒類型。

表 13 不同分類器下訓練準確度比較(training accuracy)

	R5	X4	R5X4	Overall
SVM	98.55%	75.22%	52.87%	90.58%
ANN	99.68%	90.83%	72.41%	95.59%
C4.5	99.20%	82.60%	69.00%	93.76%

資料來源：利用 14 個物化特性在另外兩種分類器下操作，得到的訓練結果

說明：比較在不同分類器下這 14 個物化特性的分類能力如何。在訓練的部分，會有過度適應的情形，不過都會有相當好的準確率。

表 14 不同分類器下測試準確度比較(test accuracy)

	R5	X4	R5X4	Overall
SVM	98.06%	81.82%	46.51%	90.44%
ANN	96.44%	81.82%	30.23%	87.47%
C4.5	96.10%	67.30%	32.60%	85.50%

資料來源：利用 14 個物化特性在另外兩種分類器下操作，得到的測試結果

說明：比較在不同分類器下這 14 個物化特性的分類能力如何。在測試的部分，儘管沒有使用 SVM 要來的高，但是整體的準確度都比之前 Lamers[15] 的方法要來的好，整體準確度也都有 85%以上。

表 13 和 14 可以看到，儘管在其他分類器中，會有過度符合的現象 (overfitting)，但是最後測試的準確率都有不錯的結果。雖然R5X4 型的病毒由於含有 R5 和 X4 兩種病毒的特性，因此在預測的準確度上會稍微差一點。但我們

所挑選的物化特性組在不同的分類器底下，都有不錯的分類能力，之後我們針對這一組的物化特性作更深入的分析。

5.2.4 獨立測試

我們在 2009 年 7 月的時候有再從資料庫獲得新的資料，將之前所用過的序列刪除之後，得到了一批全新的獨立測試資料(independent test data)。我們便使用這批資料來做預測，以檢驗 SVM-PCP 以及所挑選出來的物化特性在新的資料下，是否還是有好的分類能力。表 15 為新資料的資料數以及所得到的獨立測試結果。

表 15 獨立測試結果

	病毒類別			Overall
	R5	X4	R5X4	
Numbers	290	34	79	403
Test accuracy	97.93%	52.94%	21.52%	79.16%

資料來源：從 HIV 序列資料庫得到的一組新的序列資料，為我們取得序列資料後到 2009 年 7 月之間更新的全部資料內容。

說明：這組新的資料 R5X4 類型的病毒數量相當多，對照我們之前所使用的病毒數量，幾乎和我們那來作訓練的數量是一樣多的，序列數量的不平均，也導致了預測準確度降低，這部分還有改進的空間。

由表 15 可以看到，R5 類的病毒預測已經達到一個很穩定的狀態，幾乎都不會有預測錯誤的情形。但是 R5X4 類的病毒卻只有較低的準確度，我們推測可能是因為原本這類病毒的資料較少，導致訓練結果可能不足以完全得到他的物化特性，並且 R5X4 型病毒的特性是介於另外兩種病毒之中，也是較難分類的一類。此外我們在訓練的過程，是利用全部序列的準確率(overall accuracy)當作評估函數(fitness function)，會導致 R5X4 以及 X4 這兩類資料比較少的病毒預測準確度的降低。

5.3 生物發現

我們分析了分類模型中物化特性得分最高的那一組，並把其中可以直觀、簡單就被一般生物學家了解的一些特性，像親水性和疏水性、帶電量、...等拿出來討論並研究相關的生物意義。

5.3.1 可變性

因為 V3 環狀序列的多變性序列組成，所以 V3 環狀序列的結構非常的具有可變性(flexible)。在之前的研究中指出，V3 環狀序列的可變性會影響 V3 環狀序列和協同受體的連結規則，進而改變所使用的協同受體[20]。儘管 V3 環狀序列

的可變性會影響連結的協同受體，但是在之前的研究中還沒有找到確切的規則來說明可變性和病毒所使用的協同受體之間的關係[21]。在我們所挑出分類的物化特性組中，KARP850101 為可變性的物化特性[22]，但是在主效果分析中此物化特性的影響力為相對較小的。KARP850101 在 R5 和 X4 資料中的平均值為 0.232 及 0.040。

