Chap. 3. Base substitution Pattern

- DNA 염기서열 내에 축적된substitution의 양과 특성을 파악하는 것이 molecular evolution연구의 핵심.

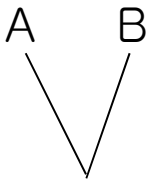
Patterns of Substitutions within Genes

"If it is not broken, don't fix it"

← 거의 모든 유전자들은 이 생명체가 서식하는 환경조건에 대하여 최적인 상태임.

- "mistake" 가 생명체에 미치는 영향
  1) disadvantageous or deleterious 해로운
  2) advantageous 이로운
  3) neutral 중립적

→Advantageous changes are rare



Mutation Rates

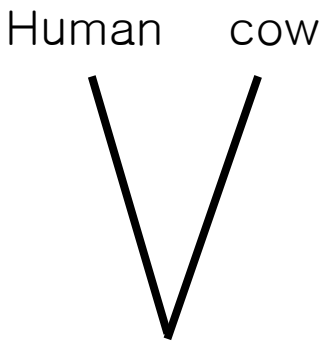
r = substitution rate 치환율

K = number of substitutions two sequences have undergone since they last shared a common ancestor 두 서열이 공동조상 이래로 축적한 치환의 양

T = divergence time 분기 시간

$r = K/(2T)$

치환율을 계산하려면 적어도 두 종 이상의 데이터가 있어야 가능함.



Q: 인간과 소의 한 유전자의 mutation rate 구하기

- 대상 유전자는 1,000bp
- 인간과 소의 이 유전자는 40bp의 차이가 있음
- 인간과 소는 6,300,000년 전 (6.3MYBP)에 분지하였다고 가정

$r = K/(2T)$

$r = K/(2T) = 40 / (2 \times 6{,}300{,}000) = 3.17 \times 10^{-6}$

← 유전자 전체에 대한 rate.

한 site 당 치환율은?

$(3.17 \times 10^{-6}) / 1{,}000 = 3.17 \times 10^{-9}$

단위는?

3.17×10^{-9} bp/site/year

참조: http://www.timetree.org/index.php

Functional Constraint

- DNA 서열 중 “functional constraint” 상태에 있는 부분들
  → 진화적인 변화가 매우 느리게 축적되는 경향이 있음.

- 하나의 유전자 내에도 변이 축적 속도가 다른 부분들이 있음.

TABLE 3.1 Average pairwise divergence among different regions of the human, mouse, rabbit, and cow beta-like globin genes.

Region	Legth of Region(bp) in Human	Average Pairwise Number of Changes	Standard Deviation	Substitution Rate (substitutions/ site/10^9 years)
Noncoding, overall	913	67.9	14.1	3.33
Coding, overall	441	69.2	16.7	1.58
5′ Flanking sequence	300	96.0	19.6	3.39
5′ Untranslated sequence	50	9.0	3.0	1.86
Intron 1	131	41.8	8.1	3.48
3′ Untranslated sequence	132	33.0	11.5	3.00
3′ Flanking sequence	300	76.3	14.3	3.60

Note: No adjustment is made for the possibility that multiple changes may have occurred at some sites.

