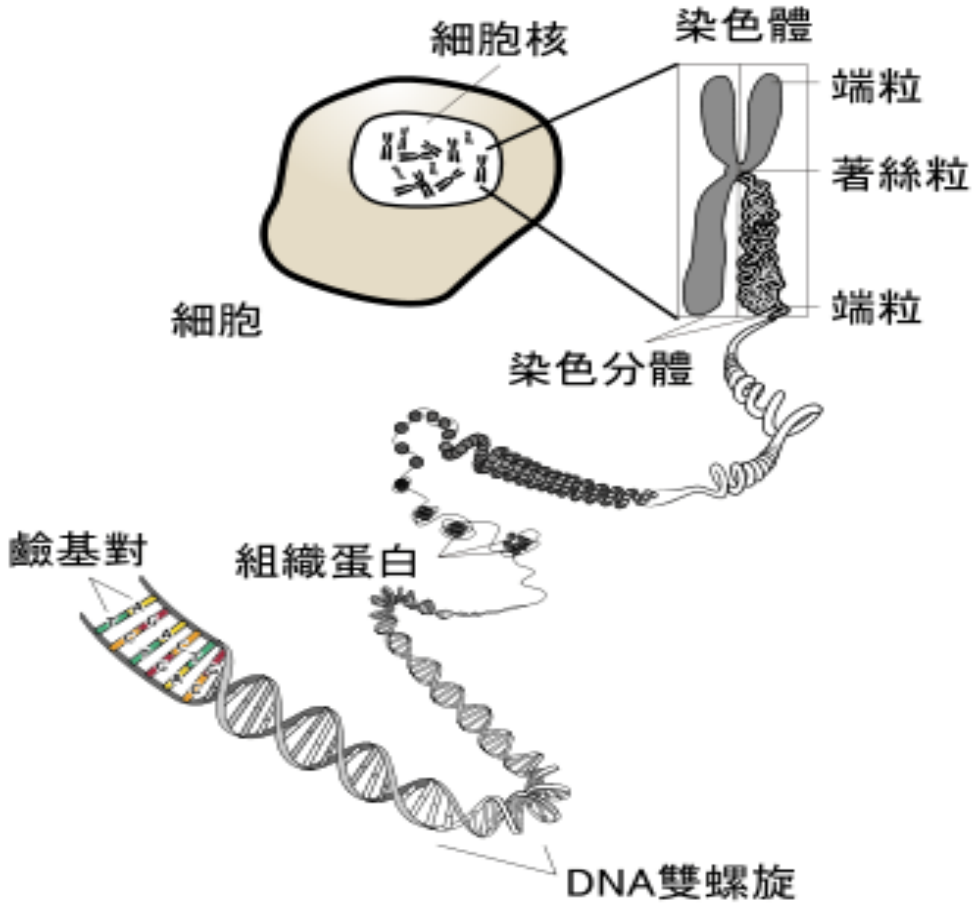
Mathematical Models of Molecular Evolution

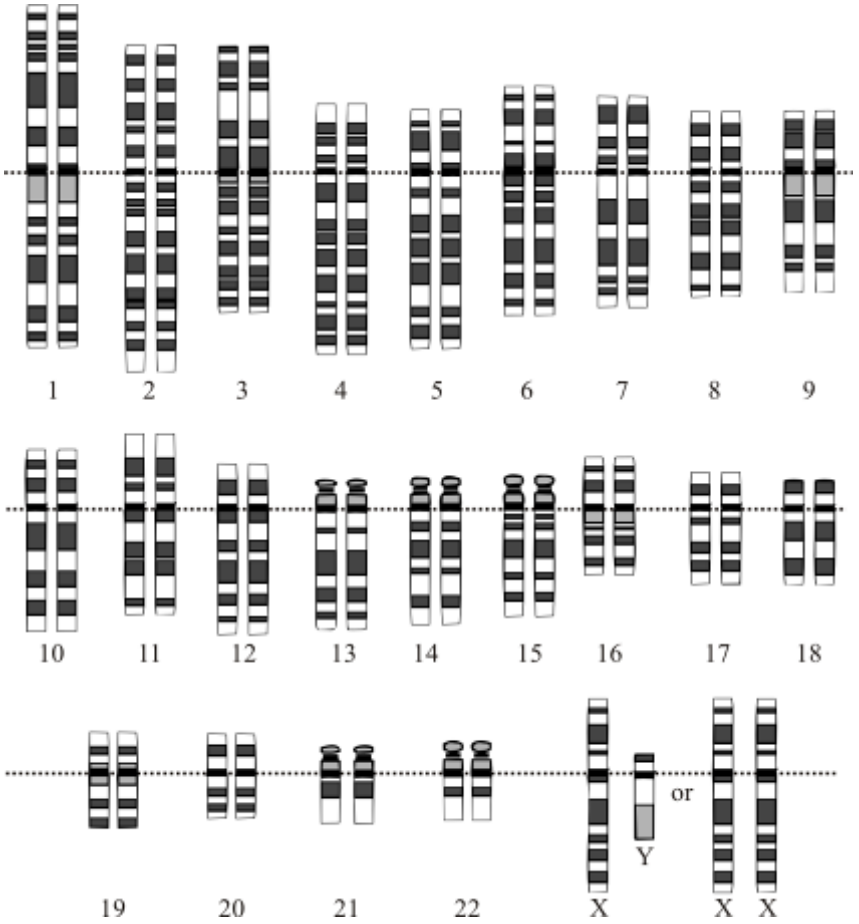
Terminology

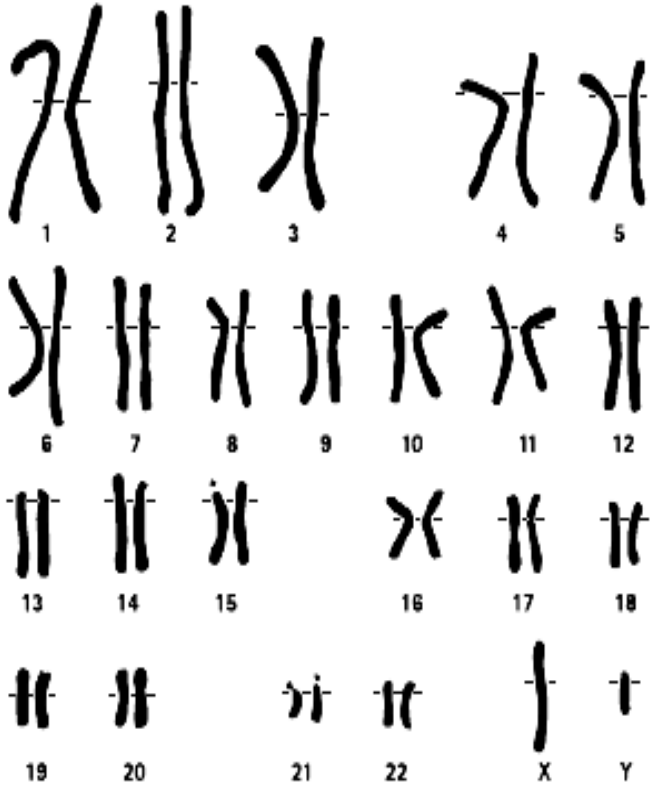
- DNA(deoxyribonucleic acid): A(Adenine), G(Guanine), C(Cytosine), T(Thymine)
  - ✓ Double helix
  - ✓ A=T, C≡G
- Genome (基因體): the set that contains all chromosomes of a single species
  - Human genome = {22對常染色體 + 2條性染色體}
- Molecular evolution (分子演化)
- Phylogeny (演化樹): evolutionary tree