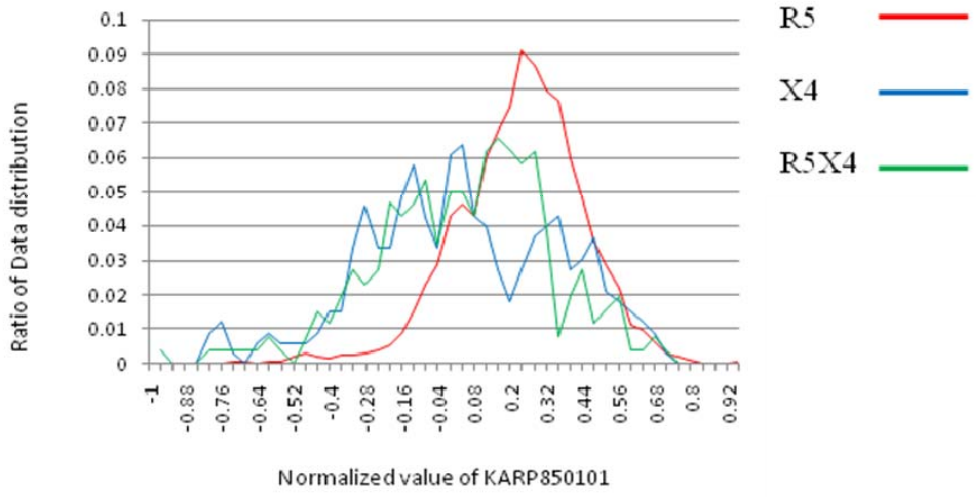

圖 8 KARP850101 特性的資料分佈圖

資料來源：將Data1225中不同類型病毒的KARP850101特性值取出來，並依照病毒類型分別作分布曲線。

說明：圖中的縱軸(Y)為資料的分布比例，橫軸(X)為一般化(normalized)後所代表的物化特性值。

圖 8 是利用 KARP850101 特性來看資料分佈，可以發現在 R5 和 X4 兩種類別的病毒資料中，主要分布區域是有區隔的。而 R5X4 型的病毒資料分佈，則是會介於另外兩類病毒之間。

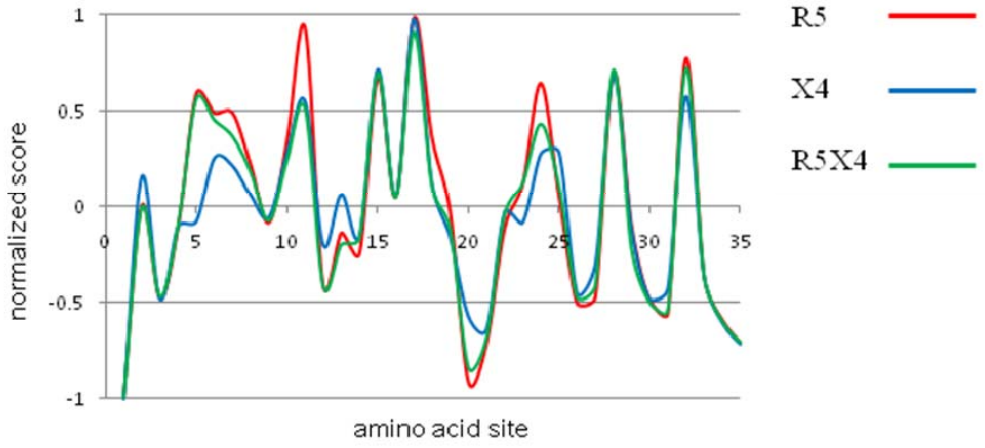

圖 9 KARP850101 特性在序列上的資料分佈圖

資料來源：先將Data1225作多重序列比對，並利用之前研究提供的標準序列HXB2 [1]來作比對，並得到長度為35個胺基酸的序列。之後將Data1225中不同類型病毒序列上，每個位置的KARP850101特性值取出來，經過平均之後，依照病毒類型分別作分布曲線。

說明：圖中的縱軸(Y)為一般化(normalized)後所得到的物化特性平均值，橫軸(X)為序列的位置(site)。

在圖 9 中可以發現，在胺基酸位置中，第 5~14 個位置和第 18~25 個位置這兩個區間的資料分布是差異較大的，而這兩個區間所代表的 V3 環狀序列位置，是 V3 環狀序列的結構中主要用來和協同受體作連結的中央主幹部分(stem)，而我們又作了圖 10 來分別看不同類型的病毒，在中央主幹部分的結構情形。

圖 10 是將 R5 和 X4 類型的病毒序列各選出 3 條，讓這些序列的結構重疊之後比較兩者之間的可變性差異。由圖中可發現，隨機選出相同數量的序列時，X4 的可變性會比 R5 的要來的大。圖 8 同樣是以 swiss-model 來預測挑選序列的結構，並用 RasMol 來繪圖。