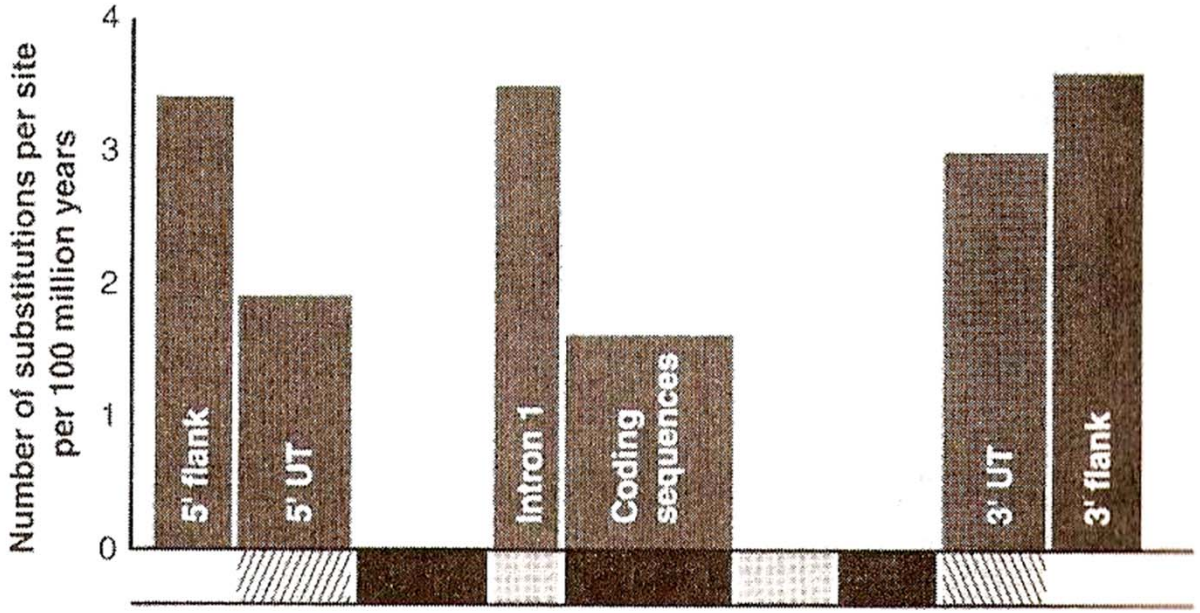


FIGURE 3.1 *Structure and relative rate of change within the beta-like globin gene in four mammals. Three hundred base pairs of the 5' and 3' flanking sequences are shown as open boxes; the 5' transcribed but untranslated (5' UT) sequence is represented as a forward slash-filled box; the 3' transcribed but untranslated (3' UT) sequence is shown as a backward slash-filled box; exons are shown as black boxes; and introns are shown as gray boxes. Relative rates of change are taken from Table 3.1.*

Synonymous vs. Nonsynonymous Substitution

- Synonymous substitution동의치환: nucleotide changed but amino acid does not changed
- Nonsynoymous substitution이의치환: both nuclotide and amino acid changed

Glycine: GGG, GGA, GGU, GGC
- Nondegenerate sites
- Twofold degenerate sites
- Fourfold degenerate sites

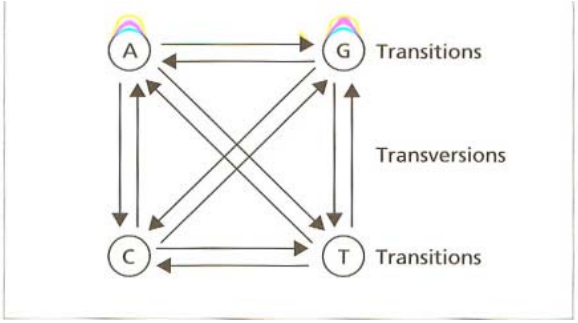


Fig. 5.10 The possible substitutions among the four nucleotides.

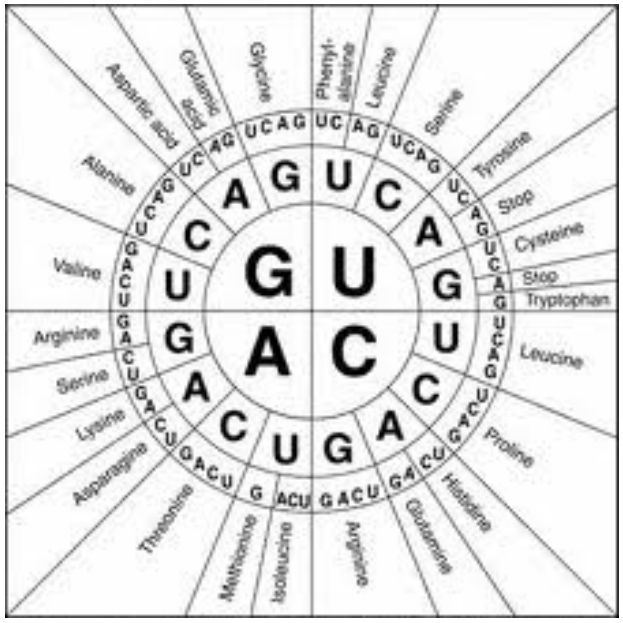
- four fold degenerate sites 의 substitution rate는 선택압이 전혀 없는 3' flanking sequences 와 거의 비슷한 수준임.

TABLE 3.2 Divergence between different kinds of sites within the coding sequence of the human and rabbit beta-like globin genes.

Region	Number of Sites(bp)	Number of Changes	Substitution Rate (substitutions/ site/10^9 years)
Nondegenerate	302	17	0.56
Twofold degenerate	60	10	1.67
Fourfold degenerate	85	20	2.35

Note: Sequences used are available from GenBank (accession numbers V00497 and V00879, respectively). No adjustment is made for the possibility that multiple changes may have occurred at some sites. A divergence time of 100 million years is assumed.

Degenerate Code



W = A or T
S = C or G
R = A or G
Y = C or T
K = G or T
M = A or C
B = C, G, or T (not A)
D = A, G, or T (not C)
H = A, C, or T (not G)
V = A, C, or G (not T)
N = A, C, G, or T

Indels and Pseudogenes

- Indels 은 단순히 base substitution이 일어나는 것 보다 약 10배 이상 일어날 확률이 낮다. → frame을 바꾸어 많은 변화를 일으키기 때문.

