國　立　交　通　大　學

生物資訊及系統生物研究所

碩 士 論 文

辨識口蹄疫病毒抗原決定位之計算系統

Computational system for identifying antigenic determinant site of foot-and-mouth diseases virus

研 究 生：黃泰欽

指導教授：何信瑩 教授

中　華　民　國　一百　年　七　月

辨識口蹄疫病毒抗原決定位之計算系統

Computational system for identifying antigenic determinant site of foot-and-mouth diseases virus

研 究 生：黃泰欽　　Student：Tai-Chin Huang

指導教授：何信瑩　　Advisor：Shinn-Ying Ho

國 立 交 通 大 學

生物資訊及系統生物研究所

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 2011

Hsinchu, Taiwan, Republic of China

中 華 民 國 一 百 年 七 月

辨識口蹄疫病毒抗原決定位之計算系統

學生：黃泰欽　　指導教授：何信瑩

國立交通大學生物資訊研究及系統生物研究所所碩士班

摘　要

口蹄疫疾病主要影響偶蹄類動物是一個高度傳染性的疾病，被認為是世界上經濟方面最重要的家畜動物疾病。此疾病由口蹄疫病毒所造成，相關的研究認為疫情爆發時使用接種疫苗方式被視為一種更合理的手段。為了達到上述的目標，定義出具意義的抗原決定位及瞭解其重要的物化性質扮演一個重要的角色在保護性疫苗的設計、免疫的診斷及抗體的生產上。由於決定抗原上之 B 細胞表頂有賴於實驗的方式定義出來，但是此工作是耗時且昂貴的，因此發展計算的方式來幫助病原體自抗原的序列中定義出可靠性的表頂是必須地。

在此研究欲建立一個免疫的模型基於 B 細胞特異性的次分類群。我們擷取來自 IEDB 註解有關口蹄疫病毒病原體 B 細胞表頂之實驗數據做為我們的訓練資料。將訓練序列轉換成物化特性的特徵空間指標，並結合特徵選取的方式來改善預測準確度。使用繼承式雙目標基因演算法，其可最大化我們研究問題分類的準確度，同時最小化選取特徵數，來幫助擷取重要的資訊自我們的目標資料集之中。然後擴展所選出的物化性質組合來定義出抗原決定位的熱點藉由掃描病原體蛋白質。使用滑動窗口給予每一個詢問片段中心位點一個抗原性質傾向依據所選出的物化性質組合，結合投票的方式及利用智慧型基因演算法調整參數來達到一致性的預測結果。此外分析所選出物化性質來找尋重要的生物意義幫助改善疫苗的設計。

在這個研究中，我們發展一個計算的系統對於預測抗原決定位基於使用病原體特異性的次分類群及特徵選取的策略。結果顯示出，此建立的預測模型不僅能達到較高的預測準確度(訓練 89.33%及測試 72%)，也能自病原體蛋白質序列中定義出抗原決定位的熱點。此外基於特徵選取也能提供有用的生物訊息供分析之用。此系統不僅可以被使用當做研究口蹄疫病毒新興病原體的工具而且可以提供一個概念對於改善 B 細胞表頂的預測上。

關鍵字：口蹄疫病毒、FMDV、B 細胞表頂、抗原決定位

Abstract

Foot-and-mouth disease (FMD) is a highly contagious disease affecting cloven-hoofed animals and it is deemed as economically important diseases of livestock worldwide. The causative agent is the foot-and-mouth disease virus (FMDV). The first priority of suggestion in the outbreak was to develop effective FMD vaccines. Accordingly, identifying significant antigenic determinant sites and understanding its important physicochemical properties play an important role in protective vaccine designs, immunodiagnostic tests and antibody production. The experimental methods for determining B-cell epitopes are time-consuming and expensive. Therefore, it is desirable to develop computational methods for reliable identification of putative B-cell epitopes from antigenic sequences.

This study aims to establish a computational system for identifying antigenic determinant sites of foot-and-mouth diseases virus based on specific subclasses of B-cell epitopes of FMDV. We retrieved training data from the IEDB database and used the annotation of B-cell epitopes experimental data about FMDV. We transferred the training sequences to feature vectors based on the physicochemical feature index and then combined a feature selection method to improve prediction accuracy. An inheritable bi-objective genetic algorithm is used to maximize classification accuracy of the investigated problem and minimize the number of selected features to draw out significant information from our objective dataset. Then the selected feature set was spanned to identify hot points of the antigenic determinant sites by screening pathogen proteins. This method assigns a scale tendency value using the selected feature set and sliding windows of the query fragment. Moreover, we analyze the physicochemical feature set to mine significant biological findings to aid improve vaccine designs.

The results showed that the prediction system could obtain high performance (training accuracy 89.33%, and test accuracy 72%) and identify promising putative antigenic hot points. Moreover, the feature selection method could provide much useful information for biological analysis. The prediction system is capable of identifying antigenic determinant sites from pathogen proteins. The system not only could be used as a tool for investigation of emerging pathogen strain of FMDV but also provides a conception to improve B-cell epitopes prediction effectiveness.

Key word: foot-and-mouth diseases virus, FMDV, B-cell epitopes, antigenic determinant site

誌謝

由衷的感謝我的指導教授何信瑩老師平時對我們的關心指導，當我們產生放棄題目的念頭時告訴我們不要殘墨，鼓勵並給予我們研究方向的指引及提供我們許多寶貴的實驗室資源讓我們無後顧之憂地學習研究，此外平時除了研究上，何老師也教導我們做人的道理要常懷感恩的心，我也常常放在心上做為我的座右銘，在我們眼中老師彷彿是無所不能地。還有要感謝黃慧玲老師對於我們的關心，黃老師會用較為詼諧的方式指導我們，在課堂上耐心的引導我們程式的要訣。感謝黃憲達主任對我有關於論文架構上的建議讓我的論文內容能更臻完善。還有感謝論文研討時楊進木所長、尤禎祥老師及林勇欣老師給予我的指導及鼓勵讓我有勇氣在台上報告，也感謝楊所長對我的推薦雖然覺得受之有愧但是還是由衷地感謝楊所長。最後在這個所上的所有老師，無論是直接或間接的給予我們許多寶貴的資源及平時的言行都是我所敬佩的也十分地感謝他們。

對於實驗室的所有成員感謝你們在我的研究過程中於公於私的關心及鼓勵我會永遠放在心上的。最後我要感謝我的家人給予我的支持讓我能完成學業，也願所有我關心及關心我的人在將來的日子裡都能平安幸福!

目錄

表目錄

頁次

圖目錄

頁次

一、緒論

1.1、研究動機

口蹄疫疾病(Foot-and-mouth diseases, FMD)是一個高度傳染性的疾病，其主要影響偶蹄類動物(cloven-hoofed animals)，如：牛，綿羊，山羊，豬和鹿。此疾病由口蹄疫病毒(Foot-and-mouth diseases virus, FMDV)所造成，此病毒具有七種血清型，分別為 O、A、C、SAT-1、SAT-2、SAT-3 及 Asia-1，各血清型之間無交互保護力或是其保護力有限，因此防治上難度極大。此疾病造成許多嚴重的影響，如：直接增加年幼動物的致死率、間接影響感染動物所有方面的產率以及造成重要的貿易限制及減少觀光上的價值等。基於以上的理由，口蹄疫疾病被認為是世界上經濟方面最重要的家畜動物疾病。由圖 1 可以看出口蹄疫疾病至今仍然在世界各地爆發尤其是資源較為不足的國家，所以這是一個及需解決且重要的問題[1]。

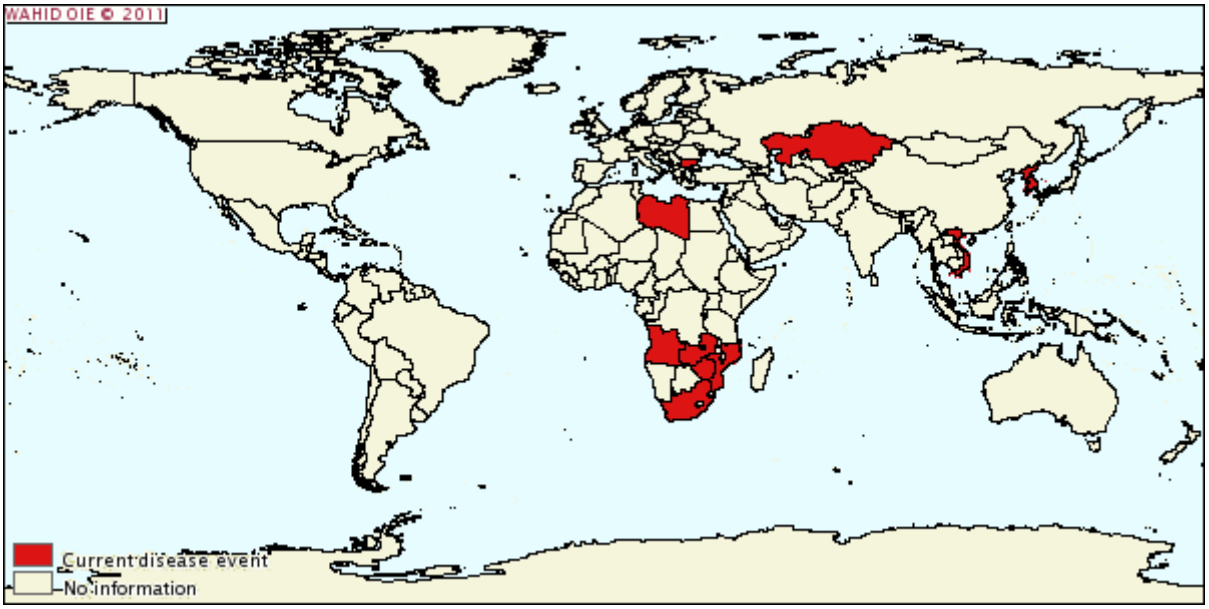


圖 1 為 2011 年 1~6 月間爆發口蹄疫疫情之區域
From World Animal Health Information Database (WAHID)

目前兩種主要的策略被用來控制疾病的擴散，分別為使用大規模的撲殺感染之動物以及在口蹄疫病爆發的期間使用疫苗提供動物保護力。在 2001 年期間英國爆發大規模的口蹄疫病疫情及其後爆發的疫情之後，相關的研究認為使用接種疫苗方式被視為一種更合理的手段來替代大規模的撲殺感染之動物，其後並建議認為應優先考慮發展改進口蹄疫病疫苗效益。然而，如果疫苗的手段被用來作為首選的控制措施，我們必須擴展我們對疫苗的理解範圍來改善疫苗的效益。尤其是傳統的不活化疫苗具有數種缺點，包括：需要具有大量培養有毒性的病毒、短的保質期(short shelf life)、需要冷凍保存以及不易於某些血清型及亞型細胞培養疫苗之生產[2]。基於這些原因，發展其不需要活病毒培養

的疫苗是必然地方向，例如：亞單位疫苗(subunit vaccines)、合成胜肽疫苗(synthetic peptides)[3-4]、DNA 疫苗(DNA vaccines)[5]以及重組病毒疫苗(recombinant virus vaccine)[6]等，均已被廣泛探討其效用在對抗口蹄疫病毒上。

以往的研究表明，合成肽或重組蛋白，其包含一個或兩個免疫原性(immunogenic)抗原決定位(antigenic determinant site)，可以誘導中和抗體(neutralizing antibodies)顯著的對抗口蹄疫病毒，並賦予充分的保護對於以口蹄疫病毒攻毒的小動物[7]。然而，這些類型的免疫疫苗在以口蹄疫病毒進行攻毒針對自然宿主(natural host)的實驗上，其效益大大低於傳統的滅活疫苗且只能給予自然宿主有限的免疫保護。有研究指出增加抗原表位的數量並找到更多具免疫抗原性的抗原表位可以增加疫苗免疫的效力[6, 8]。此外由於口蹄疫病毒具有抗原多變性的特性，可經由突變(mutation)及基因重組(gene reconstruction)兩種方式產生新型病毒，點突變(point mutation)只會造成抗原小部份改變稱為抗原漂移(antigenic drift)，此包括一連串漸進式自發的點突變，引起區域性的小流行，基因重組則為不同來源的病毒株同時感染同一宿主，複製過程中產生基因段的交換(crossover)和重新排列組合(reassortment)稱之為抗原轉移(antigenic shift)，則會引起全球性的大流行。

目前觀察病毒的變異與演化主要分為兩個方向，一為透過病毒之親緣關係，觀察不同品系(strain)之病毒與流行病學上(Epidemiology)的關聯性[9]；而另一方向為透過觀察病毒隨著時間推進，在序列各個位置上異同之處[10]。在瞭解產生變異的原因後，則需建立模型，模擬病毒演化的過程。例如：目前有一派的理論是病毒是以準種(quasi-species)的形式存在，而不是單一的病毒株，因此病毒株的擴增，是與病毒和宿主免疫系統的交互作用有關。一旦能預測到可能的病毒株，則要預測具有抗原性的區域(antigenic region)，做為以發展檢驗試劑與疫苗的參考。在病毒演化時，有些胺基酸的改變可能會與鄰近胺基酸一起變化，以維持其結構。這些區段的單獨變異可能會造成結構上大的變化，因而成為新的抗原。為了達到上述的目的，定義出重要的抗原決定位以及找出影響抗原決定位重要的物化特性(physicochemical properties)，就扮演相當重要的角色，在針對保護性疫苗的設計、免疫的診斷和測試以及抗體的生產上。

病毒感染的免疫防禦涉及既有先天(native)及適應(adaptive)免疫階段。從歷史上看，特異性的體液免疫反應(humoral immune response)被認為是最重要的因素在賦予保護對抗口蹄疫疾病上[11]。因此找出重要的抗原決定位就相當的重要在口蹄疫病毒的防治以及診斷上。然而，決定抗原上之 B 細胞表頂(b cell epitope)位置有賴於實驗的方式定義出可能的抗原決定位，但是此工作是耗時且昂貴的，尤其是面對此口蹄疫病毒多變的特性，在針對不同的病毒株上定義其抗原決定位經由實驗更是不易，因此發展計算的方式來幫助疫苗從病原體抗原的序列中定義出具可靠的推論表頂(deduced epitope)是必須地。最近，已經有研究結合實驗以及生物資訊(bioinformatics)的方式，在病原體疫苗研究上來幫助定義出口蹄疫病毒之病原體上可能的表頂位置，達到減少實驗成本的耗損的效益[12]。此外研究高變異性之病原體的研究中指出，使用生物資訊的方式來幫助疫苗的研

究及擴大我們對病原體的理解似乎是一個必然地方向[13]。因此，在此我們想要利用病原體相關之免疫資訊，研究口蹄疫病毒 B 細胞抗原決定位的物化特性，建立此病原體的免疫模型及預測系統，更進一步地獲取更多表頂的資訊達到幫助改善相關疫苗的研究上。

1.2 研究背景

B 細胞表頂或稱為抗原決定位(b-cell epitopes or antigenic determinants)，其定義為一抗原分子上的某個區域可以被免疫球蛋白(immunoglobulin)上的結合位所辨識稱之。此蛋白上的表頂分通常劃分為連續(continuous)或不連續的(discontinuous)，主要取決於是否列入表頂的胺基酸序列是連續的鏈或非連續的(圖 2)。研究指出大部份的 B 細胞抗原決定位為非線性的[14]，但是對於疫苗的研究以及免疫診斷的應用方面，針對連續性的 B 細胞抗原決定位研究是值得關注的方向。近年來，由於生物資訊免疫學的蓬勃發展已經發展出許多可應用的預測工具，在線性 B 細胞抗原決定位預測方面[15-16]。

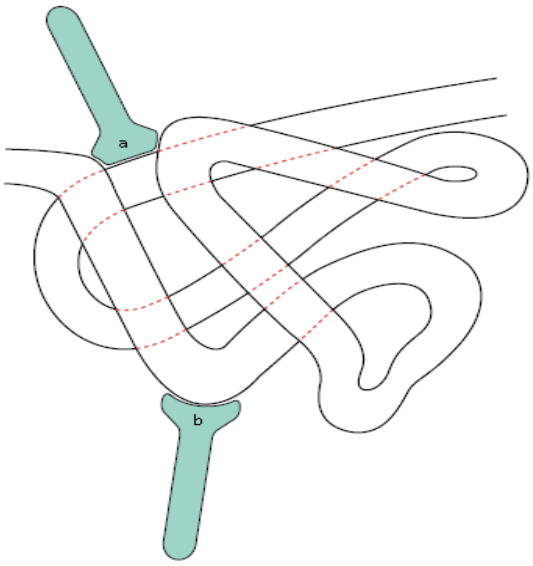


圖 2 為抗體結合至表頂與 B 細胞表頂類型示意圖[17]
註：a.為不連續型表頂、b.為連續型表頂

接下來針對我們欲研究的對象—口蹄疫病病毒，其生化背景做一個簡單探討。口蹄疫病毒是屬於小病毒科（Picronaviridiae）中的鵝口瘡病毒（Aphthovirus），其為一種無封套（non-envelop）的病毒，其結構為二十面體對稱（icosahedral symmetry）結構直徑約25nm；外面由結構上的衣殼（capsid）包覆著其基因體，其基因體由單股正股的RNA所組成，約有8500個鹼基，可直接作為mRNA轉譯成蛋白質。病毒的開放式閱讀框架(open reading framework,ORF)編碼一個大的單鏈聚合蛋白(single-chain polyprotein)，該聚合蛋白在轉譯過程中被不斷裂解產生病毒的結構性蛋白(structural protein, SP)包括1A、1B、1C和1D和非結構性蛋白（non-structural protein,NSP）包括2A、2B、2C、3A、3B、3C、

3D等，所以實際上並不存在一個完整的聚合蛋白，所以ORF可分為L、P1、P2、P3 四個區(如圖3)。

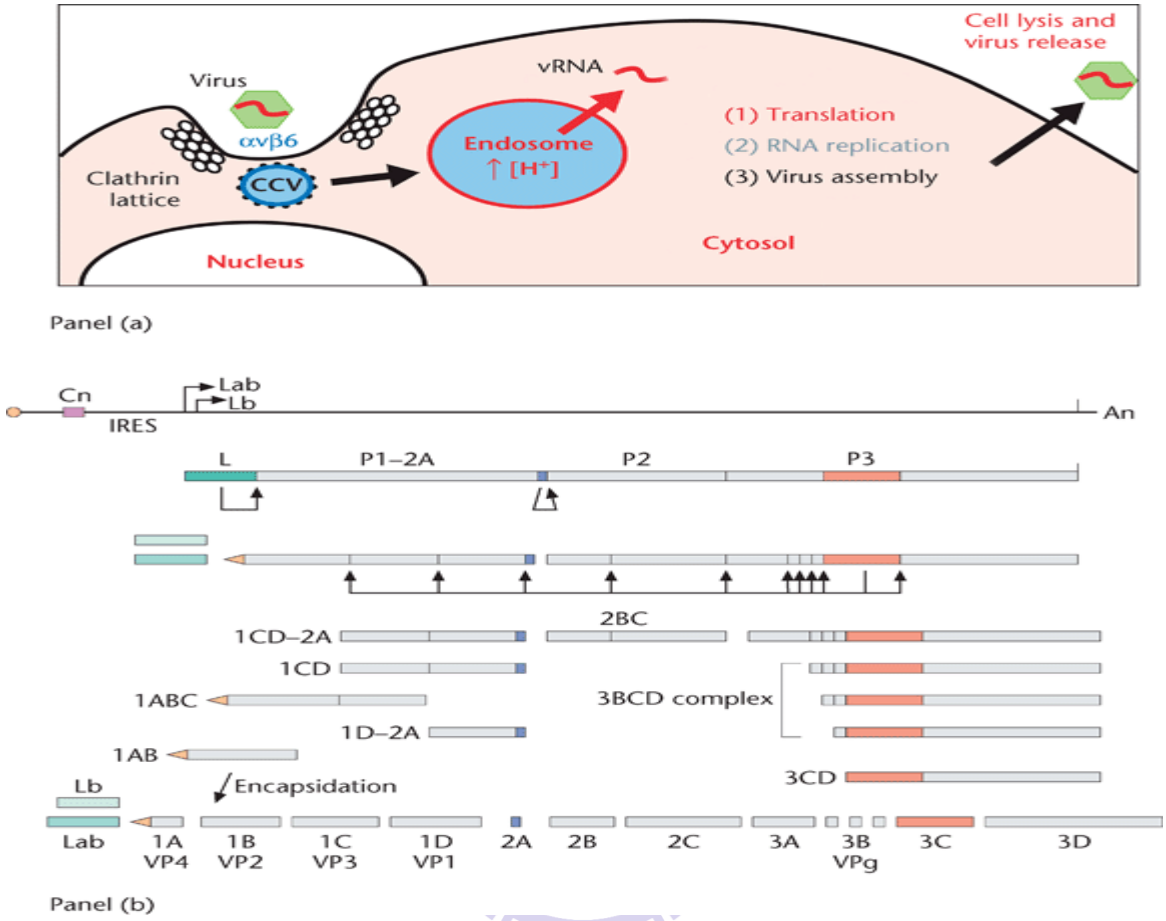


圖3 (a)口蹄疫病毒在單一細胞之生活史 (b) 口蹄疫病毒蛋白質體組成[18]

L區編碼LB和LaB兩種重疊的非結構性蛋白，這兩種蛋白是從2個不同的功能性起始密碼子轉譯而成的；LB和LaB催化自身從單鏈聚合蛋白上裂解下來，並裂解宿主細胞eIF24F和eIF24G，而阻止宿主細胞蛋白質合成，其中eIF24F是cap結合蛋白複合體中的成分，是宿主細胞蛋白質轉譯過程中所必需的起始因子。L區為病毒複製的非必需區，缺失L區的突變株仍能在宿主細胞內複製。

P1區依次編碼VP4(1A)、VP2(1B)、VP3(1C)和VP1(1D)四種結構性蛋白，最後組裝成病毒的衣殼蛋白。口蹄疫病毒的結構性蛋白（SP）組成的衣殼由60個複本組成，每個複本由四個蛋白質所組裝而成，分別為VP1(1D)、VP2(1B)、VP3(1C)、VP4(1A)組成一蛋白質次單元（protomer），其為構成蛋白質衣殼的最小單位，而蛋白質次單元再組成衣殼蛋白亞單位（capsomere）進而組裝成蛋白質衣殼（capsid）。而VP1(1D)、VP2(1B)、VP3(1C)在病毒顆粒表面組成八股的β-Barrels結構，而VP4(1A)則在內部(如圖4)。

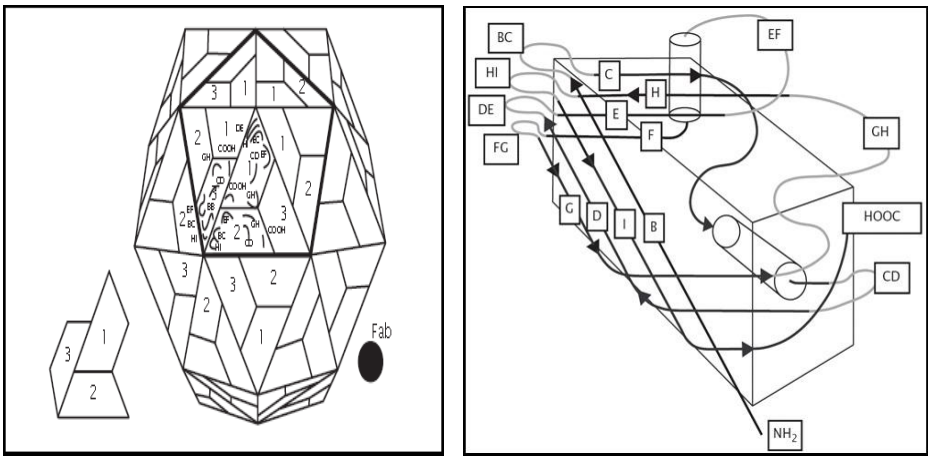


圖4 口蹄疫病毒的protomer及β-Barrels結構單元[19]

