遺傳, 環境與基因

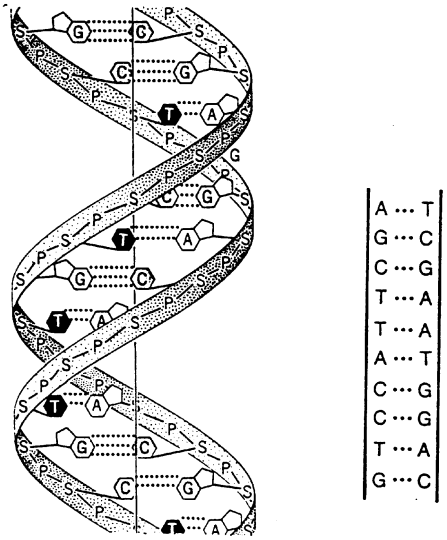
楊照崑

一、前言

近幾十年來多項科技的進步都是一日千里。但很少有比遺傳學爲人類的未來帶來更多的震撼。從疾病的防止到治療, 從胎兒性別及遺傳疾病的檢定到基因的移植與改造, 每一項都表現了一個令前人所夢想不到的事實: 我們正在改變生命的本身。這與飛機、電視、登陸月球有著本質不同的效應。在不久的將來, 也許地球上看到的不止是高樓大廈, 聲光電化的革新, 而是新生物品種, 甚至新人類的繁殖。不久以前生物學家已經把螢火蟲的基因移植到菸草的細胞中而使得整棵的菸草發出了藍色的螢光。這使我們不禁想到有一天我們可以把這個基因移植到路旁的樹裡而形成了一盞盞天然的路燈。但這仍然是身外之物, 我們將不但可以把人身上殘疾的基因拿走, 我們也可以把人性中殘忍的基因除去! 我們不但可以對唐氏呆癡症說拜拜, 我們不也可以爲父母製造一個個聰明的孩子嗎? 地球還有五十億年的生命, 但在此之前我們決不可以坐以待斃。也許我們是宇宙中惟一有智慧去瞭解宇宙的生物呢。我們要離開地球到宇宙中去, 但什麼樣的形體才適合太空旅行不也需要基因工程的改造嗎? 這一切都等待著一個好奇的你。你可以從生物醫學方面上路, 也可以從物理化學方面研究, 更可以從數學統計上去追求。本文將從統計學的眼光來看遺傳問題。我要介紹的有三項: 1. 基因如何找到? 2.遺傳與環境的關係如何測定, 3. 如何試解開遺傳字碼的奧秘。

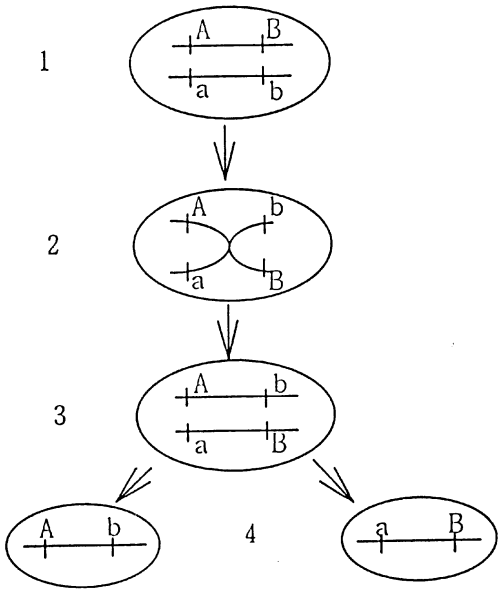
二、近代遺傳學的基本知識

現在我們已知遺傳的信息是由細胞核中染色體上的去氧核醣核酸 (DNA) 裡的字碼所傳出的。這個 DNA 分子是由兩條互補的鍊子所組成 (如圖一)。這鍊子上的珠子有四種分子, 用 ATCG 來代表。DNA 像兩條平行鐵軌, 其中 A 與 T 相扣, C 與 G 相扣 (如圖 1 右圖)。不同的字碼就構成了不同的信息, 主要是用來製造生命的基本成分: 蛋白質及酵素。這並不奇怪, 因爲大型計算機的信息是由一連串更簡單的字碼 0101 所構成的。其實遺傳字碼就是一個計算機程式。這點我們以後會談到。