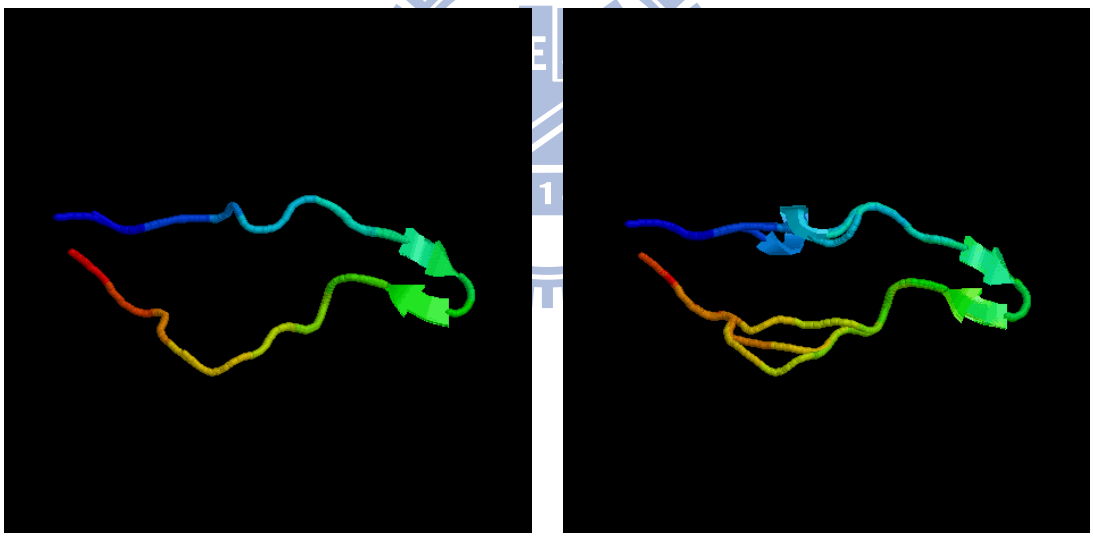

圖 10 R5 型和 X4 型病毒的可變性比較

(A)R5 型病毒 V3 序列　　(B) X4 型病毒 V3 序列

資料來源：將Data1225中R5和X4類型病毒的KARP850101特性值取出來，並將病毒資料中位於分布曲線頂點的三條序列取出來。利用Rasmol作圖使三條序列的結構重疊在一起，以比較結構的形狀。

說明：圖中有明顯結構差異的中央主幹部分，所對應到的胺基酸約在第 5~14 個位置和第 18~25 個位置這兩個區間，正好符合我們對序列中個別胺基酸位置的物化特性平均值分布所作的分析。

5.3.2 疏水性及電荷交換能力

疏水性矩(hydrophobic moment)和電荷交換能力(charge transfer capability)也是常常用來分析蛋白質序列的物化特性。疏水性矩會受到胺基酸中親水性和疏水性的影響，所以我們藉由分析親水性和疏水性來看 EISD860102 這一特性[23]。電荷交換能力是取決於胺基酸的官能基部分，不同的胺基酸所含有的電荷交換能力也會不同。在 CHAM830107 特性中主要以帶負電的胺基酸(Aspartic acid and Glutamic acid)為主，此種胺基酸的電荷交換能力也會較強[24]。

在 V3 環狀序列中第 29 個位置的天冬胺酸(Asp 324)是影響和 CCR5 協同受體之間連結能力的胺基酸。這個位置的天冬胺酸如果被天冬醯胺(Asn)取代的話，對於和協同受體的連結變化不大。但是如果是被帶電價或是疏水性的胺基酸取代的話，則是會減少和 CCR5 連結的能力[25]。而第 7 個位置的天冬醯胺(Asn 302)也是 CCR5 協同受體連結的主要位置之一[5]。V3 環狀序列的尖端及冠狀部分，電性和親水性也會影響協同受體的使用。如組成冠狀區域的 GPG 序列(Gly-Pro-Gly)，為不易被取代的疏水性胺基酸(conserved hydrophobic residues)，這類型的胺基酸會有助於病毒和 CXCR4 協同受體的連結[26]。