P2區依次編碼2A、2B、2C三種非結構性蛋白，2A為一16個胺基酸所組成的多肽，能夠自我催化P1-2A與2C解離，且從胺基酸序列分析推估2A可能是一種Tyr-Gly特異性的蛋白酶(protease)。2B和2C的功能尚不清楚，2C可能與RNA的複製有關，2B可能與病毒的感染有關。(雖然2A屬於P2區，但是有證據顯示其先形成P1-2A前驅物後再由3C所轉譯的蛋白酶將之裂解開)。

P3區編碼4種非結構性蛋白分別為3A、3B、3C和3D。3A被認為是小RNA病毒複製複合體與膜架構結合的錨定蛋白，與病毒誘導的細胞病理效應和阻斷細胞內蛋白的分泌有關。3A似乎與FMDV的致病性有密切的關係，不同血清型FMDV之3A編碼區的改變或缺失均會減弱其致病性。3B蛋白是由編碼區3個串聯重複的非等同的基因編碼，因而可以產生3種不同的3B蛋白(VPg)。VPg基因的複本數與RNA病毒的感染力有關。攜帶有pUpU架構的VPg蛋白可以與3′端的poly(A)結合作為病毒RNA合成時的引導蛋白，這種特殊的RNA複製模式與宿主mRNA轉錄不同，因而在宿主RNA合成受到抑制時，並不影響病毒RNA的合成。3C蛋白為Gln-Gly特異的蛋白酶，催化大部分聚蛋白的裂解；3C也參與組蛋白H3的裂解，可能對宿主細胞的基因轉錄有抑制的作用。3D為RNA倚賴的RNA聚合酶(RNA polymerase)，又稱病毒感染相關抗原(VIAA)催化病毒RNA的合成。病毒RNA就以VPg為引導物，以3D為RNA聚合酶，以3A、2B、2C及一些宿主蛋白形成複製複合體，附著於胞質內的膜架構上合成負鏈RNA，再以負鏈RNA為模板合成正鏈RNA，新合成的正鏈RNA包被完整的衣殼即成為一個完整的病毒顆粒[20]。

基於上述分子生物學方面的研究，蛋白質衣殼由四個部分VP4(1A)、VP2(1B)、VP3(1C)及VP1(1D)組成，由文獻可以得知對於定義出口蹄疫抗原決定位主要的研究在其結構蛋白方面，利用抗體定義出具保護能力的B細胞抗原表頂，現今已經有許多相關的研究定義出許多此病原體上保護性抗原決定位，例如：在結構上GH loop、BC loop、EF loop、C β-sheet、H β-sheet 等VP1上二級結構上[21-22]。其次，有必要提高診斷能力，對於感染動物及免疫動物的辨別上，因此近來也針對非結構蛋白部位單株抗體辨識抗原

表頂的辨識上用來幫助檢測感染與否[23]。

表 1 為相關研究中 1 不同血清型所定義出來的抗原決定位

O	A	C	Asia 1
G-H loop (RGD) VP1 133-160	G-H loop (RGD) VP1 133-160	G-H loop (RGD) VP1 133-160	G-H loop (RGD) VP1 133-160
VP1 C terminus VP1 198-213	VP1 C terminus VP1 198-213	VP1 C terminus VP1 198-213	
5 fold axis B-C loop VP1 43-48	H-I loop around VP1 170		
G-H loop (RGD) VP1 135~167			
B-C loop VP2 70-79	B-C loop VP2 70-79	B-C loop VP2 70-79	B-C loop VP2 70-79
E-F loop VP2 131-134 3 fold axis			
B-B "knob" VP3 56-61	B-B "knob" VP3 56-61	B-B "knob" VP3 56-61	B-B "knob" VP3 56-61
			C terminus VP3 56-61

註:表表示不同血清型(O、A、C 及 Asia1)其實實驗中被定義出的 B 細胞表頂，如：G-H loop 表示其結構上的名稱，VP1 表示其在蛋白質體上的部位，133-160 則是序列的位置在 VP1 上。

迄今為止，大多數商業化的抗病毒疫苗已被來自滅活或減毒活病毒或重組蛋白。雖然他們至今已成功對一某範圍的病毒病原體，但是此類型疫苗的有效性對高變病毒(hypervariable viruses)是有限的。這樣的高變RNA病毒逃避免疫機制的宿主物種由於高突變率的基因組上，它導致的轉錄錯誤。免疫系統無法趕上變化率與病毒逃避識別和控制，從而導致疾病或死亡。因此，迫切需要改進的疫苗能夠解決這些僵局是顯而易見的。事實上，正在採取各種辦法在發展這種改進的疫苗，其中的策略是採用反向疫苗的方法。逆向疫苗學涉及使用的計算方法來識別所有潛在候選人的免疫原基因組序列內的病原體[13]。

在有關定義出重要的口蹄疫抗原決定位來對抗新爆發的疫情，使用生物資訊的方式來對幫助增加實驗的效益及減少成本的耗損已經被使用，此研究篩選並確定B細胞抗原表頂的結構蛋白的口蹄疫病毒血清型Asia1，採用生物信息學方法結合分子生物學方法，以取得進一步的了解抗原的蛋白質結構中的應用設計的疫苗。在此研究中，描述了分析完整結構的胺基酸序列的使用DNAStar，使用多項物化性質找出潛在的B細胞抗原表頂，利用表現載體表達出此序列，然後用西方點墨法(Western Blot)檢測和酶聯免疫吸附試驗(ELISA)法分別評價其抗原性。此預測出潛在的17個 B細胞抗原表頂，找出了6個可能具有誘導免疫反應的序列，待進一步實驗進行確認其免疫原性[12]。

1.3 研究目標定義

當動物的免疫系統遭遇到一個新的抗原時，免疫系統會針對抗原上的抗原決定位(epitope)的數目，產生許多相對應的抗體。動物抗體被誘發後，可藉由採血後經由被誘導免疫動物之血清(serum)部分純化後獲得，亦即所謂的抗血清(antiserum)，此抗血清為多株抗體(polyclonal antibodies)，在許多免疫方面的研究，利用此抗血清來當作決定抗原決定位的位置以及抗原上是否有抗體能辨識的位置。一般來說可以被此靠血清所辨識的片段，我們可以合理地認為其具有抗原性(antigenicity)，但是是否具有能真正誘導欲免疫動物產生免疫原性(Immunogenicity)則是需要更進一步的動物實驗來認定。

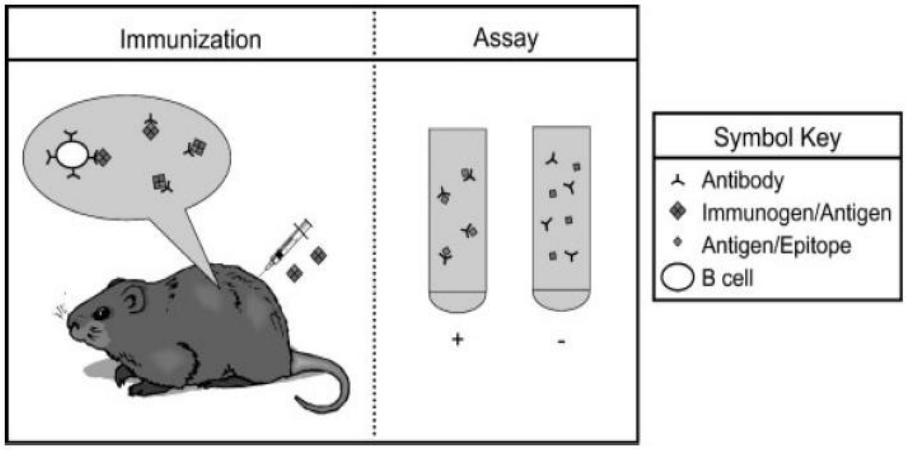


圖 5 為 B 細胞表頂映射概念[24]

圖 5 為利用圖表式描繪抗血清定義出病原上抗原決定位實驗，宿主暴露在某受質下產生抗體由 B 細胞，此受質即為免疫原(immunogen)，此受質被抗體體辨識可被當做一抗原，當抗原或免疫原被抗體特異性地辨識出來此位置為表頂。我們的題目為使用此經由表頂映射實驗所定義出來的實驗數據，找出此類型資料的物化特性來達到預測的效果。

1.4 研究目標

我們的目標是利用相關研究所建立的知識下，使用病原體(FMDV)相關B細胞抗原表頂免疫資訊，結合機械學習的方式，建立口蹄疫病毒的免疫模型來改善預測的準確度在此B細胞抗原表頂的子群當中。為了達到上述的目標必須從病原體資訊中萃取重要的特徵及降低特徵的維度達到減小有害的訊息及冗贅及不恰當的特徵，此需要針對不同次分類資料庫中，擷取出免疫相關重要資訊並使用此資訊來幫助免疫問題的研究。

我們必須瞭解我們所面對及欲解決的問題－即預測 FMDV 線性的抗原決定位自免疫

序列的資訊。由於此病毒具有易突變及多種血清型的特性，此趨同或趨異性演化缺乏顯著的序列相似性，但是其可能分享相同的結構及生物性質。在這種情況下，使用一般序列相似度的方式例如：排序(alignment)基礎的方式可能會導致預測的偏差。此外，抗原性是一種性質，也許被一種微妙的且深奧的方式編碼成序列，因此無法應用排序的方式來處理這類的問題。另一方面，由流行病學方面的研究顯示出，雖然大部分突變雖然傾向於負選擇(隨機突變)，但是口蹄疫病毒其抗原的變異受到達爾文正選擇的影響(定向突變)，其胺基酸變異上傾向於突變或取代成具有相似物化性質的胺基酸來逃避宿主的免疫系統[25]。

由此，在此我們使用物化特性(physicochemical properties)當做我們的特徵向量，我們轉換訓練序列成特徵空間基於物化特性的指標，並結合特徵選取(feature selection)的方式來改善預測準確度。由於此類問題需要由大量特徵向量中選擇出一組最佳的解集合，此為一 NP-hard 問題。為了解決此大量參數最佳化問題，我們使用 IBCGA 可最大化我們研究問題分類的準確度，同時最小化選取特徵數，來幫助擷取重要的資訊自我們的目標資料集之中。我們的目標是找出一組重要的物化特性特徵結合 SVM 分類器，達到改善分類準確度在口蹄疫病毒之 B 細胞抗原表頂的預測上。更進一步地把我們在此選出的物化特性特徵進行分析比較與先前相關的研究及找出其重要的生物意義。

另一方面為了擴展我們所選出物化性質的應用，我們使用此物化性質來定義出表頂的位置自病原體蛋白質體的序列中，我們使用不同長度的滑動窗口(sliding windows)結合免疫實驗數據去找出病原體上具此組物化性質所涵蓋的位置，然後再利用智慧型基因演算法找出一組最佳的滑動窗口組合達到結果的一致性，來決定可能的表頂位點。這部分的重點是我們使用一組由相關抗原物化性質來幫助找到重要的抗原決定位之熱點，此不同於各別物化性質的統計結果所定義出的位點而是使用綜合物化性質決定的結果。

在我們的研究中，我們發展一個計算的系統對於預測抗原決定位基於使用病原體特異性的次分類群及特徵選取的策略，找出重要的一組物化特性。此系統可以給予一段胜肽片段對其是否為具 B 細胞表頂性質做判別，也可以給予一蛋白質體序列利用此投票方式找出抗原的熱點。我們的結果顯示出，我們所建立的預測模型不僅能達到較高的預測準確度在口蹄疫病毒 B 細胞表頂的預測上，也能提供有用的資訊對於病原體蛋白質體抗原決定位的定位上。使用測試資料針對我們所預測的結果結合其它預測抗原決定位的工具所得到的結果，證實可以定義出可能的抗原決定位來減少實驗的成本及幫助疫苗的設計上。此外藉由分析此組物化性質可以更進一步瞭解及確認病原體抗原決定位的特性。

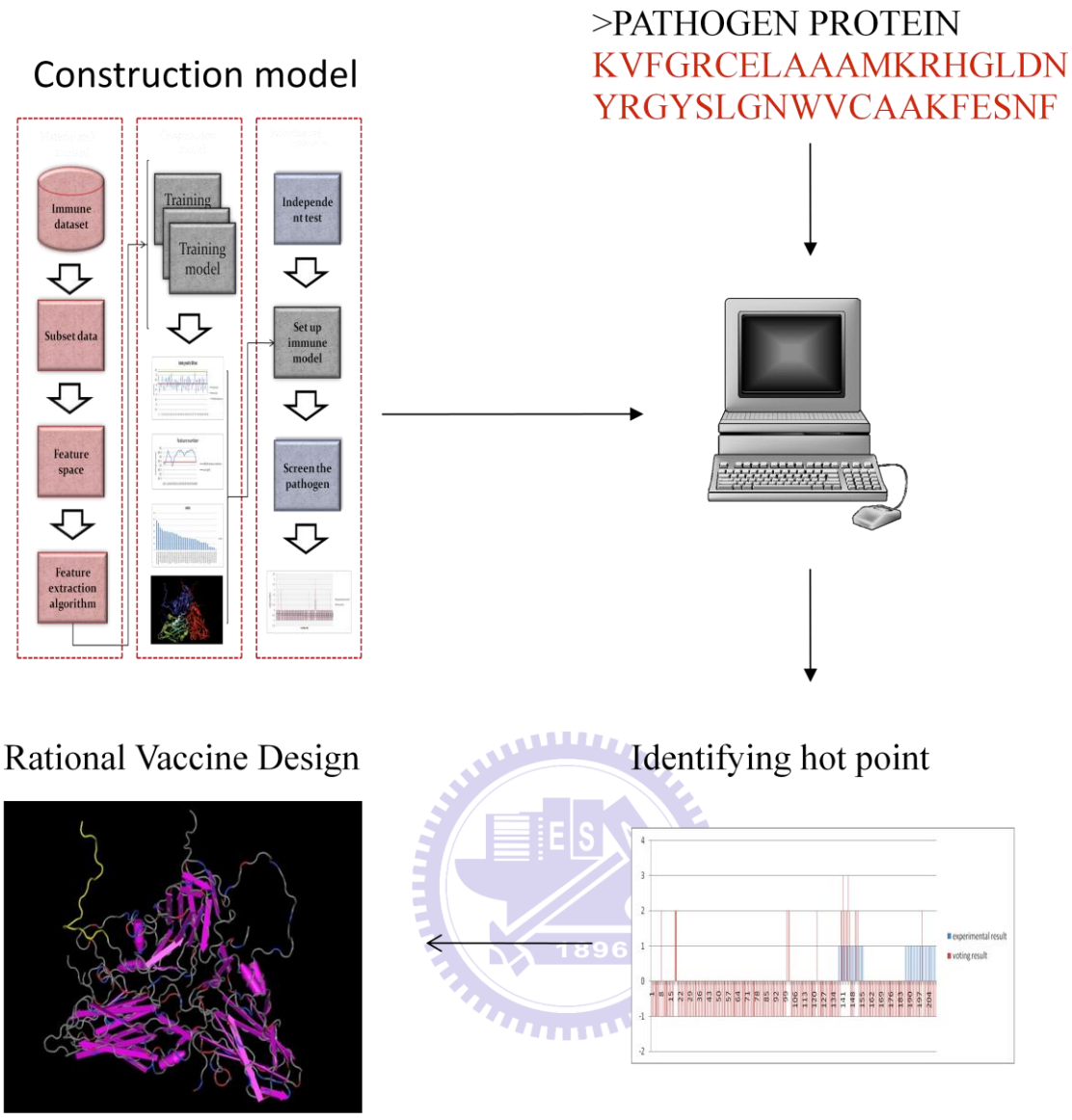


圖 6 FMDV 免疫計算的系統建構概念圖

二、計算上相關研究

2.1 資料庫

最初在 B 細胞表頂的的研究及預測基於研究 B 型肝炎表面抗原(hepatitis B surface antigen)、流感血凝素(influenza hemagglutinins)、雞瘟病毒的血凝素(fowl plague virus hemagglutinin)、人類組織相容複合體抗原(human histocompatiBility antigen HLA-B7)、人類干擾素(human interferons)、大腸桿菌(Escherichia coli)、霍亂腸毒素(cholera enterotoxins)、豚草過敏原(ragweed allergens Ra3 and Ra5)及鏈球菌(streptococcal M protein)，針對抗原胺基酸組成物化性質[26]。其後，發展了許多資料庫用來研究線性 B 細胞表頂的研究，在此我們介紹兩個主要用來發展線性 B 細胞表頂預測的資料庫。

Bcipep：

Bcipep 被建立用於發展線性 B 細胞表頂相關研究，Bcipep 是一個資料庫使用實驗的方式來測定線性 B 細胞表頂，包含免疫原性資料收集自不同文獻和其他公開資料庫中而來，此資料庫提供的有關單抗與多抗的抗體所產生的表位資料。其包含的 2479 個項目，每個項目包含胺基酸序列來源蛋白病原體分類免疫原性、中和性、實驗方法、模式生物、資料參考來源、抗體抗原結構等，並且涵蓋了範圍廣泛的病原微生物如病毒，細菌，原生動物和真菌[27]。此資料庫建立於 2005 年，其後許多 B 細胞表頂預測研究接使用此資料並用不同的機械學習型是建立模型。

IEDB：

IEDB (Immune Epitope DataBase, http://www.immuneepitope.org)此資料庫是由美國 LIAI 研究所(La Jolla Institute for Allergy and Immunology)協同其他學術研究機構共同合作，動員數十名研究人員，花費數年時間閱讀文獻蒐集、整理建製而成，為目前世界上資料量最豐富的抗原決定位資料庫，此外其會定期更新資料內容[28]。

2.2 研究方法

最初企圖預測B細胞抗原決定位研究建立在由相關實驗結果所得到物理化學特性(physicochemical properties)的分析上，其使用性質尺度的方法(propensity scale method)，將序列轉換為性質尺度值來衡量每個胺基酸的傾向，為了減少結果的波動，每個目標胺基酸序列在在滑動窗口(sliding window)計算平均性質傾向的胺基酸值，然後給予滑動窗口的中心目標的胺基酸殘基此性質尺度分數的值，以此為基礎預測是否給予胺基酸殘基序列可能是線性B細胞表頂(如圖7)。

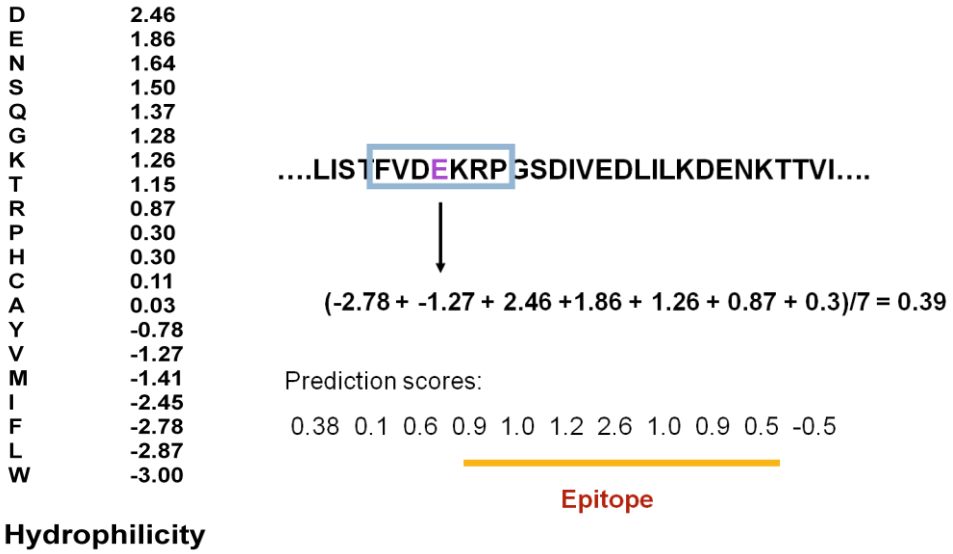


圖7 表示使用性質指標預測的概念

第一個傾向規模預測方法線性B細胞表頂是由Hopp及Woods利用Levitt所提出的親水性性質尺度(hydrophilicity propensity)分配的傾向值給予每個胺基酸[26]。這種方法是基於假設抗原決定位對應的蛋白質序列通常含有大量的帶電荷性和極性以及親水性殘基。其後，其他幾個傾向量表陸續被提出對於預測線性B細胞表頂，包括：親水性、易曲性、轉位、以及溶液的可接近性質[29-32]。之後，PREDITOP、PEOPLEE、PITOPE及BcePred預測線性B細胞表頂基於組合多項物理化學性質指標，而不是依賴個別性質指標的方式[33-35]。(如圖8)

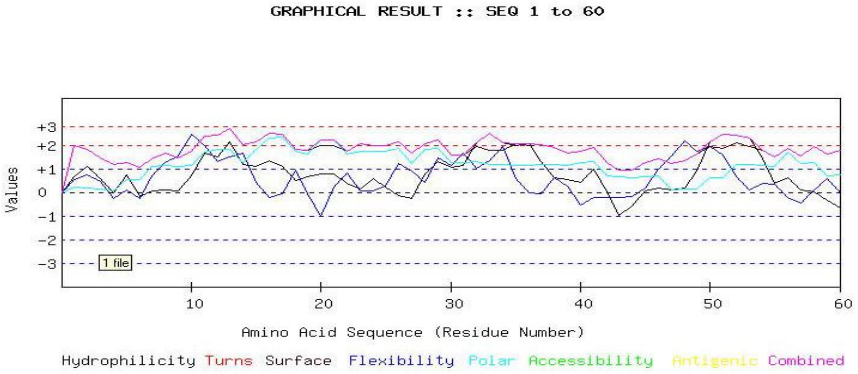


圖8 為BcePred server提供圖型化的輸出結果

註：圖縱座標表示其性質指標給的性質數值，橫坐標表示在抗原上的胺基酸位置

對於上述的研究根據Blythe and Flower研究中指出性質尺度的方法預測的結果只比隨機猜測略佳，其評估484個胺基酸的傾向量表之間的相關性研究傾向規模為基礎的輪廓和位置的線性B細胞表頂的數據集50蛋白質，其證實使用單一種胺基酸的性質傾向不能用來預測表頂的位置，並建議使用人工智能的技術幫助改善預測的方式[36]。根據相

關研究的瞭解，可以知道使用單一性質無法準確地預測出可能的B細胞表頂但是如果使用多項指標來找出可能的B細胞表頂，又不易確認出確切的邊界，因而產生過多的預測結果。

為了改善以性質指標的方式預測的準確度提出了一些改善的方式，例如：提出根據序列傾向的指標使用AAP指標[37]。BepiPred則是建立一預測模型使用隱藏馬可夫模型(HMM)結合親水性質指標的方式改善預測的準確度[38]。ABCPred使用倒傳遞類神經網路(RNN)對於預測線性表頂使用700個B細胞表頂及700個非B細胞表頂使用5倍交叉驗證訓練，輸入序列長度從10至20個胺基酸殘基，進行了測試和評估效益，得到66%的訓練準確率，在使用一個長度大小的16個胺基酸下為最佳的預測結果[39]。BCPred及FBCPred則是預測線性及彈性長度的表頂各自地，使用向量學習機使用string Kernels達到預測的準確度為67.90%及65.49%各自地[40-41]。COBEpro使用兩階段的過程預測線性B細胞表頂，在第一階段一個SVM分類器被使用來給予分數對於詢問的抗原片段，此SVM的輸入一個向量相似度對於在輸入的片段以及所有訓練的片段。在第二階段，一個預測的分數被產生關聯性與每一殘基在詢問抗原性質基於SVM分數對於每一片段。然而，COBEpro無法被使用針對分辨抗原從非抗原[42]。對於B細胞表頂預測的工具缺少一個一致性的比較方式，難以比較預測工具預測結果的優劣，如何定義出真正非表頂的特徵是此難以比較的原因之一，在無實驗驗證預測結果的研究之下[15]。

