計畫編號：DOH95-DC-1405

行政院衛生署疾病管制局九十五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、鐘翊方、楊曉珮、楊永正、李佩珊、林鈺倍、黃櫻雪、羅偉軒

執行期間： 95 年 1 月 1 日至 95 年 12 月 31 日

目　　錄

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。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。其次，核心設施將為流感疫苗研究發展計畫，建立各種有助於協調的工具，讓計畫下不同的團隊能有更好的互動，產生一加一大於二的效果。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。核心設施將根據其他的團隊的需求，建立自動化分析流程，以處理大量的資訊。我們在今年度所完成之工作如下:

1.已完成利用資訊代理人(web agent)收集下列最新流感病毒序列資料:
- NCBI (National Center for Biotechnology Information) 之 Influenza Virus Resource (IRV)
- LANL (Los Alamos National Laboratory)之 Influenza Sequence Database (ISD)
- 台灣疾管局所提供之流感病毒序列資料

2.已完成流感病毒序列資料之整合資訊庫及使用者搜尋介面之軟體架構. 此整合性資訊庫並具備下列分析功能:
- BLAST 序列相似性搜尋
- CLUSTAL W 多序列比對與 PHYLIP Genotype 親緣關係分析
- Proteotype 比對與親緣關係分析
- PDB (Protein Data Bank)蛋白質立體結構資訊與聯結

3.已完成利用 RSS 技術自動提供流感疫苗研究發展計畫的相關即時資訊功能.

4.已實際提供本生物資訊之核心設施服務予流感疫苗研發團隊. 並以實例說明本計畫已提供之核心設施服務。

關鍵詞：流感資訊核心設施、資訊代理人、資訊整合、親緣分析、分子演化、Proteotype 比對、蛋白質立體結構

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. 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. Another role of this core facility will be coordinating the interactions among different team projects, and getting the most out of the collaborations. The core will also seek for the capacity of processing large dataset and forming pipeline and quality control for data analysis. In the first year, we have accomplished the following works:

1. completed the collection of the following three influenza sequence data resources through the use of web agent and other tools:
   - Influenza Virus Resource (IRV) of NCBI (National Center for Biotechnology Information), USA
   - Influenza Sequence Database (ISD) of LANL (Los Alamos National Laboratory), USA
   - Influenza data from the Center of Disease Control (CDC) Taiwan
2. completed the integrated influenza sequence database and the graphical user interface (GUI) of its query and analysisfunctions:
   - BLAST sequence similarity search
   - CLUSTAL Wmultiple sequence alignment and PHYLIP phylogenetic Genotype analysis
   - Proteotype comparison and phylogenetic analysis
   - Protein structure information through PDB (Protein Data Bank)
3. completed the instant information and news system through the RSS technology
4. already started to provide bioinformatics analysis service to team members of the other related influenza research projects.

Keywords: influenza information core facility, web agent, integrated information, phylogenetic analysis, molecular evolution, proteotype comparison, protein structure

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。

在預測流感演化的趨勢的過程中，必須考慮到現有的數據在品質外，還可能有取樣的誤差，例如現在收集到的病毒株序列，可能是在疫苗的天擇壓力下被選擇出的變異株。換言之，這些數據可能不代表病毒株真正的演化趨勢，有一些具有流行潛力的新變異，已在人類族群中蘊釀，只是它們在取樣時並未被取到。因此區辨產生變異的假說就非常重要，有了正確的假說，才能設計正確的取樣方法。流病學的專家們在區辨假說的過程中，會需要蒐集最新的資訊，並嘗試用不同的方法將資料分組，再做親緣分析。因此流感資訊核心設施(Flu

Informatics Core)會使用各種資訊技術，例如資訊代理人（web agent）等，自動蒐集資訊。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。

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。

本研究的總目標在於

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
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
4. 建立分析流感病毒序列的自動化流程
5. 安裝或發展模擬軟體，預測流感病毒的演化趨勢
6. 利用資訊探採(data mining)技術，找到人工不易發現的關聯性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. 繼續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. 建立分析流感病毒序列的自動化流程
3. 安裝或發展模擬軟體，預測流感病毒的演化趨勢
4. 利用資訊探採(data mining)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

第三年目標

1. 繼續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. 利用資訊探採(data mining)技術，持續分析最新資訊。找到新的關聯性後，提供疾病管制局與計畫中的其他團隊參考
3. 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

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。

本　　文

(3) 結果

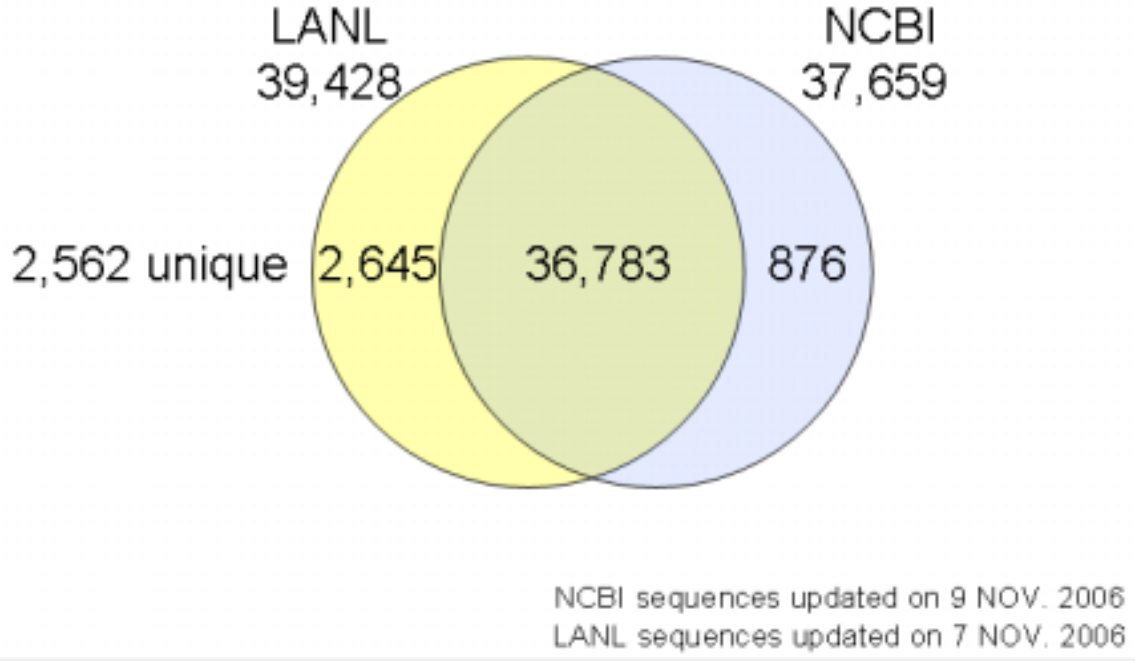
1. 利用資訊代理人(web agent)收集最新資訊

雖然在現有之主要公開流感病毒序列資料中，美國 NCBI (National Center for Biotechnology Information)所提供之 Influenza Virus Resource 流感病毒序列(http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html)可以下載使用，但是美國 Los Alamos National Laboratory (LANL)所提供之 ISD (Influenza Sequence Database)流感病毒序列(http://www.flu.lanl.gov/)則無法下載使用，因此我們使用智慧型代理人(Intelligent agent)又稱為軟體機器人(Software Robot)協助我們收集這些流感病毒資料。簡單來說，是一種能在使用者指定的環境下持續並自動執行指令，且能在不需使用者干預的情況下針對環境的改變做出適當的動作及回應的軟體系統。在理想情況下，智慧型代理人必須能夠從持續不斷的運作中學習到如何適應環境的變動，並能和同在此環境中的其他智慧型代理人溝通及合作，進而達成預定目標。為了蒐集、整合網路上的資訊，可以佈署成群的資訊代理人(web agent)來『代理』使用者瀏覽特定之網路資訊來源。搜尋引擎用來蒐集網頁的 Spider 或 Web crawler 程式也可算是簡單的資訊代理人，但這些程式通常只是把網頁整頁下載下來，而且只能沿著有 URL 網址的超連結來蒐集網

頁。較先進的資訊代理人，能自動執行填入關鍵字、勾選選項等動作，因此也能蒐集到動態產生的網頁。同時，先進的資訊代理人還能『理解』網頁的內容，將網頁內容加上特定的標記(Tag)以便於其他應用程式可以利用。

在本計畫中，我們將仿照在遺傳疾病與肝癌研究上的合作模式，依流行病專家的建議，設定流感疫苗研發計畫所需的「資訊代理人」，讓使用者將可透過資訊代理人取得最新的資訊。核心設施也將利用此工具收集相關資訊，整合到加值資料庫中。

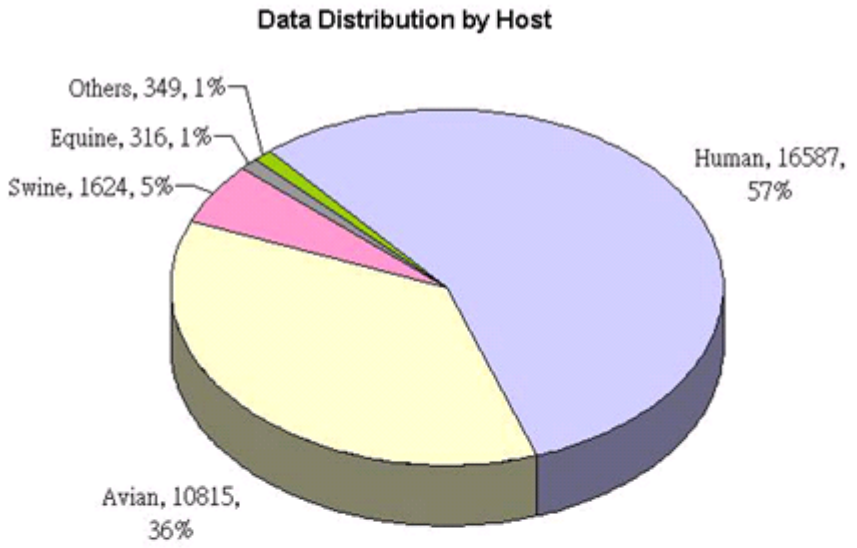
美國 NCBI 之 Influenza Virus Resource 與 LANL 之 Influenza Sequence Database 所提供流感病毒序列數目:

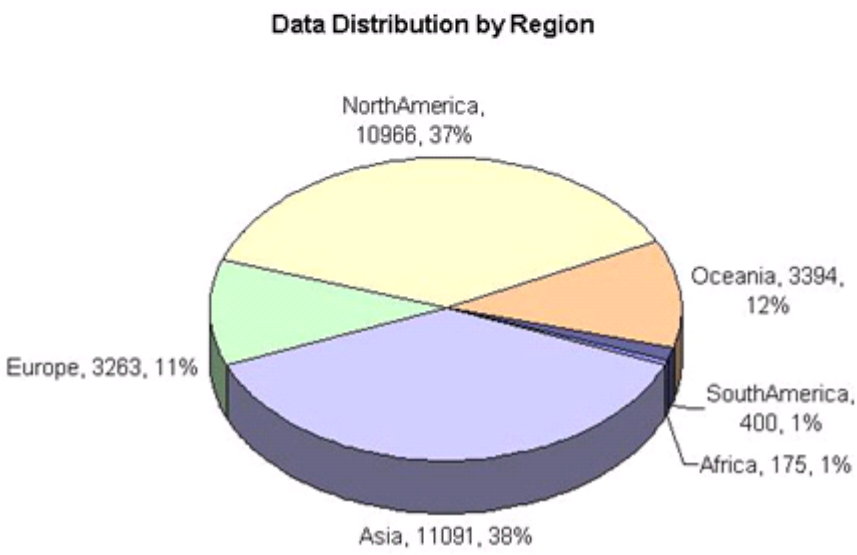


另有台灣疾管局所提供之流感病毒序列資料共 3779 筆，其資料格式如下

表所示: (序列內容因保密原因，在此隱藏)