  ※ frame shift mutation

- 새로운 유전자는 주로 gene duplication에 의해 일어남.
- 유전자의 duplication 결과 한 copy의 유전자는 원래의 기능을 담당하지만 다른 한 copy의 유전자는 굳이 기능을 담당하지 않아도 됨으로(free of selective constraint) base substitution을 축적할 수 있게 됨.

  → 새로운 유전자 또는 pseudogene (위유전자; 가유전자)를 만듬.

※ pseudogene: 원래의유전자와 비슷한 염기서열을 갖지만 transcription 되지 않거나 기능이 없음.

- 포유류 유전체는 pseudogene들로 가득 차있고 매우 빠른 속도로 substitution을 축적함. → 4 / site /100 million years

Subsitutions vs. Mutations

- Mutation돌연변이: DNA 복제 또는 복구 과정에서 오류로 인한 염기서열의 변화.
- Subsitution염기치환: 몇 세대에 걸쳐서 선택이 일어나 고착화된 mutation들.

- Ks: synonymous substituion rate
  ← 실질적인 mutation rate를 나타냄
- Ka: nonsynonymous substitution rate
  ← “natural selection”의 의미

TABLE 3.3 Ratios of synonymous differences per synonymous site (K_s) and nonsynonymous differences per nonsynonymous site (K_a) for a variety of mammalian genes.

Gene	Codons (in human)	Human/mouse		Human/cow		Human/rabbit		Mouse/cow		Mouse/rabbit		Cow/rabbit		Averages	
		K_s	K_a	K_s	K_a	K_s	K_a	K_s	K_a	K_s	K_a	K_s	K_a	K_s	K_a
Erythropoietin	194	0.481	0.063	0.242	0.068	0.394	0.070	0.495	0.076	0.480	0.058	0.342	0.071	0.406	0.068
Growth hormone	217	0.321	0.100	0.236	0.106	0.220	0.113	0.380	0.046	0.396	0.027	0.244	0.048	0.299	0.073
Prolactin receptor	621	0.304	0.082	0.249	0.122	0.321	0.072	0.358	0.124	0.413	0.088	0.300	0.114	0.324	0.100
Prolactin	226	0.364	0.098	0.368	0.085	0.395	0.064	0.382	0.112	0.307	0.131	0.521	0.064	0.390	0.092
Serum albumin	610	0.528	0.062	0.329	0.067	0.324	0.075	0.477	0.065	0.500	0.065	0.327	0.067	0.414	0.067
Alpha globin	143	0.584	0.022	0.236	0.025	0.204	0.038	0.505	0.025	0.539	0.041	0.242	0.048	0.385	0.033
Beta globin	148	0.324	0.033	0.271	0.046	0.294	0.015	0.263	0.062	0.392	0.039	0.333	0.059	0.313	0.042
Prothrombin	608	0.033	0.687	0.033	1.040	0.075	1.602	0.196	0.887	0.037	1.442	0.078	0.318	0.075	0.996
Apolipo-protein E	317	0.199	0.148	0.132	0.117	0.108	0.114	0.187	0.160	0.165	0.144	0.125	0.126	0.153	0.135
Carbonic anhydrase I	336	0.255	0.159	0.203	0.149	0.207	0.138	0.338	0.113	0.284	0.115	0.187	0.117	0.246	0.132
P53	392	0.372	0.059	0.351	0.061	0.382	0.045	0.457	0.067	0.412	0.054	0.378	0.056	0.392	0.057
Histone 2A	115	0.967	0.057	1.110	0.057	0.174	0.034	0.298	0.006	1.176	0.025	1.192	0.025	0.820	0.033
Column averages		0.394	0.131	0.313	0.162	0.258	0.198	0.361	0.145	0.425	0.186	0.356	0.093	0.351	0.152

Fixation

Allele: 대립유전자 (종 내에서 축적된 유전적 변이들)
q = 한 유전자에 대한 새로운 변이가 생기는 빈도
N = 한 개체군내에서 생식 가능한 이배체 생명체들의수

q = 1 / 2N = 1/N X 1/2

- 생명체의 변식력을 감소시키는 돌연변이는 자연선택과정을 통해서 gene pool로부터 제거되는 경향이 있기 때문에 이들의 빈도는 결국 0이 된다.
- 이로운 allele가 생겨나면 q는 점차 1로 접근함.

- 자연상태에서 나타나는 allele는 어떻게 설명?
  ← 이 변이가 selectively natural 하기 때문임.

- 일반적으로 새로운 중립적 돌연변이가 고착되는 평균시간은 대상 개체군의 4N 세대에 해당함.

- Saturation mutagenesis포화돌연변이법: 인위적으로 mutation을 만들어 그 유전자의 기능을 알아보는 방법

http://signal.salk.edu/tabout.html

한 분류군은 지속적으로 base substitution들을 축적하고 어느 지점 이후로는 saturation이 일어남. Why?

