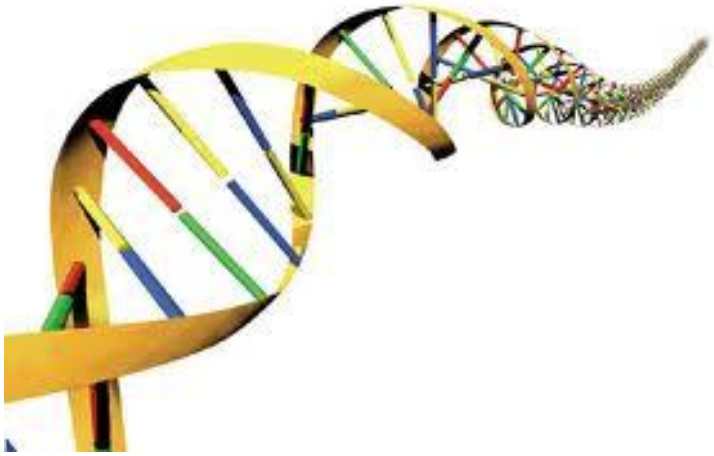
第十一屆旺宏科學獎

成果報告書



參賽編號： SA11-175

作品名稱：病毒基因解碼

——利用短序列分析工具進行多基因體病毒之探討

姓名：侯辰翔

關 鍵 詞：多基因體病毒、重組演化、短序列

摘要

隨著生物的基因體資訊量快速的成長，龐大的基因資料便依賴快速的分析與計算，生物資訊學也應運而生，利用電腦快速的從大量數據中提取有用的資訊，進行系統性的分析，或是對於實驗的設計進行輔助性的預測，這些都已成為生命科學中非常重要的研究方法。本研究的目的便是利用現有的資訊分析工具，探討多基因體病毒(Multi-component virus)之基因重組的現象。我們收集了所有多基因病毒的代表種(Type Species)，並著眼於基因體之間相似的短序列(代稱 Motif)，進行分析。由於這些共同的短片段可能是病毒快速演化中殘留之共同遺傳片段，而這些短片段又難以用傳統的序列比對工具尋找，因此我們利用短序列分析程式 MEME 來尋找多基因體病毒中，不同基因體(component)之間共有的短序列，並且計算這些片段占該段基因體的比例，以圖示的方式表現不同基因體之間的相關程度，然後再將病毒以不同分類進行歸納比較，由此方法觀察可能的演化現象。

壹、研究動機

無論是 DNA 病毒、RNA 病毒，只要是成功的感染宿主後，很快就會開始進行複製，病毒利用宿主細胞內原本就有的材料，並且借助宿主細胞的複製能力來位自己製造新的遺傳物質與蛋白質，短則數分鐘，長則數小時，就能產生數量可觀的後代。比起二三十年才繁衍下一帶的人類而言，病毒的世代交替的速度非常快，所以病毒演化的速度比人類要快速許多，是研究演化機制非常好的觀察對象。

因為生物的演化(包括病毒演化)，主要是由於遺傳物質(也就是 DNA 或是 RNA)產生變化所致，例如在遺傳物質複製的過程中發生錯誤，或是遺傳物質遭到破壞(例如輻射線照射)而發生變異。遺傳物質變化的規模可小可大，小規模的變化可能僅止於 DNA 或是 RNA 上一個基本單位(DNA 或是 RNA 的鹼基)的變化，稱為「單點突變」；大規模的改變，則可能導致遺傳物質片段的「重組(Recombination)」，原因有可能是插入了外來的新片段、遺失原本的片段，或是本身數個片段的重新組合。

如果只是發生小規模的單點突變，那麼產生的新病毒的 DNA 或是 RNA 一時之間不會有太大的改變，但若發生重組而導致的整個遺傳物質序列的損失(Defection)、交換(Exchange)、重排(Rearrangement)、或是外來序列插入(Insertion)的話，對於原本的基因底的改變就會比單點突變來得大。生物體在演化的歷程上，基因物質複製的過程中會不斷的發生重組與突變，有些會造成下一帶性狀的改變，有時這些改變會造成生物功能的缺失而導致疾病，反之也會有機會付與後代有利的生物功能，並且使他在物競天擇的演化過程中被留下，可以想見在許久之後，原本來自祖先且應該很相似的序列，因為不斷重組與突變，而變成了不連續、前後順序對調，甚至是方向顛倒的短片段，因此病毒的後代就會跟原本的祖先差異很大，甚至完全認不出來了。像這樣的改變可能需要數萬年的世代才會發生，但是在複製極為快速的病毒上，便有機會進行觀察。

貳、研究目的

了解了對於基因突變或重組可能帶來的影響之後，我們便提出一個問題：如果有兩條遺傳物質看起來十分不相像的病毒，然而他們在演化上確有其共通性，那麼他們是真的完全不相干的病毒嗎？或者他們的遺傳物質中其實存在著相同的演化移機，只是因為經過重組而喪失了相似性呢？ 我們在生物資料庫中發現有基因體恤列資料的病毒是數量非常多的，因此本研究將專注於分析具有兩條基因體以上的多基因體病毒上，因為多基因體病毒有幾個特性：

1. 多基因體病毒具有數條獨立的遺傳基因體，這些基因體在病毒感染的時候，必須同時進入一個相同的細胞，並且會在同一個細胞內進行複製，因此有較高的機會進行序列的重組。換句話說，多基因體病毒的多條遺傳基因體在演化上是有關聯的。
2. 大部分的多基因體病毒，如果使用序列比對(Sequence Alignment)的方式去分析其所屬的遺傳基因體，其不同的遺傳基因體看起來都很不相似，在生物學上也經常被認為是彼此相異的序列。

取得了多基因體病毒的列資料之後，我們便提出一個假設，多基因體病毒的不同遺傳基因體看起來之所以這麼不同，或許是因為在演化過程中累積的很多次的重新排列，導致原本可能看起來類似的序列，並成的許多順序顛倒、分散、易位的小片段，由於這些小片段已經知去的位置上的連續性，因此在序列比對法(Alignment)中可能無法顯示其相關性，因此本研究的目的，便是尋找同一隻多基因體病毒內，不同遺傳基因體之間是否存在相同的短小片段，**如果有這些小片段存在，就能猜測這兩個基因體之間或許發生過重組，甚至擁有共同的祖先。這些短片段的數量越多，就表示這兩個基因體在演化上的關係可能越密切。**