Virtual_ID	Col_lab	City_name	Town_nam	Age	hap_date	serotype	subtype	VIRUS_name	DNA_RIGHT
00043	彰基	彰化縣	大村鄉	4	2002/12/26	A	H3	INFAH3	**********
00044	彰基	彰化縣	鹿港鎮	38	2003/1/8	A	H3	INFAH3	**********
00045	彰基	彰化縣	花壇鄉	30	2003/1/26	A	H3	INFAH3	**********
00046	彰基	彰化縣	花壇鄉	29	2003/1/26	A	H3	INFAH3	**********
00047	彰基	彰化縣	埔鹽鄉	4	2003/2/1	A	H3	INFAH3	**********
00048	彰基	彰化縣	二水鄉	6	2003/2/12	A	H3	INFAH3	**********
00049	彰基	雲林縣	虎尾鎮	2	2003/2/18	A	H3	INFAH3	**********
00050	彰基	彰化縣	溪湖鎮	3	2003/3/2	A	H3	INFAH3	**********



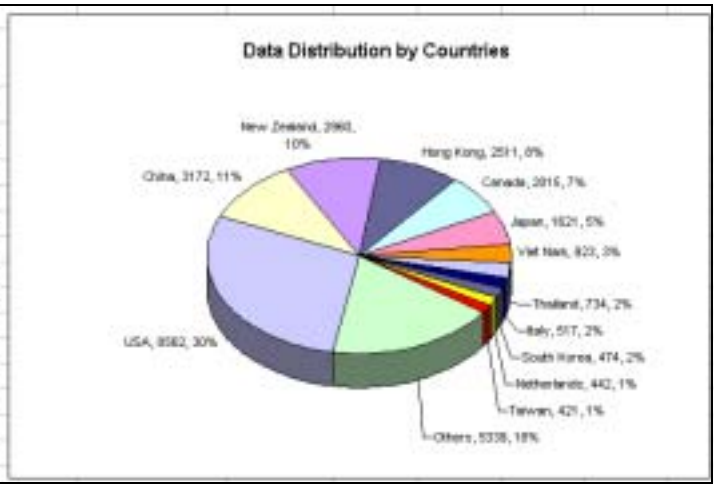


A. 資料內容依流感病毒宿主之區分

B. 資料內容依流感病毒在全球六大洲發生之區分

C. 資料內容依流感病毒發生國家之區分

Country	Region	Counts
USA	NorthAmerica	8662
China	Asia	3172
New Zealand	Oceania	2960
Hong Kong	Asia	2511
Canada	NorthAmerica	2015
Japan	Asia	1621
Viet Nam	Asia	923
Thailand	Asia	734
Italy	Europe	517
South Korea	Asia	474
Netherlands	Europe	442
Taiwan	Asia	421
Others		5338
Totals		29691



2. 利用資料庫整合各種資訊

本資訊網將收集各種有關流感病毒的資訊，包括下列各項:

1. 基因體序列資料庫：除本土的流感基因資訊外，收集 GenBank、LANL、CDC (Taiwan)等單位所提供之流感病毒株之基因體序列。
2. 蛋白序列資料庫：收集上述各序列資料庫網站所提供之流感病毒株之蛋白質序列。
3. 抗原相關基因資料庫：收集已知的流感病毒抗原相關基因序列以提供生物資訊分析。
4. 資料下載工具：提供使用者查詢及下載介面，以下載流行病研究所需之流感病毒基因體序列或蛋白質序列。
5. 分析工具：提供生物資訊分析工具如親緣分析、分子演化分析、序列資料庫比對及特定基因之抗原性分析等工具。

已完成

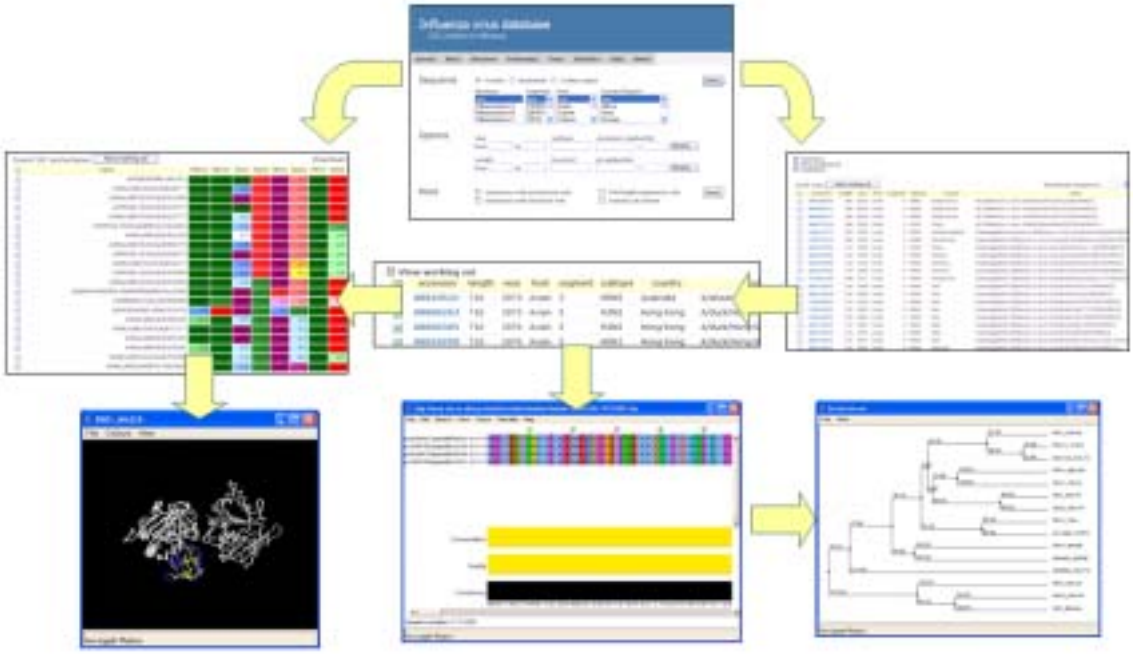
1. 美國 NCBI 之 Influenza Virus Resource (IRV) 與

2. 美國 LANL 之 Influenza Sequence Database (ISD) 及

3. 台灣疾病管制局

流感病毒資料序列資料之整合.

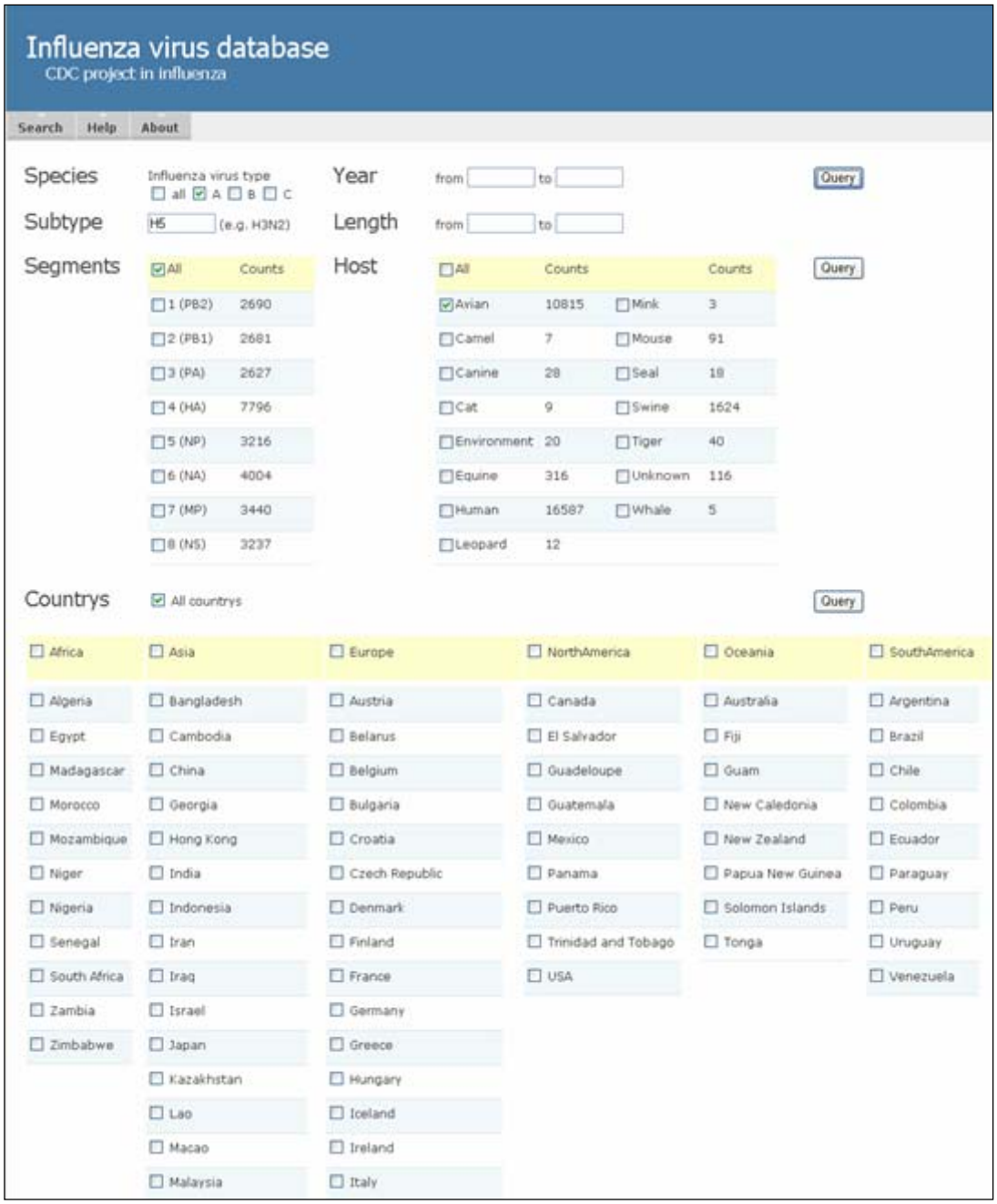
2.1.自動收集網際網路上所有公開的流感病毒資訊，並與臺灣特有的流感病毒資訊整合



