DNA序列與數學分析簡介

黃俊雄

一. 前言:

近代的分子生物研究起始於 1953 年 Watson 及 Crick 發現一 DNA 之結構爲雙線螺旋形。這一革命性的發現導致許多 DNA data 被一一的從各種生物的細胞核內讀出來。至1970中期讀 DNA 之技術更有進一步突破, 費用進一步降低。至今大約有 3×10^7 nuclestides (構成 DNA 之基本分子, 有四種) 存於美國 Los Alamos National Laboratory 的 GenBank(電腦資料庫) 及歐州德國 Heidelberg 之 EMBL databank (European Molecular Biology Laboratory)。日本亦已加入他們的資料庫系統。由於 DNA 與生物細胞功能、進化、遺傳、疾病等有非常密切之關係, DNA 可說是生物的基本因子。如今有這麼大量的 data要分析、處理, 這可需要數學家的參加大力幫忙。因爲生化學家天天忙於實驗整理資料, 大多 DNA 序列長達數萬個 nuclestides, 要把每個 nuclestide 讀出無誤可眞費時費神, 他們實在沒時間做數學分析。因此數學家們應多多參與幫忙。我們正面臨一生物革命時代, 正如數百年前伽利略發現行星運動三大定律。當時他必須面對 Tycho Brahe 花了一輩子時間所收集來的天文資料來分析。同樣地, 如今有大量的 DNA data 被讀出來, 這與細胞的功能、遺傳、疾病、進化等有什麼關係, 可眞需要聰明的數學家來研究解釋並找出定律來。有一點現在數學家比伽利略佔優勢的是: 如今有快速電腦幫忙計算, 幫忙繪圖顯像, 幫忙分析各種情形。

二. 生物背景:

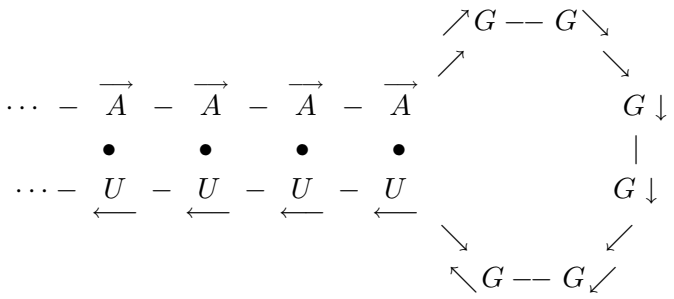
每一生物細胞內有細胞核, 核內有染色體 (chromosomes) 及核酸 (nucleic acids)。一 genome大致可說是染色體部份集合, 內含有很多基因 (gene), 一 gene 可說是 DNA 序列的部份序列。一 DNA(deoxyribonucleic acid 去氧核糖核酸) 是由四種型式的 nucleotides (叫 A (adenine), G (guanine), C (cytosine), T(thymine)) 組成, 其長度有的長達百萬單位 (每一 nucleotide 有一定分子式, 由碳、氫、氧、氮等組成)。其結構有雙線螺旋型, 也有雙線捲圈圓型, 或單線一條型, 或單、雙線圓型等 (見表一)。核酸有兩種, 一爲 DNA, 另一爲 RNA (ribonuclei acid 核糖核酸)。RNA 也是由四種型式的 nucleotides(叫 A,

G, C, U(uracil)) 組成, 其長度大多爲數千個單位而結構更複雜。在單一有機體內存在著數千個不同的蛋白質 (protein)。每一蛋白質是由一序列 amino acids(氨基酸) 組成, 大多長達數百個 (有的更長)。每一 amino acid 是由三個 RNA 之 nucleotides(即 A, G, C, U) 組成, 這組成單位叫 codon。數學上應有64種 codons, 但實際上存在的只有20種 (見表二)。已讀出的 protein 序列皆存在一資料庫, 叫 Protein Identification Resources(PIR) data base, 至今約有3000個序列。

由物理化學特性上, nucleotides 傾向於結合成一對一對存在著: G 與 C 結合, A 與 T 或 U 結合。一例如下:

$$\begin{array}{ccccccccccccccccccc}
\overrightarrow{A} & - & \overrightarrow{G} & - & \overrightarrow{C} & - & \overrightarrow{T} & - & \overrightarrow{T} & - & \overrightarrow{T} & - & \overrightarrow{C} & - & \overrightarrow{T} & - & \overrightarrow{A} & - & \overrightarrow{A} & - \cdots \\
\bullet & & \bullet & & \bullet & & \bullet & & \bullet & & \bullet & & \bullet & & \bullet & & \bullet & & \bullet & \\
T & - & C & - & G & - & A & - & A & - & A & - & G & - & A & - & T & - & T & - \cdots \\
\longleftarrow & & \longleftarrow & & \longleftarrow & & \longleftarrow & & \longleftarrow & & \longleftarrow & & \longleftarrow & & \longleftarrow & & \longleftarrow & &
\end{array}$$

很多 DNA 是由雙線 DNA 結合一起, 其結合規則即如上述。因此叫此雙線爲 complementary pairs(互補對)。在 RNA 序列中, 通常爲單線的, 但常有部份序列與其他部份序列結合一起, 而沒結合的部份就成圓圈狀, 這叫 secondary structure(見下圖)。這種結構在分子的形狀 (3-dimensional, 很複雜) 及功能佔有很重要的地位。



一實際已讀出的 RNA 例子是 E. coli 16S RNA, 列於圖1。請注意, 大多結合成對的是合於上述規則 (用短直線連接), 也有不合於上述規則 (用打點表示)。

一生命有機體會吸收養份, 使一 DNA 能複製自已成更多 DNA, 一些 DNA 經由一轉錄過程 (transcription) 可轉變成 Untranslated RNA(tRNA) 及 messenger RNA (mRNA), 而 messenger RNA 可經由一 translation 過程轉變成 protein。在

衆多 cellular RNA 中, messenger RNA 大約5%, 佔最多的 (約80%) 叫 ribosomal RNA(rRNA) 其正確功能未知, 但有證據指出它對 protein 合成時某些部件之結合而成爲 ribosome 極爲重要。

表一. **COMPLETELY SEQUENCED GENOMES**

Organism[a]	**Genome type**[b]	**Sequence length**[c]	**Accession no.**[d]
Organelles			
Mouse mitochondrion	ds-DNA, circular	16295	J01420
Bovine mitochondrion	ds-DNA, circular	16338	J01394
Human mitochondrion	ds-DNA, circular	16569	J01415
X. laevis mitochondrion	ds-DNA, circular	17553	M10217
Eukaryotic plasmids			
S. cerevisiae 2 μm plasmid	ds-DNA, circular	6318	J01347
K. lactis 2 Kl plasmid	ds-DNA, circular	8874	X00762
Prokaryotic plasmids			
pSN2	ds-DNA, circular	1288	J01763
pC194	ds-DNA, circular	2910	J01754
pBR327	ds-DNA, circular	3273	J02549
pE194	ds-DNA, circular	3728	J01755
pVH51	ds-DNA, circular	3847	K03114
pBR329	ds-DNA, circular	4150	J01753
pJD1	ds-DNA, circular	4207	M10316
pBR322	ds-DNA, circular	4363	J01749
pT181	ds-DNA, circular	4437	J01764
Co1E1	ds-DNA, circular	6646	J01566
RSC13	ds-DNA, circular	7894	J01783
Animal viruses			
Duck hepatitis B virus	ms-DNA, circular	3021	K01834
Human hepatitis B virus(ayw)	ms-DNA, circular	3182	J02203
Human hepatitis B virus(adr)	ms-DNA, circular	3188	V00867
Human hepatitis B virus(adw)	ms-DNA, circular	3200	V00866
Woodchuck hepatitis virus(WHV1)	ms-DNA, circular	3308	J02442
Woodchuck hepatitis virus(WHV2)	ms-DNA, circular	3320	M11082
Ground squirrel hepatitis virus	ms-DNA, circular	3311	K02715
Avian sarcoma virus Y73	ss-RNA, linear	3718	J02027
FBR murine osteosarcoma virus	ss-RNA, linear	3791[e]	K02712
FBJ murine osteosarcoma virus	ss-RNA, linear	4026[e]	J02084
Black beetle virus	ss-RNA, 2 linear segments	4504	K02560
Adeno-associated virus 2	ss-RNA, linear	4675	J01901
Fujinami sarcoma virus	ss-RNA, linear	4788[e]	J02194
Human polyomavirus BK(MM)	ds-DNA, circular	4963	J02039
Minute virus of mice	ss-DNA, linear	5081	J02275
Human polyomavirus JC	ds-DNA, circular	5130	J02226
Human polyomavirus BK(Dunlop)	ds-DNA, circular	5153	J02038
Parvovirus H1	ss-DNA, linear	5176	J02198
Simian virus 40	ds-DNA, circular	5243	J02400
Lymphotropic papovavirus	ds-DNA, circular	5270	K02562
Polyoma virus (a3)	ds-DNA, circular	5296	J02289
Polyoma virus (a2)	ds-DNA, circular	5297	J02288
Simian sarcoma virus	ss-RNA, linear	5319[e]	J02394
Crawford small-plaque polyomavirus	ds-DNA, circular	5350	K02737
Abelson murine leukemia virus	ss-RNA, linear	5659[e]	J02009
Moloney murine sarcoma virus(1)	ss-RNA, linear	5828[e]	J02266
Moloney murine sarcoma virus(124)	ss-RNA, linear	5833[e]	J02263
Spleen focus-forming virus	ss-RNA, linear	6296[e]	K00021
Human rhinovirus type 14	ss-RNA, linear	7212	K02121
Polivirus type 3	ss-RNA, linear	7431	K01392
Polivirus type 3 attenuated	ss-RNA, linear	7432	K00043