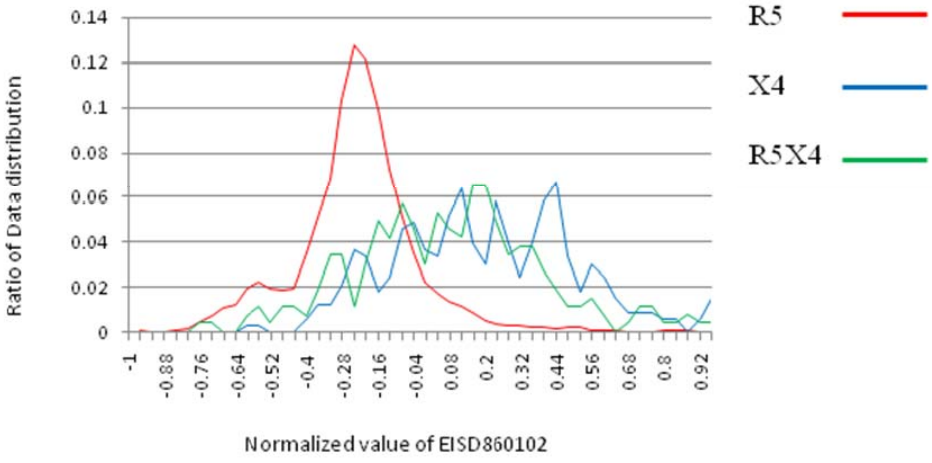

圖 11 EISD860102 特性的資料分佈圖

資料來源：將Data1225中不同類型病毒的EISD860102特性值取出來，並依照病毒類型分別作分布曲線。

說明：圖中的縱軸(Y)為資料的分布比例，橫軸(X)為一般化(normalized)後所代表的物化特性值。

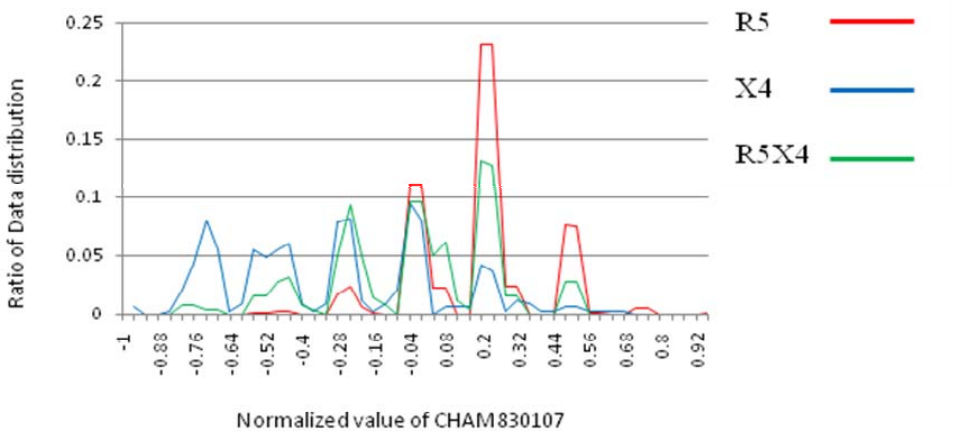

圖 12 CHAM830107 特性的資料分佈圖

資料來源：將Data1225中不同類型病毒的CHAM830107特性值取出來，並依照病毒類型分別作分布曲線。

說明：圖中的縱軸(Y)為資料的分布比例，橫軸(X)為一般化(normalized)後所代表的物化特性值。

圖 11 和圖 12 為單獨看疏水性矩及電荷交換能力兩種特性的資料分佈。主要的 R5 和 X4 型病毒的分布是在不同的區域，我們可以說單獨看這兩種特性時，都有可以分開 R5 及 X4 型病毒的能力。

圖 13 為 R5 型和 X4 型病毒的疏水性比較，圖中紅色部分為疏水性矩值較高的胺基酸，藍色部分為疏水性矩值較低的胺基酸，白色為中間值的胺基酸。由圖 12 可以發現兩種類型病毒的 V3 環狀序列中，X4 型比較偏好疏水性矩高的胺基酸，而 R5 則是偏好值較低的胺基酸。這點由 R5 序列的平均值為-0.229，比 X4 序列得到的平均值 0.210 低中可以得到統計上的解釋。