人類染色體結構圖

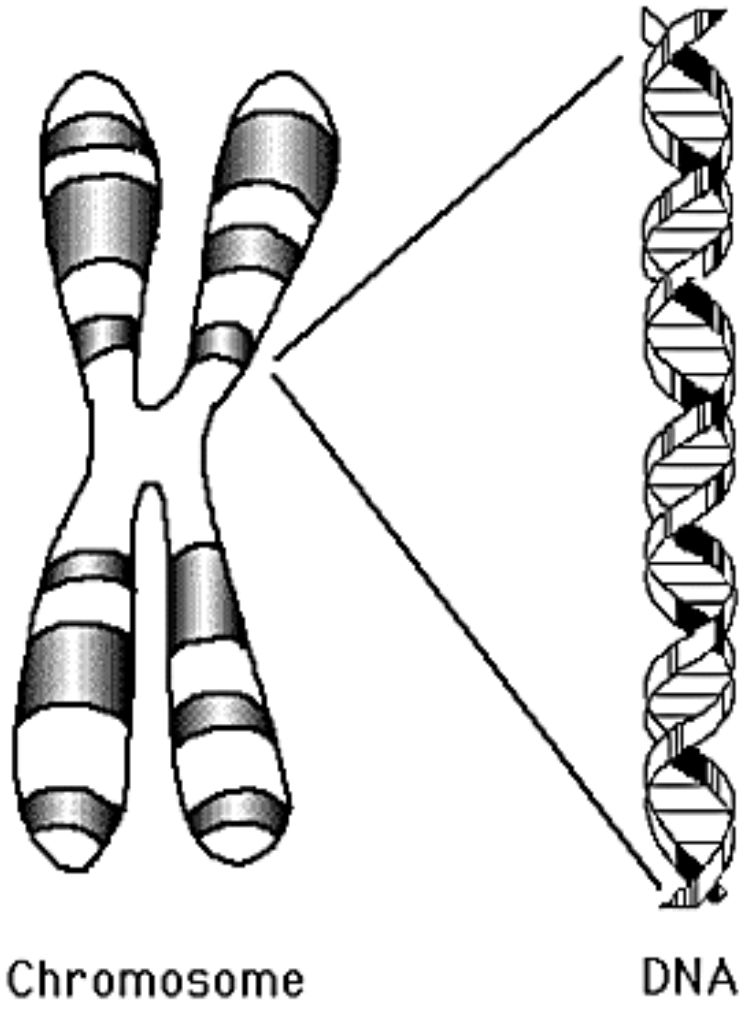


人類染色體


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
or
X
X


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

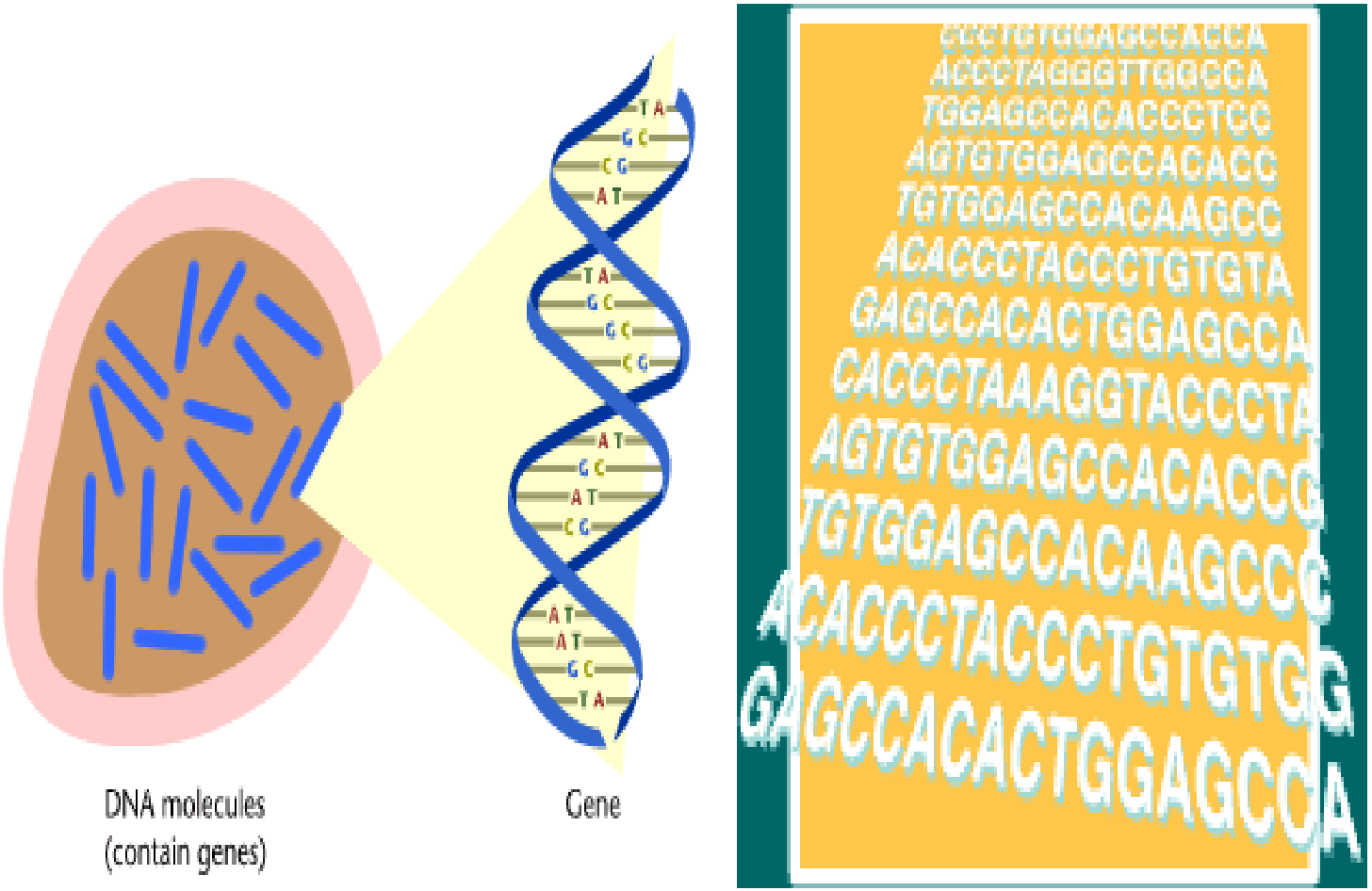
基因組 (Genome) –四種“字母”編成的生命書



DNA - 兩條大分子串形成的雙螺旋

四種大分子 - A,C,G,T

大分子串 - 四種“字母”編成的文章


DNA molecules
(contain genes)
Gene

基因組
非常巨大

A stretch of genome from the X chromo-some of Homo sapien

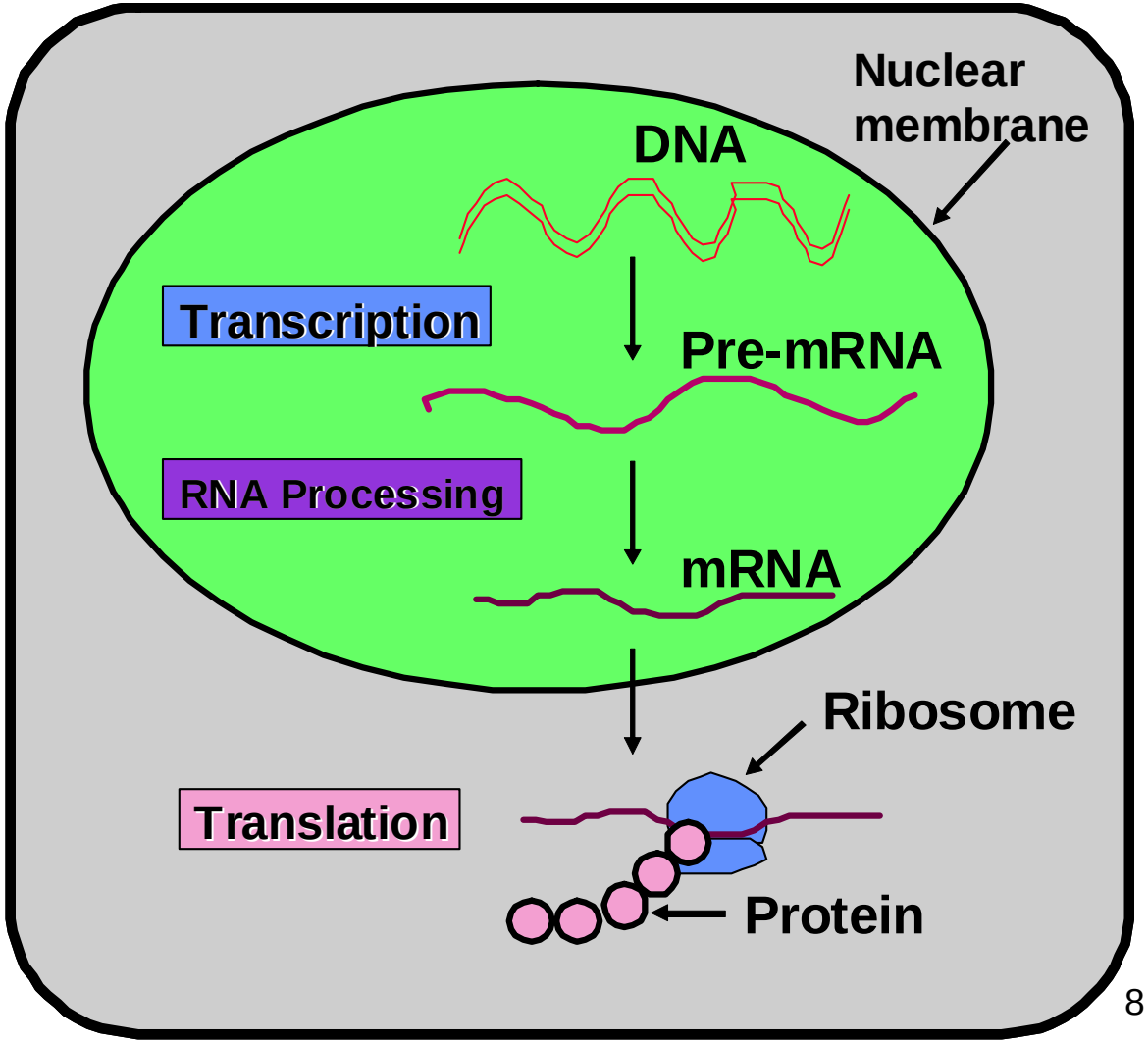
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=2276452&db=Nucleotide&dopt=GenBank

