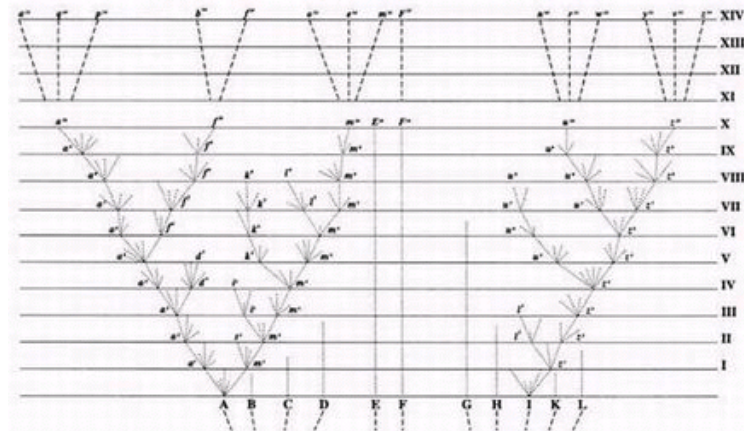
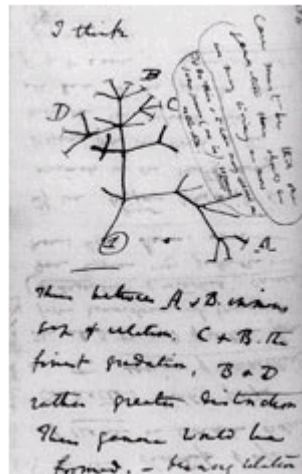


2024년도 한국진화학회 겨울학교

(분자진화학분야)

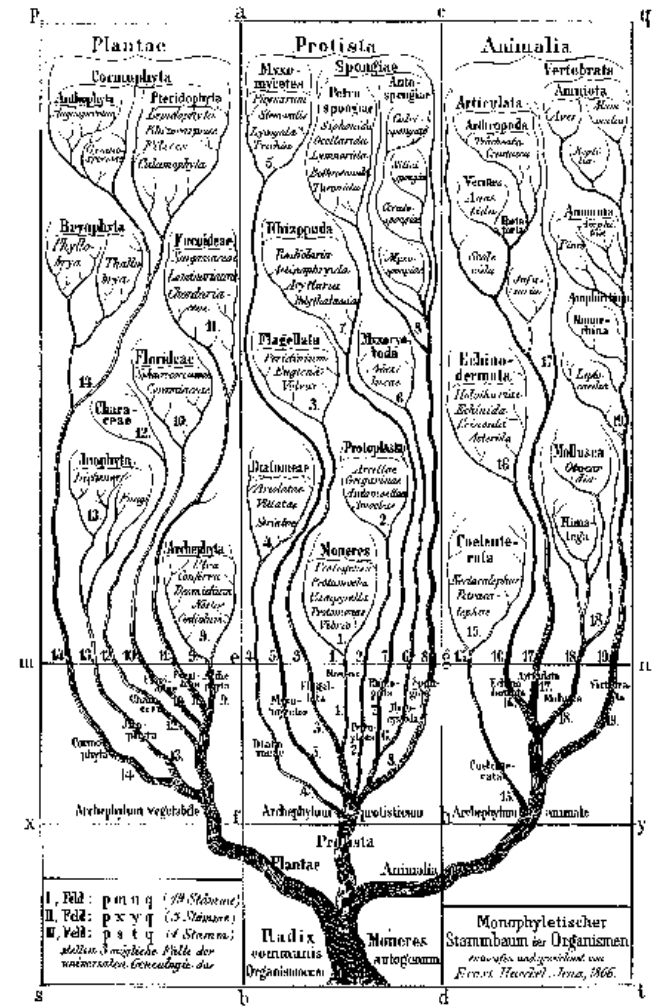
극지연구소
서태건

계통수(phylogenetic tree)와 생물의 진화/유연 관계



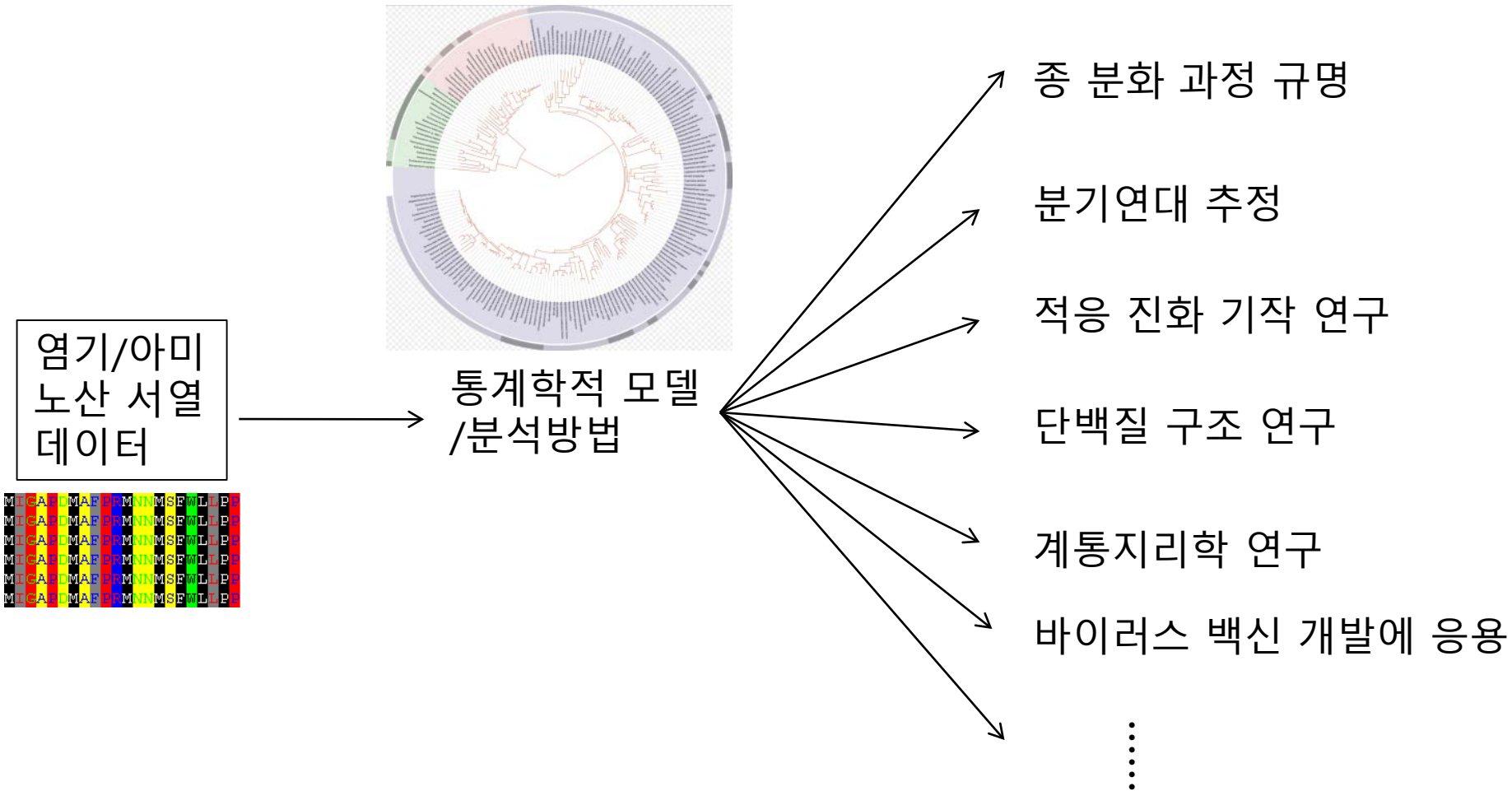
Darwin's notebook (1837)

On the Origin of Species by Natural Selection (Darwin, 1859)

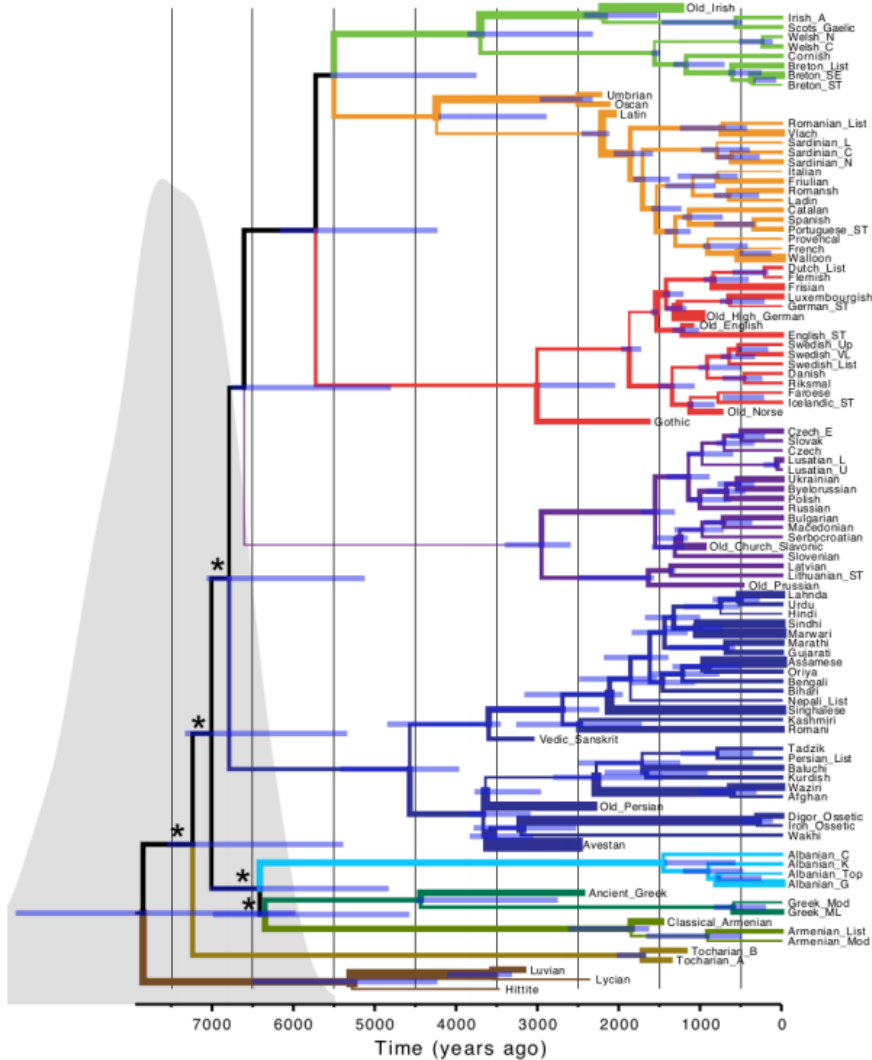


Generelle Morphologie der Organismen (Haeckel 1866) 2

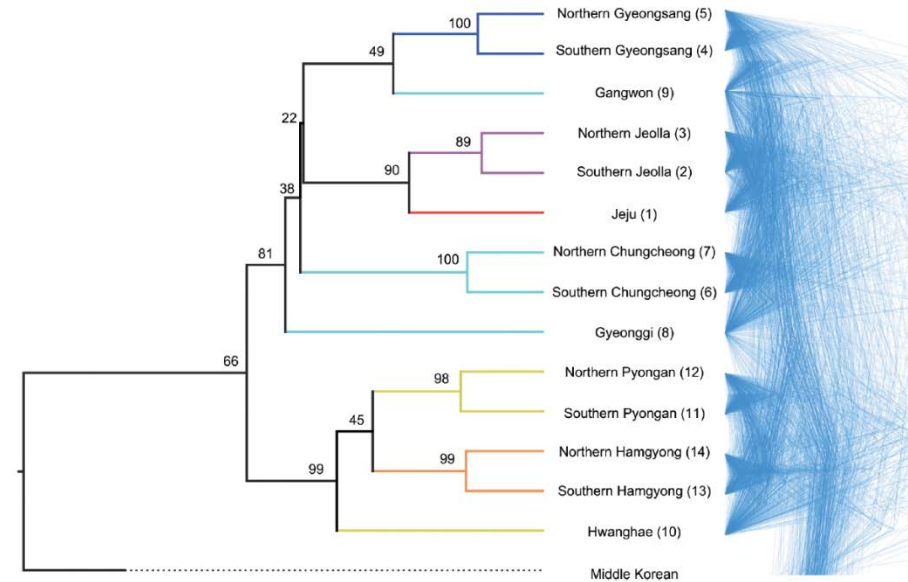
분자계통수의 추정이 다른 연구에 미치는 영향은 매우 크다



계통수 추정은 무생물 (언어)의 진화과정 규명에도 이용된다

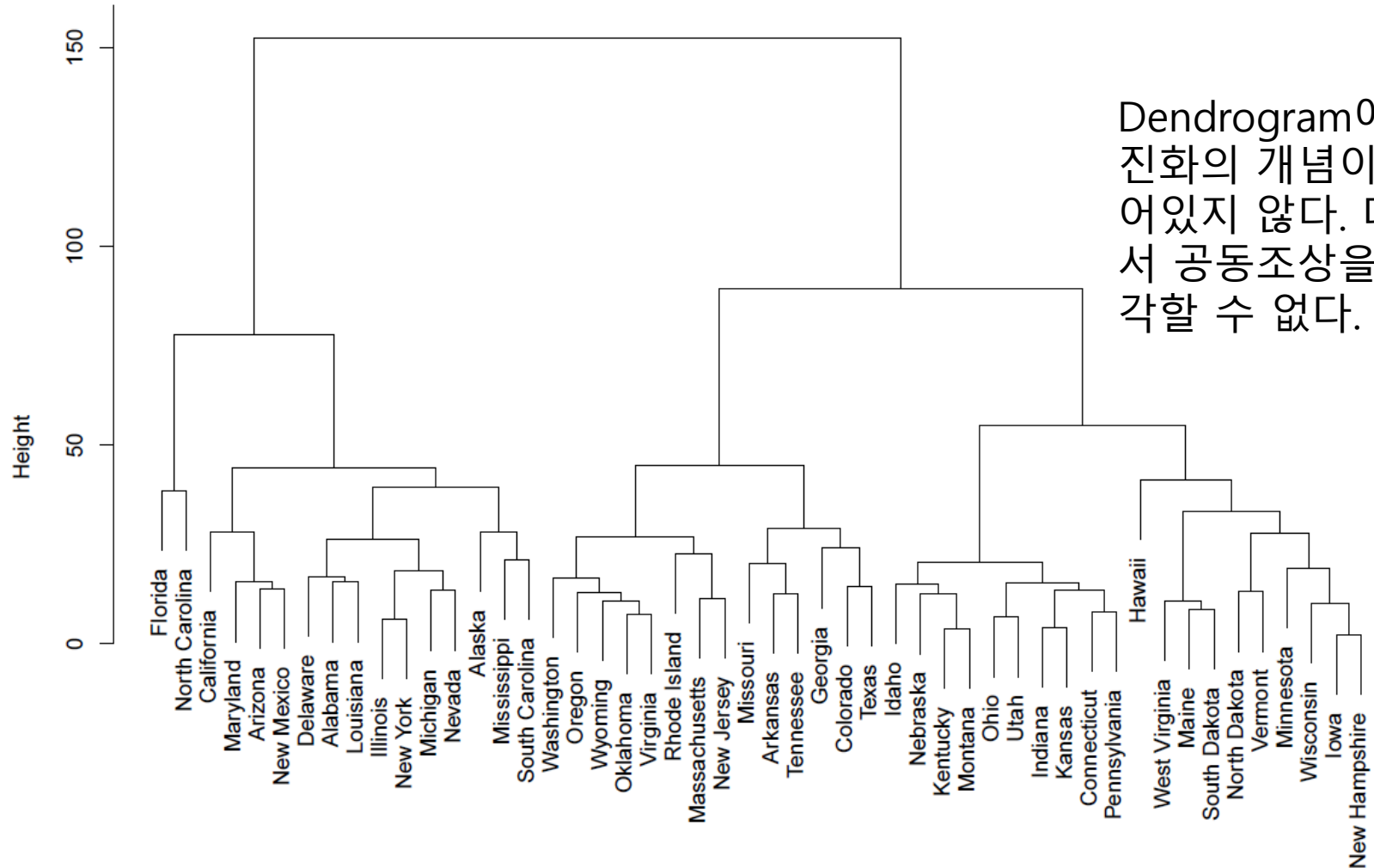


(Bouckaert et al. 2012)