並且試著觀察不同分類的多基因體病毒，其不同遺傳基因體之間的相關性是否具有一致性。

參、研究設備及器材

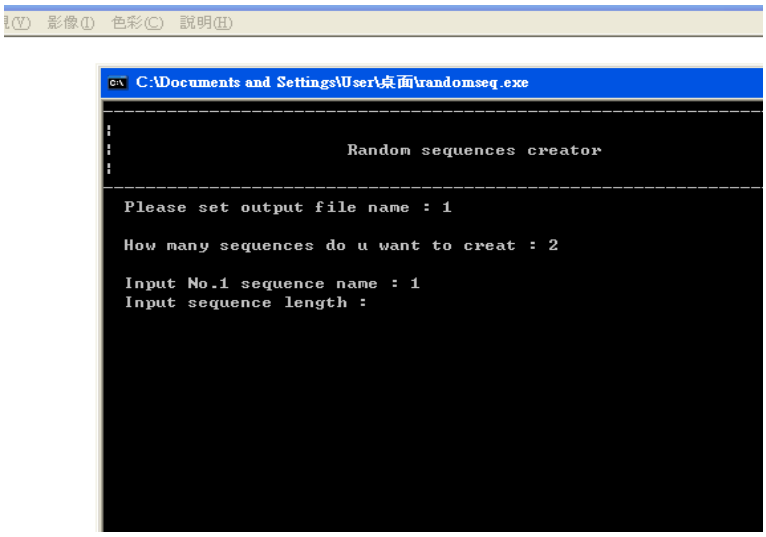
一、電腦

二、隨機序列製造器(Random sequence creator)

此程式可以隨意製造出以 A、T、C、G 四個基本字元所排列之文字列，且可以自由選擇各個序列的條數、長度，用以製造隨機之病毒(隨機樣本)序列並進行統計分析，如圖二所示。

註：隨機樣本舉例說明

Nodamura virus 有兩條基因體，分別為 3024 個鹼基對、1336 個鹼基對所組成。則 Random sequence creator 可以製造出兩條基因體，分別為 3024 個鹼基對、1336 個鹼基對，但是鹼基對的 A、T、C、G 則是亂數排列的。此一動作所得到的檔案，**即為一個條數、長度都相同的隨機樣本**。

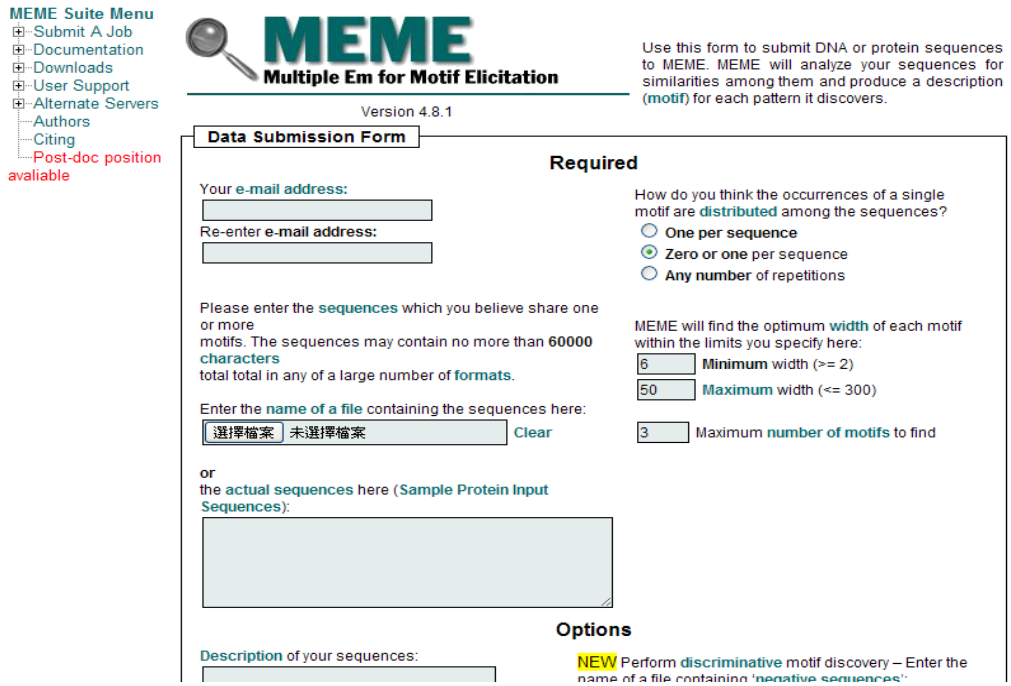


(圖二) Random sequence creator 的執行視窗畫面

三、**MEME suite**　(Multiple Em for Motif Elicitation)

(http://meme.nbcr.net/meme/intro.html)

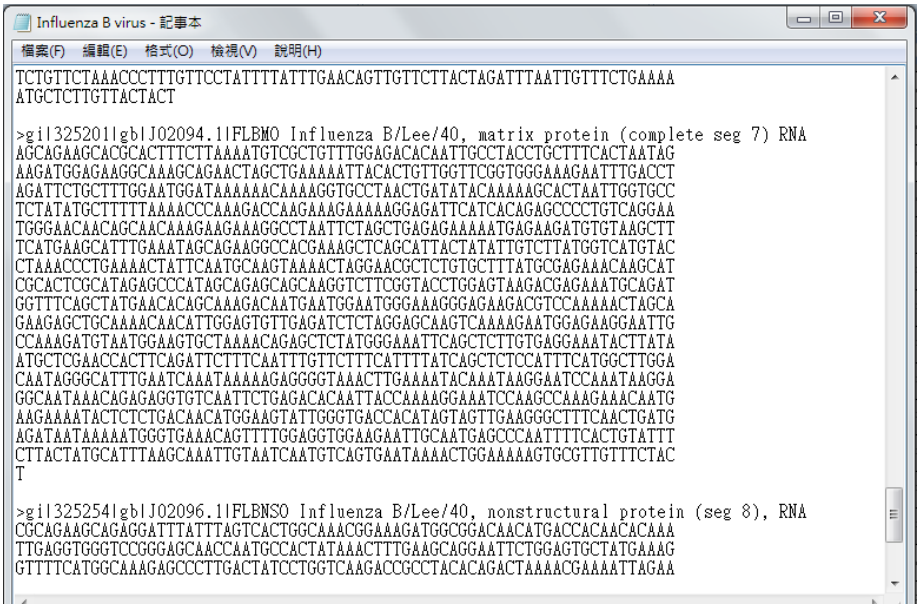
短序列分析程式 MEME，是用於從基因序列或是蛋白質序列中搜索相似短序列的工具，這些短序列很有可能具有相同的功能，或是在演化中具有共同的來源，其資料輸入畫面如圖三所示。



(圖三) MEME 的輸入視窗

MEME 的輸入為 FASTA 格式的序列文件，FASTA 格式的文件是現在生物科學界用來記錄 DNA、RNA 和蛋白質序列最常使用的標準格式，在 FASTA 檔案中，每一條序列使用〝>〞標注序列的開始，在〝>〞符號後的文字，都用來對這一條序列作註解，例如：序列名稱以及說明性的文字。而下一行開始，為序列的內容，直到下一個注釋符號為止。