The complete genome has 2,000,000 such Pages

```
   1 tgctgagaaa acatcaagctg tgtttctcct tccccaaag acacttcgca gcccctcttg
  61 ggatccagcg cagcgcaagg taagccagat gcctctgctg ttgccctccc tgtgggcctg
 121 ctctcctcac gccggccccc acctgggcca cctgtggcac ctgccaggag gctgagctgc
 181 aaaccccaat gaggggcagg tgctcccgga gacctgcttc ccacacgccc atcgttctgc
 241 ccccggcttt gagttctccc aggcccctct gtgcacccct ccctagcagg aacatgccgt
 301 ctgccccctt gagctttgca aggtctcggt gataatagga aggtctttgc cttgcaggga
 361 gaatgagtca tccgtgctcc ctccgagggg gattctggag tccacagtaa ttgcagggct
 421 gacactctgc cctgcaccgg gcgccccagc tcctccccac ctcctcctc catccctgtc
 481 tccggctatt aagacggggc gctcaggggc ctgtaactgg ggaaggtata cccgccctgc
 541 agaggtggac cctgtctgtt ttgatttctg ttccatgtcc aaggcaggac atgaccctgt
 601 tttggaatgc tgatttatgg attttccagg ccactgtgcc ccagatacaa ttttctctga
 661 cattaagaat acgtagagaa ctaaatgcat tttcttctta aaaaaaaaaa aaaccaaaaa
 721 aaaaaaaaaa aaaccaaaaa actgtactta ataagatcca tgcctataag acaaaggaac
 781 acctcttgtc atatatgtgg gacctcgggc agcgtgtgaa agtttacttg cagtttgcag
 841 taaaatgaca aagctaacac ctggcgtgga caatcttacc tagctatgct ctccaaaatg
 901 tatttttct aatctgggca acaatggtgc catctcggtt cactgcaacc tccgcttccc
 961 aggttcaagc gattctccgg cctcagcctc ccaagtagct gggaggacag gcacccgcca
1021 tgatgcccgg ttaatttttg tatttttagc agagatgggt tttcgccatg ttggccaggc
1081 tggtctcgaa ctcctgacct caggtgatcc gcctgccttg gcctcccaaa gtgctgggat
1141 gacaggcgtg agccaccgcg cccagccagg aatctatgca tttgcctttg aatattagcc
1201 tccactgccc catcagcaaa aggcaaaaca ggttaccagc ctcccgccac ccctgaagaa
1261 taattgtgaa aaaatgtgga attagcaaca tgttggcagg atttttgctg aggttataag
1321 ccacttcctt catctgggtc tgagcttttt tgtattcggt cttaccattc gttggttctg
1381 tagttcatgt ttcaaaaatg cagcctcaga gactgcaagc cgctgagtca aatacaaata
1441 gatttttaaa gtgtatttat tttaaacaaa aaataaaatc acacataaga taaaacaaaa
1501 cgaaactgac tttatacagt aaaataaacg atgcctgggc acagtggctc acgcctgtca
```

DNA → RNA → Protein

Eukaryotic Cell



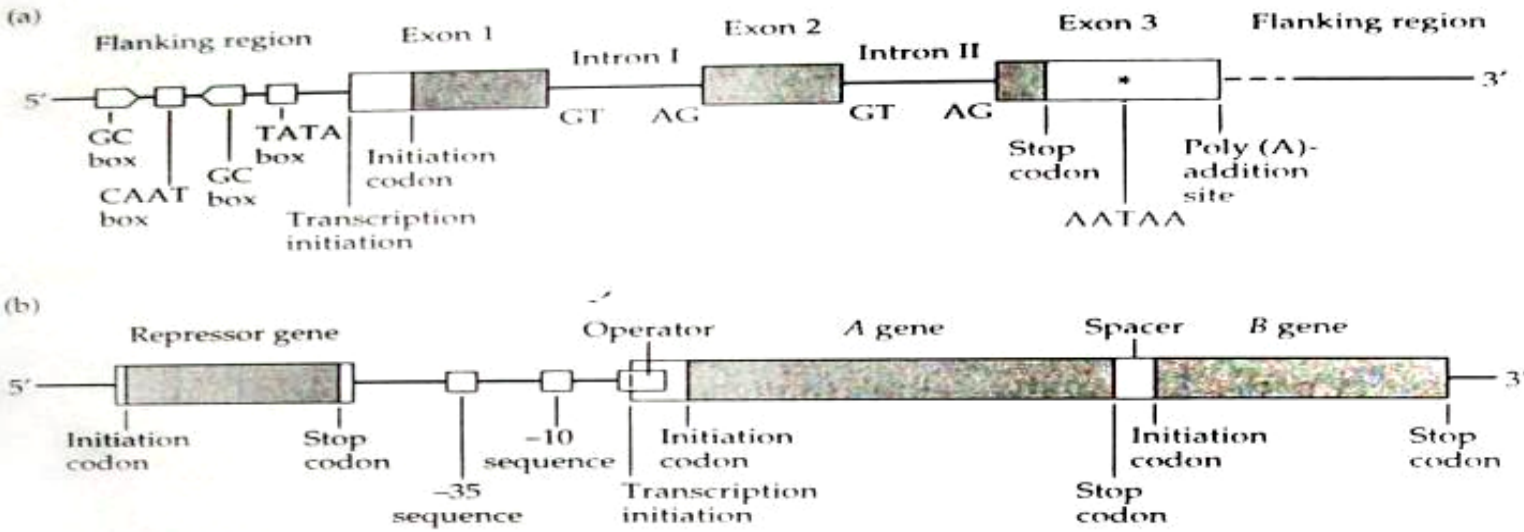


Figure 1.3 (a) Schematic structure of a typical eukaryotic protein-coding gene. Note that, by convention, the 5′ end is at the left. Rectangles denote exons; a white area in a rectangle denotes a transcribed but untranslated region, while a shaded area denotes a translated region. * denotes the site of the poly(A)-addition signal AATAA. (b) Schematic structure of an induced prokaryotic operon. Genes *A* and *B* are protein-coding genes and are transcribed into a single messenger RNA. The repressor gene encodes a repressor protein, which binds to the operator and prevents the transcription of the structural genes by blocking the movement of the RNA polymerase. The operator is a DNA region with at least 10 bases, which may overlap the transcribed region of the genes in the operon. By binding to an inducer (a small molecule), the repressor is converted to a form that cannot bind to the operator. Then RNA polymerase can initiate the transcription of the genes *A* and *B* in the operon (see Lewin 1994). In both (a) and (b), the regions are not drawn according to scale. From Li and Graur (1991).

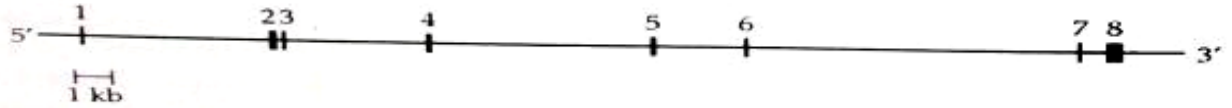


Figure 1.4 The localization of the eight exons in the human factor-IX gene. The vertical bars represent the eight exons. Only the transcribed region is shown. The exons and introns are drawn to scale. The total length of the exons is 1,386 nucleotides, as opposed to a total of length of introns of 29,954 nulceotides. From Li and Graur (1991).

From Wen-Hsiung Li. 1997. Molecular Evolution

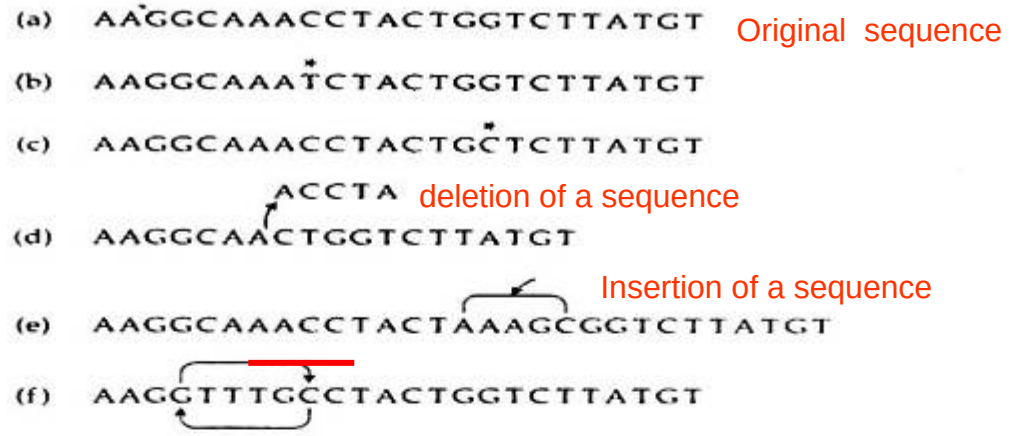


Figure 1.11 Types of mutations. (a) Original sequence; (b) a transition from C to T; (c) a transversion from G to C; (d) a deletion of the sequence ACCTA; (e) an insertion of the sequence AAAGC; (f) an inversion of 5′—GCAAAC—3′ to 5′—GTTTGC—3′. From Li and Graur (1991).

