基因體-盲自我複製者

中央大學系統生物與生物資訊所及物理系　李弘謙

達爾文於 1859 年提出物種演化論[1]。這是一個極爲大膽的劃時代的想法。在此之前創造論是唯一對物種稍有系統的論說。1866 年孟德爾斯從豆子的變種實驗中總結物種的特徵由上一代遺傳到下一代[2]，1910 年摩根證明果蠅的染色體是遺傳機制的載體[3]，1944 麥克琳托克從玉米的實驗中推論傳授遺傳特徵的工作由染色體中某種單元執行，稱之爲「基因」[4]。但是人們對染色體及基因的微觀結構仍然一無所知。同年，大物理學家，量子物理的共同發明者薛定格在「什麼是生命？」一書中，基於統計概念大膽的假設遺傳與演化的機制必然取決於染色體及基因分子層次結構[5]。這個假設引發了瞭解染色體分子-DNA-結構的激烈競爭，最後由華生與克理克二人於 1953 贏得[6]，同時也創造了嶄新的分子生物領域。上世紀 90 年初，結合分子生物與尖端光電科技的人類基因工程啓動。經過十餘年的發展，人類基因工程成功的完成了包括人類在內的數百種物種全因體的測序[7,8]。

現在我們知道基因體是生命遺傳密碼的載體，也是生命運作的藍圖，而染色體是包裝之後的基因體。每一種物種的每一個細胞裡都藏著一條特定於該物種的基因體。以人類爲例，世界上六十多億人每一個人的億萬個細胞裡都藏著一條相同度約爲 99.9%的人類基因體。

基因體是一種泛稱爲去氧核糖核酸（DNA）的大分子，每條 DNA 由兩條互補的去氧核糖核糖核酸亞基鏈組成，每鏈各爲一條核苷酸序列。核苷酸依它們所含的鹼基分類。鹼基有四種，分爲腺嘌呤（A）、胸線嘧啶（T）、胞嘧啶（C）、鳥嘌呤（G），各由它們英文名稱的第一個字母代表。一條核苷酸序列可以簡單化的由核苷酸中所含的鹼基代表。比如：ACCGTA 代表一條由六個所含鹼基分爲 A、C、C、G、T、A 的核苷酸序列。鹼基可以分類爲兩類互補對：A 與 T 爲一對，C 與 G 爲另一對。如果一個 DNA（分子）的其中一鏈核苷酸序列有一段是 ACCGTA，則另一鏈（互補的）核苷酸序列在相同的位置必爲 TGGCAT。也就是說，因爲有互補的關係，一條鏈序列完全決定了另一條鏈的序列。因此，DNA 本身也可以簡單化的由一個鹼基或字母序列代表。以這種簡化之後的方法描述，人類基因體是一個長約三十億字母的序列。

過去十多年人們對如何解讀這篇密碼雖然尚未完全的瞭解，但是已經取得很可觀的進展。就我們已知，這篇密碼的細膩、微妙、多樣性非常驚人。源源不斷的有關基因體令人嘆爲觀止的新發現令我們讚嘆造物神奇之餘不禁要問：這種神奇的編碼是如何發生的？如果我們還認識以下幾點我們會覺得更不可思議了：所有物種的基因體都共有一個或極少數幾個共同祖先、這個（些）共同祖先在不長於四十六億年前（地球剛冷卻的年代）只是一些毫無生命意義的化學分子、使這個祖先由無生命意義的化學分子逐漸演化成今日的基因體的機制非常簡單－隨機的突變及天擇的篩選。

說到這兒，對物理與統計不太熟悉的朋友們或要問，這在哪兒奇怪了？關鍵在這兒：如上所說，基因體是以 A、C、G、T 四個字母編寫成的一篇生命密碼。亦即它必然不是一篇亂碼。亂碼就是四個字母以最高的亂度，或最低的有序度，排列（有幾近無數種這類排列）。亂碼無法承載任何訊息。與此相反，基因體載著生命的訊息，所以必然要有較高的有序度。因此，基因體演化的方向，必然是從較無序朝較有序的方向走。現在問題來了，隨機的突變，由於是隨機的，看似會增加基因體的亂度，減低它的有序度。天擇的篩選可以部分逃避這個問題。天擇表示不是每一個隨機的突變都會存活。我們可以想像天擇只讓使基因體有序度增加的突變存活。顯然這是一個使基因體訊息不斷增加的可能的演化機制。然而問題又來了。如果所有的突變都要經過天擇的篩選，那麼只有很小一部份的突變能存活。這會大大的減低演化的速率，那麼，就很難