其範例如圖四所示。



(圖四)FASTA 檔案的一個範例

肆、研究過程及方法

一、實驗步驟

(一) 尋找病毒

為了尋找用來分析病毒的資訊，我們首先尋找了 NCBI(National Center of Biotechnology Information)的網站，NCBI 是美國全國生物技術信息中心，這個網站收集了各種分子生物資料(如圖五所示)，是現今生物基因體資料最豐富的資料庫。我們首先在 NCBI 上找尋哪些是屬於多基因體病毒，並且下載其序列資料為 FASTA 檔案。 因為全部的病毒資料數量龐大，而且同種的病毒也收錄了不同地域性的亞種，以至於資料的重複性高，因此我們決定將分析的範圍所小至分類中各個屬的代表種(Type Species)，根據 Virus Taxonomy(參考資料)中的 Order of Presentation of Virus Taxonomic Descriptions 章節中，找尋到各屬之代表種。並將其基因序列下載下來。

(圖五) NCBI 的網站(http://www.ncbi.nlm.nih.gov)

(二)序列比對演算法(Alignment)與 MEME 短序列分析工具

一般分析序列的相似性時，最常使用的是序列比對演算法(Alignment)，大概的演算過程，是將兩個或是兩個以上的生物序列排序在一起(可能是 DNA 序列、序列或是蛋白質序列)，並且按照順序逐一的去標明其相似的位置，例如 DNA 中，兩條或是多條序列的相同位置上若都是 A，便標名為相似，在序列比對演算法中，也允許排列位置上有少許的調整彈性，也就是在排序時可以插入間隔

(gap，通常以"-"表示)，但是基本上是沿著續料的順序依次完成排列與標示。但是當比對的序列出現可能的片斷重組時，由於序列的連續性遭到破壞，因此序列比對演算法就可能出現低估其關連性的結果，如下面的兩條 DNA：

第一條：AAAAAAAAAAATTTTTTTTTTTTTCCCCCCCCCCCC

第二條：AAAAAAAAAAACCCCCCCCCCCCTTTTTTTTTTTTT

第一條與第二條序列的差別，可能僅僅是後半部"連續的 T"與"連續的 C"發生一次的片段互換，但是如果使用序列比對演算法，很可能會得到這樣的結果：

AAAAAAAAAAATTTTTTTTTTTTCCCCCCCCCCCC
OOOOOOOOOO OOOOOOOOOOO
AAAAAAAAAAA----------CCCCCCCCCCCCTTTTTTTTTTTT

面對可能發生重組的序列，使用序列比對分析可能遺失掉重要的資訊，因此我們決定使用專門尋找短序列的程式 MEME 來分析，MEME 是一個應用 EM ALGORITHM 而設計的程式，它可以從多條序列中尋找短序列，不僅可以尋找「每一條序列上都有的短序列」，甚至是尋找「只有在某幾條序列上才有的短序列」。由於 MEME 並非依賴序列比對演算法，因此可以尋找順序、位置甚至方向顛倒的短序列，十分符合我們的目標。

(三)篩選

分析工具 MEME 可以找到可能的短序列，這些短序列根據找尋的結果排列為不同的名次，並且列出讓使用者挑選，然而我們到底要挑選哪些短序列分析才可信？為了決定哪些才是具有統計意義的片段。我們必須設計一個篩選的機制，我們的想法是，以隨機(Random)的方式創造出許多仿造多基因體病毒之基因體長度的亂數序列(也就是 A、T、C、G 隨意的排列)，然後讓 MEME 去尋找短序列，由於亂數序列中沒有刻意要保留或是刪去任何短序列。因此我們便能知道，在這

樣的亂數的環境中，可能存在的短序列會多麼相似。表示若同樣相似程度的短序列可在亂數環境中就可能出現的，那麼短序列就不具生物學上的意義。

1.隨機樣本

既是多基因體病毒也是代表種的病毒約有五十多隻，我們取其長度大小前10%的病毒，即鹼基對序列總長前5短的病毒5隻為範本。對每一隻都用 Random sequence creator 仿造出100隨機樣本(共5*100=500個隨機樣本)。此處，取鹼基對總長度最短的理由是因為，在最小的亂數環境中，最難存相似的短片段，因此以最小規模的亂數懷環境來決定標準，是較為嚴格的方式；反之，如果有很長的鹼基對總數，則在使用 Random sequence creator 時，就會有更大的可能包含全部的狀況。

2.選取標準

製造出來的隨機樣本的功用在於設定出"選取標準"。因此我們使用一種相當直觀的評估方式——漢明距離(Hamming distance)來做為我們計算的標準。漢明距離的計算方式是單純計算兩個**短序列**之間單元的差異數量。例：

短序列一：AGGTTTTATATCCGAG

XX

短序列二：AAATTTTATATCCGAG

短序列一與短序列二，只有在第二與第三個位置上有所不同，因此其漢明距離就是2。我們計算時則**採用相同數量**，在實際病毒所找到的每一群短序列，都計算一個平均的 X**(相同數量)**，這個數據反應的是這一群短序利彼此的相似程度，X 越大，相似性越高，越小則越不相似。

如果在已知的多基因體病毒中找到的短序列，其 X 值與亂數環境中的的平均值相近，代表這一組短序列並不可信，因為在亂數環境中也找得到同樣相似程度的狀況。相反的，如果在已知的多基因體病毒中找到的短序列，其 X 遠大於亂數環境的平均值，就代表這樣相近的短序列，幾乎不可能是亂數產生的結果，因此就具有很高的獨特性。我們選擇了一個較為嚴苛的方式，即是在病毒中找到

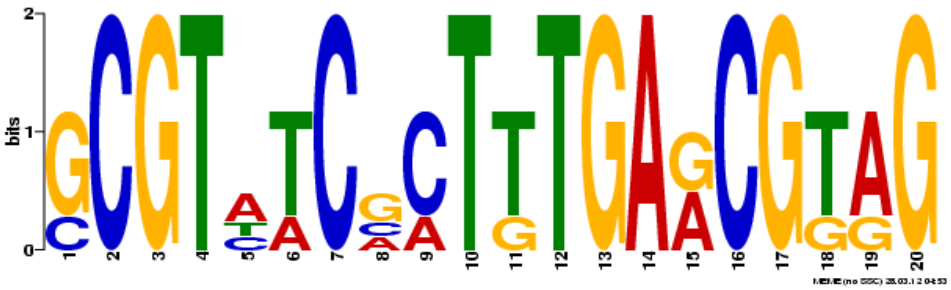
的短序列組，其平均的 X 值要比隨機製造 X 值得 5%都還要大才會採納(Hamming distance 要比隨機製造的漢明距離的 95%還要小)。

總共有 30(個短序列/每一隨機樣本)*500 個隨機樣本，前 5%就是 750 名。因此病毒中找到之短序列組，其 X 值要比第 750 名之 X 值還要大才行。

決定了"選取標準"，就可以將實際病毒以 MEME 分析出來的短序列，都用"選取標準"篩選過，留下合乎標準的短序列。

3.MEME 分析，舉例說明

例：由 MEME 找出的其中一段短序列為：



而這個相似短序列在 4 個位置有找到

Sites

Click on any row to highlight sequence in all motifs.

