以計算理論探討病毒疫苗設計之可行性評估

胡裕仁 [1,2]
[1] 國立中科實驗高中
[2] 國立中興大學應用數學系博士班
huyujen@gmail.com

王舜德
中國醫藥大學 MD. PhD.
wst917@gmail.com

胡裕華
獨立研究者
yuhhuahu@gmail.com

柯志斌
國立中興大學應用數學系教授
jbke@amath.nchu.edu.tw

王翰霖
國立中科實驗高中
smart850601@gmail.com

林柏毅
國立中科實驗高中
linus.only1535@yahoo.com.tw

林義傑
國立中科實驗高中
jason851124@gmail.com

溫英華
國立中科實驗高中
shadow10230@gmail.com

摘要

本研究主要利用最大概似法、動態規劃演算法及近鄰相接法來嘗試縮短生醫領域在抗體研發的時程。透過序列比較的計算方式加速找出病毒序列具有專一性的有效區段。使科學家可以減少盲目測試的實驗。我們期望找出經過電泳之後，可以判斷具有可製造抗體的最佳生物序列區段。藉由已知流感病毒的基因序列來分析現有流感病毒的演化親緣關係。嘗試由已知流感病毒疫苗來設計未知的流感病毒疫苗之建議。

關鍵詞：生物序列、動態規劃、抗體研發

Abstract

This study use maximum likelihood method and dynamic programming algorithm and neighbor-joining method to shorten the biomedical field in antibody research and development. Through the sequence comparison calculations, we speed up to identify viral sequences specific section. And thus scientists can reduce the blind to test experiments. We expect to find out that can be judged the optimal biological sequence segment and produce antibodies after electrophoresis. By influenza viruses are known gene sequence analyze the evolution of the relationship between unknown influenza virus. We tried to propose use of the known influenza virus vaccine to design unknown influenza virus vaccine.

Keywords: Biological sequences、Dynamic programming、Antibody R&D.

1. 前言

概似函數 (Likelihood Function, L(θ))[3] $L(\theta)= f(X_1,\cdots,X_n;\theta) = f(x_1;\theta)\cdot f(x_2;\theta)\cdot\ldots\cdot f(x_n;\theta)$，本函數將在參數 θ 的所有可能取值上，使這個函數最大化。這個使 $L(\theta)$ 可能性最大的 $\hat{\theta}$ 值即被稱為 θ 的最大概似估計。由於生物序列中基因呈現離散資料型態，即假設目前有三組生物基因序列 $A=\{atcatgctacgtcag\}$、$B=\{ctgatcgatgcatgc\}$、$C=\{agtctggatcgagtc\}$、$\theta = \frac{f}{n}$。字母 f 代表在不同基因字串中被觀察到相同字元的次數，n 代表各基因字串的總數。$\theta_{AB} = \frac{3}{15}$(A 和 B 序列)、$\theta_{BC} = \frac{6}{15}$(B 和 C 序列)、$\theta_{AC} = \frac{5}{15}$(A 和 C 序列)。因此可以令 $\hat{\theta}=\theta_{BC} = \frac{6}{15}$，為概似函數取得最大值，即為 $\hat{\theta}$ 的最大概似估計[2]。

只要有生命的地方，就有病毒存在；病毒可能在第一個細胞進化時就存在了[3]。雖說病毒起源於何時尚不清楚，且因病毒不會形成化石，因此沒有外部參照物來研究其進化過程，同時病毒的多樣性顯示它們的進化很可能是來自多方式的並非單一演化[4]。

分子生物學技術是目前較有效的找尋病

毒起源的方法[5]；但這些技術需要獲古代病毒 DNA 或 RNA 的樣品，但目前儲存在實驗室中最早的病毒樣品也不過百年內[6][7]。

已知第一個被發現的病毒是煙草花葉病毒，由馬丁烏斯·貝傑林克於 1899 年發現並命名[8]。病毒是一種具有細胞感染性的微粒子，它是由一個保護性的外殼包裹的一段 DNA 或者 RNA，藉由感染的機制進行自我複製[4]，但無法獨立生長和複製。同時，它也能在細胞外保持極強的生命力。病毒可以感染所有具有細胞的生命體。如今已有超過 5000 種以上類型的病毒得到鑒定[9]。