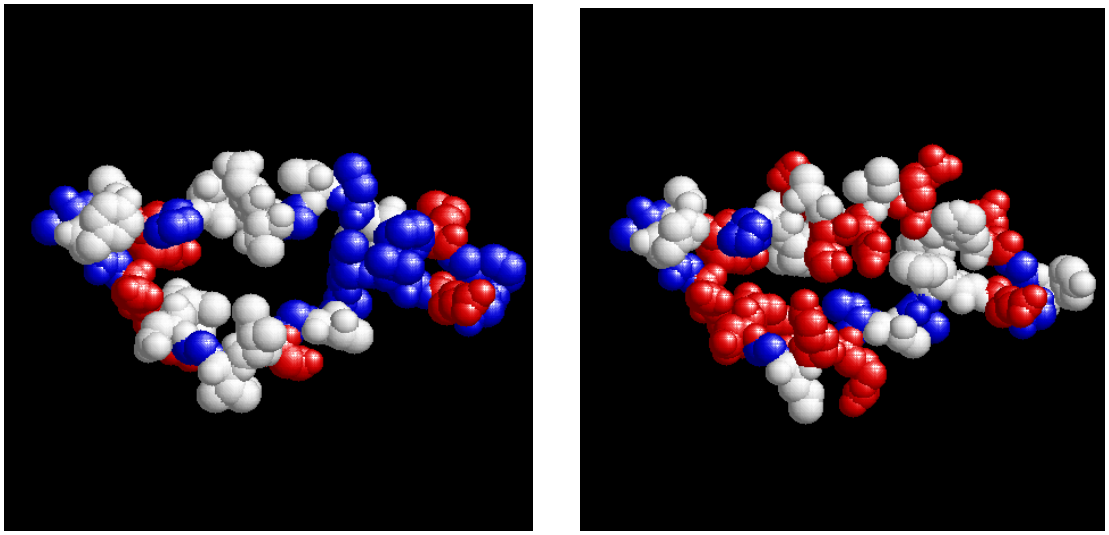

圖 13 R5 型和 X4 型病毒的疏水性比較

(A)R5 型病毒 V3 序列　　(B) X4 型病毒 V3 序列

資料來源：將Data1225中R5和X4類型病毒的EISD860102特性值取出來，並將病毒資料中位於分布曲線頂點的序列取出來。利用Rasmol作圖，比較兩種病毒疏水性矩大小的分佈情形。

說明：圖中利用紅藍色來代表疏水性矩值的高低，由圖可以明顯的看出來，兩種類型病毒的疏水性矩分布有很大的不同。

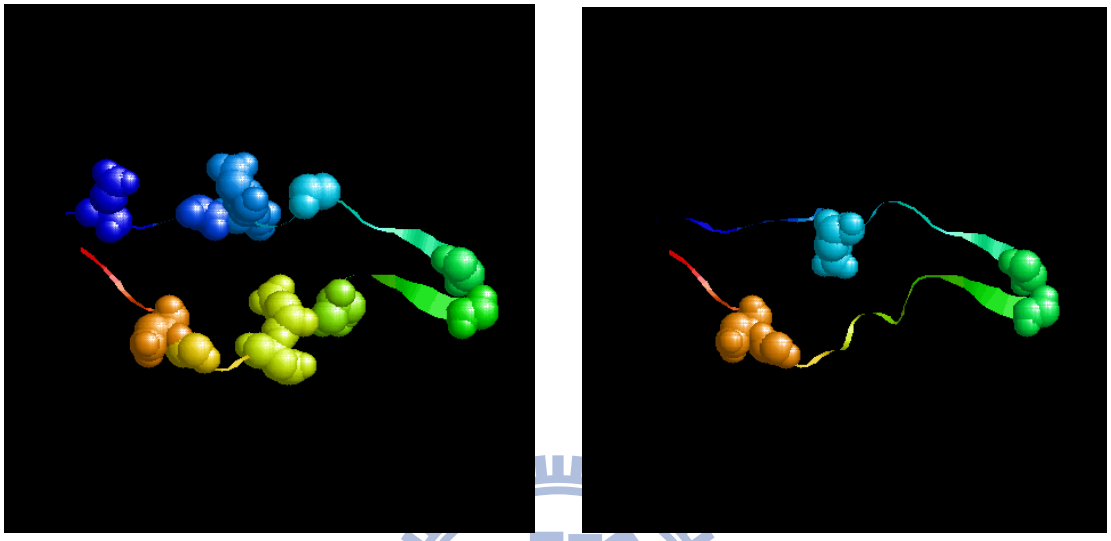

圖 14 R5 型和 X4 型病毒的電荷交換能力比較
(A) R5 型病毒 V3 序列　　(B) X4 型病毒 V3 序列

資料來源：將Data1225中R5和X4類型病毒的CHAM830107特性值取出來，並將病毒資料中位於分布曲線頂點的序列取出來。利用Rasmol作圖，比較兩種病毒電荷轉換能力大小的分佈情形。

說明：圖中利用紅藍色來代表電荷轉換能力值的高低，由圖可以明顯的看出來，兩種類型病毒所含有電荷轉換能力的胺基酸數量有很大的不同，其中 R5 型病毒含有較多高電荷轉換能力的胺基酸。