Multiple hit (다중치환)가 일어나기 때문

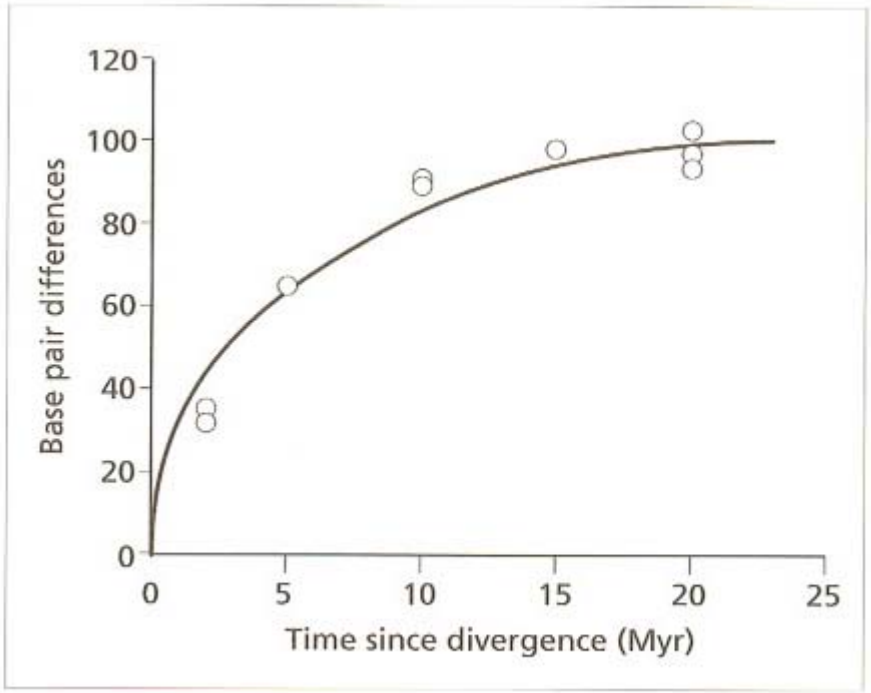


Fig. 5.11 Number of nucleotide substitutions between pairs of bovid mammal mitochondrial sequences (684 basepairs from the *COII* gene) against estimated time of divergence. Notice that the observed number of substitutions is not linear with time but curvilinear. Data from Janecek *et al.* (1996).

Models of sequence evolution

Saturation된 구간에서는 관찰되는 거리(차이; difference)가 실제 변화율과 차이가 생기게 되어 "correction (보정)"이 필요함.

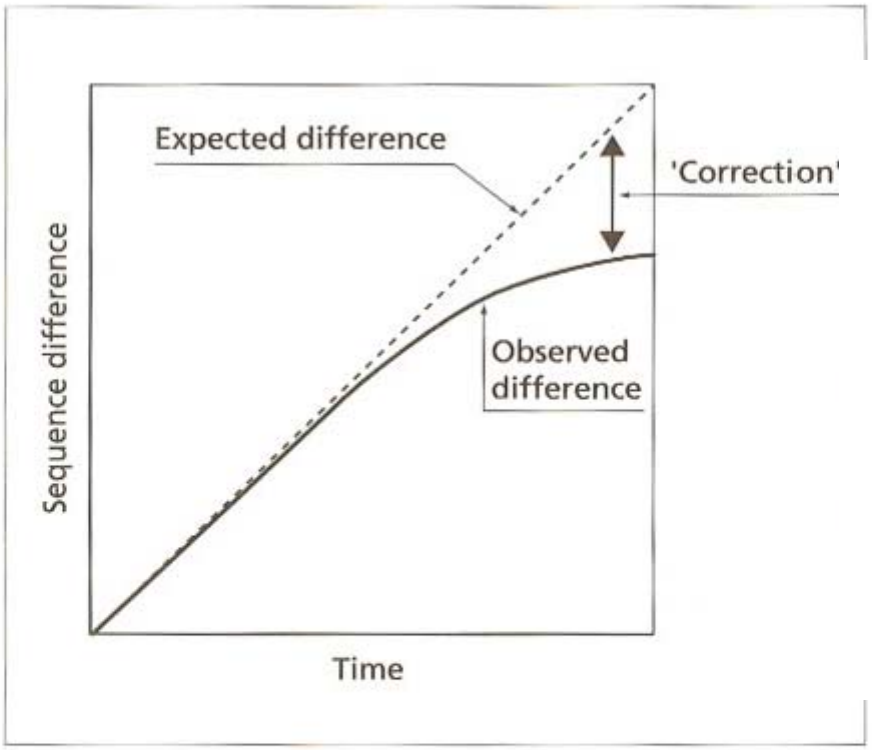


Fig. 5.12 The need to correct observed sequence differences. The extent of observed differences between two sequences is not linear with time (as we would expect if the rate of molecular evolution is approximately constant) but curvilinear due to multiple hits. The goal of distance correction methods is to recover the amount of evolutionary change that the multiple hits have overprinted and to 'correct' the distances for unobserved hits. In effect, the methods seek to 'straighten out' the line representing observed differences.

