計畫編號：DOH95-DC-1405

行政院衛生署疾病管制局九十七年度科技研究發展計畫

流感病毒生物資訊系統之建立

研 究 報 告

執行機構：國立陽明大學 生物資訊研究中心

計畫主持人：張傳雄

研究人員：張傳雄、鍾翊方、楊永正、蘇俊泓、蔡毓舜、陳毅誠、周坤億、李曉玟

執行期間： 97 年 1 月 1 日至 97 年 12 月 31 日

目　　錄

共　（ 69 ）頁

摘　　要

(1)中文摘要

世界各國在推動大型計畫時，通常都會建立核心設施，以增加分析的效率與品質。該核心設施必須對許多計畫的成員服務，為求時效，為流感疫苗研究發展計畫建立專屬之流感資訊核心設施，支援此計畫下之各團隊做即時的生物資訊分析。第一年度流感資訊核心設施首先將依各團隊的需求，利用資訊技術蒐集所有與流感相關的數據與文獻，在整合後以 RSS 的技術，主動提供給各研究團隊參考。在第二年度並增加知識管理的部份，利用 ontology 整合、分類流感病毒相關研究成果，加入下列生物學特性: 抗原、抗藥性、HI titer、病毒 recombination and reassortment 等分析於 IVBS。在資料量增大時，資料品質的控管，與自動化的分析流程就成為重要的問題。今年度核心設施根據其他的團隊的需求，建立分析流感病毒序列的自動化流程於 Influenza Virus Bioinformatics System (IVBS): 增加流感病毒全基因體之基因型(genotyping)之預測與分佈功能、增加使用者個人流感病毒基因序列資料之輸入儲存與後續分析功能。並建立與增加新的生物資訊學技術預測流感病毒的演化趨勢，協助選擇適合用來做流感疫苗的流感病毒株: B 細胞、CTL 細胞抗原決定位胜肽片段預測程式及系統。我們在今年度所完成之工作如下:

1. 已完成 IVBS 流感病毒序列資料之自動更新
2. 已完成以 sequence logo 及階層方式同時呈報大規模流感病毒蛋白質序列排列中氨基酸之相似程度
3. 已完成流感病毒全基因體之基因型(genotyping)預測與分佈功能
4. 已完成增加使用者個人流感病毒基因序列資料之輸入儲存與後續分析功能
5. 已完成依據抗原決定位之胺基酸特性開發預測 B-cell epitope 序列之程式
    a. 已完成 IVBS 現有 B-cell epitope 的胺基酸組成分析
    b. 已完成現有預測 B-cell epitope 方法之研究調查
    c. 已完成根據調查的 B-cell epitope prediction 預測方法撰寫預測程式

關鍵詞：流感資訊核心設施、流感病毒基因型、個人序列資料、細胞抗原決定位胜肽片段預測

摘　　要

(2) 英文摘要

Core facilities are usually set up to increase both the analysis efficiency and the work quality when a large-scale research project is initiated in many countries around the world. The prupose of these core facilities is to serve the needs of other related projects. First year, this influenza information core facility collects all the available influenza data and information, and provides them to the other team projects after integrating with the RSS technology. Second year, this core facility analyze the viral biological characteristics such as antigene, drug resistance, HI titer, virus recombination and genetic reassorment etc., to further expand the knowledge base. The core will also seek for the capacity of processing large dataset and forming pipeline and quality control for data analysis. Third year, according to the request of other related projects, the influenza A virus genotype prediction tool and personal workingset database were added. By using bioinformatics techniques, we develop tools to predict trend of influenza virus evolution, and facilitates the prediction of vaccine targets.

In the third year, we have accomplished the following works:

1. Automatic update of influenza virus sequence
2. Use sequence logo to show the conserved region of similar influenza protein sequences
3. Influenza A virus genotype database and prediction tool
4. Personal workingset database for saving and analysing personal sequence data
5. B-cell epitope prediction tool
    a. Simialrity analysis of B-cell epitope nucleotide sequences
    b. Investigation of available tools for B-cell epitope predtion
    c. Development of in-house B-cell epitope predtion tools

Keywords: influenza information core facility, genotyping, Personal workingset database, epitope prediction tool

本　　文

(1) 前言

流感疫苗研究發展計畫的目的是要在三年內建立預測會在臺灣流行的病毒株，並在我國自行開發、製造流感疫苗。目前世界衛生組織有標準作業程序，設計與製造流感疫苗，所以我國至少要能做到世界衛生組織的預測水準。不過世界上對流感病毒究竟如何演化，並無定論，因此對於猜測未來會流行的病毒株並無最佳對策。我國的流感疫苗研發計畫將利用本土資訊，改進預測的結果。國內的疾病管制局與流行病學專家將主導流感疫苗的研發，但需生物資訊學者之配合，以增加工作效率，如期達到目標。

流行性感冒要能在人類中造成大流行，必須滿足三個條件：(1)必須為全新的病毒，方能逃避人體免疫機制、(2)有能力造成人類致病與(3)能有效率的在人與人之間傳播 (Viboud et al., 2006) 。因為流感病毒具有抗原多變性的特性，可經由突變及基因重組兩種方式產生新型病毒，點突變只會造成抗原小部份改變，引起流感的小流行，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和重新排列組合(reassortment)，則會引起全球性的大流行(epidemic)，reassortment 的發生對流感病毒的演化與造成週期性流行的重要性目前尚未明朗，但有可能影響其演化機制。為了瞭解不同年

代中流感基因體可能發生的基因位移及變異，我們建立了流感病毒生物資訊系統 (Influenza Virus Bioinformatics System , IVBS) ，該系統整合 NCBI IVR 與 LANL ISD 資料庫中的所有公開病毒序列資訊，使用各種資訊技術，例如資訊代理人（web agent）等，自動蒐集資訊，並提供與流感病毒相關的泛用分析工具，包含相似度比對、多序列排比、親緣樹分析以及抗藥性分析等。此外，為減少重複以同一策略分析不同組的數據的人力，核心設施將建立自動化的分析流程，增加分析的效率。若流病學專家不知道應選用哪一種分析方法，核心設施將提供諮詢與建議。

目前提供觀察流感變異與演化主要分為兩個方向，一為透過病毒之親緣關係，觀察不同品系之病毒與流行的關係;而另一方向為透過觀察流感病毒隨著時間推進，在序列各個位置上異同之處。

在瞭解產生變異的原因後，則需建立模型，模擬流感病毒演化的過程。例如目前有一派的理論是病毒是以 quasi-species 的形式存在，而不是單一的病毒株。因此哪一株會擴增，是與病毒和人體免疫系統的交互作用有關。若各團隊沒有自己的模型，則核心設施會安裝或撰寫模擬所需的程式，將常用的模型建立起來，協助流病學專家進行模擬。在這過程中，核心設施亦將比較各模型之優缺點，並提出可能的改進方案。一旦能預測到可能的病毒株，則要預測具有抗原性的

區域，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。因此如何利用基因變異的資訊與 HA、NA 的三級結構來預測抗原區就非常重要。核心設施將建立適當的方法，預測具有抗原性的區域。

流感病毒生物資訊系統(IVBS)之前已經整合了 IEDB (Immune Epitope Database, http://www.immuneepitope.org)流感病毒的 B 細胞抗原決定位(B-cellepitope)資料，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。然而，相較於 IVBS 中的流感病毒蛋白序列數量，IEDB 的 epitope 資料量仍遠遠不足。針對這樣的情況，我們分析了 B-cell epitope 胜肽(peptide)片段胺基酸的理化學性質(physico-chemical properties)，根據這些特性建立規則，設計分析 B-cell epitope 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-cell epitope 位置，提供研究人員進一步的參考資訊，協助加速流感病毒疫苗研究的進展。

本研究的總目標在於

1. 自動收集網際網路上所有公開的流感病毒資訊，並與臺灣特有的流感病毒資訊整合。
2. 利用 RSS 技術，自動提供疾病管制局、流感疫苗研究發展計畫的其他團隊即時資訊
3. 開放 Influenza Virus Bioinformatics System (IVBS)上線使用予疾管局相關人員使用，並提供教育訓練課程。
4. 增加知識管理的部份，利用 ontology 整合、分類流感病毒相關研究成果。
5. 加入感染者流行病學資料及下列生物學特性:抗原、抗藥性、HI titer、病毒 recombination and reassortment 等分析於 IVBS 。
6. 建立分析流感病毒序列的自動化流程。
7. 增加所建立之分析流感病毒序列的自動化流程於 Influenza Virus Bioinformatics System (IVBS) 。
8. 開放 Influenza Virus Bioinformatics System (IVBS)上線使用予其他子計畫相關人員使用，並提供教育訓練課程。
9. 利用資訊探採技術找到人工不易發現的關聯性，提供疾病管制局與計畫中的其他團隊參考。
10. 發展最新的生物資訊學技術，協助預測適合用來做流感疫苗的病毒株。

第一年目標

1. 提供流感疫苗研究發展計畫的其他計畫，各種生物資訊分析的需求
2. 自動收集網際網路上所有公開的流感病毒資訊，並與臺灣特有的流感病毒資訊整合
3. 利用 RSS 技術，自動提供疾病管制局、流感疫苗研究發展計畫的其他團隊即時資訊

第二年目標

1. 開放 Influenza Virus Bioinformatics System (IVBS) 予疾管局相關人員使用，並提供教育訓練課程 (2007.03)
2. 增加知識管理的部份，利用 ontology 整合、分類流感病毒相關研究成果 (2007.06)
3. 加入感染者流行病學資料及下列生物學特性: 抗原、 抗藥性、HI titer、病毒 recombination and reassortment 等分析於 IVBS (2007.09)
4. 建立分析流感病毒序列的自動化流程 (2007.12)

第三年目標

1. 建立與增加分析流感病毒序列的自動化流程於 Influenza Virus Bioinformatics System (IVBS)
    i. 增加以 sequence logo 及階層方式同時呈報大規模流感病毒蛋白質序列排列中氨基酸之相似程度(2008.03)
    ii. 增加流感病毒全基因體之基因型(genotyping)之預測與分佈功能(2008.03)
    iii. 增加使用者個人流感病毒基因序列資料之輸入儲存與後續分析功能(2008.06)
2. 建立與增加新的生物資訊學技術預測流感病毒的演化趨勢，協助選擇適合用來做流感疫苗的流感病毒株
    i. 開發流感病毒蛋白質 B 細胞抗原決定位(B-cell epitope)胜肽片段預測程式及系統 (2008.06)
    ii. 開發流感病毒蛋白質 CTL (cytotoxic T lymphocyte)細胞抗原決定位胜肽片段預測程式及系統 (2008.09)
    iii. 開發完整流感病毒(whole virus) Hemagglutination activity 之 seropositivity 生物資訊預測程式及系統 (2008.12)

本　　文

(2) 材料與方法

硬體設備

本計畫陽明大學的生物資訊研究中心先前已由其他計畫及陽明大學所補助購置之 IBM p690 高性能電腦，10-個節點的 Mac 電腦叢集、16-個節點的刀鋒伺服器。在有了高性能的計算設備後，生資中心一方面利用自動化的備份系統來維護資訊的安全，另一方面則利用良好的機房環境來維持系統的穩定性。電腦機房不僅有自己的防火牆、不斷電系統，還有緊急備用的電源，所以在停電時，所有的電腦還是可以繼續運作。除此之外，空調系統不但有不斷電系統，也有自動化的切換模式以保持衡溫。

其後為進一步節省人力資源和時間，生資中心藉由參與亞太先進網路協會 (Asia-Pacific Advenced Network，APAN)的協助，將常用的資料建立為 Bio-mirror，讓不同實驗室的程式設計人員可以很容易地取得資料。在設備上則採用儲存區域網路(storage area network，SAN)來儲存資料，讓多台電腦叢集 可以共用資料。

因為生資中心擁有這些國家級的電腦設備，所以有權調度計算的優先次序。在因應緊急狀況時，將可隨時支援流感疫苗研究發展計畫的所有計算需求。

軟體系統

生物資訊學的研究，除分析工具外，最重要的是資訊的收集。雖然許多數據可直接下載，也有許多序列只能以網頁形式瀏覽。若能利用智慧型代理人，自動收集資訊，將可節省許多人力。目前團隊成員所製作的智慧型代理人軟體已成功地使用在單核酸多型性分析，與肝癌資訊網。

在資料庫整合之建構部份，我們採取 PHP 語法與流程控制及 MySQL 進行增、刪、改、查功能，自定資料記錄的規則，將資料庫以固定的架構來組成資料正規化，用來表示資料庫如何組成的架構資料模型，建構出功能強大的流感病毒生物資訊系統互動 Linux 資料庫網站。建立 MySQL 資料庫、MySQL 資料型別、資料庫增刪作業、資料庫查詢作業，資料的匯入與匯出、PHP 連接 MySQL、資料表結合、處理日期時間資料。

在利用 RSS 技術自動提供指定的資訊部份，我們透過 XML 特性所制定的格式，將網頁內容抽取出來， 讀者訂閱 RSS 後，只要透過 RSS 閱讀器，就可看到。XML 是 eXtensible Markup Language 的簡稱，它的其中一個主要功能就是作 Data Exchange，而 Content Feed 正是一種 Exchange Data 的應用， RSS 只是 Content Feed 的另一種型式。RSS 規格沒有對 title 標記是 plain text 或是 html，RSS 是在 XML 裡包 HTML，所以我們有 XML encoding、HTML encoding。RSS 是 news 為導向的網站公佈的 XML 文件格式，它列出它們目前的標題，提供連結到相關文章的 URL。

序列資料取得

現有提供新定序流感病毒序列以及相關資訊的公用資料庫主要有 NCBI 所建立與維護的 Influenza Virus Resource (IVR, 網址：http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html) 及 Los Alamos National Laboratory (LANL) 所建立的 Influenza Sequence Database (ISD 網址：http://www.flu.lanl.gov/) (Macken et al., 2001)。IVR 的資料源自於 GenBank，除序列外，該資料庫並提供流感病毒基因體列表，供使用者參考。IVR 每日都會進行內容更新，並提供 FTP 下載服務，如有更新資料，會立即置於 FTP 站台上 (網址：ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/) 。

ISD 所儲存的序列資料，除有大部份 IVR 序列資料外，還有研究機構直接提供的序列資料。以截至 2007/7/2 為止，ISD 所含有的 50,836 筆核酸序列為例，有 3,908 筆為使用者直接提交序列資料，佔 7.6%。因此我們只希望擷取出這些序列。該資料庫中雖有提供序列下載功能，但免付費之使用者下載序列的數目受到限制。為此我們撰寫程式，先自搜尋頁面中擷取出 ISD 特有的存取序號 (accession number) ，編碼原則為 ISD 後加七碼，再依據其存取號碼連結至屬於該序列的內容頁，分析其內容，並將其資訊儲存於資料庫中。

除國際上公用資料庫外，亦於台灣疾病管制局取得 3,779 筆台灣

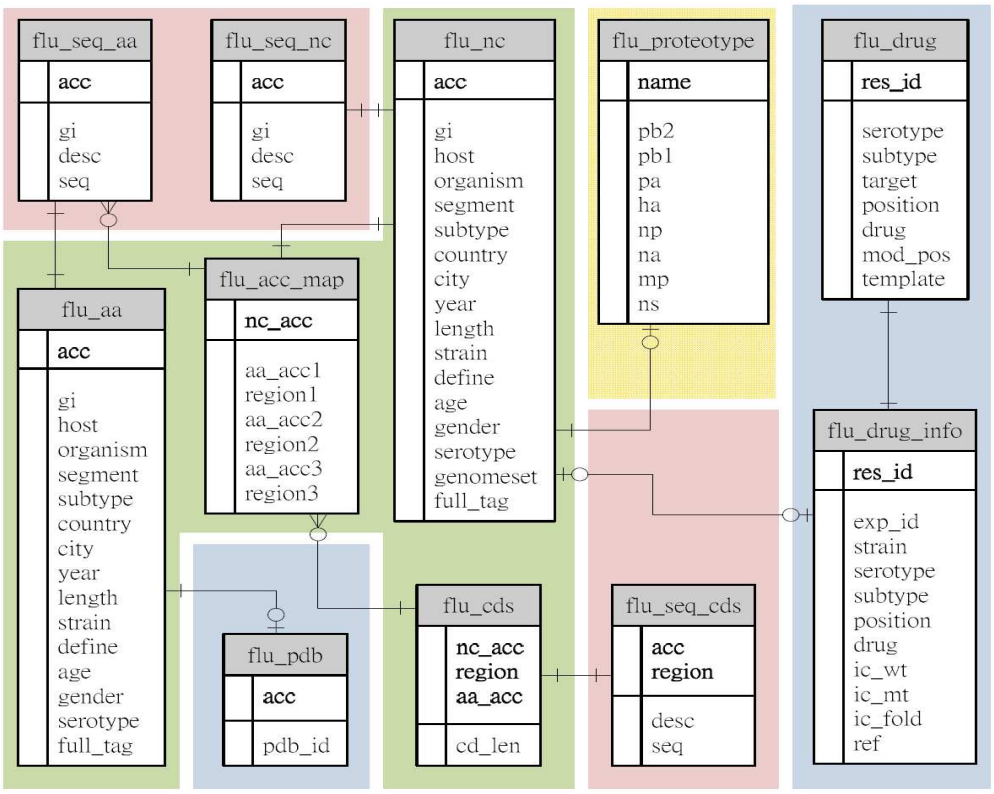
地區所分離出之流感病毒序列，並經轉入資料庫中，與國際序列整合。目前資料庫中共有 83,544 筆核酸序列，98,509 筆蛋白質序列及 97,206 筆編碼區序列。

流感病毒基因型資料取得

系統中所提供的感病毒基因型資料是由 FluGenome 蒐集整合加入 IVBS 中，共 3,678 株流感病毒全基因體之基因型; 及 47,169 筆各 segment 之基因型的資料。

資料庫設計

由於需整合兩資料庫所得到的序列資料，因此我們建立了六個資料表：flu_aa、flu_cds、flu_nc 用於儲存流感序列相關資訊，flu_seq_aa、flu_seq_cds、flu_seq_nc 用於儲存蛋白質序列、轉錄區序列以及核酸序列內容，flu_acc_map 連結核酸序列、蛋白質序列的存取號碼以及轉錄範圍，flu_pdb 連結 PDB id 及蛋白質序列，flu_proteotype 記錄各病毒株中各基因的 proteotype，flu_drug 紀錄抗藥病毒株於蛋白質序列上發生突變的相關位置，flu_drug_info 紀錄該與抗藥性有關的突變位置其參考文獻以及實驗數據。資料庫採關聯式資料庫設計，實體聯繫模式圖 (Entity-relationship modeling) 如下圖。



系統環境

本系統架設於類 Debian Linux 3 中，網頁服務使用 PHP 5.0.5 版開發，關聯式資料庫使用 MySQL 5.0.32 版，網頁伺服器 Apache 2.2.3，與外部程式串接透過 Perl 版本 5.8.8，Bioperl 模組版本為 1.5 版。

抗原決定位

我們以 PHP 語言撰寫網頁型式的 B-cell epitope 預測工具，預測結果使用 PEAR IMAGE_GRAPH 圖形套件繪製圖形呈現給使用者，並將 IVBS 資料庫中的序列資料與此工具連結，可讓使用者對 IVBS 中的病毒蛋白序列進行 B-cell epitope 預測。而利用 SVM 預測 B-cell

epitope 則使用 LibSVM 工具(www.csie.ntu.edu.tw/~cjlin/libsvm/)進行預測模型的建立，並使用 PHP 將結果呈現於網頁上。

表一、B-cell epitope 預測工具所使用的程式技術與語言

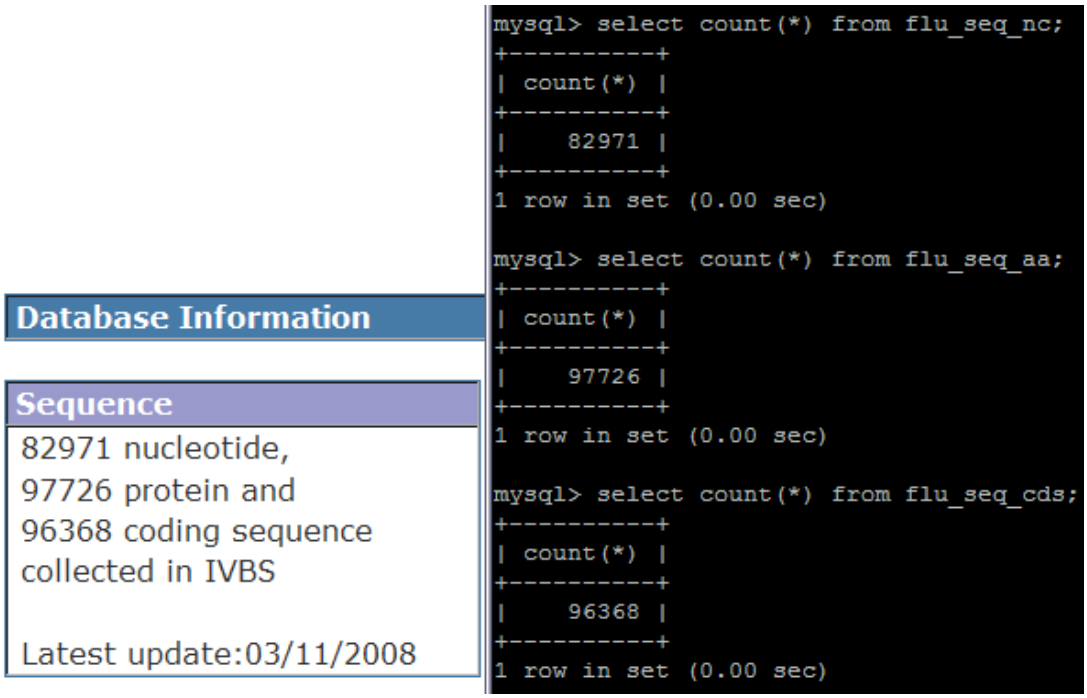
資料處理技術	程式語言名稱與版本
資料庫	MySQL
網頁程式語言	PHP
網頁圖形套件	PEAR:IMAGE_GRAPH
SVM 機器學習套件	LibSVM v2.88