Name	Strand	Start	*p*-value	Sites
gi|14794973|gb|AF389463.1|	+	2446	2.54e-11	TCTGCTTGCG GCGTTTCACTTTGAGCGTAG ATAACAACTT
gi|14794973|gb|AF389463.1|	+	2615	4.58e-11	GGAGTATCCT GCGTATCCATTTGAGCGTAG AGACGGCACA
gi|14794973|gb|AF389463.1|	-	957	6.18e-11	CTCGTTTTGC GCGTAACGCTGTGAACGTAG CTGACACAAA
gi|14794977|gb|AF389464.1|	-	1458	1.99e-10	TGCAAATATA CCGTCTCGCTTTGAACGGGG GCCTGCCTAT

此一短序列之 X 值為 11(4 段短序列的 20 個鹼基對中有 11 個完全對應到。)

Name 為病毒基因體的 FASTA 檔案代稱

Start 為起始的鹼基對。例：2446 為這段短序列從第 2446 個鹼基對開始

p-value 數值愈小，代表分析的結果越準確

Sites 則將短序列直接表示出來

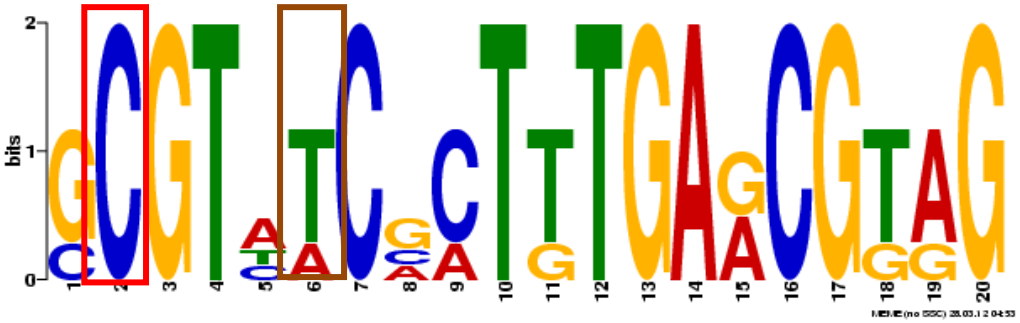
英文字母大小則代表：

GCGTTTCACTTTGAGCGTAG

GCGTATCCATTTGAGCGTAG

GCGTAACGCTGTGAACGTAG

CCGTCTCGCTTTGAACGGGG



在互相對應後，如果 4 條全部一樣(例：紅框內的 C)，則為最大的字母；而如果有不一樣(例：咖啡色框內的 A、T)，則由大小差異之英文字母組成，則這一欄就不完全一樣，我們就不計算在 X 值內，這個短序列的 X 值為 11。

我們一一將每隻亂數製造器(Random sequence creator)的病毒，都輸入 MEME 分析，且以這種方式計算出來所有的 X 值，並統計成表一，由表一所得的結果以圖六表示，並據圖六決定"選取標準"。

(四)分類

1.算出比例值

將全部的實際病毒都以 MEME 分析過，將合乎標準的短序列留下。我們決定計算一個病毒內每一個基因體相似片段之總短序列的全長佔該基因體長度的比例值。

例：

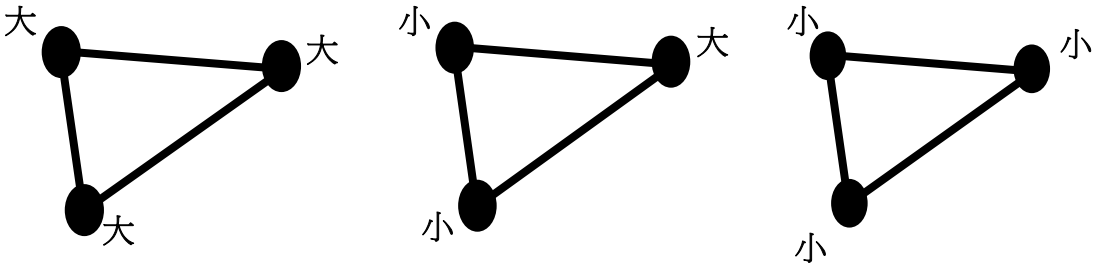
Atkinsonella hypoxylon virus 的 RNA2 鹼基對總長為 2135，與其他基因體共用的 30 個短序列(共有 30*20=600 個鹼基對)，比例值為 600/2135*100%=28.10%