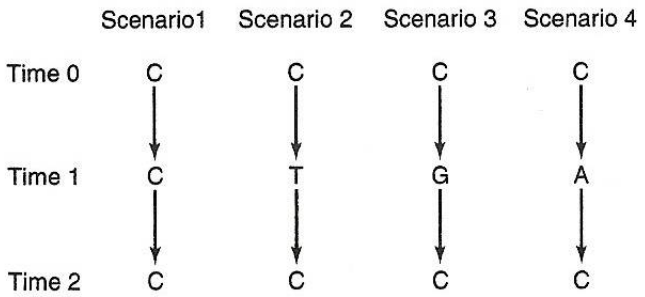


FIGURE 3.5 *Four possible routes by which a site appears to have been unchanged after two time intervals have passed.*

Estimating Substitution Numbers

K: number of substutions observed in an alignment between two sequences.
Jukes-Cantor Model, Kimura's Two-parameter Model

Jukes-Cantor Model

$P_{C(t)}=1/4 + (3/4)e^{-4\,\alpha t}$

P: 주어진 시간에 한 site에 어떤 하나의 염기가 있을 확률
- 한 사이트에 여러 번의 변화가 일어난 것을 가정
- α: 한 염기가 다른 세 개의 염기중 하나로 치환될 확률

$K = -3/4 \ln[1-(4/3)(p)]$

p: 두 서열 사이에 서로 다른 염기수

K: 다중치환이 일어날 경우 서열 위치당 발생한 실제 치환 수

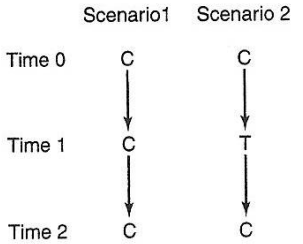


FIGURE 3.2 *Two possible scenarios where multiple substitutions at a single site would lead to underestimation of the number of substitutions that had occurred if a simple count were performed.*

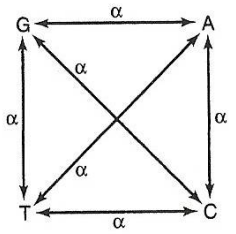


FIGURE 3.3 *Diagram of the Jukes-Cantor model of nucleotide substitution. For their model, Jukes and Cantor assumed that all nucleotides changed to each of the three alternative nucleotides at the same rate, α.*

Kimura's two parameter Model

- transitions (α) 과 transversions (β)을 고려

$K = 1/2 \ln[1/(1-2P-Q)]+1/4 \ln[1/(1-2Q)]$

P: a simple count reveals to be transitions
Q: a simple count reveals to be transversions

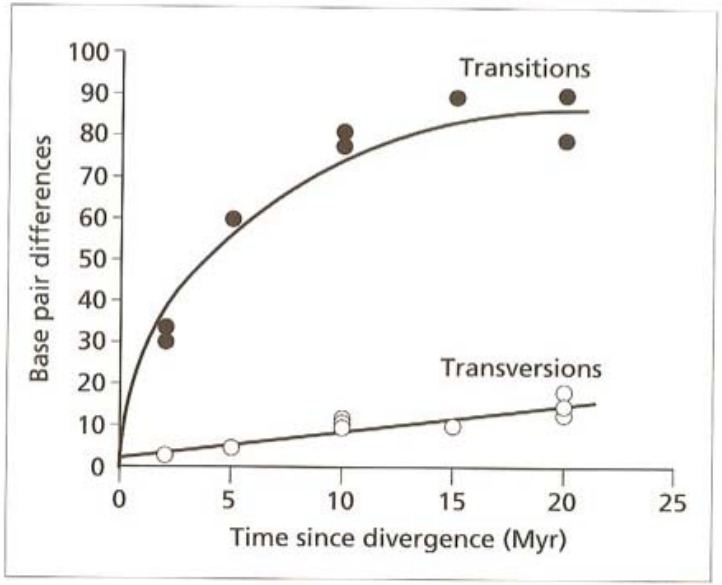


Fig. 5.13 The number of transitions and transversions between the same bovid mammal sequences used in Fig. 5.11. Transitions accumulate much more rapidly than transversions and become saturated, whereas transversions accumulate more slowly and show no evidence of saturation.

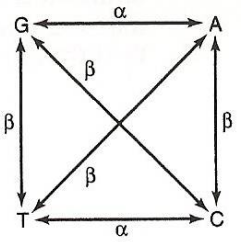


FIGURE 3.4 *Diagram of Kimura's two-parameter model of nucleotide substitution. Kimura assumed that nucleotide substitutions occurred at essentially two different rates: α for transitions (i.e., changes between G and A or between C and T), and β for transversions (changes between purines and pyrimidines).*

Models with Even More Parameters

- 12 different types of substituions are possible
- Using 12-parameter model + GC contents parameter

TABLE 3.4 Relative frequencies of nucleotide substitutions in Alu-Y (Sb) sequences throughout the human genome.