(3) 結果

1. IVBS 流感病毒序列資料之自動更新

→已完成 IVBS 流感病毒序列資料之自動更新

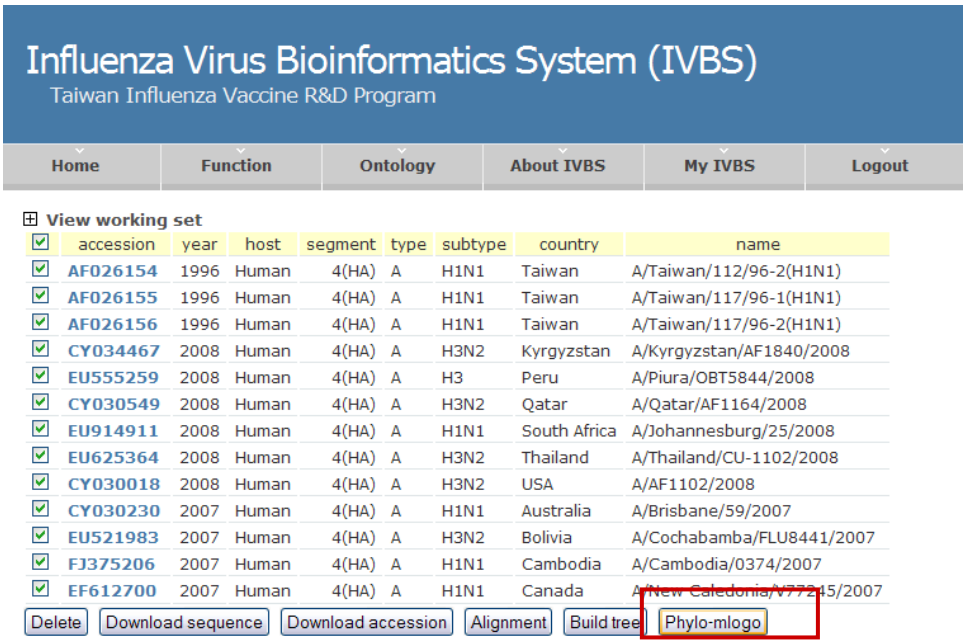
說明: 目前為每天自動更新 NCBI 上流感病毒的序列資料到 IVBS 資料庫中並即時呈現在網頁上。

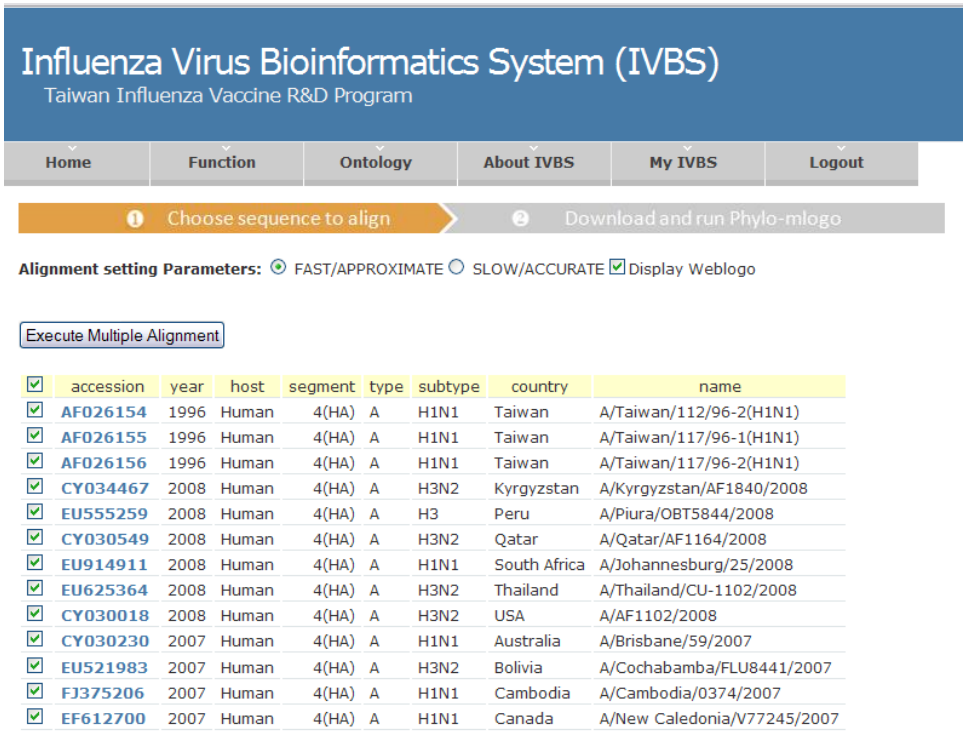


2. 增加以 sequence logo 及階層方式同時呈報大規模流感病毒蛋白質序列排列中氨基酸之相似程度

→已完成 IVBS 與 Phylo-mlogo 連結，增加以 sequence logo 及階層方式同時呈報大規模流感病毒蛋白質序列排列中氨基酸之相似程度。

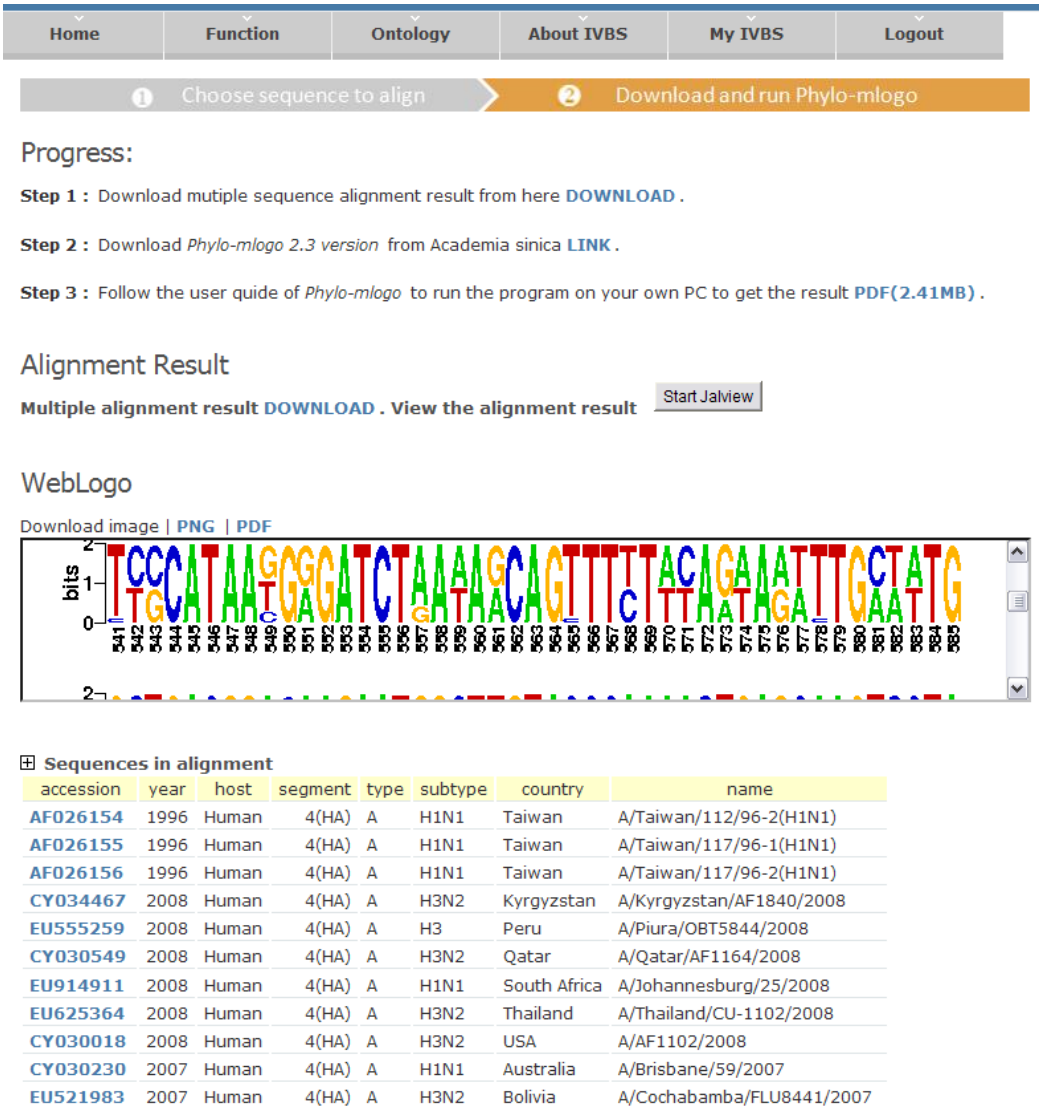
說明: IVBS 系統在現有的 Proteotype 分析，及 Weblogo 呈現 alignment 的結果外，將增加以 sequence logo 及階層方式同時呈報大規模流感病毒蛋白質序列排列中氨基酸之相似程度。目前 IVBS 與 Phylo-mlogo 連結的方式是以外部網頁連結，當使用者查詢完資料並加入到個人工作區後，有一個按鍵「Phylo-mlogo」(如下圖)，由於 Phylo-mlogo 適用的檔案為 alignment 後的結果，所以第一步驟為多序列排比(如下下圖)。



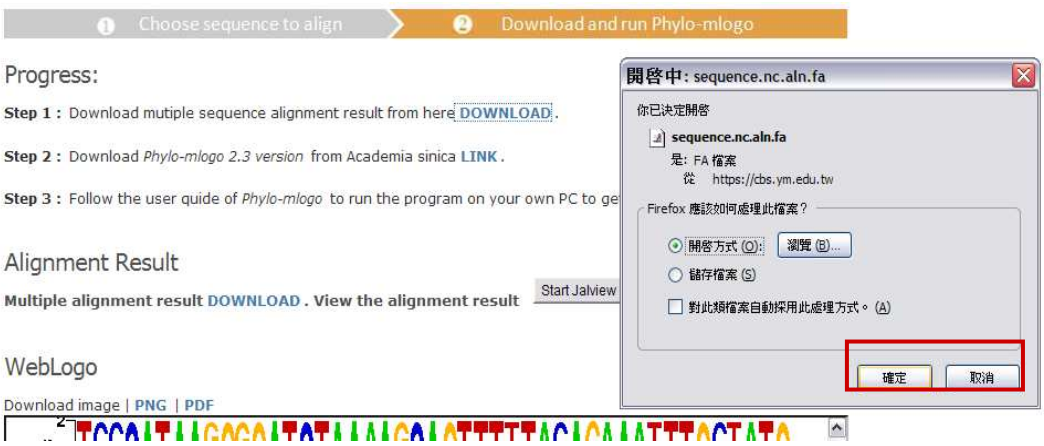


accession	year	host	segment	type	subtype	country	name
AF026154	1996	Human	4(HA)	A	H1N1	Taiwan	A/Taiwan/112/96-2(H1N1)
AF026155	1996	Human	4(HA)	A	H1N1	Taiwan	A/Taiwan/117/96-1(H1N1)
AF026156	1996	Human	4(HA)	A	H1N1	Taiwan	A/Taiwan/117/96-2(H1N1)
CY034467	2008	Human	4(HA)	A	H3N2	Kyrgyzstan	A/Kyrgyzstan/AF1840/2008
EU555259	2008	Human	4(HA)	A	H3	Peru	A/Piura/OBT5844/2008
CY030549	2008	Human	4(HA)	A	H3N2	Qatar	A/Qatar/AF1164/2008
EU914911	2008	Human	4(HA)	A	H1N1	South Africa	A/Johannesburg/25/2008
EU625364	2008	Human	4(HA)	A	H3N2	Thailand	A/Thailand/CU-1102/2008
CY030018	2008	Human	4(HA)	A	H3N2	USA	A/AF1102/2008
CY030230	2007	Human	4(HA)	A	H1N1	Australia	A/Brisbane/59/2007
EU521983	2007	Human	4(HA)	A	H3N2	Bolivia	A/Cochabamba/FLU8441/2007
FJ375206	2007	Human	4(HA)	A	H1N1	Cambodia	A/Cambodia/0374/2007
EF612700	2007	Human	4(HA)	A	H1N1	Canada	A/New Caledonia/V77245/2007

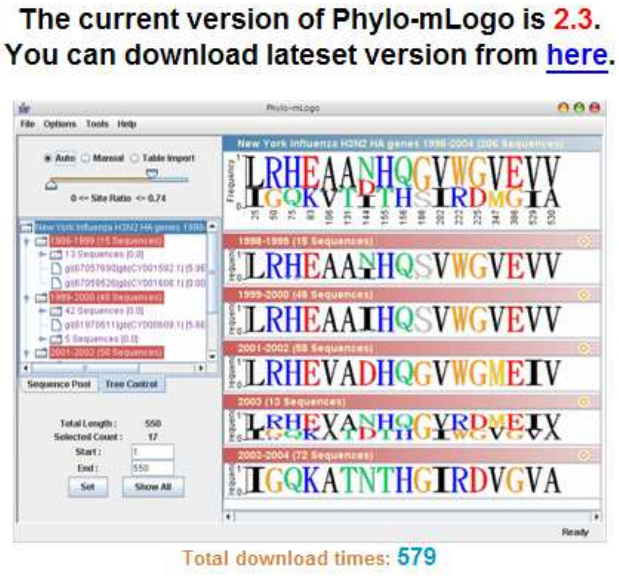
在第二步驟中:



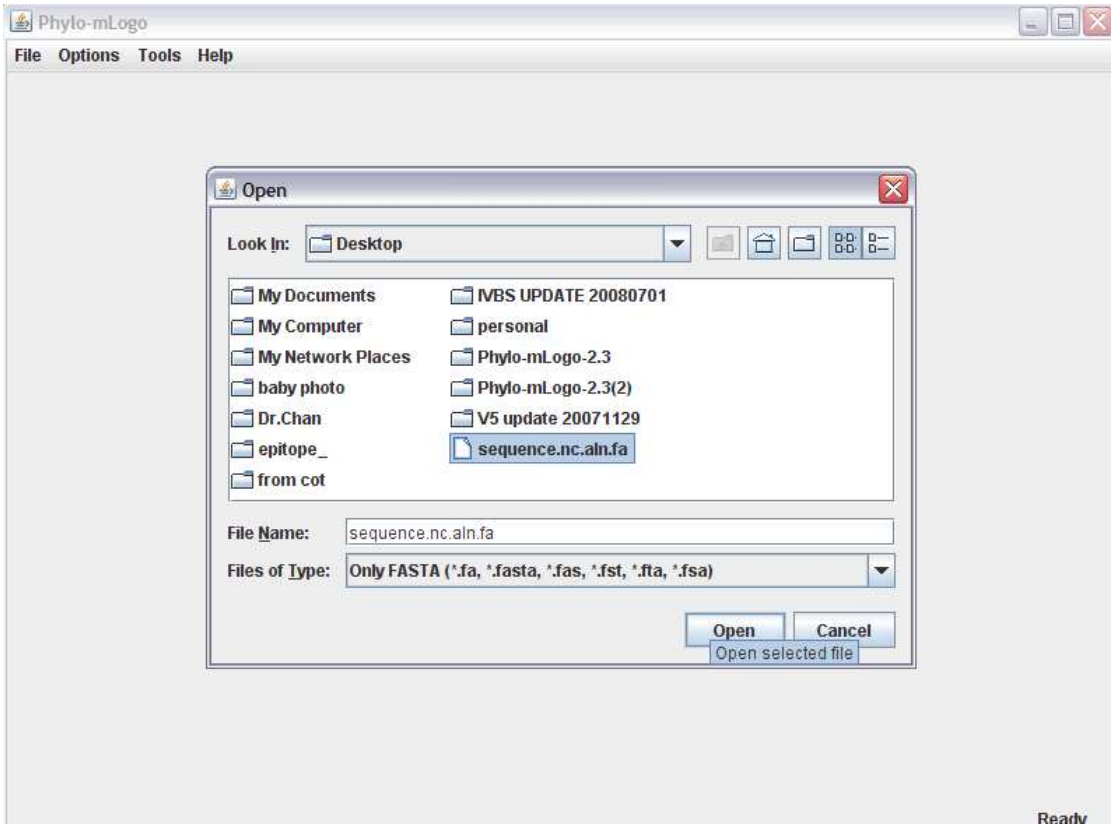
Step 1: 按下 download 鍵，下載在 IVBS 做完多序列排比後的結果

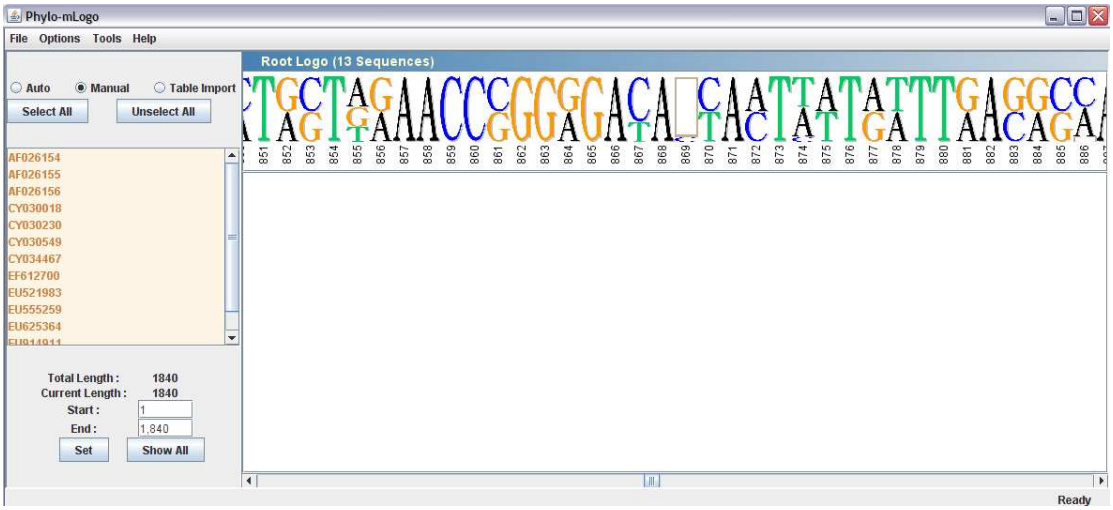


Step 2: 按下 link 鍵，連結到中研院網站下載 Phylo-mlogo 軟體



Step 3: 多序列排比後的結果 upload 到 Phylo-mlogo，按照網頁上所提供的步驟即可看到結果。





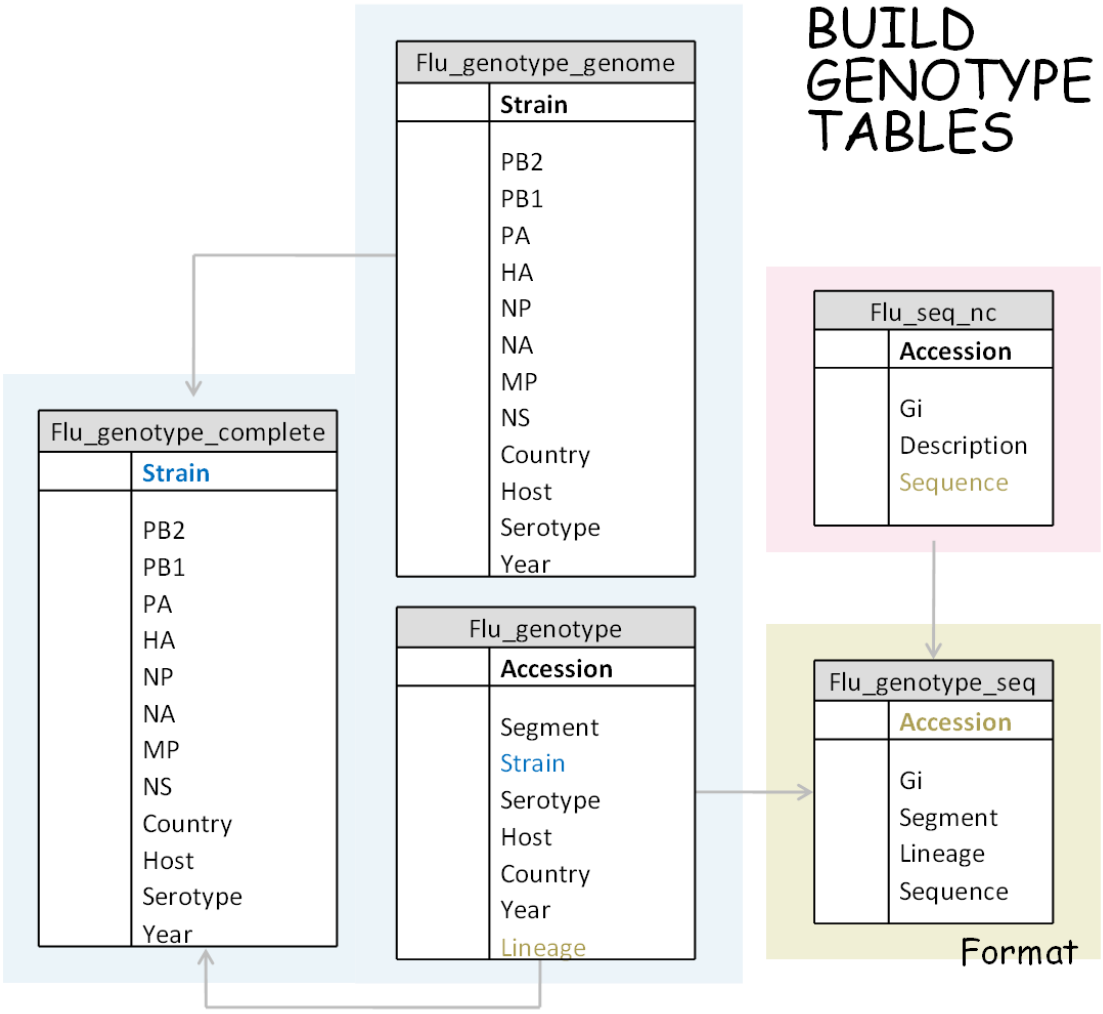
3. 增加流感病毒全基因體之基因型(genotyping) 之預測與分佈功能

→已完成增加流感病毒全基因體之基因型(genotyping) 之預測與分佈功能

說明： IVBS 系統在現有的流感病毒基因序列分析之外，也自 FluGenome 蒐集流感病毒之基因型資訊整合加入 IVBS 中，共 3,678 株流感病毒全基因體之基因型; 及 47,169 筆各 segment 之基因型的資料。其資料如下表所示。