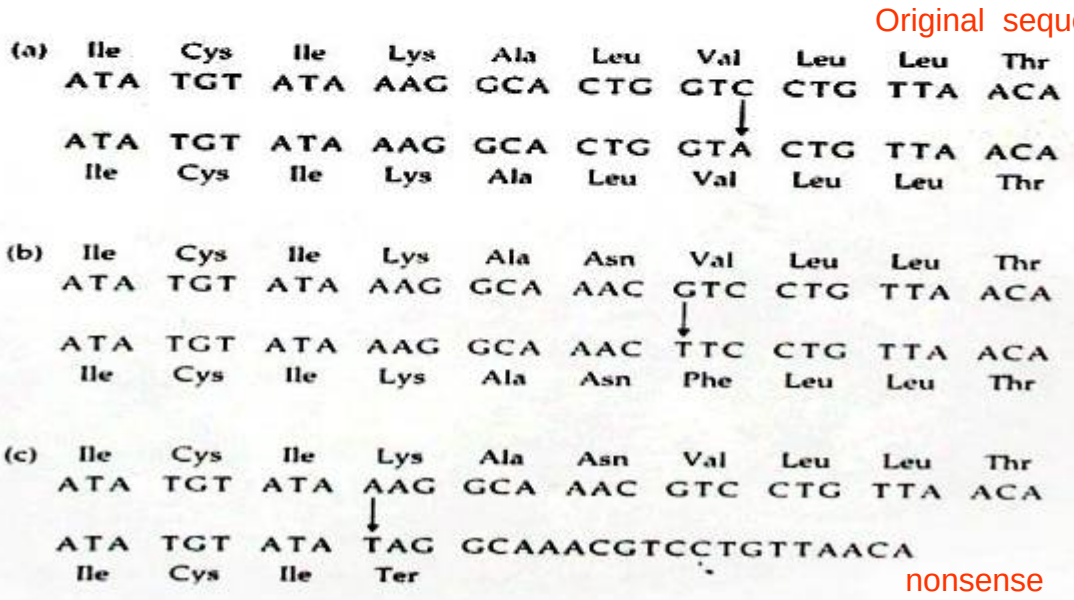


Figure 1.12 Types of point mutations in a coding region: (a) synonymous, (b) missense, and (c) nonsense. From Li and Graur (1991).

From Wen-Hsiung Li. 1997. Molecular Evolution

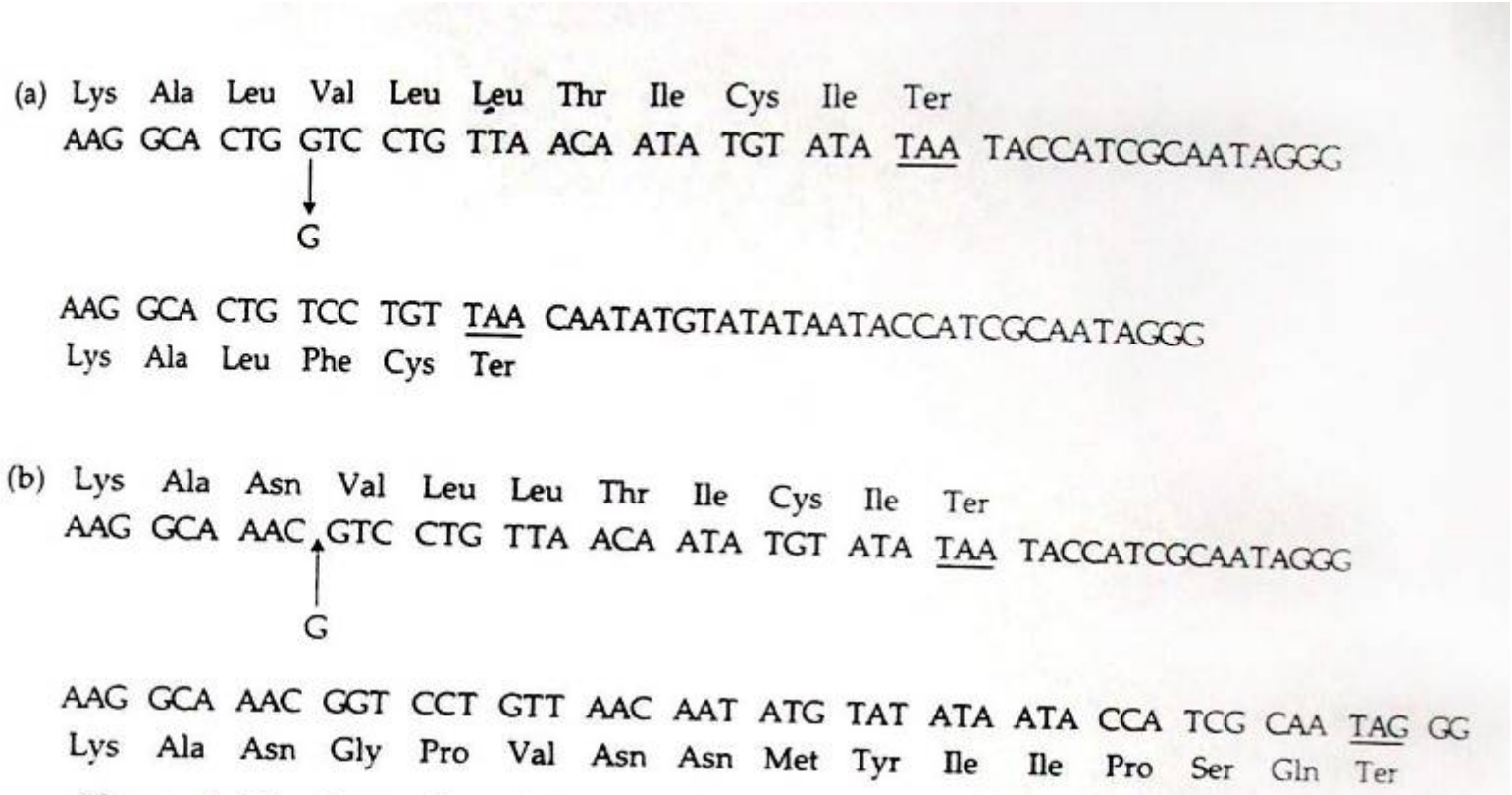


Figure 1.16 Examples of frameshifts in reading frames caused by deletion or insertion. (a) A deletion of a G causes premature termination and (b) an insertion of G obliterates a stop codon. From Li and Graur (1991).

From Wen-Hsiung Li. 1997. Molecular Evolution

How to reconstruct molecular phylogenetic tree?

- Sequence selection and alignment: to determine site-by-site homologies and to detect DNA or amino acid differences

  example:

  C T T G A C T – A G A
  C T – – A C T G T G A

- Build a mathematical model describing the evolution in time of the sequences
  - estimation of the genetic distance between two homologous sequences
  - measured by the expected number of nucleotide substitutions per site that have occurred on the evolutionary lineages between them and their most recent common ancestor
  - Such distances may be represented as branch lengths in a phylogenetic tree

- Apply an appropriate statistical method to find the tree topology and branch lengths that best describe the sequences' phylogenetic relationships

- interpretation of results

Nucleotide substitution models

Outline

- Background
- Evolutionary distance --- nucleotide substitution rates
  - Non-coding sequence
  - Protein-coding sequence
- Human Genome Project

DNA & RNA

- DNA (A(Adenine), G(Guanine), C(Cytosine), T(Thymine))
  - ✓ A=T, C≡G
  - ✓ Double helix

- RNA (A, G, C, U(Uracil))