	To				
From	A	T	C	G	Row Totals
A	—	4.0	4.6	9.8	18.4
	—	(1.5)	(1.7)	(3.6)	(6.7)
T	3.3	—	10.4	2.7	16.4
	(1.2)	—	(3.8)	(1.0)	(6.0)
C	7.2	17.0	—	6.2	31.1
	(5.0)	(33.2)	—	(4.5)	(42.6)
G	23.6	4.6	6.0	—	34.2
	(37.7)	(3.2)	(3.7)	—	(44.7)
Column totals	34.1	26.3	21.0	9.0	
	(44.0)	(37.8)	(9.2)	(18.7)	

Note: Members of the *Alu* repeat family are approximately 260 base pairs in length. They are derived from one or a small number of ancestral sequences that have been duplicated almost 1 million times during primate evolution.

The relative frequencies of substitutions observed involving each of the four nucleotides within 403 Alu-Y (Sb) repeat sequences scattered throughout the human genome excluding those involving CpG dinucleotides. Values in parentheses were obtained when substitutions at CpG dinucleotides were not excluded. A total of 7,433 substitutions (2,713 of which were at sites other than CpG dinucleotides) were accumulated by the 403 Alu-Y (Sb) repeats included within this analysis since they were propagated roughly 19 million years ago.

Models for estimating the number of nucleotide substitutions among a pair of DNA sequences