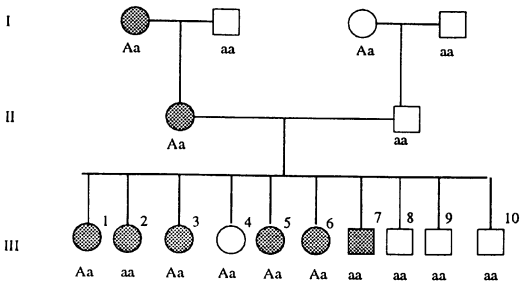


圖一. DNA 分子的模型

一個人有二十二對染色體及一對性別染色體。每對染色體一個來自父親一個來自母親。在人類染色體上的 DNA 約含有 3×10^9 個字碼。基因是一組可以製造一個蛋白質的信息碼, 其長短不一, 大約都在幾千左右。若以同等的比例, 一個基因大約相當於在300公里的長線上找一個特定10cm 長的線段。但這個線段並不好找, 要由它對人體的表現 (phenotype) 去判斷它的位置。譬如說某人有色盲, 但我並不能知道控制色盲的基因在那裡。即使我們把 3×10^9 個字碼排出來了與沒有色盲的人相比較, 我們仍然找不到色盲的基因, 因爲兩人不同的部分太多了。從指紋到臉型, 從身高到智力都會是基因的影響, 那段不同之處會產生色盲不是用這種硬比的方法可以確定的。必須用統計的方法。



圖二. 減數分裂時, 兩條染色體互換位置的模型。1.原細胞, 2. 染色體互換位置, 3. 互換完畢, 4. 形成生殖細胞 (精子或卵)。解開基因位置的主要線索在於當生物生殖細胞分裂形成單個染色體精子或卵細胞之時會有一個交換過程 (見圖二)。原來二條由父母來的染色體會糾纏在一起, 將某些信息互換之後, 再傳至下一代。這種過程在生物上叫減數分裂。在圖二中, Aa表示同一個位置上的兩個等位基因 (即在 A 位置上可以有 A 或 a, 像孟德爾實驗中之黃色與青色豆子之基因, 而 Bb 表示另一個基因, 可以是碗豆花的顏色。) 我們很容易看到若 A 與 B 距離愈近, 則其可以交換成爲圖二 (4) 的機會就愈小, 因此我們可以從 A 與 B 的交換機率 (文中以 θ 表示) 以測定 A 與 B 間的距離。

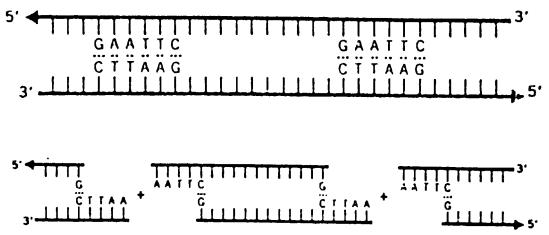


圖三. 小型家譜, 其中方型表男, 圓型表女, 黑色表帶病者, 空白表健康者。圖三是一個具用代表性的小家譜。假設某疾病是由一顯性基因所造成, 其位置我們不知道。但我們知道 A 基因的位置及各人所帶有之等位基因 AA,Aa 或 aa。用圖二的符號來表, 我們令 B 爲帶病基因 (顯性) 且 b 爲正常基因 (隱性)。從圖三中第 I 代與第 II 代之關係可以看出第 II 代母親身上 B 必與 A 同源於 I 代母親, 而 b 與 a 同源於 I 代父親, 即圖二 (1) 之情形。因第 II 代父親必爲二條 ab 之基因, 故第 III 代子女中若有 Ab 或 aB 者必經過了交換過程。由圖可知2,4,7三人的 AB 成份是由交換而來。因此得此種情形之可能性爲

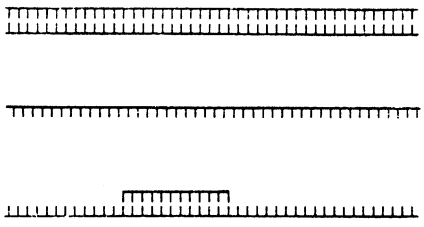
$$L(\theta) = \binom{10}{3}\theta^3(1-\theta)^7。 \qquad (1)$$