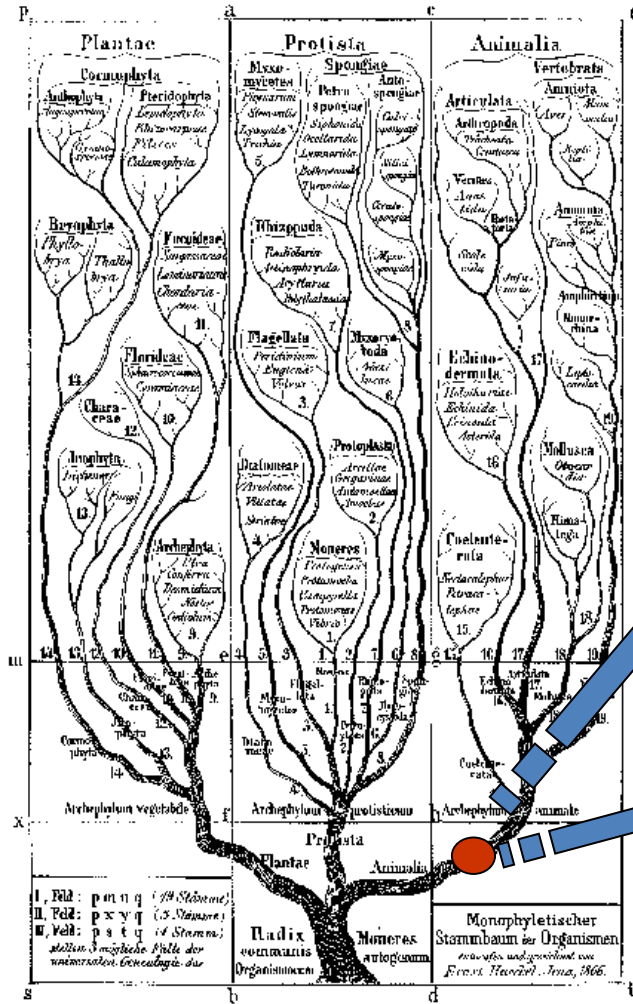
(Lee 2015)

Phylogeny 와 dendrogram(clustering)은 다르다

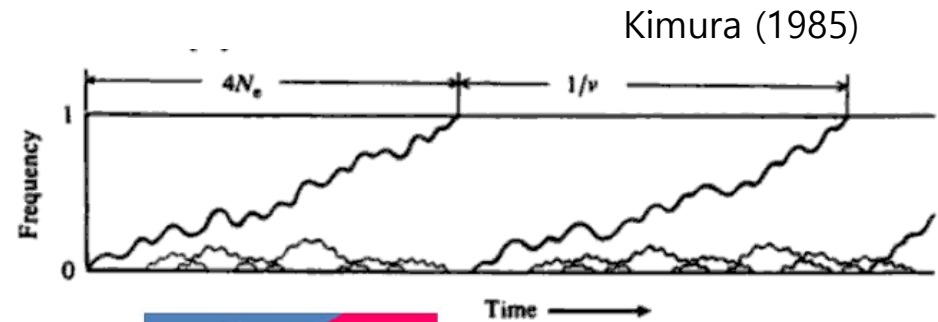


범죄발생 유사도에 의해 미국50개 주를 clustering 한 결과

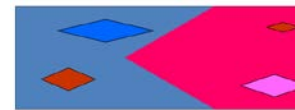
계통수와 집단유전학과의 관계



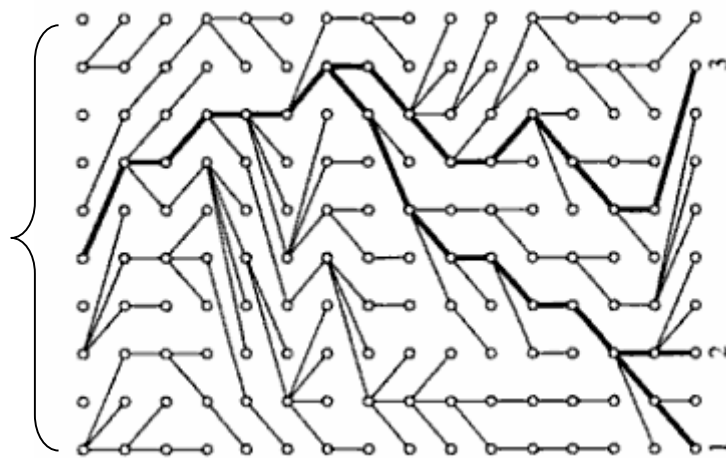
Haeckel (1866)



Kimura (1985)



$\theta(=4N\mu)$



time
Coalescent theory, Hein et al. (2005)

돌연변이의 발생과 집단내에서의 고정

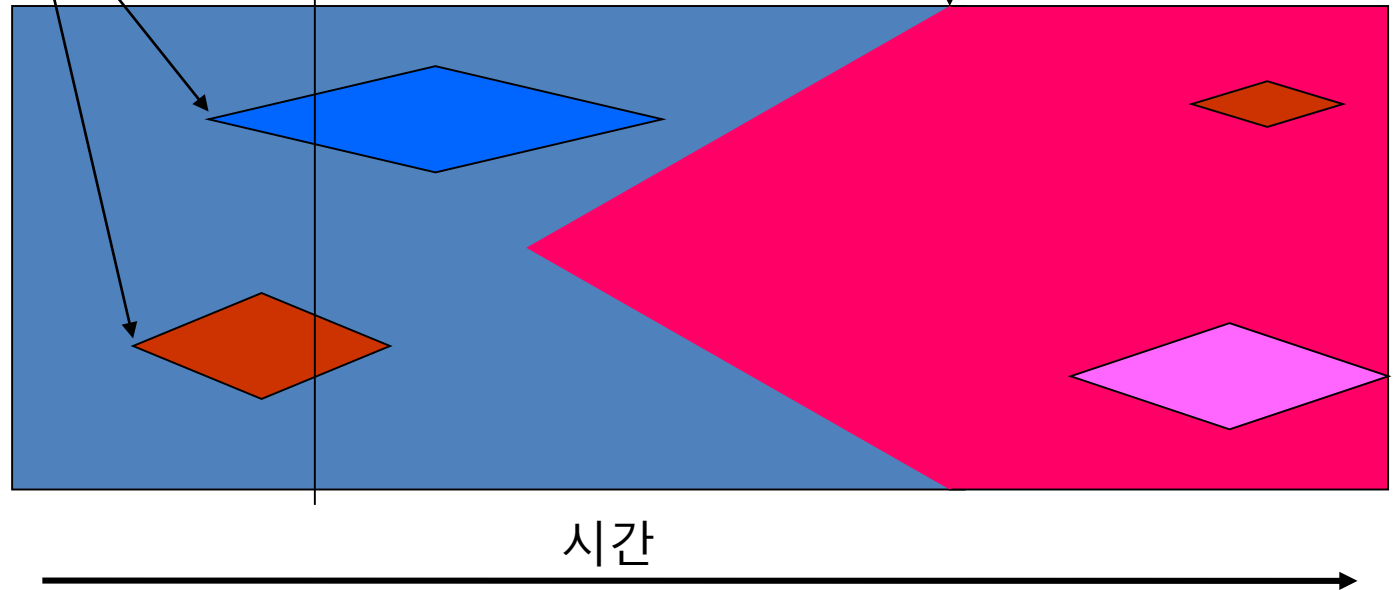
새로운 돌연변이 발생

이 시점에서 개체들의 유전자 타입은 다양성을 보임 (→Polymorphic site)

고정(Fixation): 이 시점 이후의 모든 개체들은 이 타입의 돌연변이를 가짐. 생물종간의 비교에는 고정된 돌연변이만 고려

집단내에서의 돌연변이 타입을 서로 다른색으로 표현

세로축의 폭은 집단의 크기를 나타냄



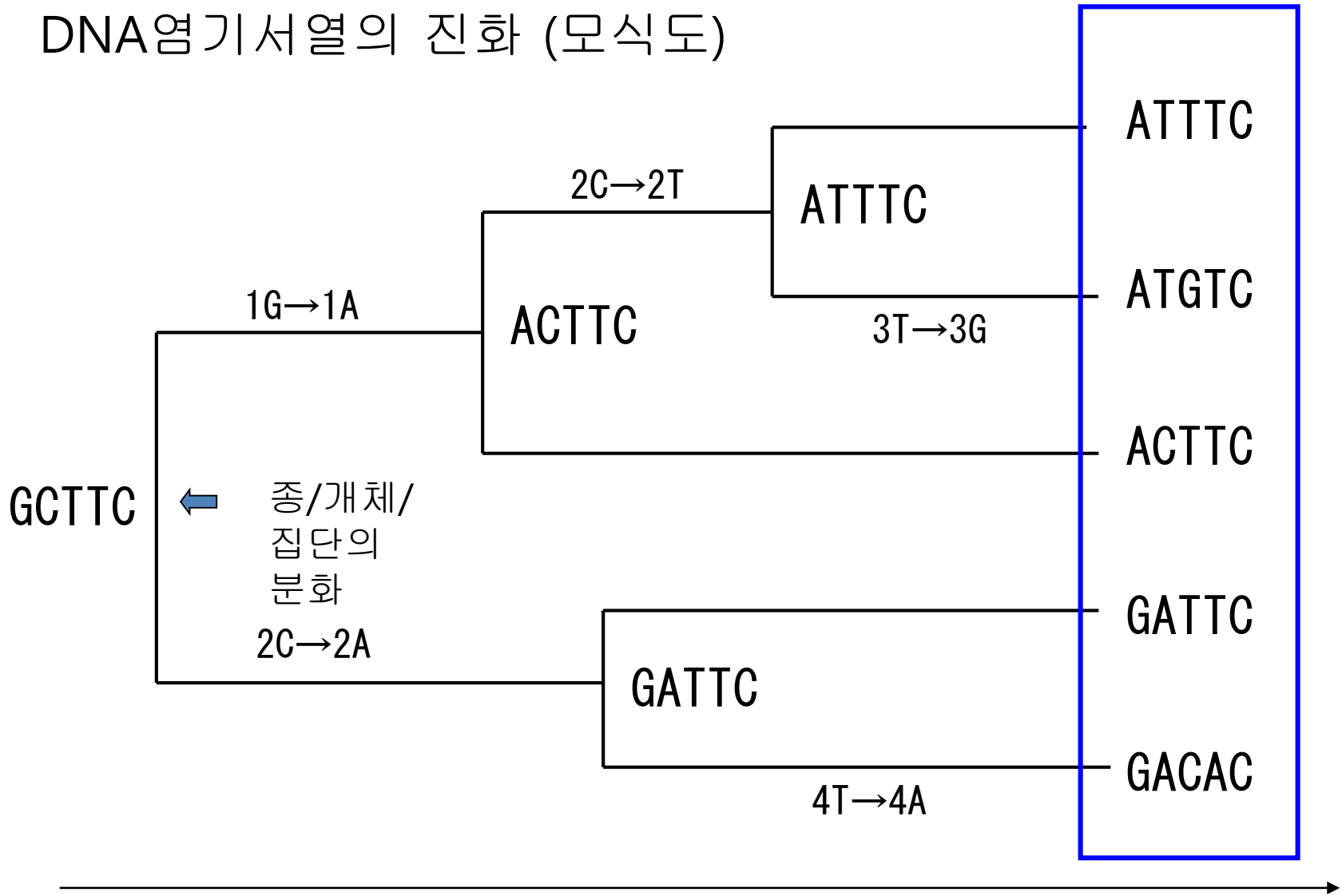
집단의 다양성을 나타내는 척도

$$\theta (= 4N\mu)$$

N: effective population size (집단의 크기)

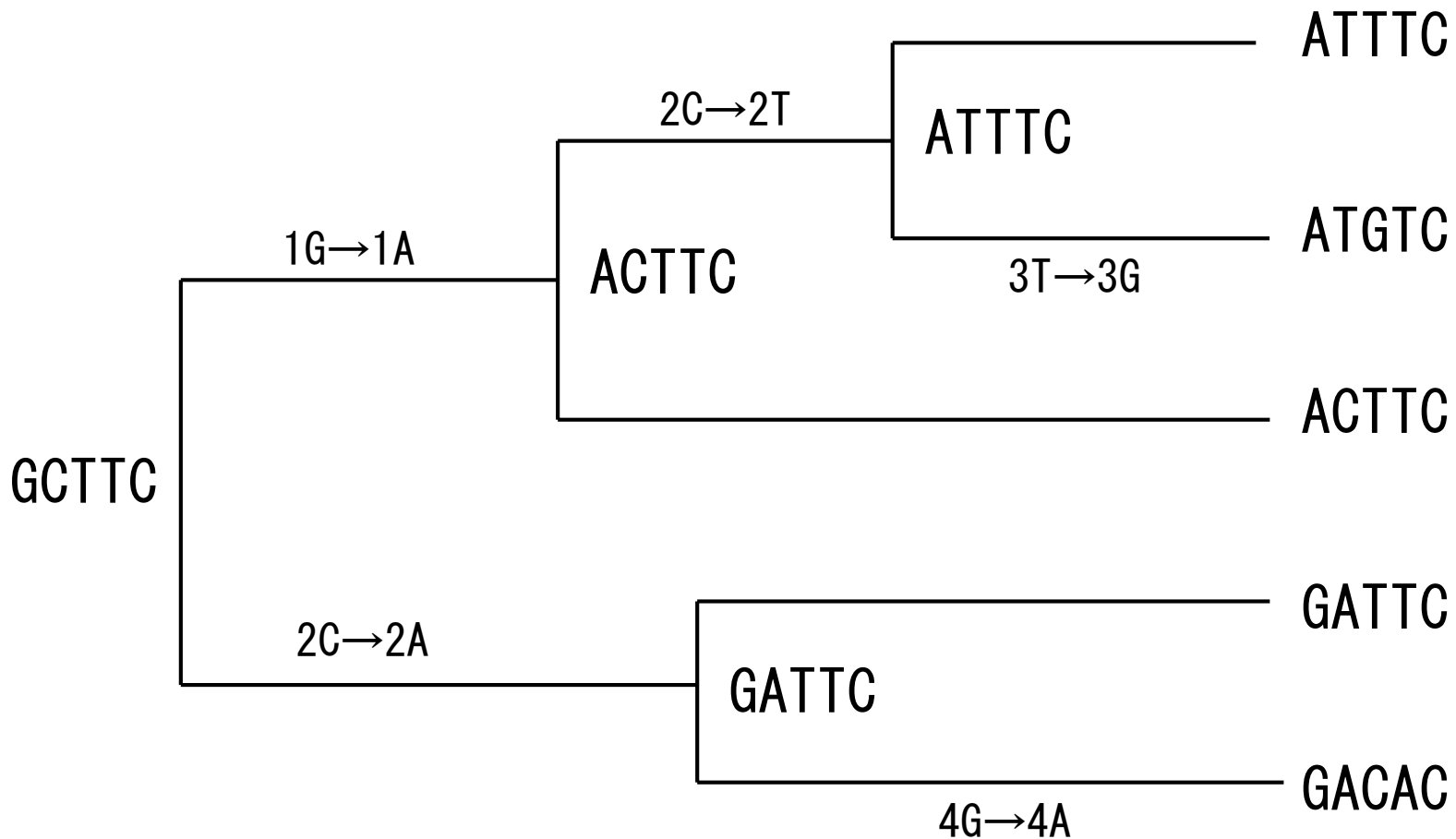
μ : mutation rate per generation (세대당 돌연변이율)

DNA염기서열의 진화 (모식도)



실제 관측가능한것
은 현재의 데이터

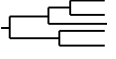





DNA염기서열의 진화 (모식도)