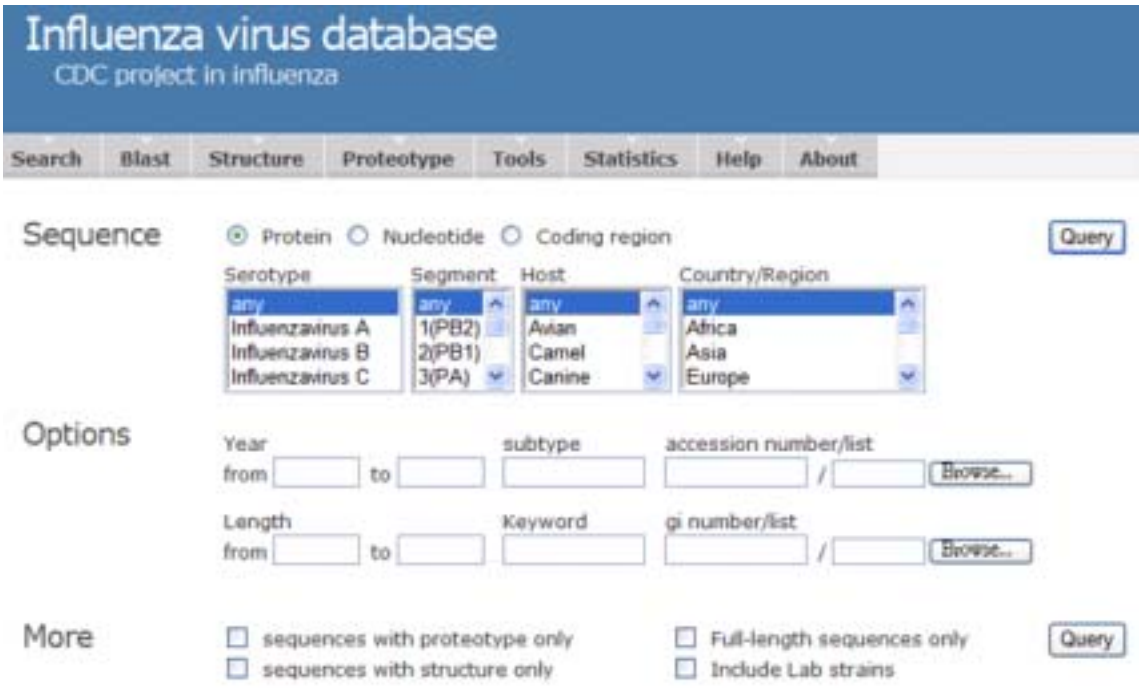
此為流感病毒資料庫之整體架構內容。

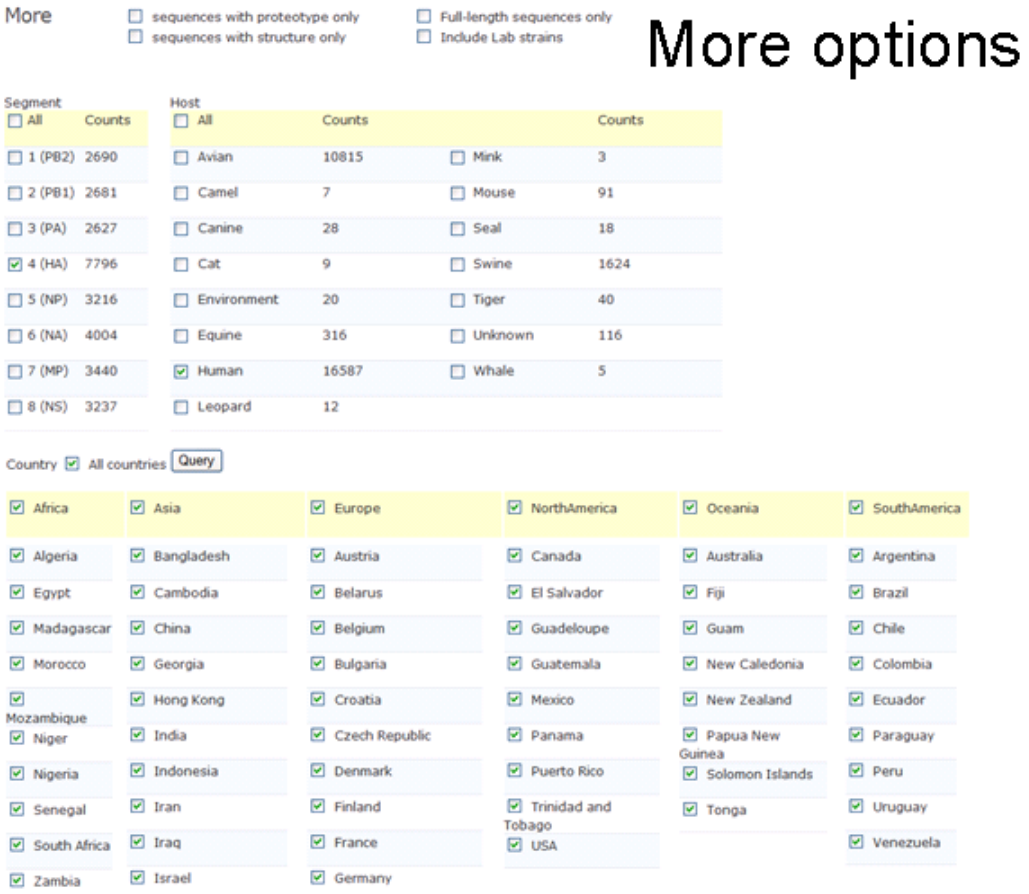
2.1.1.已完成流感病毒序列資料庫及使用者搜尋介面之軟體架構.

A. 流感病毒序列資料庫

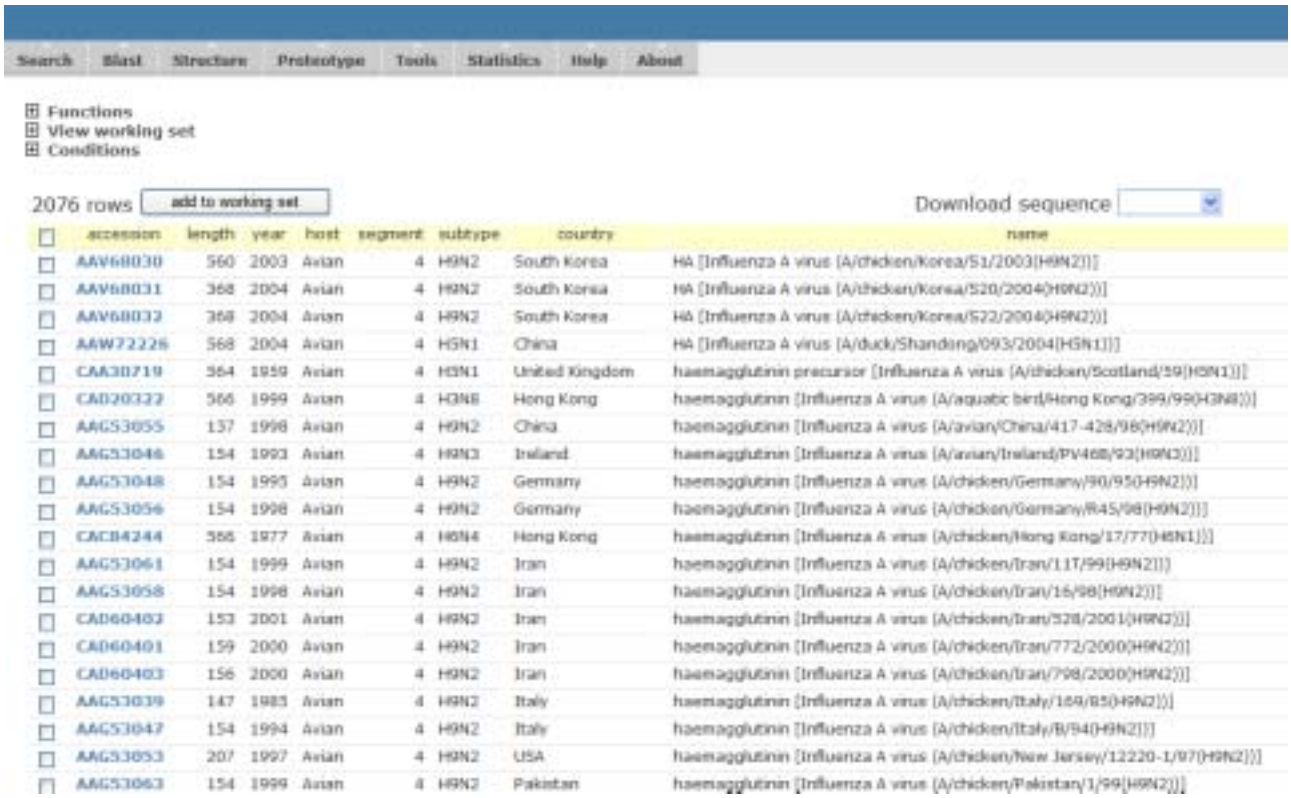


已完成之流感病毒序列資料庫及使用者搜尋介面選項:

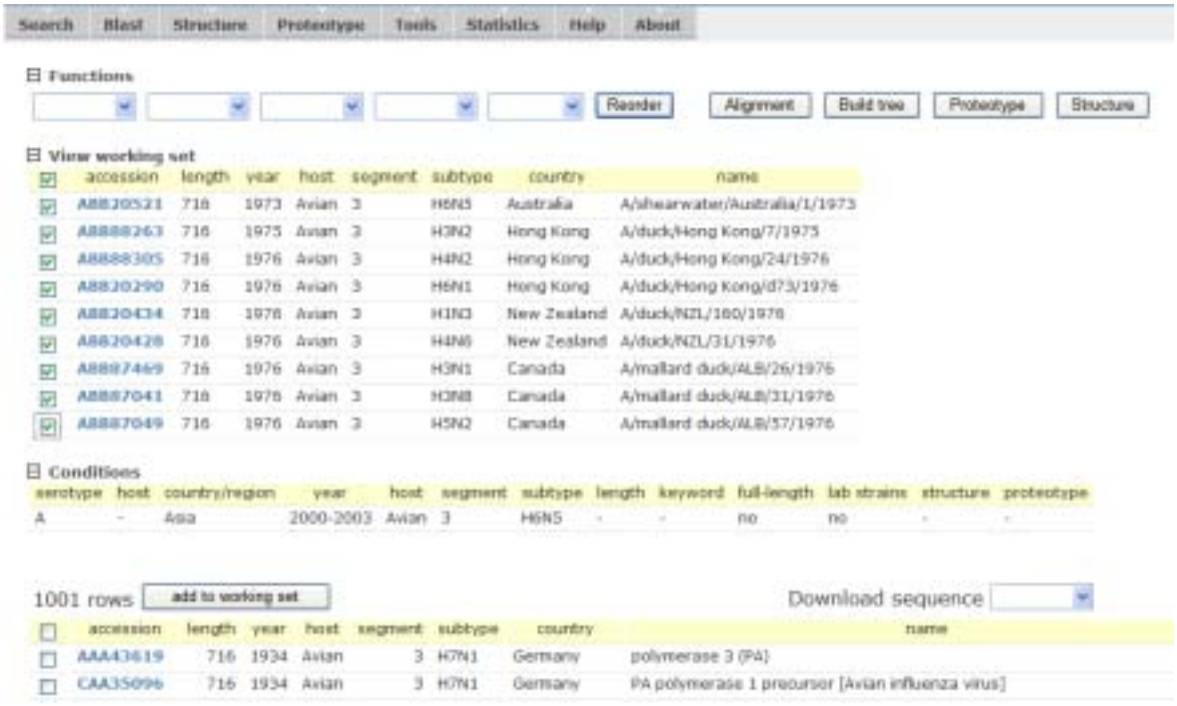

Influenza virus database
CDC project in influenza
Search
Blast
Structure
Proteotype
Tools
Statistics
Help
About
Sequence
Protein
Nucleotide
Coding region
Query
Serotype
any
Influenzavirus A
Influenzavirus B
Influenzavirus C
Segment
any
1(PB2)
2(PB1)
3(PA)
Host
any
Avian
Camel
Canine
Country/Region
any
Africa
Asia
Europe
Options
Year
from
to
subtype
accession number/list
Browse...
Length
from
to
Keyword
gi number/list
Browse...
More
sequences with proteotype only
sequences with structure only
Full-length sequences only
Include Lab strains
Query


More
sequences with proteotype only
sequences with structure only
Full-length sequences only
Include Lab strains
More options
Segment
All
Counts
1 (PB2) 2690
2 (PB1) 2681
3 (PA) 2627
4 (HA) 7796
5 (NP) 3216
6 (NA) 4004
7 (MP) 3440
8 (NS) 3237
Host
All
Counts
Avian 10815
Camel 7
Canine 28
Cat 9
Environment 20
Equine 316
Human 16587
Leopard 12
Mink 3
Mouse 91
Seal 18
Swine 1624
Tiger 40
Unknown 116
Whale 5
Country
All countries
Query
Africa
Algeria
Egypt
Madagascar
Morocco
Mozambique
Niger
Nigeria
Senegal
South Africa
Zambia
Asia
Bangladesh
Cambodia
China
Georgia
Hong Kong
India
Indonesia
Iran
Iraq
Israel
Europe
Austria
Belarus
Belgium
Bulgaria
Croatia
Czech Republic
Denmark
Finland
France
Germany
NorthAmerica
Canada
El Salvador
Guadeloupe
Guatemala
Mexico
Panama
Puerto Rico
Trinidad and Tobago
USA
Oceania
Australia
Fiji
Guam
New Caledonia
New Zealand
Papua New Guinea
Solomon Islands
Tonga
SouthAmerica
Argentina
Brazil
Chile
Colombia
Ecuador
Paraguay
Peru
Uruguay
Venezuela