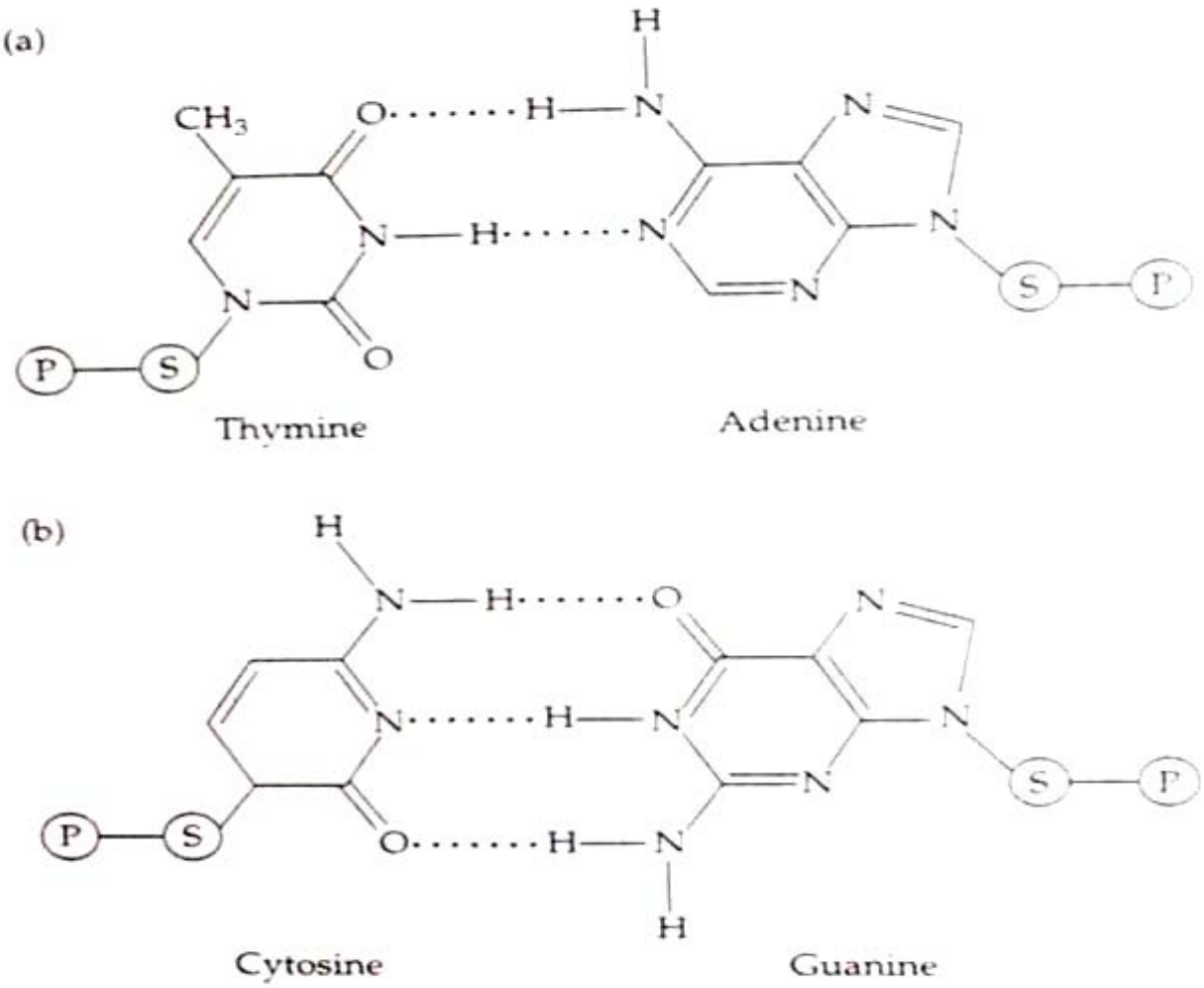


Figure 1.1 Complementary base pairing by means of hydrogen bonds (dotted lines) between (a) thymine and adenine (weak bond), and (b) cytosine and guanine (strong bond). Ⓟ, phosphate; Ⓢ, sugar. From Li and Graur (1991).

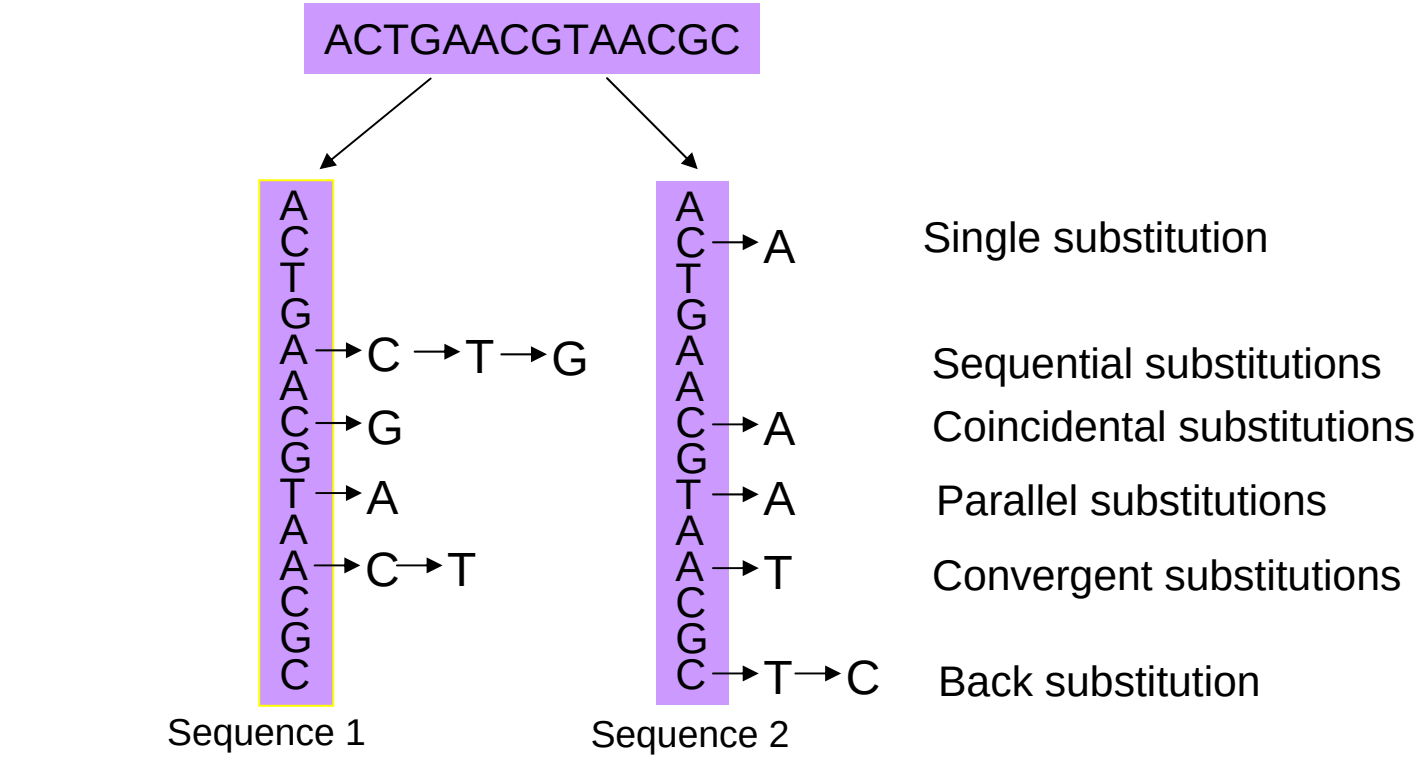
From Wen-Hsiung Li. 1997. Molecular Evolution

Estimation for Nucleotide Substitution Rates --- Noncoding Sequence

Evolutionary Changes in Nucleotide Sequences

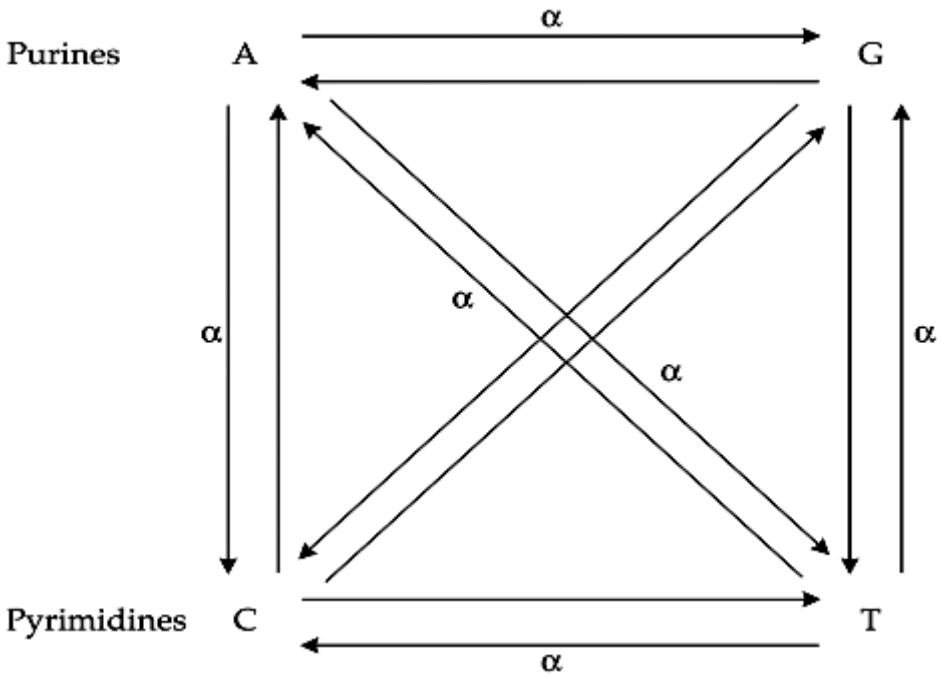
- Number of nucleotide substitutions between sequences