ds: double stranded

ms: mixed (single+double)

ss: single stranded

表一. **COMPLETELY SEQUENCED GENOMES**

Organism[a]	Genome type[b]	Sequence length[c]	Accession no.[d]
Poliovirus type 1	ss-RNA, linear	7440	J02281
Poliovirus type 1 attenuated	ss-RNA, linear	7441	V01150
Human hepatitis A virus	ss-RNA, linear	7478	K02990
Human papillomavirus 1A	ds-DNA, circular	7811	V01116
Cottontail rabbit papillomavirus	ds-DNA, circular	7868	K02708
Human papillomavirus 6b	ds-DNA, circular	7902	X00203
Human papillomavirus type 16	ds-DNA, circular	7904	K02718
Bovine papillomavirus type 1	ds-DNA, circular	7945	J02044
Maloney murine leukemia virus	ss-RNA, linear	8332	J02255
AKV murine leukemia virus	ss-RNA, linear	8371	J01998
Bovine leukemia virus	ss-RNA, linear	8714	K02120
Human T-cell leukemia virus type II	ss-RNA, linear	8952[c]	M10060
Human T-cell leukemia virus type I	ss-RNA, linear	9032[c]	J02029
Lymphadenopathy-associated virus	ss-RNA, lincar	9193	K02013
Visna lentivirus	ss-RNA, lincar	9202	M10608
Rous sarcoma virus	ss-RNA, linear	9625[c]	J02342
AlDS-associated virus-2[f]	ss-RNA, linear	9737[c]	K02007
Human T-cell leukemia virus type III[f]	ss-RNA, linear	9751[c]	K02083
Yellow fever virus	ss-RNA, linear	10862	K02749
Vesicular stomatitis virus	ss-RNA, linear	11162	J02428
Sindbis virus	ss-RNA, linear	11703	J02363
Influenza type A	ss-RNA, 8 linear segments	13588	J02143
Adenovirus 2	ds-DNA, linear	35937	J01917
Epstein-Barr virus	ds-DNA, linear	172282	V01555
Plant viruses			
Coconut cadang-cadang viroid (fast)	ss-RNA, circular	246	J02050
Avocado sunblotch viroid	ss-RNA, circular	247	J02020
Coconut cadang-cadang viroid (slow)	ss-RNA, circular	287	J02051
Hop stunt viroid	ss-RNA, circular	297	X00009
Cucumber pale fruit viroid	ss-RNA, circular	303	X00524
Chrysanthemum stunt viroid (CSV2)	ss-RNA, circular	354	M19506
Chrysanthemum stunt viroid (CSVI)	ss-RNA, circular	356	M19505
Potato spindle tuber viroid	ss-RNA, circular	359	J02287
Tomato apical stunt viroid	ss-RNA, circular	360	K00818
Tomato planta macho viroid	ss-RNA, circular	360	K00817
Citrus exocortis viroid (C)	ss-RNA, circular	371	J02053
Citrus exocortis viroid (DE25)	ss-RNA, circular	371	K00964
Citrus exocortis viroid (DE26)	ss-RNA, circular	371	K00965
Satellite tobacco necrosis virus	ss-RNA, linear	1239	J02399
Maize streak virus	ss-DNA, circular	2687	K02026
Tomato golden mosaic virus	ss-DNA, 2 circular segments	5096	K02030
Bean golden mosaic virus	ss-DNA, 2 circular segments	5233	M10070
Cassava latent virus	ss-DNA, 2 circular segments	5503	J02057
Tobacco mosaic virus (vulgare)	ss-RNA, linear	6395	J02415
Cauliflower mosaic virus (D/H Hungary)	ds-DNA, circular	8016	J02047
Cauliflowver mosaic virus (Strasbourg)	ds-DNA, circular	8024	J02048
Cauliflower mosaic virus (CM1841)	ds-DNA, circular	8031	J02046
Brome mosaic virus	ss-RNA, 3 linear segments	8213	K02706
Alfalfa mosaic virus	ss-RNA, 3 linear segments	8274	J02000
Cowpea mosaic virus	ss-RNA, 2 linear segments	9370	X00206
Bacteriophage			
MS2	ss-RNA, linear	3569	J02467
ϕX174	ss-DNA, circular	5386	J02482

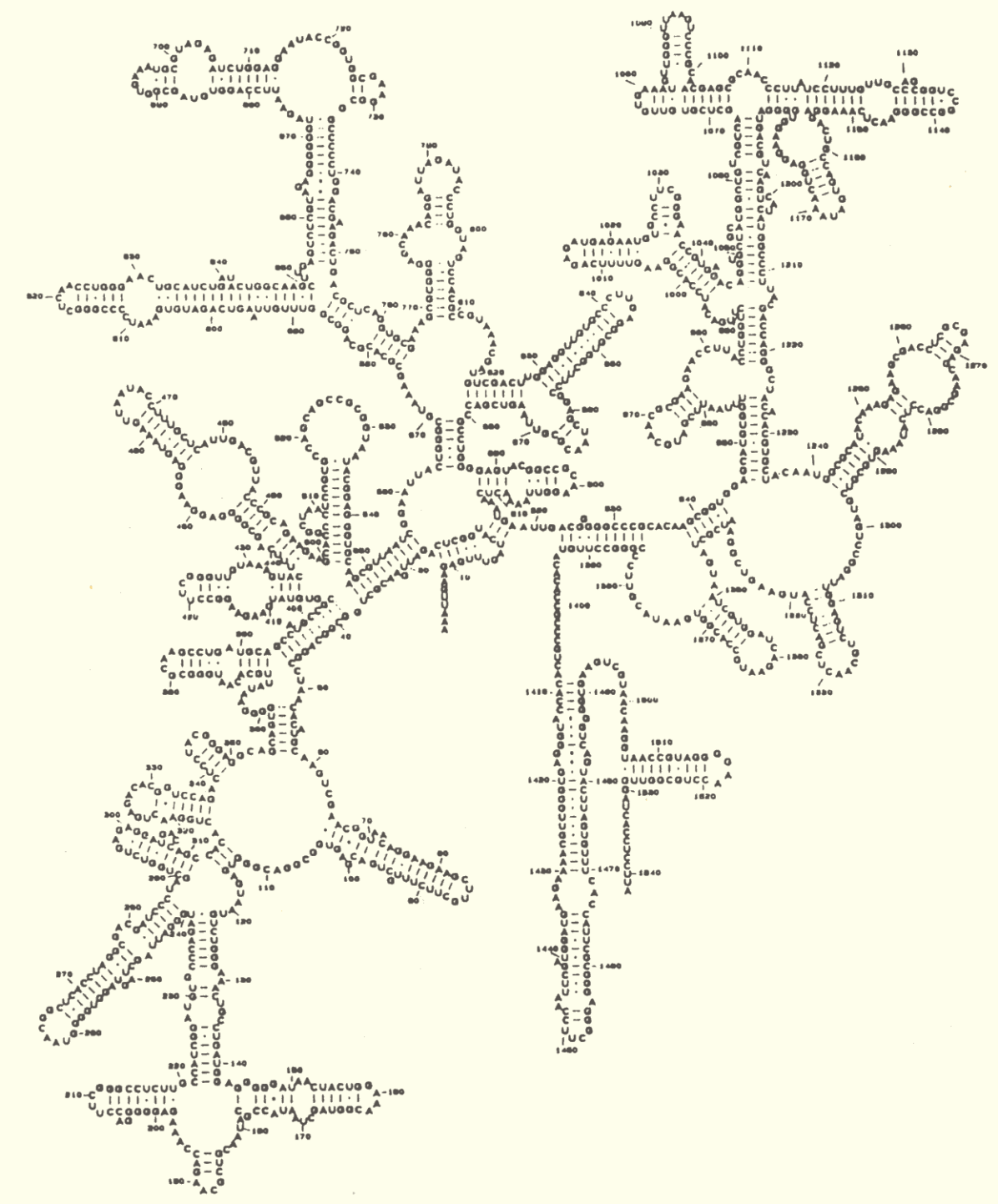


圖1. Secondary structure of E. coli 16S RNA

	U		C		A		G	
U	UUU UUC	Phenyl-alanine	UCU UCC UCA UCG	Serine	UAU UAC	Tyrosine	UGU UGC	Cysteine
	UUA UUG	Leucine (1)			UAA UAG	TERMINATE	UGA	TERMINATE
							UGG	Tryptophan
C	CUU CUC CUA CUG	Leucine (1)	CCU CCC CCA CCG	Proline	CAU CAC	Histidine	CGU CGC CGA CGG	Arginine (3)
					CAA CAG	Glutamine (2)		
A	AUU AUC AUA	Isoleucine	ACU ACC ACA ACG	Threonine	AAU AAC	Asparagine	AGU AGC	Serine
	AUG	Methionine			AAA AAG	Lysine	AGA AGG	Arginine (3)
G	GUU GUC GUA GUG	Valine	GCU GCC GCA GCG	Alanine	GAU GAC GAA GAG	Aspartic acid Glutamic (2) acid	GGU GGC GGA GGG	Glycine