在人類基因體計畫完成後，科學家發現生物可表現的基因，都會產生其所對應的蛋白質；因此由基因體學 (genomics) 衍生了蛋白質體學 (proteomics) ，開始了蛋白質體學的應用[1]。不論是基因體或蛋白質體，科學家必須處理數量龐大的基因或蛋白質序列，並以高產能(high-throughput)的觀念來進行基因或蛋白質序列的分離或檢定。

上述過程中科學家多半是利用電腦計算分析病毒和宿主 DNA 的序列資訊，希望得到對不同病毒之間的進化關係有更好的了解，但是在生化實驗操作中如何有效利用電腦計算的討論實則不多，不過借用資訊技術確實可以有助於發現現代病毒的祖先。

2. 文獻探討

由於病毒是由核酸及蛋白質結合組成，其 DNA 有單股及雙股，同樣的 RNA 也是如此；然而不論是 DNA 或 RNA 都需經過 mRNA 的階段，只是增殖的機制卻不一樣。因此，病毒很可能是由細胞在演化過程中突變出來。

病毒依其種類之不同，有的終生不改變其穩定性，而有些則非常不安定如流行性感冒病毒。因此病毒具有經常性的突變，即原本病毒的基本分子結構發生變異產生了生物序列結構的新組合，因此生成新型病毒[14]。病毒起源的理論之一『共進化理論』指出：病毒可能進化自蛋白質和核酸複合物，與細胞同時出現在遠古地球，並且一直依賴細胞生命生存至今[6][7]。不同的病毒很可能是通過一種或多種機制在不同的時期產生[9]。

國際病毒分類委員會（The International Committee on Taxonomy of Viruses，ICTV）制定了相關病毒分類表，根據病毒的細胞生物宿主，分為「細菌病毒」、「真菌病毒」、「植物病毒」、「無脊椎動物病毒」、及「脊椎動物病毒」[10]。而病毒的組成都含有遺傳物質（RNA 或 DNA）[11]；所有的病毒也都有蛋白質形成的外殼，用來包裹和保護其中的遺傳物質。在正常情況下，病毒感染會引發免疫反應，消滅入侵的病毒。而這些免疫反應能夠通過注射抗體來產生。

3. 研究猜想

本研究提出如何處理數量龐大的基因體或蛋白質體的計算實驗，並透過「流感病毒」來測試此方法。我們希望找出可以快速且正確地在實驗中來檢定一個病毒是否具有可發展為抗體的生物序列區段。並且判斷它是否具有發展為抗體的關鍵位置之可能。而如何順利地找出此區段，正是抗體研究實驗是否可以成功的一個最關鍵步驟[1]。

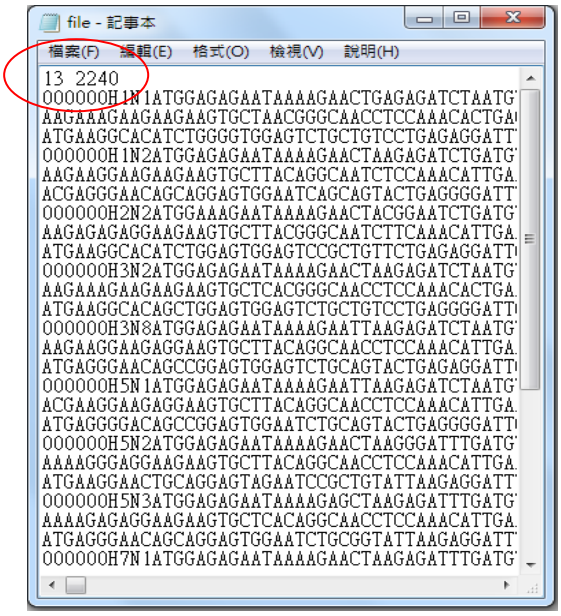
本研究嘗試找出檢測流感病毒中的核酸序列表現，為了發展快速有效地製造抗體方法，本研究乃實作以流感病毒來找出可以製作抗體的核酸區段進行研究。研究先自 NCBI 資料庫取得相關的流感病毒 mRNA 序列。