圖 14 為 R5 和 X4 兩種病毒的電荷交換能力比較圖，其中以空間球狀態顯示的部分是電荷交換能力較高的胺基酸。R5 型的病毒序列組成明顯擁有較多高電荷交換能力的胺基酸。在 CHAM830107 特性中所得到的值越高，表示含有較多電荷交換能力好的胺基酸，R5 型病毒序列資料的平均值為 0.193，遠大於 X4 型病毒的-0.289，表示 R5 型病毒的 V3 環狀序列會含有較高電荷交換能力的胺基酸。Luo 在 2007 年的研究中也有提到，使用 CCR5 作為協同受體的病毒，病毒中的 V3 環狀序列會比較偏好於帶負電的胺基酸組成[27]，也就是由有較高電荷交換能力的胺基酸組成。

蛋白質的疏水性及帶電荷在先前的研究便有被廣泛的提到，但是直接提到疏水性矩(EISD860102, Atom-based hydrophobic moment)[23]和電荷交換能力

(CHAM830107, A parameter of charge transfer capability)[24]這兩種特性的研究卻沒有，利用我們所挑選出來的物化特性，可以提供給生物學家新的研究方向。

5.3.3 端點序列組成

主效果分析中分數最高的物化特性 CHOP780206，主要是敘述 N 端的非螺旋結構區域的相關特性[18]，而 N 端的相關特性在之前的人類免疫缺失病毒的研究中，也常常被提到。V3 環狀序列中 N 端和 C 端的序列區域，是會和協同受體 CCR5 作連結的區域，而這部分的序列組成也會影響到協同受體的連結[25]。除此之外，可以利用 CCR5/CXCR4 兩種協同抗體的人類免疫缺失病毒，在 N 端的序列缺失(deletion)也會影響到幫助病毒連結的協同受體[28]。其他研究也說到 N 端的突變會導致 V3 環狀序列端點的結構不穩定，使得原本會和 C 端密合的 N 端結構產生移動(shift)，這種結構改變會造成所使用的協同受體由 CCR5 轉換成 CXCR4[29]。

圖 15 是看 CHOP780206 特性的資料分佈情形，可以看到 R5 和 X4 兩組資料有明顯的分開，而 R5X4 型的剛好位在兩者之間。

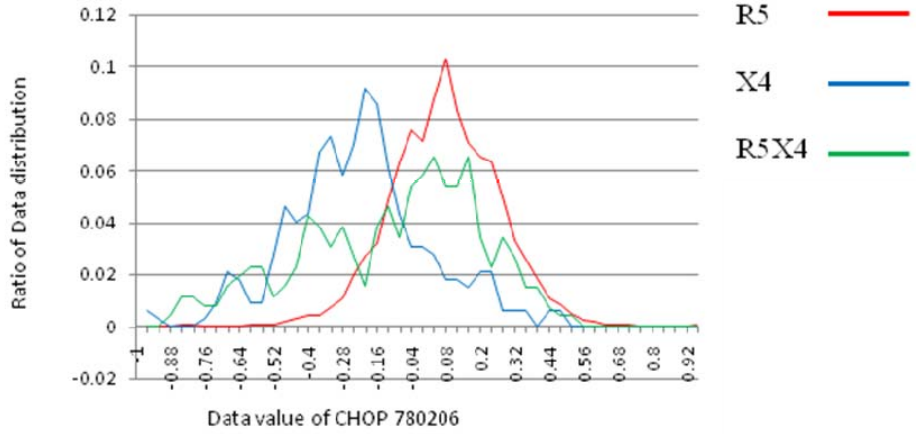

圖 15 CHOP780206 特性的資料分佈圖

資料來源：將Data1225中不同類型病毒的CHOP780206特性值取出來，並依照病毒類型分別作分布曲線。

說明：圖中的縱軸(Y)為資料的分布比例，橫軸(X)為一般化(normalized)後所代表的物化特性值。

圖 16 是比較 R5 和 X4 的 V3 環狀序列中，N 端的胺基酸組成。圖左的是 R5 型的病毒，在端點的部分會形成穩定的平板狀結構(strand)，而圖右 X4 型病毒因為端點組成胺基酸不同，導致端點結構較不穩定，不容易形成平板狀結構。CHOP780206 特性在標準化過後，R5 型和 X4 型病毒資料的平均值分別為 0.078

和-0.226。圖 16 是由 CHOP780206 特性中 R5 和 X4 兩類別資料中最靠近資料平均值的序列，利用 swiss-model 網站作結構預測，再經由 RasMol 軟體繪圖而成。主要是利用卡通方式(cartoon style)呈現出 V3 環狀序列的骨架結構，在端點的部份特別將不同的胺基酸標示出來，以作為比較。