所以由以上相關研究的瞭解知道主要分為兩種類型的預測方式，一種為根據抗原的性質指標的方式給予每一個滑動窗口的中心點一個分數然後再利用訓練的結果決定表頂的位置及範圍，另一種則是給予一片段決定可能為表頂或非表頂。

以上相關研究的努力下已經有很大的改善在B細胞表頂的預測上，但是對於改善保護性的B細胞表頂在減少實驗成本的效益上仍值得努力。Söllner et al.最近研究了實用的預測的抗原性、序列變異和保護轉譯後修飾模體在預測具保護性線性B細胞表頂，他們的分析表明在查詢蛋白質序列使用重點領域的一個子集有可能提高預測性能線性B細胞表頂預測結果。這指出針對B細胞表頂的研究結合演化方面的資訊亦即設計分類器結合使用特異性次分類當做訓練資料，具有改善預測效率的潛力[15, 43]。

另一方面，有研究指出為了增加敏感度及特異性，一致性的方法被利用結合各種工具。下面的表格示範此一致性如何產生效益，理論上來說這是簡單及容易瞭解的概念。但是實際上，很多問題存在在這方面的研究，像是電腦計算的限制，計算預測的速度，不同的輸出格式以及困難的整合預測的結果。例如：一個序列片段在預測中被給予高的分數，但此預測結果與實驗結果所認同的表頂不同，困難地對於決定是否有高的分數為表頂或是預測中多次出現是否應當被包括在此一致性的輸出中。以上所述均增加困難在建構一致性的方法上，此外很少一致性預測軟體具有增加預測準確度可以被使用[16]。

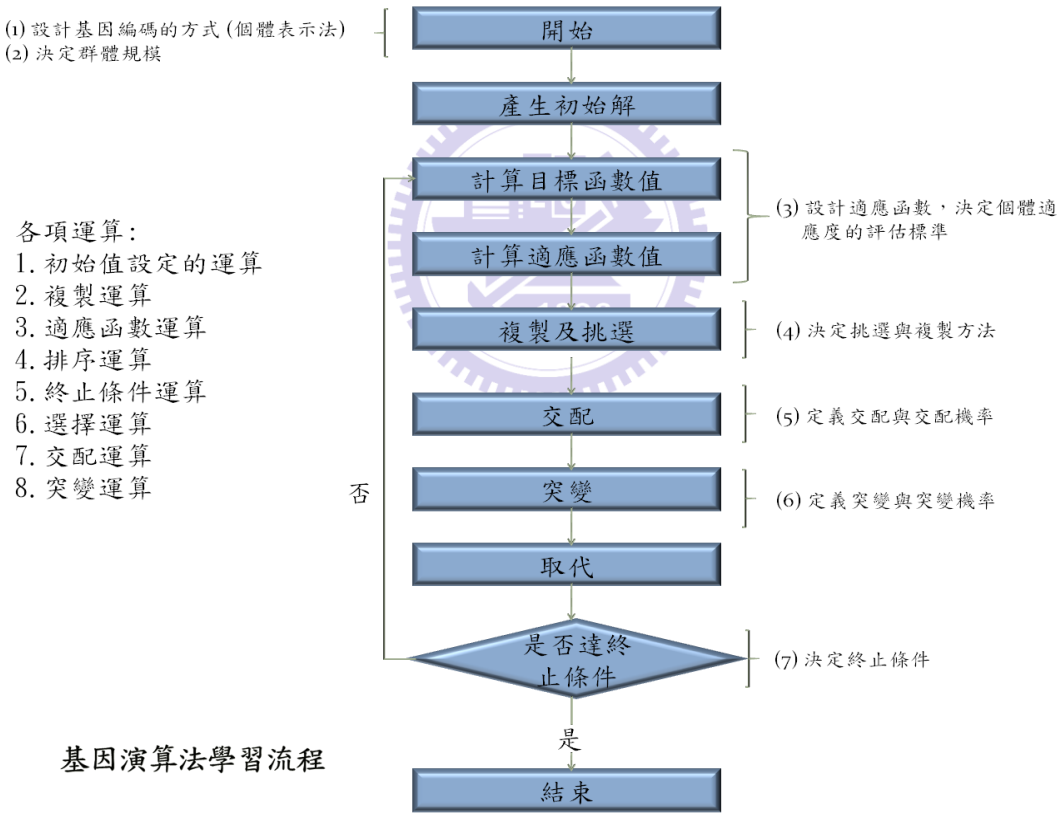
表 2 為表示使用表頂預測結果一致性的概念

Prediction method	M1	M2	M3	M4	Consensus method
Prediction result	Non-epitope	Non-epitope	Non-epitope	Non-epitope	0% epitope
	epitope	Non-epitope	Non-epitope	Non-epitope	25% epitope
	epitope	epitope	Non-epitope	Non-epitope	50% epitope
	epitope	epitope	epitope	Non-epitope	75% epitope
	epitope	epitope	epitope	epitope	100% epitope

三、最佳化演算法

3.1 基因演算法

基因演算法(Genetic Algorithm)是 John Holland 及其學生於 1970 年代左右共同研究出來的。經過近年來的不斷發展，現今已被廣泛的應用到求解最佳化問題、資料搜尋、人工智慧以及機器學習的領域上[44]。其以達爾文演化論為基礎，模擬生物界的“適者生存，不適者淘汰”的生存演化法則，每一物種在生存環境中會彼此相互競爭，只有適應性強的物種才得以存活與繁衍，這種自然淘汰機制會逐漸發展出來最優秀的品種。Holland 認為自然界的演化是發生在生物染色體的基因中，每一種生物的特徵係來自於該種生物上一代的基因排列，而演化是指每一代間的基因所發生的變化情形。所謂適者生存是指這一代的基因排列優於上一代的基因排列而產生比上一代更能適應環境生存的世代。



基因演算法學習流程

圖 9 基因演算法示意圖

因此基因演算法是強調基因型的轉變，將問題的可能解答經過編碼成為基因型式，利用遺傳運算子演化來找到最佳解，其中基因型的改變主要經由交配、突變達到演化的效果。在此部分我們使用實數型(real-code)基因演算法來設計染色體，此具有減低如二進位(binary)及符號型(symbol)基因演算法實際操作時需要經過編碼及解碼的步驟，只要針對問題的解空間進行染色體設計便可達到演化式計算的效果。

3.2 直交實驗設計與因素分析

3.2.1 直交表

直交表 (orthogonal array) 是由 R. A. Fisher 最先提出的，直交所代表的意思是平衡 (balance) 而不混合 (mix)，亦即統計上的獨立 (statistically independence)，因此直交表中每一欄的各水準值 (level) 出現次數是相同的，使用直交表事實上僅是進行部份因素實驗 (fractional-factorial experiment) ，因此能較完全因素實驗 (full-factorial experiment) 節省大量執行的時間，且直交實驗具有系統推理的特性，因此只需進行部份因素實驗就可以求得最佳解的近似解 (near optimum)[45]。

以兩水準，三因素直交表說明，若要進行完全因素實驗，需要八次實驗($2^3=8$)，如果需要在八次實驗中選擇只作四次實驗，如何選擇能達到均勻取樣的目的，建構一個 $L_4(2^3)$ 的直交表，如下表所示，可以縮減成四次實驗，這樣的取法使用均勻且對稱的取樣來推測全部實驗的最佳解。

表 3 為兩水準三因素完全實驗

實驗批次	因素		
	F1	F2	F3
1	0	0	0
2	0	0	1
3	0	1	0
4	0	1	1
5	1	0	0
6	1	0	1
7	1	1	0
8	1	1	1

表 4 為兩水準三因素直交實驗表

實驗批次	因素		
	F1	F2	F3
1	0	0	0
2	0	1	1
3	1	0	1
4	1	1	0

3.2.2 直交實驗設計

直交實驗設計(Orthogonal Experimental Design, OED)，實驗的過程中通常利用假設來減少實驗結果的因素個數，以節省成本的損耗但是又要接近真實的結果，直交實驗設計是為了解決此問題而被提出的方法之一。其主要包含兩個重要的部份：直交表(Orthogonal Array, OA)與因素分析(Factor Analysis)來達到此效果[45-47]。首先透過直交表，產生出獨立且均衡的每一個因素，再藉由每一個因素分析出的主效果(Main Effect)，由主效果推論每一個因素對於該實驗結果的優劣。因此以直交實驗來解最佳化問題時，問題的一個參數可視為直交實驗中的一個因素，而參數視為因素的水準(Level)值。完全因素實驗(Complete Factorial Experiment)會以全部水準值的排列組合進行實驗，而 OED 僅取全部排列組合中的一部份來進行分析實驗，也就是部份因素實驗(Fractional Factorial Experiment)，因此直交陣列系統推理化的特性只需進行部份因素實驗就可以推測出所有搜尋空間中最佳的近似解(Near Optimum)，可節省大量執行的時間。

	f1	f2	f3	f4	f5	f6	f7	y(fitness)
1	1	1	1	1	1	1	1	Y1
2	1	1	1	2	2	2	2	Y2
3	1	2	2	1	1	2	2	Y3
4	1	2	2	2	2	1	1	Y4
5	2	1	2	1	2	1	2	Y5
6	2	1	2	2	1	2	1	Y6
7	2	2	1	1	2	1	1	Y7
8	2	2	1	2	1	1	2	Y8
Sj1	Calculate Sj1							
Sj2	Calculate Sj2							
$\|Si1 - Si2\|$	Calculate $\|Si1 - Si2\|$							
Factor rank	Rank the $\|Si1 - Si2\|$ value							

圖 10 為直交實驗設計

例如：f1 的$|S_{j1} - S_{j2}|$計算 S_{j1} 的值為 $Y_1+ Y_2+ Y_3+ Y_4$ ，S_{j2} 的值為 $Y_5+ Y_6+ Y_7+ Y_8$ 的計算出相對應的$|S_{j1} - S_{j2}|$值，然後以此數值大小決定此因素的影響力。

3.3 智慧型基因演算法

為解決得到最佳的 window size 組合投票結果符合真實的實驗結果，此為參數最佳化問題，必須使用最佳化演算法幫助解決此問題，在此本文使用智慧型基因演算法(Intelligent Genetic Algorithm, IGA)[48]做為最佳化的工具，此最佳化演算法具有收斂速度快，精確度高的優點。

智慧型基因演算法與傳統基因演算法最大之不同，乃是以智慧型交配取代一般的單點交配或多點交配，因為傳統基因演算法的交配方式無法評估染色體中參數個別的優劣，且交配點是由隨機方式產生，得到的後代染色體品質不容易有顯著地提昇。此智慧型基因演算法將染色體交配過程視為一種因素實驗(factor experiment)。將來自親代的兩個染色體中已經被切割好將要進行交配的片段，視為直交實驗設計的因素，並以此染色體片段「互換」或「不換」做為兩種水準值(1/0)，以此兩水準直交實驗產生出優良品質染色體的機率便可大幅提昇，進行步驟如下：

1.令產生染色體中的交配點所切割出的基因片段為實驗因素；假設因素數目為 n，即有 n 個基因片段，選擇 $L_\beta(2^{\beta-1})$直交表的前 n 欄作為實驗之用，其中 $\beta = 2^{[\log(n+1)]}$。

2.令因素 j 的水準 1 與水準 2 分別表示來自親代染色體 P_1 與 P_2 第 j 個基因片段。

3.根據直交表，計算各因素組合實驗的評估值 y_t，t =1, 2,….,β。

4.計算主效果 S_{jk}，其中 $j = 1, 2,\ldots,n$，$k = 1, 2$。

5.決定各因素的最佳水準。在評估函數望大時，則選擇主效果值較大之水準；在評估函數望小時，則各因素的最佳水準為主效果值較小之水準。如評估函數望大且 $S_{j1}>S_{j2}$，則因素 j 的最佳水準為 1；反之則最佳水準為 2。

6.根據各因素的最佳水準，選擇對應親代染色體中的基因片段，組合成第一個子代染色體。

7.將各因素的主效果差值($|S_{j1} - S_{j2}|$)排名，差值越大者排名越高。

8.以類似第一個子代染色體的方式來組合因素，將差值排名最差的因素，選擇與第一個子代相反的水準，則可產生第二個子代染色體。

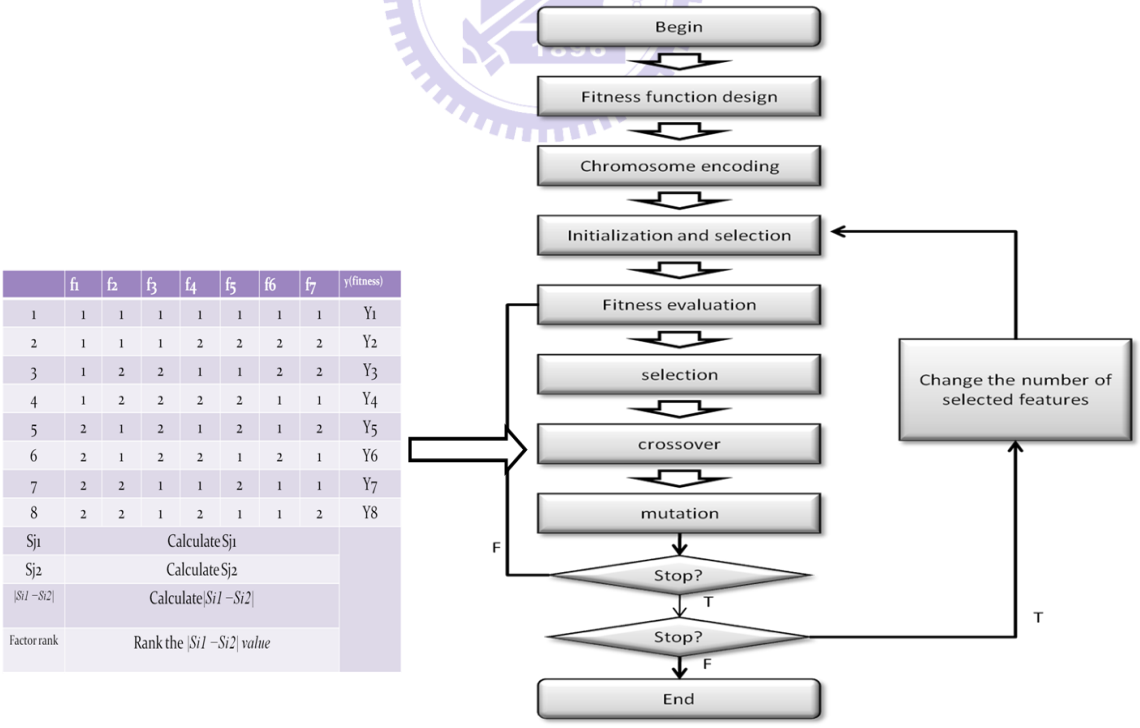


圖 11 表示智慧型基因演算法架構

3.3.1 智慧型交配

智慧型交配運算結合了兩水準直交實驗與交配運算，能夠有效率的產生具有優秀評估值的新染色體。假設我們使用 $L_{N+1}(2^N)$ 直交表來做智慧型交配運算，本染色體交配運算的詳細進行步驟如下：

步驟一：假設即將進行交配運算的兩條染色體為 P_1、P_2，比對 P_1、P_2 內的基因，並將重複出現在兩條染色體內的基因移動至染色體末端，其相對應的控制基因。

步驟二：隨機將基因切割成[$N/2$]個基因片段。每個基因片段及代表直交表的一個因素。P_1 中第 j 個基因片段，及代表因素 j 的第一個水準值；P_2 則代表第二個水準值。

步驟三：計算直交表中每個染色體排列組合方式的適應值 f_t ，$t = 1,2,3,\ldots,N+1$。

步驟四：計算各因素之主效果 S_{jk}，$j = 1,2,3,\ldots N$，$k = 1,2$。

步驟五：決定各因素的最佳水準。在評估函數望大時，則選擇主效果值較大之水準；在評估函數望小時，則各因素的最佳水準為主效果值較小之水準。如評估函數望大且 $S_{j1}>S_{j2}$，則因素 j 的最佳水準為 1；反之則最佳水準為 2。

步驟六：根據各參數的最佳水準，選擇對應親代染色體中的參數，組合出第一個子代染色體。

步驟七：將各參數的主效果差值排名，差值越大者排名越高。

步驟八：以類似第一個子代染色體的方式組合參數，除了盤明最差的參數選擇的水準與第一個子代染色體相反，可產生第二個子代染色體。

3.3.2 突變運算及演化終止條件

突變運算使用任意合理之亂數運算。假設一條染色體編碼之總長為 N 個位元，突變率為 P_m，則每次的突變運算，隨機由染色體中選出「$N*P_m$ 」個位元，然後以亂數變化所選中的物化性質參數。

基因演算法所需的演化時間必須視問題的複雜度而定，較一般化的作法是設定演算法的評估次數與問題中的參數數目成正比。一但終止條件達到設定值，演化即停止，並輸出所搜尋過的最佳解、用最佳解當模型參數。

3.4 繼承式雙目標基因演算法

針對欲選擇一組最小的可提供資訊的特徵向量(informative feature)，但是又要最大化預測的準確度此為一個雙目標最佳化的問題。一個有效率的繼承式雙目標基因演算法(Inherit Bi-oBjective Genetic Algorithm, IBCGA)被[49]我們用來解決這問題。

IBCGA 包含一個智慧型基因演算法具有繼承式的機制。此智慧型基因演算法使用分割解決問題的策略(divide-and-conquer strategy)及直交表(orthogonal array)針對基因的互換步驟解決大尺度參數最佳化的問題。在這個研究中智慧型基因演算法可以有效地探索及利用搜尋空間 C(n, r). IBCGA 可以搜尋 C(n, $r\pm1$)藉由繼承一個好的解空間 C(n, r)。所以，IBCGA 可以獲得一組高品質的解在單一批次實驗下，在這裡 r 是指一個我們有興趣的範圍內。

染色體編碼：

此被提出來的染色體被編碼成一個二進位的基因用來特徵選取之用，此外此染色體還包括一參數基因用來調整 SVM 的參數，在這裡基因(gene)及染色體(chromosome)即為一般被使用在基因演算法中所描述的。此染色體由 n=531 個二進位的基因所組成，每一個基因(b_i)代表可被選擇含有訊息的性質，此外還包含 4 位元的基因此被用來調整 SVM 的參數 C 及 γ。假如 b_i=0 此第 i 項的性質被排除在 SVM 的分類器中，反之則被包含在內。對於 SVM 的γ及 C 參數編碼為 16 個值 $2^{-7}, 2^{-6}\ldots, 2^{8}$

- 1 , 2, 3, 4 , 5 , 6 , 7…….....,531,532,533

圖 12 為 IBCGA 染色體編碼的方式

此特徵向量對於SVM分類器，被編碼成一個含多個基因的染色體使用下面的步驟。確認一個蛋白質或胜肽序列，然後將此序列轉換成該序列內含胺基酸所對應 aaindex 中物化性質胺基酸的數值，然後使用平均的方式把此相關的指標轉成此染色體中的基因此代表此序列所包含物化性質的數值，由於每一個 aaindex 中所對應不同的物化性質，所以每一條序列可以得到 531 個物化性質的特徵向量。然後最後使用正規化的方式把以上得到各染色體的特徵向量正規化成數值[-1, 1]之間。

然後此適應函數針對 IBCGA 得到令人滿意的解，此外選擇適當的參數對於 SVM 在 10 倍的交互驗證下(10-CV)，因此此最終染色體被獲得，IBCGA 使用適應函數 f(X)可以同時獲得一組解 Xr 在這裡 r=r_{start}, r_{start}+1, …, rend 在一次運算中同。此 IBCGA 演算法被給予 r_{start} 及 r_{end} 依照下列描述：

步驟一：起始，隨機產生一起始族群 N_{pop} 所有染色體中 n 個二位元的基因具有 r 1's 及 n-r 0's 在此 r = r_{start}。

步驟二：評估此適應函數值對於個體使用 f(X) 。

步驟三：選擇傳統的競爭選擇其在隨機選取的兩個中，選擇較優良的自交配池中。

步驟四：選擇 $p_c \cdot N_{pop}$ 親代自交配池中執行直交表互換在選擇一對親代當中，在此 p_c 為互換的機率。

步驟五：隨機選擇 $p_m \cdot N_{pop}$ 應用在互換的個體進行突變的操作，在此 p_m 為突變的機率。為了避免最好的適應函數值被破壞，此步驟不針對最好的適應函數值個體進行。

步驟六： 終止條件，假設停止條件對於獲得解 X_r 被滿足，輸出最好的個體 X_r。否則回到步驟二。

步驟七：繼承是假如 $r < r_{end}$ 隨機改變一個位元在二位元的基因上對於每一個個體由 0 到 1，增加數量對於 r 為 1 然後繼續步驟 2 否則停止演算法。

3.5 向量學習機

向量學習機(Support vector machine, SVM)，是一個學習的模型處理兩種分類的問題。SVM是藉由找出一個超平面(hyperplane)，使之將兩個不同的集合分開使用最大距離在兩個向量組成之間[50]。為了使要本更容易地分離開來，SVM使用各種功能的kernel來轉換樣本至更高維度的空間或是特徵空間(feature space)。在這邊我們使用常用的radial based function (RBF)當作我們投射樣本的kernel。此RBF為一非線性的Kernel，定義如下面所示：

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

此 kernel 參數 γ 決定樣本如何轉換到更高維度的空間。而成本參數 C(cost)為給有誤差的資料一點懲罰，讓它多一點成本，亦即 C 是我們決定要給多少懲罰的權重。此兩個參數 C 及 γ 必須被調整找到最佳的值在得到最好的預測效益上。一般的做法是把訓練資料拆成兩個部分，一個部分用來訓練，另一個部分用來驗證準確度，若準確度不夠的話，換參數再做一次。在這部份我們是利用 IBCGA 在利用此 SVM 分類器情況下，同時選取 feature 下幫助選取此 feature set 的最佳的參數值。

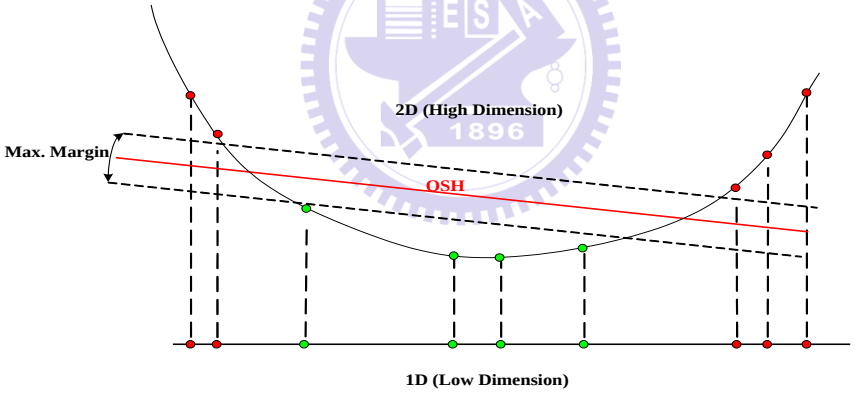


圖 13 為二維 SVM 概念示意圖

四、資料的建構及物化性質指標

4.1 使用資料的建立

在這個研究中，我們使用了三個資料集，分別為口蹄疫病毒資料集、獨立測試資料集以及 EL-Manzalawy 資料集(FMDV-dataset、independent test dataset、EL-Manzalawy dataset)分別做為免疫模型建立、獨立的測試資料及用來評估演算法的資料各自地。其中口蹄疫病毒資料集及獨立測試資料集由網路上的免疫資料庫(IEDB)及相關研究文獻中收集得到，EL-Manzalawy 資料集則是使用 Bcipep 中的資料而來(如表 5)。