比例值越高代表基因體與其他基因體重組程度越高

比例值越低代表基因體與其他基因體重組程度越低

2.歸類

分成 3 種，圖中每一個點代表一個基因體，”大”、”小”則是代表比例值(20%)為界。



A **B** **C**

A 型：代表基因體之間可能有較多的相似短序列，彼此之間的相關性很大，可能具有共同的演化來源。

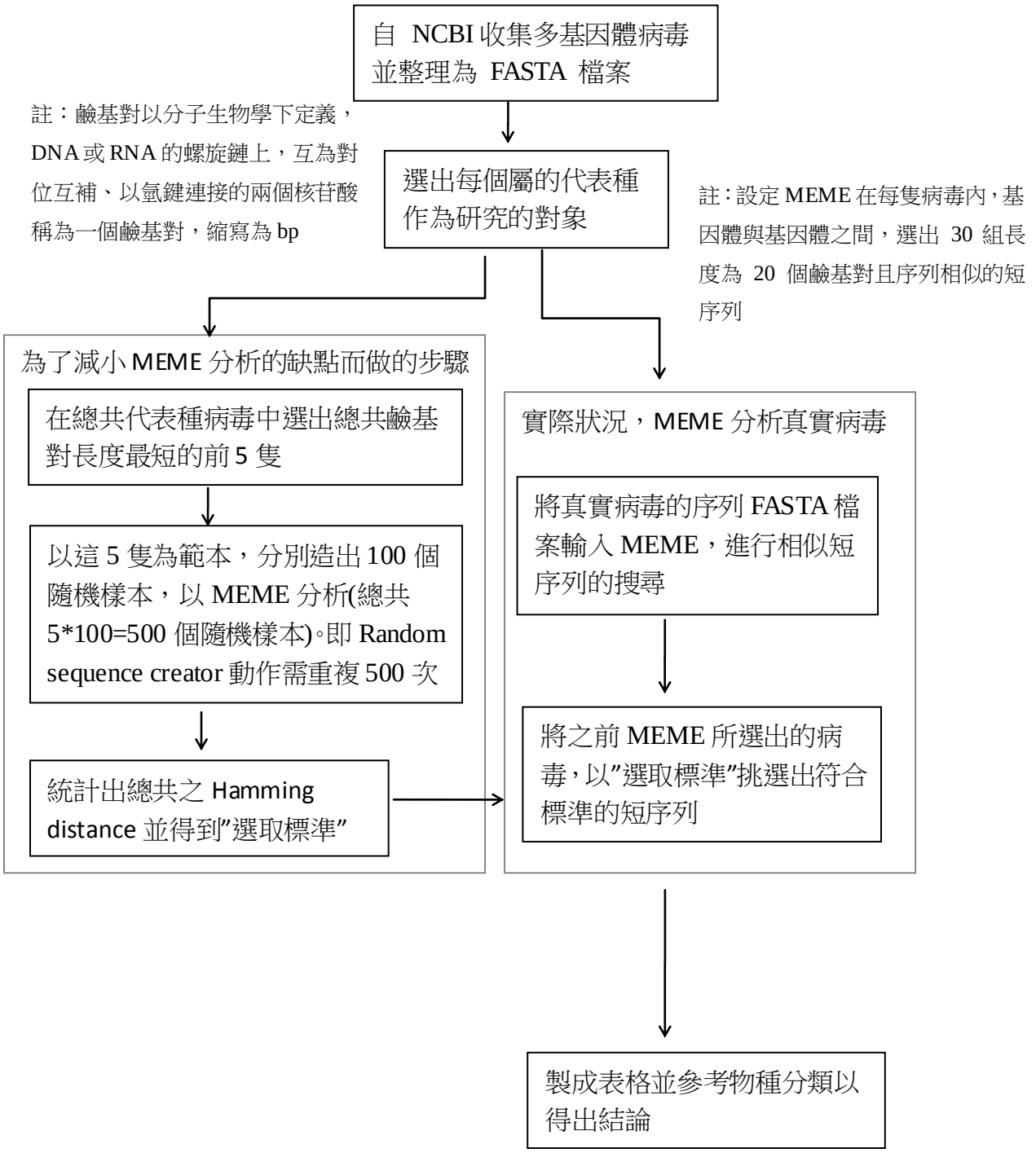
B 型：並不是所有的基因體，其總短序列的鹼基對占該基因體總長的比例都高。推測只有某些基因體可能是在演化上扮演較重要的角色。

C 型：幾乎所有的比例都是偏低。代表基因體之間的重組不明顯，互相沒有太大的影響以及關連。

(五)比例排序

將每一隻病毒每個基因體的比例值算出，並以分類完的表作為依據討論，針對以上三種型態，選出其中一隻代表性病毒，進行更細部的分析。

二、研究方法流程圖：


自 NCBI 收集多基因體病毒並整理為 FASTA 檔案
註：鹼基對以分子生物學下定義，DNA 或 RNA 的螺旋鏈上，互為對位互補、以氫鍵連接的兩個核苷酸稱為一個鹼基對，縮寫為 bp
選出每個屬的代表種作為研究的對象
註：設定 MEME 在每隻病毒內，基因體與基因體之間，選出 30 組長度為 20 個鹼基對且序列相似的短序列
為了減小 MEME 分析的缺點而做的步驟
在總共代表種病毒中選出總共鹼基對長度最短的前 5 隻
以這 5 隻為範本，分別造出 100 個隨機樣本，以 MEME 分析(總共 5*100=500 個隨機樣本)。即 Random sequence creator 動作需重複 500 次
統計出總共之 Hamming distance 並得到"選取標準"
實際狀況，MEME 分析真實病毒
將真實病毒的序列 FASTA 檔案輸入 MEME，進行相似短序列的搜尋
將之前 MEME 所選出的病毒，以"選取標準"挑選出符合標準的短序列
製成表格並參考物種分類以得出結論

伍、研究結果

找出既是多基因體也是代表種的病毒，選出前五短基因體長度最小的前五隻病毒，分別為

Bean golden yellow mosaic virus

Nodamura virus

Ourmia melon virus

Striped Jack nervous necrosis virus

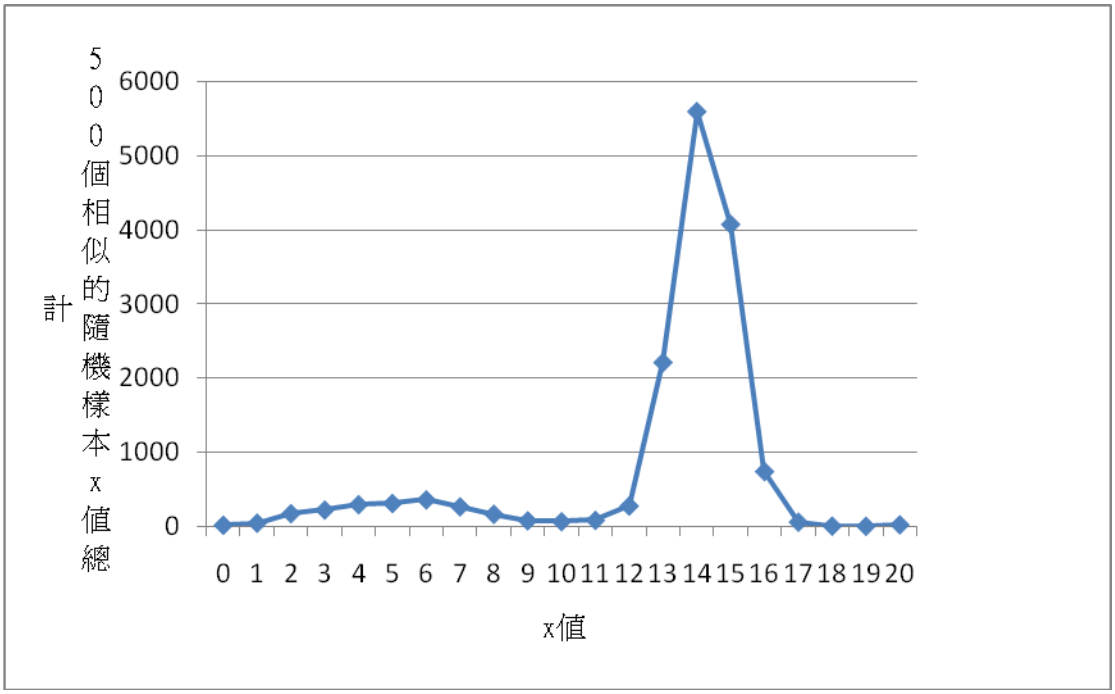
White clover cryptic virus 1

以這 5 隻為範本，利用 Random sequence creator 分別造出 100 個相似的隨機樣本，並利用 MEME 進行分析。

表一：500 隨機樣本 X 值總計，將 X 值為 0、1、……、20 時，分別加總，繪成圖六的折線圖。

病毒名稱 / x 值	*Bean golden yellow mosaic virus*	*Nodamura virus*	*Ourmia melon virus*	*Striped Jack nervous necrosis virus*	*White clover cryptic virus 1*	500 相似的隨機樣本 x 值總計
0	1	3	7	0	3	14
1	2	7	14	10	9	42
2	37	30	45	32	26	170
3	47	39	45	42	45	218
4	66	62	56	56	52	292
5	67	69	75	55	41	307