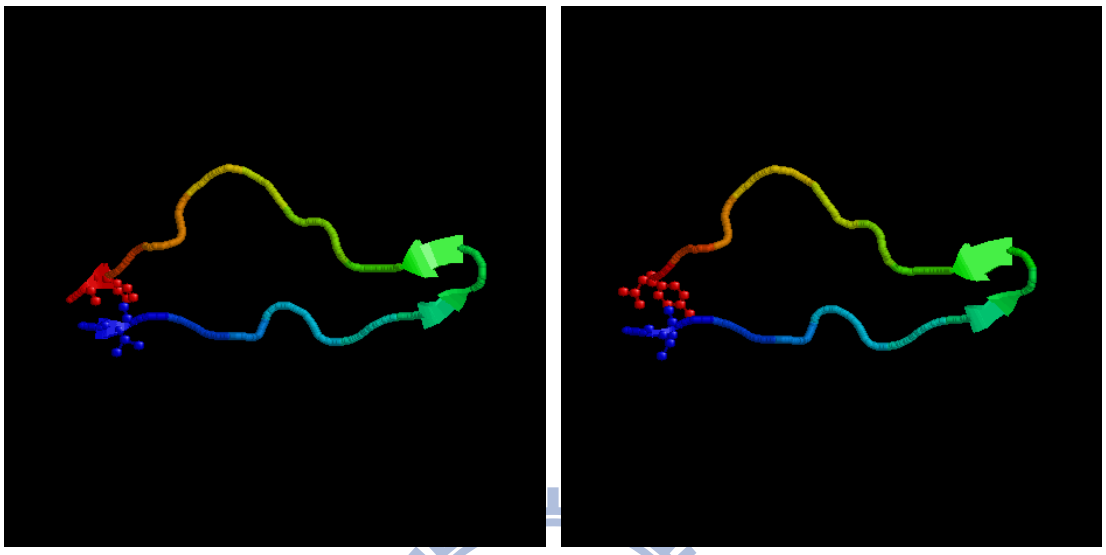

圖 16 R5 及 X4 型病毒 N 端結構比較圖
(A)R5 型病毒 V3 序列　　(B) X4 型病毒 V3 序列

資料來源：將Data1225中R5和X4類型病毒的CHOP780206特性值取出來，並將病毒資料中位於分布曲線中點的序列取出來。利用Rasmol作圖，比較兩種病毒N端結構的不同。

說明：由圖可以看出來，儘管兩種序列的結構會差不多，但是在 N 端結構的部分，R5 型病毒較容易形成平板狀的二級結構(sheet)，而 X4 型病毒則不易形成平板狀結構。

六、結論

6.1 結論

本研究希望提出一個新的方法，使得我們可以利用 V3 環狀序列，準確預測出人類免疫缺失病毒的類型，並且希望所挑選出來的物化特性組，能夠應用到不同的資料及分類器中。我們利用 SVM-PCP 挑選 V3 環狀序列的物化特性，並預測人類免疫缺失病毒是屬於哪一種類型的病毒。不管在病毒分成兩類型或三類型的情況下，利用 SVM-PCP 所得到的結果，都可以有效的提升預測的準確度。並且對我們挑出來的 14 個物化特性作評估，不僅在物化特性組所使用的 data1225 中有不錯的分類能力，搭配使用另外兩組資料時，也可以有好的預測能力。甚至在不同的分類器下，這一個物化特性組也能有相當好的預測結果。

而進一步分析這組物化特性，我們也找出了許多相關的研究，這些研究印證了此組物化特性的可信度。並且有些新的特性，可以提供給生物學家作為進一步研究人類免疫缺失病毒的其他方向。

6.2 未來展望

未來的研究方向，不僅可使用 V3 環狀序列的物化特性來做預測，還能夠加入其他周遭的序列一同分析，以找出更好的結果。或是將此技術應用到其他蛋白質間反應的問題上，以改進預測準確度以及輔助挑選蛋白質相關的物化特性，讓生物資訊的技術不只有單純的數字結果，能夠更直觀的去了解相關的生物意義。