→ Two homologous DNA sequences that descended from an ancestral sequence



Mathematical Model

- $K = N/L$, N = # of nucleotide substitutions, L = the length of DNA sequence
- Jukes-Cantor's one-parameter model (1969)



讓我們先從一個鹼基 A 出發，來解釋它會如何的變化。我們定義

$p_{A(t)}$ = 經過 t 時間後，這個位置仍然是 A 的機率

很明顯的 $p_{A(0)} = 1$。在時間 $t = 1$ 時，這個位置仍然是 A 的機率是 1 減去所有可能變化的機率，所以是

$$p_{A(1)} = 1 - 3\alpha$$

而在時間 $t = 2$ 時，這個位置仍然是 A 的可能有兩種：

(i) 在 $t = 1$ 時是 A，在 $t = 2$ 時也是 A

(ii) 在 $t = 1$ 時不是 A (即可能是 G 或 T 或 C)，然後在 $t = 2$ 時又變回 A

所以

$$p_{A(2)} = (1 - 3\alpha)\, p_{A(1)} + \alpha(1 - p_{A(1)})$$

因此，對所有的 t，我們可以得出一個一般式，

$$p_{A(t+1)} = (1 - 3\alpha)\, p_{A(t)} + \alpha(1 - p_{A(t)})$$

經過整理之後，我們可以得到，

$$\Delta p_{A(t)} = p_{A(t+1)} - p_{A(t)} = -3\alpha p_{A(t)} + \alpha(1 - p_{A(t)}) = -4\alpha p_{A(t)} + \alpha$$

因爲這是時間離散型的方程式，當時間間隔夠小時，我們可以用時間連續型的方程式來逼近。所以我們可以得到一個一階線性的微分方程，

$$\frac{dp_{A(t)}}{dt} = -4\alpha p_{A(t)} + \alpha \qquad (1)$$

這是一個起始值問題，它的起始條件是 $p_{A(0)} = 1$。

經過一些計算，我們可以解出

$$p_{A(t)} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \qquad (2)$$

這就是在時間等於零時是 A，經過 t 時間後仍然是 A 的機率。對相同的微分方程式 (1) 來說，我們如果將起始條件改變成爲 $p_{A(0)} = 0$，則我們可以得到在時間等於零時不是 A，經過 t 時間後是 A 的機率

$$p_{A(t)} = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} \qquad (3)$$

因爲這是 one-parameter model，AGCT 彼此間都是對稱的關係，所以方程式 (2) 和 (3) 對 G 或 C 或 T 來說，都是適用的。所以我們定義

$p_{ij(t)}$ = 某個位置的鹼基 (nucleotide) 一開始是 i，在經過 t 時間後會變成 j 的機率

在這裡 i， j = A，G，C，T，

然後我們可以得到

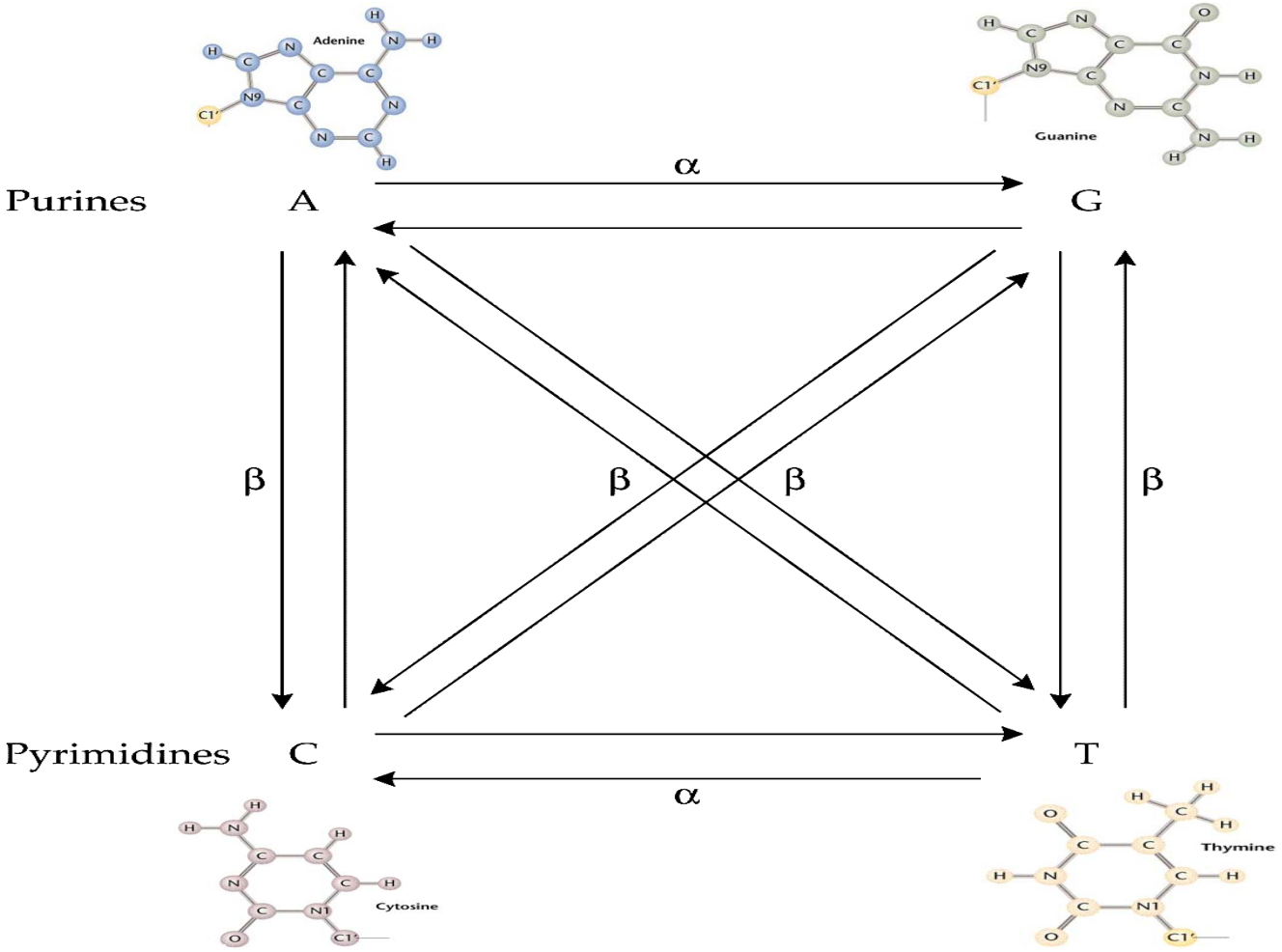
$$p_{ii(t)} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \tag{4}$$

以及對所有 $i \neq j$

$$p_{ij(t)} = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} \tag{5}$$

方程式 (4) 和 (5) 便可以用來描述 Jukes-Cantor’s one-parameter model，並且用來做預測。

Kimura’s two-parameter model (1980)



我們先將四種鹼基 ATCG 分別用數字 1234 來表示，並令 M 是一個 4 乘 4 的矩陣。m_{ij} = 某個位置的鹼基 (nucleotide) 一開始是 i，在下一個 generation 會變成 j 的機率。對 Kimura's two-parameter model 來說，

$$M = \begin{bmatrix} 1-\alpha-2\beta & \beta & \beta & \alpha \\ \beta & 1-\alpha-2\beta & \alpha & \beta \\ \beta & \alpha & 1-\alpha-2\beta & \beta \\ \alpha & \beta & \beta & 1-\alpha-2\beta \end{bmatrix}$$

我們令 $p_{k(t)}$ = 某個位置的鹼基 (nucleotide) 在經過 t 時間後會變成 k 的機率，在這裡 k = A，T，C 和 G。然後再定義向量 $P(t) = (p_{A(t)}, p_{T(t)}, p_{C(t)}, p_{G(t)})$。我們就可以得到用來描述 Kimura's two-parameter model 的方程式是

$$P(t+1) = P(t)M \tag{6}$$

藉由線性代數的知識，我們可以解出方程式 (6)。舉例來說，我們可以得到

$$p_{ii(t)} = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t}$$

在這裡 i = A, T, C, 和 G。而 $p_{ii(t)}$ 和之前定義過的一樣，是某個位置的鹼基 (nucleotide) 一開始是 i，在經過 t 時間之後還是 i 的機率。