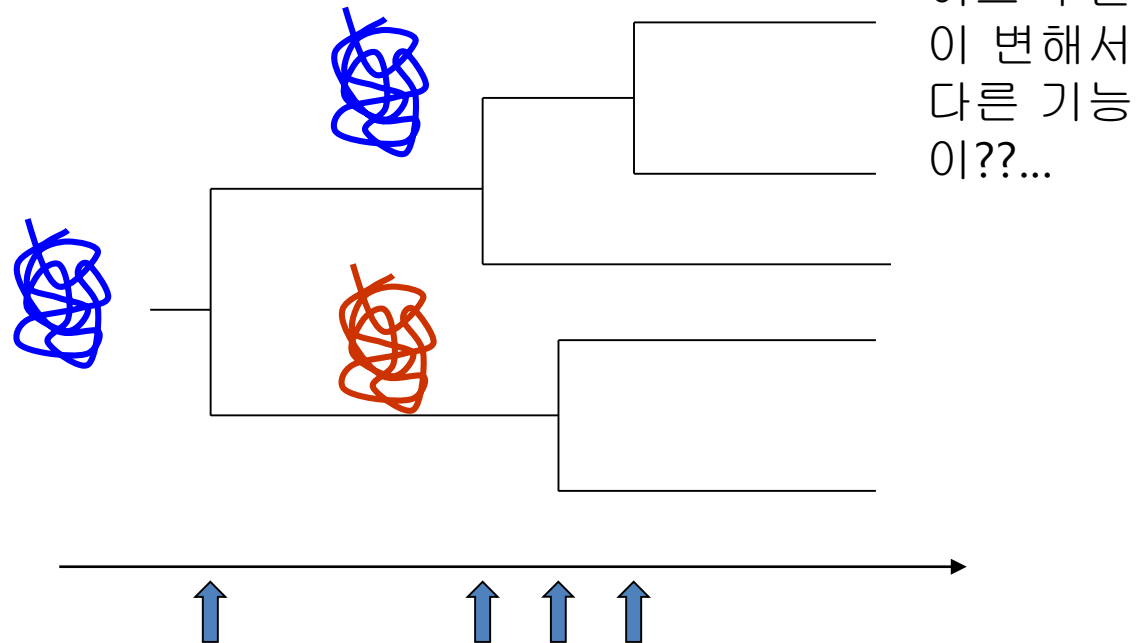


Time

계통관계를 추정

실제 관측가능한것
은 현재의 데이터

- 분자진화학의 데이터분석으로 가능한것들의 예
 - 계통관계의 추정 ()
 - 분기연대의 추정 ()
 - 조상의 유전자배열, 단백질 구조의 추정 ()
 - 기능의 변화 과정의 추정 ( →  )



분자계통수로 추정하는 생명의 기원
(HIV의 예)

美의료계 非常

했다. 원인이 밝혀지지 않은 불
가사의 한 별과 과거에 자취를

이후

감추었던 별이 다시 나타나기
 슬슬 부리고있기 때문이다.
 후천성면역결핍증(AIDS)은
 지금까지 원인이 알려지지 않은
 별과운데 가장 치명적이며 무
 서운 속도로 번지고있어 미국
 의학계의 골칫거리로 나타나고

[illegible]

사망을 38%의 「면역결핍증」: 輸血감염 추정

점액 여성들에 「毒性중독증」 발생 79명 사망

한때 자취감했던 나병·말라리아·노예노동

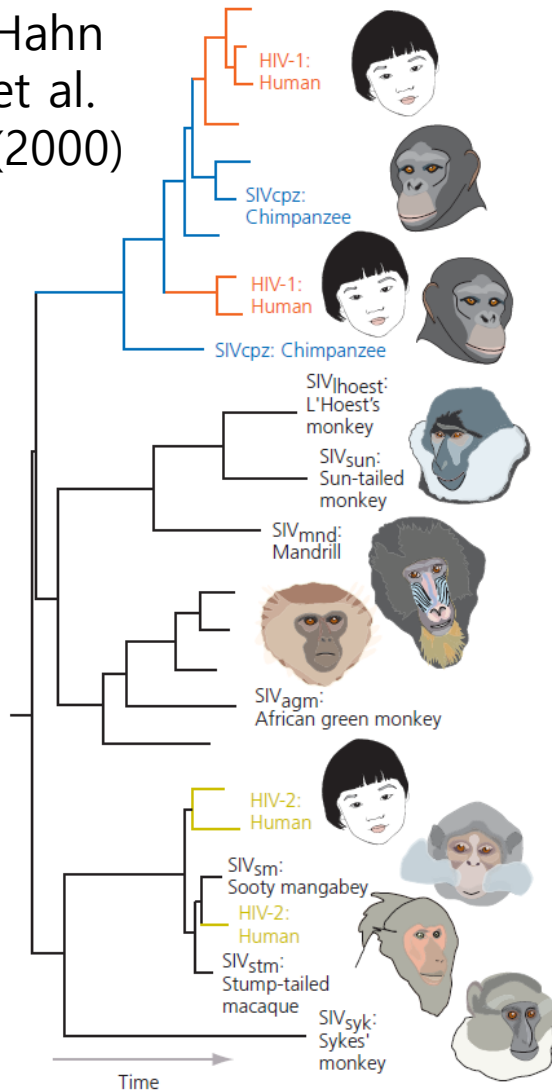
[illegible][illegible][illegible]

한자 漢字 漢字은 50
타 100까지
여종의 새로운 배신이 개발
이지만 앞으로는 어떤으로
지않는 새군이 나타날 모
다수 우월로

- HIV (Human Immunodeficiency Virus)가 AIDS를 유발
- 20세기 이전에 AIDS (후천성 면역결핍증)는 지구상에 존재하지 않았던 질병
- HIV도 20세기 이전에 존재하지 않았던 바이러스
- HIV가 어떻게 세상에 등장하게 되었을까?

DNA 염기서열을 이용한 HIV의 기원 추정

Hahn
et al.
(2000)



- HIV 은 침팬지를 감염시키는
SIV (Simian Immunodeficiency Virus)와
진화적으로 가깝다

- HIV-1, HIV-2는 단계통이 아님

Zoonosis (=zoonotic transmission): 바이러스
의 숙주의 영역을 넘어서 다른 숙주를 감염시키
게 되는 현상

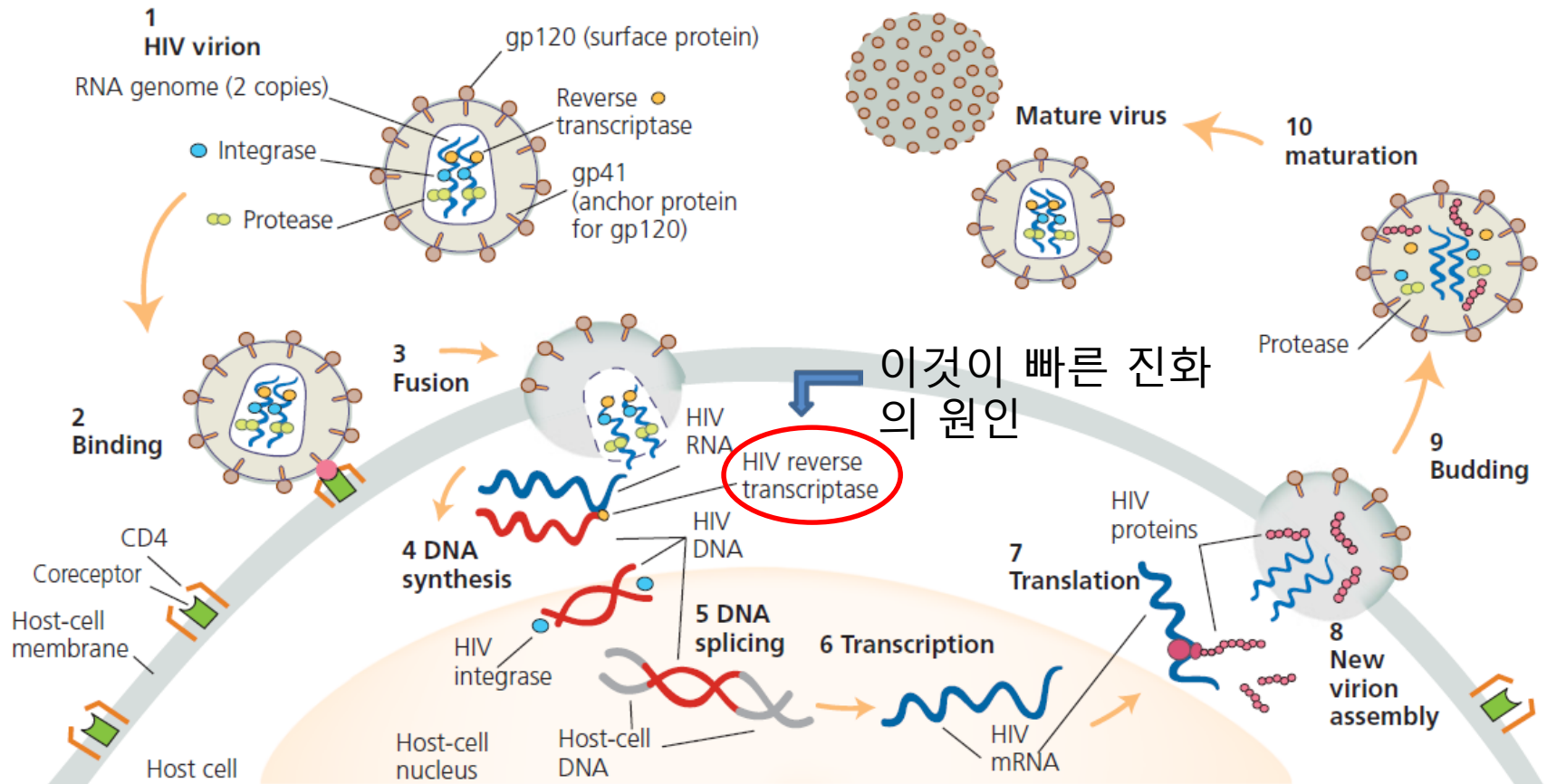
FIV (Feline ~) → 고양이과를 감염시킴

SIV (Simian ~) → 원숭이 부류를 감염시킴

↘ **HIV → 인간에게 감염시킬수
있는 능력획득**

(그림 출처 : Herron and Freeman 2014 Evolutionary Analysis 5/e)

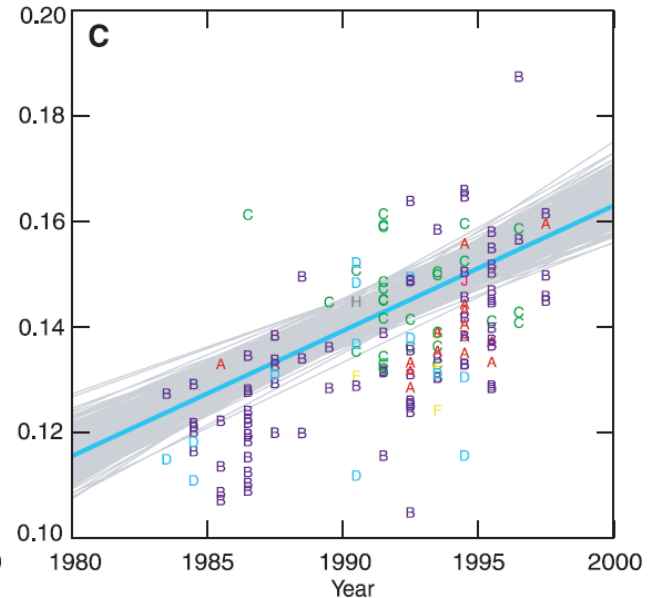
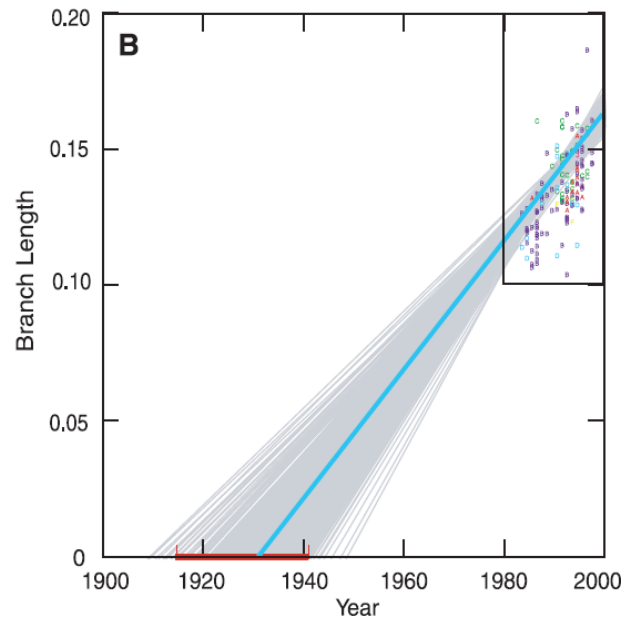
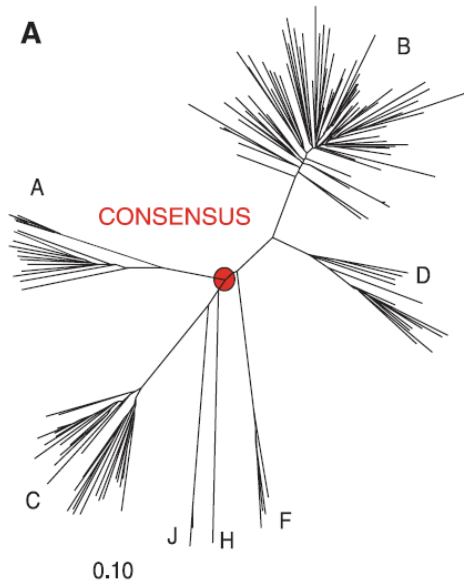
HIV는 왜 빨리 진화하는가?