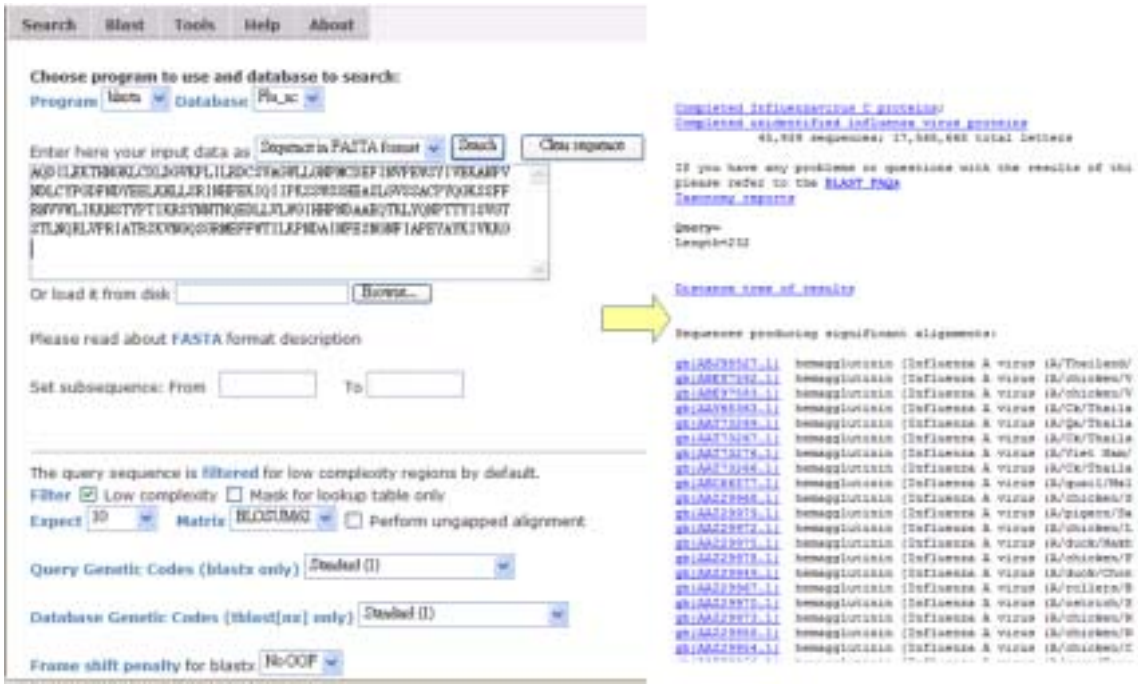
已完成之流感病毒序列資料庫搜尋結果顯示:



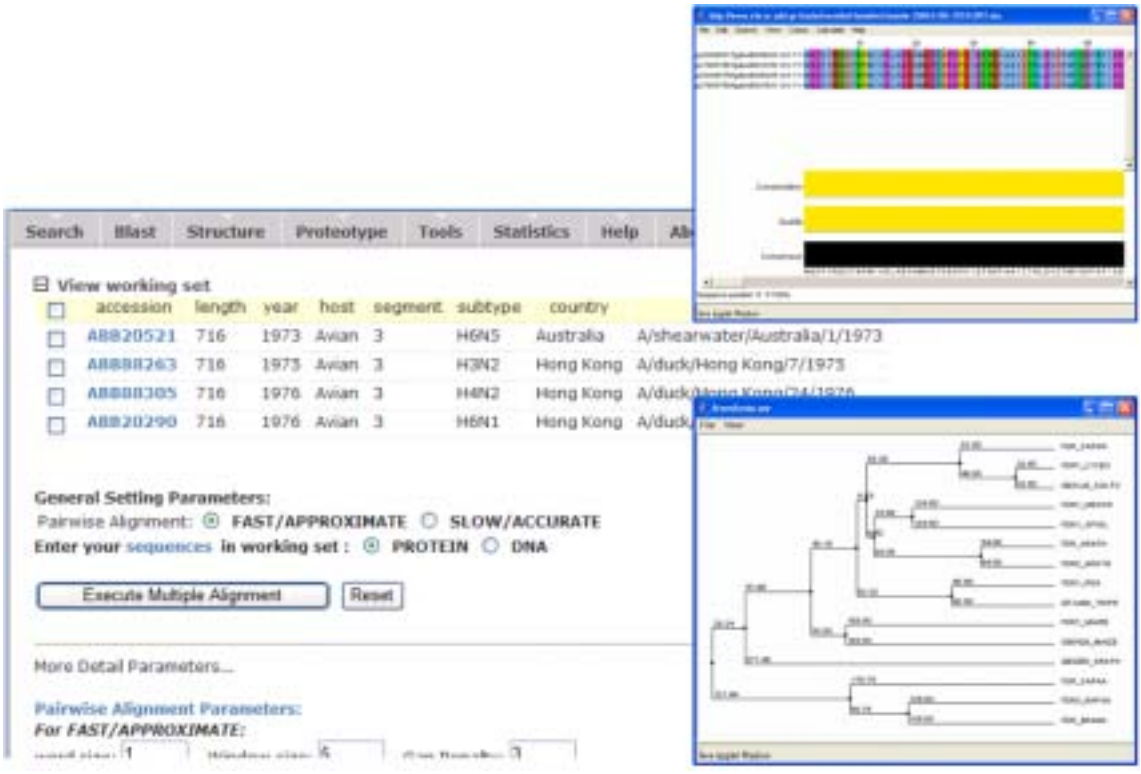
已完成之流感病毒序列資料庫所提供之各式功能選項:



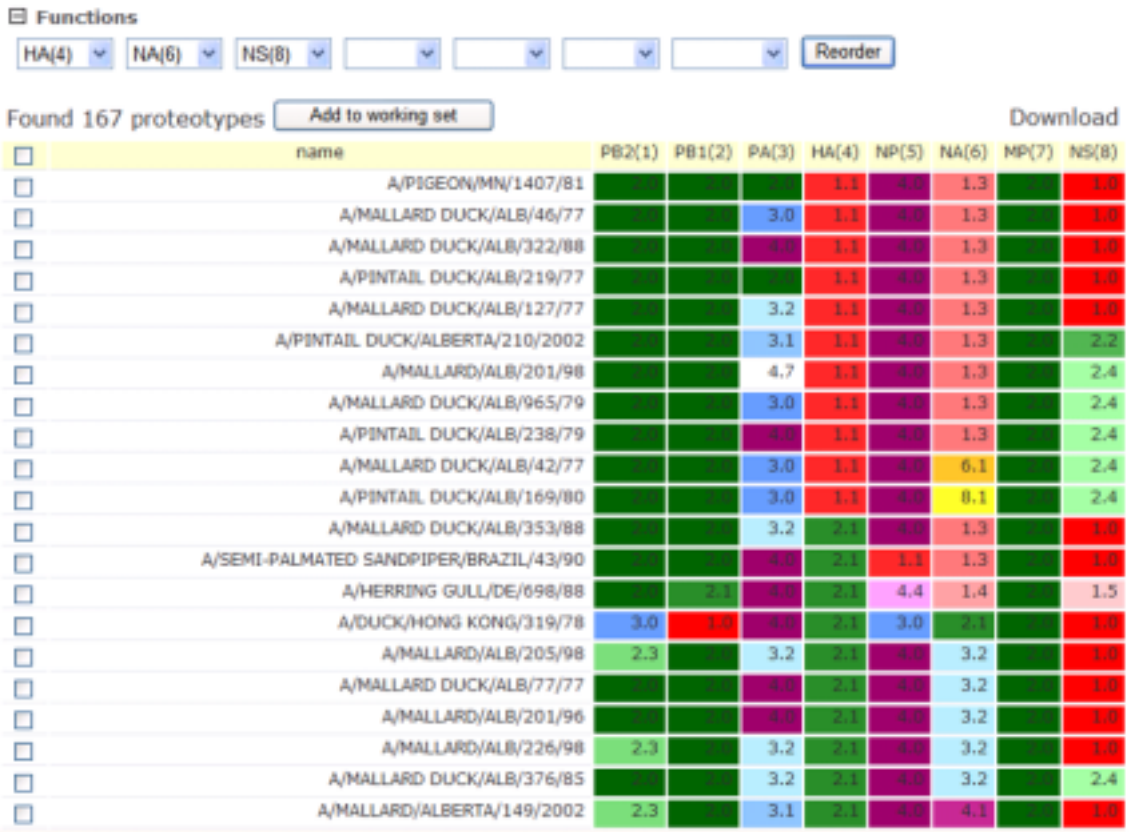
已完成之流感病毒序列資料庫所提供之 BLAST 序列相似性搜尋功能:



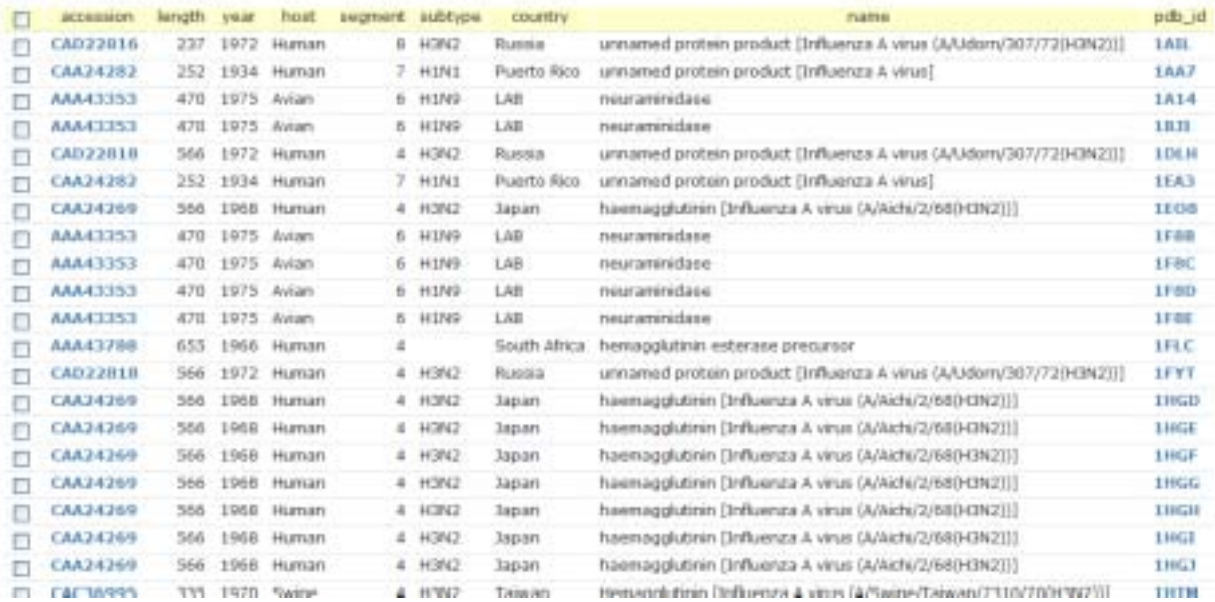
已完成之流感病毒序列資料庫所提供之 CLUSTAL W 多序列比對與 PHYLIP Genotype 親緣關係分析功能:



已完成之流感病毒序列資料庫所提供之 Proteotype 比對與親緣關係分析功能:



已完成之流感病毒序列資料庫所提供之 PDB (Protein Data Bank)蛋白質立體結構資訊與聯結功能:



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

RSS 為 XML 語言的應用，它利用 XML 許多預先定義好的標籤來呈現資訊的內容；透過這些標籤，讀者可以很快地掌握像是作者、發表時間、標題、描述等詮釋資料，也可以用這些詮釋資料來檢索、比對、排序、重組等。它被廣泛使用於將網站最新頭條訊息的內容，有效率地整理出來，讓讀者能夠直接讀取真正的資訊內容，進一步地能夠將資訊加以匯聚。這項功能如同網站上的「最新消息」功能，目前許多網站都使用 RSS 當作其訊息傳達的工具，各大新聞網站也相繼採用 RSS 取代新聞信件，成為新時代的媒體形式。