4. 實驗分析

4.1 實驗過程

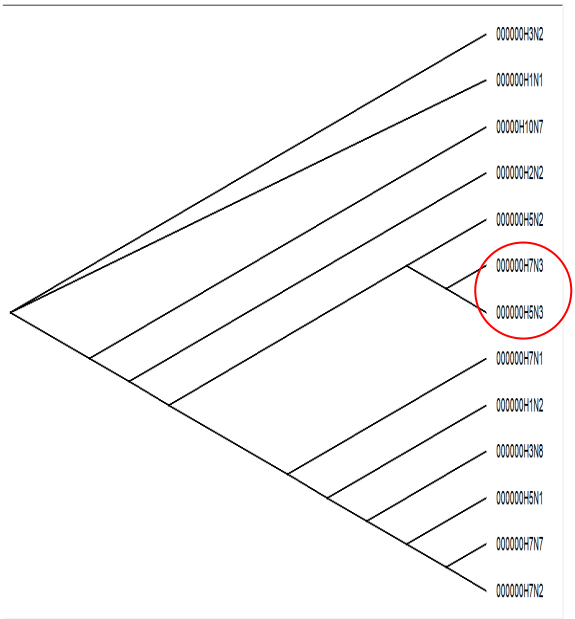
為了比對流感病毒基因序列之間的親疏關係，我們從 NCBI (National Center for Biotechnology Information)網站的資料庫中擷取了 H1N1、H1N2、H2N2、H3N2、H3N8、H5N1、H5N2、H5N3、H7N1、H7N2、H7N3、H7N7、H10N7 共 13 個流感病毒的基因序列，然後利用 PHYLIB3.69 的演化分析軟體的 dnaML[6]來計算。

在軟體 dnaml 的輸入檔中，因為格式限定，每個染色體序列的長度都要等長，但事實上每個流感的基因序列的長度都不盡相同。以本例而言最短的是 H3N8，共有 2271 個字元。最長的是 H10N7，有 2301 個字元，為了能降低到最小誤差折衷擷取了 2240 個字元並將其整理排列後貼上記事本。(請見下圖 4-1)

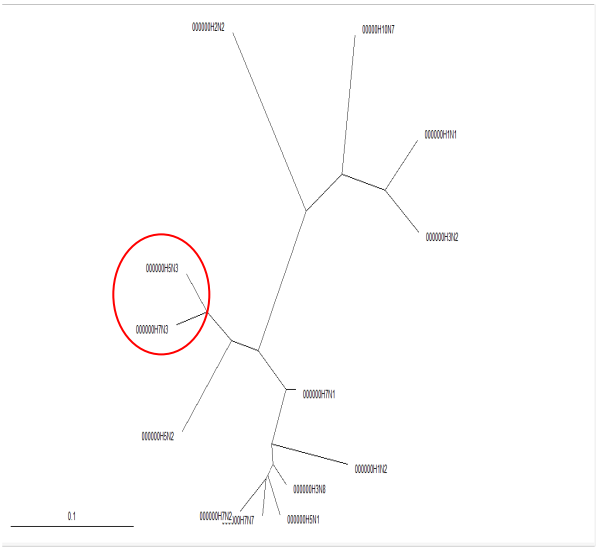


（圖 4-1:在記事本中檔案的處理）

研究使用演化分析軟體 PHYLIB 3.69 軟體中的最大概似法 dnaml 軟體來計算 13 種現有流感的核酸序列。接著我們以 PHYLIB 3.69 軟體中分別進行 dnaml 計算並得到如圖 4-2 的樹枝圖及 4-3 親緣關係圖。利用這些圖形可推測出現有流感病毒間彼此之間的演化關連性。