HIV의 life cycle (Herron & Freeman 2014)

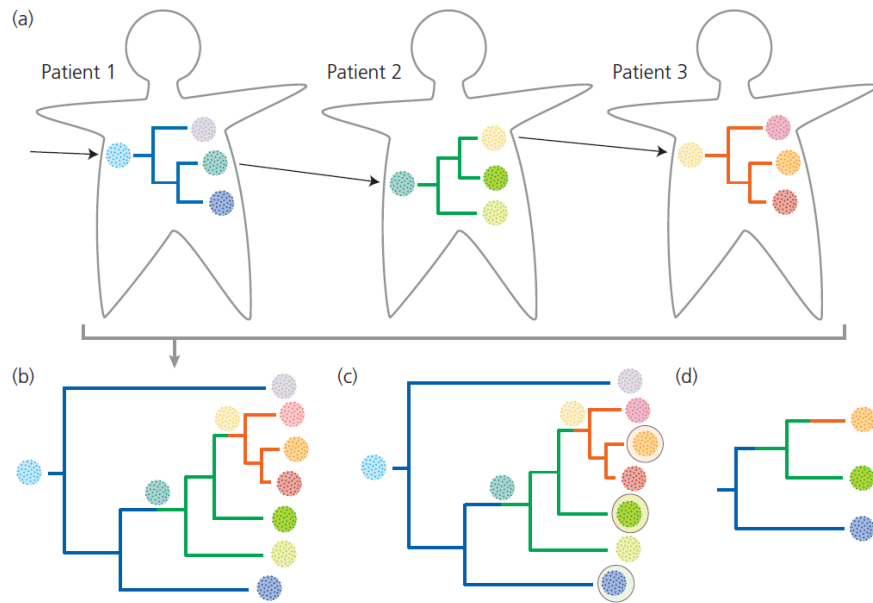
HIV-1 의 최초 등장 시기 추정

(Korber et al. 2000)

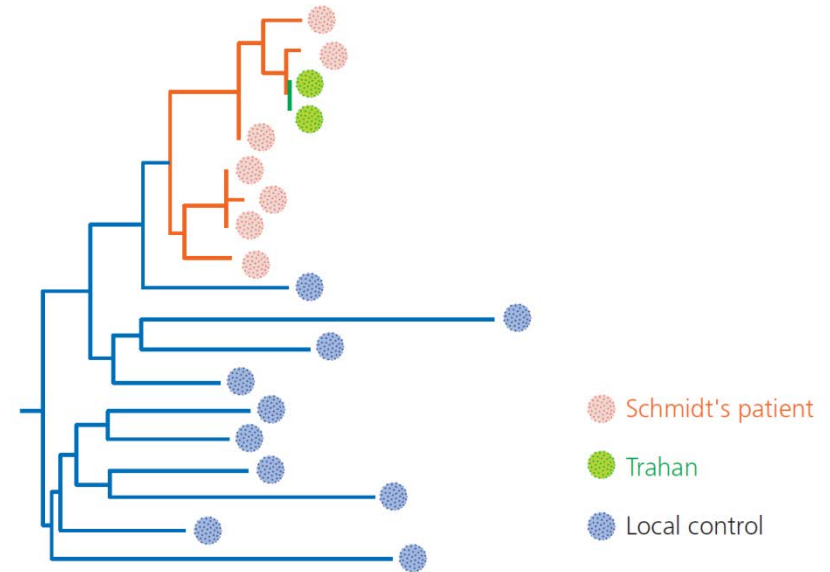


- 매년 DNA 치환량을 측정하여 회귀직선을 구하면 공동조상의 연대를 추정할 수 있다 ← 분자시계 (molecular clock) 개념 이용
- HIV가 최초 등장한 시기는 대략 1931년 (1915~1941)
- 이 결과는 OPV (Oral Polio Vaccine) AIDS 가설을 반증하는 증거의 하나가 됨

법의학분야에서도 사용되는 분자계통수



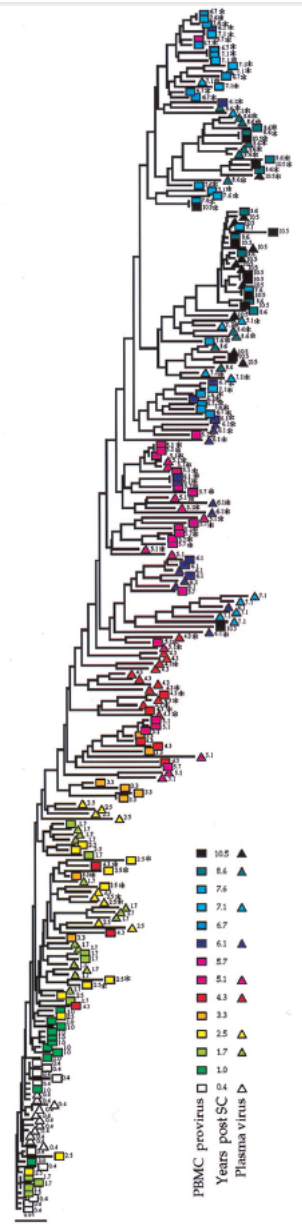
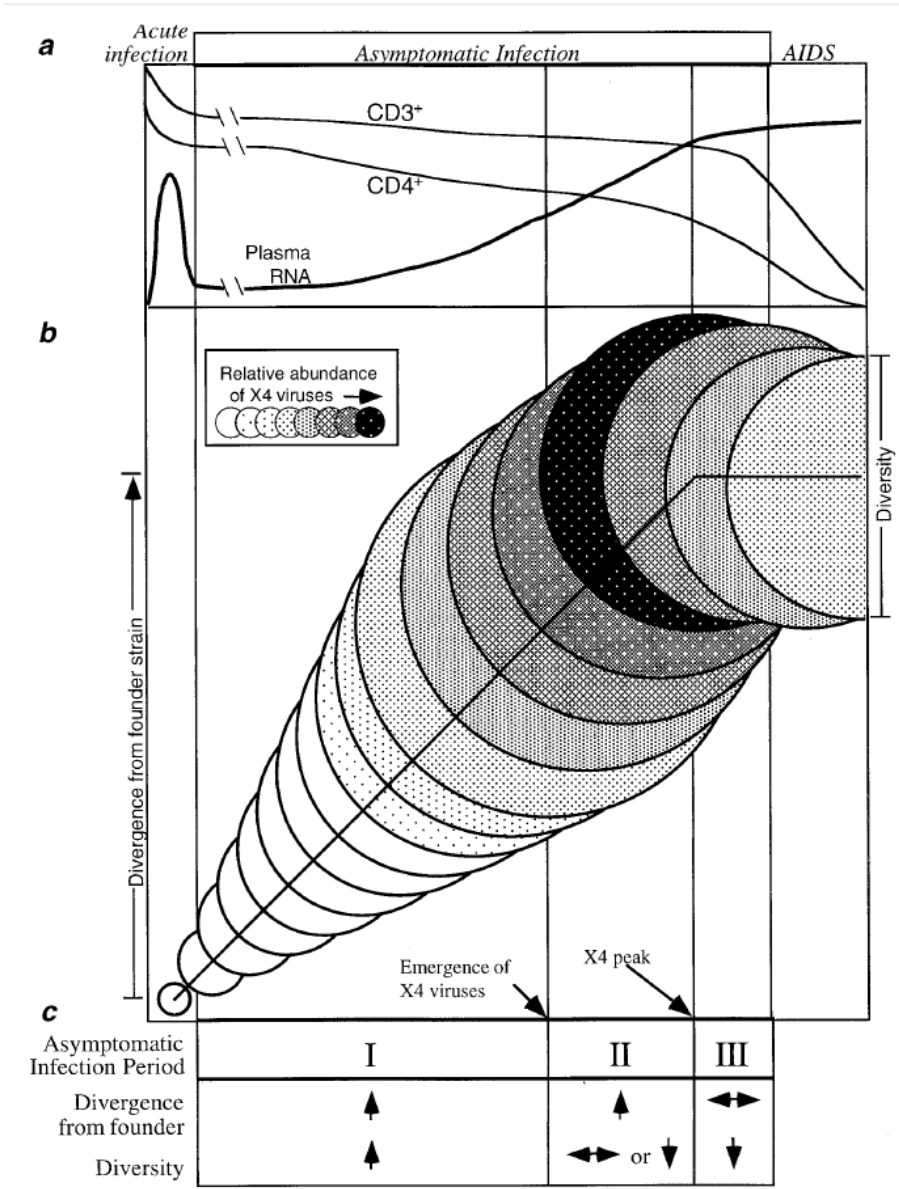
환자에서 환자로 HIV가 전이되는 양상과 각각의 분자계통수



미국 Louisiana 주의 내과 의사 Schmidt가 내연관계의 간호사 Trahan에게 고의로 HIV를 주사했다고 의심되는 사건 → 징역 50년 유죄판결 확정 (2000년)

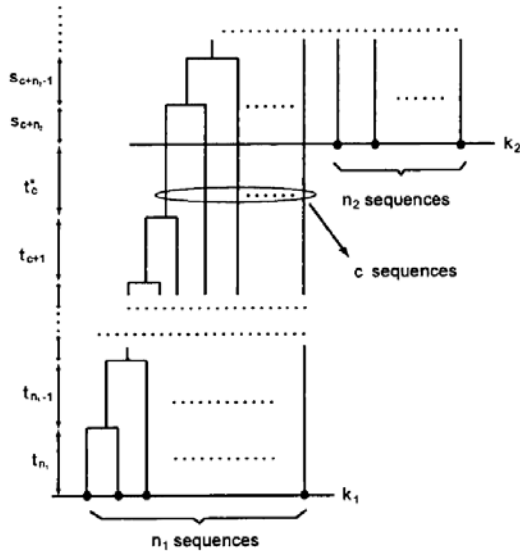
(그림 출처 : Herron and Freeman 2014 Evolutionary Analysis 5/e)

환자의 체내에서 HIV가 진화하는 방식



(Shankarppa et al 1999)

Coalescent 이론을 이용한 환자 체내의 HIV의 집단의 크기 추정 (Seo et al. 2002)



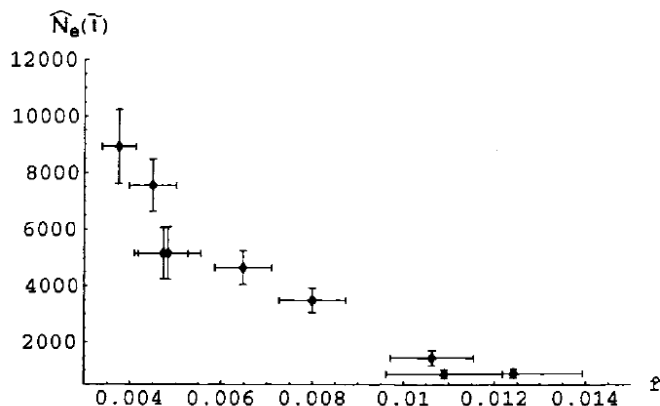
$$L_1 = \left[\prod_{i=n_1}^{c+1} p(t_i | e(t_i), N_e) \right] \times p(t_c^*, s_{c+n_2} | e(t_c^*), k_2, N_e) \\ \times \left[\prod_{i=c+n_2-1}^2 p(s_i | e(s_i), N_e) \right]$$

$$p(t_i | e(t_i), N_e) = \frac{i(i-1)}{2N_e} \exp\left(\frac{-i(i-1)}{2N_e} t_i\right)$$

$$p(t_c^*, s_{c+n_2} | e(t_c^*), k_2, N_e) = \exp\left(-\frac{c(c-1)}{2N_e} t_c^*\right) \frac{(c+n_2)(c+n_2-1)}{2N_e} \\ \times \exp\left(\frac{-(c+n_2)(c+n_2-1)}{2N_e} s_{c+n_2}\right)$$

$$p(s_i | e(s_i), N_e) = \frac{i(i-1)}{2N_e} \exp\left(\frac{-i(i-1)}{2N_e} s_i\right)$$

좌측의 계통수가 얻어지는 가능성을 likelihood 함수로 표현



진화속도와 집단의 크기사이에는 음의 상관관계가 있다. 이를 환자의 면역체계와 관련하여 설명 가능

면역력 강함 → 바이러스 집단 감소 & 바이러스 빨리 진화
면역력 약함 → 바이러스 집단 증가 & 바이러스 천천히 진화

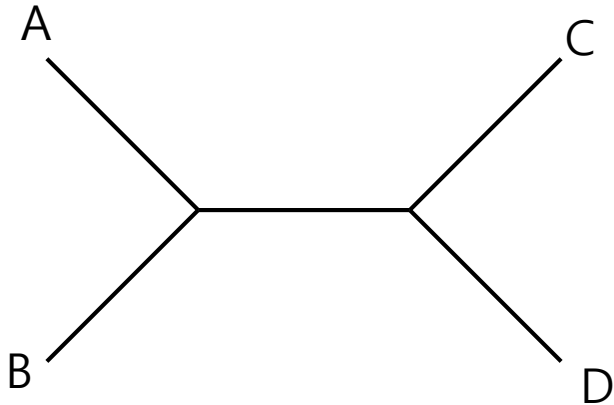
계통수 추정법의 종류

- Maximum parsimony
 - nonparametric; DNA 치환 모델 가정 X
 - long branch attraction
- Distance matrix method
 - NJ (Neighbor-Joining method), UPGMA
- Maximum likelihood method
 - parametric; DNA 치환 모델 가정
- Bayesian method
 - posterior \propto prior \times likelihood

분자계통수에 관한 기본 사항

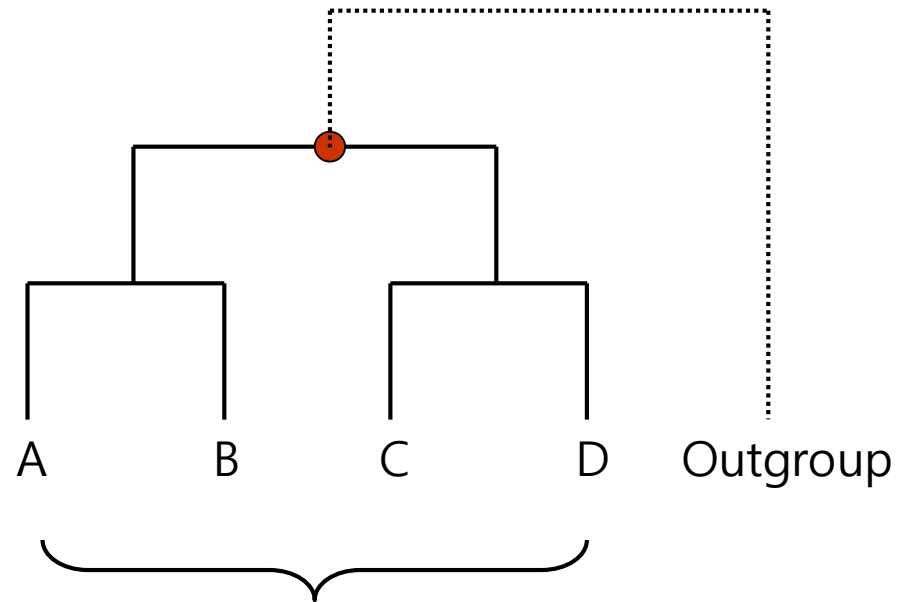
Phylogeny, phylogenetic tree (계통수)

Unrooted tree



$((A, B), C, D)$; 혹은 $((C, D), A, B)$;

Rooted tree



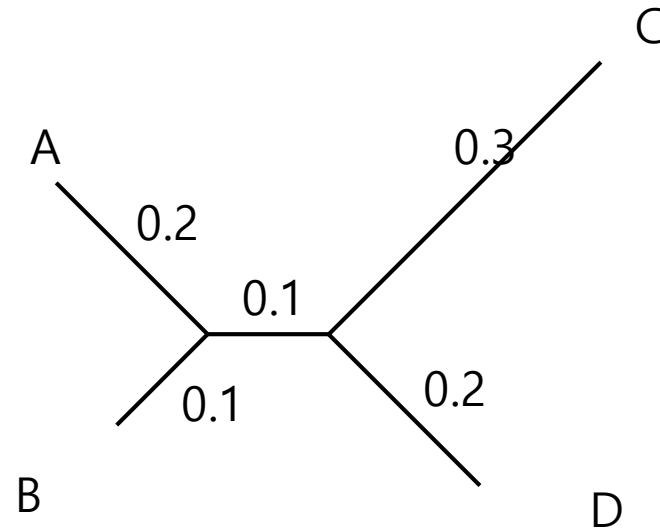
$(((A, B), (C, D)), \text{Outgroup})$;

* Newick (New Hampshire) tree format

- Evolutionary distance (진화적인 거리)

Branch length 는 진화의 거리(사이트당 평균적으로 일어나는 염기치환수) 나타냄

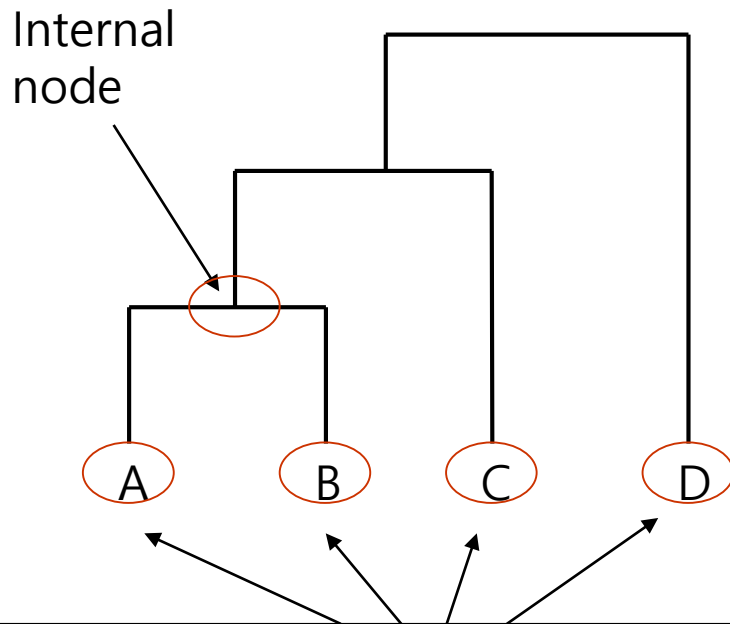
예) 0.2 : 1사이트중 평균적으로 2회 치환이 일어남을 의미



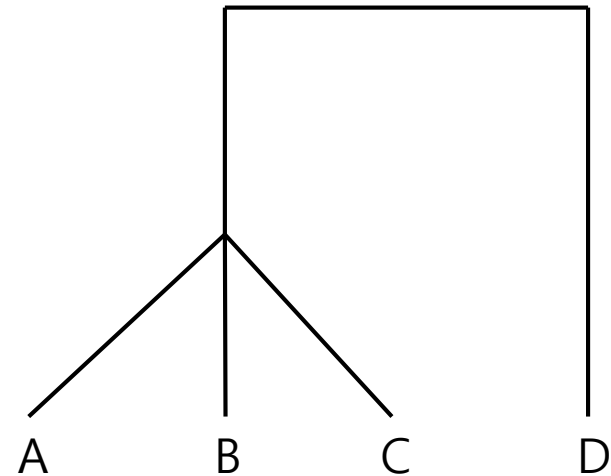
$((A:0.2, B:0.1):0.1, C:0.3, D:0.2);$ 혹은
 $((C:0.3, D:0.2):0.1, A:0.2, B:0.1);$ 로 계통수를 표시함