接下來我們便可以將 Jukes-Cantor's one-parameter model 和 Kimura's two-parameter model，應用在估計兩條 DNA 序列之間的平均每一個位置 (per nucleotide) 會發生多少次取代 (substitution) 的數目上，也就是估計 K 值。我們先定義 $I(t)$ = 兩條 DNA 序列在經過時間 t 之後，在某一個特定位置上的鹼基是相同的機率。假如某一個特定位置在時間等於零時是 A，則

$$I(t) = (p_{AA(t)})^2 + (p_{AT(t)})^2 + (p_{AC(t)})^2 + (p_{AG(t)})^2 \tag{7}$$

根據 Jukes-Cantor's one-parameter model，我們將方程式 (4) 和 (5) 帶入 (7)，可以得到

$$I(t) = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t} \tag{8}$$

事實上，當某一個特定位置一開始時是 T，C 或 G 時，Equation (8) 仍然是成立的。也就是說 (8) 式和某個特定位置一開始時是哪一個鹼基是無關的。相似地，如果我們將 Kimura's two-parameter model 應用在 (7) 式，則我們可以得到

$$I(t) = \frac{1}{4} + \frac{1}{4}e^{-8\beta t} + \frac{1}{2}e^{-4(\alpha+\beta)t} \tag{9}$$

同樣的，(9) 式和某個特定位置一開始時是哪一個鹼基也是無關的。

根據 Equation (8)，我們可以得到兩條 DNA 序列在某一個特定位置上不同的機率是 $p = 1 - I(t)$。所以，

$$p = \frac{3}{4}(1 - e^{-8\alpha t}) \tag{10}$$

因爲在 Jukes-Cantor's one-parameter model 中，α 是鹼基取代 (nucleotide substitution) 的速率，所以 $3\alpha t$ 是每一個位置在經過時間 t 之後，會發生多少次取代的數目。因爲我們是拿兩條 DNA 序列來比較，所以我們可以得到 $K = 2(3\alpha t)$。因此，從 (10) 式我們可以推導出

$$K = -(\frac{3}{4})\ln(1 - \frac{4p}{3}) \tag{11}$$

這便是從 Jukes-Cantor's one-parameter model 推導出的，用來估計兩條 DNA 序列之間平均每個位置曾發生多少次取代的公式。在應用上，p 是用我們觀察到的兩條 DNA 序列之間，平均每個位置有多少差異來代入。也就是 $p = D/L$，在這裡 L 是 DNA 序列的長度，而 D 是兩條 DNA 序列之間有多少個位置是不同的。當 L 很大時，K 的樣本變異係數會近似於

$$V(K) = p(1-p)/[L(1-4p/3)^2]$$

(Kimura and Ohta 1972)。以上是根據 Jukes-Cantor's one-parameter model 所推導出的 K 值。

如果是根據 Kimura's two-parameter model，我們可以推導出 K 值的公式是

$$K = \frac{1}{2}\ln(a) + \frac{1}{4}\ln(b) \qquad (12)$$

在這裡 $a = 1/(1-2P-Q)$ ， $b = 1/(1-2Q)$ ，P 和 Q 分別是兩條 DNA 序列之間 transitional 和 transversional 的差異的比例 (proportions)。K 的樣本變異係數是近似於

$$V(K) = [a^2 P + c^2 Q - (aP + cQ)^2]/L$$

在這裡 $c = (a+b)/2$ (Kimura 1980)。

Markov models

At any single site, the model works with probabilities

$P_{ij}(T)$ = the probability that base *i* will have changed to base *j* after a time *T*.

The subscripts *i* and *j* take the values 1,...,4 to represent the nucleotides A, T, C, G for DNA sequences and 1,...,20 for amino acid sequences.

Given a stochastic variable $X(t)$ describing the evolution through time *t* of a site in one sequence, the Markov assumption asserts that $P_{ij}(T) = Pr[X(s+T) = j \mid X(s) = i]$ is independent of $s \geq 0$.

The probabilities of transition from one base to another, $P_{ij}(T)$, can be written as a matrix $\mathbf{P}(T)$, and then we can write

$$\mathbf{P}(T+dT) = \mathbf{P}(T)(\mathbf{I} + \mathbf{Q}dT)$$

where dT represents a small time, and $\mathbf{I}$ is the identity matrix. The matrix $\mathbf{Q}$ is known as the instantaneous rate matrix and has off-diagonal entries Q_{ij} equal to the rates of replacement of i by j. (The diagonal entries, Q_{ii}, are defined by a mathematical requirement that the row sums are all zero.) This equation is solved to give

$$P(T) = e^{TQ} = I + TQ + \frac{(TQ)^2}{2!} + \frac{(TQ)^3}{3!} + \cdots$$

Spectral decomposition (also termed diagonalization) of $\mathbf{Q}$ allows us to calculate the matrix $\mathbf{P}(T)$:

$$P(T) = U \cdot \text{diag}\{e^{\lambda_1 T}, \ldots, e^{\lambda_n T}\} \cdot U^{-1}$$

where the matrix $\mathbf{U}$ contains the eigenvectors of $\mathbf{Q}$, the λ_i are the eigenvalues of $\mathbf{Q}$ and diag{} denotes the diagonal matrix of the elements contained in the braces. The components $P_{ij}(T)$ can be written as

$$P_{ij}(T) = \sum_k c_{ijk} e^{\lambda_k T}$$

where the sum is over $k = 1,\ldots,4$ for DNA sequences and over $k = 1,\ldots,20$ for amino acids; c_{ijk} is a function of $\mathbf{U}$ and $\mathbf{U}^{-1}$. Note that $\mathbf{T}$ and $\mathbf{Q}$ are confounded; $T\mathbf{Q} = (T/r)(r\mathbf{Q})$ for any $r \neq 0$ (e.g., half the time at twice the rate has the same result). Therefore, absolute times T typically cannot be used, and in practice, time is scaled to units of expected substitutions per site.

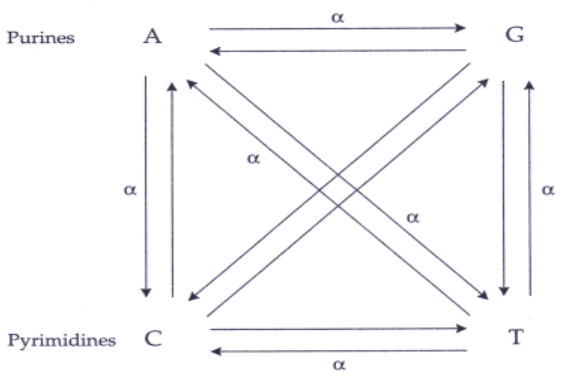
Simple Models of Molecular Evolution

- Zuckerkandl and Pauling (1965) proposed the theory of a molecular clock
  - the rate of molecular evolution is approximately constant over time for all the proteins in all lineages

- Jukes and Cantor (1969) proposed a stochastic model for DNA substitution in which all nucleotide substitutions occur at an equal rate

Jukes-Cantor's one-parameter model (1969)

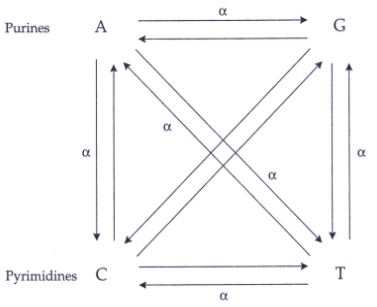
- Jukes and Cantor (1969) described above is defined by $Q_{ij} = \alpha$ for all i, j = 1,...,4; $i \neq j$, meaning that each base is substituted by any other at equal rate
- A consequence of this model is that the base frequencies (π_i) are all assumed equal to 0.25



- $K = -\frac{3}{4}\ln(1-\frac{4}{3}p)$

Reasons for more complicated models

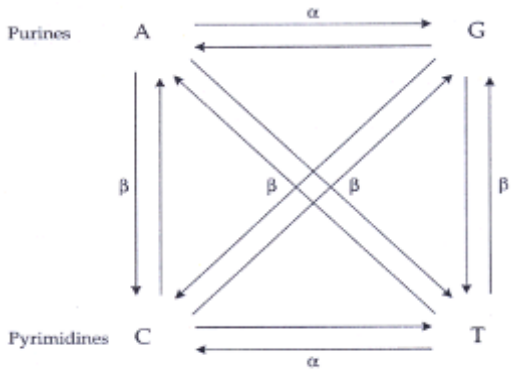
- Mutation rates affected by many factors
  - chromosomal position (Sharp et al. 1989)
  - G + C content (Wolfe 1991)
  - nearest neighbor bases (Blake et al. 1992)



- transitions occur more frequently than transversions (Brown and Simpson 1982)
  - often twice as frequently, but the ratio can be much higher

Kimura's two-parameter model (1980)

- Kimura (1980) proposed a two-parameter model that considered the difference in transition and transversion rates



$$\mathbf{Q} = \begin{bmatrix} \cdot & \beta & \beta & \alpha \\ \beta & \cdot & \alpha & \beta \\ \beta & \alpha & \cdot & \beta \\ \alpha & \beta & \beta & \cdot \end{bmatrix}$$

(the order of the bases for columns and rows are A, T, C, G)

- $$K = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right) + \frac{1}{4}\ln\left(\frac{1}{1-2Q}\right)$$