（圖 4-2：由最大概似法求出的流感病毒基因序列的演化樹關係-有根樹）



（圖 4-3：由最大概似法求出的流感病毒基因序列的演化樹關係-無根樹）

4.2 實驗分析

藉由 dnaml 利用最大概似法繪出的圖 4-2 及圖 4-3。我們發現了現有的流感病毒進化的相關性及演化上的相對距離。由於實驗分析了來自 NCBI 已公開的十三隻流感病毒的 mRNA 序列。由於生醫工程在進行病毒抗體研發時，多半是直接上 NCBI 進行 BLAST(basic local alignment search tool)的 PCR 引子設計並比對是否已有被研究出來的基因抗體。雖然利用 BLAST 搜索比對未知及已知的序列，可以由已知序列功能來猜測未知序列功能，且 BLAST 會提供序列比對的實驗建議環境假設、如：建議合成溫度等。

然而 BLAST 是採用 Smith-Waterman 的局部最佳化比對計算[2]，因此對於一條序列到底它的最佳抗體區段要從那裡開始找則未提供最好的建議。雖然 Smith-Waterman 的局部比對在進行序列比對檢查時速度快，但卻失去了全盤比對的精確度。透過現代資訊科技發展的快速成長，當科學計算問題與病毒抗體研發實驗兩者相比較；前者可能只是多花一點時間，但是後者卻可能如同大海撈針般只剩下細微的成功機率。且由於電腦設備的不斷的進步，計算能力的持續提昇。因此未來研者人員亦可以透過以 I-PAD 等各種具有多核心之平板電腦，以及其它類型的智慧型手持裝置來進行相關抗體的研發所需的序列計算。因此面對計算耗時及未來存有不確定實驗結果的機率，研究人員要考量如何取捨，結果應該不證自明了。

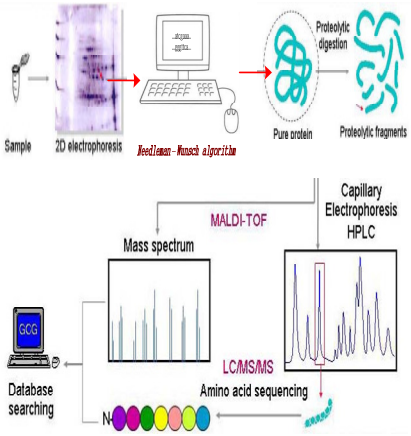
在現代的抗體研發製造流程中要進行電泳分析及純化蛋白實驗時。如果想要順利找出可當做抗體的位置序列[1][13]。就像是買樂透期望會中獎一樣無法預期一定會有好結果。雖然存在有固定實驗步驟流程卻沒有系統性的實驗操作模式。因此在實驗室進行抗體研發製造的實驗過程就如同數學中的機率問題：運氣好的科學家一次實驗就成功，運氣不好的科學家可能要重複好多年仍然是難以成功。

5. 實驗討論

生醫資訊研究若單獨使用 BLAST，可以讓研究者在其中尋找與其感興趣的序列或類似的序列。但是在進行抗體研發卻又面臨一個高度不確定的機率問題。

因此本研究提出在嘗試對未知病毒基因電泳分析之後，再針對現有病毒及抗體序列再以 Needleman -Wunsch 動態規劃演算法[15]進行序列比對計算，找尋最佳實驗起始抗體區段，並以 NEUROD4 來分析，相關流程、如圖 5-1 所示，再以鄰近相接法來適合的抗體序列，以降低抗體研發的不確定性。

圖 5-1：由研究所提方法來研發抗體流程圖



因為 Needleman-Wunsch 動態規劃演算法已被證明，可以進行任兩序列的完整比對[15]。因此本研究先將研究出已知流感的病毒序列和未研發出的流感病毒序列以 Needleman-Wunsch 動態規劃演算法逐條做完整的序列比對計算。