- 계통수의 형태

Bifurcating tree

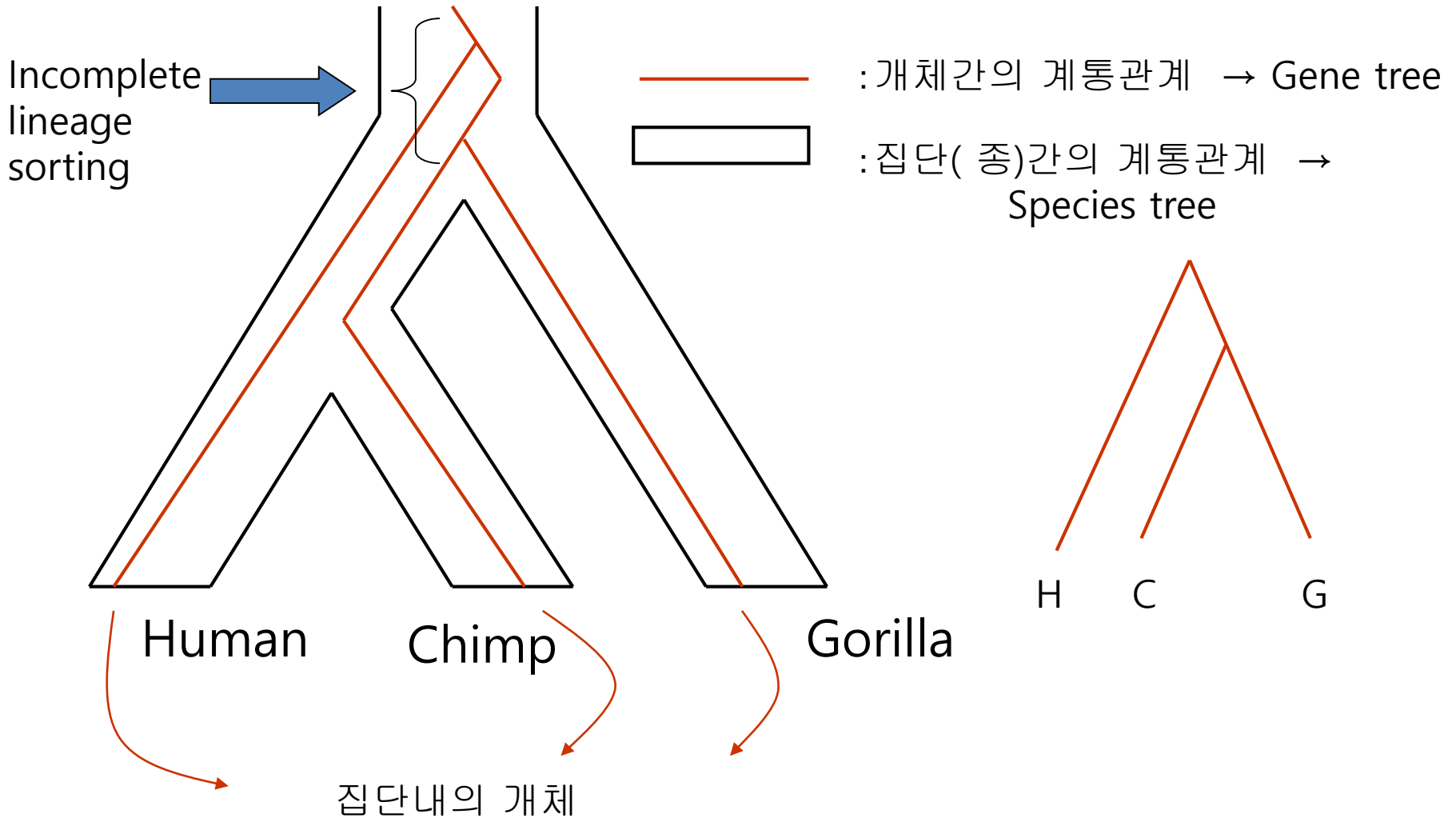


Multifurcating tree

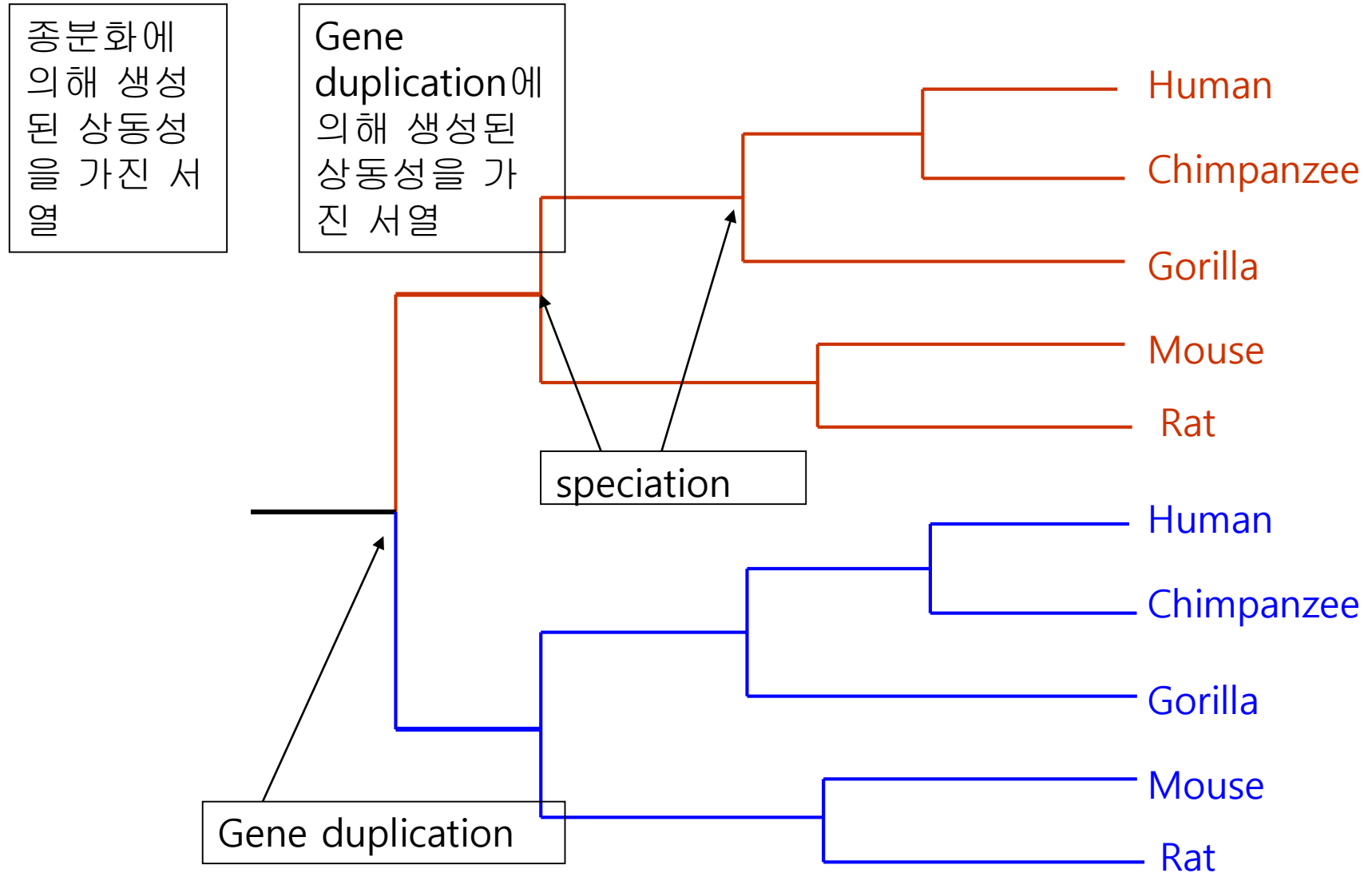


OTU (Operational Taxonomic Unit) , taxon (복수형명사: taxa), terminal node, tip

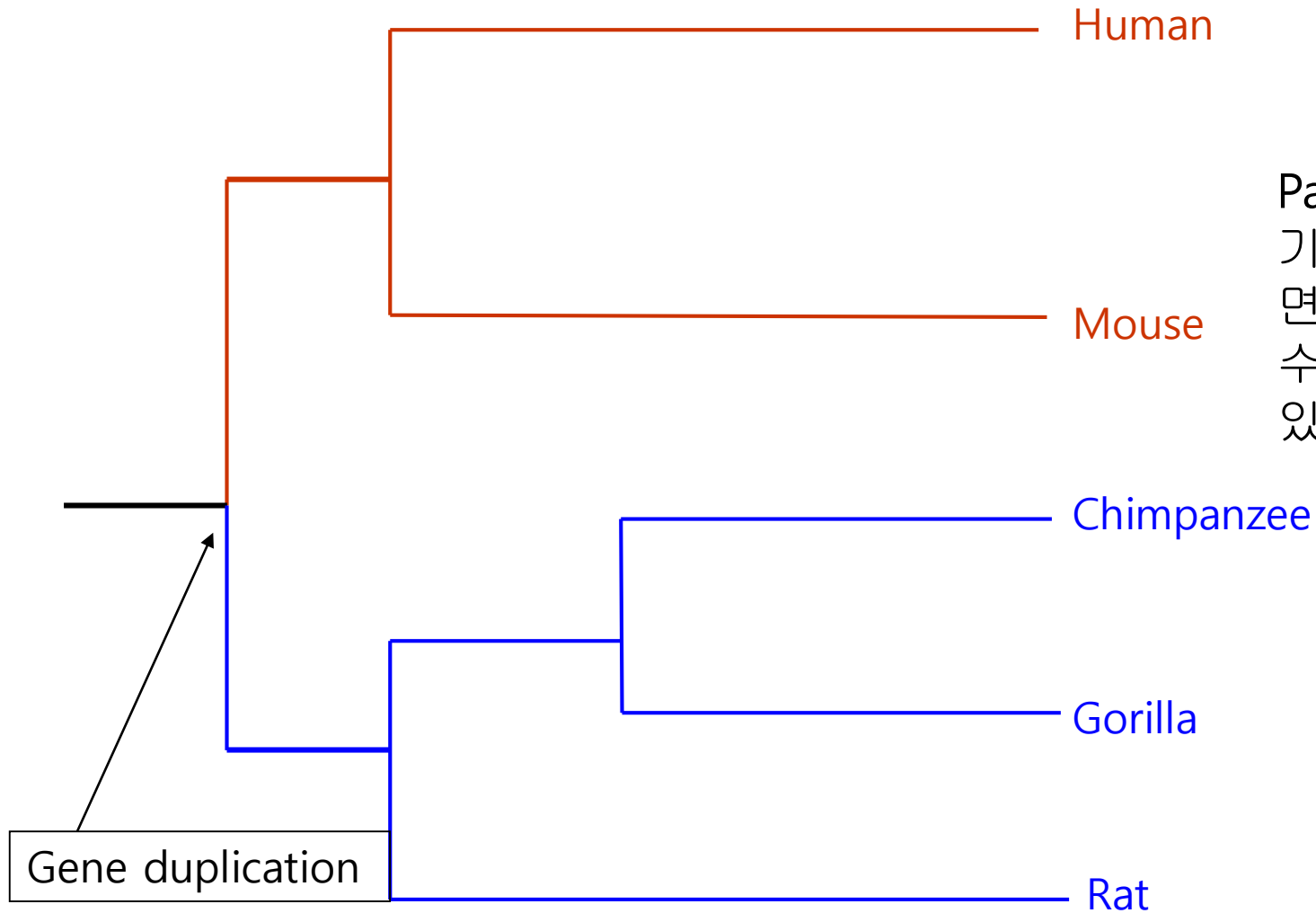
- Gene tree와 species tree는 일치하지 않을수 있으므로 주의를 요함



- Ortholog/paralog 에 주의 (種의 계통관계 추정을 위해서는 orthologous 서열을 사용해야함)



- Ortholog/paralog 에 주의 (種의 계통관계 추정을 위해서는 orthologous 서열을 사용해야함)



Paralogous 염기서열을 사용하면 잘못된 계통수가 얻어질수 있음

Maximum parsimony (MP) method

계통수상에서 일어나는 염기치환의 수를 단순히 세어 염기치환수가 최소가 되는 계통수를 구하는 방법. 염기치환의 중복을 고려하지 않는다.

사이트	1	2	3	4	5	6	7	8	9
서열1	A	A	T	T	C	G	C	C	A
서열2	A	A	T	T	C	T	C	C	T
서열3	G	A	C	G	C	T	C	G	G
서열4	A	A	T	G	C	G	C	C	T

1,3,4,6,9 : variable sites

4,6 : informative sites

Informative sites : 2종류 이상의 염기가 각각 2회이상 등장하는 사이트.

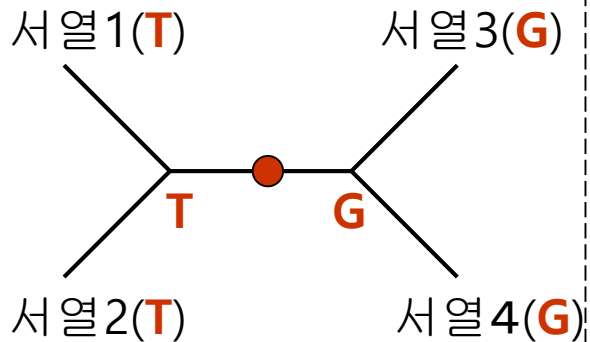
MP method에서는 informative sites 만 고려됨

4개의 염기서열에 대하여 가능한 Unrooted tree는 3가지. 3가지의 계통수에 대하여 parsimony score를 계산.