RSS 發展目的是想將資訊以 XML 為基礎的方式，附上不同的後設資料描述來提供資訊。換句話說，便是網站透過 RSS 來發佈消息，讓夥伴網站或讀者可透過簡單的程式或軟體，即可獲得想要的資訊。簡單而言，RSS 是屬於一種半主動式的傳播方式，讀者可以設定好自己所要蒐集的資訊，然後一次收回，以得知最新狀況。此一行為近似於使用者在收取電子郵件一般；相對於電子報收發的差異，在於電子報係以主動傳播方式，定時發佈消息給讀者，RSS 則是透過網站或軟體來進行。所以 RSS 的出現解決了網站管理者所必須面臨的許多問題，網站的郵件伺服器也毋須負擔發送大量信件，使得在網路上蒐集與傳遞訊息更加容易，同時也增進網路傳輸與連線品質，減少不必要

的人工作業。

目前許多網站都提供了 RSS feed，讓你可以「餵食」各式各樣的 RSS 閱讀器(Reader)；如此一來對於讀者來說，就不用擔心電子郵件地址外流、收到垃圾信，訂閱跟取消訂閱也更為便捷（祇需要在自己的 RSS 閱讀器裡設定即可，不用再麻煩地處理認證信函）。藉由 RSS 閱讀器能夠讓讀者得以自行篩選資訊，祇看想看的內容，不收不想看的東西，藉此而對資訊有更大的掌控權。

RSS 發佈系統之建構

為了收集世界上流行感冒或禽流感最新發佈消息或文章，我們試著建立自己的 RSS 發佈系統，希望讓計畫相關人員能快速與方便檢索到所需要的即時相關整合資訊。這邊依照 CDC 專家的建議及找尋現今流行感冒及禽流感資料發佈資訊網站，將所蒐集到的資料型態分成九個類別而分屬於六個網站即時資訊，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，這些網站資訊將進一步描述於後。這邊建議使用 Firefox 網頁瀏覽器(http://www.moztw.org/) 以及配合 Sage RSS 瀏覽功能(http://sage.mozdev.org/)觀看所建構之即時資訊發佈系統(此系統已掛

載於本計畫之流感序列資料庫的網頁上)。以下就所收集到相關網頁即時資訊的呈現方式分別描述如下：

(1) 整合資訊

圖一顯示所收蒐集到的六個網頁全部資訊，而為了讓使用者容易檢視，我們在每筆資料註解其來源的網站，如圖一中紅色匡線標示資料來自 NCBI、WHO 或 GoogleNews。再者提供每篇文章的標題、發表日期、作者、摘要介紹等，並提供直接連結到相對應的網頁，以及每日定時更新所有連結資訊，請見圖一中藍色匡線所示。不過由於六個網站所提供資訊的多樣化，為方便使用者能分別檢視不同型態資料以及作為日後進一步分類的依據，我們亦提供個別檢視不同資料的即時資訊。

(2) NCBI 即時資訊

由於目前美國 NCBI 的 PubMed 收集相當多文獻摘要以供檢索，因此我們以其做為收集流感與禽流感之學術文章的參考依據 這邊我們透過使用 NCBI 所提供的 Entrez Programming Utilities 來得到我們所需文章之訊息，然後再結合我們自己發展的格式轉換工具，將資料作適當的轉換。本系統暫定 RSS 只抓取最新 30 筆流感及禽流感相關文章發表狀況，以後可針對需求調整搜尋筆數以及定義不一樣找尋之主題，成果請見圖二所示。

(3) WHO 即時資訊

這邊將世界衛生組織 WHO 新聞站台中關於禽流感的 RSS 資訊納入本系統中，請見圖三所示。事實上 WHO 網站中還提供其他主題或疾病爆發之 RSS 資訊，以及網站中每週流行病學的紀錄等，都是值得我們未來考慮納入的即時發佈資訊。

(4) GoogleNews 即時資訊

現在許多新聞網站都提供不同類別的主題或讓使用者自行定義關鍵字，而搜尋所感興趣的內容以進行 RSS 新聞訂閱的動作。這邊我們以 avian flu 或 bird flu 當作關鍵字收集 Google 網頁中的相關即時新聞資訊，而展現於所建構之發佈系統，其結果請見圖四所示。未來我們可針對其他新聞網站如 BBC、CNN 等做相關資訊的搜尋，而亦可定義不同的關鍵字以找尋所需之資訊。

(5) PandemicFlu、ProMEDmail 及 CIDRAP 即時資訊

從 CDC 專家那邊得知這三個網站皆是收集流感及禽流感資訊之重要來源之處，因此進一步收錄這些網站提供之相關即時資訊。這邊我們將 PandemicFlu 網頁中(http://www.pandemicflu.gov/)所發佈的 News 之 RSS 即時訊息納入本發佈系統，請見圖五所示；而對 ProMEDmail 網頁中(http://www.promedmail.org/)關於 Latest Information on Avian influenza 的相關資訊轉化成本發佈系統之

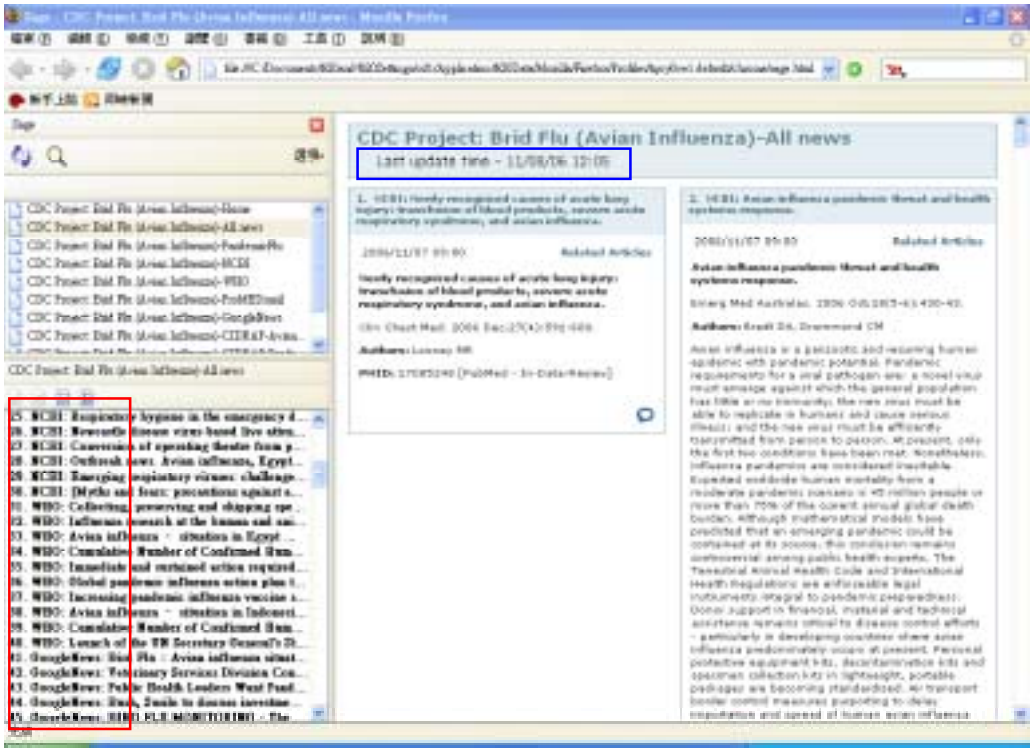
RSS 新聞發佈格式，請見圖六所示；至於 CIDRAP (http://www.cidrap.umn.edu/)則因其網頁提供了流感、禽流感及疫苗等最新訊息，因此我們分成三部分將其資訊轉化成本發佈系統之 RSS 新聞發佈格式，請見圖七　圖九所示。

持續進行之目標

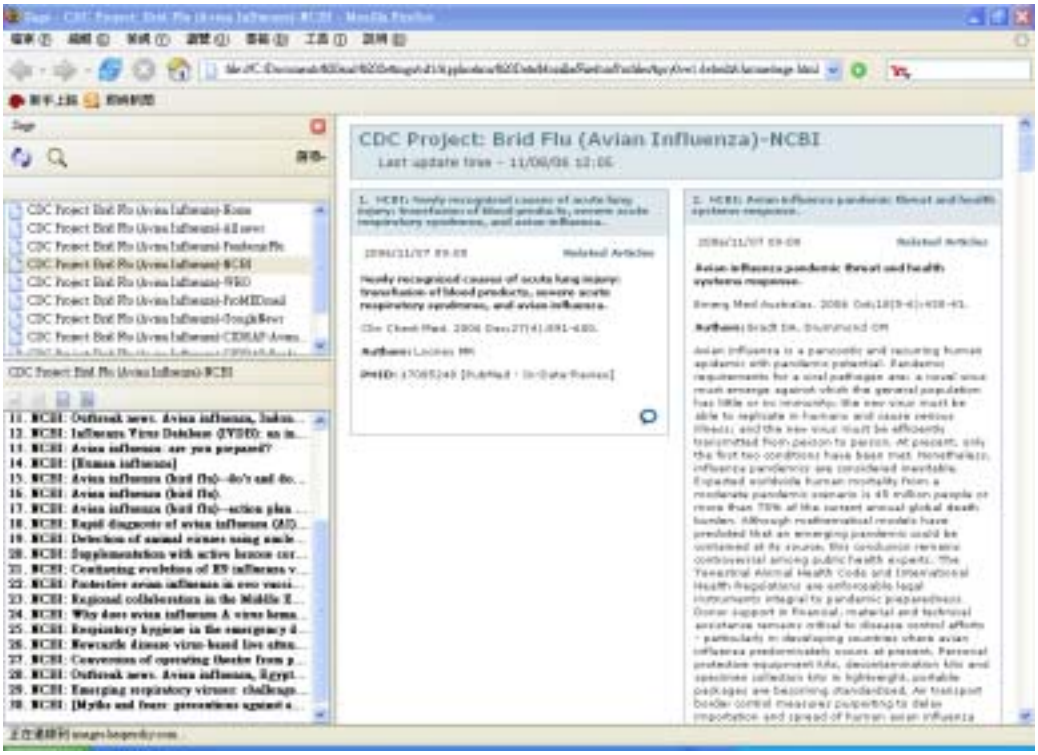
(1) 建立完善之 RSS 發佈系統，這邊除了建立易於使用的人機介面外，希望進一步蒐集更多相關流感與禽流感的資料，並將針對不同需求，建立更多樣化的 RSS 發佈系統。

(2) 將 RSS 即時資訊加入資料庫中，以供未來查詢相關舊有資訊。

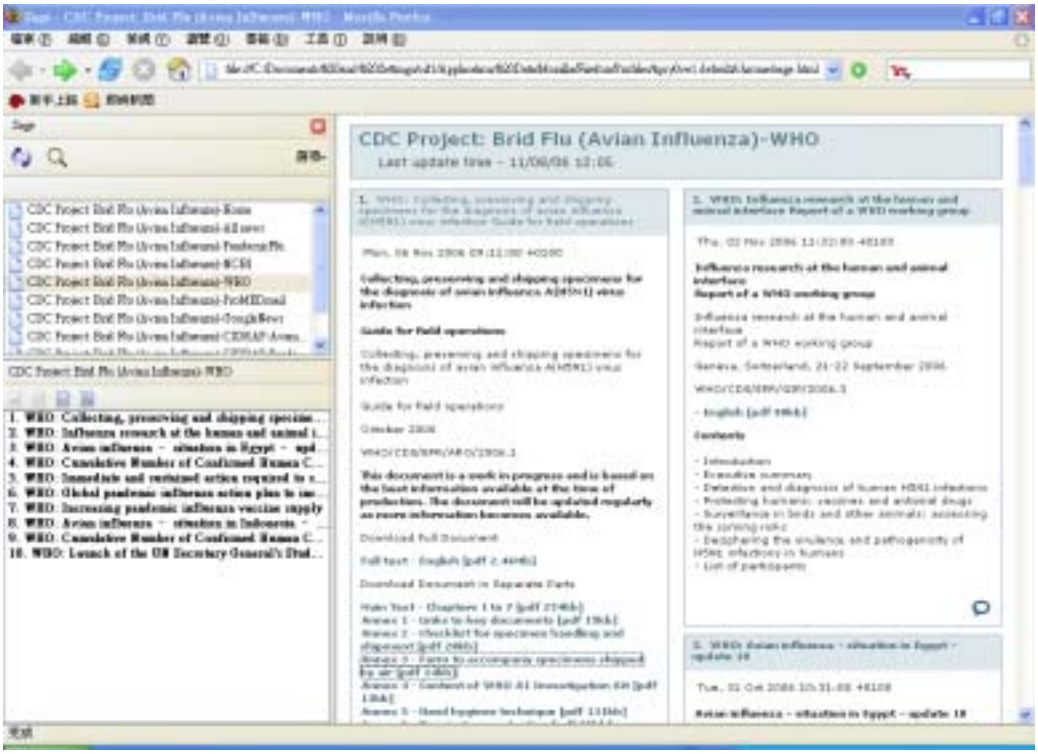
(3) 希望加入文字探勘(text-mining)技術，將我們感興趣的關鍵字及關連性標示出來，讓讀者更快掌握此學術文章或新聞文章所探討的內容。