表 5、此研究所用的資料集

資料集名稱	序列來源生物體	資料數目(正反應/負反應)	序列長度(分別為最短長度、平均長度以及最長長度)	資料庫描述
口蹄疫病毒資料集	口蹄疫病毒	806(214/592)	4,9.76,50	此病原體(FMDV)免疫實驗數據來自於 IEDB
獨立測試資料集	口蹄疫病毒	118(83/35)	6,15.87,41	由此病原體研究所得到的數據及 IEDB 得到(此不包含於口蹄疫病毒資料集中之序列)
EL-Manzalawy 資料集	數種不一樣病原體範圍的表頂，包含口蹄疫病	1868(934/934)	4,14.19,56	此為降低同源性的資料集由 EL-Manzalawy et al 自 Bcipep 得到

4.1.1 口蹄疫病毒資料

此病毒主要分為七種血清型，各血清型之間無交互保護力或是其保護力很有限，在此我們不欲針對每一個血清型做詳細的分類，而是採取找出可能的共同的物化性質針對病毒抗血清辨識特異性序列上。找出抗血清抑或單株抗體中，所能辨識出病原體上抗原決定位的物化性質是我們的目標。於此我們自 IEDB 資料庫提取出相關 B 細胞表頂的實驗資料，包含所有有關口蹄疫病毒此病原體生物體，但不包含非此病毒病原體上的序列。此資料集被建立用來建構我們想要的免疫模型。

我們在此研究中使用此病原體抗原上能與此病毒抗血清在免疫實驗上產生反應的序

列當作我們免疫正反應資料(positive data)為表項的資料。另一方面，先前在免疫研究中針對 B 細胞表項的研究使用分類方式來建立免疫的模型，其使用表項的部分是基於抗體抗原交互作用免疫實驗上有反應的序列，而非表項部分是利用結構的訊息選擇非表面蛋白的部分當做訓練的資料，有文獻指出此並非真正的負反應資料(negative data)[15]。在此我們同樣地使用基於實驗所得到的結果，使用此病原體中經由免疫實驗無法與此抗血清產生特異性結合的當作我們的負反應資料為非表項資料，在這樣的設定之下，我們自 IEDB 此重要的免疫資料庫中取出我們的訓練集來建立我們此病原體相關的 B 細胞表項免疫模型。

FMDV 資料篩選及建立步驟如下：

我們依照 IEDB Curation Manual 2.0 所描述來獲得我們的訓練資料避免錯誤的擷取資料。來源生物體(Source Organism)表項的識別是基於病毒抗血清或單株抗體來辨識此病原體序列。由 Browse 選項由來源生物體選項(by source organism)，進入後可以得到一個生物體分類的樹狀表，此表為生物學上的分類，其由上而下的層次依照分類學的次序，依照對於口蹄疫病毒分子生物學背景的瞭解，使用此表選擇依序由病毒-單股 RNA 病毒-單股正股 RNA 病毒-小核醣體病毒-小核醣體病毒科-鵝口瘡病毒屬-口蹄疫病毒。
(ssRNA viruses - ssRNA positive-strand viruses - Picornavirales- Picornaviridae - Aphthovirus - Foot-and-mouth disease virus)，選擇 Foot-and-mouth disease virus 此項為與口蹄疫病毒相關的免疫資訊。

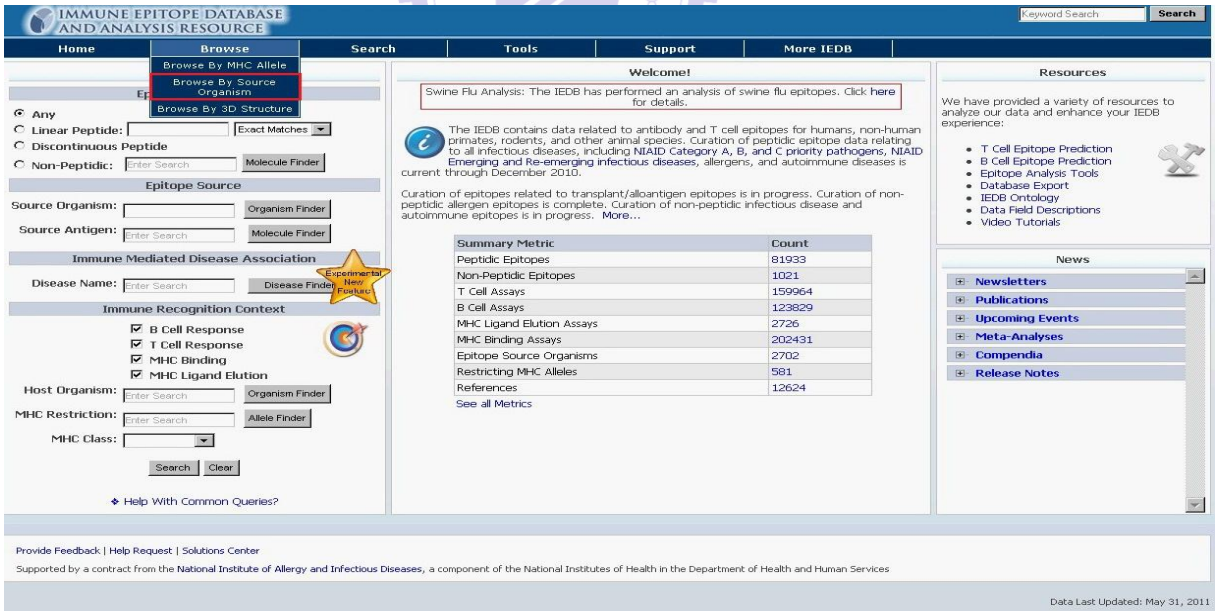


圖 14 為 IEDB 網站

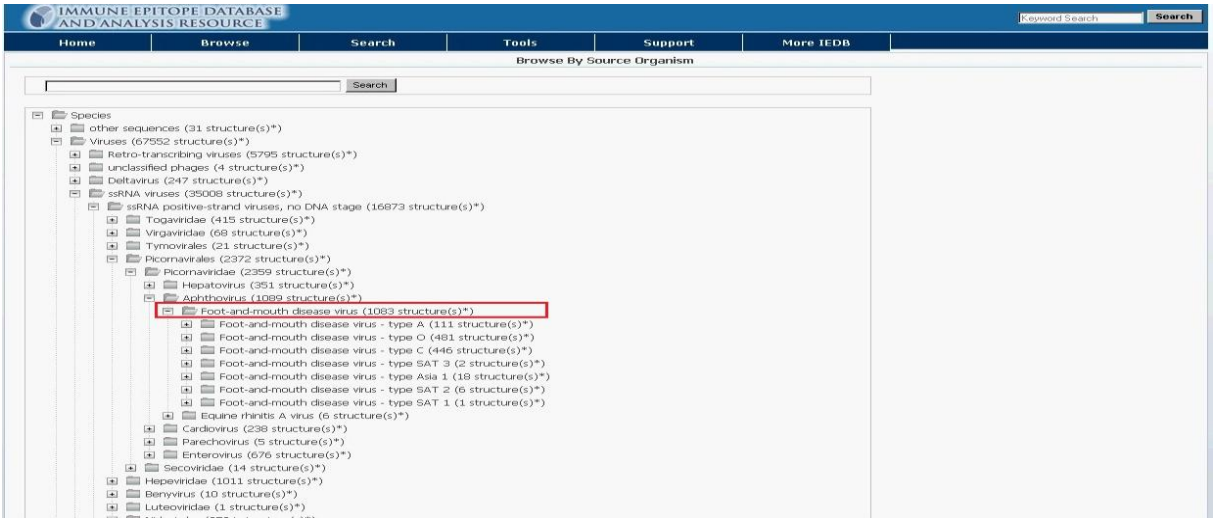


圖 15 為分類學上的層級

進入此 FMDV 病毒選項後，此與 FMDV 相關免疫研究資料項目包含 structure(1083)、reference(140)、source antigen(75)、MHC Binding(21)、B cell arra(2540)及 T cell array(590)，每一項目後面的數字代表其所包含資料數量。標示紅色的部分為我們所主要使用的資訊部分。分別簡單說明其所包含的內容 structure 代表此病毒免疫實驗所代表的序列此序列有一個獨一無二的表頂序號(epitope ID)，選擇 structure 可以看見此但是此獨一無二的表頂相關的資訊，此表頂可能來自不同的免疫接受器所定義出來的表頂，包含 T 細胞、B 細胞及 MHC Binding，例如：若為 T 細胞免疫陣列所辨識具有實驗上正反應我們可以稱之為 T 細胞表頂，每一表頂可能具有多種特性，若選擇進入某表頂 ID 後，可查閱其為何種免疫實驗所鑑定出來的表頂以及參考文獻。Reference 則是此資料項目相關文獻的彙集。source antigen 則是這個資料庫中用來實驗用的病原體不同的品系(strain)數目。MHC Binding(21)、B cell arra(2540)及 T cell array(590)則各是用不同的免疫接受器來定義出相關抗原的項目。

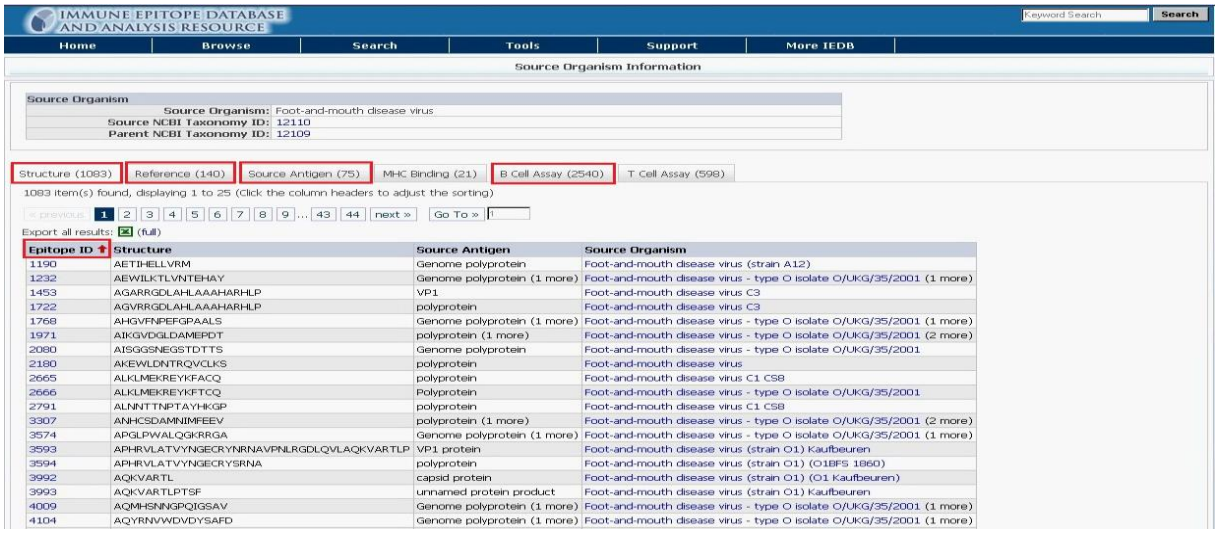


圖 16 為主要使用的資料項目

由 structure 的 Epitope ID 選項進入，我們可以看到有幾個重要的訊息，包含 source accession 此代表序號為此 FMDV 在 NCBI 中的序號、B cell assay 在 B 細胞微陣列實驗中與此序列有關的實驗數目、3D 同源結構、source antigen 指在此病原蛋白質為其蛋白質體上的那一部分，source organism 則是此病原體名稱及品系，reference 則是此序列所參考的文獻。

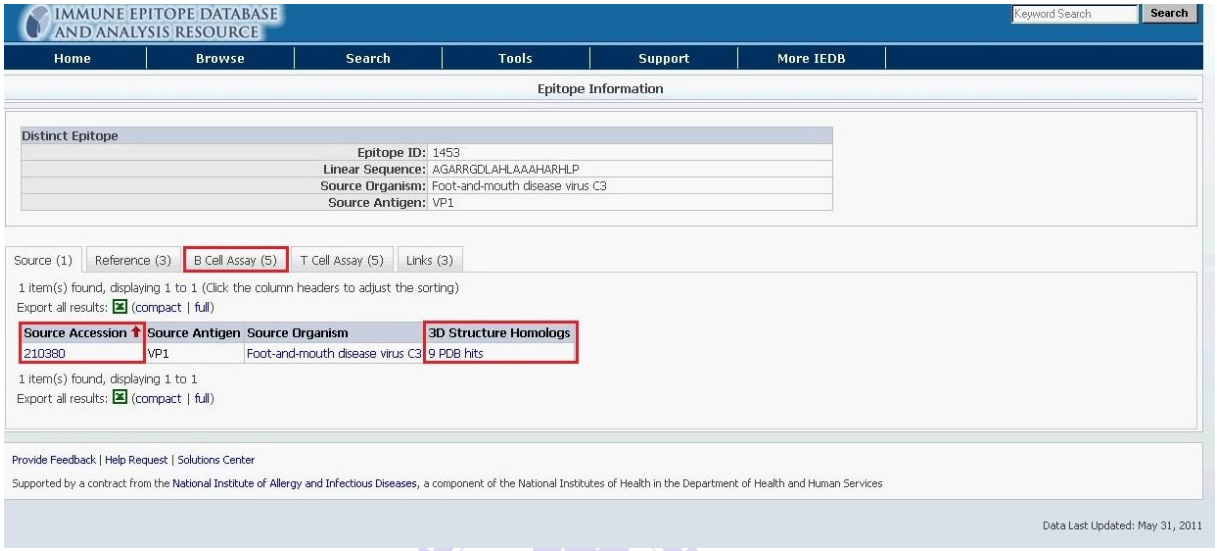


圖 17 為進入 Epitope ID 的項目

由下圖可以看出進入 Epitope ID 的項目選擇 B cell assay 可以看到多個實驗結果每個 B cell 實驗均有一個 B cell ID，在這邊我們可以看到此表頂實驗結果均為正值，但是在其它表頂可能有正有負，在此我們假設只要多次實驗中有正反應的實驗其為具有表頂可能性，我們把此序列歸為正反應的資料在我們的訓練的資料中。

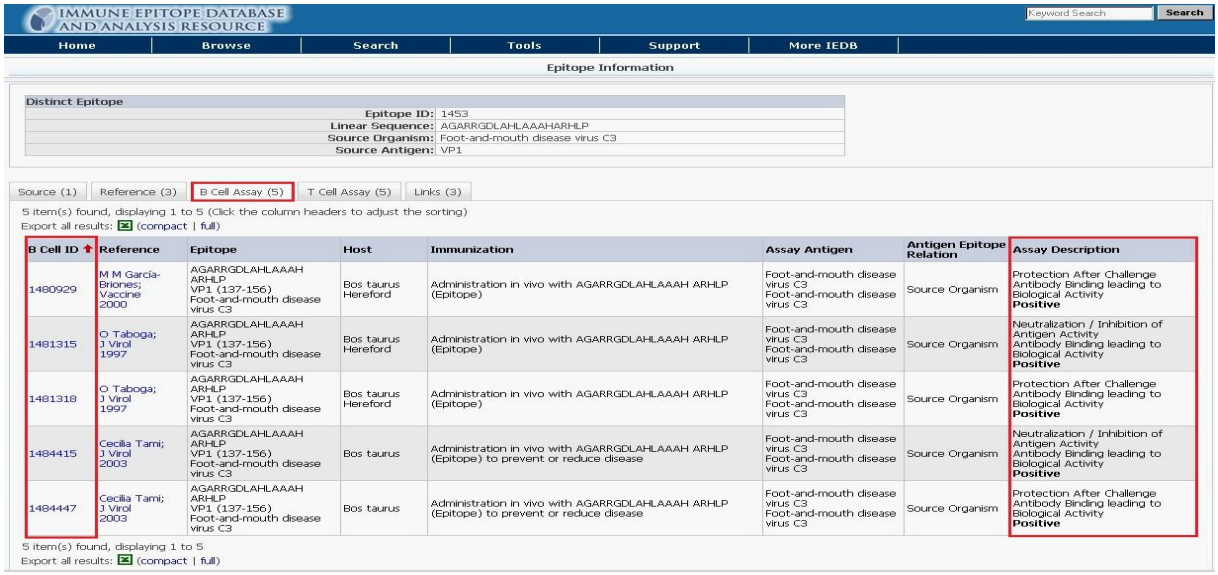
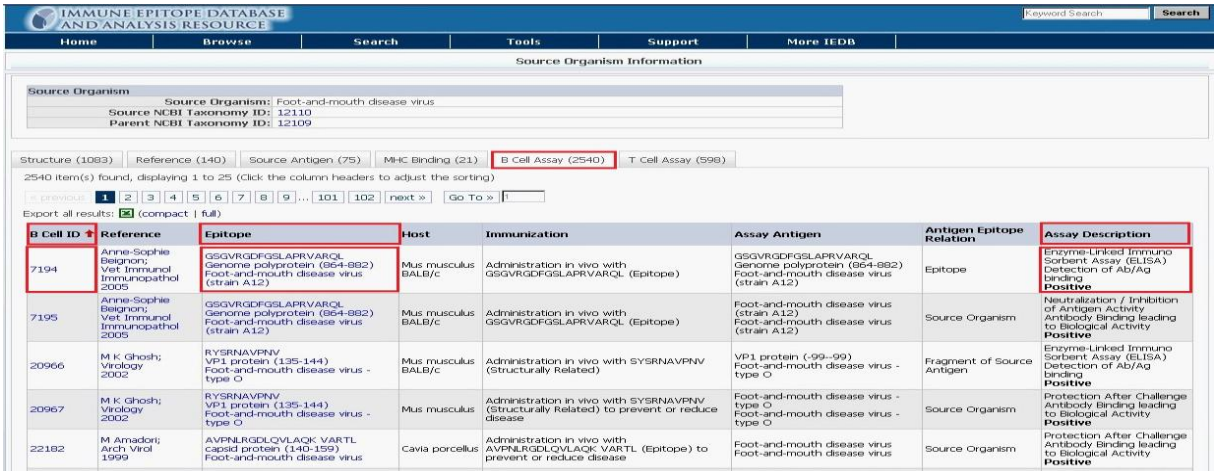


圖 18 為 Epitope ID 內含的項目

此外也可以直接進入 B cell assay 中由 B cell ID 去得到我們所要的資訊，但是要移除重複的序列資訊及確認實驗的結果。



B Cell ID	Reference	Epitope	Host	Immunization	Assay Antigen	Antigen Epitope Relation	Assay Description
7194	Anne-Sophie Beignon; Vet Immunol Immunopathol 2005	GSGVRGDFGSLAPRVARQL Genome polyprotein (864-882) Foot-and-mouth disease virus (strain A12)	Mus musculus BALB/c	Administration in vivo with GSGVRGDFGSLAPRVARQL (Epitope)	GSGVRGDFGSLAPRVARQL Genome polyprotein (864-882) Foot-and-mouth disease virus (strain A12)	Epitope	Enzyme-Linked Immuno Sorbent Assay (ELISA) Detection of Ab/Ag binding **Positive**
7195	Anne-Sophie Beignon; Vet Immunol Immunopathol 2005	GSGVRGDFGSLAPRVARQL Genome polyprotein (864-882) Foot-and-mouth disease virus (strain A12)	Mus musculus BALB/c	Administration in vivo with GSGVRGDFGSLAPRVARQL (Epitope)	Foot-and-mouth disease virus (strain A12) Foot-and-mouth disease virus (strain A12)	Source Organism	Neutralization / Inhibition of Antigen Activity Antibody Binding leading to Biological Activity **Positive**
20966	M K Ghosh; Virology 2002	RYSRNAVPNV VP1 protein (135-144) Foot-and-mouth disease virus - type O	Mus musculus BALB/c	Administration in vivo with SYSRNAVPNV (Structurally Related)	VP1 protein (-99--99) Foot-and-mouth disease virus - type O	Fragment of Source Antigen	Enzyme-Linked Immuno Sorbent Assay (ELISA) Detection of Ab/Ag binding **Positive**
20967	M K Ghosh; Virology 2002	RYSRNAVPNV VP1 protein (135-144) Foot-and-mouth disease virus - type O	Mus musculus	Administration in vivo with SYSRNAVPNV (Structurally Related) to prevent or reduce disease	Foot-and-mouth disease virus - type O Foot-and-mouth disease virus - type O	Source Organism	Protection After Challenge Antibody Binding leading to Biological Activity **Positive**
22182	M Amadori; Arch Virol 1999	AVPNLRGDLQVLAQK VARTL capsid protein (140-159) Foot-and-mouth disease virus	Cavia porcellus	Administration in vivo with AVPNLRGDLQVLAQK VARTL (Epitope) to prevent or reduce disease	Foot-and-mouth disease virus Foot-and-mouth disease virus	Source Organism	Protection After Challenge Antibody Binding leading to Biological Activity **Positive**

圖 19 為表示 B cell assay 項目

由上所述我們確認我們要收集的 B 細胞表頂資訊，由表頂序號進入，確認每一序列實驗的類型及實驗的反應，擷取我們要的 B 細胞表頂相關資訊。此部分，由 B 細胞表頂序號(B cell ID)去下載其此病原體相關所有與 B 細胞免疫反應相關的序列，此 B 細胞表頂序號是指不同實驗下或不同文獻所鑑定的 B 細胞表頂，因此同一表頂序號序列下可能有不同的實驗數據。此相關的資料中移除冗贅的資料然後留下 845 個 B 細胞表頂資料。由於線性表頂的長度是不明確的，我們移除了只一個胺基酸的序列為 39。最後剩下 806 個獨一無二的序列當作我們的訓練集，此包含了 214 個具有抗體反應的及 592 個無抗體反應的資料，並確認與表頂序號所得相同序列資訊，交叉確認，使用此資料當作我們的訓練集來進行接下來物化特性特徵萃取的工作。此外為了瞭解實驗資料定義出表頂序列的分佈在此病原體蛋白質體上，我們也統計了用來訓練的資料庫中的分佈數量(如表 6)。

表 6 口蹄疫病毒訓練資料分佈

蛋白質體	組成次單元	序列數量
病毒的結構蛋白	VP4	5
	VP2	17
	VP3	14
	VP1	662
病毒的非結構蛋白	2B	108
	2C	
	3A	
	3C	
	3D	

4.1.2 獨立測試資料

為了評估我們所建立的免疫模型獨立的測試是必須的，我們收集了相關研究由免疫實驗方式定義出 FMDV 上之 B 細胞表頂之數據當作我們的測試資料集，此包含兩個部分。第一個測試集，此實驗數據包含 FMDV 結構及非結構蛋白中定義出的 B 細胞表頂，共 50 條相關的序列，這部分得到的資料是以關鍵字：FMDV、B cell epitope 來進行資料的探索。所以本質上與我們得到的訓練資料一致，但是其建立於我們免疫模型建立之後，目的是為了測試是否可以預測出病原體中具有免疫特性的片段。此外，也收集 IEDB 中研究 FMDV 但是非天然來源序列(no natural source)，所得到的資料當做我們第二個測試集子集，此包含 71 個序列我們移除包括非 20 個胺基酸之序列，最後得到 68 條。第二個測試集則是是否能預測出利用此致免疫部位設計的胜肽疫苗是否具有免疫反應(如表 5)。此兩個子集的最大差異在於是否與病原體上的序列完全一致，但是相同地是其序列實驗的結果乃基於病毒的抗血清來確認此是否具有抗原性。