若以最大可能性估計法估計 θ 可得其估計值 $\hat{\theta} = 3/10 = 0.3$。不過這僅爲一估計值, θ 的眞值可能在相當大的範圍之外。

從上例可以看出要找基因 B 的位置, 我們必須知道 A 的位置及如何測得在 A 位置二個等位基因的組成: AA Aa 或 aa。這在以往是很困難的事。但在近三四十年來, 由於生物化學及微生物學上的快速發展, 我們已在染色體上找到了許多已知位置的標記, 它們與已知位置的基因在基因的定位上有同樣的功能。這些標記稱之爲「限制碼片段長度多樣性」(RFLP, restriction fragment length polymorphism) 及多樣長度重複片段 (VNTR, variable number tandem repeats)。相信大家或許已看過了這兩個英文的縮寫。由於它們對近代遺傳學太重要了。我必須簡單的介紹一下。



圖四. 限制碼ECORI 切割 DNA 之模型, 上圖切割前, 下圖切割後。

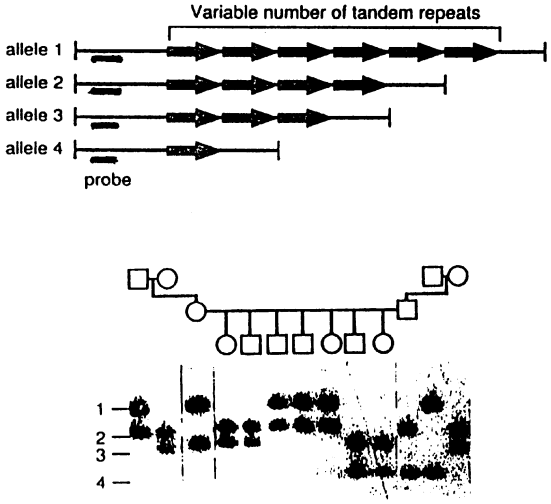


圖五. DNA 雙螺旋被分開及被探針 (小片者) 附著之模型。

當人類 (或其他生物) 的細胞經過化學處理之後, 可將染色體中之 DNA 成分分離出來。用一種叫限制碼的酵素可將 DNA 的分子在固定的位置上加以切割。這種碼原來是細菌用來切割它們的敵人, 濾過性毒, 入侵的 DNA 之用的。因其可以限制病毒之生長, 所以稱之爲限制碼。圖四代表一種叫 ECORI 的限制碼, 它專切 GAATTC 的

地方。因此碡含有6個字碼, 又 DNA 字碼之排列大體而言不會含什麼規則, 因此六個字碼可有 $4^6 = 4096$ 個組合。也就是說平均每4096處就有一處被切, 因此人類全部的 DNA 可被切成約 $3 \times 10^9/4096 =$ 一百萬個片段。當然各段並不一樣長, 因 DNA 的字碼並非週期性的排列著。而且每個人被切的各段也不全相同, 因人與人之間的 DNA 雖大同但有小異。下一步是用一個適當的方法 (通常是加熱) 使得雙條 DNA 分子分離成單條 (如圖五所示)。這時若使用一種半面的探針 (圖五下圖的小片段) 則它就會附在它互補的位置上。譬如這隻探針是 ATCGT, 則它必附著在 TAGCA 上面。若探針含有20個字碼, 則在 DNA 的排列中只有 $1/4^{20} = 9 \times 10^{-13}$ 的機會找到同樣的組合, 因人類 DNA 之總長只有 3×10^9。因此碰上了, 在全染色體中, 大概就只有這麼一個片段了。也就是說探針可以找到每個人含有此探針組合的片段。但這也有一個條件, 因人與人的字碼大同小異, 這個探針必須在大同的部分而不在小異的部分上。這樣的探針已找到了很多, 但卻花了相當的工夫來確定它的使用的價值。