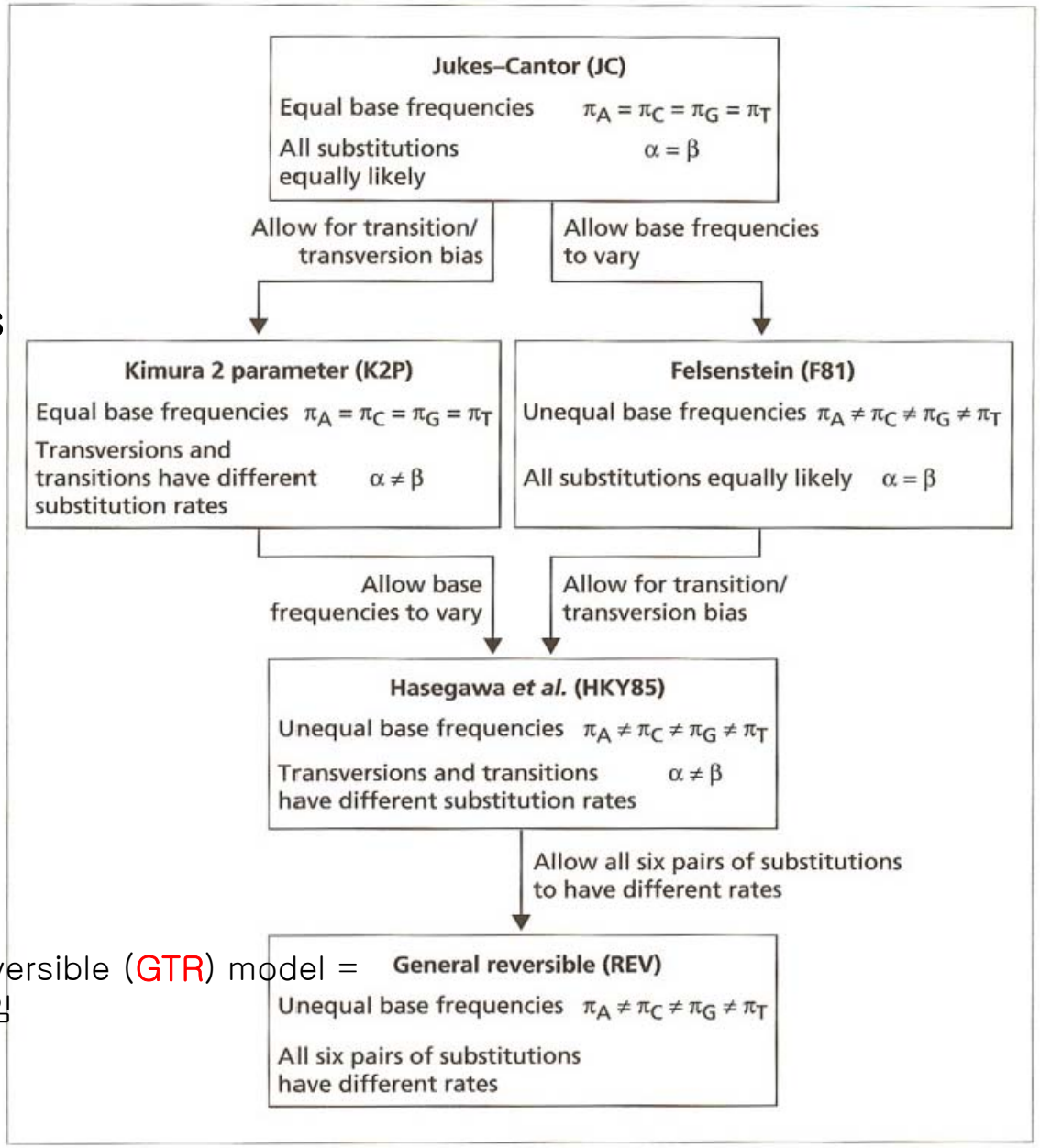


Fig. 5.14 Interrelationships among five models for estimating the number of nucleotide substitutions among a pair of DNA sequences. The JC, K2P, F81 and HKY85 models can all be generated by constraining various parameters of the REV model.

General Time Reversible (GTR) model = 가장 복잡한 모델임

원의 크기: base frequency
원의 색(gray, black, white): 다른 substitution들의 비율

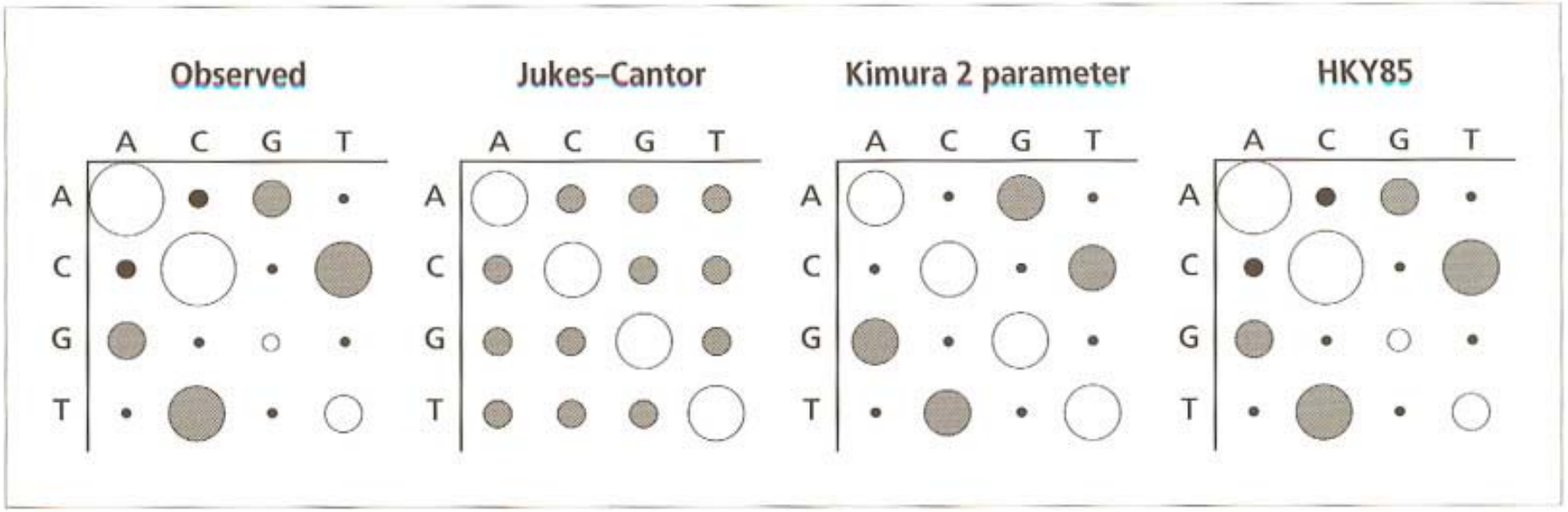


Fig. 5.15 Observed and expected numbers of nucleotide pairs between human and chimpanzee mtDNA sequences for three different models. As the models add parameters they more closely approximate the observed pattern. Data from Tamura (1994).

MODEL TEST: a program finding the best model for given data
http://darwin.uvigo.es/software/modeltest.html