表 7 為獨立測試資料集

抗原來源	數量(正反應資料/負反應資料)		參考資料
合成胜肽	68(57/11)		IEDB
	病毒結構蛋白	病毒非結構蛋白	參考文獻
病原體	VP0~VP3 17 (6/11)	X	[12]
	VP1 10 (8/2)	X	[51]
	X	2C 13(2/11)	[23]
	VP1 (4/0)	3B 1/0	[52]
	VP1 (2/0)	X	[53]
	VP1 (1/0)	X	[54]
	VP2 (1/0)	X	[55]
	VP1 (1/0)	X	[56]

4.1.3 病原體中表頂位置的資料

欲針對 FMDV 使用我們所選出的物化性質來辨識出可能的表頂位置由病原體蛋白質體序列中，因此我們必須自所得到的訓練及測試的資料中延伸相關資訊找出其原始病原體抗原位置的資訊，目的是確認出實驗所得到表頂序列位置的資訊，用來進行進一步地測試並決定我們所使用的投票機制結合智慧型基因演算法。

這個部份可以由我們所擷取出來的表頂來進行資料的收集依照下列的步驟：

1. 首先，藉由表頂序號我們可以得到一組 source accession，此為此表頂針對此病原體株。
2. 由此 source accession 可以連接到 NCBI 原始資料建立的序列，我們依此原則建立一個檔案其包含 Source Molecule Accession(GI)、GeneBank No，並針對實驗上所使用的序列在此病原體蛋白質體上結構部份及非結構部份及對象予以註解。
3. 接下來由表頂序號中的 B 細胞表頂序號註解可以用來找出此段表頂序列是針對此病原體蛋白質體上的那一個位置。
4. 把以上的病原體序列自 NCBI 中下載 fasta 格式，並加上免疫研究位置的註解，最後得到我們所要的序列及實驗的位置。

4.1.4 評估演算法的資料

這部份的資料是想瞭解在綜合的病原體資料下此演算法分類的效益是否有顯著地差異，並比較此資料之下基於SVM分類器的研究比較彼此的分類效率。

由於 B 細胞表頂序列長度非固定，我們使用資料集來自 EL-Manzalawy et al.，此包含彈性長度之表頂(flexible length)。此 EL-Manzalawy et al.所包含的 B 細胞表頂之序列來自於 Bcipep。基於此資料庫提取的資料集來比較分類的準確度與相關研究中，評估所使用演算法效益，此彈性長度的資料及建立的準則如下：

1. 在Bcipep中所有與B細胞表頂相關的當作正反應資料，但是要取大於4個胺基酸長度的長度的資料因為大部分表頂長度範圍在此資料庫中為4~20內得到1223條序列。
2. 移除重複的序列及降低其相似性54至80%下，因此最後剩下934條序列。
3. 隨機自SwissProt選取等量於正反應資料當作負反應資料。
4. 最後得到1868(934/934)條序列。

以上所描述的部分來自此Bcipep資料庫引用的文獻中處理資料的方式[40]。此為我們EL-Manzalawy dataset的資料取得方式。

4.2 物化性質指標

針對此一目的分析了 B 細胞表頂胜肽(peptide)片段胺基酸的物理化學性質(physicochemical properties)，根據這些特性建立規則，設計分析 B 細胞表頂發生位置的預測程式，預測沒有實際實驗資料驗證的病毒蛋白的 B 細胞表頂位置，提供研究人員進一步的參考資訊，協助加速病毒疫苗研究的進展。

我們使用 AAindex 當作我們的物化特性指標，此胺基酸指標資料庫 Amino Acid Index Database (AAIndex)，包含數以百計的各種物理化學性質及生物的胺基酸性質，每一個指標表現出一種胺基酸的性質其使用數位矩陣的型式呈現此性質。

在最新版的 AAIndex 9.0 包含 562 個指標。這些物化性質包含分子量(molecular weight)、疏水性(hydrophobicity)、親水性(hydrophilicity)、水合潛力(hydration potential)、易接近表面面積(average accessible surface area)、自由能轉移(free energy transfer)、彈性(flexibility)、殘基體積(residue volume, mutability)、熔點(melting point)、光學活性(optical activity)、側鏈體積(side chain volume)、極性(polarity)及等電點 isoelectric points。在此我們在這些物化性質指標中移去缺少 20 個胺基酸的部分，最後剩下 531 個物化性質，使用這些物化性質當作我們的物化特性指標[57]。

五、免疫模型建構

此章針對本研究中所收集之材料及如何應用最佳化演算法建立免疫模型做一詳細的說明，並描述我們選取之物化性質所建立的模型與相關研究比較的準則，接著說明如何使用我們所選出的物化性質來定義出表頂位置自病原體中以及其評估的方式。最後討論我們所分析物化性質的準則。

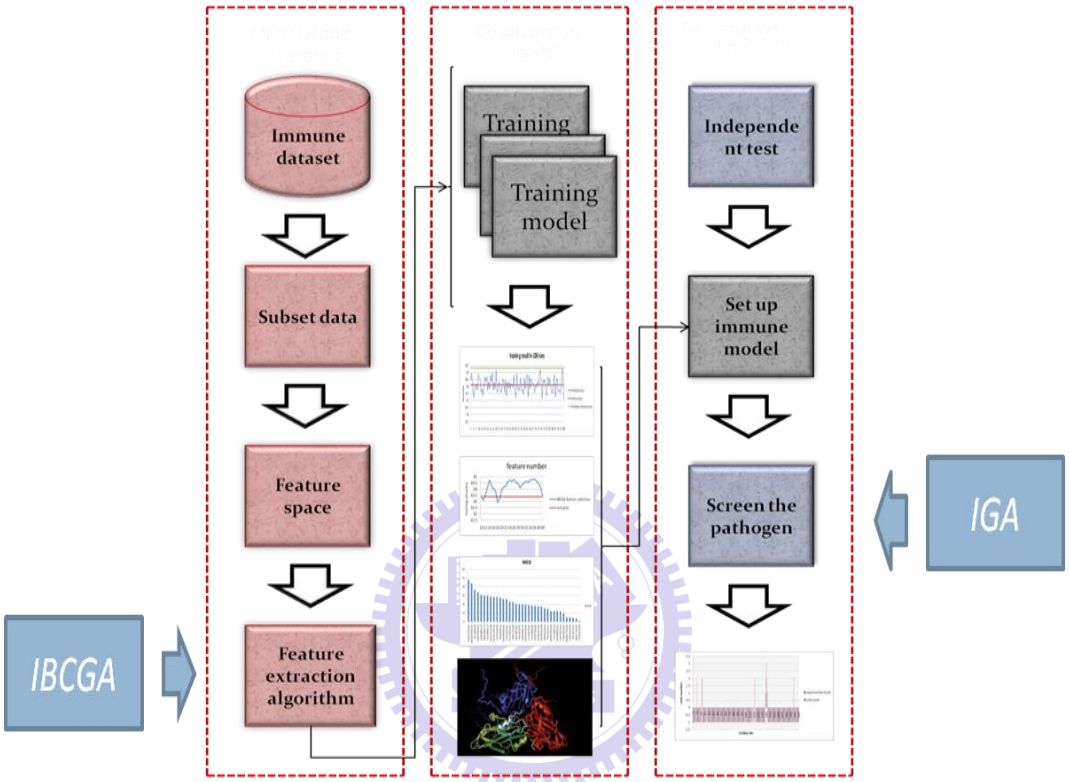


圖 20 表示本研究最佳化演算法使用的部分

5.1 免疫模型建構方法及評估的方法

在這個研究中我們利用由免疫資料庫所得到的由此口蹄疫抗血清所定義出病毒病原體蛋白質中表頂的部位，當作我們的訓練資料，依照下列步驟建立我們的免疫模型：

（1） 將所得到的序列資料依註解 positive 定為 1 類，negative 定為 0 類。

（2） 把所得到的訓練資料轉成 aaindex 中的 531 個物化特性。

（3） 由於每一物化特性值轉換後大小不一，接著將其利用正規化公式轉換每一物化特性值為-1~1 之間。

（4） 以上前處理完後，使用繼承式雙目標基因演算法幫助選取重要的物化特性，設定的參數如下: $N_{pop} = 30$、$P_c = 0.8$、$P_m = 0.05$、$r_{start} = 40$ 及 $r_{end} = 10$ 及 10 倍的交互驗證下(cross-validation)。

（5） 在以上的相同的設定之下，跑 100 批次實驗，最後選取訓練準確度最高的當作我們的免疫模型。

我們認為在最好的訓練結果中所代表的含意，是指在使用機械學習方式分類上，此組特徵集合可以正確地幫我們分辨出此基於實驗所的訓練資料中表頂及非表頂的序列。另一方面，由於高的準確度(accuracy)其所代表的物化性質，我們合理的相信可以更加接近 FMDV 病原體中 B 細胞表頂所蘊含的物理化學性質。

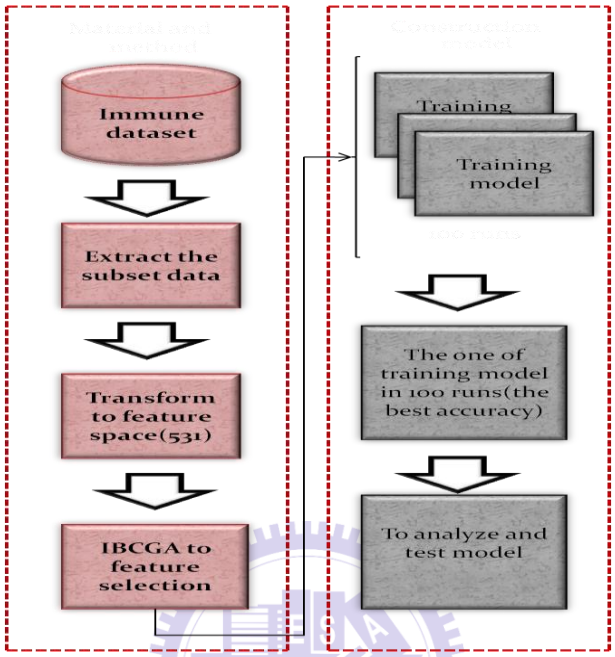


圖 21 利用 IBCGA 結合物化性質建立免疫模型示意圖

依照所得到訓練的免疫模型，依照得到的值計算下面的值，可做為評估比較的標準：

敏感度 (sensitivity or true positive rate, TPR)：$TPR = TP / P = TP / (TP + FN)$

準確度 (accuracy, ACC)：$ACC = (TP + TN) / (P + N)$

特異性 (specificity or True Negative Rate, TNR)：$SPC = TN / N = TN / (FP + TN) = 1 - FPR$

Matthews correlation coefficient (MCC)：$MCC=(TP*TN-FP*FN)/(PN*P'N')^{1/2}$

此外為了比較我們基於 B 細胞表頂次分類群所萃取出重要的物化性質所建立的模型之實際效益，我們使用獨立測試資料集來測試建立模型的效益，並與其它兩篇相關研究基於不同病原體表頂範圍使用分類的方式所建立的預測 B 細胞表頂的模型做比較 (BCPred 及 FBCPred)。

5.2 自病原體蛋白質序列中定義出表頂位置及評估的方法

此研究的另一個重點為，利用我們所挑出病原體中重要的物化性質來幫助定義出病原體上表頂的位置。由先前研究中可以瞭解欲定義某病原體蛋白質序列中表頂的位置，一般使用與表頂相關的物化性質來定義出來，但是有研究指出其只略佳於隨機選取，因為表頂所具有之物化性質無法以單一物化性質定義出來，也有研究使用多種物化性質來

交互比對定義出表頂的位置，然後利用訓練的結果決定閾值(threshold)，以此來決定抗原決定位的位置。

在此我們使用所選出的物化性質來定義出表頂的位置自病原體序列中，針對每個目標胺基酸序列在在滑動窗口(sliding window)計算平均性質傾向為表頂或非表頂，然後給予滑動窗口的中心目標的胺基酸殘基此性質尺度的傾向值，由於多種不同的滑動窗口所定義出的抗原位置波動的結果，我們欲從變動的結果中找尋一致性的結論並且降低偽陽性的結果因此在此使用訓練資料中原始實驗位置的資訊結合投票的機制並使用智慧型基因演算法幫助找出最佳的投票組合。

決定病原體蛋白質序列中抗原決定位依照下列的三個步驟來進行：
步驟一：資料的處理，此部份要得到我們步驟二最佳化演算法所需要的格式。

（1） 利用訓練資料中的序列資訊，找到此抗原決定位在病原體序列中的位置，由先前建立訓練資料庫中，自 NCBI 中下載病原體蛋白質序列，切成不同的滑動窗口分別為 3、5、7、9、11、13、15、17、19、21，此滑動窗口範圍大小被決定依照大部份抗原決定位範圍，為小於等於 20 長度大小。
（2） 將切好的序列轉成所選出的物化性質，然後假設每一序列均具有抗原決定位性質。
（3） 將處理完成的序列，使用我們所建立 FMDV 免疫模型來決定其具抗原性質或不具抗原性質。依照假設每一滑動窗口均假設具有抗原性質，其與 FMDV 免疫模型預測相符合給予該滑動窗口中心位點一個值 1，若不符合給予值 0，此時在相同位點上由於不同滑動窗口大小下，其胺基酸殘基組成的差異將造成物化性質的差異，因此在同一位點不同滑動窗口此位點上有不同的性質尺度傾向值(圖 22)。

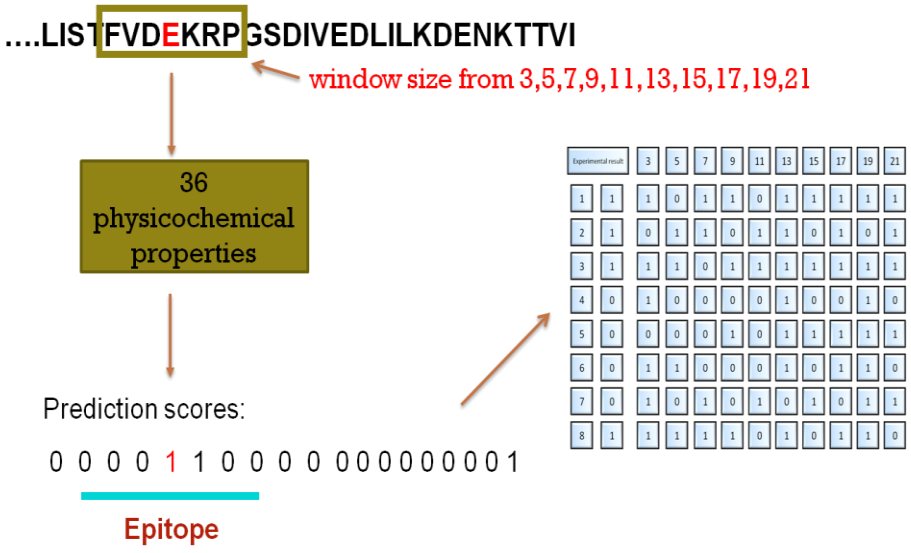


圖 22 為物化性質判斷抗原位點的概念

（4） 然後再依照真實實驗中定義出 FMDV 表頂的位置給每一位置相對應的值，具有抗原性質給 1，不具抗原性及無實驗結果均給予 0，最後可以得到如下列例子的資料格式(圖 23)。

Experimental result		3	5	7	9	11	13	15	17	19	21
1	1	1	0	1	1	0	1	1	1	1	1
2	1	0	1	1	0	1	1	0	1	0	1
3	1	1	1	0	1	1	1	1	1	1	1
4	0	1	0	0	0	0	1	0	0	1	0
5	0	0	0	0	1	0	1	1	1	1	1
6	0	1	1	0	0	0	1	0	1	1	0
7	0	1	0	1	0	1	0	1	0	1	1
8	1	1	1	1	1	0	1	0	1	1	0

圖 23 為處理完成的資料格式
實驗結果左側部分為位置右側為真實實驗的結果
不同滑動窗口的值為其評估物化性質的結果

步驟二：決定適當的抗原位點，我們必須瞭解，在同一位點上所得到抗原性質及非抗原性質物化性質的差異是由於不同滑動窗口下胺基酸序列組成的不同導致的差異，而此差異乃基於此物化性質綜合的結果，為了避免過度變動的結果，我們必須決定適當的方式確認表頂的位置，此外我們欲確認出抗原表頂位置，但是又要減少偽陽性的值，因此我們利用在訓練集中多個實驗結果所定義的位置資訊，以及我們物化性質的組合相近的位點來找出最好的組合，並由多組不同滑動窗口找出最一致性的結果來確認表頂的位置(圖 22)。

（5） 在此我們使用投票的方式來決定哪幾個滑動窗口值之一致性的結果與真實實驗結果相符合，使用智慧型基因演算法幫助決定最佳的組合。

（6） 此智慧型基因演算法的適應函數(fitness function)在此設定為針對陽性預測的值 PVV=TP/TP+NP，針對所得到預測結果與真實結果的一致性為我們此投票結果所需求的。

（7） 我們設定得到投票結果相對高票數的位置，其為表頂的可能性最大，此外若投票得到最高票且其相近位置也都有相對高票我們認為此位置為表頂的機會相對地也很大當作我們的假設。

（8） 由以上結果可以定義出我們要的抗原熱點(hot point)。此位點意味著自多組滑動窗口中決定可能的抗原決定位之位置，並減少偽陽性情況下得到的位點。

步驟三：定義出此重要位點之後，決定適當的表頂範圍，我們依照所決定抗原位點投票的滑動窗口當作我們決定的範圍。

（9） 接下來我們利用此熱點為中心以決定出最大的範圍，將所得到的位點以所決定投票的滑動窗口大小當作我們決定的大小，假設為 5,7,21 看此熱點中滑動窗口投票滑動窗口為 5,7，則此 7 當作我們此部分預測熱點抗原決定位的大小。

（10）依照此方式來定義出序列的範圍，得到抗原範圍可能為 7,21 其中一種以所推得的結果。因此我們由具抗原性的熱點，並依我們投票所使用來投票的滑動窗口提出的當作推論的表頂。

（11）但是依照實驗成本及需求，建議可以利用此熱點的位置延伸其範圍，使用我們決定抗原性質的模型來決定是否選擇的範圍具有抗原性質。

假設此位點彼此具重複的部份但是又無完全重疊，我們視其為獨立的抗原。例如: TGESADPVTTTV 及 TTGESADPVTT，黃色顯示重疊的部分，但是在抗原決定位來說我們不能說這兩個一樣，但是可以說兩個都是。因此很難比較誰的正確在無實驗測試，及無完全一樣的預測結果情況下，原則上符合我們這組物化性質可以視為具有抗原性。

針對評估定義出表頂位置的效能上，在無實驗測試下不易比較預測結果正確與否，在無完全一樣的預測結果情況下，即使有接近相似度的序列仍然無法明確的認定此為病原體上的表頂，但是為了比較其它預測工具定義出來的表頂之優劣，因此在此我們使用相關文獻所提出比較的方式，在使用獨立測試資料中病原體蛋白序列之訊息，比較測試資料中使用不同預測工具定義出表頂之序列，此文獻設定若有 4 個胺基酸殘基相同視為預測正確，若無視為預測錯誤[58]。

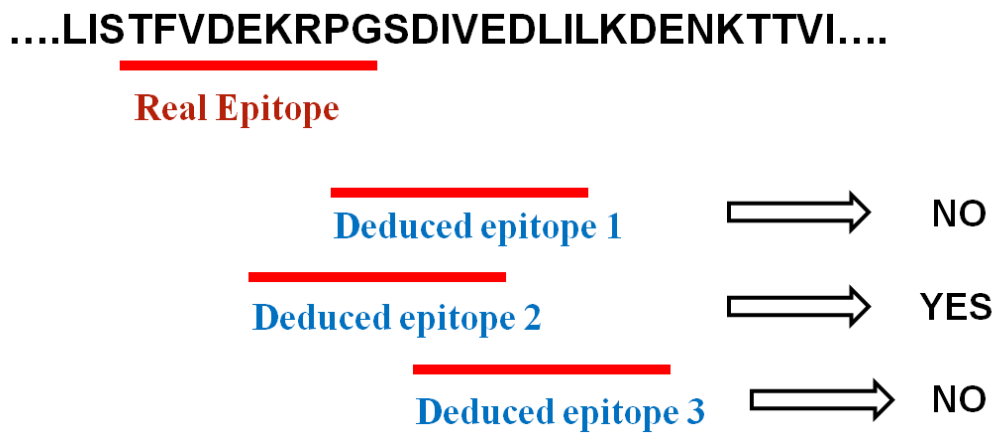


圖 24 為評估定義出表頂位置效能的概念

5.3 物化特性的分析準則

針對使用特徵選取方式找出的物化性質，其所隱含的生物意義是我們所關心的課題之一，對於如何分析出重要的生物意義，我們在此設定幾個分析物化性質的準則來針對我們所選出的物化性質做進一步地分析：

1. 統計100批次(100 runs)下物化性質出現的頻率：這部分針對531個物化性質中被此100個批次所挑出的的物化性質中，物化性質出現頻率大於30的比較的與我們所挑選出的那組物化性質間的關聯性。
2. 使用主效果分析分析(Main Effective Different, MED)各物化性質對於分類上的影響：此主效果分析結果可瞭解每一個物化性質影響此模型分類上的效益的大小，針對分數高的物化性質進行分析。此部分主要利用直交實驗設計幫助分析各物化性質的影響力。

3. 針對各物化性質文獻探討：針對物化性質由其aaindex所提供的data of aaindex1 format所註解的項目，針對D及T項目找出其資料來源文獻針對其建立的準則方法目標做一瞭解，另外由C項目瞭解其它與此值相關的物化性質指標，找大於0.8的值如表。

4. 免疫研究中物化性質討論：由免疫研究所提出幾個重要的物化性質做一通盤討論，比較與我們所選出來的物化性質之異同之處。

5. 病原體表頂研究與物化性質的關聯性：比對FMDV病原體上表頂性質的研究，可以知道我們利用特徵選擇的方式所資料探勘(data mining)出的生物資訊與近年來的研究之關聯性，利用交叉比較的方式，確認出我們的物化特性的意義。

表 8 aaindex 註解格式

(Data Format of AAindex1)

```
**********************************************************************
*                                                                    *
* Each entry has the following format.                               *
*                                                                    *
* H Accession number                                                 *
* D Data description                                                 *
* R LITDB entry number                                               *
* A Author(s)                                                        *
* T Title of the article                                             *
* J Journal reference                                                *
* * Comment or missing                                               *
* C Accession numbers of similar entries with the correlation        *
```

```
*      coefficients of 0.8 (-0.8) or more (less).                    *
*      Notice: The correlation coefficient is calculated with zeros  *
*      filled for missing values.                                    *
* I Amino acid index data in the following order                     *
*      Ala       Arg       Asn       Asp       Cys       Gln       Glu       Gly       His       Ile *
*      Leu       Lys       Met       Phe       Pro       Ser       Thr       Trp       Tyr       Val *
* //                                                                                                 *
**********************************************************************************
```

最後整合上述分析的結論，這部份除了想證明我們的特徵選擇選出的物化性質與相關研究中是相互驗證之外，更進一步地要找出是否有其它重要的生物意義。

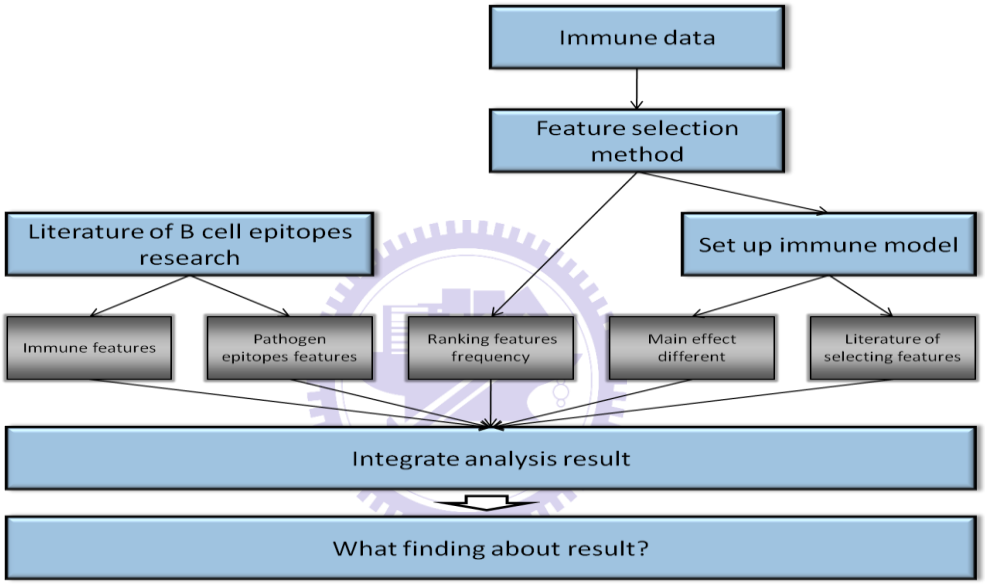


圖25 表示物化性質分析的流程

5.4 評估最佳化演算法效益

這部分我們欲瞭解繼承式雙目標基因演算法選取特徵向量來幫助分類的問題上與其它免疫研究中基於大範圍病原體中的序列在分類上問題上效益如何。雖然我們設定基於 B 細胞表頂的次分類群來建構我們的免疫模型，但是在使用免疫研究中與我們所研究相關問題主要想要比較 IBCGA 選出特徵幫助分類的訓練效果是否一樣有不錯的效益在不同範圍的病原體資料上，比較以 SVM 分類器為基礎的研究之差異，我們選擇免疫研究中針對 EL-Manzalawy 資料集所包含的 B 細胞表頂之序列來自於 Bcipep 做為比較的基礎[40]。

在此設定參數使用 fivefold 的訓練方式來比較分類的效果，使用此彈性長度資料集來進行此最佳化演算法特徵選取及分類的工作，並比較文獻中與我們的訓練結果，選擇最佳的序列結果比較其敏感度(sensitivity)、特異性(specificity)、準確度(accuracy)及 Matthews correlation coefficient (MCC)。

六、結果與分析

6.1 評估最佳化演算法與相關研究之效益

基於此資料庫提取的資料集來比較分類的準確度與相關研究中，評估所使用演算法效益。得到的結果顯示出 IBCGA 分類的效益可以有效的幫助分類上在不同的問題上。

表 9 IBCGA 基於 EL-Manzalawy 資料的分類比較結果

Performance of different method on homogy-reduced dataset of EL-Manzalawy et al flexible length data set using fivefold cross-validation.