表二. **Codons and their amino acids. The coodons are shown in their mRNA form.**

三. 序列排對 (Sequence Aligments):

序列間的比對是非常的重要, 因爲這有助於了解很多分子的進化、結構及功能。功能相同的巨分子們及不同生物類中具有同名稱的巨分子們通常皆具有某一程度不同的序列, 而其不同程度隨著進化的距離而加大。功能不同的序列們 (如 hemoglobin 及 myglobin) 常常是由同一祖先序列分差發展而來的, 在相關序列中其對應部位常具有相似的生化活動。

(3.1) 排對的數目:

令兩個序列各爲 $\underline{a} = a_1 a_2 \dots a_n$ 及 $\underline{b} = b_1 b_2 \dots b_m$, 其中 a_i 及 b_j 爲 A, G, C, T 中之一個 (當我們考慮 DNA 時); 或爲20種符號之一 (當考慮 protein 時)。因爲可解 $a_1 = b_1$, $a_2 = b_2$, $a_3 = b_3$, 但 $a_4 = b_5$, $a_5 = b_6$, $a_6 = a_7$, (即 $n = 6$, $m = 7$) 我們稱 $\underline{b}$ 爲由 $\underline{a}$ 插入 b_4 而來或 $\underline{a}$ 爲由 $\underline{b}$ 刪除 b_4 而來。如此在比對兩序列時我們插入空元素中於序列中使兩序列同長。即新的排對序列爲

$$a_1^* a_2^* \dots a_L^*$$
$$b_1^* b_2^* \dots b_L^*,$$

a_i^* 及 b_i^* 是由原來 $\underline{a}$ 及 $\underline{b}$ 插入多個 ϕ 而來; $n + m \geq L \geq \max\{n, m\}$。例子: $\underline{a} = ATAAGC$, $\underline{b} = AAAAACG$, 則排對之一如下:

$$\underline{a}^* = \phi ATAAGC\phi$$
$$\underline{b}^* = AAAAA\phi CG。$$

這樣排對的數目太多了 (通常 $n \geq m \geq 1000$)。因此要加入規範:$\binom{\phi}{\phi}$ 不允許發生, $\binom{C\phi}{\phi G}$ 及 $\binom{C\phi}{G\phi}$ 看成與 $\binom{C}{G}$ 一樣。如此可能排對的數目爲

$$g(n,n) = \binom{2n}{n} \approx a^{2n}(4\sqrt{n\pi})^{-1} \quad \text{as } n \to \infty。$$

當 $n = m = 1000$ 時, $g(100,100) \sim 10^{600}$, 太大不能一一硬是排對! 假如進一步限制在一定長度範圍內不許有插入發生, 只許有突變發生 (即 $\phi \neq a_i^* \neq b_i^* \neq \phi$), 則此數目可進一步縮小。

當考慮 k 序列排對時, 我們可知其可能排對數目會更大。令 $f_k(n)$ 爲 k 序列 (其長度爲 n) 之排對數目, 則可證明出 for $k \geq 2$,

$$\lim_{n\to\infty} \log f_k(n)/n = \log C_k,$$

where $C_k = (2^{1/k}-1)^{-k} \approx \frac{1}{\sqrt{2}}[\frac{k}{\log 2}]$。當 $n = 1000$, $k = 3$ 時, $f_3(1000) \approx 10^{1755}$, 太大了。

(3.2) 用動態規畫做兩序列之排對:

我們要比對兩序列之相似性首先要定義兩序列之距離函數D : $D(\underline{a},\underline{b}) = \min \sum_{i=1}^{L} d(a_i^*, b_i^*)$ 此處 minimum 是考慮所有可能的排對。雖然可能的排對數目很大, 但我們可用動態規畫 (Dynamic Programming) 來解決此問題。首先定義函數 d 如同成本函數:

$$d(a_i^*, b_i^*) = \begin{cases} 2 & \text{if } a_i^* = \phi \text{ or } b_i^* = \phi \\ 1 & \text{otherwise } (a_i^* = b_i^* = \phi \text{ 不允許發生})。 \end{cases}$$

則很容易證出如下定理:

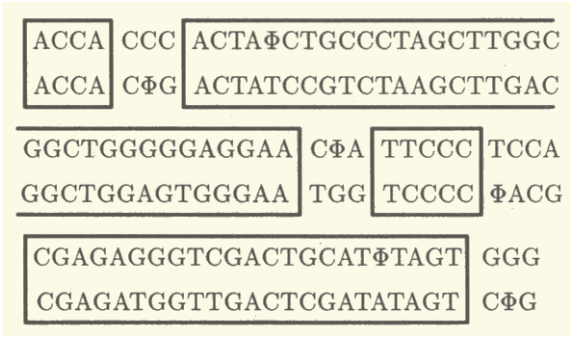
定理1: 若 $\underline{a} = a_1 a_2 \ldots a_n$ 且 $\underline{b} = b_1, b_2 \ldots b_m$, 定義 $D_{ij} = D(a_1 a_2 \ldots a_i,\ b_1 b_2 \ldots b_j)$。並設

$$D_{oo} = 0,\ D_{oj} = \sum_{k=1}^{j} d(\phi, b_k),$$
$$\text{且 } D_{io} = \sum_{k=1}^{i} d(a_k, \phi)。$$

則

$$D_{ij} = \min\{D_{i-j,j} + d(a_i,\phi),\ D_{i-1,j-1} + d(a_i,b_j),\ D_{i,j-1} + d(\phi,b_j)\}。$$

此即動態規畫的式子。當 $d(a_i,b_j) = 1$, $d(\phi,b_j) = d(a_i,\phi) = 2$ 時, 這演算法則具有計算複雜性 $O(nm)$。一個例子是, 我們排對兩個 Escherichia coli tRNA 序列: threonine tRNA(即 $\underline{a}$) 及 valine tRNA (即 $\underline{b}$)。利用定理1我們得到72個最佳排對, 其中之一顯示如下 (與72 最佳排對共通部份則用方格框起來):



定義距離函數 $D(\underline{a},\underline{b})$ 再設法重覆的找最小 D 是一常用法則。另一常用法則是定義相似函數 $S(\underline{a},\underline{b})$, $S(a_i,b_j) < 0$ 若 $a_i \neq b_j$ 且 $S(a_i,b_j) > 0$ 若 $a_i = b_j$。如此我們可把

定理1中之D及 d 全改成S而 minimum 則改成 maximnm (即找最大相似度)。

假如事先給定一短的模型 $\underline{a}$, 我們想找出在 $\underline{b}$ 中那些地方很近似 $\underline{a}$。此問題可寫成: 找出 i 及 j $(i \le j)$使

$$\begin{aligned} & S(\underline{a}, b_i b_{i+1} \dots b_{j-1} b_j) \\ = & \max_{k \le \ell} S(\underline{a}, b_k b_{k+1} \dots b_\ell)。 \end{aligned}$$

解法就是把原來的動態規畫中式子修改一下就行了。舉例如下: 在 E. coli promoter sequence 中, 模型 TATAAT 是已知具有顯著的特殊功能。故令 $\underline{a} = TATAAT$。$S(a,a) = 1$, $S(a,b) = -1$ for $a \ne b$, and $S(a,\phi) = -2$。令 $\underline{b}$ =E. coli promoter sequence。利用動態規畫我們可找出在 (6,43) 位置上有最佳排對:

TATAAT
CATGAT。

此處 CATGAT 在 $\underline{b}$ 內, 爲一 canonical -10 pattern。在 (6,13) 位置上有:

TATAAT
TCGAAT,

此處 TCGAAT 在 $\underline{b}$ 內, 其功能在研究中, 不過在排對上倒一樣好。

有些全然相似性很小的序列們, 竟然可發現一些令人驚奇的關係來。比如在 viral (病毒) 及 host DNA 中有令人出乎意料外的長段列非常的相似。從數學上講, 我們要找出 $a_i a_{i+1} \dots a_j$ 及 $b_k b_{k+1} \dots b_\ell$ 使

$$\max_{\substack{1 \le i < j \le n \\ 1 \le k < \ell \le m}} S(a_i a_{i+1} \dots a_j, b_k b_{k+1} \dots b_\ell)。$$

同樣地我們可用修飾後的動態規畫來解此問題。基本構想如下: 先定義

$$\begin{aligned} H_{ij} = & \max\{0, S(a_x a_{x+1} \dots a_i, b_y b_{y+1} \dots b_j) \\ & : 1 \le x \le i,\ 1 \le y \le j\}。 \end{aligned}$$

記錄x及y。令 $H_{io} = H_{oj} = 0$ for $1 \le i \le n$, $1 \le j \le m$。令 $S_{ok} = S(\phi, b_1 \dots b_k) = -\hat{g}_k$。則易證出

$$\begin{aligned} H_{ij} = & \max\{0, H_{i-1,j-1} + S(a_i, b_j), \\ & \max_{1 \le k \le i}\{H_{i-k,j} - \hat{g}_k\}, \\ & \max_{1 \le \ell \le j}\{H_{i,j-\ell} - \hat{g}_\ell\}\}。 \end{aligned}$$