解釋從沒有生命的化學分子演進至今日的基因體只需要四十六億年了。事實上今日的分子演化學家大致同意大部分存活的突變是中性的，並沒有經過天擇[9,10]。

我們如何看基因體的有序度呢？簡單的說，基因體中的字母排列愈均勻亂度就愈高，排列愈不均勻有序度就愈高。我們用一個簡單的例子來闡明這個觀念。假設台灣男女各半的兩千三百萬人口被排成一列，我們要測試這個排列的亂度。方法之一是從排頭到排尾數相鄰的一對人有多少對是男男、男女、女男、或女女。如果排列是隨機（無序）的，那麼任何一人之後的下一個人是男或女的機會相等，所以四類「二人行」的出現次數應該相同，亦即各五百七十五萬次。另一個極端是最有序的排法：所有的女生先排，之後再排所有的男生。這樣就會有女女、男男各出現一千一百五十萬次，女男一次，男女零次。如此，四類「二人行」在任何一排列中的出現次數，是該排列有序度的一種度量。我們可以把四個出現次數簡化而用標準誤差(SD)單量代表，則隨機排列的 SD 爲零，女先男後排列（男先女後亦同）的 SD 爲五百七十五萬。SD 是有序度的很粗粒化的一個代表。我們還可以量各類「三人行」（女女女、女女男、女男女、男女女、、、男男男等共八類）的出現次數，或它們的 SD，對排列的有序度作進一步的測量。我們更可以對「四人行」、「五人行」等等作測量。如是我們可以對排列的有序度作漸進的瞭解。

剛才說隨機排列「二人行」得 SD 爲零並不嚴格的正確。隨機排列嚴格的 SD 與平均出現次數 f 的根成正比，而 f 等於排列長度 L 除以類數（「二人行」類數爲 4、「三人行」類數爲 8、等等）。今定義變化係數 CV＝SD/f，則隨機排列的 CV 平方與 L 成反比：

$$CV^2 = a/L \quad \text{(隨機排列)} \qquad (1)$$

其中 a 爲已知理論常數。我們現在可以用 CV^2 衡量一個排列的有序度。一個長爲 L 的排列有序度的最小值爲 a/L（排列隨機時）。如果長度維持不變，則排列的有序度隨著 CV^2 量的增加而增加；如果排列是隨機的，則排列的有序度隨著長度的增加而減少。公式(1)也可以倒著用，把它寫成

$$L = a/CV^2 \qquad (2)$$

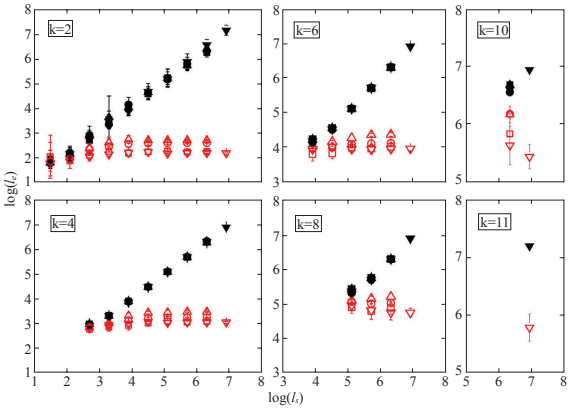
這表示如果我們已知一排列爲隨機，不知其長度但知其 CV，則從（2）可求出其長度。因爲 CV^2 的大小顯示排列的有序度，現在我們用(2)給予有序度一個更直觀的表示如下：如果某排列（相對於「n 人行」）的 CV 已知（或已求出），則**定義**該排列**有效長度**（或有序長度）爲

$$L_e = a/CV^2 \qquad (3)$$

對此方程式我們應該作以下的詮釋：任何一隨機排列的有效長度爲其眞實長度；非隨機排列的有效長度短於其眞實長度；一非隨機排列的有效長度如爲 L_e，則其有序度（或亂度）相等於長度爲 L_e 之隨機排列之有序度（亂度）。我們可以用(3)任意製造長度爲 L 但有效長度爲 L_e 的排列（假設 $L>>L_e$）：先造一個 L_e 長的隨機排列，然後把它的 L/L_e 個拷貝串聯起來即可。

以上藉男女排列介紹用有效長度量化一個序列的有序度的觀念。引用的範例可以直接套用在任何二元化的序列上。我們現在把這個觀念應用在討論量化基因序列的有序度上。這非常簡單，只要把二元的男女以四元的 A、C、G、T 替代，把「n 人行」以 k 個字母的字串（簡稱爲「k 字串」[11]）替代即可。與之前不同的地方是之前「n 人行」有 2^n 類，現在「k 字串」有 4^k 類。比如說「二人行」有 4 類，「二字串」則有 16 類。替代再去算 CV 之後(1)、(2)、(3)就可直接套用（註：有些略去的計算細節很重要；請參看[12,13]）。