透過 NCBI 中建置了 Needleman-Wunsch 動態規劃演算法的 Needle 計算程式軟體，我們使用全序列計算比對功能。因為生物學上定義，當蛋白質中超過 25%的胺基酸序列相同或 DNA 中超過 75%的核酸序列相同，幾乎可以確定蛋白質或 DNA 序列具有同質性，可作為親緣判定的參考[16]。同時推測出未知流感病毒疫苗應該可由何種已知流感病毒疫苗來研發出來的可能。

上述推論成立的理由是因為基因組裡常見 DNA 突變(稱為差錯型突變)[2]，最常出現是在單核苷酸的替換。科學家稱為「單核苷酸多型性」(single nucleotide polymorphism)。Jukes-Cantor 距離(Jukes-Cantor distance)計算假設是四種核苷酸對核苷酸的替換率是一樣的(突變發生的機率皆相等)。該模型亦是以任兩個序列之間的核苷酸替換的數量進行的最大概似估計(MLE)計算。再以 Jukes 和 Cantor 模型分枝距離公式 $d=-\frac{3}{4}\log_e(1-\frac{4}{3}p)$ 計算相對距離。本研究以 Jukes 和 Cantor 分枝樹距離公式[2][17]計算了十三種的核苷酸序列相似性矩陣(結果如式 1)。再代入鄰近相接法進行二次分析。

$$\begin{bmatrix} & D1 & D2 & \cdots & D13 \\ D1 & 0 & \cdots & \cdots & \vdots \\ D2 & \vdots & 0 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D13 & \cdots & \cdots & \cdots & 0 \end{bmatrix} \text{...式1}$$

5.1 Needleman-Wunsch 動態規劃演算法

若 A、B 分別為二序列，即

STEP1：$A=(a_1,a_2,...,a_n)$、$B=(b_1,b_2,...,b_m)$，由 A、B 二序列構成的二維表如下：

表 5-1：A、B 序列的二維表

	a_0	a_1	a_2	$\cdots$	a_n
b_0	$s(0,0)$	$s(1,0)$	$s(2,0)$	$\cdots$	$s(n,0)$
b_1	$s(0,1)$	$s(1,1)$	$s(2,1)$	$\cdots$	$s(n,1)$
b_2	$s(0,2)$	$s(1,2)$	$s(2,2)$	$\cdots$	$s(n,2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
b_m	$s(0,m)$	$s(1,m)$	$s(2,m)$	$\cdots$	$s(n,m)$

STEP2：計算表 5-1 中的每一個元素 $s(i,j)=\max\{s(i-1,j-1)+s(a_i,b_j),s(i-1,j)-d,s(i,j-1)-d\}$ ---(式 2.1)，其中 $s(0,0)=0$、$s(i,0)=-i\cdot d$、$s(0,j)=-j\cdot d$；d：表示序列的連續空位分

數，$\{x,y,z\}\in\mathbb{Z}$，"–"表示序列比對時的空格，$s(a_i,b_j)=\begin{cases}x & ,a_i=b_j\\ y & ,a_i\neq b_j\\ z & ,\text{"–"}\end{cases}$。

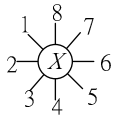
STEP3：由 STEP2 最後一次算出的$s(n,m)$來回溯計算得出最優比對路徑，並記下相對應的位置若回溯的位置是水平向左、則在縱列插空格並記為$(a_i,-)$，反之若回溯的位置是水平向上、則在橫列插空格並記為$(-,b_i)$。