如此從(i,j)位置倒回追查出 (i,j) 使

$$H_{ij} = \max_{\substack{1 \le k \le n \\ 1 \le \ell \le m}} H_{k\ell},$$

這時的 (i,j) 及其對應記錄下的 x, y 即告訴你最相似的片段。

有時有些未知的因素存在於序列中而使我們用數學方法找出的最佳排對並不正確地反映生化特性。因此我們可利用生化資料來正確估計相似分數, 或成本函數, 如此情況可改善一些。另一方法即找出所有接近最佳的排對來。此即找出所有排對而其對應的相似分數落在最佳相似分數的某一距離內。如此找出的排對中很可能某一個即合乎生化的解釋。解決此問題的演算法與以前類似只不過在計算過程中要記錄很多指標 (pointer) 及倒回追蹤 (traceback)。

利用 nucleotides 之間能量關係及動態規畫我們可以設法預估出 RNA 之 secondary structure。這問題牽涉到化學分子的物理特性, 有興趣的人可看參考資料。

(3.3) bb12 多重序列的排對:

多重序列的排對可 把二序列排對的動態規畫延伸下來加以解決之, 但其計算複雜度就成 $O(2^R n^R)$, R爲序列個數。通常n很大, $R \geq 3$ 時, 計算量太大, 不切實際。因此另一新的, 簡便方法就被提出來, 今說明如下: 令 $R = 6$及6序列排對如下:

$$\begin{aligned} Seq1 &\ \cdots A \cdots \\ Seq2 &\ \cdots A \cdots \\ Seq3 &\ \cdots T \cdots \\ Seq4 &\ \cdots A \cdots \\ Seq5 &\ \cdots \phi \cdots \\ Seq6 &\ \cdots C \cdots \end{aligned}$$

把每一縱行看成一 vector。然後計算此行 A 出現的次數除以6, C出現的次數除以6, G出現數的次數除以6, T出現次數除以6, 以及ϕ出現次數除以6。如此得到 $a = (p_A, p_C, p_G, p_T, p_\phi)$。我們要找出排對使測距

$$D = \sum_{\substack{\text{over} \\ \text{R-sequences}}} (p_A \log p_A + p_C \log p_C + p_G \log p_G + p_T \log p_T + p_\phi \log p_\phi)$$

最大。此演算法是相當複雜。令 $\underline{a} = a_1 \ldots a_n$, $\underline{b} = b_1 \ldots b_m$, $\cdots$, $\underline{r} = r_1 \ldots r_q$。則考慮 $a_1 \ldots a_i$; $b_1 \ldots b_j$; $\cdots$; $r_1 \ldots r_x$之排對, 其最後一行爲:

$$\begin{matrix} \epsilon_1 a_i \\ \epsilon_2 b_j \\ \vdots \\ \epsilon_R r_x \end{matrix}$$

此處 $\epsilon_i = 0$ 或 1, 且 $oa_i = \phi$。令距離函數 D 如上定義, 則可導出式子:

$$D_{ij\cdots x} = \max_{\underline{\epsilon} \neq \underline{o}} [D_{i-\epsilon_1, j-\epsilon_2, \ldots, x-\epsilon_R} + d(\epsilon_1 a_i, \epsilon_2 b_j, \ldots, \epsilon_R r_x)\}。$$

這式子的計算複雜度是 $O(2^R n^R)$, 記憶體 $O(n^R)$。

四. 統計分析 (Statistical Analysis):

統計分析有助於找出一些特異現象及模型, 比較兩序列之特性, 等。

(4.1) 兩序列中最長相配片段之長度估計:

首先考慮兩序列中最長相配片段的長度變化。此問題與丟銅板類似。假設連續丟一銅板, 其出現人頭的機率是 p。則連續出現人頭的最長次數 R_n 會滿足

$$P(\lim_{n\to\infty} R_n / \log_{1/p}(n) = 1) = 1。 \quad (4.1)$$

令假設兩序列的分佈情形一樣, 則 $a_i = b_i$ 時即表示出現"人頭", 其機率

$$p = p_A^2 + p_C^2 + p_G^2 + p_T^2。 \quad (4.2)$$

故最長相配片段的長度 R_n 滿足 (4.1) 式, 而其 p 值滿足 (4.2) 式。更複雜的結果可

證明出來。令兩序列長各爲 n 及 m 且 $\log(m)/\log(n) \to 1$。則其最長相配片段 (含 k 不相配單位) 具有長度 M 並滿足下式:

$$\begin{aligned} E(M) \approx\ & \log(qmn) + k\log\log(qmn) \\ & + k\log(q/p) - \log(k!) + r\log(e) \\ & - 0.5 \end{aligned}$$

此處E表期望值; $q = 1-p$; $\log \equiv \log_{1/p}$; $r = 0.577\cdots$ 爲 Euler constant; e 爲 exponential constant$= 2.718\cdots$。變異數

$Var(M(n,m)) \approx [\pi\log(e)]^2/6 + 1/12$。

(4.2) 頻率分析:

氨基酸 glutamine 可由 CAG 或 CAA 表示, 但由頻率分析中發現 CAG codon 在一些基因內發生的頻率非常高。利用 codon 偏好出現模式可減低 oligomers 用以確認蛋白質序列之次數。codon 偏好模式與 E. coli 及酵母之基因表示法有相關。codon 頻率與其對應之 tRNA 量多少有高度相關。在蛋白質編碼序列中 oligonucleotide 傾向於重覆出現其週期爲三。這種週期性在非編碼序列中不存在。

在 DNA 序列中, 四種 nucleotides 並不是平均出現其中, 其基本組合在序列內與序列對序列中皆有變異。利用近鄰居模式可幫助了解分子構造。在 eukaryotes 內 CG 一齊出現的次數非常少, 而在一般序列中 $PuPu$ 及 $PyPy$ 比 $PuPy$ 或 $PyPu$ 更偏好出現 (Pu 即 purine base (A or G), 而 Py pyrimidine base (C or T))。另一 uncleotide ordering 例子爲序列中 A 之聚合。在長序列 RNA 及 DNA 單一及重 A 出現次數比期望 (假設隨機出現) 的少, 另外長的連串 A (runs of A), 比如 AAA, 出現的次數比期望的更多。G 與 C 一齊出現的頻率很少, 但單一的 G 及 C, 及 GG, CC 出現次數的比期望的更多。長的 G 聚合與 C 聚合出現的次數期望的少很多。在單線 DNA (病毒內) 序列中其含迴語 (palindrom, 即倒起來讀也一樣之句, 如 eye, madam) 之區域比期望的少。在雙線核酸內, 迴語序列包含一雙摺對稱軸 (twofold axis of symmetry) 如 AGCT, 因此這種序列能對自己摺疊起來。迴語的少量出現應與單線 DNA 病毒的 secondary structure 的拘束 (constraints) 有關。

(4.3) 馬可夫分析:

我們可把一長序列 DNA/RNA 看成一馬可夫鏈 (Markov chain)。令 $A = 1$, $C = 2$, $G = 3$, $T = 4$。定義 order k 之馬可夫鏈爲 $\{X_t | t = 1, 1, \cdots, N\}$ 滿足

$$\begin{aligned} & P\{X_{n+1} = i_{n+1} | X_n = i_n, \ldots, \\ & \qquad X_{n-k+1} = i_{n-k+1}, \ldots, X_0 = i_0\} \\ = \ & P\{X_{n+1} = i_{n+1} | X_n = i_n, \ldots, \\ & \qquad X_{n-k+1} = i_{n-k+1}\} \quad \text{for all } n。 \end{aligned}$$

在均勻馬可夫鏈中轉移機率不隨時間變動。故有 $p_k = 3 \cdot 4^k$ 未知參數要估計 ($i_n = 1, 2, 3, 4$)。令 $n(i_1, i_2, \ldots, i_r)$ 爲轉移 $i_1 \to i_2 \to \ldots \to i_r$ 在一序列中觀察到的次數。則轉移機率之最大概似估計如下:

$$\hat{p}(i_1, i_2, \ldots, i_k; i_{k+1})$$

$$= \hat{P}\{X_{k+1} = i_{k+1} | X_k = i_k, \ldots, X_1 = i_1\}$$
$$= \frac{n(i_1, \ldots, i_k, i_{k+1})}{n(i_1, \ldots, i_k, +)},$$

此處 $n(i_1, i_2, \ldots, i_k, +) = \sum_{j=1}^{4} n(i_1, i_2, \ldots, i_k, j)$。由此我們可做一些統計檢定。我們可用貝氏訊息準則 (BIC) 來決定 order k。其方法如下: 先計算 log-likelihood $L(k)$

$$L(k) = \sum n(i_1, \ldots, i_k, i_{k+1}) \log \hat{p}(i_1, \ldots, i_k; i_{k+1}),$$

此處加號 $\sum$ 是對所有 $n(i_1, \ldots, i_k, i_{k+1}) > 0$ 之可能 $i_1, i_2, \ldots, i_{k+1}$。則 $BIC(k) = -2L(k) + p_k \log n$, 此處 $n(\leq N)$ 是算 $n(\cdot)$ 時子序行的數目。要找的 k 即是使 $BIC(k)$ 最小的 k。

通常因爲 DNA 序列的線性結構有不均勻性要找一長而均勻的序列是非常難。而要做 high order 馬可夫分析卻需要很長的序列 (因爲 p_k 很大), 這便成爲不切實際。要減低 p_k 參數數目, Raftery 提出一模型:

$$p(i_1, \ldots, i_k; i_{k+1}) = \sum_{j=1}^{k} \lambda_j q(i_j, i_{k+1})$$

此處 $Q = \{q(i,j), 1 \leq i, j \leq 4\}$ 是一 4×4 row-stochastic matrix 需估計, $\lambda_1, \ldots, \lambda_k$ 未知參數且其和爲一。所以要估計的參數個數減成爲 $p_k = 11 + k$, 比 3×4^k 少了太多。代價是需用非線性最佳化法來從事參數估計。