6	83	72	65	74	62	356
7	50	60	60	50	42	262
8	29	47	37	23	26	162
9	20	13	20	11	10	74
10	16	14	14	9	8	61
11	24	10	18	12	16	80
12	41	61	49	52	70	273
13	321	508	382	432	566	2209
14	1000	1086	1107	1129	1271	5593
15	986	747	836	861	641	4071
16	194	149	151	140	104	738
17	14	8	14	12	8	56
18	0	0	2	0	0	2
19	0	0	0	0	0	0
20	2	15	3	0	0	20



(圖六) 500 個相似的隨機樣本 X 值以及對應的總數目之折線圖

總共有 30(個短序列/每一隨機樣本)*500 個隨機樣本*5%=750。而根據表一，從左邊(X 值為 0 開始)至第 750 名時，X 值為 7。故本研究的**"選取標準"X 值為 7。**

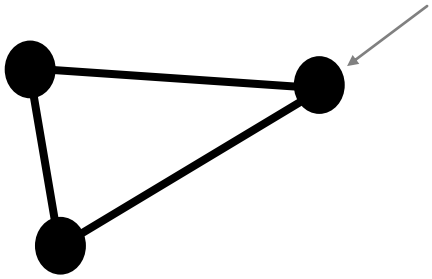
表二： 找尋到既是多基因體也是代表種的病毒(共 54 隻)

Alfalfa mosaic virus	*Infectious bursal disease virus*	*Nodamura virus*
Atkinsonella hypoxylon virus	*Infectious pancreatic necrosis virus*	*Olive latent virus 2*
Banana bunchy top virus	*Infectious salmon anemia virus*	*Ourmia melon virus*
Barley stripe mosaic virus	*Influenza A virus*	*Peanut clump virus*
Barley yellow mosaic virus	*(AGooseGuangdong196(H5N1))*	*Penicillium chrysogenum virus*
Barley yellow mosaic virus	*Influenza A virus*	*Potato mop-top virus*
Beet necrotic yellow vein virus	*(AHong Kong107399(H9N2))*	*Rice ragged stunt virus*
Bluetongue virus	*Influenza A virus*	*Rift Valley fever virus*
Broad bean wilt virus 1	*(AKorea4261968(H2N2))*	*Rotavirus A*
Brome mosaic virus	*Influenza A virus*	*Satsuma dwarf virus*
Bunyamwera virus	*(ANew York 3922004(H3N2))*	*Soil-borne wheat mosaic virus*
Carnation ringspot virus	*Influenza A virus*	*Striped Jack nervous necrosis virus*
Cherry rasp leaf virus	*(APuerto Rico 81934(H1N1))*	*Subterranean clover stunt virus*
Citrus psorosis virus	*Influenza B virus*	*Thogoto virus*
Colorado tick fever virus	*Influenza C virus (CAnn Arbor150)*	*Tobacco rattle virus*
Cowpea mosaic virus	*Lettuce big-vein associated virus*	*Tobacco ringspot virus*
Cucumber mosaic virus	*Lettuce infectious yellows virus*	*Tobacco streak virus*
Cypovirus 1	*Lymphocytic choriomeningitis virus*	*Tomato spotted wilt virus*
Drosophila x virus	*Mycoreovirus 1*	*White clover cryptic virus 1*
Dugbe virus		
Fiji disease virus		

表三：各個病毒比例排序表格

病毒種類	病毒名稱	類型	平均比例
ssDNA(單鏈 DNA 病毒)	*Banana bunchy top virus*	*A*	29.69%
	Bean golden yellow mosaic virus	*A*	27.47%
	Subterranean clover stunt virus	*B*	15.48%
dsRNA(雙鏈核糖核酸病毒)	Atkinsonella hypoxylon virus	*B*	18.22%
	Bluetongue virus	*C*	6.94%
	Colorado tick fever virus	*C*	5.47%
	Cypovirus 1	*C*	8.62%
	Drosophila x virus	*C*	17.84%
	Fiji disease virus	*C*	4.86%
	Infectious bursal disease virus	*C*	16.54%
	Infectious pancreatic necrosis virus	*C*	16.01%
	Mycoreovirus 1	*C*	6.10%
	Penicillium chrysogenum virus	*C*	6.86%
	Rice ragged stunt virus	*C*	6.36%
	Rotavirus A	*C*	6.48%
	White clover cryptic virus 1	*A*	31.44%
NSssRNA 反鏈核糖核酸病毒(The negative-sense ssRNA Viruses)	*Bunyamwera virus*	*B*	13.59%
	Citrus psorosis virus	*C*	9.53%
	Dugbe virus	*C*	9.70%
	Infectious salmon anemia virus	*C*	7.80%
	Influenza A virus (AGooseGuangdong196(H5N1))	*C*	7.60%
	Influenza A virus (AHong Kong107399(H9N2))	*C*	9.16%
	Influenza A virus (AKorea4261968(H2N2))	*C*	6.85%
	Influenza A virus (ANew York3922004(H3N2))	*C*	11.34%
	Influenza A virus (APuerto Rico834(H1N1))	*C*	11.74%
	Influenza B virus	*C*	6.76%
	Influenza C virus (CAnn Arbor150)	*C*	9.11%
	Lettuce big-vein associated virus	*C*	8.25%
	Lymphocytic choriomeningitis virus	*C*	6.16%