STEP4：統計由 STEP3 的所有回溯記錄計算出兩序列的最佳完整序列比對結果。

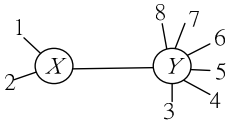
5.2 鄰近相接法（Neighbor-Joining method）

本法是一種基於距離並可以快速得到進化樹的聚類演算法，適合處理大量資料數據，演算法如下：

STEP1：計算 N 條序列的兩兩相對距離D_{ij}，並取得距離矩陣。



(a)原始星形拓樸



(b)合併節點的星形拓樸

STEP2：將 N 條序列視為 N 個節點(即:星形拓樸結構、如圖(a))。

STEP3：任選二個節點a、b，計算任二個節點合併之後的總長L_{ab}；則星形結構全長

$$S_0=\sum_{i=1}^{N}L_iX=\frac{1}{N-1}\sum_{i<j}D_{ij}\ldots\text{式 2}\ ,$$

其中X是星形結構中心唯一的內部節點，式 2 最右邊是求和時每條邊被計算了 N-1 次。若 1、2 節點合併可得圖(b)，此時節點變為兩個、分別為 X、Y，並重算各節點枝長：L_{XY}

$$\frac{1}{2(N-2)}\left(\sum_{k=3}^{N}(D_{1k}+D_{2k})-(N-2)(L_{1X}+L_{2X})-2\sum_{k=3}^{N}L_{iY}\right)\ldots\text{式3}\ ,$$

其中

$$L_{1X}+L_{2X}=D_{12}\ldots\text{式 4}\ ,$$

$$\sum_{k=3}^{N}L_{iY}=\frac{1}{N-3}\sum_{3\le i<j}D_{ij}\ldots\text{式 5}$$ 且

$$S_{12}=L_{XY}+(L_{1x}+L_{2x})+\sum_{k=3}^{N}L_{iY}\ldots\text{式 6}\ ,$$

由式 2 至式 6 得出$S_{12}=$

$$\frac{1}{2(N-2)}\sum_{k=3}^{N}(D_{1k}+D_{2k})+\frac{1}{2}D_{12}+\frac{1}{N-3}\sum_{3\le i<j}D_{ij}\ldots\text{式7}\ .$$

STEP4：比較由 STEP3 得到的各個節點總枝長，其中總枝長度最小的稱為鄰居，並且予以合併節點形成圖(b) 拓樸結構。此時

$$L_{1X}=\frac{(D_{12}+D_{1Z}-D_{2Z})}{2}\ldots\text{式8}\ 、$$

$$L_{2X}=\frac{(D_{12}+D_{2Z}-D_{1Z})}{2}\ldots\text{式9}\ ,$$

其中

$$D_{1Z}=\sum_{i=3}^{N}D_{1i}/(N-2)\ldots\text{式10}\ 、$$

$$D_{2Z}=\sum_{i=3}^{N}D_{2i}/(N-2)\ldots\text{式11}\ 。$$

STEP5：合併節點 1,2 為新節點，計算新節點至各節點的距離，即

$$D_{1-2,j}=(D_{1,j}+D_{2,j})/2\ldots\text{式12}\ 。$$

STEP6：重複 STEP3 至 STEP5 直到內部節點為 N-3 個，並得出一組無根樹。

5.3 實驗結果

本研究由傳統的抗體研發製造流程加入了最大概似估計及序列比對的 Needleman-Wunsch 動態規劃計算，並使用軟體來針對流感病毒序列進行序列的全局比對。另外在鄰近相接法的模擬部分我們利用華盛頓大學 Joseph Felsenstein 教授開發的 PHYLIP-3.69 版進行比較分析[2]。

此外計算十三種流感的 mRNA 的相似距離矩陣得到式 1 結果。我們發現進化樹(鄰接法計算)結果如圖 4-2 及圖 4-3 所示。在 H7N3 和 H5N3 兩者在演化的相似度比以 H7N3 和 H7N1 及 H7N2 還近，因此如果我們進行 H7N3 的新藥研發，若能先以 H5N3 的已知疫苗結果來設計可能為 H7N3 的疫苗區段應可加速新疫苗研發成功的機率。並充當實驗分析的初始序列區段，順利找出 H7N3 的疫苗序列的抗體可能關鍵位置區段實驗，期望協助生醫工程快速找到具有高專一性的目標抗體序列區段。