馬可夫分析之例子如下: bacteriophage 入序列其長爲 48502 nucleotides。此序列分成 5 區域, 各有其生化意義, 叫 Late, Early 1, Early 2, Control, Silent。馬可夫分析結果是 Late 區域爲 order 2, 其他爲 order 1。

我們可進一步用 Fourier Transform, Walsh Transform, 及 Correlation coefficient 等來找出週期性之模型。這時每一氨基酸需適當的給一分數。在用機率模型分析時我們很難找出明顯結構, 但用上述方法則易找出固定明顯的結構。

(4.4) 統計上顯著的模型:

給定一序列我們可從它取出一些統計量, 但如何確定這些統計量的'顯著'與否? 這就需要先建立理論模型用以爲準則。通常皆假設'隨機模型'而拿來比較。此模型又可分爲'獨立隨機'及'馬可夫依賴隨機模型'。S. Karlin 以此對 DNA 序列做了很詳細研究。另一準則爲資料混亂方法 (Data Shuffling Methods)。比如把一序列中有 A 及 G 的位置上交換排列而把 C 及 T 的位置不動。通常我們考慮 100 至 500 permutations, 統計量可從每一次 permuted data 算出。如果原來的統計量 (沒經 permutation) 比新得的一堆統計量還偏離極端, 則原來的統計量即被視爲顯著。

今定義一 k-word 爲在一序列中的連續 k 個字母。我們看一下 k-word 重複出現的頻率。令 $f_k(\nu)$ 爲那些出現 ν 次之 k-words 的個數。則 $N_k^* = \sum_{\nu \leq 2} f_k(\nu)$ 爲不同重現 k-words 的個數。有關之統計量如平均值、變異數、範圍、偏差等皆在 karlin 的研究報告。今舉例如下, 有三個 papovavirus genomes: Simian virus-40 (SV-40, 長= 5243), polyoma (長= 5293) 及

human BKV Dunlop strain (長= 5153), 其 $f_k(\nu)$, $k = 6, 8$之分佈見表三。由此表知 SV-40 及 BKV 之分很接近, 這表示他們之 genomes 是相似。而 Polyoma 在長的及高次重現的 k-words 中有較低的個數。這可區分出它與 SV-40 及 BKV 之不同。這些比較對所有 k, $4 \le k \le 15$ 皆成立。

表三. **REPEAT-OCCURRENCE DISTRIBUTION OF OLIGONUCLEOTIDES (WORDS) OF LENGTH 6 AND 8**

$f_6(\nu)$

ν	1	2	3	4	5	6	7	8	9	10	11	12	13	14	18
BKV-Dun	898	628	361	175	87	57	30	14	7	1	0	0	2	0	1
SV-40	943	627	363	183	90	47	29	11	7	5	1	3	1	2	0
Polyoma	1174	740	421	198	78	22	5	2	1	0	0	0	0	0	0
Random Polyoma	1391	899	423	132	42	11	3	1	0	0	0	0	0	0	0

$f_8(\nu)$

ν	1	2	3	4	5	6
BKV-Dun	4158	383	57	12	2	0
SV-40	4198	434	46	7	1	1
Polyoma	4709	269	11	3	0	0

Note: The entries indicate the number of distinct k-words that occurred exactly ν-times for the specified sequence. The row labeled random is obtained from the pointwise minimum of the cumulative distribution derived from 10 random permutations of the polyoma sequence.

定義 L_r 爲重現至少 r 次的最長字的長度。即

$$L_r = \max\{k : \sum_{\nu=r}^{\infty} f_k(\nu) \ge 1\}。$$

在隨機序列的情形下, L_r 之漸近分配可以導出來, 我們可以此來判斷在一般序列下所求之 L_r 是否顯著 (即夠長)。在 E. coli phages, T7 phage 顯示有一組顯著的長的重現字, 而入 phage 顯示無顯著的長的重現字。

今考慮有條件相似結構。由於 DNA 結合成對時的互補現象, 我們考慮對映 Δ^C : $A \to T, T \to A, G \to C, C \to G$; 及對映 $\Pi_k^{(I)}$: inverse permutation on k-words, 即把第 i 字母移到第 $k+1-i$ 位置。由 Δ^C 及 $\Pi_k^{(I)}$ 可定義二價關係 (dyad relation)。比如兩 5-word: ATTCG 及 CGAAT 有二價關係。令 $g_{Dk}(\nu, \mu)$ 爲長度 k 之 k-words pair (W, W^*) 之個數並滿足下列條件: W 及 W^* 具有二價關係並且 W 出現 ν 次而 W^* 出現 μ 次。一例子如表四。$k = 8$, 序列爲 SV-40, polyoma, HPV 及 BPV(human and bovine papilloma viruses)。對 polyoma 及 SV-40 而言 g_{D8} 之個數分析顯示其變異性差不多。但 Polyoma 次數較少。這些次數與 g_{D8} permutation range 指出對這兩 viruses $g_{D8}(1,1)$ 及 $g_{D8}(1,2)$ 已超出混亂集合 (shuffled sets 共經任意 30 permutations) 的最大值。這情形對 SV-40 內之所有 $g_{D8}(i,j)$, $i = 1,2,3$; $j = 1, \cdots, 6$ 一樣清楚。這表示這些統計量很顯著。HPV 及 BPV 在表四中很接近。這指出兩者間仔細生化關係需經進一步研究。

表四. COUNT OCCURRENCE DISTRIBUTION OF $g_{D,8}$ FOR SV-40, POLYOMA, HPV, AND BPV COMPARED WITH $g_{D,8}$ FOR CORRESPONDING PERMITES SEQUENCES[a]

SV40						
	1	2	3	4	5	6
1	276	68	9	3	0	1
	(195,260)	(0,25)	(0,3)			
2		15	9	1	0	0
		(0,4)	(0,1)			
3			0	0	0	0
			(0,0)			

Polyoma			
	1	2	3
1	248	36	0
	(192,229)	(0,16)	(0,3)
2		2	0
		(0,2)	(0,0)
3			0
			(0,0)

HPV				
	1	2	3	4
1	478	114	11	1
	(341,407)	(40,92)	(0,7)	(0,2)
2		12	1	0
		(0,6)	(0,1)	(0,0)
3			2	0
			(0,0)	(0,0)

	BPV				
	1	2	3	4	5
1	486	115	10	0	0
	(327,418)	(43,101)	(0,8)	(0,1)	
2		17	2	0	1
		(0,8)	(0,1)	(0,0)	
3			3	0	0
			(0,0)	(0,0)	

[a] The range of counts for 30 permutations are shown in parentheses for each (i, j).

五.DNA 之幾何及拓撲結構:

過去十多年中有關封閉性圈形 DNA 的幾何及拓撲的研究發展成一很重要的領域。主要理由是封閉性圈形 DNA 與斷了一線或雙線的 DNA 在物理及化學性質上有基本之不同。這些性質直到最近才可解釋, 理由在 (1) 這些 DNA 之各線鏈環起來而具有一鏈數 (linking number) Lk, (2) Lk 有二基本特性: (a) Lk 在 DNA 結構連續變形下具有不變性 (b) Lk 是由兩幾何量加起來, twist (旋數) Tw, 及 writhing (捲數) Wr, 即

$$Lk = Tw + Wr。$$

這兩特性可應用在許多場合, 比如鏈結缺陷及超級盤繞之分析, topisomerases 之各型酵素性質的分析等。

(5.1) 兩空間封閉曲線之鏈數Lk:

令 C_1 及 C_2 是兩連續有向封閉曲線 (在3度空間中)。今把它投影到平面上, 其中之一會棋過另一個交於數點。比如在圖2中有二例, 一有兩交點, 而另一有四交點。對每一交點可指定一指數 $+1$ 或 -1, 此數是從看上面曲線的切向量 (tangent vector) 必須轉至下面曲線的切向量之方向而定, 如果轉向是順時針, 即定爲 -1, 否則爲 $+1$。在此鏈數 Lk 即是把這些指數全加起來 (對所有交點) 再除以2, 可用 $Lk(C_1, C_2)$ 表示之。如此, 圖一中之左邊 $Lk = [-1 + (-1)]/2 = -1$, 而在右邊 $Lk = [(-1) + (-1) + (-1) + (-1)]/2 = -2$。一個有趣例子是在圖3, 每一交點具有 $+1$, 故其 $Lk = +4$。這例子描繪出兩螺旋型封閉曲線互相旋繞 (依右手法則 in right-handed sense) 爲一 DNA 之極佳模式。圖4是一很奇怪的例子, 有兩 $+1$ 交點及兩 -1 交點, 因此 $Lk = 0$。圖中央之交點不算因爲是一曲線對自己的交點, 但即使是 $Lk = 0$ 這兩曲線不能分開。這例子常稱爲 trapped figure 8。

鏈數是有四個主要特性: (1) 此數與那一投影面 (用以計算它) 無關, 這點很重要因爲 Lk 是空間的特性而不是投影面的; (2) 假如任一曲線連續地變形但不斷裂, 則 Lk 是不變 (參見圖 5); (3) 假如某一曲線之方向倒反過來則 Lk 之符號就改變, 如圖6所示; (4) 假如此對曲線經由鏡面反映, 則 Lk 也是改變符號, 如圖7所示。

只外 Lk 也可用空間平面曲線交接方法與 Gauss 積分方法來定義, 有興趣的人請見參考資料。

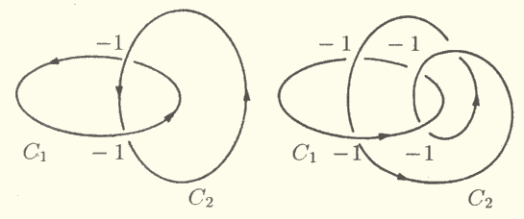


圖2. Linking numbers of pairs of curves using the index approach.