Tree 1

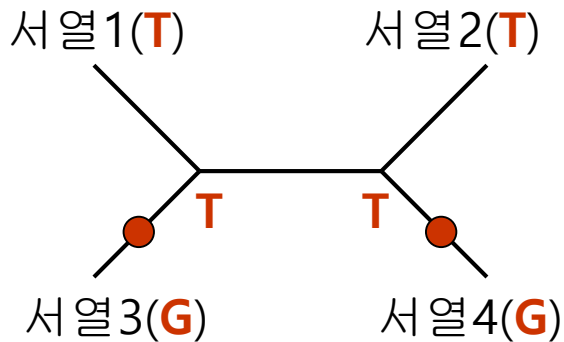
MP
tree

4th site



Score = 1

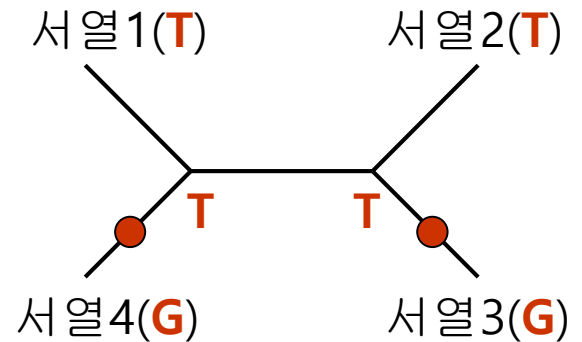
Tree 2



Score = 2

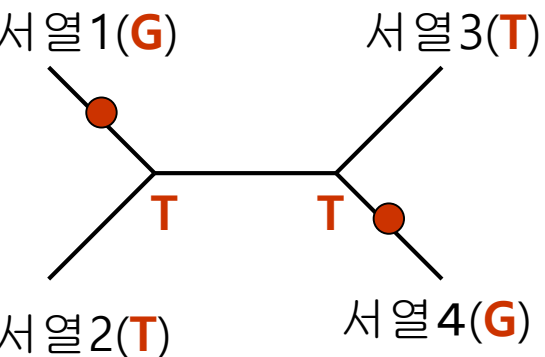
MP
tree

Tree 3



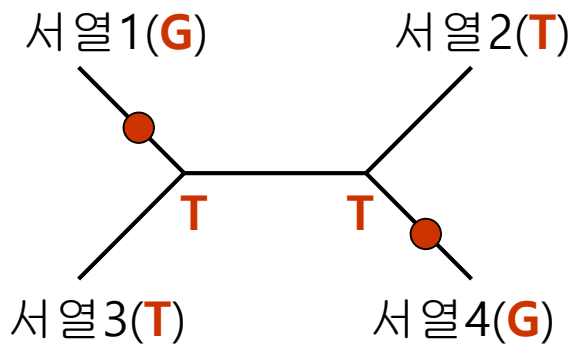
Score = 2

6th site



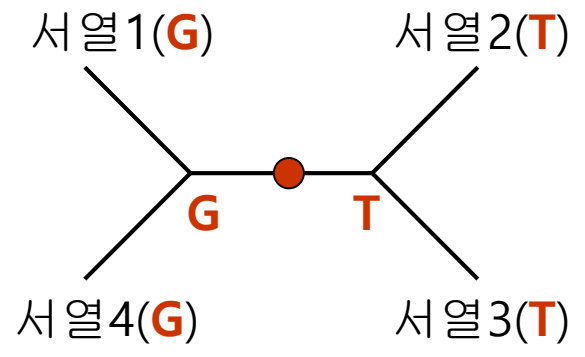
Score = 2

Score 합: 3



Score = 2

Score 합: 4



Score = 1

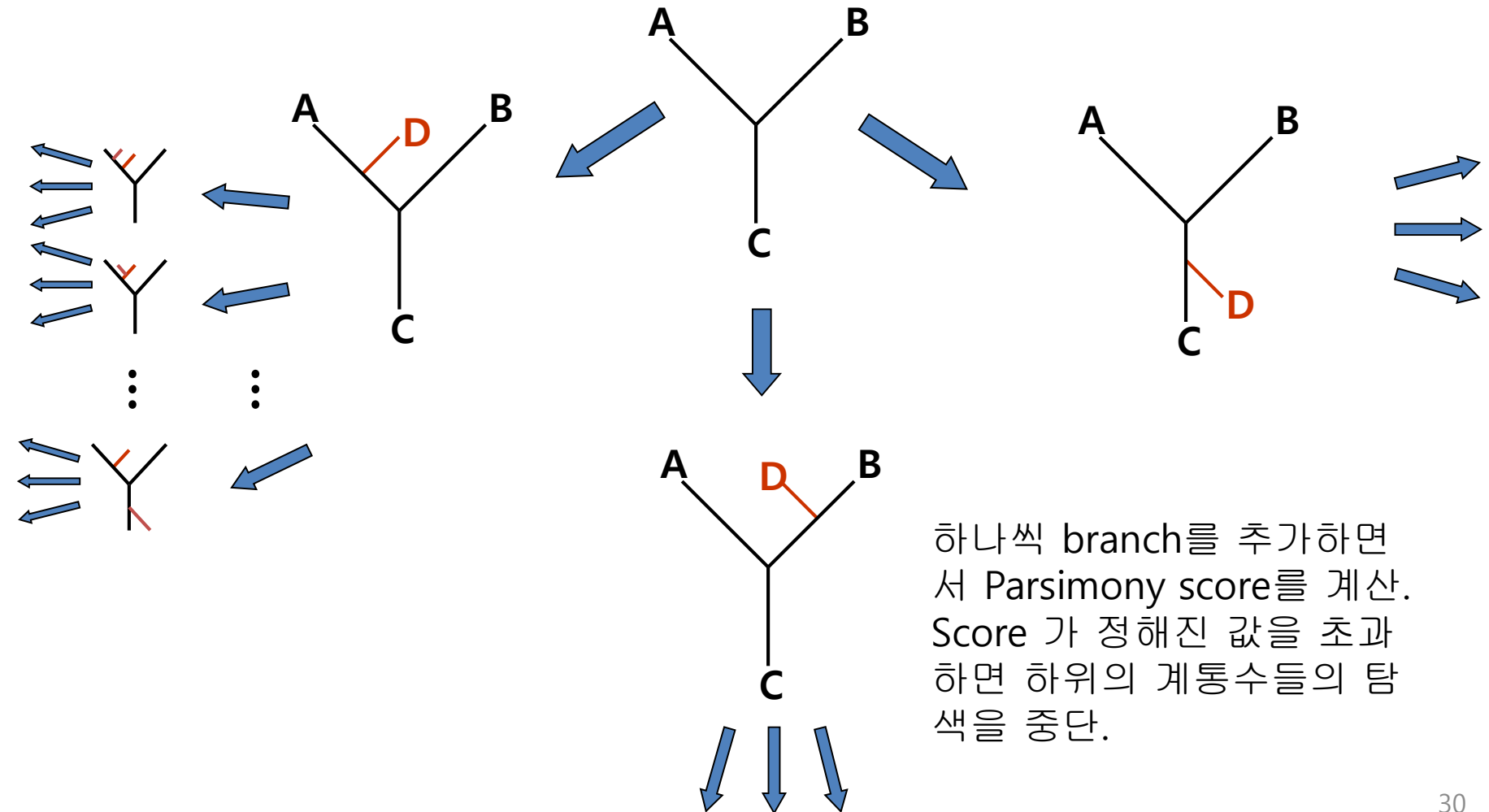
Score 합: 3

- 염기서열의 수가 증가함에 따라 가능한 계통수의 수는 급격히 증가
 → 모든 계통수에 대하여 parsimony score를 계산/비교하는 것은 불가능

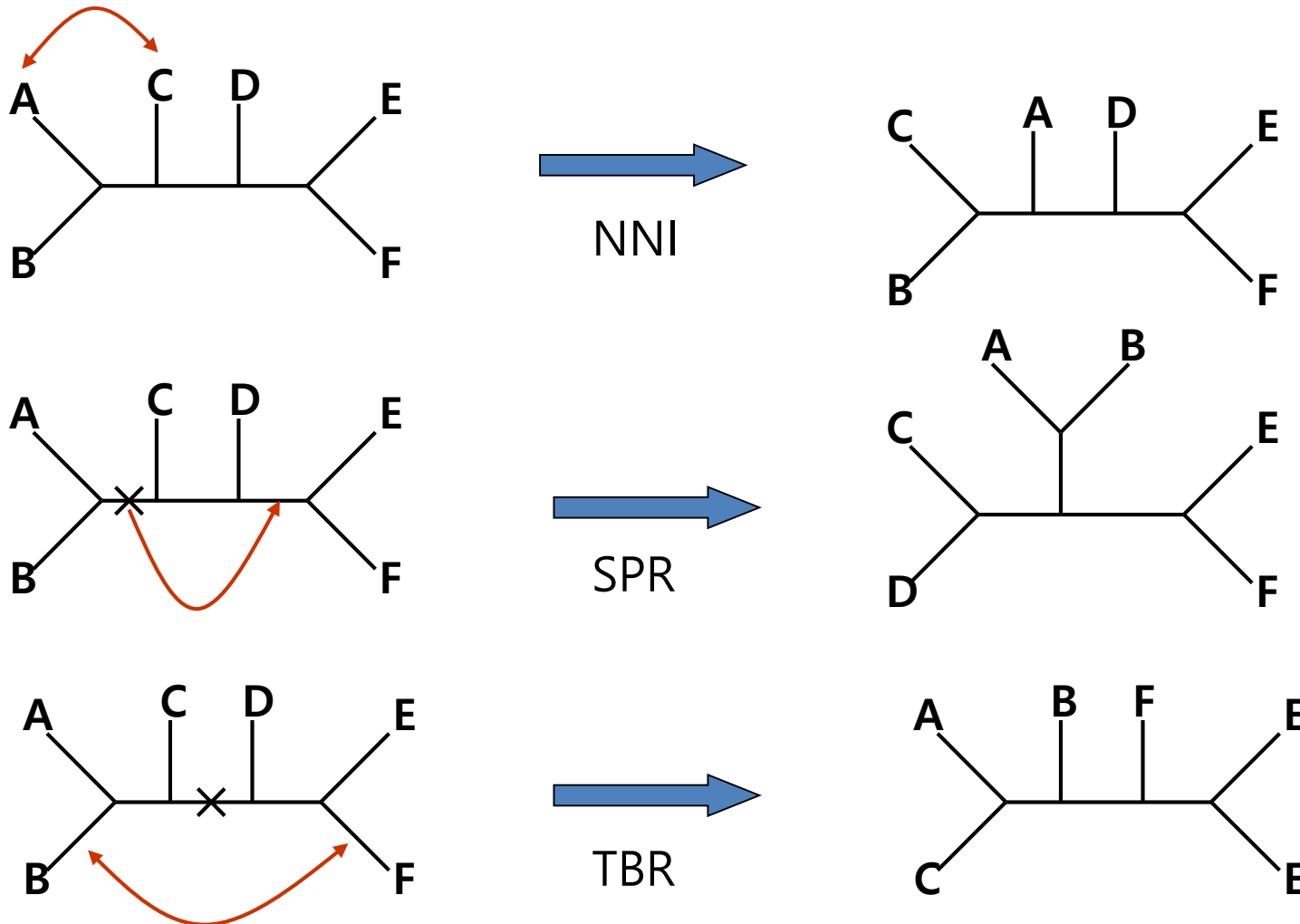
Num of OTU	Num of rooted tree	Num of unrooted tree
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025
11	654,729,075	34,459,425
12	13,749,310,575	654,729,075
13	316,234,143,225	13,749,310,575
14	7,905,853,580,625	316,234,143,225
15	213,458,046,676,875	7,905,853,580,625
...
n	$\frac{(2n-3)!}{2^{n-2}(n-2)!}$	$\frac{(2n-5)!}{2^{n-3}(n-3)!}$

- 계통수 탐색법: exhaustive search

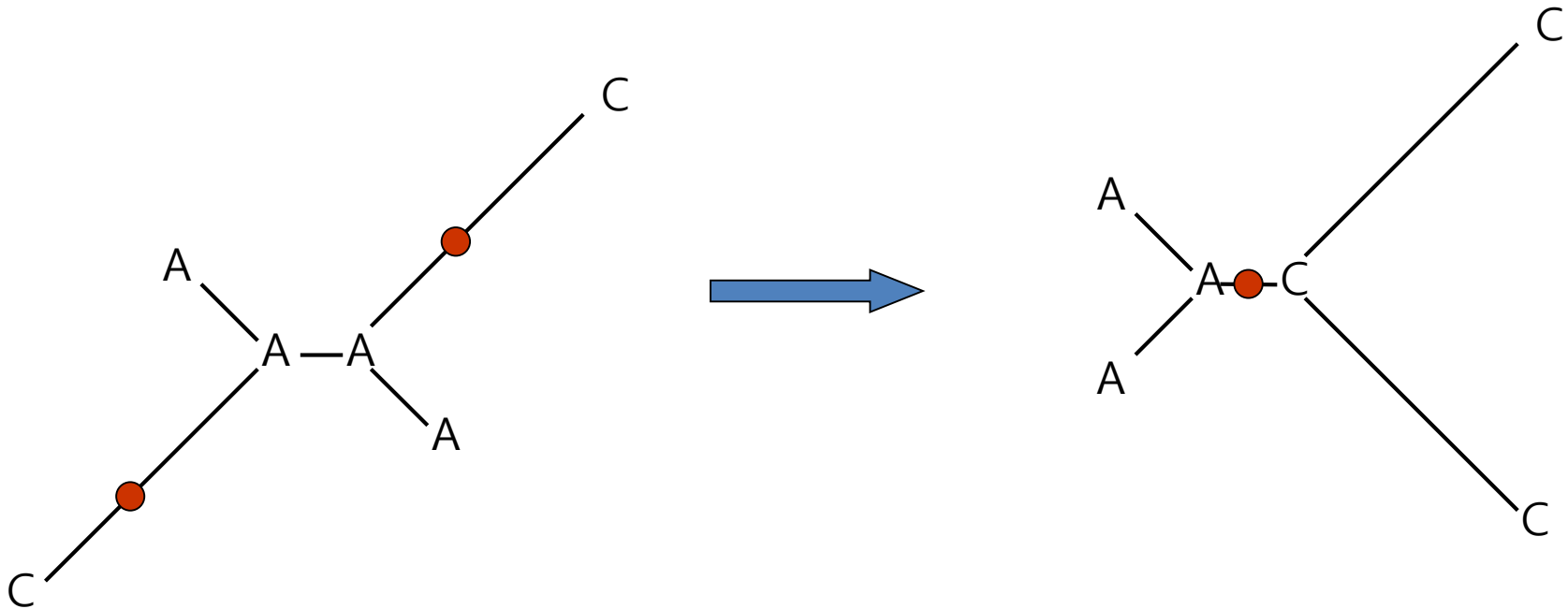
- branch-and-bound : 비현실적인 계통수를 만나면 탐색을 도중에 중단.
모든 계통수를 조사하지 않아도 전부 조사한것과 같은 결과 .



- 계통수 탐색법 : heuristic search (계통수의 일부분만 탐색)
 - Nearest neighbor interchange (NNI)
 - Subtree pruning and regrafting (SPR)
 - Tree bisection and reconnection (TBR)



- Maximum parsimony의 문제점
 - 다중치환이 고려되지 않음 (한번만 카운트됨)
 - Long branch attraction

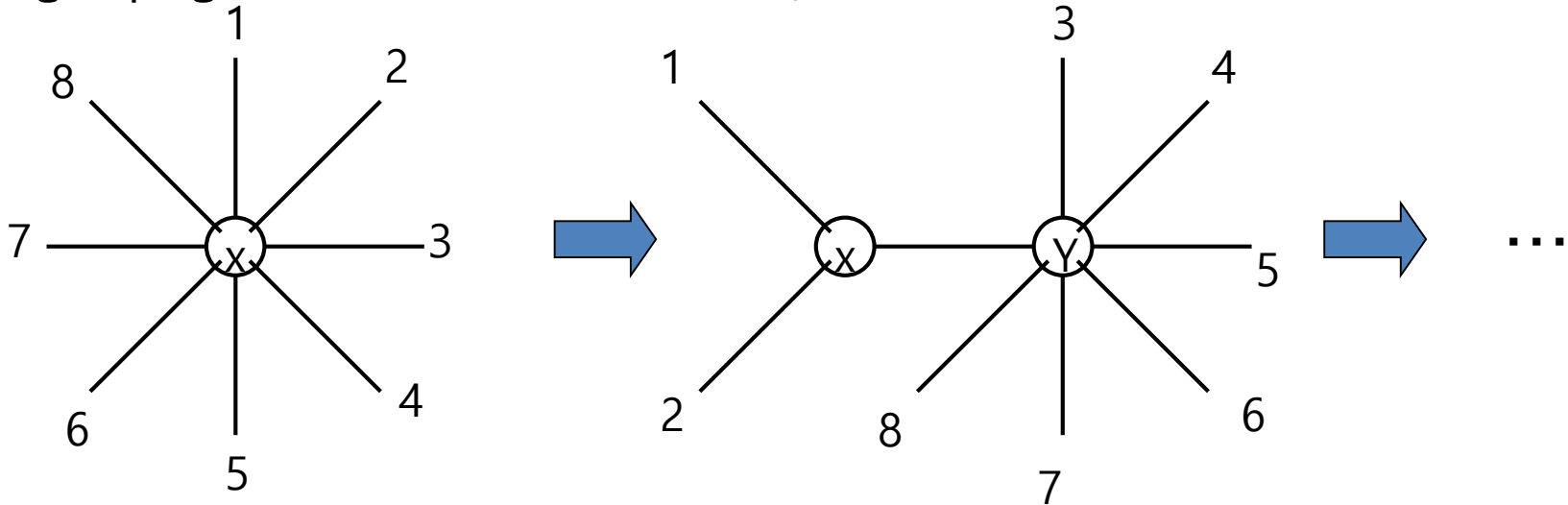


Homoplasy가 존재하는 데이터
(공동조상에서 유래한것이
아니지만 상동성을 가짐)

추정된 계통수

Neighbor-Joining (NJ) method

- Distance matrix method 중 가장 널리 쓰이는 방법. Star tree로부터 시작, branch length의 합계가 최소가 되도록, 단계적으로 grouping 한다 (Saitou & Nei 1987).