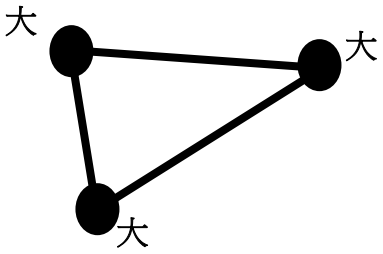
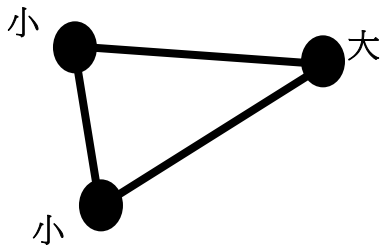
	Rift Valley fever virus	*C*	7.60%
	Thogoto virus	*C*	10.98%
	Tomato spotted wilt virus	*C*	7.75%
PS ssRNA 正鏈核糖核酸病毒(The positive-sense ssRNA Viruses)	*Alfalfa mosaic virus*	*B*	15.94%
	Barley stripe mosaic virus	*B*	18.45%
	Barley yellow mosaic virus	*C*	8.24%
	Beet necrotic yellow vein virus	*C*	10.08%
	Broad bean wilt virus 1	*C*	13.96%
	Brome mosaic virus	*C*	11.31%
	Carnation ringspot virus	*B*	20.33%
	Cherry rasp leaf virus	*C*	9.16%
	Cowpea mosaic virus	*C*	12.49%
	Cucumber mosaic virus	*C*	16.73%
	Nodamura virus	*A*	25.27%
	Lettuce infectious yellows virus	*C*	8.49%
	Olive latent virus 2	*C*	14.48%
	Ourmia melon virus	*A*	23.37%
	Peanut clump virus	*C*	13.31%
	Potato mop-top virus	*B*	16.58%
	Satsuma dwarf virus	*C*	11.83%
	Soil-borne wheat mosaic virus	*C*	12.87%
	Striped Jack nervous necrosis virus	*B*	20.09%
	Tobacco rattle virus	*C*	13.67%
	Tobacco ringspot virus	*C*	7.92%
	Tobacco streak virus	*C*	16.03%

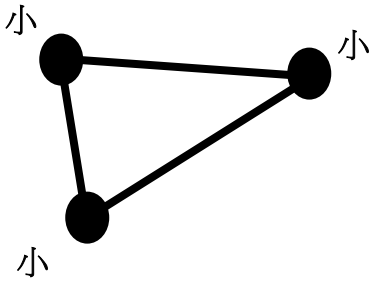


可以在其他基因體上，找到相似短序列的總長，佔該基因體長度的比例值

示意圖，只以三點呈現，因為多點過於複雜。

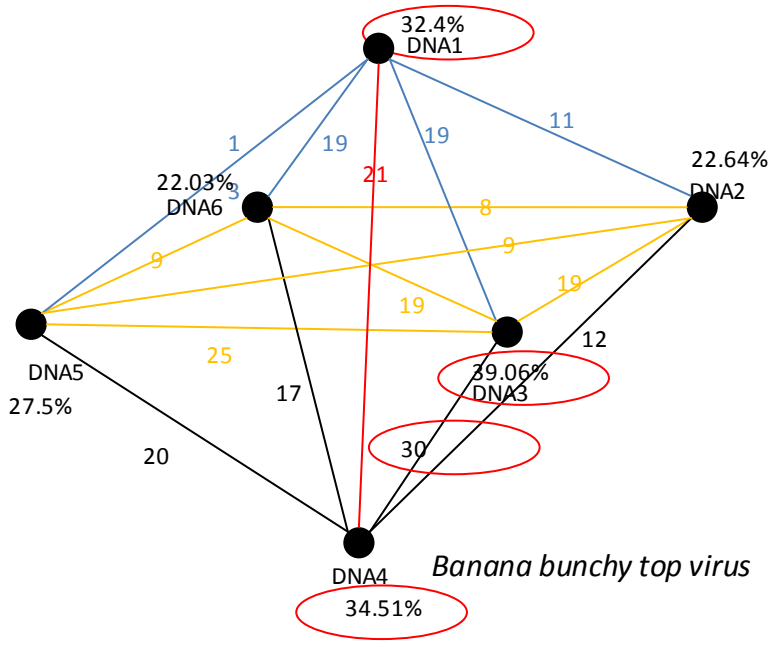
每一點代表一個基因體。”大”、”小”則是代表比例值。

<table>
<tr><td> 
</td><td>代表基因體之間可能有較多的相似短序列，彼此之間的相關性很大，可能具有共同的演化來源。(代稱 A 型)</td></tr>
<tr><td colspan="2">病毒名稱：Banana bunchy top virus、Bean golden yellow mosaic virus、White clover cryptic virus 1 、Nodamura virus 、Ourmia melon virus</td></tr>
<tr><td> 
</td><td>並不是所有的基因體其總短序列的鹼基對佔該基因體總長的比例都高。只有一條可能是在演化上扮演較重要的角色。(代稱 B 型)</td></tr>
<tr><td colspan="2">病毒名稱：Subterranean clover stunt virus 、Atkinsonella hypoxylon virus 、Bunyamwera virus 、Tomato spotted wilt virus 、Barley stripe mosaic virus 、Carnation ringspot virus、Potato mop-top virus、Striped Jack nervous necrosis virus</td></tr>
</table>

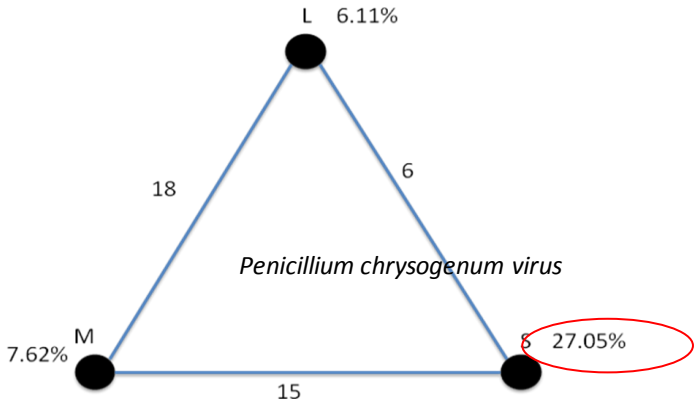
<table>
<tr><td> 
</td><td>幾乎所有的比例都是偏低。代表基因體之間的重組不明顯，互相沒有太大的影響以及關連。(代稱 C 型)</td></tr>
<tr><td colspan="2">病毒名稱： Bluetongue virus 、Colorado tick fever virus 、Cypovirus 1 、Drosophila x virus 、Fiji disease virus 、Infectious bursal disease virus 、Infectious bursal disease virus 、Mycoreovirus 1 、Penicillium chrysogenum virus 、Rice ragged stunt virus 、Rotavirus A 、Citrus psorosis virus 、Dugbe virus 、Infectious salmon anemia virus 、Influenzavirus A 、Influenza B virus 、Influenza C virus 、Lettuce big-vein associated virus 、Lymphocytic choriomeningitis virus 、Rift Valley fever virus 、Thogoto virus 、Tomato spotted wilt virus 、Barley yellow mosaic virus 、Beet necrotic yellow vein virus 、Broad bean wilt virus 1 、Brome mosaic virus 、Cherry rasp leaf virus 、Cowpea mosaic virus 、Cucumber mosaic virus 、Tobacco ringspot virus</td></tr>
</table>