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
AAC(2008)	63.31	70.90	55.73	0.269
CTD(2008)	60.32	59.66	60.98	0.206
DC(2008)	63.78	63.05	64.51	0.276
AAP (2007)	61.42	62.85	60.00	0.229
FBCPred (2008)	65.49	68.36	62.61	0.310
IBCGA	70.39	72.37	68.41	0.408

此結果中其他資料的來源，皆為此資料庫相同的序列資料，使用不一樣的特徵(feature)，AAC(amino acid composition)為使用胺基酸的組成當做特徵，CTD (Composition- Transition Distribution)使用 21 個 feature 基於不同長度的序列映射到固定長度的向量，DC(Dipeptide Composition)使用雙胜肽特徵。AAP(amino acid pairs)使用 positive 及 negative 序列組成比例當作特徵，FBCPred 使用 string Kernels。最後得到結果可以顯示出利用 IBCGA 結合物化性質可以幫助分類效率的改善。

6.2 FMDV 預測模型的訓練及測試

在我們使用物化性質指標結合特徵選取的方式來建立免疫的預測模型，使用 FMDV 免疫表項的訊息來幫助建立免疫的模型。在 100 批次實驗中，每一個 IBCGA 訓練資料中設定 IBCGA 的參數為 $N_{pop} = 30$, $P_c = 0.8$, $P_m = 0.05$, $r_{start} = 40$ 及 $r_{end} = 10$ 及 10 倍的交互驗證下(tenfold cross-validation)。對於每一個訓練模型中的特徵數 r，IBCGA 選擇一個 m 的物化性值特徵集(feature set)以及決定 SVM 的參數值。我們最終選擇第 99 組訓練模型來當作我們的免疫模型(如圖 26)。

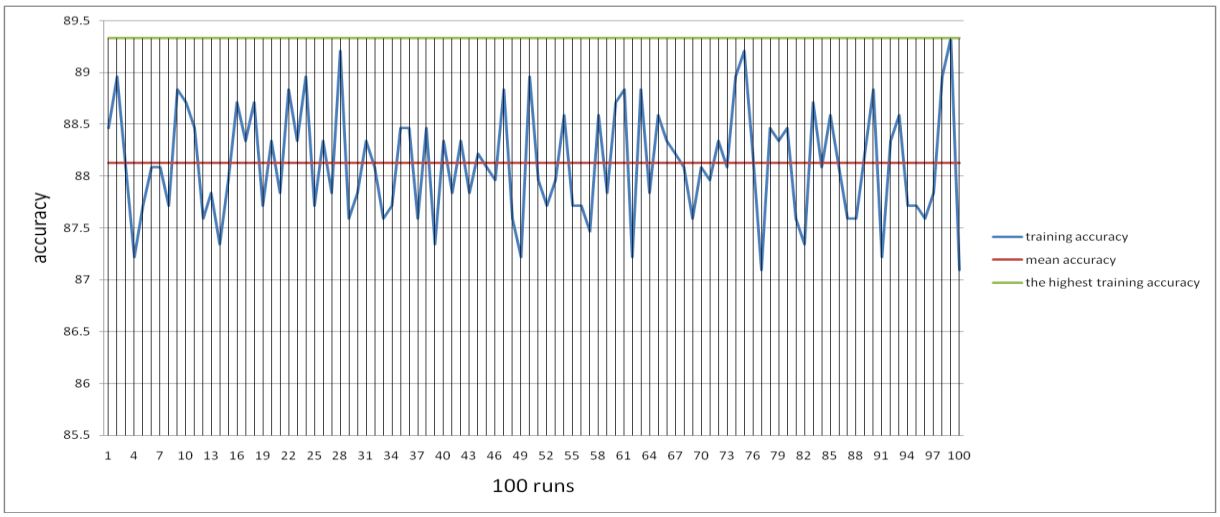


圖 26 在 100 批次實驗下所建立的免疫預測模型

根據上面的敘述在這最高準確度的模型中，最終選擇 m=36 的物化性質特徵集、得到 SVM 的 C=8.00 及 $\gamma = 0.25$，且得到最好的平均準確度(Mean of Accuracies)為 84.83% 以及整體準確度(Overall Accuracy)為 89.33%。雖然另一組特徵集 m=26 也有相當高的準確度達到 84.80%，但是基於前面的設定，選擇最高的訓練集，我們仍然選擇 m=36 特徵集當作我們分類上的物化性質(圖 27)。

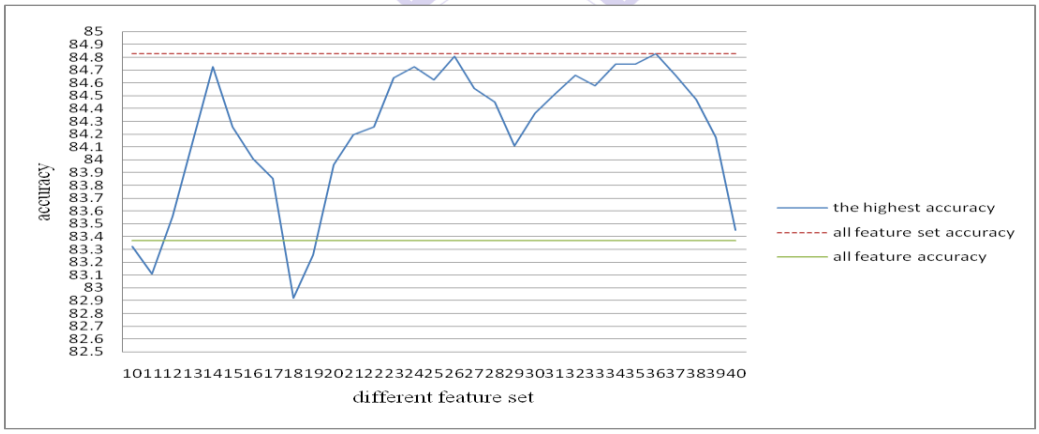


圖 27 最高訓練準確度其物化特性組別

此外可知此模型的敏感度(sensitivity)為 75.23%，特異性(specificity)為 94.43%，Matthews 相關係數(Matthews correlation coefficient, MCC)為 0.720，由此看出對於模型的敏感度低於特異性許多，亦即針對序列正反應的資料的分類準確度上低於負反應的資料。由抗體辨識抗原位置的性質可以略窺一二，既使是一個胺基酸的差異都可能導致抗體無法辨識出抗原表頂的位置，在某些免疫研究表頂位置其序列上有相似性但是僅是幾個胺基酸的差別，這也是預測 B 細胞表頂的困難之處，若基於序列的相似性來預測，可能會陷入高的 FP 值，於此我們利用物化性質的相似性可以降低 FP 值，而且由於此病原體具有高度的序列變異性，其透過突變轉變抗原上的胺基酸來逃避宿主的免疫系統，所以以序列相似性來預測可能無法預測出新品系的抗原決定位。

為了比較我們基於 B 細胞表頂次分類群所萃取出重要的物化性質所建立的模型，我們使用獨立測試集來測試模型間的效益，並且與其他相關研究所建立預測 B 細胞表頂的

模型做比較，我們選擇基於不同病原體表頂範圍所建立的模型來做比較獨立測試的結果，使用我們的所建立的獨立測試資料庫。

此獨立測試資料庫如第二章所描述，其主要包含兩個部分，第一個部分由病原體相關免疫研究文獻所收集而來，其由實驗定義出包含正反應的資料及負反應的資料各為 26 及 24 條序列。利用我們所建立的模型比較與 Bcpred、Fbcpred 基於 Bcipep 資料庫中序列建立的模型[40-41]，包含固定長度及變動長度降低同源性序列。此我們基於 FMDV 免疫資料所建立的模型所得到的測試準確度為 72 %優於 Bcpred 及 Fbcpred 的 31.25%及 45.83%各自地，此外在 FMDV 模型的敏感度與特異性上也都有相當不錯的表現，另外，可以看出在此測試資料下，基於非固定長度的模型 Fbcpred 優於 Bcpred 基於固定長度訓練集的模型。(表 10)

表 10 為獨立測試資料庫針對口蹄疫病毒抗原辨識上

	accuracy	sensitivity	specificity	TP	TN	FP	FN
Bcpred model	31.25%	41.67%	20.83%	11	5	19	15
Fbcpred model	45.83%	62.5%	29.17%	16	7	9	18
FMDV model	72%	70.83%	75%	18	18	6	8

另一個獨立的測試集是基於相關研究中，利用先前針對 FMDV 所設計的胜肽序列，利用免疫實驗測定其是否具有抗原性，可以看出相同地，FMDV 所建立的模型所得到的測試準確度為 61.76%優於 Bcpred 及 Fbcpred 的 48.53%及 48.53%各自地(表 11)。

這部分有趣的是雖然我們可以得到較佳的效果在此部分的比較上，但是卻低於基於病原體抗原上序列所得到的測試集在測試的準確度上，這部分我們試著自相關文獻解釋其差異性，有文獻指出 FMDV 病原體其表頂會受到達爾文正選擇的影響，亦即其抗原決定位的改變來逃避宿主是有其方向性的，其傾向於轉變抗原位置的胺基酸基於相似的物化性質。所以以病原體抗原上序列的測試集其物化性質相似性大於利用化學方式修飾的胜肽序列。但是既使如此，其可能仍然保有一些 FMDV 的表頂抗原的物化性質所以我們測試的結果仍然優於 Bcpred 及 Fbcpred 測試的結果。

表 11 為獨立測試資料庫針對口蹄疫病毒設計胜肽辨識上

	accuracy	sensitivity	specificity	TP	TN	FP	FN
Bcpred model	48.53%	56.14%	9.1%	32	1	10	25
Fbcpred model	48.53%	47.37%	54.55%	27	6	5	30
FMDV model	61.76%	59.65%	72.72%	34	8	3	23

此外由訓練資料的組成可以知道(如表 6)，大部分的資料為抗原的結構區域蛋白 (約 1:7)，所以我們由獨立的測試的資料中想要知道是否有辨識上的差異在於病原體結構及非結構區域的組成蛋白質中，最後得到的結果在非結構區域的測試準確度為 71.42%，結構區域的測試準確度為 73.53%，結構區域的測試準確度略優於非結構區域，但是差異極小，可以得知此免疫模型可以適用於結構區域及非結構區域抗原的預測。

6.3 決定滑動窗口使用 IGA-投票的方式及比較其它的方法

根據我們選出的物化性質，在此利用與訓練集中前面所使用的資料中序列表頂位置的訊息，這部分我們選擇的原則是，在所有 806 個 training data 中我們選取針對不同血清型及針對蛋白質體上不同部位的序列當作我們決定投票滑動窗口的訓練集，此外選擇在一條序列中有多個實驗結果的序列當作我們的訓練資料。最後選擇了 14 條包含 OAC Asia SAT1 血清型及 VP1、VP2、VP3、VP 及 nonstructural part 包含實驗的結果 556/806，有一條編號 ABI16232.1 具有雖然有 187 個實驗結果我們選擇捨棄因為其序列中包含非 20 個典型胺基酸。使用前面所描述決定表頂的方式來決定表頂的位置熱點，最後由所使用的資料，決定使用 5、7、21 當作我們 IGA-voting method 的最終滑動窗口，我們所設定的適應函數之 PPV 值最高 0.559322(TP=33,FP=26)，在三個滑動窗口物化性質綜合結果的情況下，若大於 2 票我們認為此點為抗原性質的熱點，此外在所有此 3 票的時候可以認定此部位具有極高之抗原性質。

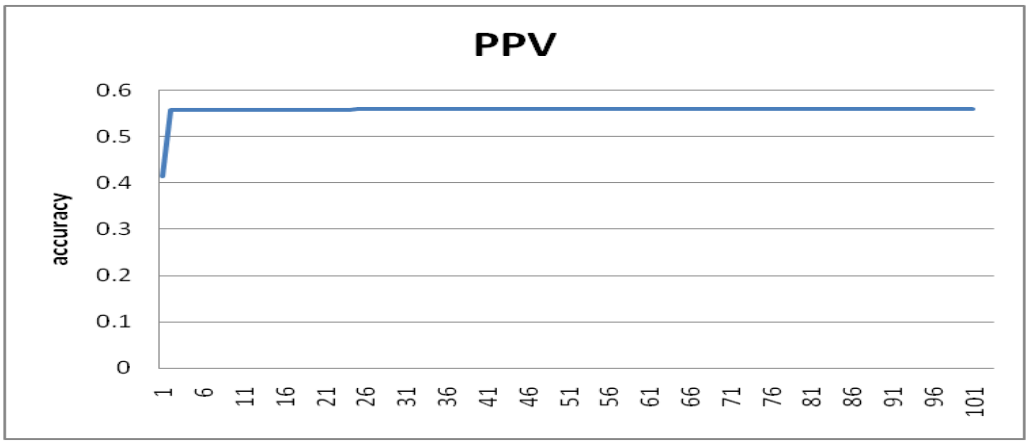


圖 28 IGA 在 100 代內的準確度

下圖顯示出利用我們所定義出的熱點結合實驗結果的示意圖，此示意圖紅色表式投票的結果，藍色表示真實實驗定義出來的位置。橫坐標表示胺基酸位置。縱座標表示票數，若實驗結果及預測結果為非表頂，我們設定顯示為-1 值。實驗結果一律顯示為+1 值，預測結果若為表頂位置則顯示其票數。

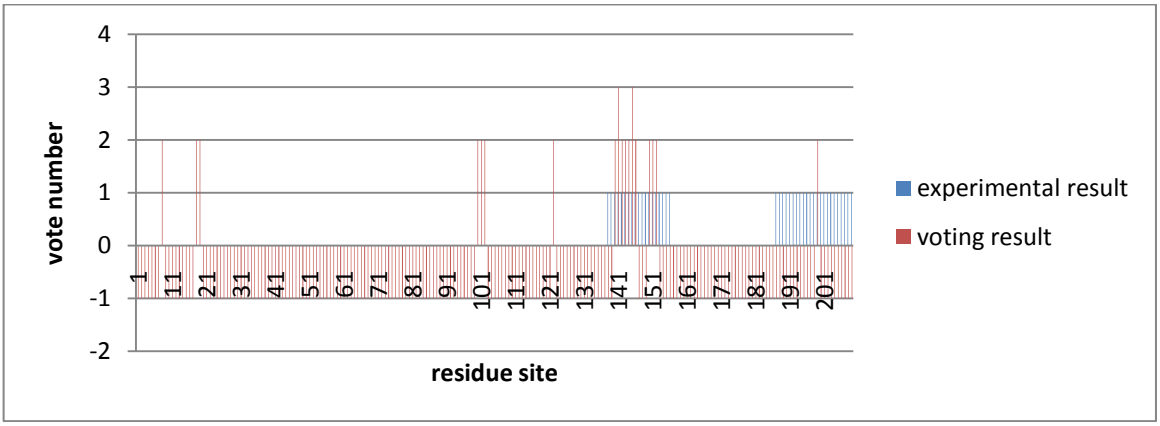


圖 29 使用 IGA 選定滑動窗口定義出的熱點

註紅色部分顯示出我們的演算法所預測的熱點，藍色的部分顯示出實驗上所定義出的位置，其中若低於 0 的位置代表其為抗原位點的機會小，此病原體編號為 no 6318188。

然後可以依照票數決定其表頂的範圍，或以此部位擴張其在實驗上所需要的範圍。例如：最高票位置為 145，我們可以以此為中央點 TYTASTRGDLAHLTATHARHL 擴張至 21 個胺基酸殘基，或選擇 GDLAHLT 擴張至 7 個胺基酸殘基，$_{143}$DLAHLT$_{148}$ 則是為實際實驗的結果。接下來，我們自 PDB 中找出 no 6318188 其結構圖，然後標示出其實驗結果的位置，利用圖形上的註解進行分析比較，可以看出是一個變動性較高的位置，形狀較為不固定且具帶電性胺基酸。

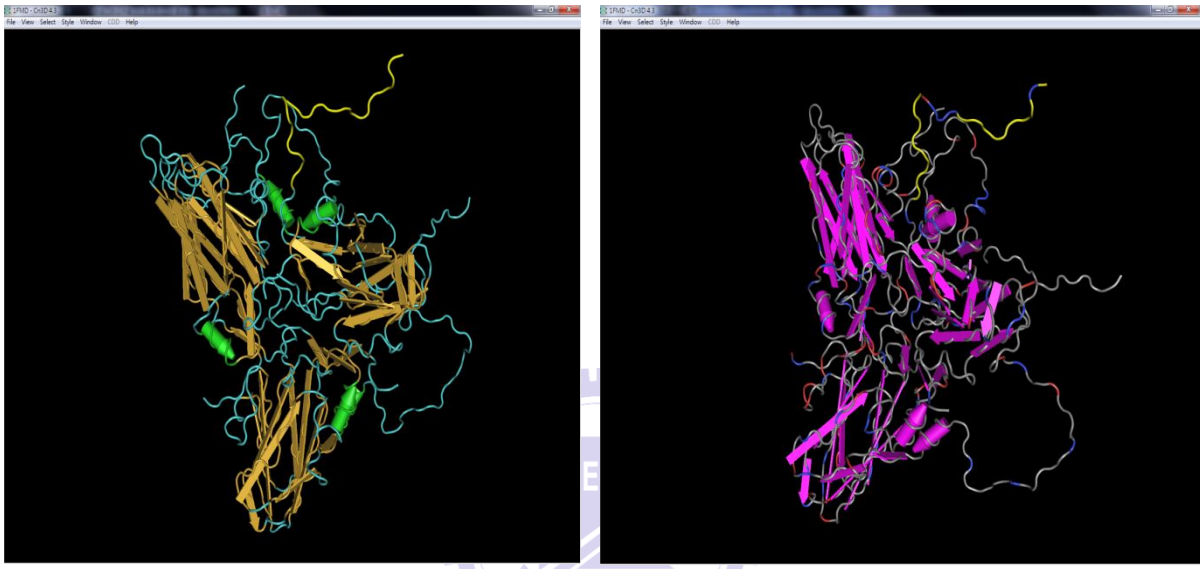


圖 30 病原體編號為 no 6318188 之結構

註：黃色部位表示實驗所得到的抗原決定位的結構位置，使用軟體[59]呈現圖型化

接下來，我們以同樣的結構標出這些被我們使用投票方式定義出的熱點，然後利用圖形上的註解進行分析比較，可以發現大部份我們計算方式找出物化性質定義出來的點在結構上較為靈活的部位為 loop，且其附近帶有帶電性胺基酸，且與實驗上得到的抗原決定位相近。

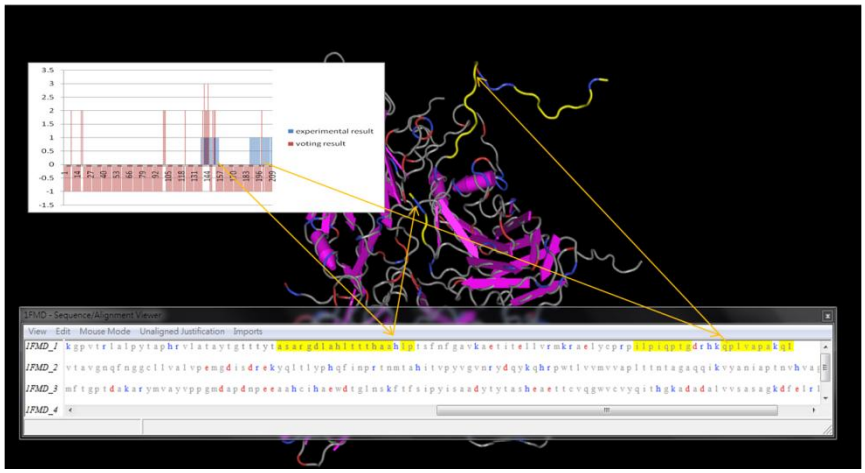


圖 31 為使用呈現病原體編號為 no 6318188 之結構

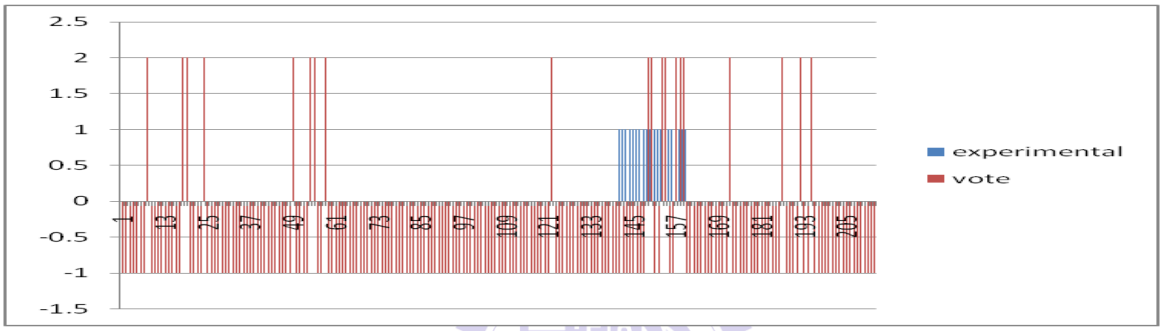
註：黃色部位表示所有熱點位置的結構位置

以上結果其利用由訓練資料中序列抗原位置及抗原性質的資訊所得到的位點及範圍，使用 IGA-投票的方式決定出最好的投票組合。接下來利用我們獨立測試資料中的資料進行測試此演算法定義出來的效果。以下顯示出測試資料中，使用相同的方式所定義的熱點與實驗得到的位置之結果。此為獨立測試集中的序列。