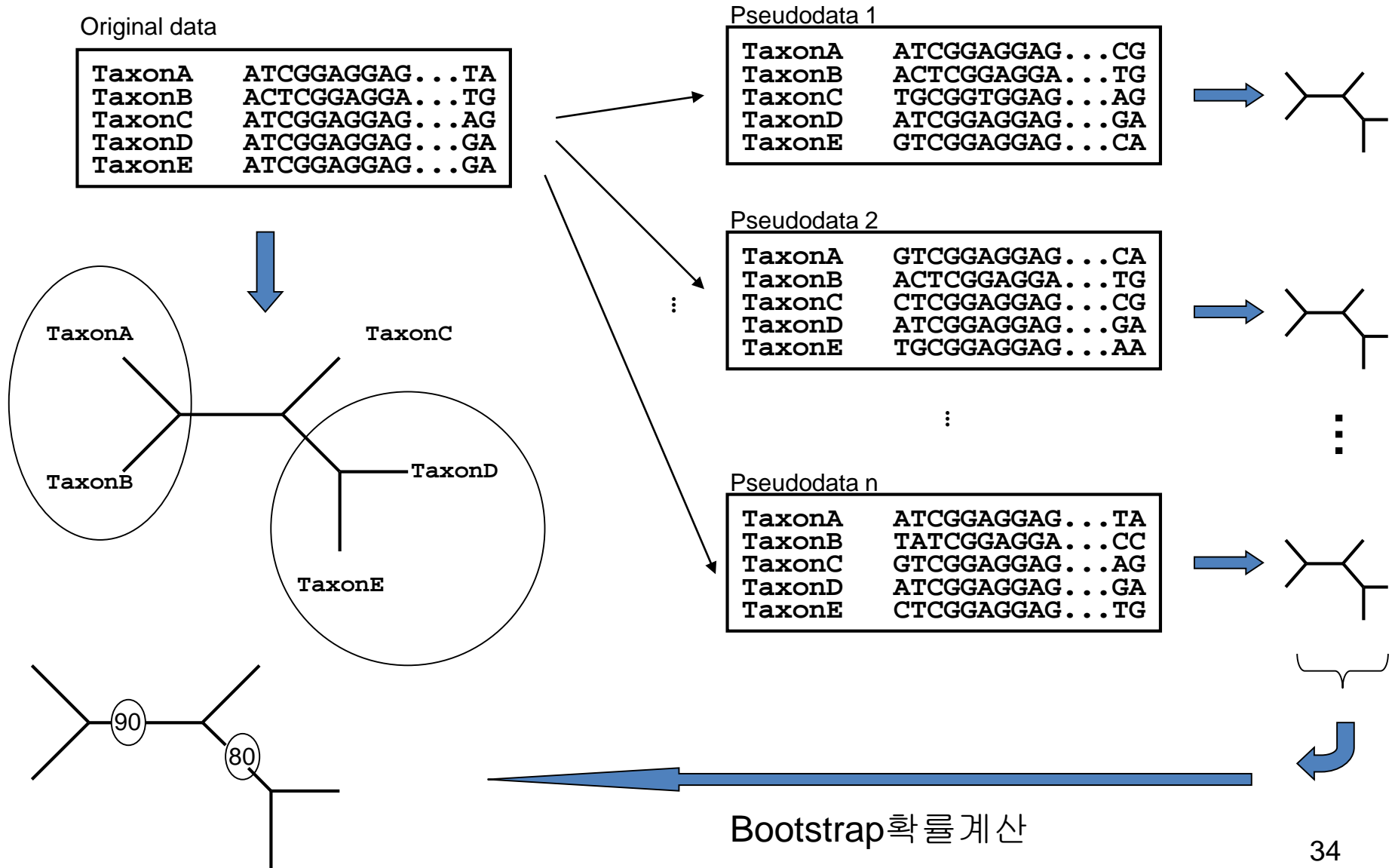


$$\begin{pmatrix} d_{12} & d_{13} & d_{14} & d_{15} & d_{16} & d_{17} & d_{18} \\ & d_{23} & d_{24} & d_{25} & d_{26} & d_{27} & d_{28} \\ & & d_{34} & d_{35} & d_{36} & d_{37} & d_{38} \\ & & & d_{45} & d_{46} & d_{47} & d_{48} \\ & & & & d_{56} & d_{57} & d_{58} \\ & & & & & d_{67} & d_{68} \\ & & & & & & d_{78} \end{pmatrix}$$

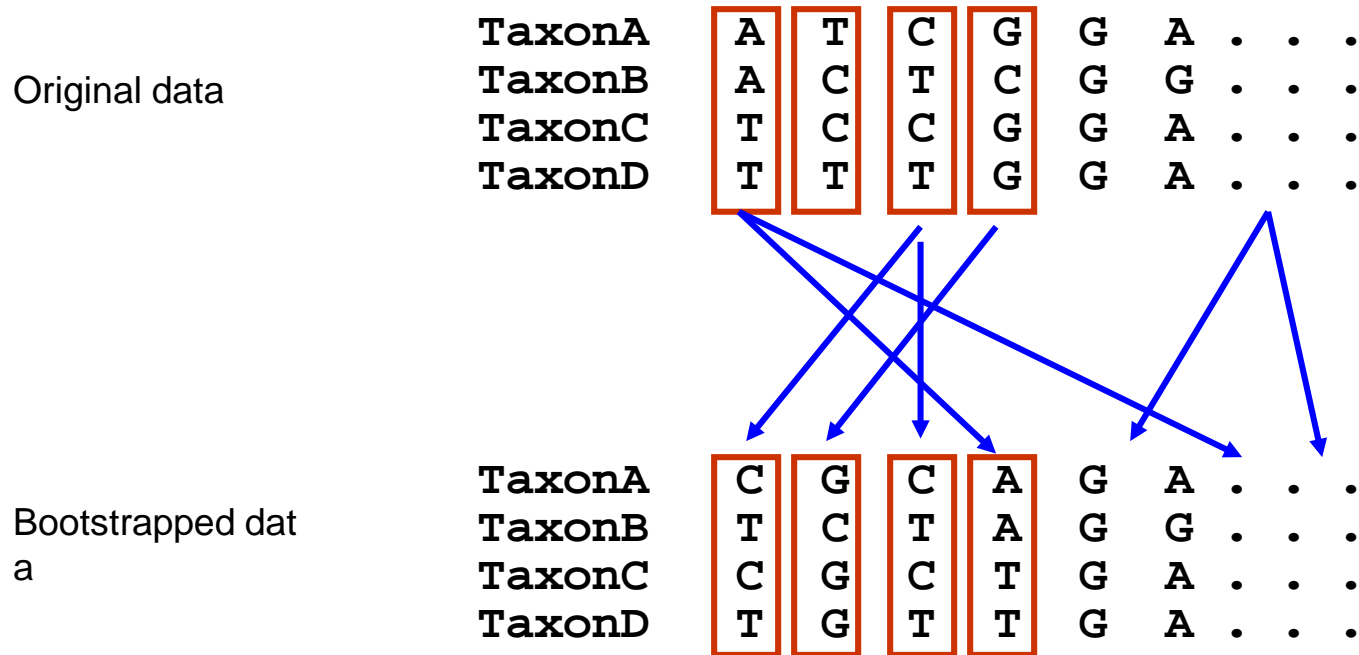
$$S_{12} = \frac{1}{2(N-2)} \sum_{k=3}^N (d_{1k} + d_{2k}) + \frac{1}{2} d_{12} + \frac{1}{N-2} \sum_{3 \leq i < j < N} d_{ij}$$

d_{ij} : 서열 i와 j 사이의 진화적인 거리. 사이트당 염기치환수를 나타냄

Bootstrap method (계통수의 신뢰도 추정)



- Generation of pseudo-data

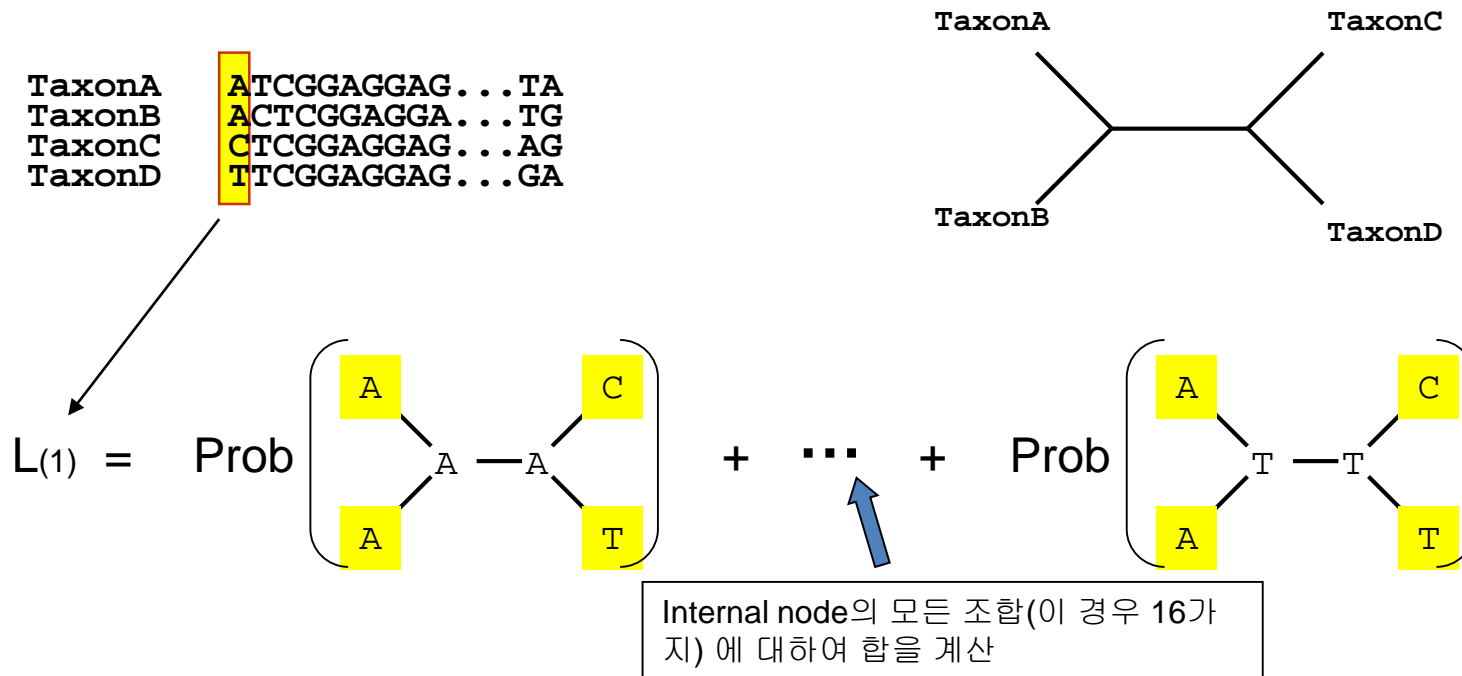


- *i.i.d.* assumption : *i*ndependently and *i*dentically *d*istributed sample
- Each alignment column is resampled

Maximum likelihood (ML) method

- 각 계통수별로 likelihood를 계산, likelihood가 최대가 되는 계통수 및 모델파라미터를 구하는 방법

(likelihood: 데이터가 모델에 어느정도 잘 맞는가 나타내는 수치. 「데이터가 생성될 확률」과 밀접한 관계가 있음.)

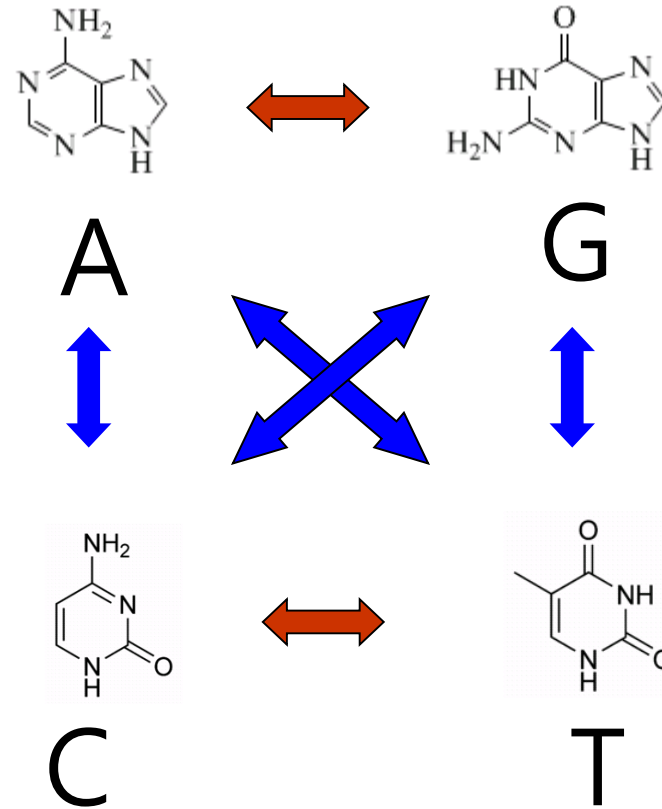
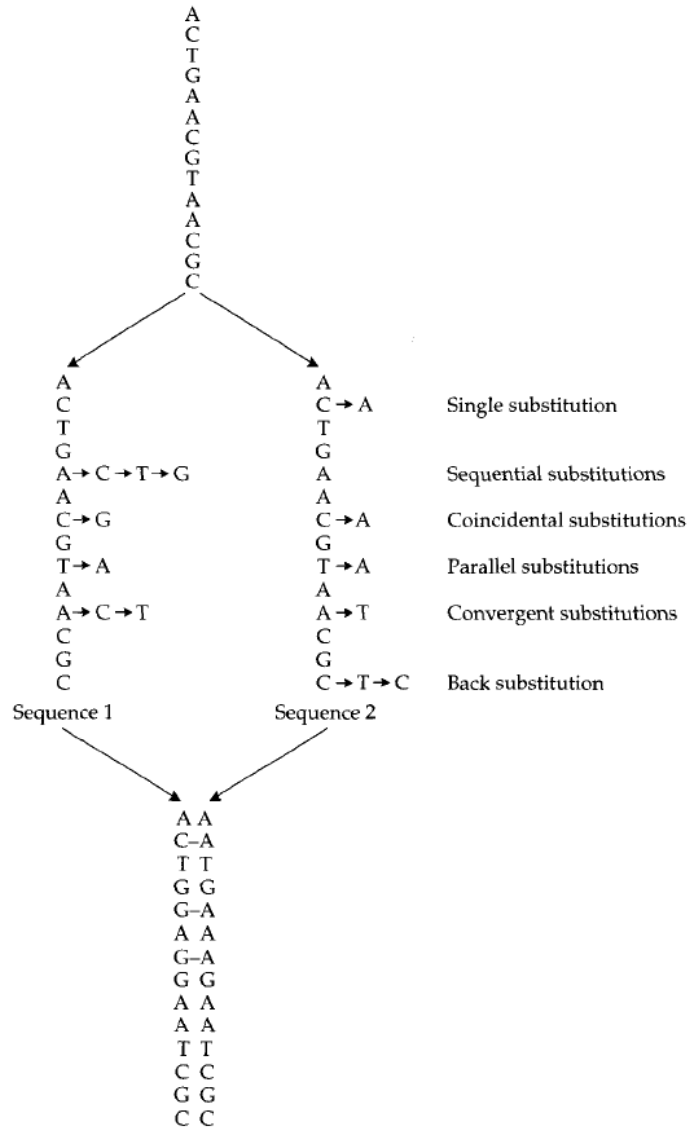


$$L = L(1) \times L(2) \times L(3) \times \dots \times L(n)$$

$$\log L = \log L(1) + \log L(2) + \log L(3) + \dots + \log L(n)$$

Log-likelihood가 최대가 되는 계통수, branch length, 모델 파라미터등을 추정한다.