6. 結論與未來展望

本研究提出的實驗猜想方法，在實驗中進

行了數值比對計算，並且得到了一些和一般純粹進行生物實驗的生醫學者不同的想法。例如：圖 4-2 及 4-3 的 H7N3 和 H5N3 兩類病毒結構由此次分析計算出的結果是最接近的。可是在生醫科學家的直覺卻不那麼想，因為此兩類病毒從分子結構組成的名稱就覺得關係比較遠。

因此本研究的結論剛好在未來可以進一步地提供生醫科學家在實驗中先以電腦計算。他們可以先檢驗研究進行的推測，再進行生物實驗分析以加速找出具有專一性的抗體。以提高實驗成功的機率。

致謝

本研究承蒙行政院國家科學委員會經費補助（NSC 101-2514-S-796-001）。

參考文獻

[1] 莊榮輝、吳裕仁，"***蛋白質體學與單株抗體應用***"，蛋白質體學，取自台大醫學院生化與分生研究所 http://juang.bst.ntu.edu.tw/Protein/proteomics&mab.htm，2003。

[2] 楊晶、胡剛、王奎、沈世鎰，***生物計算***，科學出版社，2010。

[3] Iyer LM, Balaji S, Koonin EV, Aravind L. "***Evolutionary genomics of nucleo-cytoplasmic large DNA viruses.***" Virus Res., 117 (1): 156–84, 2006.

[4] Villarreal, Luis P. "***Viruses and the Evolution of Life***", ASM Press, 2005.

[5] Liu Y, Nickle DC, Shriner D, et al., "***Molecular clock-like evolution of human immunodeficiency virus type 1***" Virology. 10;329(1):101–8, 2004.

[6] Shors, Teri. ***"Understanding Viruses."*** Jones and Bartlett Publishers., 2008.

[7] Collier, Leslie; Balows, Albert; ***"Sussman, Max, Topley and Wilson's Microbiology and Microbial Infections ninth edition."***, Volume 1, Virology, volume editors: Mahy, Brian and Collier, Leslie. Arnold., 1998.

[8] Norrby E., ***"Nobel Prizes and the emerging virus concept***" Arch. Virol., 153 (6): 1109–23, 2008.

[9] Dimmock, N.J; Easton, Andrew J; Leppard, Keith, "***Introduction to Modern Virology sixth edition,***" Blackwell Publishing, 2007.

[10] M. H. V. Van Regenmortel, C. M. Fauquet, D. H. L. Bishop, "***Classification and Nomenclature of Viruses : Seventh Report of the International Committee on Taxonomy of Viruses***." Academic Press, 2000.

[11] Carlson GA, Hsiao K, Oesch B, Westaway D, Prusiner SB., "***Genetics of prion infections.***", 7. 1991.

[12] M. Blalock., "***A beginner's guide to microarrays***." Kluwer Academic Publishers, Norwell, Massachusetts USA, 2003.

[13] Wu YJ, Chen HM*, Wu DJ*, Wu JS*, Chu RM, Juang RH, "***Preparation of monoclonal antibody bank against whole water-soluble proteins from rapid-growing bamboo shoots.***" Proteomics 6(22): 5898~5902, 2006.

[14] Leppard, Keith; Nigel Dimmock; Easton, Andrew., "***Introduction to Modern Virology***." Blackwell Publishing Limited, 2007.

[15] Needleman, Saul B., Wunsch, Christian D., "***A general method applicable to the search for similarities in the amino acid sequence of two proteins.***" Journal of Molecular Biology 48 (3): 443–53.,1970.

[16] P. W. Lisette, P. David, "***Noninvasive Genetic Sampling Tools for Wildlife Biologists: A Review of Applications and Recommendations For Accurate Aata Collection***", Journal of Wildl. Manage., 1419-1433, vol 69, 2005.

[17] Jukes TH & Cantor CR, "***Evolution of protein molecules.***" In Munro HN, editor, Mammalian Protein Metabolism, pp. 21-132, Academic Press, New York, 1969.

[18] Saitou N, Nei M. "***The neighbor-joining method: a new method for reconstructing phylogenetic trees.***" Molecular Biology and Evolution, volume 4, issue 4, pp. 406-425, 1987.