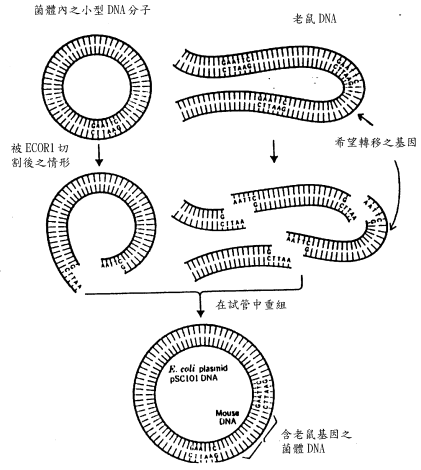
即使探針部分人人相同, 但含有探針的切割片並不一定人人相同, 因探針不過20個字碼, 而切割片的平均長度是4096。因此同一探針所得之片段可有多種的變化, 因此稱之爲「限制碡片段長度多樣性」。但又因此種長度之不同多半發生於某種字碼之不斷重複, 如 $GTGT\ldots = (GT)_n$, 而 n 可有多種變化, 故又可稱之爲「多樣長度重複片段」。圖六表示了某個被切片段受到同一探針的識別之後, 有四種不同的長度 1,2,3,4 個等位標記, 而它們在每個人身上的長度又在圖六下圖中表示出來。這種表示的工具叫電游。它是因片段重量的不同在同時間內會在膠中因靜電力移動不同的距離。而探針中所含的放射線使得它們的位置在底片上顯示出來。在圖中, 左邊第二代的母親有等位標記1與3, 父親有 2與4, 因此他們的七個子女必須有父母親各一個等位標記。若是其中一個子女的標記是 1/3 或 3/3, 則他 (她) 就不會是他們的子女了 (註一)。在此我們可以看見 RFLP 或 VNTR 完全可以看成一個已知位置的基因。而現在這種標記在人類及生物中已大量的發掘, 使得找基因有了大量定位的工具。



圖六. RFLP, VNTR 與遺傳標記之關係。上圖: DNA經限制碡切割之後, 由探針 (probe) 所指認之不等長片段。下圖: 此種片段在親子之間所形成之關係, 即每個人有二個片段, 一從父親來, 一從母親來。

圖七所表示的是如何將老鼠的一段 DNA 在試管中移植到一個細菌的 DNA 分

子中去。這種移植之後的 DNA 又可用融合, 微注射, 或由細菌或病毒帶入其他生物之生殖細胞而形成新的品種。當然這其中含著許多困難的步驟, 包括移植的基因是否能與原生物體內其他基因合作。這往往都是試了才知道。有興趣的讀者可參考資料1。



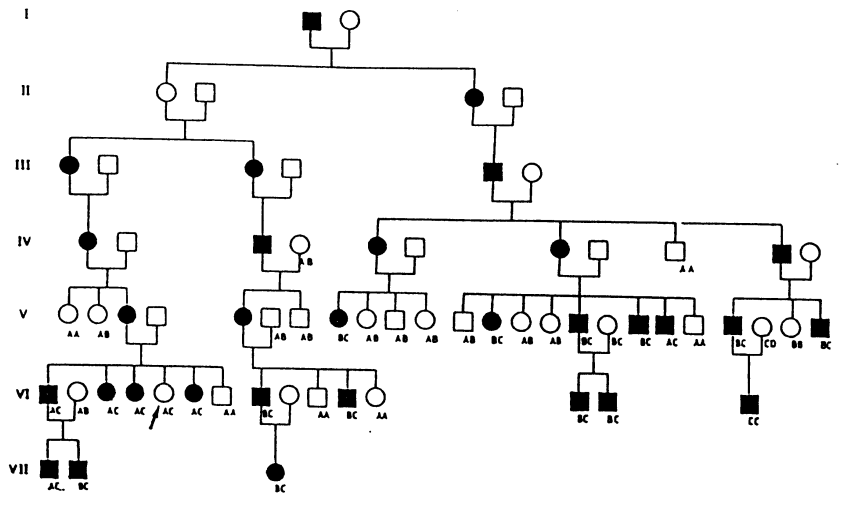
圖七. 基因轉移之步驟

三、單個基因位置之測定

到目前爲止, 如果一個屬性像身高, 智力可能被許多基因共同推動, 要找這些基因除非運氣特別好, 很難做到。但各種顯然由單一基因引起之疾病像亨丁頓舞蹈症。鐮刀紅血球貧血症, 基因往往可從家譜找出來。其原則仍如圖三, 只是家譜大得多, 可有幾百人。而致病的基因也往往是隱性的, 因而概率公式也比 (1) 要複雜很多。圖八就是一個有名的例子。要分析這樣的家譜, 沒有計算機是辦不到的。但其原則仍可以圖三爲例, 假若圖三中的第一代已經消失, 則第二代母親的一對染色體會是 AB 與 ab 或 Ab 與 aB 則無法知道。我們必須假定這兩種情形各爲 $1/2$, 因此 (1) 變成了

$$L(\theta)=\frac{1}{2}\left[\binom{10}{3}\theta^3(1-\theta)^7+\binom{10}{3}\theta^7(1-\theta)^3\right] \tag{2}$$