利用視覺化顯示出病原體序號 4038547 實驗的結果在使用黃色標示 VSNVRGDLQVLAQKAERALP，然後將熱點標示紅色大寫部位。

TTSAGESADPVTATVENYGGETQVQRRQHTDISFILDRFVKVTPKDQINVLDLMQI
PAHTLVGALLRTAAYYFSDSELAVKHKGGLTWVPNGAPETALDNTTNPTAHHKAPLT
RLALPYTAPHRVLATVYNGSCKYSDARVSNVRGDLQVLAQKAERALPTSFNYGAIK
ATRVTELLYRMKRAETYCPRPLLAIQPSDARHKQEIVASAKQLL

4038547(VP1) type O



283554648(VP1) type O

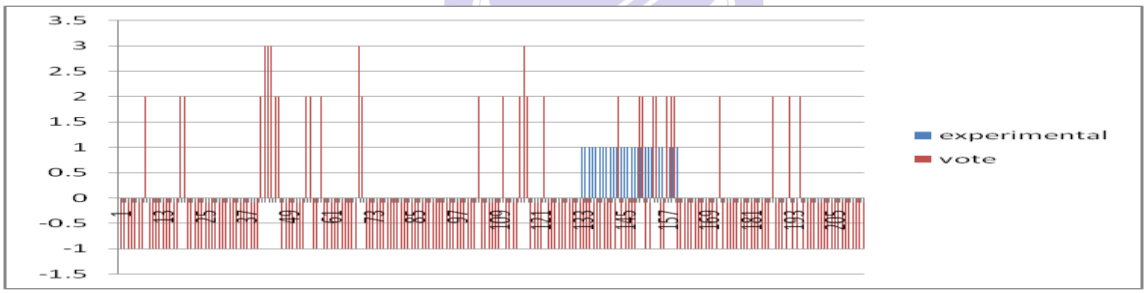


圖 32 獨立測試集中為定義出來的熱點位置

視覺化結構來看可以瞭解，實驗的序列位置為藍色部位，結果顯示出，此序列重疊性很高且其位置在於可變動性較高的位置。以下為此同源結構的註解

Annotated according to the program DSSP. "H" for alpha-helix, "G" for 3-10 Helix, "E" for beta-strand, "T" for turn, "X" for residues not in PDB, space for loop.得知此抗原決定部位是一個 loop。比較我們使用 IGA 投票結果，針對我們研究的對象進行定義出下面的位點，此顯示出與真實實驗結果得到抗原決定位的部分有許多的重疊。另一個熱點的位置的結構特徵為紅色所顯示出的部位，可以看出其為一個變動性較高的 turn 結構。

TTSAGESADPVTATVENYGGETQVQRRQHTDISFILDRFVKVTPKDQINVLDL
MQIPAHTLVGALLRTAAYYFSDSELAVKHKGGLTWVPNGAPETALDNTTNPT
AHHKAPLTRLALPYTAPHRVLATVYNGSCKYSDARVSNVRGDLQVLAQKAE
RALPTSFNYGAIKATRVTELLYRMKRAETYCPRPLLAIQPSDARHKQEIVASAK
QLL

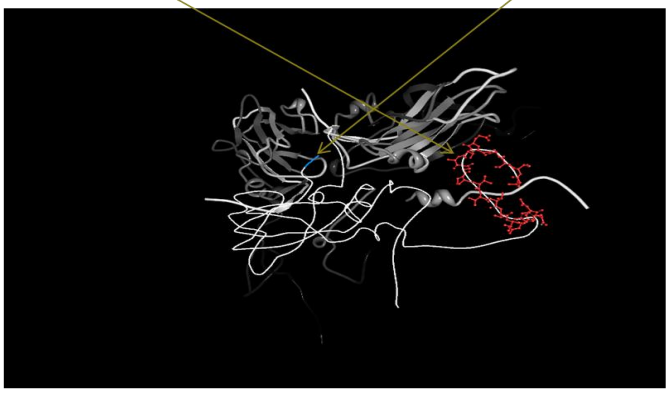


圖 33 為獨立測試集中定義出的位置顯示在同源結構下

由圖可以顯示出，我們所標示出的熱點與實驗所得之結果位置一致性很高，另外若以我們決定滑動窗口的方式，可以得到更接近實驗結果的序列覆蓋率。

另一種，比較方式為利用不同工具預測的表頂結果與我們所預測的結果與真實實驗對照，當做我們的比較方式之一，但是由於實驗結果數量不多我們比較命中率，此指所有預測結果與真實實驗結果相符合的數目。

我們比較四個與免疫預測相關的工具，為 Bepipred、 ABCpred 、BcePred 及 LEPD，分別依照其文獻中最佳參數設定，比較各個預測結果的命中率。下面顯示測試的參數設定:

ABCpred，window length of 16 residues threshold = 0.6

Bepipred， threshold=0.35

BcePred ，default is 2.38 chose hydrophilicity, flexibility, polarity and exposed surface properties

LEPD，without

IGA-Voting，5,7,21 windows size vote

表 12 為比較獨立測試集中不同病原體之序列在使用不同工具之下

	ABCpred	Bepipred	BcePred	LEPD	IGA-Voting
No.4038547	2/19	1/9	0/6	2/10	7/20
No.283554648	3/18	1/9	2/7	2/18	8/32
True positive rate	12.22%	6.3%	17.64%	13.4%	28.85%

由上面的結果可以知道由於各個病原體序列所得到的實驗結果並不多，在我們使用比較的方式中，這個假設預測結果中若無與實驗映射出位置大於等於四個相同的序列，則我們當作為 FP，因此再少的 TP 實驗結果下以及多個預測值下，可以得到如表的比較值，可以看出我們的結果優於其它預測工具。

因此根據我們所研究目標 FMDV 免疫資訊得到的物化性質來決定預測的分類模型，在我們得到的結果中顯示出的確可以較為準確地預測出 B 細胞表頂的位置，或許更進一步地使用這些物化性質來定義出表頂的位置在此病原體蛋白質體序列上，所定義出的熱點與實驗所得的結果所定義出的表頂位置有很大的一致性，顯示出利用特徵選取物化性質的方式綜合物化性質來決定表頂位置可以較為接近真實生物的特性。

6.4 物化性質分析

針對上述結果我們可以瞭解此預測系統基於特徵選取所選出的物化性質能明顯地幫助我們對於 FMDV 抗原決定位的分類及定位上，所以在這章我們將要詳細的分析這些物化性質的特徵，找出有用的生物意義幫助我們對於未來疫苗研究上物化性質的瞭解。

免疫研究中物化性質討論，基於先前相關 B 細胞表頂物化性質的研究，可以得知與二級結構的特性[32]、疏水性/親水性[26]、極性、易曲性及表面的可接近性有密切相關[30, 34, 60]。此外自抗體抗原交互作用的研究表示大部分抗原決定位在結構上的特色為其可與抗體產生分子間交互作用力，一般具有親水性、有極性產生靜電作用力、具有彈性及使抗體容易與之接觸的結構特性。如先前相關研究中所提及，利用單一抗原相關物化性質來預測結果並不理想，但是在各種相關物化參數研究下，不容易自為數眾多的這些指標中瞭解其彼此間隱晦的意涵[36]。

在此我們利用相關病原體免疫的資訊使用 IBCGA 所選出一組物化特性組合，此包含 36 個物化性質指標，在此我們針對所選出物化性質的相關文獻，進行分析歸納。由初步的文獻分析可以發現在我們所選出的 36 個物化性質中，與二級結構、疏水性/親水性、極性、易曲性上述相關研究中相同的性質，在所有 feature 中特別是與二級結構此構形相關的物化性質占了大部分在所選出來的 feature 中(17/36)。此與抗原表頂位的相關研究中指出符合，表頂常出現在 loop 或 turn 蛋白質表面突出的部分，某些則是出現在 alpha-helix 及 beta-shee 但是很少出現在 beta-shee 的中央[61]。疏水性/親水性指標則是具有認為表頂位於蛋白質表面大部分為親水性殘基，其占有第二多數量(11/36)，其它極性及易曲性則是各占 3 項，最後一項為分析蛋白質表面與內部組成之物化性質。

被選到的物化性質中不乏先前相關研究中被用來預測 B 細胞表頂的指標，如: HOPT810101 此物化性質為 Hopp 及 Woods 利用 Levitt 所提出的親水性性質尺度 (hydorophilicity propensity)[26]，BHAR880101、KARP850102 則是先前利用易曲性 (flexibility)預測預測 B 細胞表頂[30]。若看單一指標可以看出在某些相關物化性質的局部峰值，可能為的表頂位點，但並非絕對在基於實驗得到的結果下。

由此我們可以推想使用多種相似性質指標但不同數值下，共同”合作”來預測效果會比起單一指標效果更佳，在我們基於此病原體表頂免疫資料下使用特徵選取的方式所選出的物化性質。

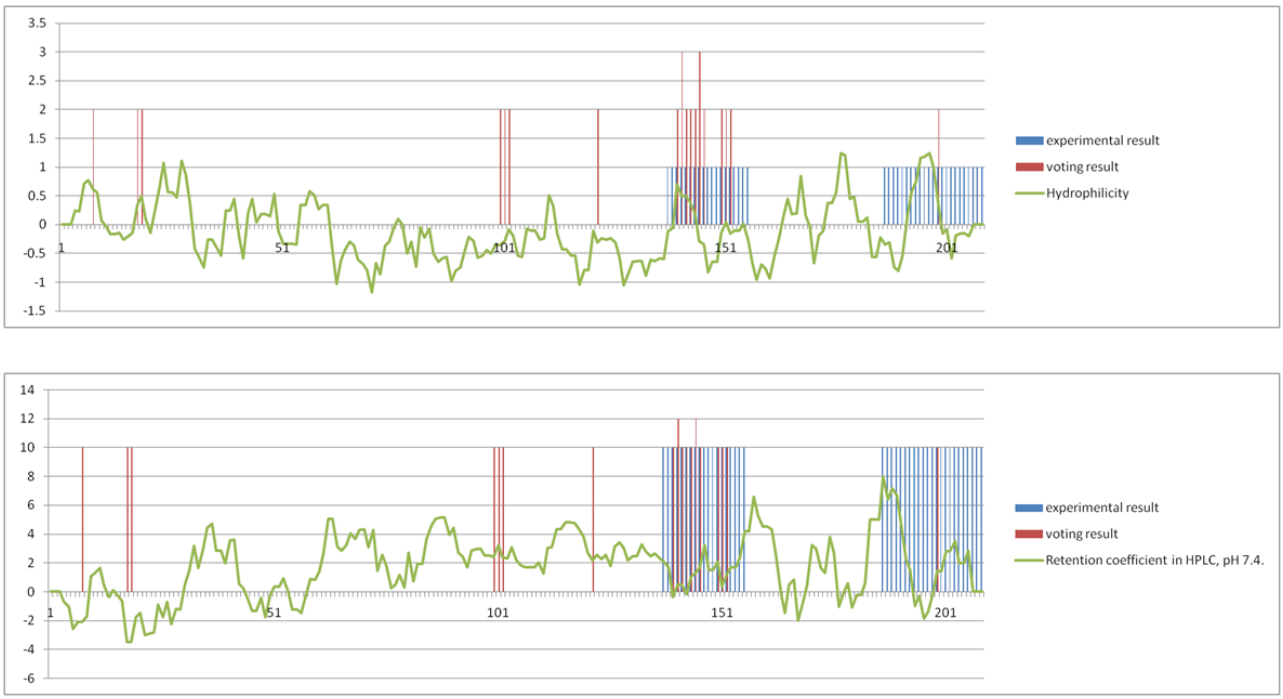


圖 34 為所選出的物化性質單一指標所得到的值

接下來我們進一步分析物化性質，使用MED針對主效果分析分析各物化性質對於分類上的影響，可以得到各組feature影響此模型分類上(如圖35)。可以發現BUNA790101具有最高的MED數值，查詢相關文獻發現此物化特性指標來自^{1}H-NMR核磁共振對於研究線性四聚胜肽在水溶液中構型的變化所得到的物化性質數值，此外BUNA790103與此文獻相關的也在MED排名中佔第9此為與構形相關的feature，特別地是與random coil相關[62]。接下來的MONM990201則是與膜蛋白間alpha-helix間turn相關的物化性質[63]，RICJ880106及ONEK900101為與alpha-helix組成相關的物化特性[64]。QIAN880110及QIAN880114則是與alpha-helix及beta-shee二級結構相關[65]。NAKH920104則是研究膜蛋白膜內膜間膜外的組成，此與親水性/疏水性指標相關。

此結果顯示出與二級結構相關的物化性質戰有很大的影響力在我們使用36個物化性質共同幫助分類上，不意外地與turn、loop、alpha-helix及beta-shee二級結構相關物化性質指標在MED排名中在前7項中佔了6項，特別地是排名第一的是與random coil相關的物化性質，此是否暗示著在FMDV抗原定位上具有random coil構型或是其抗原所在構形為結構紊亂的組態，以及FAUJ880112與負電荷相關的物化性質在MED中以所佔有的物化性質比例上排名也得到相當高的分數，這部分值得我們更進一步地討論。

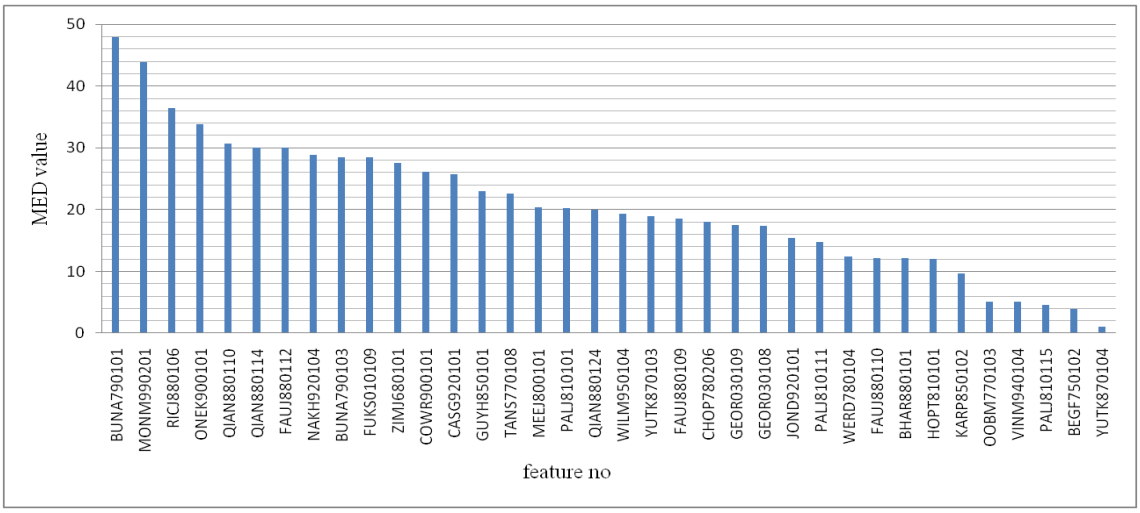


圖 35 MED 排名最高準確度訓練結果的那組物化性質

表 13 大於 MED 大於 30 的分析結果

序號	描述	文獻摘要
BUNA790101	alpha-NH chemical shifts (Bundi-Wuthrich, 1979)	利用 NMR 研究 peptide 的構型在水溶液中 random coil 傾向
MONM990201	Averaged turn propensities in a transmembrane helix (Monne et al., 1999)	分析膜蛋白上的 turn 組成(在兩 α 螺旋間)
RICJ880106	Relative preference value at N3 (Richardson-Richardson, 1988)	胺基酸偏好在末端的 α 螺旋
ONEK900101	Delta G values for the peptides extrapolated to 0 M urea (O'Neil-DeGrado, 1990)	一個熱力學規模的螺旋傾向性的經常發生的胺基酸(α 螺旋構型參數)
QIAN880110	Weights for alpha-helix at the window position of 3 (Qian-Sejnowski, 1988)	α 螺旋構型參數
QIAN880114	Weights for beta-sheet at the window position of -6 (Qian-Sejnowski, 1988)	beta-sheet 構型參數
FAUJ880112	Negative charge (Fauchere et al., 1988)	Negative charge

接下來我們統計在 100 批次下物化性質出現的頻率(如圖 36)，並且與我們所挑選出的那組物化性質間做比較，使用 aaindex 中註解相關 correlation coefficients 0.8，及文獻的相關性。我們想要知道在在 100 批次下最常被選到的物化性質與我們得到的物化性質是否有關聯性，因為我們相信能代表此問題生物意義的物化性質應該頻繁地出現在其它的以相同病原體免疫資料中。

我們比較出現30次以上的物化性質與我們選出來的此36個物化性質的部份(如表13)。其中 KLEP840101、MEEJ800101、BUNA790103 及 COSI940101 與此 36 物化性質中完全相同，特別的排名第一的 KLEP840101 此在我們利用 MED 排名的物化性質中也得到

很高的值，此外排名第二的 QIAN880138 雖然與我們此 36 物化性質中無完全相同的，但是此物化性質為與 coil 相關，此與我們利用 MED 排序中第一名的物化性質 random coil 方面相關，而另外 1 項 VINM940103 則是我們選出的物化性質 set 中性質相近，在以 correlation coefficients 0.8 前提下，分析這些結果間接提供我們文獻探索的方向在更進一步地分析上。

表 14 呈現出此統計在 100 批次下，大於 30 的物化性質

Feature NO	Data description	Frequency	the correlation coefficients of 0.8
KLEP840101	Net charge (Klein et al., 1984)	38	ˇ
QIAN880138	Weights for coil at the window position of 5 (Qian-Sejnowski, 1988)	35	*
VINM940103	Normalized flexibility parameters (B-values) for each residue surrounded by one rigid neighbours (Vihinen et al., 1994)	34	*
MEEJ800101	Retention coefficient in HPLC, pH7.4 (Meek, 1980)	32	ˇ
BUNA790103	Spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich, 1979)	31	ˇ
COSI940101	Electron-ion interaction potential values (Cosic, 1994)	30	ˇ

註ˇ表示與我們選出的 feature set 相同，*表示相似度大於 0.8，×表示不同。

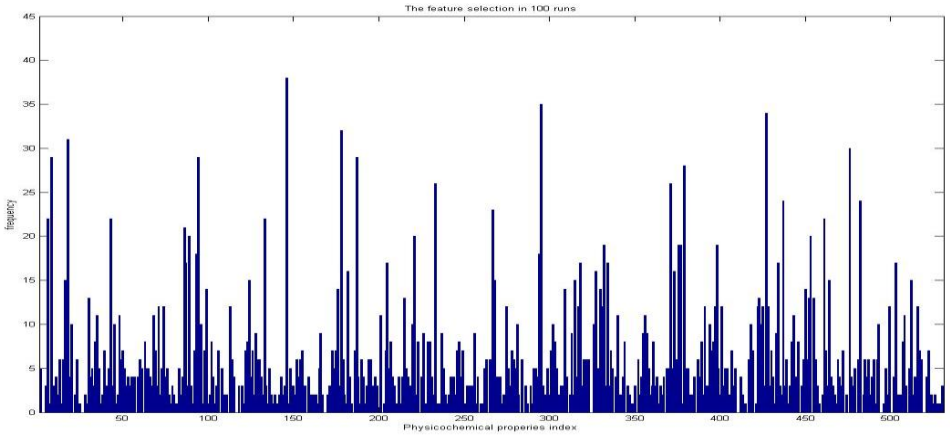


圖 36 在 100runs 物化性質出現的頻率

由先前 B 細胞表頂研究的文章中，此部分中與物化性質相關的文獻主要分為兩類，第一類由單株抗體定義出表頂位置，並利用 X-ray 及 NMR 研究抗體抗原間相互辨識的基礎[66-67]，另一種是利用單抗或多抗定義抗原決定位後，研究不同病毒株抗原決定位差異的研究。

我們可以瞭解到，在結構上 G-H loop、B-C loop、E-F loop、B-B knob、C terminus 等，二級結構上的位點為重要的抗原決定位。其中 VP1 上的 G-H loop 被視為重要的抗原結構特徵(圖 37)，其具有高度地免疫原性(immungenic)及誘導高層級的中和抗體，在 G-H loop 結構中的 loop 位於表面衣殼上的一個小凹陷，其由一個短區域的 β-sheet，隨後由精胺酸-甘胺酸-天門冬胺酸（RGD）三肽，然後再接一個 3_{10} 螺旋[68]。由 133~160 結構上可以看出分別具有 alpha-helix、β-sheet、loop 構型，此針對 FMDV 結構也指出 RGD 部位具有可移動性(flexibility)及暴露性，這結構上的特徵與我們選出的物化性質一致，在 MED 排名的前幾項中。特別地是在 VP1 的 G-H loop 結構研究中發現此 loop 的結構呈現結構的紊亂性(disorder)[67]，這暗示著此 loop 的特性似乎接近 random coil 的性質，而此物化性質在 MED 中排名第一。

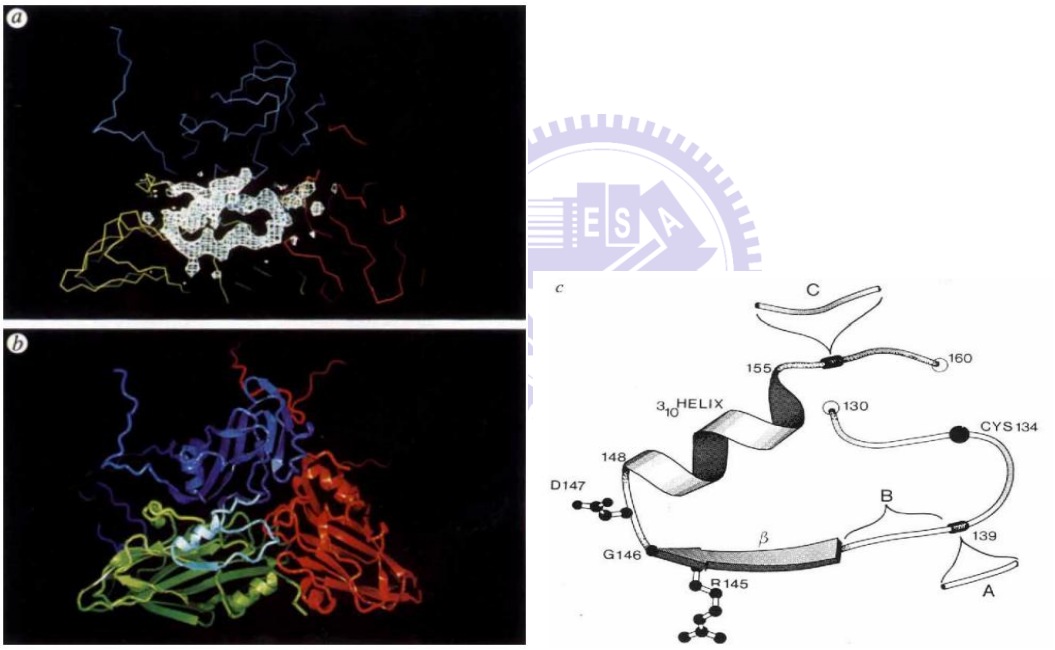


圖 37 G-H loop 結構上的特徵

註:在 alpha-helix 後面接一個 loop 然後接一個 β-sheet，在 loop 上有 RGD 序列(145~148)