我們用上述的方法計算了全基因體資料庫 2006 年 3 月所存共 747 條全基因體及其片段的有效長度。上圖是果蠅（○）、線蟲（△）、芥末（□）、及人類（▽）的部分結果。基因體（後三種物種爲第一條染色體）的長度分爲 4.6、15.1、30.3、226 百萬鹼基。圖中橫軸爲片段序列長度（鹼基數），縱軸爲有效序列長度，紅色表示基因序列的結果，黑色表示將基因序列搗亂之後所得隨機序列的結果 [14]。我們注意到隨機序列的有效長度與眞實長度相同，而基因片段序列的有效長度則很快達到一個遠比眞實長度要短的飽和值。同時，飽和長度似乎與物種無關。這兩個觀察從 747 條全基因體中得到證實，並且發現飽和有效長度滿足一個很簡單的普適性關係：



$$\ln L_e(k) = ak + B \qquad (k = 2 \text{ 至 } 11) \qquad (4)$$

其中 a 值約爲 0.97，B 值約爲 3.7 [13,14]。

關係(4)是一個不簡單也未被人們預料到的結果，它給予我們一些對基因體生長與演化的瞭解的重要啓示。從關係(3)之後的討論，知道如果一條序列的有效長度遠短於它的眞實長度，很自然的解釋是序列中有許多複製。我們根據這個啓示設計了一個極簡單的以隨機片段複製爲主要機制的基因體生長模型[12-14]，成功的解釋了許多與基因體中「k 字串」出現頻率的現像，包括基因體有效長度的普適性及關係(4)。這些結果有生物與物理量方面的意義。

先說生物。隨機片段複製是否眞的是基因體生長的主要機制呢？我們無法說這是絕對正確，但是這個假設可以解釋許多基因體已知的生物現像，也沒有已知的現像與它相違背。這些現像中最主要的是：每一個已知物種的基因體中，不論是編碼段或非編碼段都含有高量的複製片段。事實上所有物種的已知基因絕大部分都是同源的[10]。我們在這兒只舉一個有趣的實例。果蠅第三號染色體上有一個含八個同位序列基因的 HOX gene 串群，共同控制胚胎發育時體形的形成。而許多脊椎動物則有大小不一、位於不同染色體裡的四個同源串群。在人類基因體中它們分別位於第 2、7、12 及 17 號染色體上，每一串群由九到十一個基因組成。這個現像已獲共識的解釋顯而易見：在果蠅與脊椎動物的共同祖先的類基因體內有一個 HOX gene 串群。脊椎動物的祖先在與果蠅的祖先演化分歧之後把這個串群以不同的眞實度複製了三次。基因體中有不勝枚舉的這類例子。

隨機片段複製有什麼物理意義呢？這要回到本文第三段對於爲什麼基因體能在隨機突變的前提下仍能如此快的演化所提出的疑慮。關鍵是隨機片段複製與（中性的）隨機單點（或少數幾點）突變的後果大大不同。隨機突變因爲會增加序列的亂度而極不利於資訊在基因體中累積。隨機片段複製也會增加基因體大尺度的亂度，但是只要片段夠長，偶而也會把載著基因或有其它編碼的片段整個複製下來，因而增加基因體的資訊含量。我們知道靠隨機單點突變的機制編碼是萬分艱難的，所以我們基因體最早的祖先一定是用了極長的時間來累積一點原始的資訊。一旦基因體有了最簡單的複製機能，如果之後基因體的生長以隨機片段複製爲主，那基因體累積資訊的速度就會指數性的增進。所以基因體如果確實是以隨機片段複製爲主要的生長機制，那麼這件事的本身也是天擇的結果。

基因體以隨機片段複製生長其實並不奇怪。我們只要看看四周或回顧自己就會注意到複製他人或自己的文句、編碼、說詞等是累積資訊最常被用也最有效的方法。因爲我們有意識，我們的複製不至於隨機。基因體是沒有意識的，它好比一個盲的抄書匠，只能瞎抄一通。也許這也就是爲什麼基因體是由看似雜亂的片段資訊組成的，而不是像智慧人所寫的、一體成形的一本書。

參考文獻

[1] Charles Darwin, *The Origin of Species*, Penguin Books, London, (1982).
[2] Gregor Mendel, *Proc. Nat. History Soc. Brunn*, **4**, 3, (1866).
[3] T.H. Morgan, *Science*, **32**, 120 (1910).
[4] B. McClintock, *Genetics*, **29**, 478 (1944).
[5] E. Schroedinger, *What is Life*? Cambridge University Press, Cambridge, (1976).
[6] J.D. Watson and F.H.C Crick, *Nature*, **171**, 737 (1953).
[7] J.C. Venter et al, *Science*, **291**, 1304 (2001).
[8] E.S. Lander et al, *Nature*, **409**, 860 (2001).
[9] M. Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge (1983).
[10] D. Graur and W-H Li, *Fundamentals of*

Molecular Evolution, 2nd edition, Sinauer Associates, Sunderland, Massachusetts (2000).

[11] B.L. Hao et al, *Chaos, Solitons and Fractals*, **11**, 825 (2000).

[12] L.S. Hsieh et al, *Phys. Rev. Lett*., **90**, 018101 (2003).

[13] H.D. Chen et al, *Phys. Rev. Lett*., **94**, 178103 (2005); C.H. Chang et al, *J. Bioinfo. & Comp. Biology*, **3**, 587 (2005).

[14] H.C. Lee et al. (To be published.)