圖3. Linking of helically intertwined curves with a circular axis.

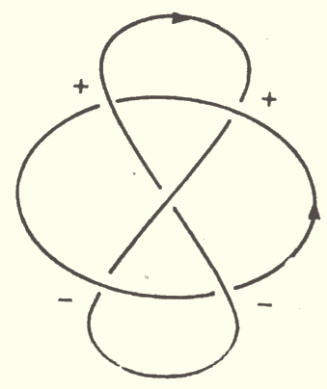


圖4. The trapped figure 8:a pair of curves with $Lk = 0$.

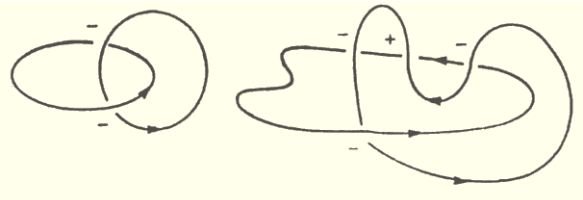


圖5. The invariance of Lk under deformation.

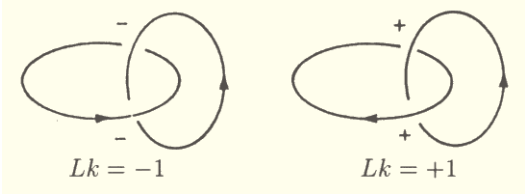


圖6. The reversal of the sign of Lk when one of the curves is reversed in orientation.

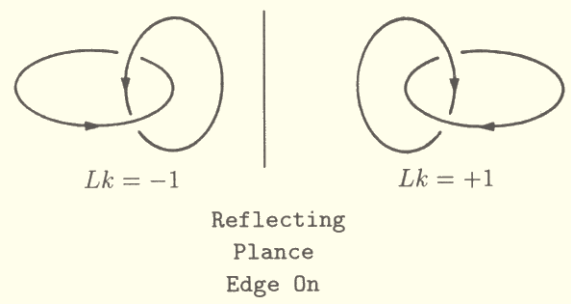


圖7. The reversal of the sign of Lk of two curves when refleeted in a plane.

(5.2) 空間封閉曲線的捲數(Writhing):

空間一封閉曲線 C 投影到平面上, 這曲線也許會自己橫過自己數次。對每一橫過的點, 指定一指數 $+1$ 或 -1, 如以前 Lk 之說明。把這些指數全加起來即是 C 的有向捲數 (directed writhing number), 方向是由於投影所致。眞正的捲數, 以 $Wr(C)$ 或 Wr 表示, 是定義成所有可能投影的有向捲數的平均值。如此在圖 8 中兩曲線對幾乎所有投影皆有一交點, 但其指數分別爲 -1 及 $+1$, 因此 $Wr = -1$ 及 $+1$。

與 Lk 不同, 如果 C 的方向改了, 有向捲數並不改變, 因爲對每一交點兩個交會線段之方向皆改了, 如此 Wr 並沒改變。更進一步分析, 不像 Lk, 當 C 變形時 Wr 是確實改變了。比如在圖8中, 任何不打結的曲

線皆可變形成一圓圈, 此時 $Wr = 0$。圖9是多餘的例子用來顯示不同的 Wr。

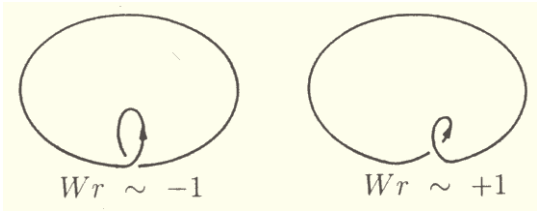


圖8. The writhing number of curves with one coil.

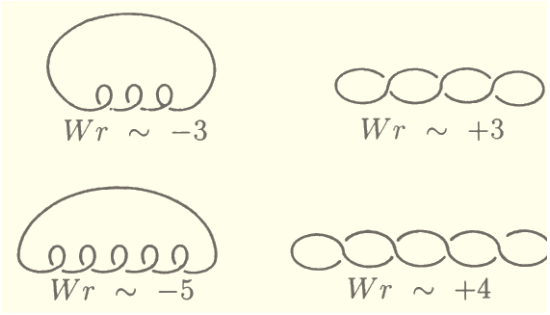


圖9. The writhing number of curves with multiple coils.

(5.3) 一曲線對另一曲線的旋數(Twist):

基本上, 一曲線 C_2 對另一曲線 C_1 的旋數, 以 $Tw(C_2, C_1)$ 或 Tw 表示之, 是測量 C_2 繞著 C_1 的旋轉大小。最簡單的例子是一螺旋線 (helix) 對自己軸旋轉, Tw 是此螺旋線對自己軸環繞的次數。此數是正的, 假如螺旋線是右手旋, 否則爲負即爲左手旋。一些例子示於圖10中。

旋數可用一向量從 C_1 指到 C_2 的旋轉來測量。在圖11中, 向量旋轉了 2π 角度, 而旋數即總轉角除以 2π。對一螺旋型曲線其繞另一封閉圓圈轉, 則旋數也可類似用向量轉角來定義。一例如圖12。但是對更一般化的例子中, 比如 C_1 不是直線或在一平面上, 旋數的定義就更複雜了, 因爲幾何觀念不是那麼顯然了。這需要 C_1 與 C_2 間向量所組成的帶狀曲面 (如圖11, 12), 叫對應曲面, 來解釋。我們假設此曲面在接近 C_1 地方可微分 (或平滑), 此即在 C_1 上之每點皆有一切平面與對應曲面相切。在圖13中一部份對應曲面給標示出來。令 T 是 C_1 在 x 點的單位切向量。令 V 是與 T 垂直的單位相量 (定在 x 點) 但與曲面相切而指向 C_2 方向。如此, T 及 V 是互相垂直而支撐在 x 點的切平面。然後其橫積 $T \times V$ 是與曲面垂直 (在 x 點上), 而旋數 Tw 即是定義成 V 在 $T \times V$ 方向上的總改變量當 x 沿著整條 C_1 上移動時。Tw 即定義如下:

$$Tw = Tw(C_2, C_1) = \frac{1}{2\pi} \int_{C_1} (T \times V) \cdot dV。$$

通常旋數並不爲整數, 而常常隨 C_1 或對應曲面的變形而改變。更進一步說, $Tw(C_2, C_1)$ 不必要等於$Tw(C_1, C_2)$。今舉例如下:

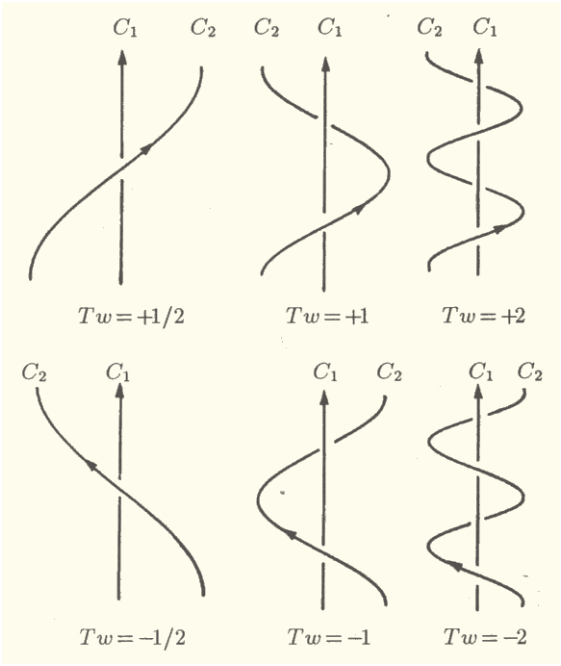


圖10. The twist of helices and about a linear axis.

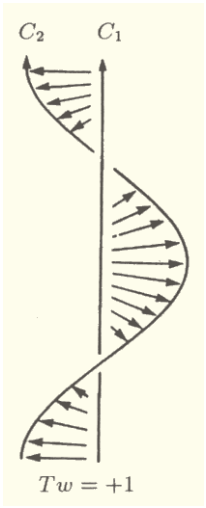


圖11. The spinning arrow approach to the twist of a helix with linear axis.

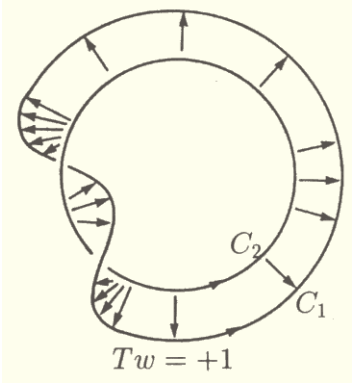


圖12. The spinning arrow approach to the twist of a helical curve with circular axis.