另一個值得關注的物化性質針對 Negative charge，針對 FMDV 抗原變異的流行病學文獻中可以發現帶電荷胺基酸在抗原決定位很重要[69]。我們知道帶電性胺基酸為 Asp、Glu、Arg、Lys、His(D、E、R、K、H)，其中 Arg、Lys、His 帶正電 Asp、Glu 帶負電。

我們試著由圖形化看出表頂帶電性胺基酸在表頂上的位置分佈情況，圖 38 表示出與帶電性質相關的物化性質可以知道局部極性部位與我們預測的熱點相互重合的機會很大，雖然並非完全的一致。

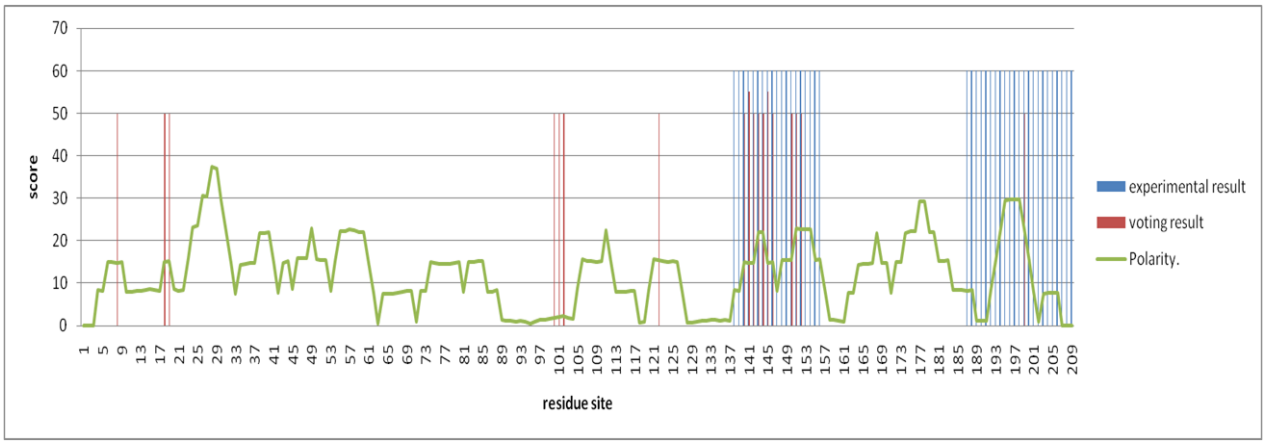


圖 38 為表示出極性(polar)物化性質的指標與實驗結果

此外分析我們訓練資料中其胺基酸組成(圖 39)可以發現帶電性胺基酸在表頂序列之組成大於非表頂序列，當然這並不意味著有帶電性胺基酸的序列即為具抗原性的序列而是此性質為抗原性序列所蘊含的物化特性之一。

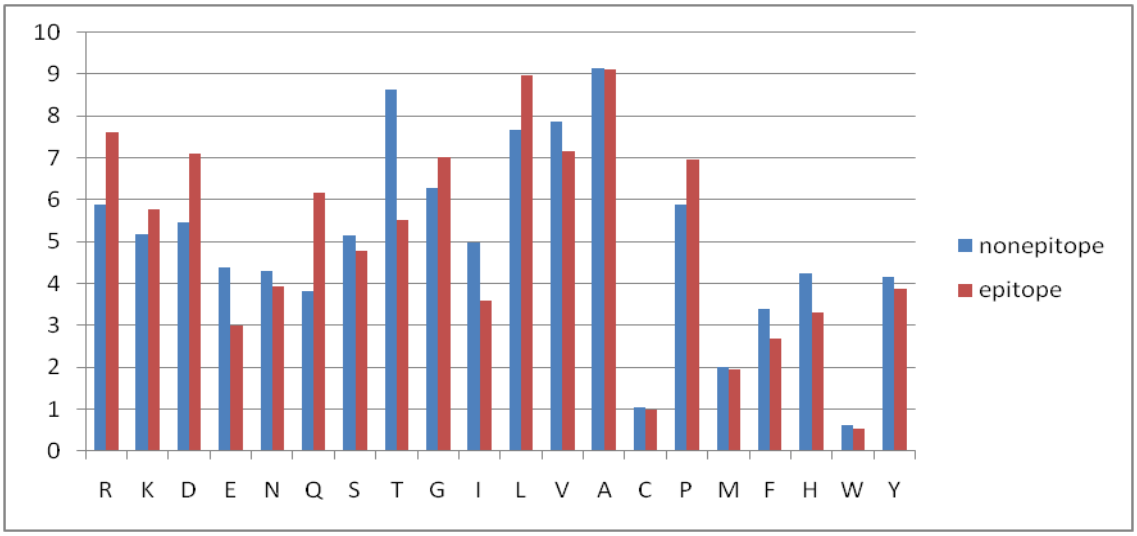


圖 39 訓練資料中表頂及非表頂組成

此外針對我們所訓練的 100 批次實驗結果中，得知每一批次選出來的 feature set 都有所差異，且最佳的 training set 得到 36 個不同的 features 指標，此變動的結果是否能得到更為一致性的結論，由最近研究中指出使用 fuzzy-c-mean 的分群方式針對此 531 物化性質做進一步的分析，顯示雖然有 531 個物化特性指標，但是主要分佈在 20 群相近的分群中。可以知道將此 100 批次實驗所得到 feature set 分群分佈的結果，雖然得到的每一批次 feature set 可能不一樣，但是主要的 feature 還是分佈在前六類的 cluster 中。

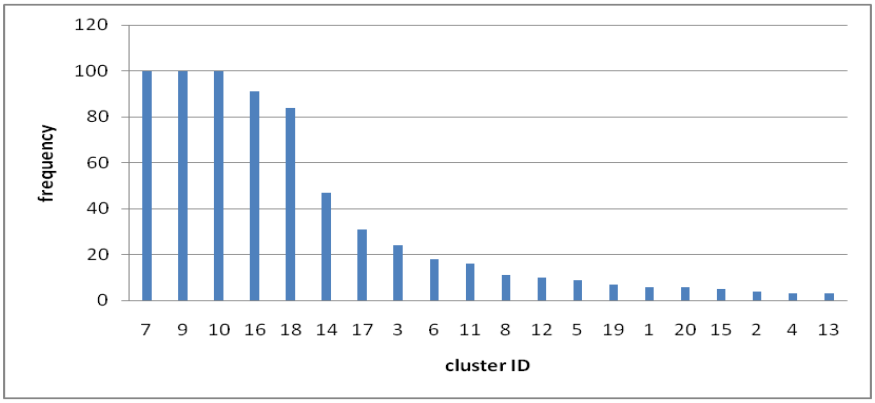


圖 40 此 100 批次實驗所得到 feature set 分群分佈的結果

最佳的 training set 雖然得到 36 個不同的 features 指標，實際上由此文獻來看此 36 個不同的 features 指標主要分為六群當中[70]。此六群各自為 C7、C9、C10、C16、C18 及 C14。

表 15 為物化性質分群

Cluster	A	B	C	H	P	O	TOTAL
C_1					1	1	2
C_2					2		2
C_3				6			6
C_4					3		3
C_5			1	2	1		4
C_6	1			3	1	1	6
C_7	47	7	2	74	14	3	147
C_8					3		3
C_9	51	1	3	50	6	21	132
C_{10}	38	30	2	42	9	2	123
C_{11}					6		6
C_{12}				2			2
C_{13}			1				1
C_{14}				12	2	1	15
C_{15}				1			1
C_{16}	1		38	4			43
C_{17}				3			3
C_{18}	3			17	8		28
C_{19}				1		1	2
C_{20}					2		2
TOTAL	141	38	47	217	58	30	531
RATE	0.266	0.072	0.089	0.409	0.109	0.056	

註黃色部位為我們的物化性質在相關文獻中所顯示分佈的位置

綜合上述物化性質分析的結果，可以知道對於 FMDV 抗原決定位來說，二級結構方面的物化性質重要性遠大於其它部分，但是抗原決定位的性質並無法簡單地以單一性質呈現出來，利用我們的物化性質組合能更加接近實驗的結果。此外可以發現 FMDV 抗原決定位的物化性質是帶電性、具有 loop 或 turn 性質但是一些也具有 random coil 性質這些性質都是結構上 flexibility 性質較高的地方。此外我們選出來的物化性質是有意義地對於預測口蹄疫病毒的 B 細胞表頂上。

七、結論

7.1 討論

在我們的研究中，我們發展一個計算的系統對於預測抗原決定位基於使用病原體特異性的次分類群及特徵選取的策略。我們的結果顯示出，我們所建立的預測模型不僅能達到較高的預測準確度在與 B 細胞表頂的預測工具中，也能提供有用的資訊對於抗原決定位的分析上，基於物化性質的分析上可以提供未來設計疫苗及研究病原體表頂位置的一個方向。

對於定義出抗原的位置很難有評估結果優劣的方式，在沒有實驗的基礎之下，我們使用了估計的方式來評估此結果，但是我們認為使用綜合預測結果的方式來定義可能的位置是更好的方式。至於使用我們所提供的工具來決定抗原的位置及序列，建議可以依照個人的需求以此熱點擴增序列的長度或縮減序列的長度。

由結果表示出此針對 B 細胞表頂次分類群來訓練得到的物化性質，在使用來幫助定義出病原體上重要的生物意義是較為有效的益的方式，對於高度變異下的病原體預測序列抗原性質是一件相當不易的事，我們的假設是基於物化性質的保守性，亦即其病原體具有達爾文正選擇的情況下其突變傾向於置換相似物化性質的胺基酸[25]，以此原則下達到預測其它品系的病毒株，因此假設此病毒株傾向於達爾文負選擇，在這個情況下，此模型預測的效果可能不盡理想，因為抗病毒血清可能無法辨識出專一性結合的表頂。此外，此免疫的資訊是基於利用抗血清所得到的序列資訊，在我們的訓驗資料中大部分為 A、O 及 C 血清型(29/38)所得到的實驗結果，合理地認為其在此三類的預測效果可能會優於使用在其它血清型上。或許將來更多的實驗結果定義出來可以使用相同的方式幫助此模型更加地完善。

7.2 未來展望

對於的未來發展上，由此研究結果可以瞭解到利用特徵選取的方式可以幫助找尋並確認實驗上所得到的資訊，這個架構或許可以提供未來針對感興趣的病原體之免疫研究的一個方向，比起大尺度地研究表頂之物化特性，似乎針對相關病原體所得到的物化性質可以更加接近真實的生物意義，此外若結合多種預測工具一致性的結果(如表 2)[16]，或許可以更進一步地篩選出我們所欲得到的預測結果。此概念若能結合先進的免疫微陣列實驗技術，快速且大量的針對病原體進行抗原篩選，將可幫助遏止及病的擴散在面對新爆發疾病病毒疫苗的設計。

參考文獻

1. Grubman, M.J. and B. Baxt, *Foot-and-mouth disease.* Clinical Microbiology Reviews, 2004. 17(2): p. 465-+.
2. Doel, T.R., *FMD vaccines.* Virus Res, 2003. 91(1): p. 81-99.
3. Rodriguez, L.L., et al., *A synthetic peptide containing the consensus sequence of the G-H loop region of foot-and-mouth disease virus type-O VP1 and a promiscuous T-helper epitope induces peptide-specific antibodies but fails to protect cattle against viral challenge.* Vaccine, 2003. 21(25-26): p. 3751-3756.
4. Taboga, O., et al., *A large-scale evaluation of peptide vaccines against foot-and-mouth disease: lack of solid protection in cattle and isolation of escape mutants.* J Virol, 1997. 71(4): p. 2606-14.
5. Kim, S.A., et al., *DNA vaccination against foot-and-mouth disease via electroporation: study of molecular approaches for enhancing VP1 antigenicity.* J Gene Med, 2006. 8(9): p. 1182-91.
6. Shao, J.J., et al., *Promising multiple-epitope recombinant vaccine against foot-and-mouth disease virus type O in swine.* Clin Vaccine Immunol, 2011. 18(1): p. 143-9.
7. Su, C.X., et al., *Heterologous expression of FMDV immunodominant epitopes and HSP70 in P.pastoris and the subsequent immune response in mice.* Veterinary Microbiology, 2007. 124(3-4): p. 256-263.
8. Du, Y., et al., *Enhanced immunogenicity of multiple-epitopes of foot-and-mouth disease virus fused with porcine interferon alpha in mice and protective efficacy in guinea pigs and swine.* J Virol Methods, 2008. 149(1): p. 144-52.
9. Klein, J., *Understanding the molecular epidemiology of foot-and-mouth-disease virus.* Infect Genet Evol, 2009. 9(2): p. 153-61.
10. Feigelstock, D.A., et al., *Emerging foot-and-mouth disease virus variants with antigenically critical amino acid substitutions predicted by model studies using reference viruses.* Vaccine, 1996. 14(2): p. 97-102.
11. Cox, S.J. and P.V. Barnett, *Experimental evaluation of foot-and-mouth disease vaccines for emergency use in ruminants and pigs: a review.* Vet Res, 2009. 40(3): p. 13.
12. Zhang, Z.W., et al., *Screening and identification of B cell epitopes of structural proteins of foot-and-mouth disease virus serotype Asia1.* Vet Microbiol, 2010. 140(1-2): p. 25-33.
13. Sirskyj, D., et al., *Innovative bioinformatic approaches for developing peptide-based vaccines against hypervariable viruses.* Immunol Cell Biol, 2011. 89(1): p. 81-9.
14. Reineke, U. and M. Schutkowski, *Epitope mapping protocols*. 2nd ed. Methods in

molecular biology,. 2009, New York: Humana Press. xiii, 456 p., 16 p. of plates.
15. El-Manzalawy, Y. and V. Honavar, *Recent advances in B-cell epitope prediction methods.* Immunome Res, 2010. 6 Suppl 2: p. S2.
16. Yang, X. and X. Yu, *An introduction to epitope prediction methods and software.* Rev Med Virol, 2009. 19(2): p. 77-96.
17. Rao, K.V., *antigen.* ENCYCLOPEDIA OF LIFE SCIENCES 2001. .
18. King, A.M.Q., G.J. Belsham, and A.I. Donaldson, *Foot-and-mouth Disease*, in *eLS*. 2001, John Wiley & Sons, Ltd.
19. Frank, S.A., 2002.
20. Mason, P.W., M.J. Grubman, and B. Baxt, *Molecular basis of pathogenesis of FMDV.* Virus Res, 2003. 91(1): p. 9-32.
21. Lea, S., et al., *The structure and antigenicity of a type C foot-and-mouth disease virus.* Structure, 1994. 2(2): p. 123-39.
22. EPM Teurlings, A.K., Foot-and-Mouth Disease Virus:Structure and Antigenicity. 1998.
23. Oem, J.K., et al., *Development of synthetic peptide ELISA based on nonstructural protein 2C of foot and mouth disease virus.* J Vet Sci, 2005. 6(4): p. 317-25.
24. Greenbaum, J.A., et al., *Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools.* J Mol Recognit, 2007. 20(2): p. 75-82.
25. Haydon, D.T., et al., *Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates.* Genetics, 2001. 157(1): p. 7-15.
26. Hopp, T.P. and K.R. Woods, *Prediction of protein antigenic determinants from amino acid sequences.* Proc Natl Acad Sci U S A, 1981. 78(6): p. 3824-8.
27. Saha, S., M. Bhasin, and G.P. Raghava, *Bcipep: a database of B-cell epitopes.* BMC Genomics, 2005. 6: p. 79.
28. Vita, R., et al., *The immune epitope database 2.0.* Nucleic Acids Res, 2010. 38(Database issue): p. D854-62.
29. Huang, J., S. Kawashima, and M. Kanehisa, *New amino acid indices based on residue network topology.* Genome Inform, 2007. 18: p. 152-61.
30. Parker, J.M., D. Guo, and R.S. Hodges, *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.
31. Kolaskar, A.S. and P.C. Tongaonkar, *A semi-empirical method for prediction of antigenic determinants on protein antigens.* FEBS Lett, 1990. 276(1-2): p. 172-4.
32. Pellequer, J.L., E. Westhof, and M.H.V. Vanregenmortel, *Correlation between the Location of Antigenic Sites and the Prediction of Turns in Proteins.* Immunology Letters, 1993. 36(1): p. 83-100.
33. Saha, S. and G.P.S. Raghava, *BcePred: Prediction of continuous B-cell epitopes in*

antigenic sequences using physico-chemical properties. Artificial Immune Systems, Proceedings, 2004. 3239: p. 197-204.

34. Alix, A.J.P., *Predictive estimation of protein linear epitopes by using the program PEOPLE.* Vaccine, 1999. 18(3-4): p. 311-314.
35. Pellequer, J.L. and E. Westhof, *Preditop - a Program for Antigenicity Prediction.* Journal of Molecular Graphics, 1993. 11(3): p. 204-&.
36. Blythe, M.J. and D.R. Flower, *Benchmarking B cell epitope prediction: underperformance of existing methods.* Protein Sci, 2005. 14(1): p. 246-8.
37. Chen, J., et al., *Prediction of linear B-cell epitopes using amino acid pair antigenicity scale.* Amino Acids, 2007. 33(3): p. 423-428.
38. Odorico, M. and J.L. Pellequer, *BEPITOPE: predicting the location of continuous epitopes and patterns in proteins.* J Mol Recognit, 2003. 16(1): p. 20-2.
39. Saha, S. and G.P.S. Raghava, *Prediction of continuous B-cell epitopes in an antigen using recurrent neural network.* Proteins-Structure Function and Bioinformatics, 2006. 65(1): p. 40-48.
40. El-Manzalawy, Y., D. Dobbs, and V. Honavar, *Predicting flexible length linear B-cell epitopes.* Comput Syst Bioinformatics Conf, 2008. 7: p. 121-32.
41. El-Manzalawy, Y., D. Dobbs, and V. Honavar, *Predicting linear B-cell epitopes using string kernels.* Journal of Molecular Recognition, 2008. 21(4): p. 243-255.
42. Sweredoski, M.J. and P. Baldi, *COBEpro: a novel system for predicting continuous B-cell epitopes.* Protein Eng Des Sel, 2009. 22(3): p. 113-20.
43. Sollner, J., et al., *Analysis and prediction of protective continuous B-cell epitopes on pathogen proteins.* Immunome Res, 2008. 4: p. 1.
44. Holland, J.H., *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence*. 1st MIT Press ed. Complex adaptive systems. 1992, Cambridge, Mass.: MIT Press. xiv, 211 p.
45. Hedayat, A., N.J.A. Sloane, and J. Stufken, *Orthogonal arrays : theory and applications*. Springer series in statistics. 1999, New York: Springer. xxii, 416 p.
46. Dey, A., *Orthogonal fractional factorial designs*. 1985, New York: Wiley. viii, 133 p.
47. Taguchi, G.i., S. Konishi, and American Supplier Institute., *Orthogonal arrays and linear graphs : tools for quality engineering*. 1987, Dearborn, Mich.: American Supplier Institute. vii, 72 p.
48. Shinn-Ying, H., S. Li-Sun, and C. Jian-Hung, *Intelligent evolutionary algorithms for large parameter optimization problems.* Evolutionary Computation, IEEE Transactions on, 2004. 8(6): p. 522-541.
49. Shinn-Ying, H., C. Jian-Hung, and H. Meng-Hsun, *Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications.* Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 2004. 34(1): p.

609-620.

50. Chih-Chung Chang and Chih-Jen Lin, L.a.l.f.s.v.m.A.T.o.I.S.a.T., 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
51. Yang, D., et al., *Identification of a conserved linear epitope on the VP1 protein of serotype O foot-and-mouth disease virus by neutralising monoclonal antibody 8E8.* Virus Res, 2011. 155(1): p. 291-9.
52. Cooke, J.N. and K.M. Westover, *Serotype-specific differences in antigenic regions of foot-and-mouth disease virus (FMDV): a comprehensive statistical analysis.* Infect Genet Evol, 2008. 8(6): p. 855-63.
53. Gao, S.-d., et al., *B cell epitopes within VP1 of type O foot-and-mouth disease virus for detection of viral antibodies.* Virologica Sinica, 2010. 25(1): p. 18-26.
54. Ma, L., et al., *Preparation and characterization of neutralizing monoclonal antibodies against FMDV serotype O with synthetic peptide antigen.* Hybridoma (Larchmt), 2010. 29(5): p. 409-12.
55. Yu, Y., et al., *Fine mapping of a foot-and-mouth disease virus epitope recognized by serotype-independent monoclonal antibody 4B2.* J Microbiol, 2011. 49(1): p. 94-101.
56. Wang, H., et al., *Identification of a conformational epitope on the VP1 G-H Loop of type Asia1 foot-and-mouth disease virus defined by a protective monoclonal antibody.* Vet Microbiol, 2011. 148(2-4): p. 189-99.
57. Kawashima, S., et al., *AAindex: amino acid index database, progress report 2008.* Nucleic Acids Res, 2008. 36(Database issue): p. D202-5.
58. Chang, H.T., C.H. Liu, and T.W. Pai, *Estimation and extraction of B-cell linear epitopes predicted by mathematical morphology approaches.* J Mol Recognit, 2008. 21(6): p. 431-41.
59. Wang, Y., et al., *Cn3D: sequence and structure views for Entrez.* Trends Biochem Sci, 2000. 25(6): p. 300-2.
60. Rubinstein, N.D., et al., *Computational characterization of B-cell epitopes.* Mol Immunol, 2008. 45(12): p. 3477-89.
61. Barlow, D.J., M.S. Edwards, and J.M. Thornton, *Continuous and discontinuous protein antigenic determinants.* Nature, 1986. 322(6081): p. 747-8.
62. Bundi, A. and K. Wüthrich, *1H-nmr parameters of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH.* Biopolymers, 1979. 18(2): p. 285-297.
63. Monne, M., et al., *Turns in transmembrane helices: determination of the minimal length of a "helical hairpin" and derivation of a fine-grained turn propensity scale.* J Mol Biol, 1999. 293(4): p. 807-14.
64. O'Neil, K.T. and W.F. DeGrado, *A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids.* Science, 1990. 250(4981): p. 646-51.

65. Qian, N. and T.J. Sejnowski, *Predicting the Secondary Structure of Globular-Proteins Using Neural Network Models.* Biophysical Journal, 1988. 53(2): p. A98-A98.
66. Acharya, R., et al., *The three-dimensional structure of foot-and-mouth disease virus at 2.9 A resolution.* Nature, 1989. 337(6209): p. 709-716.
67. Logan, D., et al., *Structure of a major immunogenic site on foot-and-mouth disease virus.* Nature, 1993. 362(6420): p. 566-8.
68. Burman, A., et al., *Specificity of the VP1 GH loop of Foot-and-Mouth Disease virus for alphav integrins.* J Virol, 2006. 80(19): p. 9798-810.
69. Mohapatra, J.K., et al., *Analysis of the leader proteinase (L(pro)) region of type A foot-and-mouth disease virus with due emphasis on phylogeny and evolution of the emerging VP3(59)-deletion lineage from India.* Virus Res, 2009. 141(1): p. 34-46.
70. Huang, H.L., et al., *Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties.* BMC Bioinformatics, 2011. 12: p.