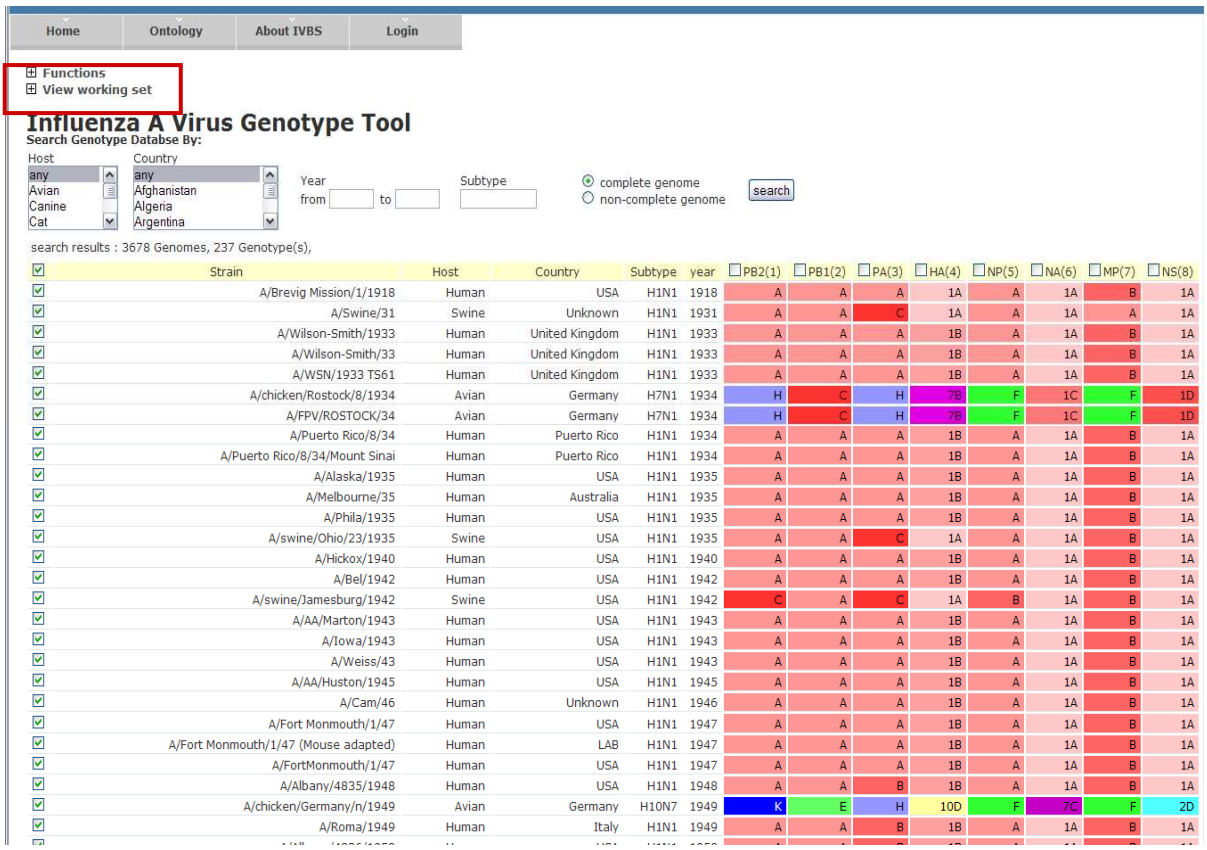
表、IVBS 中各 8 segment 基因型資料之筆數

total	PB2	PB1	PA	HA	NP	NA	MP	NS
47169	4420	4198	4242	10454	4426	6650	5754	5831

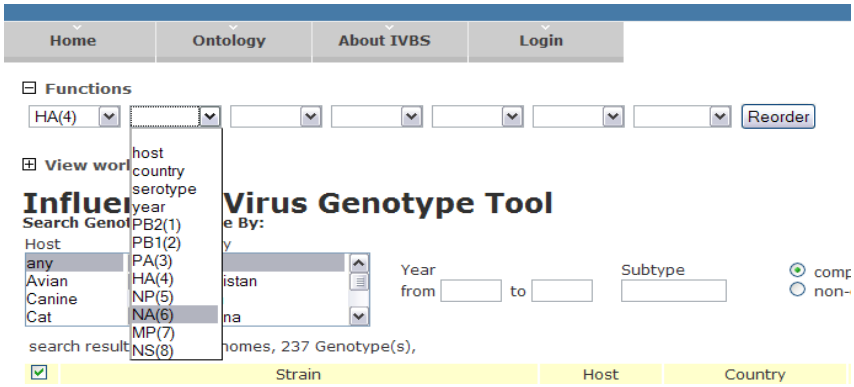


我們利用結合 FluGenome 兩個資料庫: Flu_genotype_genome 和 Flu_genotype，建立了 Genotype 的資料庫: Flu_genotype_complete，並整合了 IVBS 本身的序列資料庫: Flu_seq_nc，建立 Genotype 序列的資料庫: Flu_gebotype_seq，有用於做 Genotype 預測。利用資料表結合、PHP 連接 MySQL，做成網頁。

我們亦完成使用者介面的設計，Genotype 頁面與 Proteotype 相似，利用顏色不同表現病毒株彼此基因型的差異性，使用者可利用頁面上方「Function」功能鍵，選擇一些欄位依照所選順序做重新排列。



舉例說明: 用 HA 為排序的第一順位，NA 為第二順位，進行排序



結果如下:

Home | Ontology | About IVBS | Login

⊞ Functions
⊞ View working set

Influenza A Virus Genotype Tool

Search Genotype Database By:

Host: any / Avian / Canine / Cat

Country: any / Afghanistan / Algeria / Argentina

Year from ____ to ____

Subtype ____

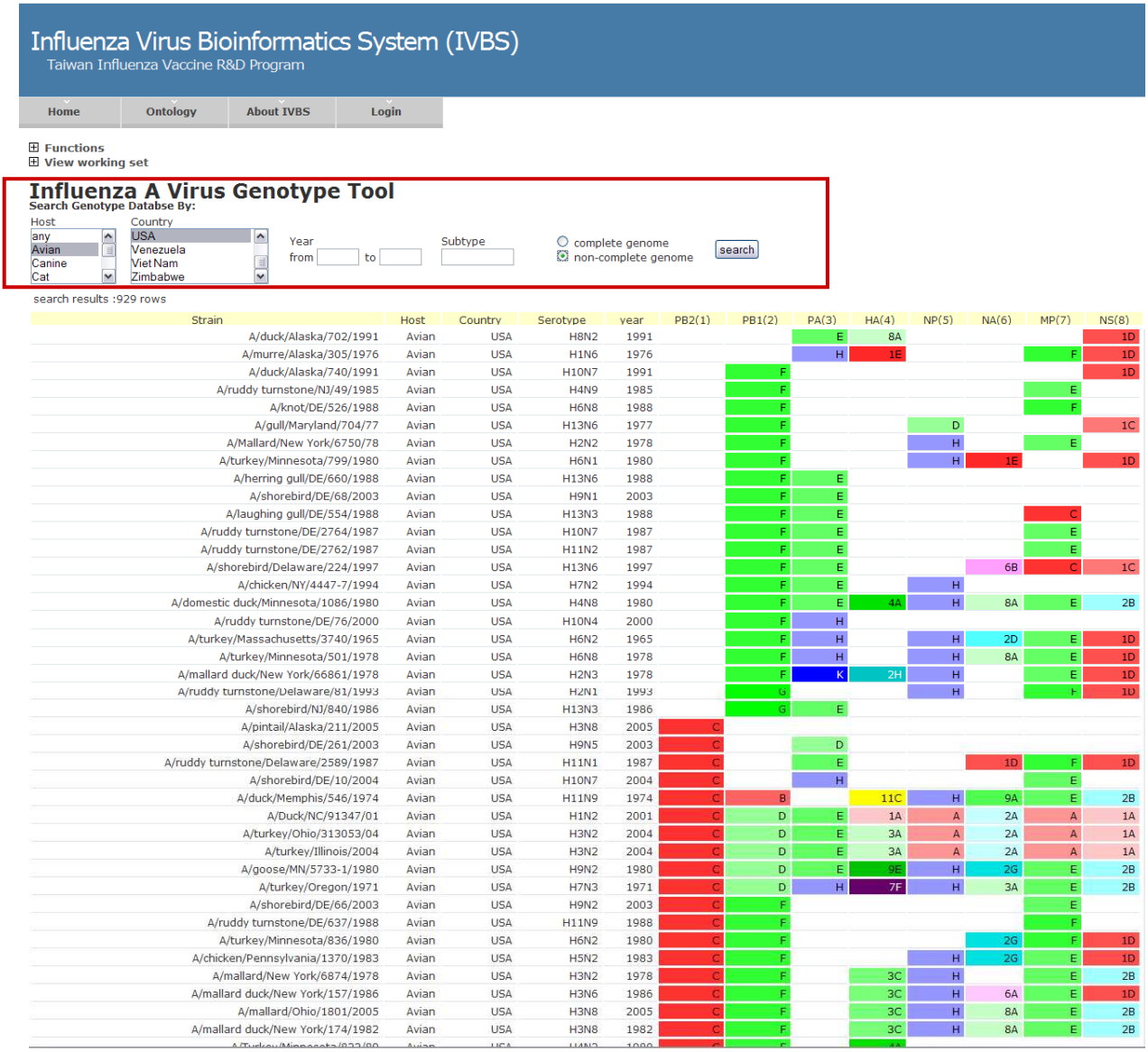
◉ complete genome
○ non-complete genome

search

search results : 3678 Genomes, 237 Genotype(s),

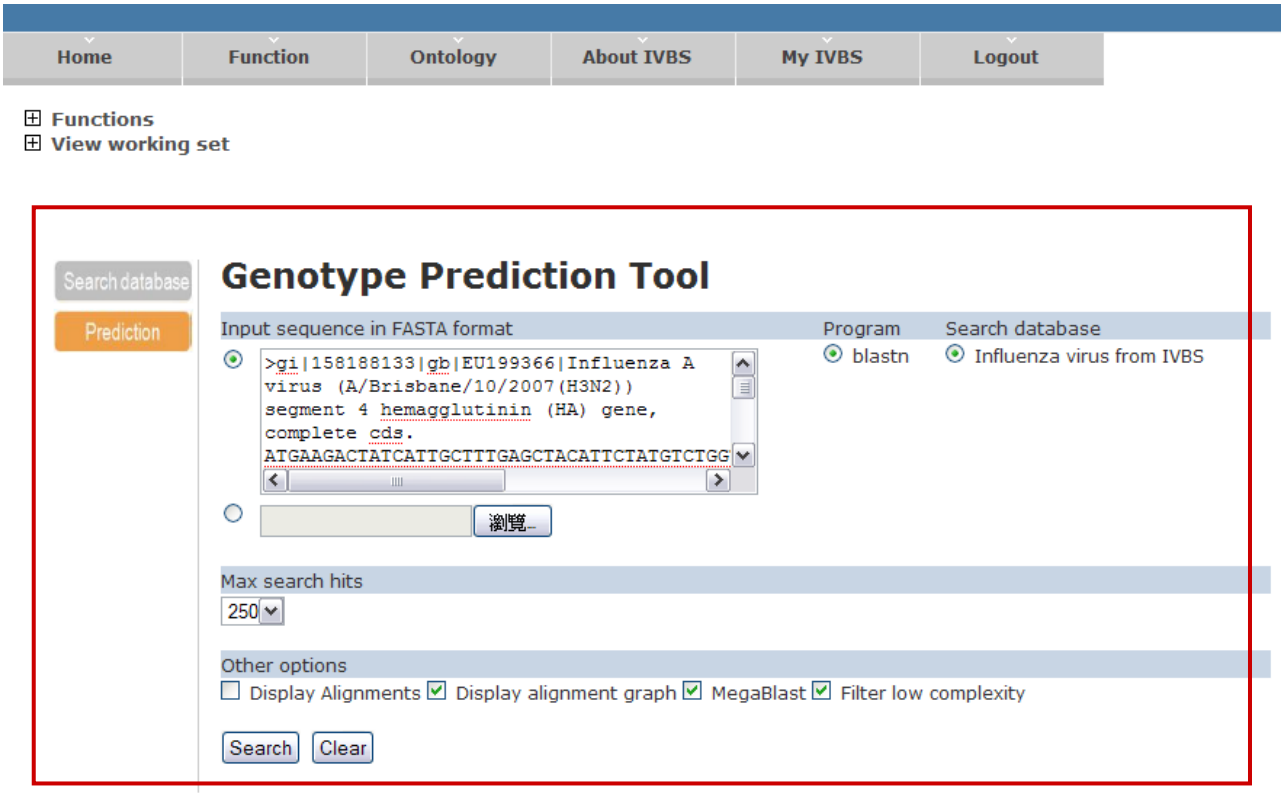
☑	Strain	Host	Country	Subtype	year	☐PB2(1)	☐PB1(2)	☐PA(3)	☐HA(4)	☐NP(5)	☐NA(6)	☐MP(7)	☐NS(8)
☑	A/mallard/Alberta/208/2000	Avian	Canada	H10N7	2000	C	F	H	10A	H	7F	E	1D
☑	A/mallard/Alberta/209/2003	Avian	Canada	H10N7	2003	C	F	H	10A	H	7F	E	1D
☑	A/pintail/Alberta/202/2000	Avian	Canada	H10N7	2000	C	F	H	10A	H	7F	E	2B
☑	A/mallard duck/Minnesota/19/1979	Avian	USA	H10N7	1979	C	F	H	10A	H	7G	E	1D
☑	A/mallard/Ohio/122/1989	Avian	USA	H10N7	1989	C	F	E	10A	H	7G	E	1D
☑	A/mallard/Ohio/99/1989	Avian	USA	H10N7	1989	C	F	E	10A	H	7G	E	1D
☑	A/common scoter/Maryland/297/2005	Avian	USA	H10N8	2005	C	F	E	10A	H	8A	E	1D
☑	A/longtail duck/Maryland/295/2005	Avian	USA	H10N8	2005	C	F	H	10A	H	8A	E	1D
☑	A/chicken/Germany/n/1949	Avian	Germany	H10N7	1949	K	E	H	10D	F	7C	F	2D
☑	A/quail/Italy/1117/1965	Avian	Italy	H10N8	1965	K	E	I	10D	F	8C	F	1D
☑	A/duck/Hong Kong/562/1979	Avian	Hong Kong	H10N9	1979	K	G	E	10E	F	9B	F	1E
☑	A/sharp-tailed sandpiper/Australia/10/2004	Avian	Australia	H11N9	2004	G	G	D	11A	F	9A	F	1E
☑	A/sharp-tailed sandpiper/Australia/6/2004	Avian	Australia	H11N9	2004	G	G	D	11A	F	9A	F	1E
☑	A/shoveler/Netherlands/19/1999	Avian	Netherlands	H11N9	1999	G	G	D	11A	F	9B	F	1E
☑	A/duck/England/1956	Avian	United Kingdom	H11N6	1956	K	E	E	11B	F	6D	F	1E
☑	A/Black Duck/Ohio/194/1986	Avian	USA	H11N1	1986	C	F	E	11C	H	1E	E	1D
☑	A/mallard/Ohio/1851/2005	Avian	USA	H11N1	2005	C	F	E	11C	H	1E	E	1D
☑	A/mallard/Ohio/1851/2005	Avian	USA	H11N9	2005	C	F	E	11C	H	1E	E	1D
☑	A/green-winged teal/Ohio/1747/2005	Avian	USA	H11N2	2005	C	F	E	11C	H	2D	E	1D
☑	A/bufflehead/Ohio/246/1986	Avian	USA	H11N2	1986	C	F	E	11C	H	2G	E	2B
☑	A/green-winged teal/Ohio/81/1999	Avian	USA	H11N2	1999	C	F	E	11C	H	2G	E	1D
☑	A/mallard/Missouri/MO130/2005	Avian	USA	H11N3	2005	C	F	E	11C	H	3A	E	2B
☑	A/mallard/Ohio/102/1986	Avian	USA	H11N3	1986	C	F	E	11C	H	3A	E	1D
☑	A/mallard/Ohio/94/1993	Avian	USA	H11N3	1993	C	F	E	11C	H	3A	E	2B
☑	A/environment/Delaware/235/2005	Environment	USA	H11N6	2005	C	F	E	11C	H	6A	E	2B
☑	A/environment/Delaware/232/2005	Environment	USA	H11N8	2005	C	F	E	11C	H	8A	E	2B
☑	A/environment/Delaware/234/2005	Environment	USA	H11N8	2005	C	F	E	11C	H	8A	E	2B
☑	A/american black duck/Ohio/1823/2005	Avian	USA	H11N9	2005	C	F	E	11C	H	9A	E	1D

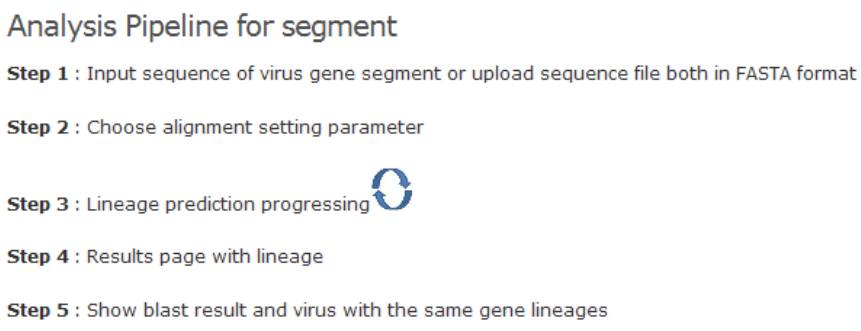
同時把 complete genome 與 non-complete genome 之所有序列的基因型整合在同一個頁面，也可利用頁面上的搜尋功能依照使用者所選擇的條件查找 genotype 資料庫內的資料且顯示出來。