More models

- Felsenstein (1981) proposed a model in which the rate of substitution to a nucleotide depends only on the equilibrium frequency of that nucleotide

$$\mathbf{Q} = \begin{bmatrix} \cdot & \mu\pi_T & \mu\pi_C & \mu\pi_G \\ \mu\pi_A & \cdot & \mu\pi_C & \mu\pi_G \\ \mu\pi_A & \mu\pi_T & \cdot & \mu\pi_G \\ \mu\pi_A & \mu\pi_T & \mu\pi_C & \cdot \end{bmatrix}$$

- Blaisdell (1985) introduced an asymmetry for some reciprocal changes

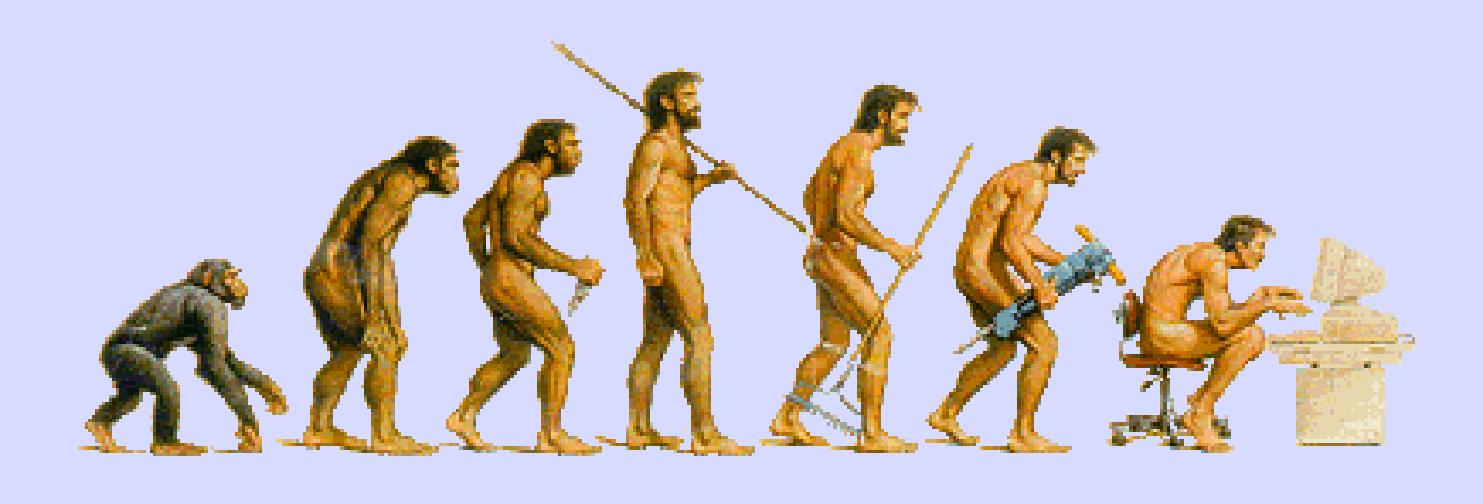
$$\mathbf{Q} = \begin{bmatrix} \cdot & \gamma & \gamma & \alpha \\ \delta & \cdot & \alpha & \delta \\ \delta & \beta & \cdot & \delta \\ \beta & \gamma & \gamma & \cdot \end{bmatrix}$$

Models of nucleotide substitution

O\S[a]	A	T	C	G
a. Two-parameter model (Kimura 1980)				
A	$1-\alpha-2\beta$	β	β	α
T	β	$1-\alpha-2\beta$	α	β
C	β	α	$1-\alpha-2\beta$	β
G	α	β	β	$1-\alpha-2\beta$
b. Four-parameter model (Blaisdell 1985)				
A	$1-\alpha-2\gamma$	γ	γ	α
T	δ	$1-\alpha-2\delta$	α	δ
C	δ	β	$1-\beta-2\delta$	δ
G	β	γ	γ	$1-\beta-2\gamma$
c. Six-parameter model (Kimura 1981a)				
A	$1-2\alpha-\gamma$	γ	α	α
T	δ	$1-2\alpha-\delta$	α	α
C	β	β	$1-2\beta-\varepsilon$	ε
G	β	β	ξ	$1-2\beta-\xi$
d. Nine-parameter model				
A	$1-g_T\beta_1-g_C\gamma_1-g_G\alpha_1$	$g_T\beta_1$	$g_C\gamma_1$	$g_G\alpha_1$
T	$g_A\beta_1$	$1-g_A\beta_1-g_C\alpha_2-g_G\gamma_2$	$g_C\alpha_2$	$g_G\gamma_2$
C	$g_A\gamma_1$	$g_T\alpha_2$	$1-g_A\gamma_1-g_T\alpha_2-g_G\beta_2$	$g_G\beta_2$
G	$g_A\alpha_1$	$g_T\gamma_2$	$g_C\beta_2$	$1-g_A\alpha_1-g_T\gamma_2-g_C\beta_2$
e. General model				
A	$1-\alpha_{12}-\alpha_{13}-\alpha_{14}$	α_{12}	α_{13}	α_{14}
T	α_{21}	$1-\alpha_{21}-\alpha_{23}-\alpha_{24}$	α_{23}	α_{24}
C	α_{31}	α_{32}	$1-\alpha_{31}-\alpha_{32}-\alpha_{34}$	α_{34}
G	α_{41}	α_{42}	α_{43}	$1-\alpha_{41}-\alpha_{42}-\alpha_{43}$

[a]O, Original nucleotide; S, substitute nucleotide.

Human Genome Project



人類基因體計畫 (Human Genome Project)

- 基因體（genome):收集某一個物種所有的染色體而組成的完整的集合
- 人類基因體: 22對常染色體 + 2條性染色體（各取一條, 共24條）
- 基因體計畫: 人類, 老鼠, 果蠅, 大腸桿菌, 酵母菌等等
- Genomics: 研究基因體的內涵和組成結構

傳統遺傳學和反方向的遺傳學

- 遺傳疾病（genetic disease）
  - 侏儒基因：成對且顯性
  - 一個正常人和侏儒結婚，至少有 1/2 的機會會生出侏儒
  - 侏儒和侏儒結婚，最多只有 1/4 的機會會生出正常人
- 傳統遺傳學：費時且費錢
- 反方向的遺傳學
  - DNA定序
  - 找出所有基因再研究功能

基因預測 (gene prediction)

- 人類的DNA序列大概包含了30億個鹼基，而其中只有小於3%的地方是有功能的基因
- 如何預測
  - 找特定結構
  - BLAST (Best Local Alignment Search Tool)
  - 比較人類和老鼠的DNA序列：coding region with *Ka*/*Ks* <<1

人類基因數目

- 1990年10月1日，由美國國家衛生院和英國衛爾康基金會（Wellcome Trust）主導的人類基因體計畫（Human Genome Project）正式展開，該計畫由美、英、德、法、日、大陸為首，共有18個國家參與

- 美國科學家Craig Venter在1998年所創立的賽雷拉公司（Celera Genomics）

- 2001年1月共同發表：人類有3~4萬個基因

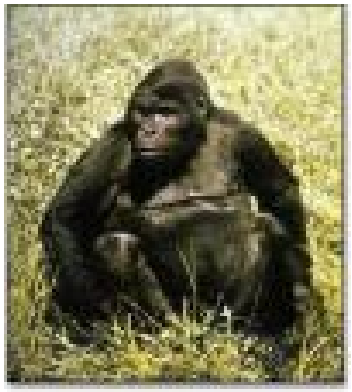
- 兩者結果只有6千多個基因有交集

已完成排序的大基因體

- 1995-2005 – About ～100 bacterial genomes(細菌) 0.5-9 Mb; hundreds to 2000 genes
- 1996 April – Yeast (Saccharomyces cerevisiae;酵母) 12 Mb, 5,500 genes
- 1998 Dec. -Worm (Caenorhabditis elegans;線蟲) 97 Mb, 19,000 genes
- 2000 March - Fly (Drosophila melanogaster;果蠅)137Mb, 13,500 genes
- 2000 Dec. - Mustard (Arabidopsis thaliana;阿拉伯芥)125 Mb, 25,498 genes
- 2000 June – Human (Homo sapiens) 1st rough draft
- 2001 Feb 15/16 – Human, “working draft” 人類 3000 Mb, 35,000~40,000 genes
- 2005 Sep 1 (*Nature*) – Chimpanzee 黑猩猩
- 老鼠、稻米、虐原蟲、還有許多

Application of the similarity

- 以前的分類學
  - 現代人：動物界、脊索動物門、哺乳綱、靈長目、人科、人屬、智人種
  - 黑猩猩(chimpanzee)和大猩猩、猩猩等一起歸進猩猩科，而且和人類是在1500萬年前就已經在演化樹上分開了
- 現在
  - 人類和黑猩猩是在700萬年前才從共同祖先分開來
  - 人類和黑猩猩的DNA序列之間的相似度高於95%
  - 黑猩猩：人科？甚至是人屬？



Reference

- Statistical Methods in Bioinformatics, WJ Ewens and GR Grant
- Molecular Evolution, Wen-Hsiung Li