由此式可求得 θ 之最大可能性估計值爲 0.318。



圖八. 用以找亨丁頓舞蹈症基因所用之委內瑞拉家譜, 圖中A,B,C,D 表示某標記之多樣性, 與圖六之1,2,3,4只是符號不同。

不過用這種方法還得有一個條件, 即我們必須由一個人的表現可以推斷出他二條等位基因的成分。譬如說若帶病基因爲顯性, 他是一個病人, 他必須是 BB 或 Bb, 而健康者必須是 bb。反之若帶病基因爲隱性, 則病者必須是 bb。但基因的表現常會不明確, 以乳癌爲例, 並不是每個帶致病乳癌的基因者都會發病。在此種情形下, 用構成 (1) 式或 (2) 式就有困難。在這種情形下, 可用一種叫「得病親屬」法。即只考慮已得病的親屬, 特別是兄弟姐妹。因得病者必定會有致病基因。這時若得病者大部分都有某個等位標記, 則這個標記與致病基因必定不遠。基因也因而可以找到。這種抽樣多半要選很多的小家庭, 與圖八的大家譜有相當的不同。但兩者在抽樣與分析上都必藉重統計知識。

四、遺傳成分的估計

當許多基因可能同時作用於某特性時, 因需用到多個標記。其組合數之大目前的計算仍望塵莫及。但直覺上我們應可從血統與某種特性的關係來判斷遺傳成分之大小。淺言之, 一個屬性常發生在親屬身上, 其遺傳性必大, 反之則必小。不過基因的表現往往會因環境而異, 像有些鱷魚, 它的性別完全決定於卵孵化時的溫度, 與性染色體的組成無關。但要把人一生的遭遇計量化非常困難。只有極少特例, 我們可以把環境與遺傳的因子分開。雙生子是一個佳例。雙生子的類性與成長過程可有下面四種情形:

$$\begin{aligned} \text{MZT} &= \text{同卵雙生, 同處長大。} \\ \text{DZT} &= \text{異卵雙生, 同處長大(同性別)。} \\ \text{MZA} &= \text{同卵雙生, 異地長大。} \\ \text{DZA} &= \text{異卵雙生, 異地長大。} \end{aligned} \tag{3}$$

若令 X_{i1}, X_{i2} 分別爲二個雙生子在屬性 X 上之表現, 則一簡單的模型爲

$$\begin{aligned} X_{ij} &= u_{ij} + e_{ij} + \epsilon_{ij}, \\ & \quad i = 1, 2, \ldots, n,\ j = 1, 2。 \end{aligned} \tag{4}$$

在式中 u 表遺傳成分, e 表環境成分而 ϵ 表其他不可測之變異, 又其中下標的 i 表示家庭, j 表示二個雙生子。故在此模型中, 對同卵雙生而言 $u_{i1} = u_{i2}$, 而對同處長大的雙生子, $e_{i1} = e_{i2}$。我們可以想像若對 (3) 中之四種雙生子都有相當的數據, 則可以測定 u 與 e 在 X 中所佔之分量。本文不介紹統計計算的細節, 只以有名的密里蘇打雙生子調查爲例 (參考資料2)。

表一. 雙生子間性格之相關性

項目	MZT	MZA	DZT	DZA
生活品質	0.58	0.48	0.23	0.18
成就	0.51	0.36	0.13	0.07
對抗外在壓力	0.52	0.61	0.24	0.27
自我控制	0.41	0.50	-0.06	0.03
保護自己	0.55	0.49	0.17	0.24

樣本大小 MZT=217, MZA=27, DZT=114, DZA=27

計算結果發現從相關性 (註二) 來看, 同卵雙生異地而養要比異卵雙生同地而養大得多, 從表一我們可以看出同卵的影響比同地的影響大得太多了。這是對遺傳重要的一

個有力的證明。大體而言, 在類似的環境下, 一個人的成就, 待人, 接物與自我約束力, 遺傳佔50%, 同養育地佔5-10%而另外40-45%則爲不可測知的變異。因爲並非在同家庭同學校長大的雙生子他們的環境影響會相同, 他們可能交不同的朋友, 看不同的漫畫, 走不同的路, 吵不同的架, 甚至在別人看不見的一起小事件會改變了一個人的一生。最近有人用同樣的雙生子模式調查遺傳與愛情關係, 發現愛情內涵及表現的方式並不遺傳, 多麼奇怪的發現! 我把它留給你去思考玩味 (參考資料3)。