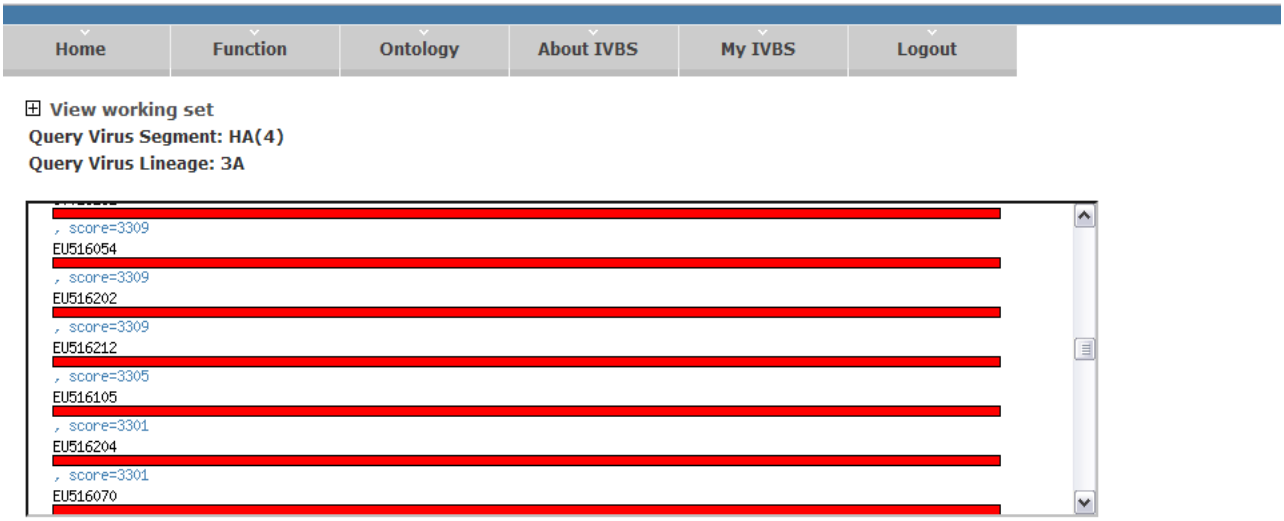
Strain	Host	Country	Serotype	year	PB2(1)	PB1(2)	PA(3)	HA(4)	NP(5)	NA(6)	MP(7)	NS(8)
A/duck/Alaska/702/1991	Avian	USA	H8N2	1991			E	8A				1D
A/murre/Alaska/305/1976	Avian	USA	H1N6	1976			H	1E			F	1D
A/duck/Alaska/740/1991	Avian	USA	H10N7	1991		F						1D
A/ruddy turnstone/NJ/49/1985	Avian	USA	H4N9	1985		F					E	
A/knot/DE/526/1988	Avian	USA	H6N8	1988		F					F	
A/gull/Maryland/704/77	Avian	USA	H13N6	1977		F			D			1C
A/Mallard/New York/6750/78	Avian	USA	H2N2	1978		F			H		E	
A/turkey/Minnesota/799/1980	Avian	USA	H6N1	1980		F			H	1E		1D
A/herring gull/DE/660/1988	Avian	USA	H13N6	1988		F	E					
A/shorebird/DE/68/2003	Avian	USA	H9N1	2003		F	E					
A/laughing gull/DE/554/1988	Avian	USA	H13N3	1988		F	E				C	
A/ruddy turnstone/DE/2764/1987	Avian	USA	H10N7	1987		F	E				E	
A/ruddy turnstone/DE/2762/1987	Avian	USA	H11N2	1987		F	E				E	
A/shorebird/Delaware/224/1997	Avian	USA	H13N6	1997		F	E			6B	C	1C
A/chicken/NY/4447-7/1994	Avian	USA	H7N2	1994		F	E		H			
A/domestic duck/Minnesota/1086/1980	Avian	USA	H4N8	1980		F	E	4A	H	8A	E	2B
A/ruddy turnstone/DE/76/2000	Avian	USA	H10N4	2000		F	H					
A/turkey/Massachusetts/3740/1965	Avian	USA	H6N2	1965		F	H		H	2D	E	1D
A/turkey/Minnesota/501/1978	Avian	USA	H6N8	1978		F	H		H	8A	E	1D
A/mallard duck/New York/66861/1978	Avian	USA	H2N3	1978		F	K	2H	H		E	1D
A/ruddy turnstone/Delaware/81/1993	Avian	USA	H2N1	1993		G			H		F	1D
A/shorebird/NJ/840/1986	Avian	USA	H13N3	1986		G	E					
A/pintail/Alaska/211/2005	Avian	USA	H3N8	2005	C							
A/shorebird/DE/261/2003	Avian	USA	H9N5	2003	C		D					
A/ruddy turnstone/Delaware/2589/1987	Avian	USA	H11N1	1987	C		E			1D	F	1D
A/shorebird/DE/10/2004	Avian	USA	H10N7	2004	C		H				E	
A/duck/Memphis/546/1974	Avian	USA	H11N9	1974	C	B		11C	H	9A	E	2B
A/Duck/NC/91347/01	Avian	USA	H1N2	2001	C	D	E	1A	A	2A	A	1A
A/turkey/Ohio/313053/04	Avian	USA	H3N2	2004	C	D	E	3A	A	2A	A	1A
A/turkey/Illinois/2004	Avian	USA	H3N2	2004	C	D	E	3A	A	2A	A	1A
A/goose/MN/5733-1/1980	Avian	USA	H9N2	1980	C	D	E	9E	H	2G	E	2B
A/turkey/Oregon/1971	Avian	USA	H7N3	1971	C	D	H	7F	H	3A	E	2B
A/shorebird/DE/66/2003	Avian	USA	H9N2	2003	C	F					E	
A/ruddy turnstone/DE/637/1988	Avian	USA	H11N9	1988	C	F					F	
A/turkey/Minnesota/836/1980	Avian	USA	H6N2	1980	C	F				2G	F	1D
A/chicken/Pennsylvania/1370/1983	Avian	USA	H5N2	1983	C	F			H	2G	E	1D
A/mallard/New York/6874/1978	Avian	USA	H3N2	1978	C	F		3C	H		E	2B
A/mallard duck/New York/157/1986	Avian	USA	H3N6	1986	C	F		3C	H	6A	E	1D
A/mallard/Ohio/1801/2005	Avian	USA	H3N8	2005	C	F		3C	H	8A	E	2B
A/mallard duck/New York/174/1982	Avian	USA	H3N8	1982	C	F		3C	H	8A	E	2B

接下來又新增流感病毒基因型 genotyping 預測功能，利用 BLAST 來比對序列相似度以預測序列之 genotype。第一步，使用者可以輸入一段流感病毒序列，或是上傳一段序列的 FASTA 檔；第二步，選擇 BLAST 搜尋結果最多幾個，是否顯示排序後的圖示結果..等；第三步，按下搜尋，genotype 預測進行中；第四步，預測結果產生頁面。頁面下方也有預測 genotype 的 pipeline。



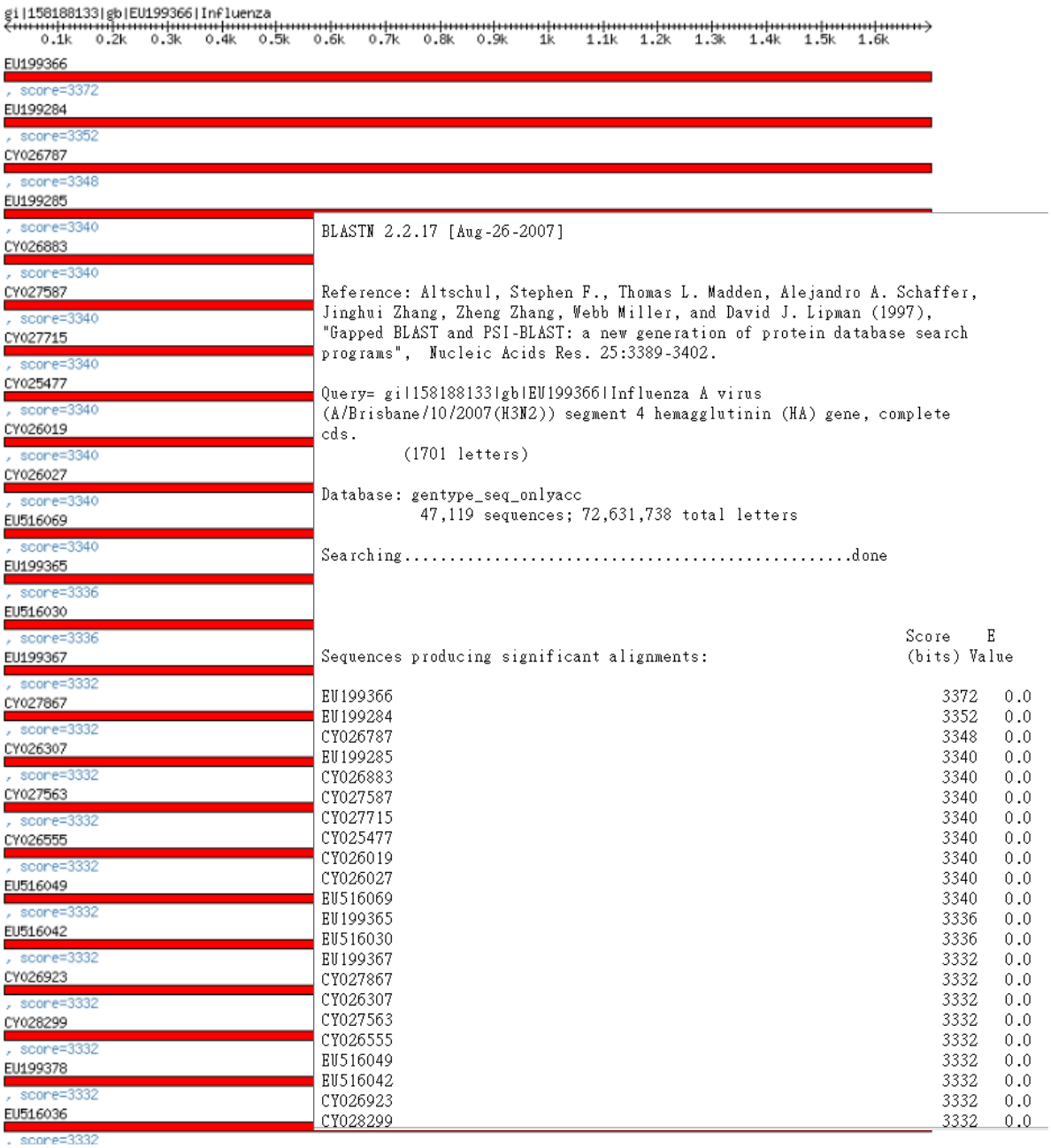


預測結果如圖顯示，這條序列是屬於 HA(4)，序列之 lineage 為 3A。頁面中間為 BLAST 的結果，頁面下方則是列出 IVBS 中之相似序列列表。同時，搜尋結果可透過加入工作區，依需要下載其序列，或以系統提供的工具如執行多序列排比、親緣樹建立、抗藥性再做進一步的分析。

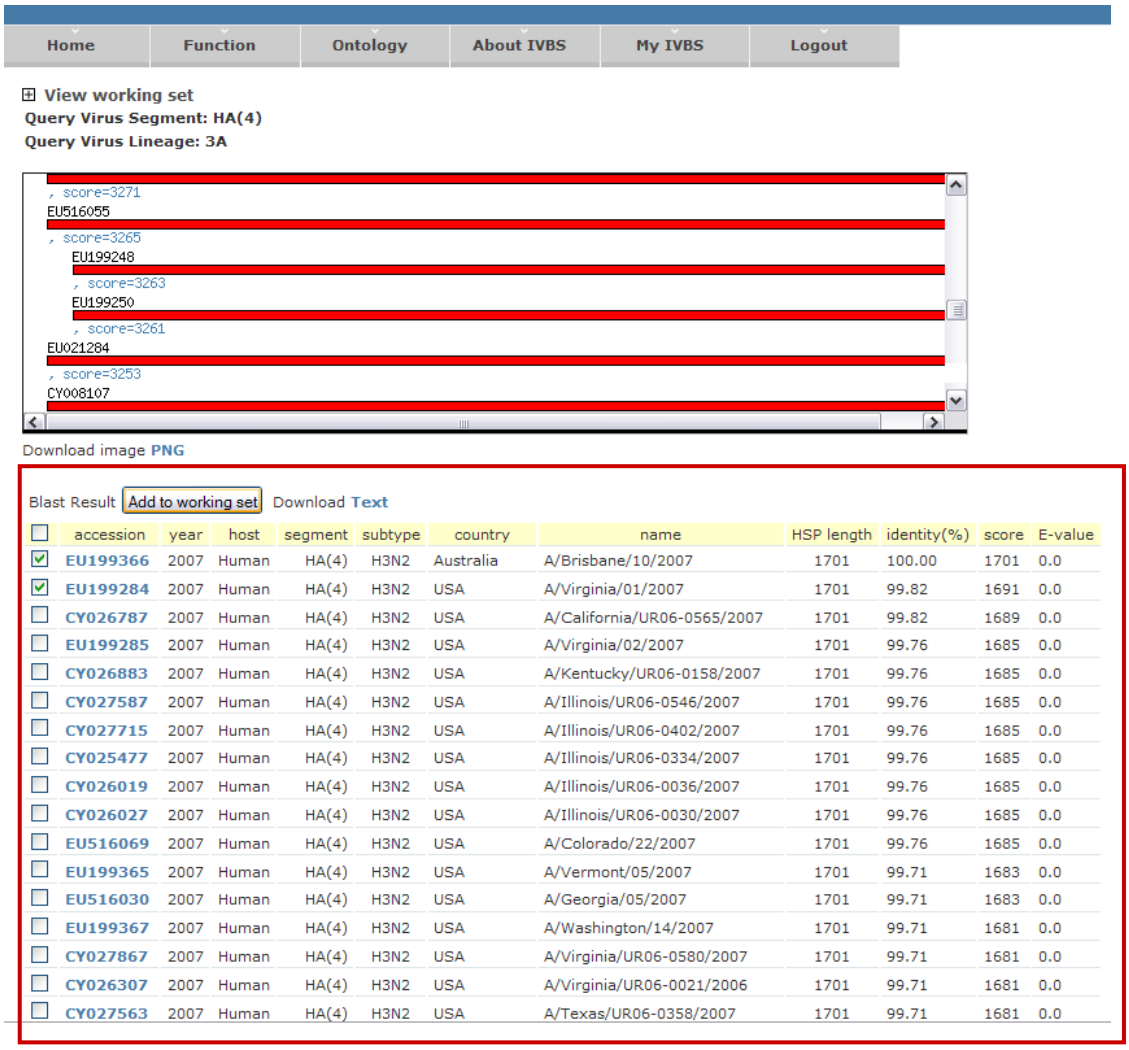


Blast Result Add to working set Download Text

accession	year	host	segment	subtype	country	name	HSP length	identity(%)	score	E-value
EU199366	2007	Human	HA(4)	H3N2	Australia	A/Brisbane/10/2007	1701	100.00	1701	0.0
EU199284	2007	Human	HA(4)	H3N2	USA	A/Virginia/01/2007	1701	99.82	1691	0.0
CY026787	2007	Human	HA(4)	H3N2	USA	A/California/UR06-0565/2007	1701	99.82	1689	0.0
EU199285	2007	Human	HA(4)	H3N2	USA	A/Virginia/02/2007	1701	99.76	1685	0.0
CY026883	2007	Human	HA(4)	H3N2	USA	A/Kentucky/UR06-0158/2007	1701	99.76	1685	0.0
CY027587	2007	Human	HA(4)	H3N2	USA	A/Illinois/UR06-0546/2007	1701	99.76	1685	0.0
CY027715	2007	Human	HA(4)	H3N2	USA	A/Illinois/UR06-0402/2007	1701	99.76	1685	0.0
CY025477	2007	Human	HA(4)	H3N2	USA	A/Illinois/UR06-0334/2007	1701	99.76	1685	0.0
CY026019	2007	Human	HA(4)	H3N2	USA	A/Illinois/UR06-0036/2007	1701	99.76	1685	0.0
CY026027	2007	Human	HA(4)	H3N2	USA	A/Illinois/UR06-0030/2007	1701	99.76	1685	0.0
EU516069	2007	Human	HA(4)	H3N2	USA	A/Colorado/22/2007	1701	99.76	1685	0.0
EU199365	2007	Human	HA(4)	H3N2	USA	A/Vermont/05/2007	1701	99.71	1683	0.0
EU516030	2007	Human	HA(4)	H3N2	USA	A/Georgia/05/2007	1701	99.71	1683	0.0
EU199367	2007	Human	HA(4)	H3N2	USA	A/Washington/14/2007	1701	99.71	1681	0.0
CY027867	2007	Human	HA(4)	H3N2	USA	A/Virginia/UR06-0580/2007	1701	99.71	1681	0.0
CY026307	2007	Human	HA(4)	H3N2	USA	A/Virginia/UR06-0021/2006	1701	99.71	1681	0.0
CY027563	2007	Human	HA(4)	H3N2	USA	A/Texas/UR06-0358/2007	1701	99.71	1681	0.0

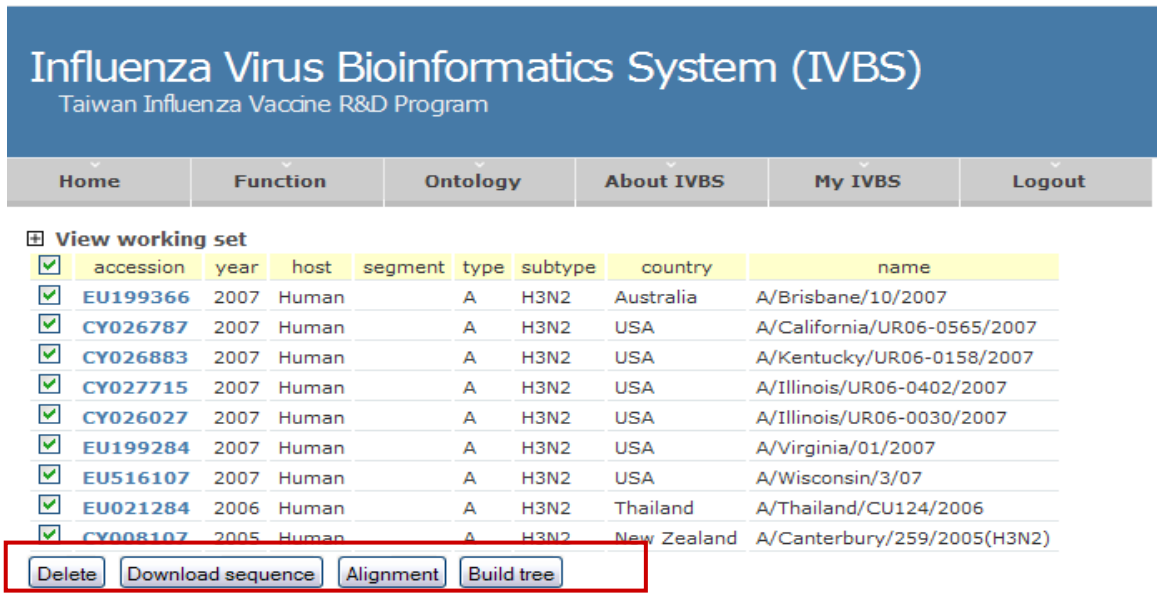


BLAST 後的結果都可經由網頁上的超連結下載來看。



accession	year	host	segment	subtype	country	name	HSP length	identity(%)	score	E-value
EU199366	2007	Human	HA(4)	H3N2	Australia	A/Brisbane/10/2007	1701	100.00	1701	0.0
EU199284	2007	Human	HA(4)	H3N2	USA	A/Virginia/01/2007	1701	99.82	1691	0.0
CY026787	2007	Human	HA(4)	H3N2	USA	A/California/UR06-0565/2007	1701	99.82	1689	0.0
EU199285	2007	Human	HA(4)	H3N2	USA	A/Virginia/02/2007	1701	99.76	1685	0.0
CY026883	2007	Human	HA(4)	H3N2	USA	A/Kentucky/UR06-0158/2007	1701	99.76	1685	0.0
CY027587	2007	Human	HA(4)	H3N2	USA	A/Illinois/UR06-0546/2007	1701	99.76	1685	0.0
CY027715	2007	Human	HA(4)	H3N2	USA	A/Illinois/UR06-0402/2007	1701	99.76	1685	0.0
CY025477	2007	Human	HA(4)	H3N2	USA	A/Illinois/UR06-0334/2007	1701	99.76	1685	0.0
CY026019	2007	Human	HA(4)	H3N2	USA	A/Illinois/UR06-0036/2007	1701	99.76	1685	0.0
CY026027	2007	Human	HA(4)	H3N2	USA	A/Illinois/UR06-0030/2007	1701	99.76	1685	0.0
EU516069	2007	Human	HA(4)	H3N2	USA	A/Colorado/22/2007	1701	99.76	1685	0.0
EU199365	2007	Human	HA(4)	H3N2	USA	A/Vermont/05/2007	1701	99.71	1683	0.0
EU516030	2007	Human	HA(4)	H3N2	USA	A/Georgia/05/2007	1701	99.71	1683	0.0
EU199367	2007	Human	HA(4)	H3N2	USA	A/Washington/14/2007	1701	99.71	1681	0.0
CY027867	2007	Human	HA(4)	H3N2	USA	A/Virginia/UR06-0580/2007	1701	99.71	1681	0.0
CY026307	2007	Human	HA(4)	H3N2	USA	A/Virginia/UR06-0021/2006	1701	99.71	1681	0.0
CY027563	2007	Human	HA(4)	H3N2	USA	A/Texas/UR06-0358/2007	1701	99.71	1681	0.0

針對 BLAST 搜尋後的結果，對於有興趣的序列資料，也可經由勾選加入個人工作區做後續的多序列排比及建立親緣樹等分析。



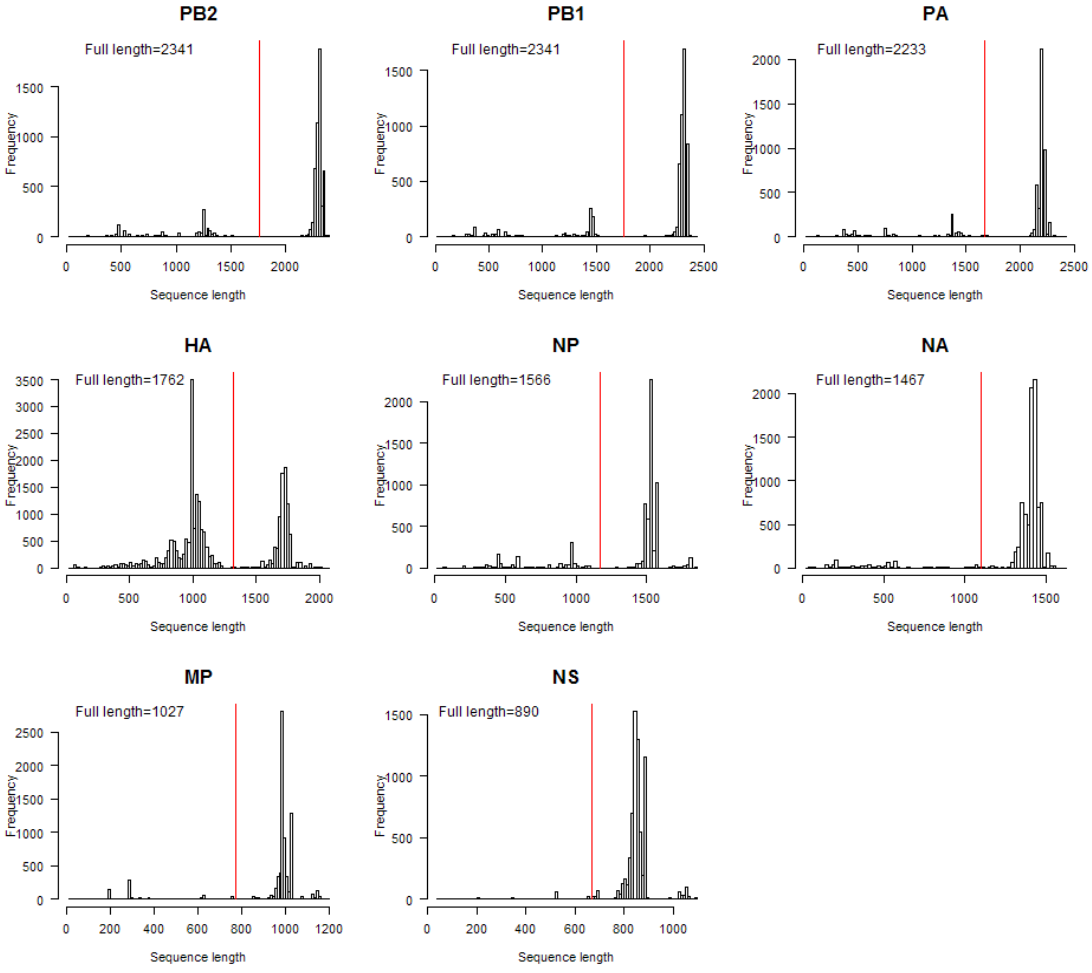
accession	year	host	segment	type	subtype	country	name
EU199366	2007	Human		A	H3N2	Australia	A/Brisbane/10/2007
CY026787	2007	Human		A	H3N2	USA	A/California/UR06-0565/2007
CY026883	2007	Human		A	H3N2	USA	A/Kentucky/UR06-0158/2007
CY027715	2007	Human		A	H3N2	USA	A/Illinois/UR06-0402/2007
CY026027	2007	Human		A	H3N2	USA	A/Illinois/UR06-0030/2007
EU199284	2007	Human		A	H3N2	USA	A/Virginia/01/2007
EU516107	2007	Human		A	H3N2	USA	A/Wisconsin/3/07
EU021284	2006	Human		A	H3N2	Thailand	A/Thailand/CU124/2006
CY008107	2005	Human		A	H3N2	New Zealand	A/Canterbury/259/2005(H3N2)