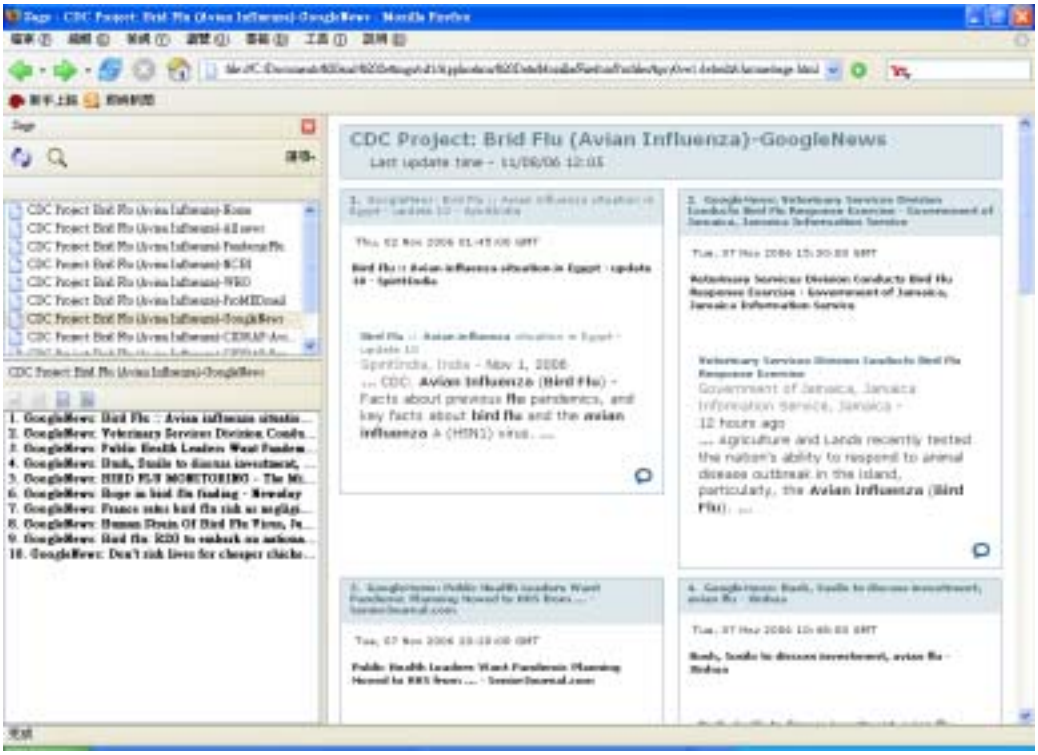
圖一：整合收集到的流感與禽流感即時資訊，所建立之 RSS 發佈系統



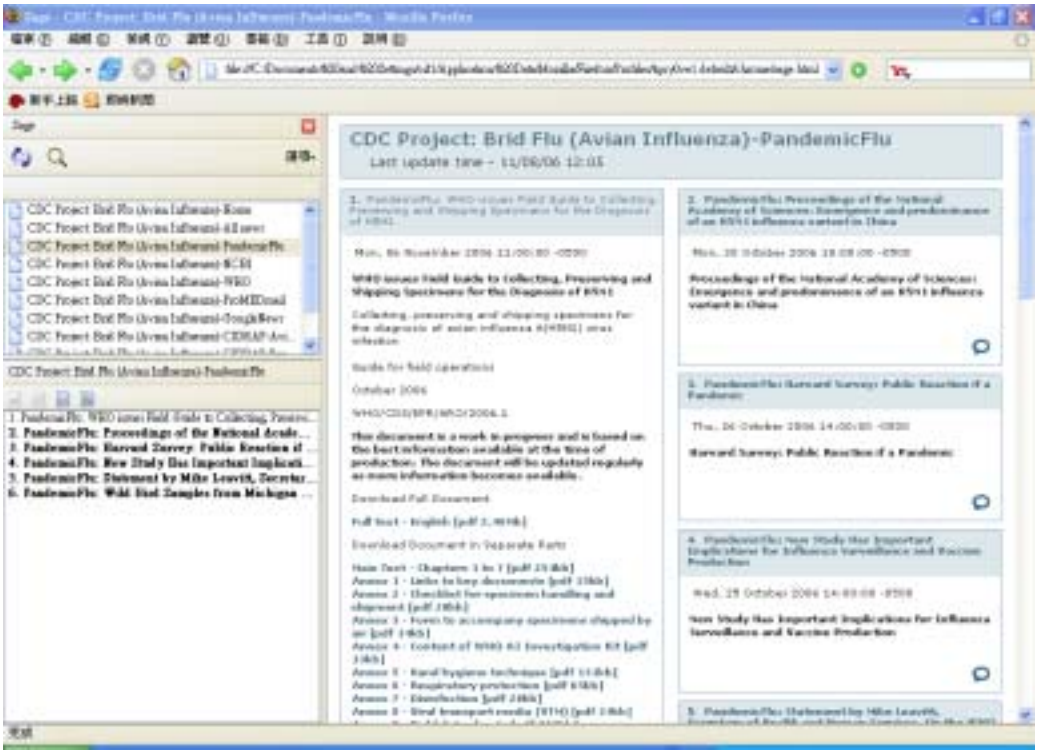
圖二：以 NCBI 的 PubMed 提供的流感與禽流感文獻資訊，所建立之 RSS 發佈系統



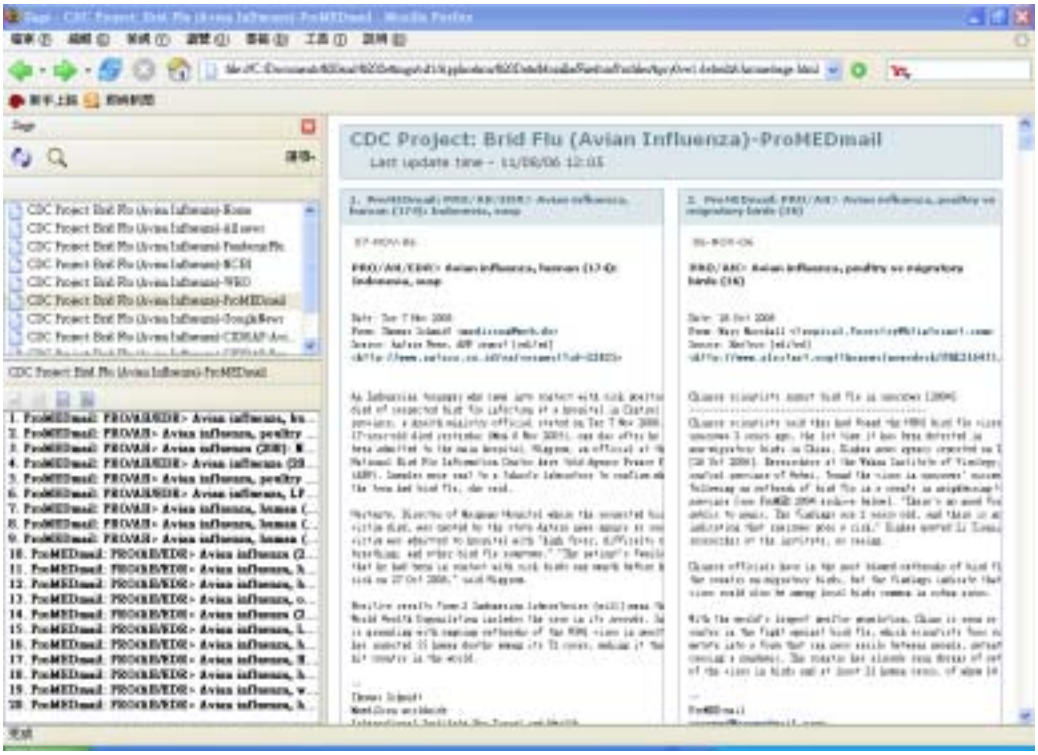
圖三：納入 WHO 禽流感 RSS 資訊於所建立之 RSS 發佈系統



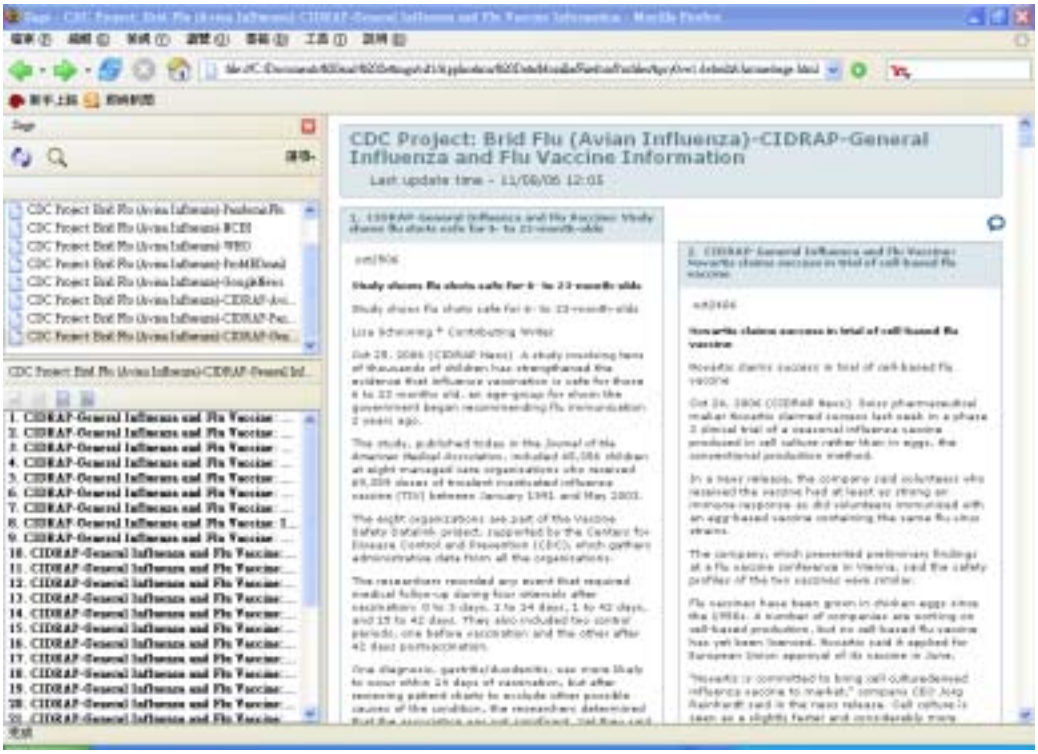
圖四：擷取 GoogleNews 關於 Avian Flu 及 Bird Flu 之相關訊息於所建立之 RSS 發佈系統