環境對人固然不易控制, 但對某些生物, 特別是家禽家畜, 我們可以控制到每隻動物 (如養雞房中的雞) 幾乎在完全相同的情況下長大。對這些生物而言, 遺傳比例就容易測定, 但也會因飼養的環境而不同。譬如說, 雞蛋大小的遺傳比例, 各種報導可從40%到80%。但這對農業仍有極大的幫助。尤其是在選種的時候, 我們想找到最能育出優良後代的個體。一般選種都用這樣的模型。令 X 表示一可量之屬性, 如蛋之大小或乳之多少, u 表遺傳成分, ϵ 表不可測成分, i 表個體指標, 則

$$X_i = u_i + \epsilon_i, \quad i = 1, 2, \ldots, n。 \qquad (5)$$

在此因環境全相同, 故不用放入。我們的目的是找最好的 u_i, 即 i 個體有最好的遺傳。請注意 (5) 式中之 u_i 是不可能用代數中聯立方程式去解的, 因在 n 個方程中有 $2n$ 個未知數, 即 u_i 與 ϵ_i, $i = 1, 2, \ldots n$ 都不知道。但若假定 ϵ_i 滿足一定的分佈 (在生物界中, 以常態分佈爲主) 而 u_i 之間的親屬關係可以確定, 如1與2爲父子關係, 3與4爲兄弟關係, 則每個 u_i 的值都可以以統計的方法測定的。譬如說某個個體的 X_i 值很好, 但牠父母子女的 X 值都不好, 則這個好的 X_i 可能是非遺傳的 ϵ_i 所造成的。因此用牠去繁殖後代效果可能不好。此種模型在農業已用了幾十年了, 我們今天能看到的肉蛋奶羊毛等產品都用育種在逐年的改進之中。但只有父子, 兄弟來決定遺傳成分是不夠精確的。親子代標記的轉移 (即第二節內所談的 RFLP 或 VNTR), 更可確定標記與 X 值之關係, 進而更精確的指示出最佳的品種。

五、試解遺傳的密碼

不久的將來, 人類全部染色體上的字碼即可公佈, 即所謂的 Human Genome Project。當然這只是一個骨桿, 人人的字碼不全相同。但基本上, 人的奧秘全在這裡了。如何從一個受精卵發展成一體嬰兒而進而成爲萬物之靈的全部製作法全在這裡了。但這有 3×10^9 個字碼。若以平常的書來計, 一個字的位置放一個 A,T,C 或 G, 每頁1000字 (很密) 每書1000頁 (很厚) 也可以裝成三千册書。我們如何從這樣龐大的數據中去解開生命的奧秘會是二十一世紀人類最大的挑戰之一。

首先我們要瞭解, 一個受精卵之所以能用DNA上的信息逐漸形成一個生物乃是因爲DNA所寫的是一個類似電腦用的程式。它的寫法是:

IF(A存在)$THEN$激活基因 B 而生產蛋白質 C

IF(D存在)$THEN$激活基因 E 而產生酵素 F
IF(F存在)$THEN$壓制基因 H 使之不作用
IF(H存在)$THEN$激活基因 I 而產生抗體 J
(6)

在此程式中的 A,D,F,H 可以是外來的抗原, 可以是自己的荷爾蒙, 也可以是小的化學分子。當然我們在 DNA 上看不見 IF 與 $THEN$ 的字樣, 它們都是用酵素來代替的。以1965年得諾貝爾獎的 Jacob 與 Monod 的實驗爲例。大腸菌有一個基因 B 可以製造消化乳醣的碡 C, 但它平時不起作用, 因有另一個基因製造一個蛋白質可以抑制基因 B。但若大腸菌的環境中有乳醣 A, 這個抑制蛋白就會與乳醣結合而失去了抑制 B 基因的功能, 因此 B 基因被激活而產生了乳醣消化碡。這正是 (6) 程式中的第一個步驟。其中 $IF-THEN$ 的功能是用那個抑制基因 B 的蛋白質。但到底要如何從這些 DNA 的字碼中去掘發出這個程式, 正是一個重要的研究方向。各方好手都已食指大動, 讓我們來看看一個有趣的發現, 作爲一個起程。