4. 分析 IVBS 現有序列資料訂出 genotype

→即將分析 IVBS 現有序列資料訂出 genotype

說明: 目前流感病毒之基因型資訊(genotype)是自 FluGenome 蒐集整合加入 IVBS 中。但由於 IVBS 有涵蓋 ISD 及 96 年度向 CDC 疾管局申請的流感序列資料，所以現在利用 IVBS 中的序列自己訂出 genotype，可分析台灣流感病毒的 genotype 之特性。

Genotyping 步驟: 第一步，蒐集流感病毒之序列(5/12 止)



從 IVBS 中，我們只取序列長度大於 75%的流感病毒序列，並扣除掉 100% 一模一樣的序列。

表、自 IVBS 取得各 8 segment 資料做基因型之筆數								
	PB2	**PB1**	**PA**	**HA**	**NP**	**NA**	**MP**	**NS**
total	6045	5913	5835	19090	6503	8682	7617	6803
>75%	4966	4840	4857	11899	5366	8334	7067	6764
-100% identical	4890	4713	4730	8121	5210	7928	6713	6608

第二步，多序列排比，使用 ClustalW。

第三步，建立親緣關係樹，使用 PHYLIP 套件中的”dnadist”和”neighbor”。

1. ”dnadist” : Molecular sequence methods。 Program to compute distance matrix from nucleotide sequences
2. 和”neighbor” : Distance matrix methods。This program implements the Neighbor-Joining method of Saitou and Nei (1987) and the UPGMA method of clustering

第四步，訂 Genotype，將與 Flugenome 做比較。

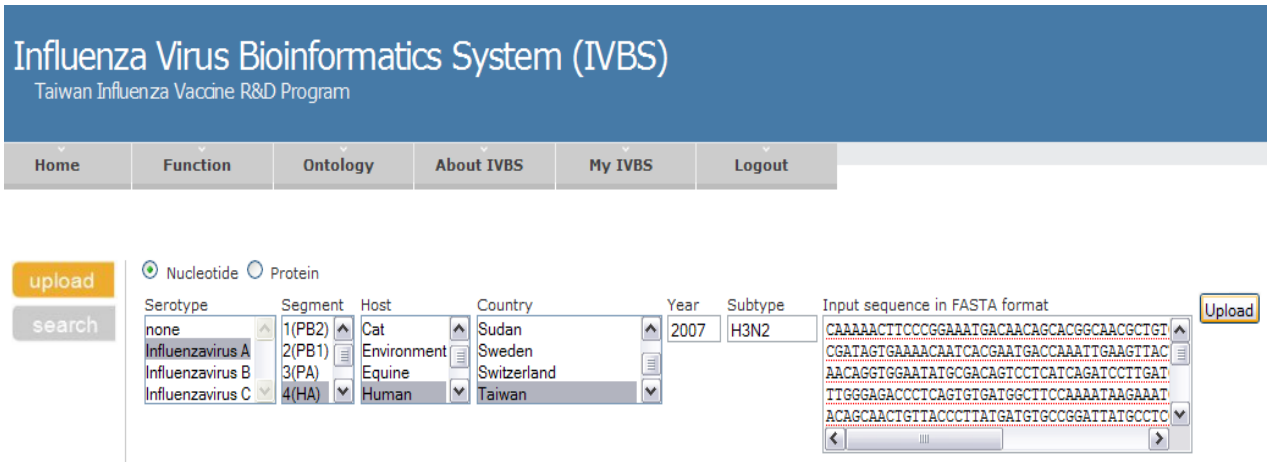
5. 增加使用者個人流感病毒基因序列資料之輸入儲存與後續分析功能

→已完成增加使用者個人流感病毒基因序列資料之輸入儲存與後續分析功能

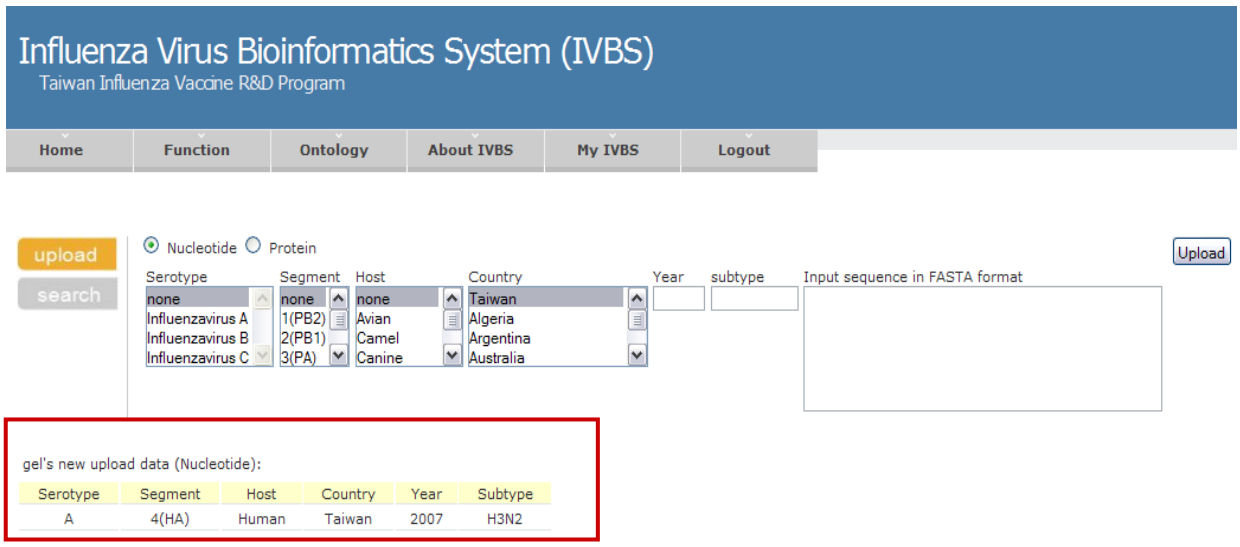
說明: 現有之 IVBS 流感病毒基因序列資料庫蒐集網際網路上所有公開的流感病毒序列資料包括 NCBI(IRV)、LANL(ISD)及台灣疾病管制局，我們增加使用者個人的資料庫輸入儲存與後續分析功能，讓使用

者擁有屬於自己**個人**的資料庫。

在 My IVBS 的頁面中有兩個部分，第一部份，讓使用者可以自行輸入自己實驗室分離出來的流感病毒序列到個人的資料庫中，輸入關於序列的一些資訊: Serotype、Segment、Host、Country..，按下「Upload」。

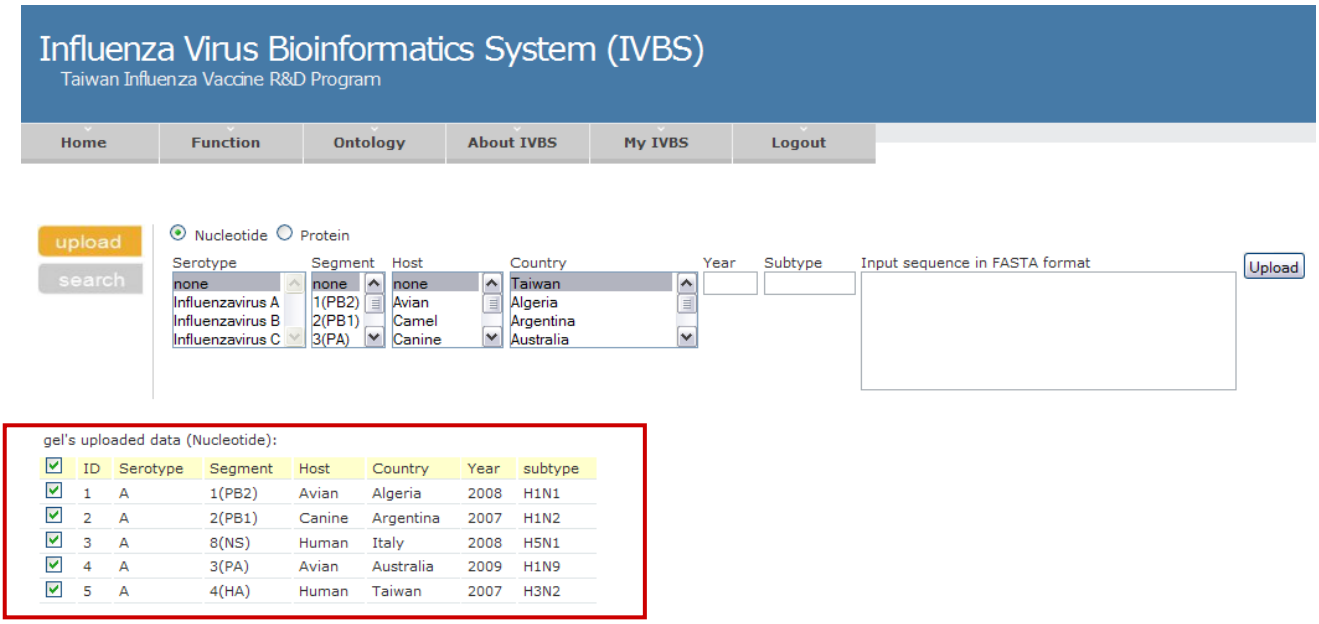


舉例說明：輸入一條 HA segment，Influenza A virus (A/Taiwan/56/2007(H3N2))的序列，按下「Upload」。

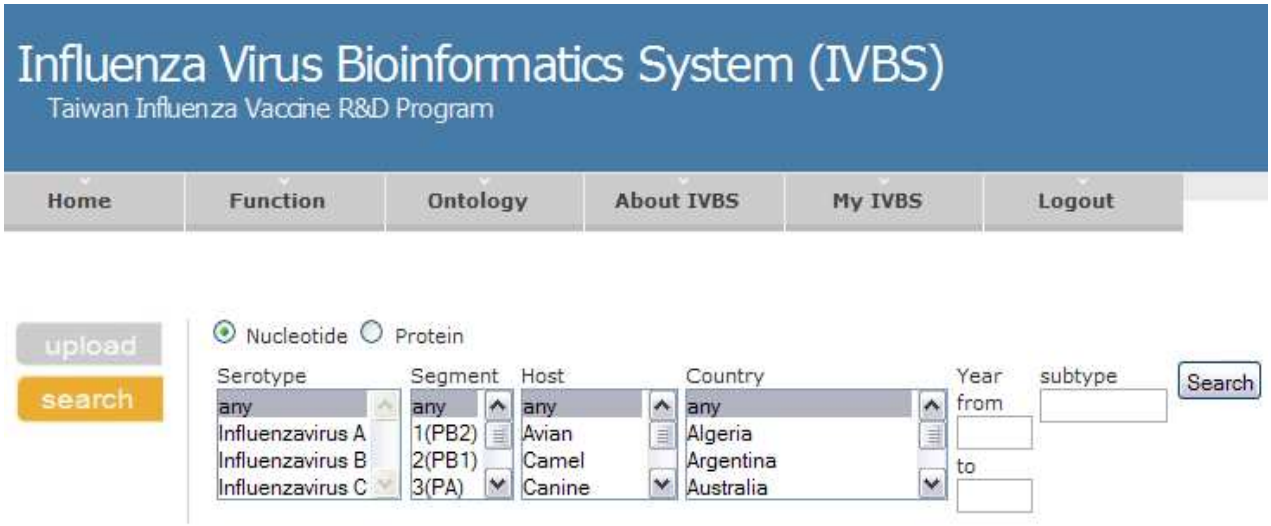


即可將輸入之序列寫入您**專屬的序列資料庫**以做後續分析，頁面下方將會顯示輸入成功的序列資料。

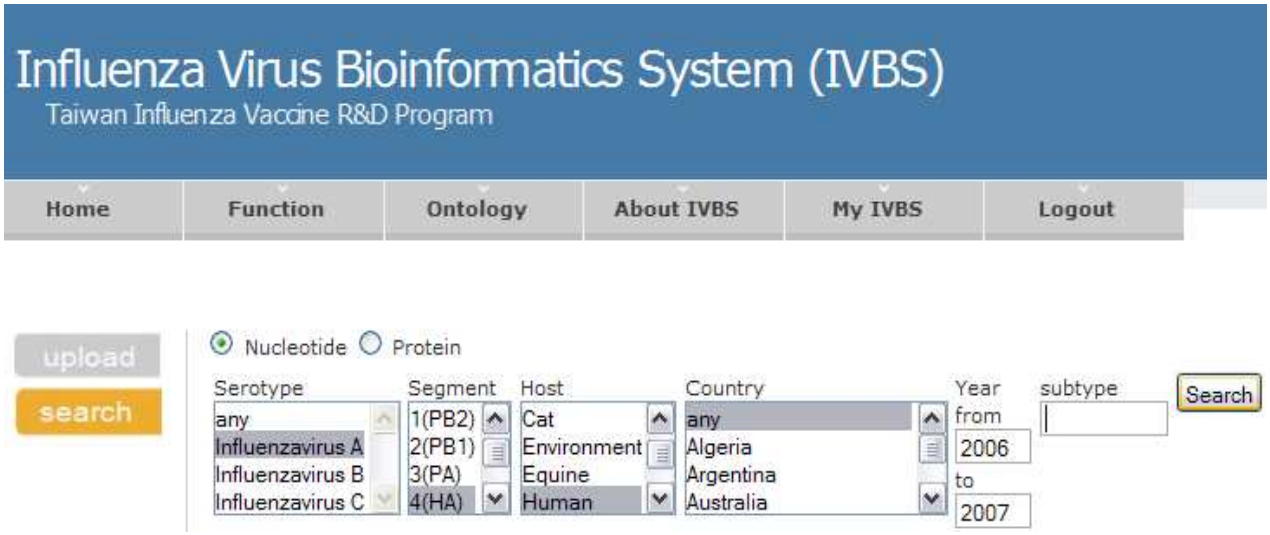
若點回 My IVBS 的首頁，將可看到您**個人**資料庫中所有輸入之序列。

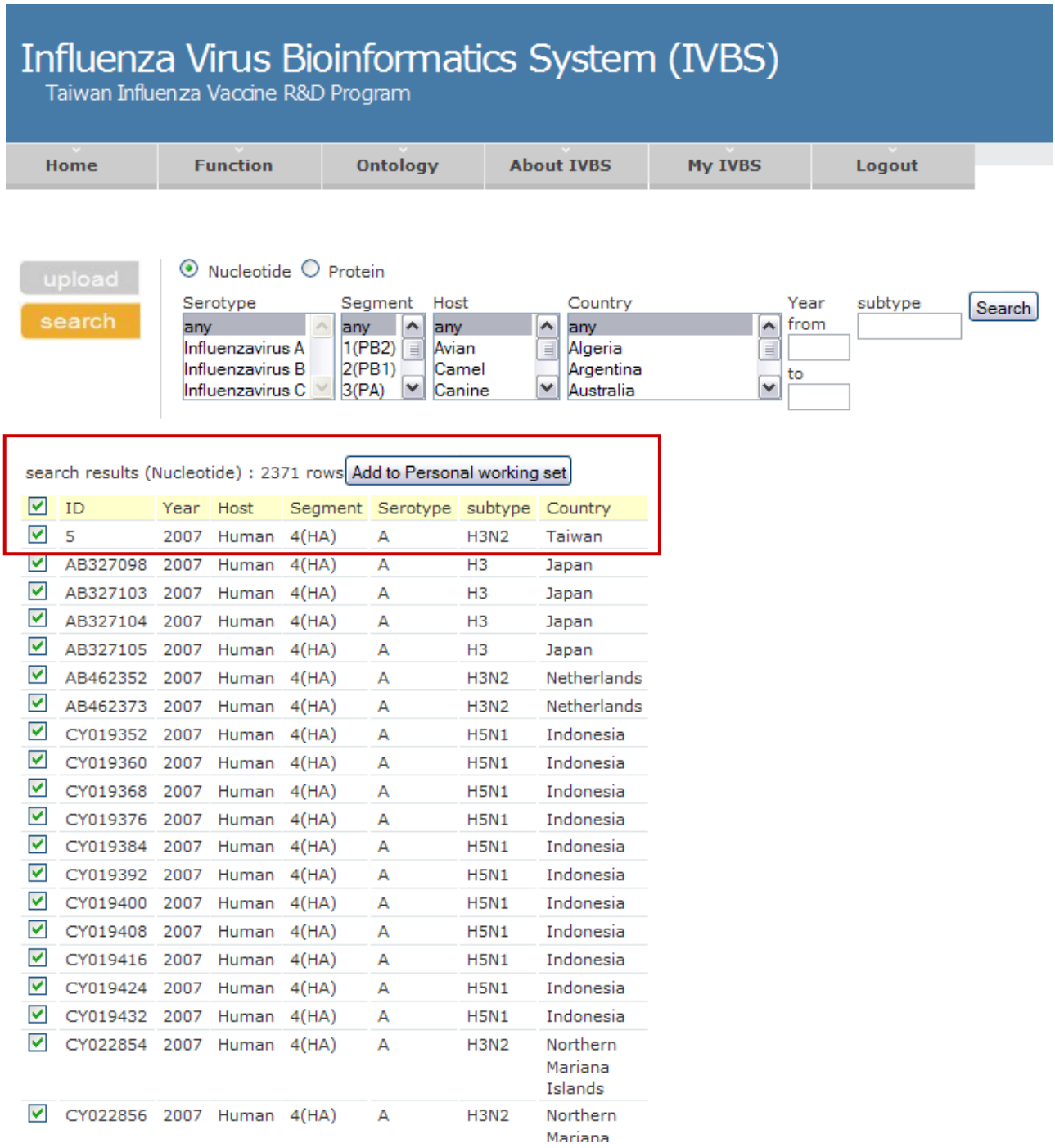


在 My IVBS 頁面中的另一部份，可以讓使用者**同時**查詢個人資料庫內及公開網域中的流感病毒序列資料，利用後續多序列排比及建立親緣關係樹之分析，可以比較公開流感病毒的序列及自己實驗室分離出來的還沒有 release 的序列。



舉例說明: 依據特定的條件年份、亞型、等特定條件搜尋，即可同時查找個人資料庫內序列及公開網域中的流感病毒序列資料。





搜尋結果為 2,371 筆序列資料，ID 5 為使用者個人資料庫內之序列資料，後續可將有興趣之序列透過加入個人工作區，進行多序列排比及建立親緣關係樹之分析比較，新開發之此功能可進一步將個人資料與公開網域中的流感病毒序列資料一起做分析比較。

6. 分析 IVBS 現有 influenza B-cell epitope 的胺基酸組成

→已完成 IVBS 現有 influenza B-cell epitope 的胺基酸組成分析

說明: 我們先前已將 IEDB 中的 Influenza Virus 的 epitope 資訊整合加入 IVBS 當中，為了瞭解組成 B-cell epitope 胜肽片段(Peptide)的胺基酸具有何種的理化傾向(Propensity)，以利後續預測工具之研發，我們首先針對 IEDB 中的 epitope 資訊進行重複序列篩選、過濾的工作，同時將各種不同的 epitope 種類分門別類，最後計算各個抗原蛋白中 B-cell epitope 的胺基酸出現頻率，以了解其胺基酸組成情況。

如表列示 IVBS 所整合的 Influenza epitope 資料筆數的概況。原始的 IEDB 中共有 7159 筆各種的 epitope 資訊，而標示為 linear 型式的共有 6942 筆。在去除重複之相同序列之後，B-cell epitope 有 118 筆資料，T-cell epitope 有 2289 筆，而 MHC 有 1819 筆，MHC Ligand 有 22 筆。

表、IVBS 中 IEDB 資料庫各型態 epitope 的資料筆數

Total	Linear epitope	Non-redundant linear epitope			
		B-cell epitope	T-cell epitope	MHC	MHC Ligand
7159	6942	118	2289	1819	22

B-cell epitope 是與抗體分子作用的抗原蛋白區段，因此抗原分子通常必須暴露於蛋白質分子表面才有機會與抗體產生交互作用。Influenza Virus 中共有 HA、NA 及 Matrix Protein (M1 & M2)等蛋白能

暴露於病毒表面。因為 IEDB 中並沒有 NA 的 B-cell epitope 資訊，故 IEDB 中各個病毒蛋白的 B-cell epitope 個數整理如下表。

表、IVBS 中 IEDB 資料庫 B-cell epitope 的資料筆數		
HA	**M1**	**M2**
100	6	12

在統計出序列獨一不重複的 B-cell epitope 個數後，為瞭解這些 epitope 的組成，我們分別計算 HA、M1 與 M2 這三種蛋白質的胺基酸組成，其結果如下表所示。