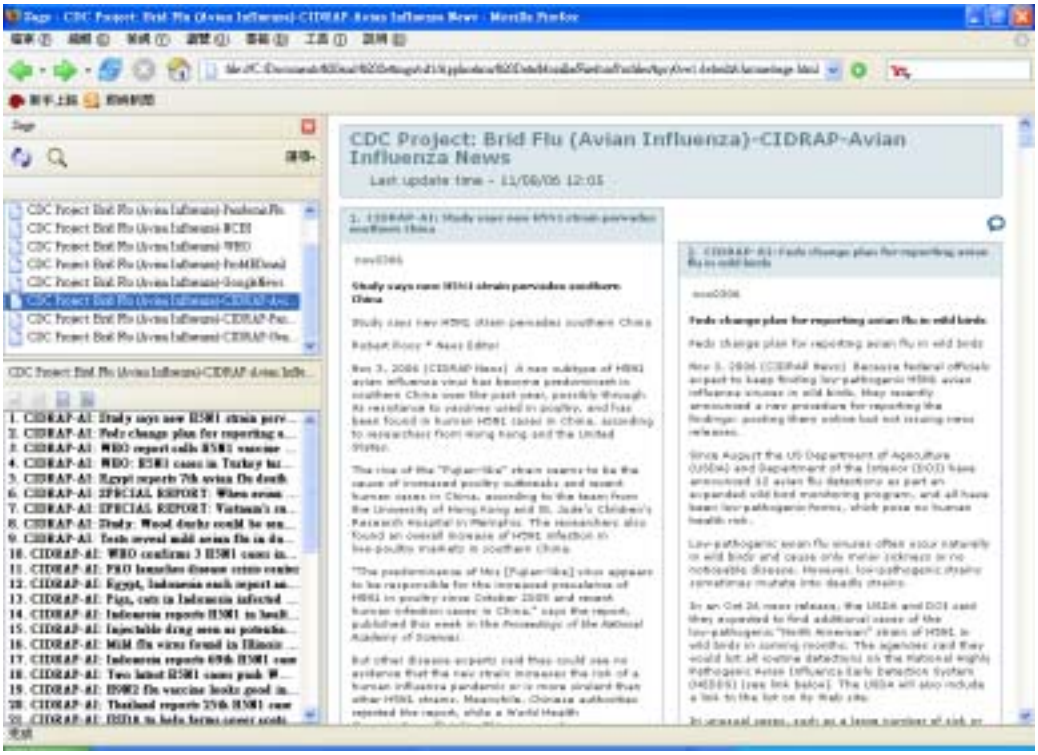
圖五：納入 PandemicFlu 網頁 News 之 RSS 即時訊息於所建立之 RSS 發佈系統



圖六：將 ProMEDmail 網頁中關於 Latest Information on Avian influenza 的相關資訊顯示於所建立之 RSS 發佈系統



圖七：將 CIDRAP 網頁中關於疫苗相關資訊顯示於所建立之 RSS 發佈系統



圖八：將 CIDRAP 網頁中關於禽流感相關資訊顯示於所建立之 RSS 發佈系統

圖九：將 CIDRAP 網頁中關於流感相關資訊顯示於所建立之 RSS 發佈系統

4. 核心設施服務

流感資訊核心設施集合了各個領域的專家，以團隊合作的方式提供生醫實驗室有策略的、有用的、有效率的生物資訊學的解決方法。根據生醫研究人員對於生物資訊的了解程度分成三類:

1. 知道如何利用生物資訊學方法去解決問題，但是對於執行大量的分析有

技術上的困難，對於這類的研究人員，可以利用例行程序平台分析來達到他們的需求；

2. 對於生物資訊有某種程度上的了解，也能夠自行去執行生物資訊的分

析，對於這類的研究人員，可以提供諮詢或訓練課程，以便能夠利用工具和加值資料庫去解決所面臨的問題；

4. 對於生物資訊了解甚少，他們需要學習如何使用生物資訊來加速他們的

   研究，對於這類的研究人員，可以藉由建立研究合作的方式，來提供有用的解決方法，一但與我們計畫的人員合作之後，他們可以學習到更多的生物資訊的方法，最後甚至能夠自行去執行。

核心技術

我們核心主要的技術有以下幾種：

1. 資訊的自動蒐集與整合
2. 各種生物資訊工具的使用與數據解讀
3. 自動化分析流程的設計與建立
4. 主題資訊網的建立
5. 資訊比對與資訊探採
6. 生物資訊工具的開發

以下舉一實例說明本計畫已提供之核心設施服務:

材料：

疾管局此流感疫苗計畫下之其中一團隊提供自己定序的 5 株流感病毒的 nucleotide sequence 請我們分析，其資料如下，列出的 strain name

為原始名稱，後面接著為自行計算之長度。從 NCBI 序列中條件設定為 host: human, subtype: H3, hemagglutinin, year from 2003-2006 取得 927 筆序列。

方法：

為了比較全球的流感病毒 strain 與此團隊所提供的 strain 在 protein sequence 特定位置上的差異，我們先把此團隊所提供的序列轉成 protein sequence。決定 frame 的方法: 將序列做 BlastX，取出 BlastX 結果裡最佳 hit 中的 frame，並檢查是否合理。接著用 EMBOSS 中的 Transeq 轉成 protein 序列。Public domain 取得的序列因為效能的因素選用 MUSCLE 作為 alignment 工具。並將 alignments 用 BioEdit 做進一步的調整跟修正以得到較佳的結果。我們撰寫了一個程式來比較氨基酸變異頻繁位置（Positive selection site)、唾液酸受器接合位置（Sialic acid receptor binding site）與抗原決定位置（Epitope/Antibody binding site）在全球流感病毒 strain 與此團隊提供的 strain 的變異情形。該程式會回報在不同位置上不同地區與年份的變異情形。我們用 WebLogo 將這些變異情形重新呈現成圖形化表示，方便我們進一步分析。

結果：

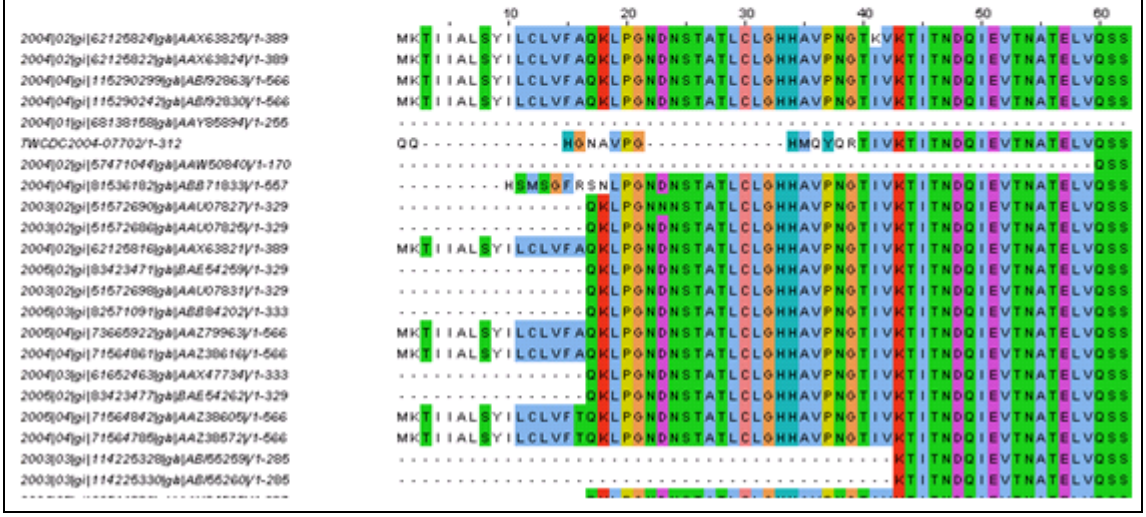
以下列圖示說明此核心設施服務之氨基酸變異頻繁位置（Positive selection site）分析部份結果。

Translation

```
                                                                  Score     E
Sequences producing significant alignments:                       (Bits)  Value

gi|68138162|gb|AAY85896.1|  hemagglutinin [Influenza A virus (A/Y   349     3e-95
gi|68138156|gb|AAY85893.1|  hemagglutinin [Influenza A virus (...   349     3e-95
gi|68138168|gb|AAY85899.1|  hemagglutinin [Influenza A virus (A/T   349     3e-95
gi|62125805|gb|AAX63816.1|  hemagglutinin [Influenza A Virus (...   349     3e-95
gi|62125809|gb|AAX63818.1|  hemagglutinin [Influenza A Virus (A/F   349     3e-95
gi|47230945|gb|AAT12674.1|  hemagglutinin [Influenza A virus (A/D   348     6e-95
gi|62125811|gb|AAX63819.1|  hemagglutinin [Influenza A Virus (A/F   348     7e-95

 Score =  349 bits (895),  Expect = 3e-95
 Identities = 169/169 (100%), Positives = 169/169 (100%), Gaps = 0/169 (0%)
 Frame = +3

Query  3    VTMPNNEKFDKLYIWGVHHPGTDSDQINLYAQASGRITVSTKRSQQTVIPNIGSRPRVRD  182
            VTMPNNEKFDKLYIWGVHHPGTDSDQINLYAQASGRITVSTKRSQQTVIPNIGSRPRVRD
Sbjct  71   VTMPNNEKFDKLYIWGVHHPGTDSDQINLYAQASGRITVSTKRSQQTVIPNIGSRPRVRD  130

Query  183  VPSRISIYWTIVKPGDILLINSTGNLIAPRGYFKIRSGKSSIMRSDAPIGKCNSECITPN  362
            VPSRISIYWTIVKPGDILLINSTGNLIAPRGYFKIRSGKSSIMRSDAPIGKCNSECITPN
Sbjct  131  VPSRISIYWTIVKPGDILLINSTGNLIAPRGYFKIRSGKSSIMRSDAPIGKCNSECITPN  190

Query  363  GSIPNDKPFQNVNRITYGACPRYVKQNTLKLATGMRNVPEKQTRGIFGA  509
            GSIPNDKPFQNVNRITYGACPRYVKQNTLKLATGMRNVPEKQTRGIFGA
Sbjct  191  GSIPNDKPFQNVNRITYGACPRYVKQNTLKLATGMRNVPEKQTRGIFGA  239
```

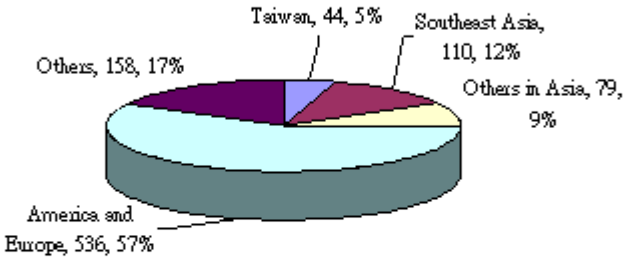
Multiple sequences alignment

- Using Muscle and ClustalW
- Manually curate

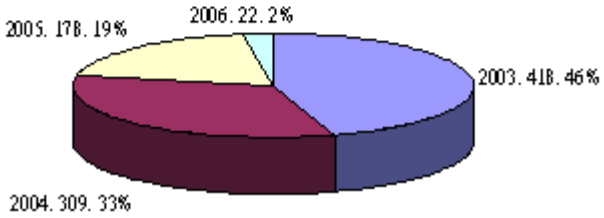


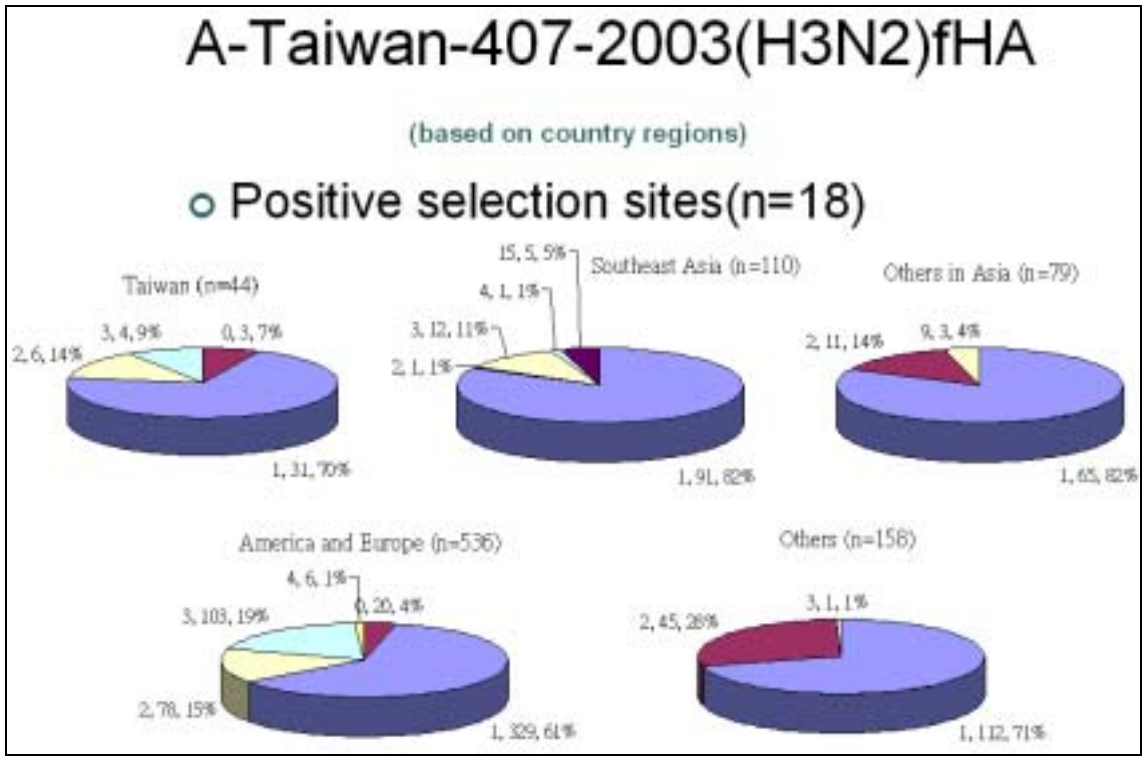
Data set

- Countries distribution (from NCBI total 927)