在語言學中有一個重要的定律叫齊夫定律 (Zipf's Law)。這是齊夫在做了大量語言統計後所歸納的一個通則。如果我們將一本大書中的字, 依用的次數多少排起來, 譬如說某中文小說共一千頁, 其中"的"出現次數最多, 共5562次, "是"第二名, 共5021次... 若用 $f(r)$ 表示排行前 r 次用字的次數, 即 $f(1)=5562$, $f(2)=5021\ldots$ 則

$$f(r) \propto r^{-\delta}, \tag{7}$$

式中 δ 爲一常數, 與文字及書都可能有關。但對文字而言, δ 接近於1。因 DNA 也是語言的一種, 它也會滿足齊夫定律嗎? 當然 DNA 中找不到"字", 只有混在一起的 ATCG 字碼。那麼它應該像一個電腦程式轉換成0101之後的情形吧。表二中有著各種"語言"在齊夫定律中的表現。這已是1995年 Nature 上的結果了。它代表著什麼? 我們能用生物化學上的知識再加上統計的方法及快速的計算機解釋這生命的語言嗎?

表二. Mantegna 等對各種類語言滿足齊夫定律的調查

"字"型	δ
百科全書 (字, 英語)	0.85
百科全書 (每3-5英文字母)	0.57
計算機程式 (十二個"0" "1"碼)	0.77
亂碼 (十二個"0' "1"碼)	0.0
哺乳動物 DNA(六字碼)	0.250
無脊框動物 (六字碼)	0.340

在 ( ) 中表示"字"的定義; 字即一個個的英文字, 3-5字母即以3-5個子母當作字, 例如 I go to school today 的4個子母"字"爲 igot osch oolt oday。計算機程式取十二個 01碼爲一個字而 DNA 取六個字碼爲一"字"。

「銀燭秋光冷畫屏, 輕搖小扇撲流螢, 天階夜色涼如水, 臥看牽牛織女星。」當杜牧爲我們創作這幽美的詩句的時候, 他可曾想到那微弱的星光與螢光之中, 隱藏著天與地的故事? 隱藏著生命的奧秘, 隱藏無限的未來? 我們又好像回到了四百年刻卜勒的書桌前。當他面對著布拉三十年火星位置記錄, 要從這亂麻似的數據中找出規則, 他懷的是什麼心情? 他會預知這數據裡藏著宇宙運行的奧秘嗎? 而我們會又有他那樣的機緣在短短的十幾年內解開這個 DNA 的奧秘嗎?

註釋

註一: 用此法判定親子關係往往需要幾個 RFLP 以防突變的情形發生 (一代一個字碼的突變率約爲 10^{-7}, 因此在兩個以上的RFLP同時突變幾乎不可能。)

註二: 即correlation coefficient 定義爲一般公式

$$\rho = \frac{\sum(X_{i1} - \bar{X}_{.1})(X_{i2} - \bar{X}_{.2})}{\sqrt{\sum(X_{i1} - \bar{X}_{.1})^2 \sum(X_{i2} - \bar{X}_{.2})^2}}$$

參考文獻

(文中引用)

1. Kingsman, SM, & Kingsman, AJ, *Genetic Engineering*, Blackwell, 1988.
2. Tellegen, A. et al.,(1988) Personality similarity in twins reared apart and together, *J. Personality Soc. Psycho.*, 54, 1031-39.
3. Waller, N.G. and Shaver, P. R., (1994) The importance of onogenetic influences on romantic love styles: A twin-family study, *Psycholo. Sci.*, 5, 268-274.
4. Mantegna, R. N. et al.(1995) Linguistic features of non-coding DNA sequences, (to appear in *nature*)
(未特別引用)
5. Ott, J., *Analysis of Human Genetic Linkage*, Johns Hopkins U. Press, 1991.
6. Weir, B. S., *Genetics Data Analysis*, Sinauer Assoc. Inc., 1990.
7. Russell, P. J., *Genetics*, 2nd Ed, Harper, 1990.
8. Weaver & Hedrick, *Genetics*, WMC Brown, 1989.
9. Genome (Chinese Translation: 基因聖戰), 1994.
10. Lander, E.S. and D. Botstein, (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, *Genetics*, **121**, 185-199.

—本文作者現爲清華大學統計所客座教授—