表、IVBS 中 IEDB 資料庫 B-cell epitope 胺基酸組成分析結果			
Protein Type Amino Acid	**HA**	**M1**	**M2**
A (Alanine)	81 (4.99 %)	13 (6.81%)	0 (0.00 %)
C (Cysteine)	52 (3.20 %)	4 (2.09 %)	14 (7.22 %)
D (Aspartate)	101 (6.22 %)	6 (3.14 %)	13 (6.70 %)
E (Glutamate)	70 (4.31 %)	8 (4.19 %)	28 (14.43 %)
F (Phenylalanin)	68 (4.19 %)	9 (4.71 %)	0 (0.00 %)
G (Glycine)	134 (8.26 %)	17 (8.90 %)	9 (4.64 %)
H (Histidine)	36 (2.22 %)	5 (2.62 %)	0 (0.00 %)
I (Isoleucine)	76 (4.68 %)	10 (5.24 %)	10 (5.15 %)
K (Lysine)	90 (5.55 %)	9 (4.71 %)	0 (0.00 %)
L (Leucine)	108 (6.65 %)	21 (10.99 %)	15 (7.73 %)
M (Methionine)	8 (0.49 %)	3 (1.57 %)	3 (1.55 %)
N (Asparagine)	125 (7.70 %)	10 (5.24 %)	17 (8.76 %)
P (Proline)	105 (6.47 %)	7 (3.66 %)	11 (5.67 %)
Q (Glutamine)	68 (4.19 %)	10 (5.24 %)	0 0.00 (%)
R (Arginine)	59 (3.64 %)	18 (9.42 %)	17 (8.76 %)
S (Serine)	131 (8.07 %)	9 (4.71 %)	21 (10.82 %)
T (Threonine)	118 (7.27 %)	15 (7.85 %)	18 (9.28 %)
V (Valine)	87 (5.36 %)	13 (6.81 %)	10 (5.15 %)
W (Tryptophan)	27 (1.66 %)	0 (0.00 %)	8 (4.12 %)
Y (Tyrosine)	79 (4.87 %)	4 (2.09 %)	0 (0.00 %)

7. 研究調查目前預測 B-cell epitope 的方法

→已完成現有預測 B-cell epitope 方法之調查

說明: B-cell epitope 為抗原(antigen)蛋白上被抗體(antibody)辨識及鍵結的胜肽片段。通常此種片段需位於蛋白質的表面方具有較大與抗體分子接觸的機會。反之，包埋於蛋白質內部之中的胜肽片段，其與抗體分子產生交互作用及接觸的機會則較低。然而位於蛋白質表面的胺基酸多屬較為親水性的(Hydrophilic)，基於這樣的現象，我們可藉由計算抗原分子中各個胺基酸親水傾向(Hydrophilicity) 的大小，估算其位於蛋白質表面的可能性，作為預測其是否可能構成 epitope 的依據。此外根據文獻調查，我們亦發現及了解 B-cell epitope 多出現在蛋白質構型(Conformation)較具彈性(Flexible)的 β-turn 二級結構(Secondary Structure)中，藉由分析胜肽片段中胺基酸構成 β-turn 結構的可能性，亦可作為預測 B-cell epitope 的方法。於是最後我們採用推測抗原蛋白上的胺基酸出現在蛋白質表面與抗體接觸的傾向(Hydrophilicity 及 Accessibility)、形成 β-turn 的機率及構成 epitope 的胺基酸的組成比率(Antigenicity)等理化性質作為開發預測 B-cell epitope 工具的準則。

8. 根據調查的 B-cell epitope 預測方法撰寫預測程式之一

→已完成依據抗原決定位之胺基酸特性開發預測 B-cell epitope 之程式

說明: 我們依據文獻調查的結果，採用 Beta-Turn (Chou & Fasman,

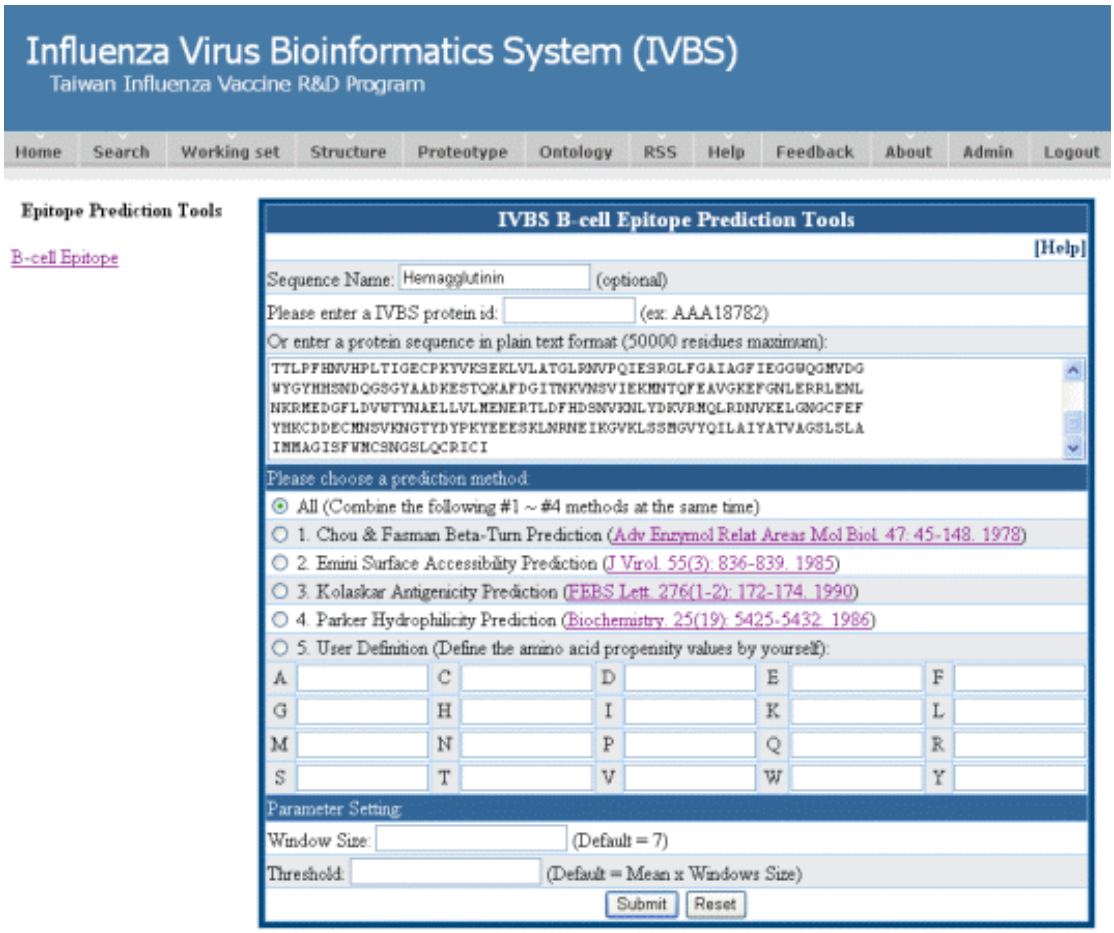
1978)、Surface Accessibility (Emini, 1985)、Antigenicity (Kolaskar, 1990)及 Hydrophilicity (Parker, 1986)等四種與組成 B-cell epitope 較具關聯性的胺基酸特性，撰寫 B-cell epitope 的預測程式。首先由文獻中取得衡量每一種胺基酸性質(amino acid propensity)的實驗或估算數值，並經正規化(normalization)處理，以使不同性質之數值能在相同的尺度上(scale)進行比較。然後計算蛋白質序列上某一固定長度內(預設值是 7 個胺基酸)該性質的平均數值，並將此平均值賦予最中間的胺基酸，以考慮周遭相鄰胺基酸對構成該項性質的影響，如此從頭至尾每次位移一個胺基酸重複地計算每一段序列的數值，數值愈大者，表示具有該項性質的潛勢(potential)愈大，同時這些性質又是與構成 B-cell epitope 有關的特性，因此藉由閾值(threshold)的篩選，數值愈大的胺基酸即可視為可能構成 B-cell epitope 的元素。此外我們亦提供使用者自定胺基酸性質數值的功能，讓使用者可彈性利用其他的性質進行 B-cell epitope 的預測與分析。整體結果呈現請見圖一~圖三所示。

9. 根據調查的 B-cell epitope 預測方法撰寫預測程式之二

→已完成以 SVM 預測 B-cell epitope 之程式

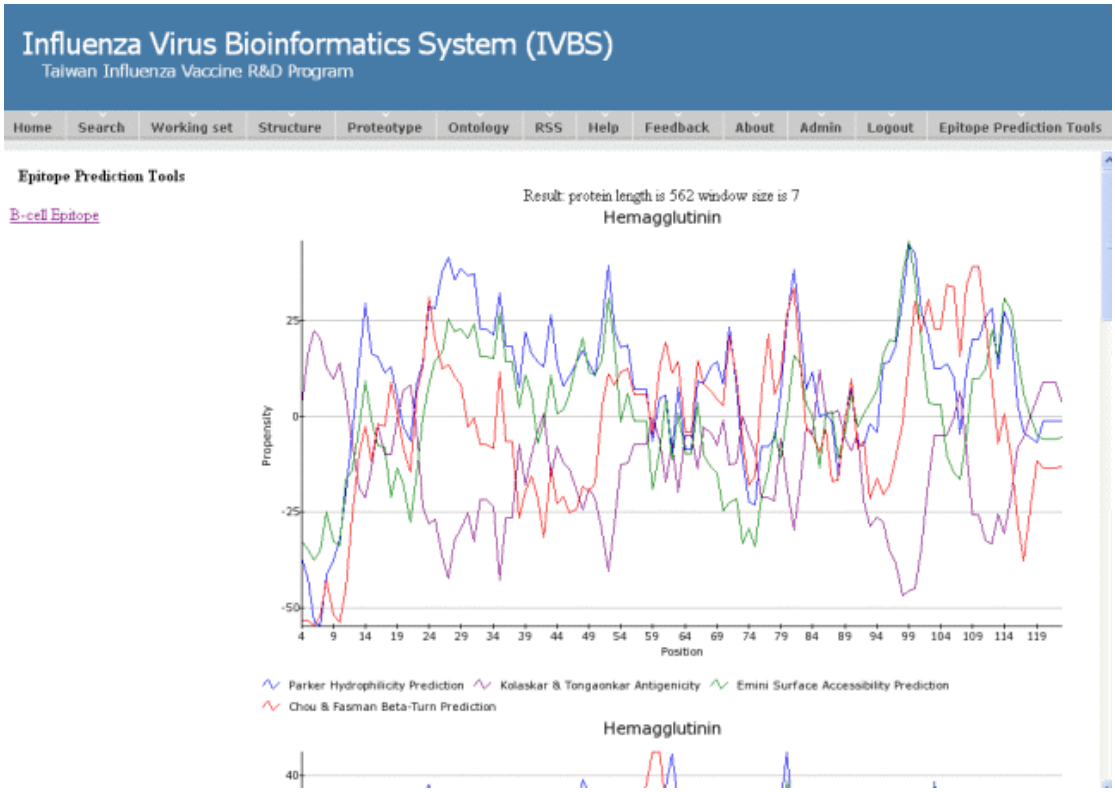
說明： 此部份是利用監督式學習概念(supervised learning)而將不同機器學習方式(machine learning)帶入以建立更有效力之 B-cell epitope 預測程式。此處我們採用長度為 20 之 872 筆 linear B-cell epitope 片段及

872 筆非 linear B-cell epitope 片段(Chen, et al., 2007)當作建立預測模式所需之正向與負向資料，而機器學習方式採用目前常用作法 Support Vector Machine (SVM)。而為了將各 872 筆正向與負向序列資料轉換成數值資訊以餵入 SVM 裡，我們採用五種編碼模式以供使用者選擇：以二位元編碼方式(Binary Coding)或利用上一段所提到的四種胺基酸理化性質以轉換胺基酸資訊。整體結果呈現請見圖四~圖六所示。

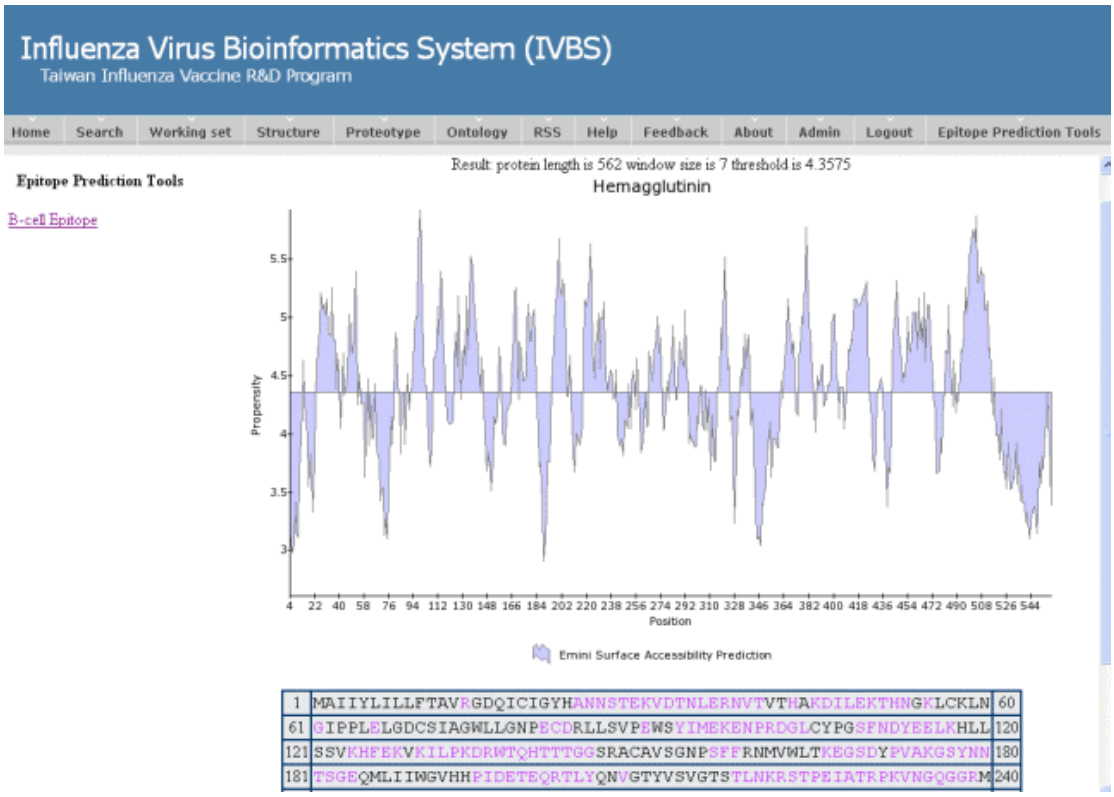


圖一：B-cell epitope 預測程式操作介面。使用者可輸入欲分析的蛋白質的 IVBS 編號或序列，選擇欲採用的胺基酸性質及設定序列掃描長

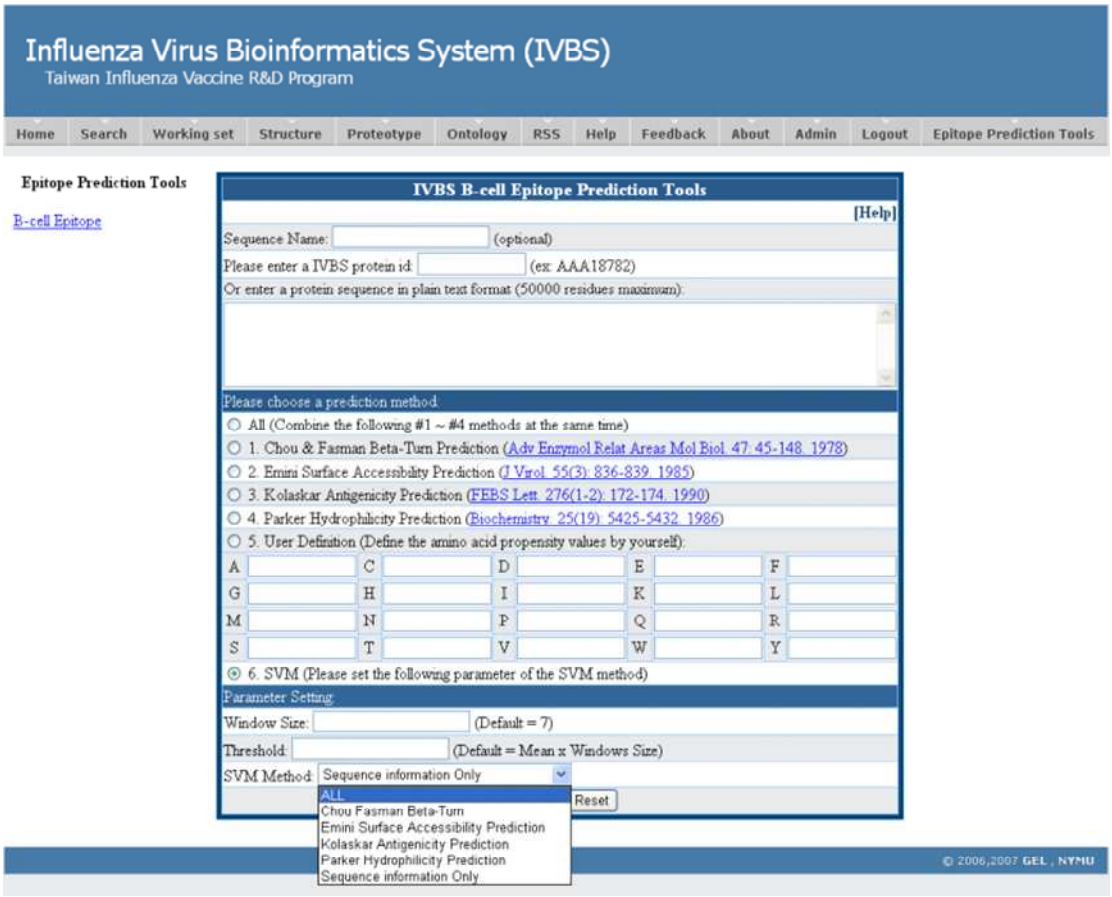
度(window size)與篩選閾值(threshold)的大小。



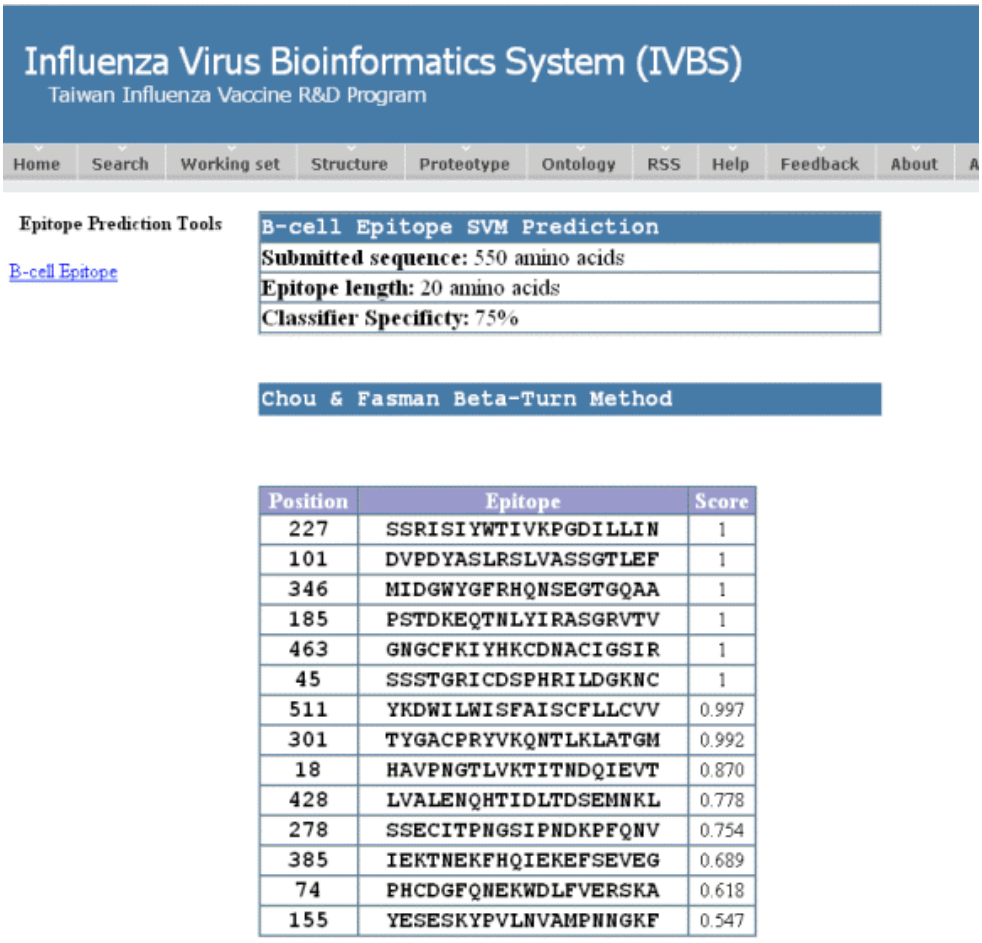
圖二：A/Japan/305/1957 H2N2 流感病毒 Hemagglutinin 蛋白質序列(SP: P03451) 綜合 Beta-Turn(紅色折線)、Accessibility(綠色折線)、Antigenicity(紫色折線)及 Hydrophilicity(藍色折線)四種胺基酸性質的 B-cell epitope 預測結果。X 軸為胺基酸位置；Y 軸胺基酸性質大小，預設序列掃描長度為 7，篩選閾值為 0。