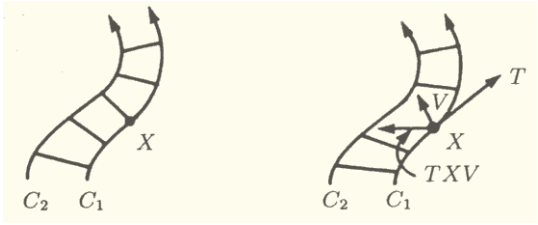


圖13. A is a graphical representation of a correspondence surface. B pictures the spanning vectors T and V to the correspondence surface at a point x of C_1.

(5.3.1) 具有直線軸的螺旋線: 令 C_2 是螺旋線, 其圓半徑爲 r, 主軸 C_1 是一直線其長爲 L (見圖14)。可假設 C_1 是在 z - 軸上, 如此

曲線C_2 : $y(s) = (r\cos(\alpha s),\ r\sin(\alpha s),\ p\alpha s)$,

此處 $y(s)$ 是從原點指向 C_2 上一點之向量, $\alpha = 2\pi\, n/L$, s 是沿主軸量的長度, $0 \le s \le L$。當 s 從 0 走到 L, 這螺旋線繞著 C_1 軸 n 次 with pitch $2\pi p$, 即 $2\pi pn = L$。要算 $Tw(C_2, C_1)$ 就先要建立起對應曲面。先由 C_1 上一點 $(0,\ 0,\ p\,\alpha\, s)$ 指向 C_2 上對應點 $(r\cos(\alpha s),\ r\sin(\alpha s),\ p\alpha s)$, 此向量即爲 $(r\cos(\alpha s),\ r\sin(\alpha s), 0)$ 叫對應向量。這些對應向量產生出對應曲面。從圖 14 上, 知 $T = (0, 01)$, 且 $V = (\cos(\alpha s),\ \sin(\alpha s), 0)$。所以 $T \times V = (-\sin(\alpha s),\ \cos(\alpha s), 0)$。最後

$$\frac{dV}{ds} = (-\alpha\sin(\alpha s),\ \alpha\cos(\alpha s), 0)$$

因此

$$(T \times V) \cdot \frac{dV}{ds} = \alpha。$$

如此

$$\begin{aligned} Tw(C_2, C_1) &= \frac{1}{2\pi}\int_{C_1} (T \times V)\cdot dV \\ &= \frac{1}{2\pi}\int_0^L (T \times V)\cdot \frac{dV}{ds} ds \\ &= \frac{1}{2\pi}\int_0^L \alpha ds = \frac{n}{L}\int_0^L ds = n。 \end{aligned}$$

特別興趣的是計算 $Tw(C_1, C_2)$, 經一些計算 (原理同上) 可得 $Tw(C_1, C_2) = \frac{np}{(r^2+p^2)^{1/2}}$, 不同於 $Tw(C_2, C_1)$。

(5.3.2) 對稱螺旋線: 上面的分析(叫半帶狀模式) 可用於全帶狀模式, 即對稱螺旋線。這可把螺旋線 C_2 對其軸 C_1 做鏡面反映,

得新的螺旋線 C_2', 而 C_2' 對 C_2 之旋數, $Tw(C_2', C_2)$ 是一樣等於 $Tw(C_1, C_2)$ 因爲單位向量 V 是在此二情形下一樣的 (見圖 15)。如此

$$Tw(C_2', C_2) = \frac{np}{(r^2 + p^2)^{1/2}}。 \qquad (5.1)$$

此即一螺旋線繞著對稱的相對伴線 n 次, 然而其旋數並不爲 n; 螺旋線對其軸之旋數是與他們之間的幾何架構有關。如果 pitch 低 (即 p 接近 0) 則 $Tw(C_2', C_2)$ 很小。反過來, 如 p 很大 (與 r 比較) 則 $Tw(C_2', C_2)$ 趨近於 n。

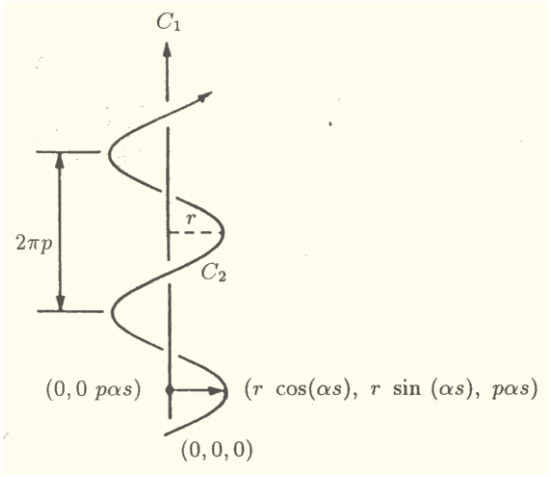


圖14. A circular helix of radius r and pitch $2\pi p$ about a linear axis.

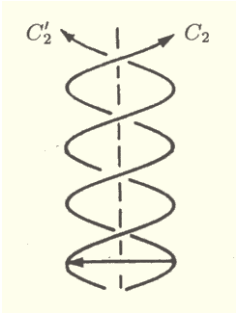


圖15. Symmetric solenoidal or circular or circular helixes about a linear axis.

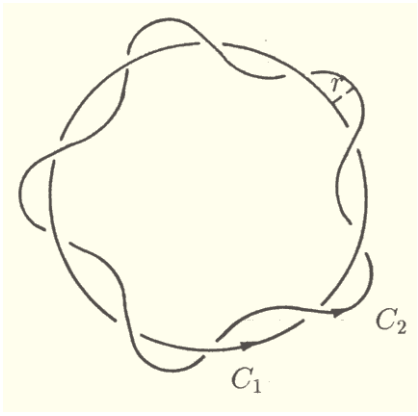


圖16. A helical curve wrapping around a torus of radius r whose axis is a circle.

(5.3.3) 具有封閉圓形軸之螺旋線: 令 C_2 是一右手環繞一封閉圓形軸 C_1 的螺旋線 (見圖 16)。C_1 的半徑是 R, 因此曲率 $k = 1/R$, 長度 $L = 2\pi R$。想像一內輪胎, 其半徑 $r < R$ 且以 C_1 爲中心軸。曲線 C_2 是一螺旋線, uniform pitch 於內胎上。假設 C_2 繞 C_1 n 次。即 $Lk = (C_2, C_1) = n$。又直接的計算可得

$$Tw(C_2, C_1) = n。$$

在此, 一對應向量即爲從 C_1 之一點 x 到 C_2 上之一點, 且落在 C_1 在 x 之向心向量與垂直 C_1 平面之向量所組成的空間上。此倒可看成 (5.3.1) 中把直線軸彎成一圓的情況來研究。

反過來要算 $Tw(C_1, C_2)$ 就很困難, 由定義可得

$$Tw(C_1, C_2) = \frac{1}{2\pi}\int_0^{2\pi n}\{(1 - rk\cos\theta)^2 + (rnk)^2\}^{\frac{1}{2}}d\theta。$$

這是一困難的橢圓積分, 要查表才行。但當 $rk \ll 1$ 時 (即內胎之半徑比起 C_1 之半徑小的太多), $rk\cos\theta$ 可忽略而得

$$\begin{aligned} Tw(C_1, C_2) &\approx \frac{1}{2\pi}\int_0^{2\pi n}(1+(rnk)^2)^{1/2}d\theta \\ &= n(1+rnk)^2)^{11/2}。 \end{aligned}$$

這解答與 $Tw(C_2, C_1) = n$ 成很明顯對比。假如定義 pitch 爲 $2\pi p$, 即 $2\pi pn = L = 2\pi R = 2\pi/k$。則

$$Tw(C_1, C_2) \approx \frac{np}{(r^2+p^2)^{1/2}}。$$

此與前面結果一樣。

如果 $rk < 1$ 但不非常小, 則

$$Tw(C_1, C_2) = (\frac{2ng}{\pi\mu})E(\epsilon),$$

此處

$$\begin{aligned} \frac{g}{\mu} &= ((r^2k^2(n^2+1)+1)^2 - 4r^2k^2)^{1/2}, \\ \epsilon^2 &= \frac{1}{2}(1-((n^2-1)r^2k^2+1) \\ &\quad ((n^2+1)r^2k^2+1)^2 - 4r^2k^2)^{1/2}, \end{aligned}$$

及 $E(\epsilon)$ 是皆知的 complete elliptical integral。這是可怕的表示法, 但在應用上卻非常有用。

如果C_2是左手旋轉, 則$Lk(C_2, C_1) = -n$, 且 $Tw(C_2, C_1) = -Tw(C_2, C_1)$, $Tw(C_1, C_2) = -Tw(C_1, C_2)$。

(5.4) 基本公式$Lk = Tw + Tr$:

在一些特殊情況下三個量 Lk, Tw, Tr 是由基本公式連在一起, 而就是此公式可在 DNA 分析中有很大應用。最主要興趣的情況是 C_1 與 C_2 爲兩有向曲線其包住一帶狀對應曲面, 而此曲面亦假設爲可微分 (在 C_1 上; C_1 爲邊界)。在此情況下這三個量滿足下面基本公式:

$$Lk(C_2, C_1) = Tw(C_2, C_1) + Wr(C_1)。$$

一些例子如圖17中, (a) 及 (b) 之 $Lk = Tw = Wr = 0$ 而 (c) 中 $Lk = Tw = +1$, $Wr(C_1) = 0$。

這公式的一重要結果而必須知道的是如下: 雖然 $Lk(C_2, C_1)$ 是拓撲上不變, 但 $Tw(C_2,\ C_1)$ 及 $Wr(C_1)$ 卻不是, 而在變形時會變。因此只要 C_1, C_2 不斷裂, 則在 Tw 有任何改變時, 在 Wr 也一定有對等改變, 只是符號相反, 大小卻一樣。比如在圖 18 中, 如果我們舉起帶子之上端, 則 $Wr(C_1)$ 開始減少而 $Tw(C_2, C_1)$ 開始增加, 然後最後變形爲圖 17(c) 中的樣子。在這過程中 $Lk(C_2, C_1) = +1$ 而 $Wr(C_1)$ 從 $+1$ 變成 0, $Tw(C_2, C_1)$ 從 0 變成 $+1$ (大約)。

這公式第二應用如下: 假設在圖 17(a) 帶子中, 其邊界線 C_2 有一裂口。假如此裂口補了, 則 $Lk = Tw = Wr = 0$。然而假如斷的一邊沿 C_1 轉了一圈 (右手旋) 然後再補回, 則 Tw 從 0 變成 1; Lk 也一樣。因爲 C_1 不變, 故 $Wr(C_1)$ 一樣爲0。如此本公式確定無誤。最後產品如圖 17(c)。

第三應用如下: 假如在圖 18 中帶子上頭環之整個帶狀曲面裂開並假設底下環經由此裂洞而往上伸成爲上頭環, 然後裂洞又補好了。則在此過程 $Wr(C_1)$ 從 $+1$ (大約) 變成 -1 (大約), 而 Tw 基本上不變, 且 Lk 是減少了2。如此本公式又一次確定無誤。

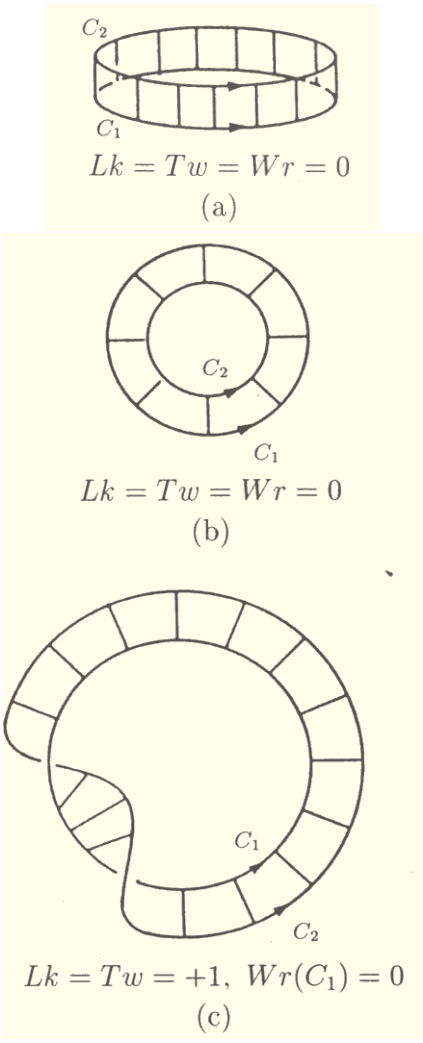


圖17. Pictorial examples demonstrating $Lk = Tw + Wr$ for ribbon models.

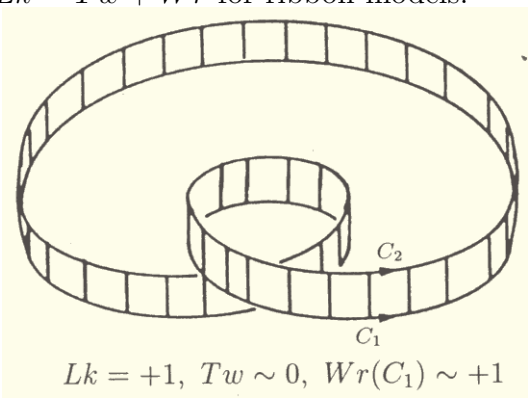


圖18. A ribbon model with $Lk = +1$, $Tw \sim 0,\ Wr \sim +1$.

(5.5) 應用於DNA分析:

在討論 Lk, Tw, Wr 之應用時, 有兩模式要用: (1) 叫全帶模式, (2) 叫半帶模式。全帶模式如前面討論之 C_1, C_2 例子。在 phosphodiester-sugar backbone chains 中, 其符號是 C 線中一 base 連到互補線 W 中之 complementary base (如圖15)。因此 Lk 寫成 $Lk(W, C)$, Tw 寫成 $Tw(W, C)$, Wr 寫成 $Wr(C)$。

半帶模式定義如下: 在 Crick-Watson 模式中描寫整個雙線螺旋結構, 其有一軸, 叫 A。在簡單線狀模式中, A 是一直線而由 C 轉成螺旋形。在圓形模式中, A 是一圓圈而 C 來轉。半帶模式即是前節所述之 C_1 及 C_2 分別看成 A 及 C (或 W)。參見圖14, 16。Tw 寫成 $Tw(C, A)$, Wr 寫成 $Wr(A)$。如果 DNA 是封閉圓形則 Lk 寫成 $Lk(C, A)$。

首先討論線性 DNA。在 Crick-Watson DNA 中, 如 C 轉 A n 次, $Tw(C, A) = n$。因此如 DNA 是長度 B bp(base-pair), 它的 $Tw = B/10.5$, 因爲要 10.5bp 來繞一圈。要算 W 線對 C 線的旋數 $Tw(W, C)$ 我們可用公式 (5.1) 來算即 $Tw(W, C) = np/(r^2 + p^2)^{1/2}$。從古典結構上, $2\pi p = 3.36nm$, $r = 1.0nm$。即 $p = 0.54nm$。如此

$$Tw(W, C) = \frac{.54n}{(1 + .54^2)^{1/2}} = 0.47n。$$

如果 DNA 長爲 B bp, $Tw(W, C) = \frac{0.47B}{10.5}$。特別有興趣的是一線對另一線的 Tw 是大約一線對軸的 Tw 之一半。在文獻中, $p/(r^2p^2)^{1/2}$ 常以 $\sin\alpha$ 表示, α 爲螺旋之

pitch angle。這例子中 $\sin\alpha = 0.47$ 即 $\alpha \approx 28°$。

其次討論封閉圓形 DNA (其軸爲圓形)。假如 DNA C 線繞 A n 次, 則 $Tw(C,A) = n$。又因 A 是平面圓, 故 $Wr(A) = 0$, 如此 $Lk(C,A) = n$(由基本公式)。即 $Lk(C,A) = Tw(C,A)$。在此情形 Lk常以 Lk_o 表示之。在此, 轉一圈約需 10.5bp 。此即如果封閉圓形 DNA (長度爲 B bp) 是在鬆弛狀態, 則 $Lk_o = B/10.5$。比如一 DNA 有 2100bp 長, 那 $Lk_o = 200$。今要算 $Tw(W,C)$, 令 R 是 C 的半徑; r 是 DNA 之半徑。假設 $r \ll R$, 由前節討論知 $Tw(W,C) = np/(r^2+p^2)^{1/2}$, $p = R/n$。對大多數 DNA 言, $r \ll R$, 因爲大多 DNA 非常長。這情形即簡化成線性情況。可易看出 $Tw(W,C) = 0.47n$, 又因 $Lk(C,A) = Lk(W,C) = n$, 因此 $Wr(C) = n - 0.47n = 0.53n$。如此對每一旋轉一圈的 DNA 片段, 其對 Tw 之貢獻是 0.47 而對 Wr 是 0.53。如果 $r < R$ 但不太小則必須用前面討論過的複雜積分式子去算 Tw。一旦算出, 可用 $Wr = n - Tw$ 算出 Wr 來。因爲 DNA 大多很長, 這種情況很少碰到。

其他應用很多, 比如一螺旋線繞一螺線的 DNA 分析以及酵素的活動 (enzymatic activity, 能改變 DNA 之結構使 Lk 改變)。有興趣的人請參見考資料。

六. 結論:

DNA 的解讀造成了生物基因的革命時代來臨。幸好我們有功能強大的電腦, 這些大量的 DNA 資料得以儲存並讓大家快速存取分析。但是最根本的一些重大問題還未解, 比如那些 DNA 模式對細胞病變或細胞功能有決定性關係。因此研究上需數學家、統計學家等大力幫忙分析。最近有一派人設法用人造神經網路來解釋分析一些問題, 但顯然還未成熟。一切有待努力。相信近十年內會有重大突破, 也許人類眞能控制基因變化, 造出新的人類來, 到時整個世界就全變了。

參考資料

1. M. S. Waterman, Ed., "Mathematical Methods for DNA Sequences", CRC Press, Boca Raton, Florida, 1989.
2. R. Doolittle, Ed., "Moleculer Evolution: Computer Analysis of Protein and Nucleic Acid Sequences", Methods in Enzymology Vol. 183, Academic Press, 1990.
3. Nucleic Acids Research, 期刊。
4. J. Molecular Biology, 期刊。
5. Annual Review on Biochemistry, 期刊。

—本文作者曾任職於中央研究院資訊研究所—