Relative Rate Test

MEASURING EVOLUTIONARY CHANGES ON A TREE

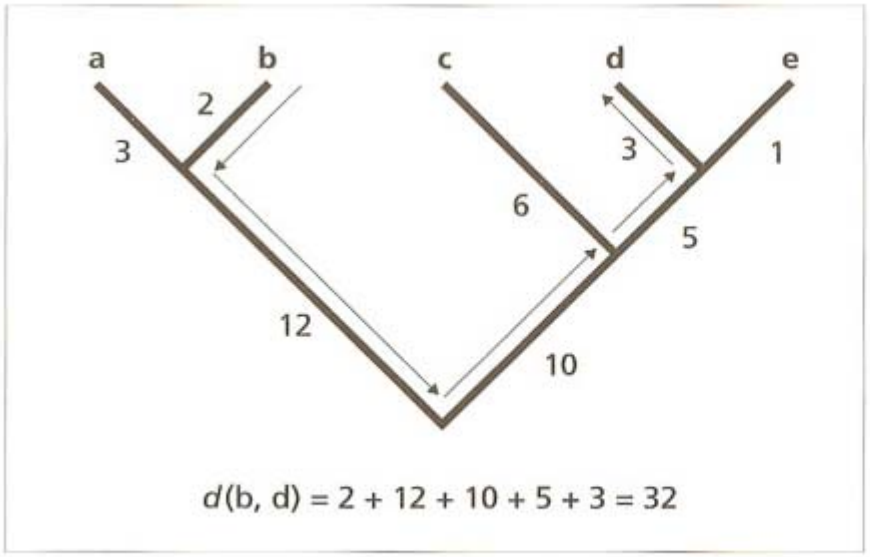


Fig. 5.21 The evolutionary distance between b and d is the sum of the edge lengths along the path in the tree between the two sequences.

Molecular Clock

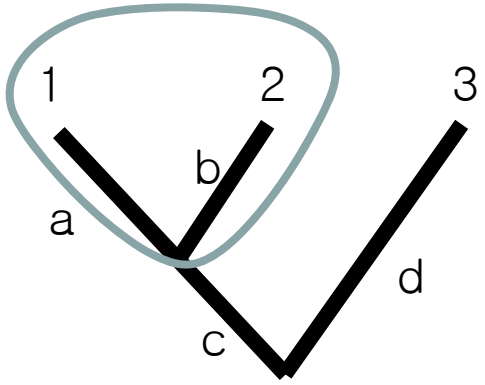
Zuckerkandl and Pauling 1960': substitution rate were so constant within homologous proteins over many tens of millions of years that they likened the accumulation of amino acid changes to the steady ticking of a **molecular clock.**

Steady rate change → apply to the determination of phylogenetic relationships
→ divergence time estimation

However, molecular clock is a controversial hypothesis.

Relative rate test to check the molecular clock in certain group

To determine relative rate of substitution between 1 and 2 (=compare "a" to "b"), we need to designate a less related species 3 as an OUTGROUP.



If d(1,3) = d(2,3),
a + c + d = b + c + d
Therefore "a = b".

Relative rate of substitution between "a" and "b" is equal.
→ Molecular clock is applicable in this group (1 and 2).

Molecular Clock

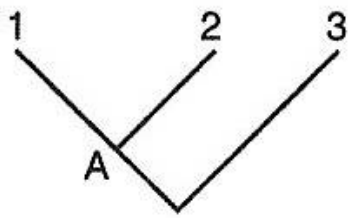


FIGURE 3.7
Phylogenetic tree used in a relative rate test. Species 3 represents an outgroup known to have been evolving independently prior to the divergence of species 1 and species 2. "A" denotes the common ancestor of species 1 and species 2.

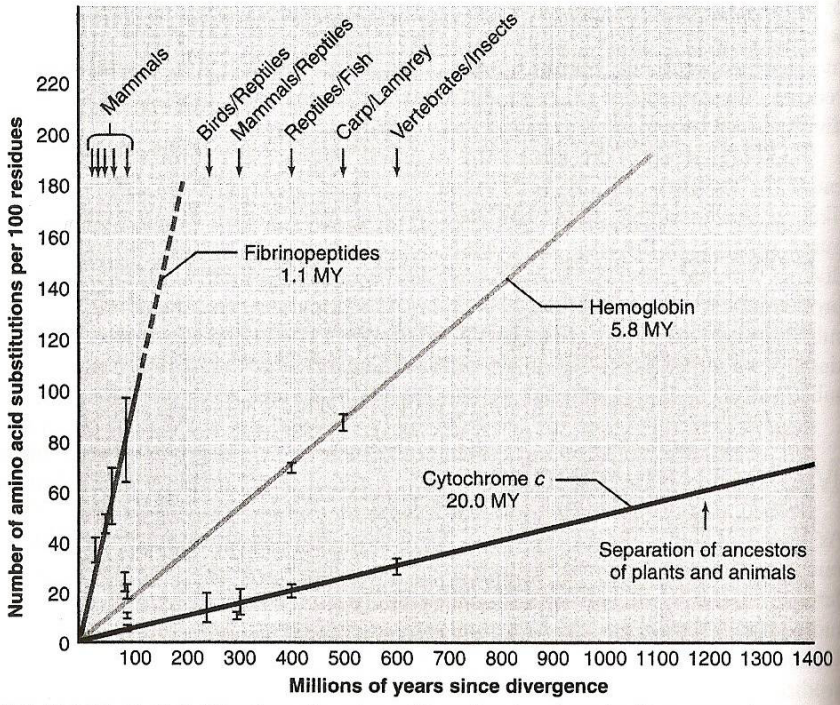


FIGURE 3.6 *Numbers of amino acids replaced and species divergence times are well correlated for a number of proteins.*