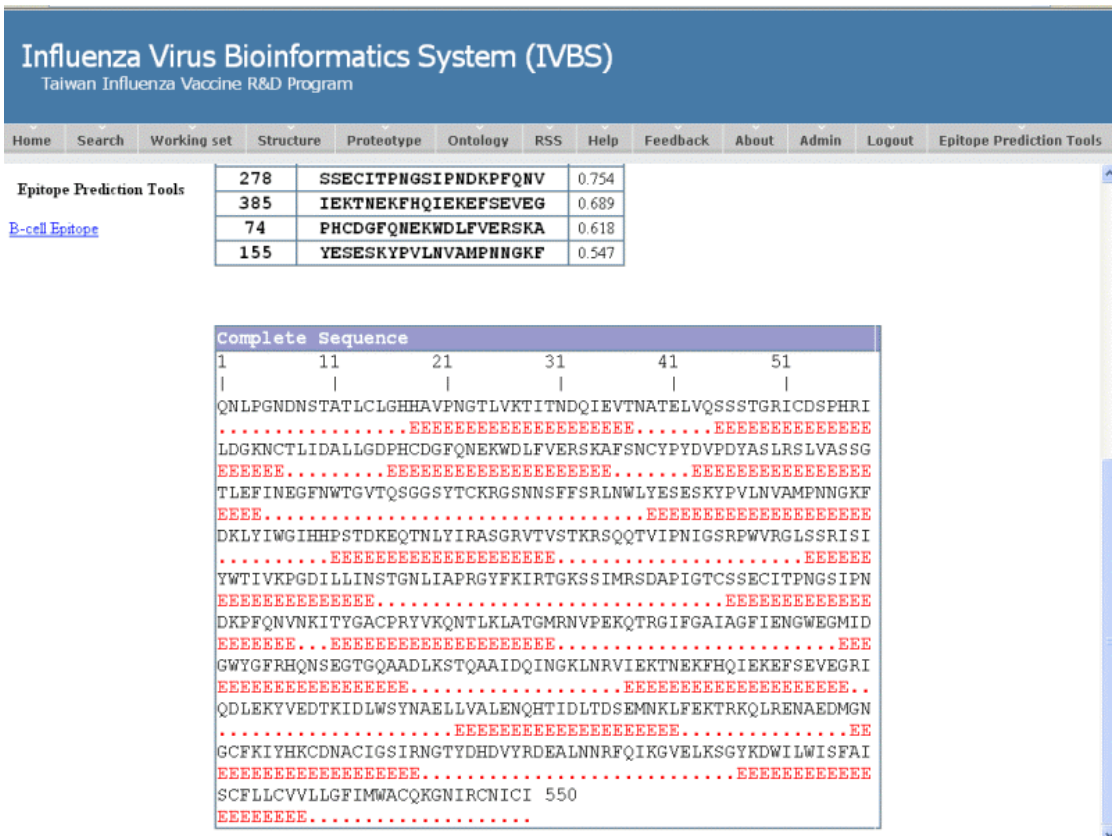
圖三：A/Japan/305/1957 H2N2 流感病毒 Hemagglutinin 蛋白質序列(SP: P03451)中胺基酸 Emini Surface Accessibility 的性質分佈圖。序列掃描長度為 7，篩選閾值為平均值乘以掃描長度。預測結果的圖譜繪於上方，折線圖分為大於閾值及小於閾值上、下二部份。下方為蛋白質的序列，大於閾值的胺基酸以桃紅色表示，即為預測可能的 B-cell epitope 胺基酸組成元素。



圖四：利用機器學習(Machine Learning)之支持向量機(Supprot Vector Machine, SVM)方法開發之 B-cell epitope 預測工具操作介面。使用者在參數設定(Parameter Setting)中，可選定是單獨針對蛋白質序列進行二位元編碼(Binary Coding)或將序列合併胺基酸理化性質進行編碼來執行 SVM 方法的 B-cell epitope 預測工作。



圖五：SVM 預測結果中會顯示原始的蛋白質序列長度、預測 B-cell epitope 的長度和預測分類器的專一性(Specificity)等。此外，預測的結果會根據 SVM 分類器預測為 B-cell epitope 的機率進行排序，逐一顯示所預測的 B-cell epitope 片段。



圖六：我們亦將每一段 SVM 所預測的 B-cell epitope 片段(圖中標紅色 E 字者)依發生位置標示於原始蛋白質序列中，方便使用者觀察整個預測的概況。

8. 分析 IVBS 現有 CTL epitope 的胺基酸組成

→已完成分析 IVBS 現有 CTL epitope 的胺基酸組成

說明： 我們首先將 IEDB 中的 epitope 資訊進行篩選及過慮重複序列的工作，同時按 epitope 的種類分門別類，最後將各個抗原蛋白中 CTL epitope 的胺基酸組成進行分析，其結果如下各表所示。

表、IVBS 中 IEDB 資料庫 CTL epitope 的資料筆數

	PB2	PB1	PA	HA	NP	NA	M1	M2	NS1	NS2	Nonstructural protein
T-cell	153	153	143	492	260	217	119	30	46	23	16

epitope											
MHC	**26**	**39**	**29**	**342**	**133**	**66**	**131**	**11**	**--**	**--**	**31**
MHC Ligand	**--**	**--**	**--**	**6**	**7**	**--**	**1**	**--**	**--**	**--**	**--**

表、IVBS 中 IEDB 資料庫 CTL epitope 胺基酸組成分析結果(一)

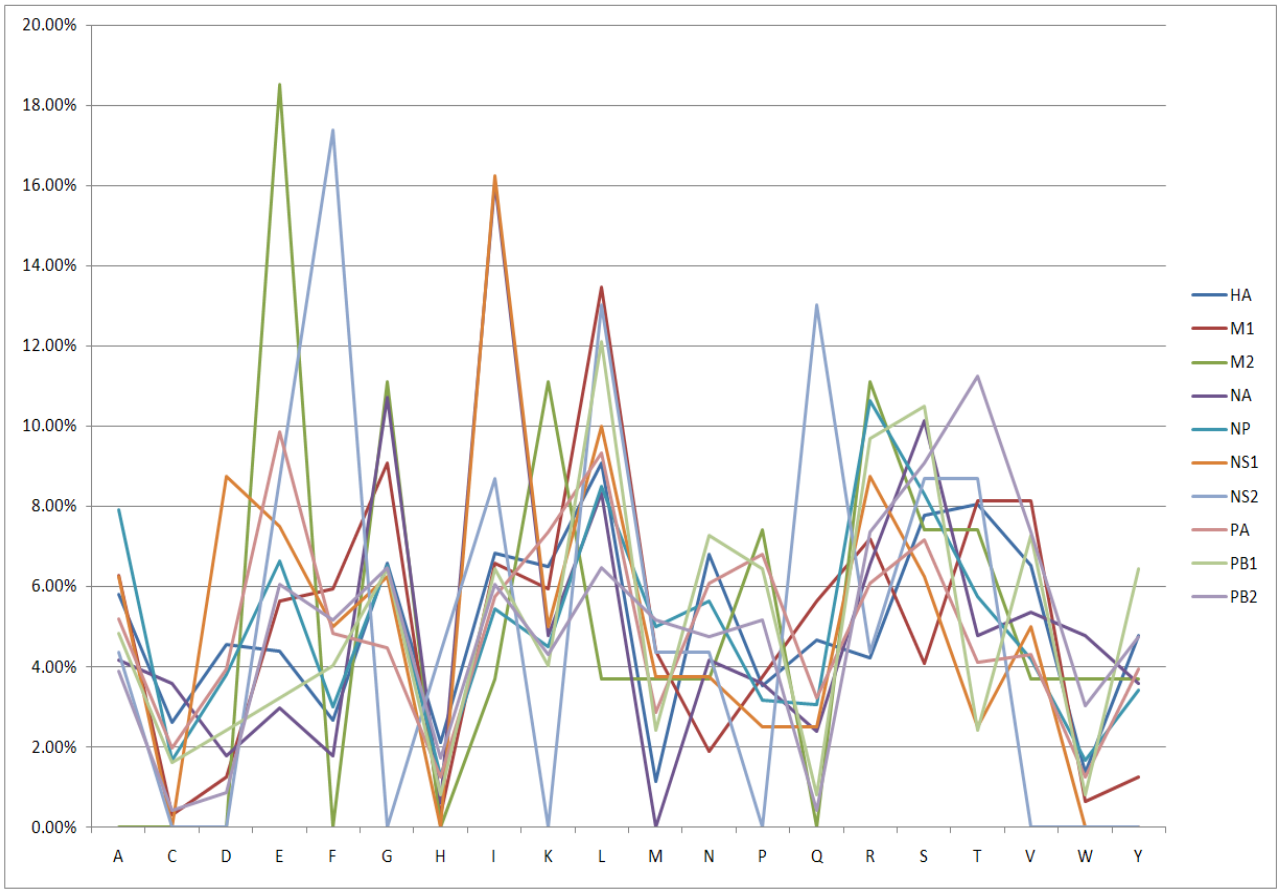
	HA	M1	M2	NA	NP
A (Alanine)	173 (5.80%)	20 (6.27%)	0	7 (4.17%)	125 (7.92%)
C (Cysteine)	78 (2.61%)	1 (0.31%)	0	6 (3.57%)	26 (1.65%)
D (Aspartate)	136 (4.56%)	4 (1.25%)	0	3 (1.79%)	60 (3.8%)
E (Glutamate)	131 (4.39%)	18 (5.64%)	5 (18.52%)	5 (2.98%)	105 (6.65%)
F (Phenylalanin)	80 (2.68%)	19 (5.96%)	0	3 (1.79%)	47 (2.98%)
G (Glycine)	196 (6.57%)	29 (9.09%)	3 (11.11%)	18 (10.71%)	103 (6.52%)
H (Histidine)	63 (2.11%)	1 (0.31%)	0	1 (0.6%)	20 (1.27%)
I (Isoleucine)	204 (6.84%)	21 (6.58%)	1 (3.7%)	27 (16.07%)	86 (5.45%)
K (Lysine)	194 (6.50%)	19 (5.96%)	3 (11.11%)	8 (4.76%)	71 (4.5%)
L (Leucine)	271 (9.08%)	43 (13.48%)	1 (3.7%)	14 (8.33%)	134 (8.49%)
M (Methionine)	34 (1.14%)	14 (4.39%)	1 (3.7%)	0	79 (5%)
N (Asparagine)	203 (6.8%)	6 (1.88%)	1 (3.7%)	7 (4.17%)	89 (5.64%)
P (Proline)	105 (3.52%)	12 (3.76%)	2 (7.41%)	6 (3.57%)	50 (3.17%)
Q (Glutamine)	139 (4.66%)	18 (5.64%)	0	4 (2.38%)	48 (3.04%)
R (Arginine)	126 (4.22%)	23 (7.21%)	3 (11.11%)	11 (6.55%)	168 (10.64%)
S (Serine)	232 (7.77%)	13 (4.08%)	2 (7.41%)	17 (10.12%)	131 (8.3%)
T (Threonine)	240 (8.04%)	26 (8.15%)	2 (7.41%)	8 (4.76%)	91 (5.76%)
V (Valine)	195 (6.53%)	26 (8.15%)	1 (3.7%)	9 (5.36%)	66 (4.18%)
W (Tryptophan)	41 (1.37%)	2 (0.63%)	1 (3.7%)	8 (4.76%)	26 (1.65%)
Y (Tyrosine)	143 (4.79%)	4 (1.25%)	1 (3.7%)	6 (3.57%)	54 (3.42%)

表、IVBS 中 IEDB 資料庫 CTL epitope 胺基酸組成分析結果(二)

	NS1	NS2	PA	PB1	PB2

A (Alanine)	5 (6.25%)	1 (4.35%)	29 (5.21%)	6 (4.84%)	9 (3.9%)
C (Cysteine)	0	0	11 (1.97%)	2 (1.61%)	1 (0.43%)
D (Aspartate)	7 (8.75%)	0	22 (3.95%)	3 (2.42%)	2 (0.87%)
E (Glutamate)	6 (7.5%)	2 (8.7%)	55 (9.87%)	4 (3.23%)	14 (6.06%)
F (Phenylalanin)	4 (5%)	4 (17.39%)	27 (4.85%)	5 (4.03%)	12 (5.19%)
G (Glycine)	5 (6.25%)	0	25 (4.49%)	8 (6.45%)	15 (6.49%)
H (Histidine)	0	1 (4.35%)	7 (1.26%)	1 (0.81%)	4 (1.73%)
I (Isoleucine)	13 (16.25%)	2 (8.7%)	32 (5.75%)	8 (6.45%)	14 (6.06%)
K (Lysine)	4 (5%)	0	41 (7.36%)	5 (4.03%)	10 (4.33%)
L (Leucine)	8 (10%)	3 (13.04%)	52 (9.34%)	15 (12.1%)	15 (6.49%)
M (Methionine)	3 (3.75%)	1 (4.35%)	16 (2.87%)	3 (2.42%)	12 (5.19%)
N (Asparagine)	3 (3.75%)	1 (4.35%)	34 (6.1%)	9 (7.26%)	11 (4.76%)
P (Proline)	2 (2.5%)	0	38 (6.82%)	8 (6.45%)	12 (5.19%)
Q (Glutamine)	2 (2.5%)	3 (13.04%)	18 (3.23%)	1 (0.81%)	1 (0.43%)
R (Arginine)	7 (8.75%)	1 (4.35%)	34 (6.1%)	12 (9.68%)	17 (7.36%)
S (Serine)	5 (6.25%)	2 (8.7%)	40 (7.18%)	13 (10.48%)	21 (9.09%)
T (Threonine)	2 (2.5%)	2 (8.7%)	23 (4.13%)	3 (2.42%)	26 (11.26%)
V (Valine)	4 (5%)	0	24 (4.31%)	9 (7.26%)	17 (7.36%)
W (Tryptophan)	0	0	7 (1.26%)	1 (0.81%)	7 (3.03%)
Y (Tyrosine)	0	0	22 (3.95%)	8 (6.45%)	11 (4.76%)

利用上表抗原蛋白中 CTL epitope 的胺基酸組成進行分析畫成樹狀圖。



9. 研究調查目前預測 CTL epitope 的方法

→已完成研究調查目前預測 CTL epitope 的方法

說明: 抗原被抗体辨識或結合的地方叫抗原決定位。T-細胞所辨認的是一連串胺基酸組成的抗原決定位，當細胞受到感染的時候，處理過的抗原會跟主要組織相容抗原(major histocompatibility complex, MHC)上的 classI/II 結合成複合體並呈現在細胞膜表面，而被 T-細胞上的 T-細胞受器(Cytotoxic T cell receptor, TCR)辨識。一個抗原通常有好幾個抗原決定位，構造越複雜，分子量越大，它的抗原決定位愈多。

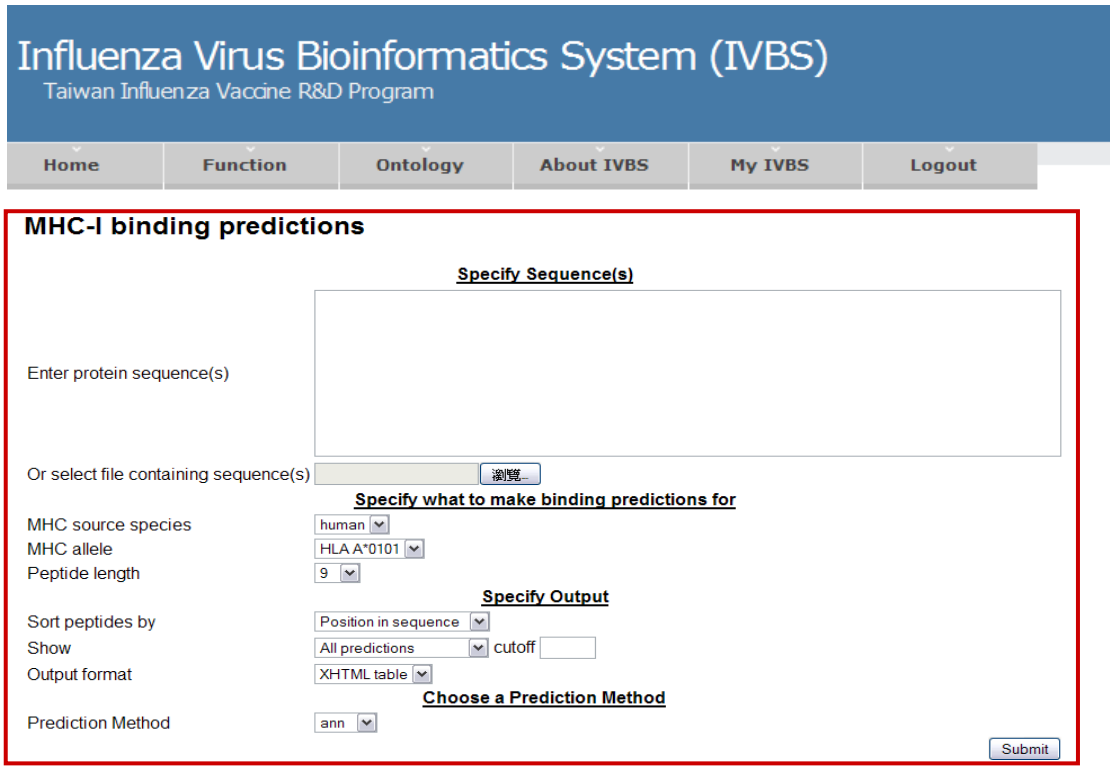
表、目前預測 MHC class I 抗原決定位的工具	
Methods	Reference
ANN(Artificial neural network)	(2003)Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. May;12(5):1007-17.
ARB(Average relative binding)	(2005)Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. Immunogenetics. Jun;57(5):304-14. Epub 2005 May 3.
(SMM)Stabilized matrix method	(2005)Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. BMC Bioinformatics. May 31;6:132.
BIMAS	(1994)Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains, *J Immunol* **152** , pp. 163–175.
SYFPEITHI	(1999)SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics. Nov;50(3-4):213-9. (2007) SYFPEITHI: database for searching and T-cell epitope prediction. Methods Mol Biol. 409:75-93.
RANKPEP	(2002) Prediction of MHC class I binding peptides using profile motifs. Hum Immunol. Sep;63(9):701-9.
PROPRED-I	(2003)PROPRED-I: prediction of promiscuous MHC class I binding sites, *Bioinformatics* 19, pp. 1009–1014.

表、目前預測 MHC class II 抗原決定位的工具	
Methods	Reference
ARB(Average relative binding)	(2005)Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. Immunogenetics. 2005 Jun;57(5):304-14. Epub 2005 May 3.
SMM-align	(2007)Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics. Jul 4;8:238.
Sturniolo	(1999)Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotechnology. Jun;17(6):555-61.
RANKPEP	(2004)Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles, *Immunogenetics* 56, pp. 405–419.
PROPRED	(2001)PROPRED: prediction of HLA-DR binding sites, *Bioinformatics* 17, pp. 1236–1237.