我們針對以上三種型態，選出其中一隻代表性病毒，進行更細部的分析，下圖中每一個邊(Edge)上的數值代表兩個基因體之間所具有的相似短序列數目，這個數值越大，代表兩者之間共享了越多的相似序列，也表示在演化上的相關性可能較大。

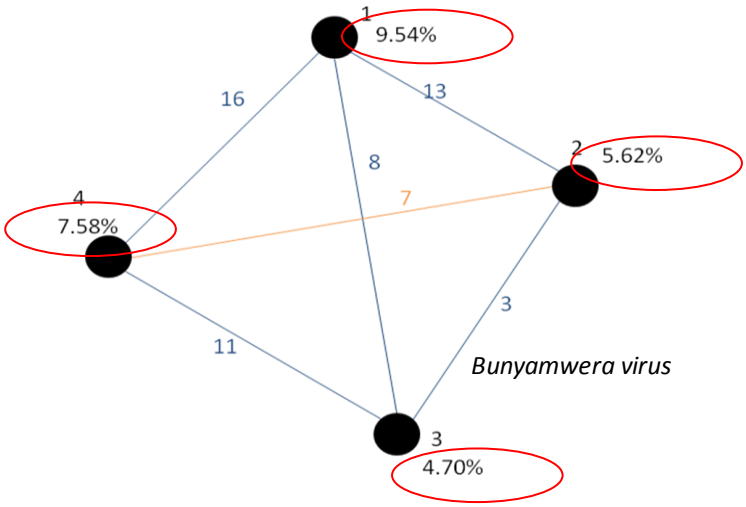
A 型代表 *Banana bunchy top virus*



B 型代表 *Penicillium chrysogenum virus*



C 型代表 *Bunyamwera virus*



陸、討論

在圖六，圖形左側有微凸起的現象，而右邊的突起還是非常的高， 序列總長最短的 5 隻病毒各有 100 個隨機樣本，且在此 500 個隨機樣本中，都以 MEME 進行分析，於每個樣本中，找出 30 個短序列，總共有 15000 個短序列，及從圖六的分析得出初步結果，但仍因為隨機樣本的數量過少，因此不夠嚴謹。

以各個病毒比例排序表格資料來看，發現只有 ssDNA 不存在既是多基因體也是代表種的病毒，而 ssDNA 內應該還有其他的多基因體病毒，只是不是代表種而已。本研究不只可以分析該病毒總平均比例與各種病毒種類的關係，也可以從完整的連結圖中，分析個別基因體之間可能的關係，我們並畫了三種模型分類的代表的實際圖，BBTV (*Banana bunchy top virus*)的 DNA3 與 DNA4 之間的線有很高的數值，顯現出它除了每個基因體之間重組頻繁之外，也有兩個特別關係密切的現象，這種現象也大多都發生在 A 型種類的病毒。

在 A 型種類的病毒內，有些只有兩條基因體，但是互換的比例卻非常高，推論此病毒在複製或需要呈現一些特性時，兩條使用到的機會都很高，所以發生重組的機會更高。從病毒比例排序表格資料最右邊一欄的結果顯示：病毒總比例的平均值，NSssRNA 病毒的平均比例比較小，而 DNA 病毒平均比例比較高，這表示基因體與基因體之間的關係較密切。

柒、結論

1. 從我們研究的病毒中，發現比例約三分之二的病毒屬於 C 型，也就是這些病毒的不同基因體之間沒有太大的相關性，代表病毒內的不同基因體之間在演化上可能沒有太大的關連，或者其關聯性已經因為快速的演化而失去，這與一般傳統對於這類病毒的認識是相同的。

2. 除了 C 型，我們在分析的病毒中也發現了較獨特的 A 型與 B 型，A 型代表基因體之間共享了較大量的遺傳物質序列，可能表示這些基因體之間或有高度重組，或者在演化的過程中可能具有共同的演化來源。B 型則顯示並不是所有的基因體其總短序列的鹼基對佔該基因體總長的比例都高，只有其中某幾條基因體具有特別高的比例值，這種情況可能顯示在此類病毒的演化中，這幾條上基因體可能佔有較重要的地位，例如可能是較為原始的基因體，而其他較不相干的基因體可能是在演化中由外來干擾而留下的。
3. 此次的研究發現：針對 ssDNA，雖然研究的總病毒數比較少，但是發生重組的機率卻高於其他的 RNA 病毒，顯示 ssDNA 多基因體病毒在演化上，其多條基因體之間的相關性很可能較 RNA 多基因體病毒具有更高的關聯性。
4. 傳統上的序列分析方法，多是利用 Alignment(將基因體直接兩兩排序對應) 來得序列之間的相似程度，但是像是病毒這類演化非常快速的生命體，很可能因為累積了太多的突變與重組，所以序列外觀已不相似，只殘留許多不連續且順序不相同的短片段，導致從 Alignment 的角度來分析，大家都認為是完全不相關的序列，但其實仍具有其相關性，例如：平均值最高的 ***White clover cryptic virus 1*** 病毒，以 Alignment 分析的結果顯示其相似度還不到 5%，但是用本研究的方法找到遠比傳統發法更多的相似序列(34.78%)。
5. 多基因體病毒的多個片段，在傳統的分析方法（主要是 alignment）內都認為不具有相關性，但我們的方法卻證明，其序列中其實蘊含了相當多的演化證據(例如 BBTV 的 DNA3 與 DNA4)。
6. 據我們收集的結果，在 NCBI 上的多基因體病毒的數量約有 396 種病毒之多，即使我們只挑選了其中代表種的病毒分析，就已經找到了許多不同以往認知的演化現象，因此我們證明了這樣的分析方法的確可行，也代表了往後我們可以使用這樣的策略來發掘多基因體病毒、甚至不同病毒之間可能的演化機制。

捌、參考資料及其他

1. Claude Fauquet, M.A. Mayo, J. Maniloff, U. Desselberger, L.A. Ball, *Virus Taxonomy*, 2005 Elsevier Inc. Order of Presentation of Virus Taxonomic Descriptions, Pages23-32

2. Claude Fauquet, M.A. Mayo, J. Maniloff, U. Desselberger, L.A. Ball, *Virus Taxonomy*, 2005 Elsevier Inc. The Single Stranded DNA Viruses Pages 277,279-287,289-299,301-341,343-369

3. NCBI(National Center for Biotechnology Information) http://www.ncbi.nlm.nih.gov/

4. ICTV(International Committee on Taxonomy of Virus) http://ictvonline.org/virusTaxonomy.asp?version=2009

5.MEME(Multiple Em for Motif Elicitation) http://meme.sdsc.edu/meme/cgi-bin/meme.cgi

6.鄭惟厚(民 87)譯。統計，讓數字會說話！台北市：天下遠見