- Years distribution (from NCBI total 927)

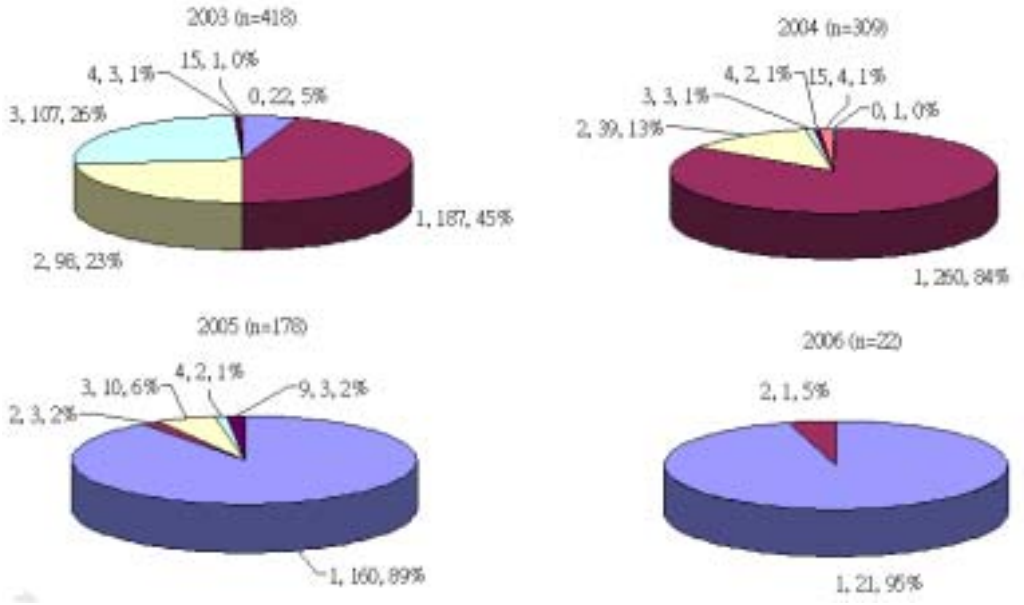




A-Taiwan-407-2003(H3N2)fHA

(based on occurred year)

- Positive selection site(n=18)



Variations shown with WebLogo

A-Taiwan-407-2003(H3N2)fHA

(based on country regions)

Variations shown with WebLogo

A-Taiwan-407-2003(H3N2)fHA

(based on occurred year)

2003

2004

2005

2006

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。

另外由於親緣分析 (phylogentic analysis)與分子演化的研究是本計畫最常需要使用的方法，所以我們也比較與評估開放軟體，並選擇適當的工具做流程分析。我們依下列步驟，連接不同的分析工具功能:

多序列排比(multiple sequence alignment)

在結果本文中，已說明流行病序列分析流程的設計原理與成果。在本計畫中我們讓流程更順暢。在面對大量序列時，我們將採用平行化的 MPI-ClustalW (Li 2003)，以增加分析的速度。

親緣分析

將 DNA sequence alignment 的結果，以親緣關係分析軟體 PHYLIP (Felsenstein 1997)以 Maximum Likelihood 的方法進行親緣關係樹的分析。我們假設 DNA 序列的演化，是依照 Kimura two-parameter model (考慮 transition 與 transversion 速率的差異(Li and Graur 1991),並假設不同的位置上的 nucleotide 取代的速率不同，其變異遵守 γ 分佈，如此我們可以評估依據現有的流感病毒 DNA 序列資料,在以上假設的演化模式下，彼此間的親緣關係以不同的樹狀圖表示的或然率。

採用 Maximum Likelihood 分析方法的優點,在於我們可以根據現有的序列資料，推斷出每一個樹狀圖的內部結點(internal nodes)的序列狀態（state），這個狀態代表結點末端的序列之共同祖徵（common ancestral states）,這對於我們分析特定位置上的序列的演化過程，提供重要的資訊. 這項結果是以演化距離為依據，而做出來的親緣關係樹，所無法做到的。此外，Maximum Likelihood 分析方法，同時提供演化距離的估算，讓我們能夠進一步地分析特定位置上序列的演化速率。

然而，以 Maximum Likelihood 進行親緣關係樹的分析的缺點,在於所需要的計算量很大，尤其是當我們有大量的 DNA 序列時，要從眾多可能的樹狀圖中，決定最佳的親緣關係樹，計算上是十分困難度的。因此，我們必須以 heuristic 的方法，找尋最佳的親緣關係樹。我們將會以

neighbor-joining 的方法，配合 tree-grafting, branch-swapping…等方式，找出 Likelihood 較高的 10~20 株 candidate trees，再進一步分析比較根據這些 candidate trees 所預測出受正向天擇壓力的序列位置的差異。

我們也額外完成了流感病毒序列資料庫之 Proteotype 比對與親緣關係分析功能。

在第二與第三年度裡，我們將完成

<u>尋找受到正向天擇壓力作用之下的可能位置(candidate sites)</u>

我們將以由 Dr. Zhi-Heng Yang 等人(Yang 2000; Yang, Nielsen et al. 2000; Yang and Bielawski 2000; Yang, Swanson et al. 2000) 所共同開發來偵測正向天擇壓力的序列位置分析方法來進行。他們的方法，首先計算每一個以 Maximum Likelihood 分析法所得的樹狀圖中的分枝上,出現了多少的同義與非同義互換(synonymous vs non-synonymous substitution)，並以此估算每一分枝上發生同義與非同義互換之速率，據此找出演化速率特別高的非同義互換序列之位置，與其出現在樹狀圖中之分佈。

根據以上的分析結果，我們能夠以不同方式所建構出的親緣關係樹狀圖，評估受到正向天擇壓力作用之下的可能位置之預測的準確度。此外利用流感病毒 HA 蛋白質的演化樹，可以進一步比較 endemic strains 與 vaccine strains 之間的差異，以評估 endemic strains 的演化方向，是否可以由前一季的流感疫苗的種類來做預測。

我們將協助流感疫苗計畫下的其它計畫，進一步結合本土流感病毒之資料，以及全球的流感病毒基因體資料庫的資料，建構整個流感病毒基因體序列親緣關係，探討本土性流感與全球性流感的暴發之關聯。此外根據整個流感病毒基因體資料的親緣分析，流行病學者可以探討 HA domain 的演化過程中變異性的來源，是受到由現有的 circulating strains，經由 antigenic drift 的機制提供;或是由流感病毒不同基因片段，經由 gene reassortment 所得到的結果。由此更可決定選取哪一株分離株，進行全基因體定序，以區辨產生基因變異的假說。

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: 流感資訊網即將完成架設提供 RSS 服務

流感資訊核心設施所開發之流感資訊網，分析流程方面，在稍做修改後，可應用其他流行性疾病的團隊合作。其中資訊代理人服務之應用、RSS 在生物資訊學上的應用、客制化的流感病毒分析工作流程等，均可發表生物資訊學領域的論文，其結果不但有助於預測流感病毒的演化趨勢上，也可預測其它流病的病毒演化趨勢。在接下來的預測特異的抗原區部份，與資訊探採的方法發展，其結果不但將有助於發展流感疫苗，也可應用在其它流行性疾病的疫苗發展上。 後面兩部份都可與合作者一起發表流病領域的論文。

在第二與第三年度裡，我們將完成

第二年目標

5. 繼續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
6. 建立分析流感病毒序列的自動化流程
7. 安裝或發展模擬軟體，預測流感病毒的演化趨勢
8. 利用資訊探採(data mining)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

第三年目標

5. 繼續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
6. 利用資訊探採(data mining)技術，持續分析最新資訊。找到新的關聯性後，提供疾病管制局與計畫中的其他團隊參考
7. 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

本　　文

(6) 計畫重要研究成果及具體建議

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。

在第二與第三年度裡，我們將完成

4. 建立分析流感病毒序列的自動化流程

5. 安裝或發展模擬軟體，預測流感病毒的演化趨勢

6. 利用資訊探採(data mining)技術，找到人工不易發現的關聯性

7. 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

具體建議

台灣疾管局進行流感疫苗研究發展計畫應主動協助流感病毒資料之取得，以利計畫之加速進行。本「流感病毒生物資訊系統之建立」計畫今年度開始之初即請求協助取得疾管局之流感病毒序列資料，但遲未獲得答覆提供，以致所需進行之流感病毒資料整合工作延後完成，因此希望疾管局之內部協調工作能確實配合計畫之執行所需。

本　　文

(7) 參考文獻

1. 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.
2. 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.
3. 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.
4. 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.
5. Cox, N. J., and Subbarao, K. (2000). Global epidemiology of influenza: past and present. Annu Rev Med *51*, 407-421.
6. 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.
7. Epperson E.S. and Tyrer H.W. (1995). Use of computer algorithms to reduce viral quasispecies sequence space. Biomed Sci Instrum. *31*, 83-88.
8. Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution *39*, 783-791.
9. Felsenstein, J. (1997). An alternating least squares approach to inferring phylogenies from pairwise distances. Syst Biol *46*, 101-111.
10. 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.
11. 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.
12. 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.
13. Francis, T. (1940). A New Type of Virus from Epidemic Influenza. Science *92*,

405-408.

14. 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.
15. 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.
16. Hughes, A. L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature *335*, 167.
17. 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.
18. Kamp C. (2003). A quasispecies approach to viral evolution in the context of an adaptive immune system. Microbes Infect. *5*, 1397-1405.
19. 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.
20. Li, K.B. (2003). ClustalW-MPI: ClustalW analysis using distributed and parallel computing. Bioinformatics *19*, 1585-1586.
21. Li, W.-H., and Graur, D. (1991). Fundamentals of Molecular Evolution: Sunderland, Mass.: Sinauer Associates).
22. 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.
23. 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.
24. Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol *3*, 418-426.
25. 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.
26. Palese, P., and Young, J. F. (1982). Variation of influenza A, B, and C viruses. Science *215*, 1468-1474.
27. 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.

28. 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
29. Smith, W., Andrewes, C., and Laidlaw, P. (1933). A virus obtained from influenza patients. Lancet *1*, 66-68.
30. 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.
31. Stewart J.J., Watts P., and Litwin S. (2001). An algorithm for mapping positively selected members of quasispecies-type viruses. BMC Bioinformatics. *2*, 1
32. 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.
33. 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.
34. 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.
35. Yang, Z., and Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. Trends In Ecology And Evolution *15*, 496-503.
36. 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.

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.