10. 根據調查的 CTL epitope 預測方法撰寫預測程式

→已完成 B-cell epitope prediction 使用者頁面之初步設計

說明: 除了完成調查預測方法之外，我們亦完成預測工具使用者介面雛型之設計(如下圖)，預測程式亦在開發測試中



本　　文

(4) 討論

本計畫配合流行病學專家的需求，協助探討流感病毒的分子演化機制。流感病毒中的 HA 與 NA 蛋白質之序列資料，是抗體主要的辨識區，因此結合收集歷年在北半球的序列數據，將有助於測試預測的方法。換句話說，世界衛生組織過去的經驗，可做為我們發展預測方法的正向控制實驗。而網際網路上公開的資訊，例如全基因體的序列，亦有助於區辨造成基因變異的假說。不過真正重要的是整合疾病防治管制局現有的、連續幾年的本土流感病毒資料，因此在第一年度我們已完成了台灣疾管局流感病毒序列資料與 NCBI 之 Influenza Virus Resource (IRV)及 LANL 之整合資訊庫。使用者以特定的流感病毒相關資訊當作篩選條件，並能提供查詢資料下載、以及後續分析，多序列排比及親緣分析 (phylogentic analysis)。

在第二年度裡，我們完成增加知識管理的部分，利用 ontology 整合，加入了流感病毒的生物學特性: 抗原、抗藥性、HI titer、病毒 recombination and reassortment 等分析於 IVBS 中。

由於流感病毒基因的高突變性，每年二月和九月 WHO 分別針對南北半球的冬季建議疫苗病毒株，利用雪貂血清進行血球凝集抑制試驗(HI titer)為免疫反應之評估標準，利用這些抗血清可以分析流行病

毒的抗原性特性，對於疫苗株選擇上有重要指標，因此蒐集了 WHO 及文獻上實驗的 HI titer 數據。

抗原被抗體辨識或結合的地方為抗原決定位(epitope)，抗原上的決定位通常含 6~8 個胺基酸，一個抗原通常有好幾個抗原決定位，研究流感病毒抗原變異的規律，將有助於對流感疫情的預測及疫苗病毒株的選用，如能找到不同流感病毒株具有相似的上百個抗原決定基，將更助未來進一步研發出單一疫苗。流感病毒抗原決定位資訊的蒐集主要以公共網域上 IEDB (http://www.immuneepitope.org/) 資料庫所提供的資訊為主。我們先把自 IEDB 中所得到的資訊轉換格式，然後匯入 IVBS 資料庫中，之後再與 IVBS 中的病毒資訊進行交互參考(Cross-reference)，以列出每株病毒中所有蛋白質的抗原決定位資訊。

IVBS 為一整合型資料庫整合了許多不同的資料來源，從 genetic property 方面，流感病毒序列上可以做多序列排比及親緣關係樹的分析，找到 reassortment and recombination linkage。從 antigenic property 方面，針對流感病毒生物學特性，經由實驗得知單點突變會造成病毒產生抗藥性，HA1 domain 產生 genome 突變會造成 antigenic sited 改變，影響免疫反應，搜集 epitope 抗原決定位的資料及 HI 力價、血清反應，則有助於疫苗選株。

此 IVBS 整合型資料庫除了提供泛用的分析工具，還包括蛋白表

現型(proteotype)上流感病毒基因體的資訊，在第三年度裡，我們增加所建立之分析流感病毒序列的自動化流程於 IVBS 中。完成了流感病毒全基因體之基因型(genotype)預測及分佈功能，可比較不同病毒株中不同基因型及蛋白表現型之差異，推斷可能發生過的基因置換。

同時也新增加了個人流感病毒基因序列資料庫，讓使用者可以擁有個人的資料庫進行資料輸入、儲存與後續分析的功能。

而流感病毒生物資訊系統(IVBS)目前整合 IEDB(Immune Epitope Database) 流感病毒的 epitope 資料，此資料庫是由美國 LIAI (La Jolla Institute for Allergy and Immunology)研究所動員十數人長時間閱讀文獻蒐集、整理而來，然其資料量仍遠少於 IVBS 蛋白質的數量。針對這樣的情況，我們將研究 B-cell 與 CTL epitope 胜肽(peptide)片段的胺基酸的物理、化學性質，根據這些特性建立規則，設計分析 B-cell 及 CTL epitope 發生位置的預測程式，提供研究人員初步的參考資訊，協助加速研究的進展。

IEDB 中 epitope 資料量的統計表，有 190 筆 B cell epitope 資料，412 筆 T cell epitope 資料。

190 B cell epitopes	75 linear epitopes
	115 conformational epitopes
412 T cell epitopes	175 CD4
	148 CD8
	89 undefined

		T cell		
Protein	Antibody	CD4	CD8	Total
HA	150	113	35	298
NP	3	44	49	96
PA	0	1	11	12
NA	24	7	8	39
M1	4	9	15	28
PB2	0	0	9	9
M2	9	0	3	12
PB1	0	0	10	10
NS1	0	1	7	8
NS2	0	0	1	1

IEDB 中針對 IEDB 所做的 epitope 資料量統計表，發現 B cell epitope 決定位較多出現在 HA 和 NA 蛋白上，因為 HA 和 NA 為流感病毒兩個表面主要的蛋白，B-cell epitope 抗原為抗體(antibody)辨識及鍵結的胜肽片段，需位於蛋白質的表面方才具有較大與抗體分子接觸的機會，所以位於病毒表面的蛋白有較大的機會成為 B cell 決定位。基於這樣的現象，我們可藉由計算抗原分子中各個胺基酸親水傾向(Hydrophilicity) 的大小，估算其位於蛋白質表面的可能性，作為預測其是否可能構成 epitope 的依據，再加入文獻的查閱，最後我們採用推測抗原蛋白上的胺基酸出現在蛋白質表面與抗體接觸的傾向(Hydrophilicity 及 Accessibility)、形成 β-turn 的機率及構成 epitope 的胺基酸的組成比率(Antigenicity)等理化性質作為開發預測 B-cell epitope 工具的準則。同時也發現 HA 蛋白中發現較多 CD4 T cell

epitope 抗原決定位，NP 蛋白中發現較多 CD8 T cell epitope 抗原決定位。

本　　文

(5) 結論與建議

在第一年度裡

1. 第一季: 開始安裝各種軟硬體需求，進行流感資訊之調查
2. 第二季: 建立流感資訊網，提供資訊代理人流感資訊收集
3. 第三季: 整合各種流感資訊，評估各種流感病毒序列與親緣分析軟體的優缺點評估，改進分析效率。
4. 第四季: 流感資訊網完成架設提供 RSS 服務

在第二年度裡

5. 第一季: 開放 Influenza Virus Bioinformatics System (IVBS) 予疾管局相關人員使用，並提供教育訓練課程 (2007.03)
6. 第二季:增加知識管理的部份，利用 ontology 整合、分類流感病毒相關研究成果 (2007.06)
7. 第三季:加入感染者流行病學資料及下列生物學特性: 抗原、 抗藥性、HI titer、病毒 recombination and reassortment 等分析於 IVBS (2007.09)
8. 第四季:建立分析流感病毒序列的自動化流程 (2007.12)

在第三年度裡

9. 增加所建立之分析流感病毒序列的自動化流程於 Influenza Virus Bioinformatics System (IVBS)
    i. 增加以 sequence logo 及階層方式同時呈報大規模流感病毒蛋白質序列排列中氨基酸之相似程度(2008.03)
    ii. 增加流感病毒全基因體之基因型(genotyping)之預測與分佈功能(2008.03)
    iii. 增加使用者個人流感病毒基因序列資料之輸入儲存與後續分析功能(2008.06)
10. 建立與增加新的生物資訊學技術預測流感病毒的演化趨勢，協助選擇適合用來做流感疫苗的流感病毒株
    i. 開發流感病毒蛋白質 B 細胞抗原決定位(B-cell epitope)胜肽片段預測程式及系統 (2008.06)
    ii. 開發流感病毒蛋白質 CTL (cytotoxic T lymphocyte)細胞抗原決定位胜肽片段預測程式及系統 (2008.09)
    iii. 開發完整流感病毒(whole virus) Hemagglutination activity 之 seropositivity 生物資訊預測程式及系統 (2008.12)

流感資訊核心設施所開發之流感資訊網，分析流程方面，已完美呈現多序列排比及親緣關係分析。資訊代理人服務方面，利用 RSS 提供計畫下人

員流感病毒資料及電子郵件寄送每日更新之流感相關文獻。印發使用者手冊並提供教育訓練課程予疾管局相關人員使用。知識管理的部分，利用 ontology 蒐集且整合公共網域中生物資料庫的流感病毒註解資訊及相關研究，並加入流感病毒之生物特性包括：抗原、 抗藥性、HI titer，將有助於預測流感病毒的演化趨勢及有助於發展流感疫苗。流行病學資料也已動態連結作業方式同步更新 TPMGD 內的資料整合於 IVBS 系統中。第三年度，我們增加了分析流感病毒序列的自動流程於 IVBS 中。序列排列中氨基酸之相似程度，除了使用 weblogo 呈現方式之外，也增加利用連結方式到中研院資訊所施純傑博士實驗室，利用他們所開發之工具 sequence logo 以階層方式同時呈報大規模流感病毒蛋白質序列排列中氨基酸之相似程度。流感病毒全基因體之基因型，從 Flugenome database 搜尋流感病毒基因型的資料，並整合 IVBS 序列資料做結合放入 IVBS 中，增加流感病毒全基因體之基因型的預測及分佈功能。個人資料庫，使用者可將個人流感病毒基因序列資料做輸入、儲存以及後續與公共網域中的流感病毒序列一起分析之功能。我們也增加新的生物資訊學技術預測流感病毒的演化趨勢，協助選擇適合用來做流感疫苗的流感病毒株。B-cell epitope prediction，開發了流感病毒蛋白質 B 細胞抗原決定位胜肽片段預測程式及系統，可以預測一段流感病毒序列為抗原決定位的可能性多大。

本　　文

(6) 計畫重要研究成果及期中報告審查委員意見之答覆

計畫重要研究成果

本計畫的總目標在於

1. 根據流感疫苗研究發展計畫的其他計畫所提出的需求，協助做各種生物資訊分析 – 因此在第一年度我們已開始提供流感病毒生物資訊分析之服務，協助其他研發團隊之人員實際解決問題。
2. 自動收集網際網路上所有公開的流感病毒資訊，並與臺灣特有的流感病毒資訊整合 – 因此在第一年度我們已完成了台灣疾管局流感病毒序列資料與 NCBI 之 Influenza Virus Resource (IRV)及 LANL 之整合資訊庫。
3. 利用 RSS 技術，自動提供疾病管制局、流感疫苗研究發展計畫的其他團隊即時資訊 - 因此在第一年度我們已可提供六個流感病毒網站即時資訊。這六個網站分別為 NCBI、WHO、GoogleNews、PandemicFlu、ProMEDmail 及 CIDRAP，其中 CIDRAP 進一步分成流感資訊、禽流感資訊以及流感疫苗資訊。
4. 開放 Influenza Virus Bioinformatics System(IVBS) 上線使用予疾管局相關人員使用，並提供教育訓練課程- 因此在第二年度我們已於三月

二十八日星期三至台灣疾病管制局昆陽分局與林森總局(透過視訊會議)進行IVBS教育訓練課程。並做帳號管理介面管理使用者使用IVBS系統情形。

5. 增加知識管理的部分，利用 ontology 整合、分類流感病毒相關研究成果- 因此在第二年度我們已完成以 UniProtKB/ Swiss-Prot 蛋白質資料庫為核心，與其他生物資料庫進行資料交互參考，整合了包括 InterPro、PRINTS、Pfam、ProtoNet、Gene Ontolog 等生物資料庫的註解內容。

6. 加入下列生物學特性：抗原、抗藥性、HI titer、病毒 recombination and reassortment 等分析於 IVBS - 因此在第二年度我們已將蒐集之流感病毒抗藥性(Drug Resistance)、抗原決定位(Epitope)、HI 力價(HI Titer)及流行病學等資料整合於 Ontology 系統之中。

7. 建立分析流感病毒序列的自動化分析– 核心設施持續在第二年度建立自動化的分析流程，增加分析的效率。

8. 增加所建立之分析流感病毒序列的自動化流程於 Influenza Virus Bioinformatics System (IVBS) -- 將建立與增加下列分析流感病毒序列的自動化流程於 IVBS 中:

    i. 在現有 IVBS 之 proteotype 分析外，增加以 sequence logo 及階層方式同時呈報大規模流感病毒蛋白質序列

排列中氨基酸之相似程度(requested by 中研院生醫所 Dr. 何美鄉及國衛院疫苗研發中心 Dr.李敏西)) (2008.03)

ii. 在現有之 IVBS 流感病毒基因序列分析外，增加流感病毒全基因體之基因型(genotyping)之預測與分佈功能 (requested by 台灣大學公衛所 Dr.金傳春)) (2008.03)

iii. 在現有之 IVBS 流感病毒基因序列資料庫外，增加使用者個人流感病毒基因序列資料之輸入儲存與後續分析功能 (requested by 成功大學分子醫學所 Dr.王貞仁)) (2008.06)

9. 建立與增加新的生物資訊學技術預測流感病毒的演化趨勢，協助選擇適合用來做流感疫苗的流感病毒株

i. 開發流感病毒蛋白質 B 細胞抗原決定位(B-cell epitope)胜肽片段預測程式及系統 (2008.06)

ii. 開發流感病毒蛋白質 CTL (cytotoxic T lymphocyte)細胞抗原決定位胜肽片段預測程式及系統 (2008.09)

iii. 開發完整流感病毒(whole virus) Hemagglutination activity 之 seropositivity 生物資訊預測程式及系統 (2008.12)

流感序列資料庫比較

流感序列、相關文獻、新聞與報告一直以來都是分散在不同的資料庫中，這些不同方面的資訊的整合，是於其他流感病毒資料庫中所沒有做到的。在流感病毒序列方面，IVBS 整合 IVR、ISD 與台灣疾管局所定序之台灣特有流感病毒，而其他之資料庫僅提供 IVR 或 ISD 所包含的序列。搜尋流感病毒序列全基因體方面，ISD 及由中國北京基因體研究所所建立的 IVDB (Chang et al., 2007) 並無提供此功能，全基因體的分析與泛用分析工具比較方面，FluGenome 並無提供多序列排比以及親緣分析工具，該資料庫著重於提供流感病毒 Genotype 的定型。於 IVBS 中，除了提供上述的分析工具外，亦提供預先計算的病毒定型資料，並且更進一步的提供 proteotype 查詢，這是目前流感病毒資料庫所僅見的。

在提供抗原決定序列 (Epitope) 資訊方面，目前 IVBS 提供已知的 epitope 資訊，同樣有提供 epitope 資訊的有 BioHealthBase，該資料庫以蒐集人類病原菌之序列為主，其特點在除提供已知 epitope 外，亦整合預測 epitope 之工具。流感病毒生物特性與流病資訊方面，本系統亦提供病毒抗藥性分析，而其他的資料庫於此方面並無著墨。

	IVBS	IVR	ISD	IVDB[3]	BioHealthBase[4]	FluGenome
Sequence source	CDC, IVR, ISD	GeneBank	GenBank, Directly summited	IVR, ISD	IVR, ISD	IVR

Complete genome sets	Y	Y	N	N	Y	Y
Alignment	Y	Y	Y[1]	Y	Y	N
Phylogenetic analysis	Y	Y	Y[1]	Y	Y	N
BLAST search	Y	Y	Y[1]	Y	Y	Y
Save working sets	Y	N	Y[1]	N	N	N
Geographic distribution	N	N	N	Y	N	N
Epitope information	Y[2]	N	N	N	Y	N
Sequence polymorphism	Y	N	N	Y	Y	N
Gene Ontology	Y	N	N	N	N	N
Proteotype	Y	N	N	N	N	N
Genotype	Y	N	N	N	N	Y
Drug resistance	Y	N	N	N	N	N
Epidemic	N	N	N	N	N	N

[1] Subscribed user only

[2] Provided epitope list

[3] BInfluenza Virus Database established by Beijing Institute of Genomics

[4] The Biodefense/Public Health DataBase (http://www.biohealthbase.org/)

本　　文

(8) 參考文獻

1. Alain J.P. Alix. (2000). Predictive estimation of protein linear epitopes by using the program PEOPLE. Vaccine 18: 311-314.
2. J.M.R. Parker, D. Guo, and R.S. Hodges. (1986). 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 25(19): 5425-5432.
3. P.Y. Chou and G.D. Fasman. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. Advances in Enzymology and Related Areas of Molecular Biology 47: 45-148.
4. E.A. Emini, J.V. Hughes, D.S. Perlow, and J. Boger. (1985). Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. Journal of Virology 55(3): 836-839.
5. A.S. Kolaskar and Prasad C. Tongaonkar. (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Letters 276(1-2): 172-174
6. P.A. Karplus and G.E. Schulz. (1985). Prediction of chain flexibility in proteins-a tool for the selection of peptide antigens. Naturwissenschaften 72: 212–213.
7. J. Chen, H. Liu, J. Yang, and K.C. Chou. (2007). Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33: 423-428.
8. L. Debelle, S. M. Wei, M. P. Jacob, W. Hornebeck, and A. J. P. Alix. (1992). Predictions of the secondary structure and antigenicity of human and bovine tropoelastins. European Biophysics Journal 21: 321-329.
9. A.D. Ghate, B.U. Bhagwat, S.G. Bhosle, S.M. Gadepalli and U.D. Kulkarni-Kale. (2007). Characterization of Antibody-Binding Sites on Proteins: Development of a Knowledgebase and Its Applications in Improving Epitope Prediction. Protein & Peptide Letters 14: 531-535.
10. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, Nemazee D, Ponomarenko JV, Sathiamurthy M, S choenberger S, Stewart S, Surko P, Way S, Wilson S, Sette A. (2005). The immune epitope database and analysis resource: from vision to blueprint. PLoS Biol. 3(3): e91.
11. Boni M.F., Gog J.R., Andreasen V., and Christiansen F.B. (2004). Influenza drift and epidemic size: the race between generating and escaping immunity. Theor

Popul Biol. *65*, 179-191.

12. Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J., and Fitch, W. M. (1999a). Predicting the evolution of human influenza A. Science *286*, 1921-1925.
13. Bush, R. M., Fitch, W. M., Bender, C. A., and Cox, N. J. (1999b). Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol *16*, 1457-1465.
14. Chang, S., Zhang, J., Liao, X., Zhu, X., Wang, D., Zhu, J., Feng, T., Zhu, B., Gao, G.F., Wang, J., *et al.* (2007). Influenza Virus Database (IVDB): an integrated information resource and analysis platform for influenza virus research. Nucleic Acids Res *35*, D376-380.
15. Cox, N. J., and Subbarao, K. (2000). Global epidemiology of influenza: past and present. Annu Rev Med *51*, 407-421.
16. Domingo E., Baranowski E., Ruiz-Jarabo C.M., Martin-Hernandez A.M., Saiz J.C., and Escarmis C. (1998). Quasispecies structure and persistence of RNA viruses. Emerg Infect Dis. *4*, 521-527.
17. Epperson E.S. and Tyrer H.W. (1995). Use of computer algorithms to reduce viral quasispecies sequence space. Biomed Sci Instrum. *31*, 83-88.
18. Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution *39*, 783-791.
19. Felsenstein, J. (1997). An alternating least squares approach to inferring phylogenies from pairwise distances. Syst Biol *46*, 101-111.
20. Ferguson N.M. and Anderson R.M.. (2002). Predicting evolutionary change in the influenza A virus. Nat Med. *8*, 562-563.Bush, R. M. (2001). Predicting adaptive evolution. Nat Rev Genet *2*, 387-392.
21. Fitch, W. M., Bush, R. M., Bender, C. A., and Cox, N. J. (1997). Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci U S A *94*, 7712-7718.
22. Fitch, W. M., Leiter, J. M., Li, X. Q., and Palese, P. (1991). Positive Darwinian evolution in human influenza A viruses. Proc Natl Acad Sci U S A *88*, 4270-4274.
23. Francis, T. (1940). A New Type of Virus from Epidemic Influenza. Science *92*, 405-408.
24. Ghedin, E., Sengamalay, N. A., Shumway, M., Zaborsky, J., Feldblyum, T., Subbu, V., Spiro, D. J., Sitz, J., Koo, H., Bolotov, P., *et al.* (2005). Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. Nature *437*, 1162-1166.
25. Gilbert, D., Ugawa, Y., Buchhorn, M., Wee, T. T., Mizushima, A., Kim, H., Chon, K., Weon, S., Ma, J., Ichiyanagi, Y., Liou, D. M., Keretho, S. and Napis, S.

(2004). Bio-Mirror project for public bio-data distribution. Bioinformatics *20*, 3238-3240.

26. Holmes, E.C., Ghedin, E., Miller, N., Taylor, J., Bao, Y., St George, K., Grenfell, B.T., Salzberg, S.L., Fraser, C.M., Lipman, D.J., *et al.* (2005). Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. PLoS biology *3*, e300.
27. Hughes, A. L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature *335*, 167.
28. Ina, Y., and Gojobori, T. (1994). Statistical analysis of nucleotide sequences of the hemagglutinin gene of human influenza A viruses. Proc Natl Acad Sci U S A *91*, 8388-8392.
29. Kamp C. (2003). A quasispecies approach to viral evolution in the context of an adaptive immune system. Microbes Infect. *5*, 1397-1405.
30. Lee, Y. H., Ota, T., and Vacquier, V. D. (1995). Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol Biol Evol *12*, 231-238.
31. Li, K.B. (2003). ClustalW-MPI: ClustalW analysis using distributed and parallel computing. Bioinformatics *19*, 1585-1586.
32. Li, W.-H., and Graur, D. (1991). Fundamentals of Molecular Evolution: Sunderland, Mass.: Sinauer Associates).
33. Lin, J., Andreasen, V., Casagrandi, R., and Levin, S. A. (2003). Traveling waves in a model of influenza A drift. J Theor Biol *222*, 437-445.
34. Maassab H.F. and Bryant M.L. (1999). The development of live attenuated cold-adapted influenza virus vaccine for humans. Rev Med Virol. *9*, 237-44.
35. Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol *3*, 418-426.
36. Nielsen, R., and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics *148*, 929-936.
37. Obenauer J.C., Denson J., Mehta P.K., Su X., Mukatira S., Finkelstein D.B., Xu X., Wang J., Ma J., Fan Y., Rakestraw K.M., Webster R.G., Hoffmann E., Krauss S., Zheng J., Zhang Z., Naeve C.W. (2006) Large-scale sequence analysis of avian influenza isolates. Science *311*,1576-1580.
38. Palese, P., and Young, J. F. (1982). Variation of influenza A, B, and C viruses. Science *215*, 1468-1474.
39. Pyhala R., Ikonen N., Haanpaa M., and Kinnunen L. (1996). HA1 domain of

influenza A (H3N2) viruses in Finland in 1989-1995: evolution, egg-adaptation and relationship to vaccine strains. Arch Virol. *141*, 1033-1046.

40. Sallie R. (2005). Replicative homeostasis II: influence of polymerase fidelity on RNA virus quasispecies biology: implications for immune recognition, viral autoimmunity and other "virus receptor" diseases. Virol J. *2*, 70-90
41. Smith, W., Andrewes, C., and Laidlaw, P. (1933). A virus obtained from influenza patients. Lancet *1*, 66-68.
42. Socolich M., Lockless S.W., Russ W.P., Lee H., Gardner K.H., and Ranganathan R. (2005). Evolutionary information for specifying a protein fold. Nature *437*, 512-518.
43. Stewart J.J., Watts P., and Litwin S. (2001). An algorithm for mapping positively selected members of quasispecies-type viruses. BMC Bioinformatics. *2*, 1
44. Terajima M., Jameson J., Norman J.E., Cruz J., and Ennis F.A. (1999) High-yield reassortant influenza vaccine production virus has a mutation at an HLA-A 2.1-restricted CD8+ CTL epitope on the NS1 protein. Virology *259*, 135-40.
45. Xu, X., Lindstrom, S. E., Shaw, M. W., Smith, C. B., Hall, H. E., Mungall, B. A., Subbarao, K., Cox, N. J., and Klimov, A. (2004). Reassortment and evolution of current human influenza A and B viruses. Virus Res *103*, 55-60.
46. Yang, Z. (2000). Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J Mol Evol *51*, 423-432.
47. Yang, Z., and Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. Trends In Ecology And Evolution *15*, 496-503.
48. Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A. M. (2000a). Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics *155*, 431-449.
49. Yang, Z., Swanson, W. J., and Vacquier, V. D. (2000b). Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol Biol Evol *17*, 1446-1455.
50. Webster, R.G., Bean, W.J., Gorman, O.T., Chambers, T.M., and Kawaoka, Y. (1992). Evolution and ecology of influenza A viruses. Microbiology and Molecular Biology Reviews *56*, 152-179.