參考文獻

1. Jensen, M.A., et al., *Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences.* J Virol, 2003. **77**(24): p. 13376-88.
2. Sing, T., et al., *Predicting HIV coreceptor usage on the basis of genetic and clinical covariates.* Antivir Ther, 2007. **12**(7): p. 1097-106.
3. Brower, E.T., et al., *Binding thermodynamics of the N-terminal peptide of the CCR5 coreceptor to HIV-1 envelope glycoprotein gp120.* Biochemistry, 2009. **48**(4): p. 779-85.
4. Crooks, G.E., et al., *WebLogo: a sequence logo generator.* Genome Res, 2004. **14**(6): p. 1188-90.
5. Huang, C.C., et al., *Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 gp120 and CD4.* Science, 2007. **317**(5846): p. 1930-4.
6. Oppermann, M., *Chemokine receptor CCR5: insights into structure, function, and regulation.* Cell Signal, 2004. **16**(11): p. 1201-10.
7. Berson, J.F. and R.W. Doms, *Structure-function studies of the HIV-1 coreceptors.* Semin Immunol, 1998. **10**(3): p. 237-48.
8. Chang, C.-C. and C.-J. Lin, *LIBSVM : a library for support vector machines.* 2001.
9. Kawashima, S., et al., *AAindex: amino acid index database, progress report 2008.* Nucleic Acids Res, 2008. **36**(Database issue): p. D202-5.
10. Huang, W.L., et al., *ProLoc: prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features.* Biosystems, 2007. **90**(2): p. 573-81.
11. Tung, C.W. and S.Y. Ho, *POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties.* Bioinformatics, 2007. **23**(8): p. 942-9.
12. Ho, S.Y., J.H. Chen, and M.H. Huang, *Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications.* IEEE Trans Syst Man Cybern B Cybern, 2004. **34**(1): p. 609-20.
13. Ho, S.Y., L.S. Shu, and J.H. Chen, *Intelligent evolutionary algorithms for large parameter optimization problems.* Ieee Transactions on Evolutionary Computation, 2004. **8**(6): p. 522-541.
14. Fouchier, R.A., et al., *Simple determination of human immunodeficiency virus*

type 1 syncytium-inducing V3 genotype by PCR. J Clin Microbiol, 1995. **33**(4): p. 906-11.

15. Lamers, S.L., et al., *Prediction of R5, X4, and R5X4 HIV-1 coreceptor usage with evolved neural networks.* IEEE/ACM Trans Comput Biol Bioinform, 2008. **5**(2): p. 291-300.
16. Sander, O., et al., *Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage.* PLoS Comput Biol, 2007. **3**(3): p. e58.
17. Fouchier, R.A., et al., *Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule.* J Virol, 1992. **66**(5): p. 3183-7.
18. Chou, P.Y. and G.D. Fasman, *Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins.* Biochemistry, 1974. **13**(2): p. 211-22.
19. Finkelstein, A.V., A.Y. Badretdinov, and O.B. Ptitsyn, *Physical reasons for secondary structure stability: alpha-helices in short peptides.* Proteins, 1991. **10**(4): p. 287-99.
20. Stanfield, R.L., et al., *Crystal structures of human immunodeficiency virus type 1 (HIV-1) neutralizing antibody 2219 in complex with three different V3 peptides reveal a new binding mode for HIV-1 cross-reactivity.* J Virol, 2006. **80**(12): p. 6093-105.
21. Watabe, T., et al., *Fold recognition of the human immunodeficiency virus type 1 V3 loop and flexibility of its crown structure during the course of adaptation to a host.* Genetics, 2006. **172**(3): p. 1385-96.
22. Karplus, P.A. and G.E. Schulz, *Prediction of chain flexibility in proteins.* Naturwiss, 1985.
23. Eisenberg, D., W. Wilcox, and A.D. McLachlan, *Hydrophobicity and amphiphilicity in protein structure.* J Cell Biochem, 1986. **31**(1): p. 11-7.
24. Charton, M. and B.I. Charton, *The dependence of the Chou-Fasman parameters on amino acid side chain structure.* J Theor Biol, 1983. **102**(1): p. 121-34.
25. Hu, Q., et al., *Restricted variable residues in the C-terminal segment of HIV-1 V3 loop regulate the molecular anatomy of CCR5 utilization.* J Mol Biol, 2005. **350**(4): p. 699-712.
26. Sundstrom, M., et al., *Mapping of the CXCR4 binding site within variable region 3 of the feline immunodeficiency virus surface glycoprotein.* J Virol, 2008. **82**(18): p. 9134-42.
27. Luo, L., et al., *Hydrophilicity dependent budding and secretion of chimeric HIV Gag-V3 virus-like particles.* Virus Genes, 2007. **35**(2): p. 187-93.

28. Nolan, K.M., A.P. Jordan, and J.A. Hoxie, *Effects of partial deletions within the human immunodeficiency virus type 1 V3 loop on coreceptor tropism and sensitivity to entry inhibitors.* J Virol, 2008. **82**(2): p. 664-73.
29. Rosen, O., et al., *Molecular switch for alternative conformations of the HIV-1 V3 region: implications for phenotype conversion.* Proc Natl Acad Sci U S A, 2006. **103**(38): p. 13950-5.