DNA서열은 어떻게 진화하는가?



↔ Transition: 염기치환이 빠름 (염기의 입체구조가 비슷)

↔ Transversion: 염기치환이 느림

DNA염기치환 모델

- 염기치환의 상대적인 속도를 파라미터로 표현
- 복잡한 분자진화 현상을 단순화 하여 몇 개의 파라미터로 설명
예) Jukes-Cantor model (JC), Kimura's two parameter model (K2, K80), Hasegawa-Kishino-Yano model (HKY85), etc.

JC

To

From

	A	C	T	G
A	-	α	α	α
C	α	-	α	α
T	α	α	-	α
G	α	α	α	-

치환의 상대적 속도가 동일(α)

K2

	A	C	T	G
A	-	α	α	β
C	α	-	β	α
T	α	β	-	α
G	β	α	α	-

Transition의 속도(β)와 transversion의 속도(α)가 다르다. 일반적으로 $\beta > \alpha$

HKY85

	A	C	T	G
A	-	π_C	π_T	$\kappa \pi_G$
C	π_A	-	$\kappa \pi_T$	π_G
T	π_A	$\kappa \pi_C$	-	π_G
G	$\kappa \pi_A$	π_C	π_T	-

염기치환속도가 염기의 빈도(π)에 비례. Transition이 transversion보다 κ 배 빠름.

Table 3.2 Models of nucleotide substitution.

	A	T	C	G		A	T	C	G
(A) Jukes-Cantor model					(E) HKY model				
A	-	α	α	α	-	βg_T	βg_C	αg_G	
T	α	-	α	α	βg_A	-	αg_C	βg_G	
C	α	α	-	α	βg_A	αg_T	-	βg_G	
G	α	α	α	-	αg_A	βg_T	βg_C	-	
(B) Kimura model					(F) Tamura-Nei model				
A	-	β	β	α	-	βg_T	βg_C	$\alpha_1 g_G$	
T	β	-	α	β	βg_A	-	$\alpha_2 g_C$	βg_G	
C	β	α	-	β	βg_A	$\alpha_2 g_T$	-	βg_G	
G	α	β	β	-	$\alpha_1 g_A$	βg_T	βg_C	-	
(C) Equal-input model					(G) General reversible model				
A	-	αg_T	αg_C	αg_G	-	ag_T	bg_C	cg_G	
T	αg_A	-	αg_C	αg_G	ag_A	-	dg_C	eg_G	
C	αg_A	αg_T	-	αg_G	bg_A	dg_T	-	fg_G	
G	αg_A	αg_T	αg_C	-	cg_A	eg_T	fg_C	-	
(D) Tamura model					(H) Unrestricted model				
A	-	$\beta\theta_2$	$\beta\theta_1$	$\alpha\theta_1$	-	a_{12}	a_{13}	a_{14}	
T	$\beta\theta_2$	-	$\alpha\theta_1$	$\beta\theta_1$	a_{21}	-	a_{23}	a_{24}	
C	$\beta\theta_2$	$\alpha\theta_2$	-	$\beta\theta_1$	a_{31}	a_{32}	-	a_{34}	
G	$\alpha\theta_2$	$\beta\theta_2$	$\beta\theta_1$	-	a_{41}	a_{42}	a_{43}	-	

GTR 모델이라고 불림

Transition probabilities

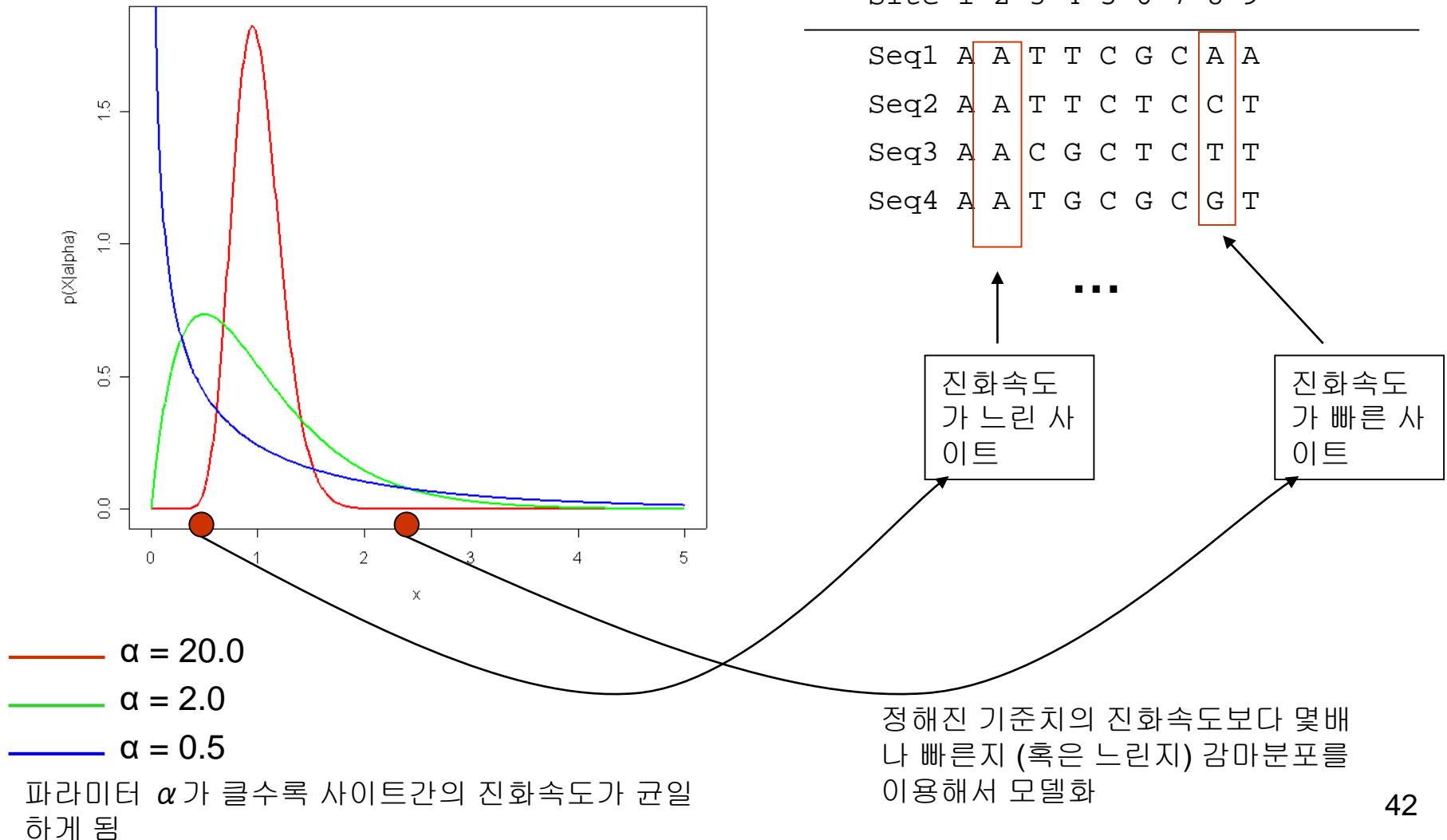
$$\mathbf{R} := \begin{array}{c} \text{A} \\ \text{C} \\ \text{T} \\ \text{G} \end{array} \begin{array}{c} \text{A} \quad \text{C} \quad \text{T} \quad \text{G} \\ \left(\begin{array}{cccc} - & \alpha & \alpha & \beta \\ \alpha & - & \beta & \alpha \\ \alpha & \beta & - & \alpha \\ \beta & \alpha & \alpha & - \end{array} \right) \end{array}$$

Transition probabilities after time t are

$$\mathbf{P}(t) := e^{t\mathbf{R}} = \begin{pmatrix} P_{AA}(t) & P_{AC}(t) & P_{AT}(t) & P_{AG}(t) \\ P_{CA}(t) & P_{CC}(t) & P_{CT}(t) & P_{CG}(t) \\ P_{TA}(t) & P_{TC}(t) & P_{TT}(t) & P_{TG}(t) \\ P_{GA}(t) & P_{GC}(t) & P_{GT}(t) & P_{GG}(t) \end{pmatrix}$$

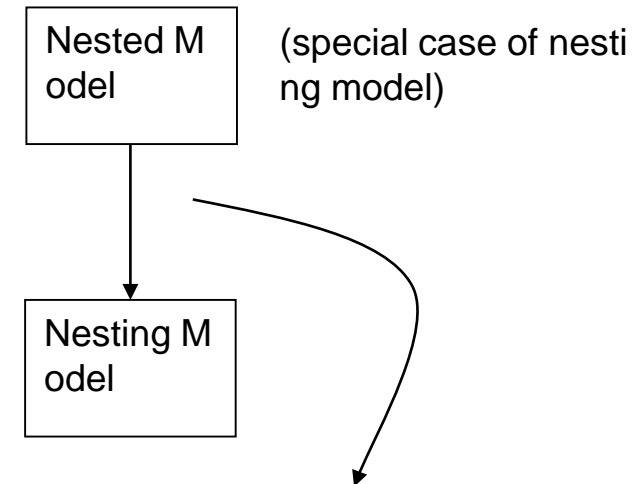
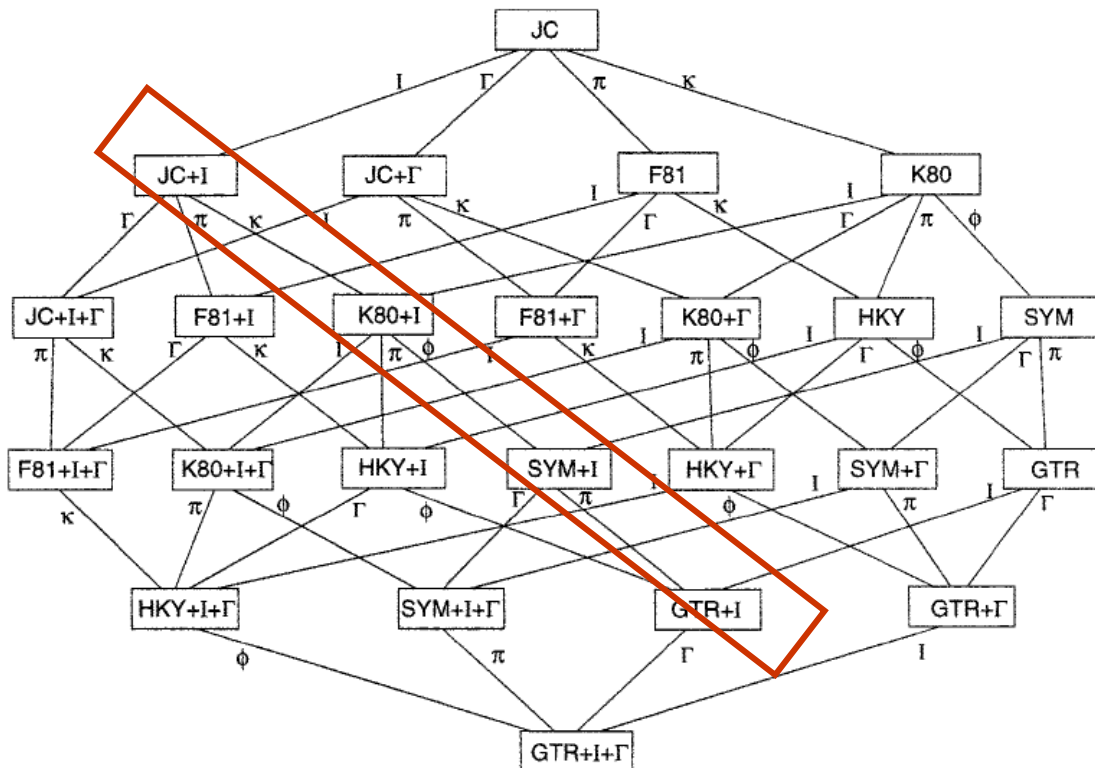
- Rate heterogeneity among site (Yang 1993)

감마분포: 염기서열 데이터의 사이트간의 진화속도의 모델



Likelihood 스코어를 이용한 모델 비교 (likelihood ratio test; LRT)

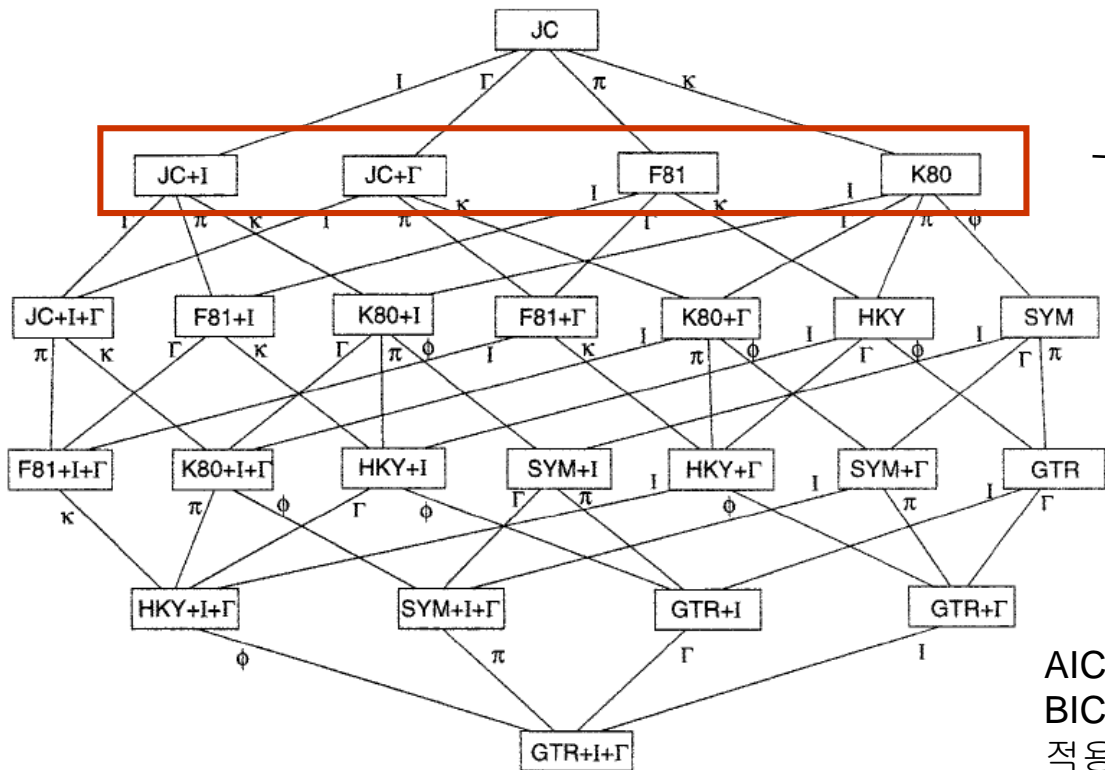
Hierarchy of DNA models



두 모델의 비교에 있어서 복잡한 모델의 파라미터를 특정한 값으로 고정시키면 단순한 모델이 될 경우 Likelihood ratio test (LRT) 를 적용할 수 있다

Likelihood 스코어를 이용한 모델 비교 (AIC, BIC, 등)

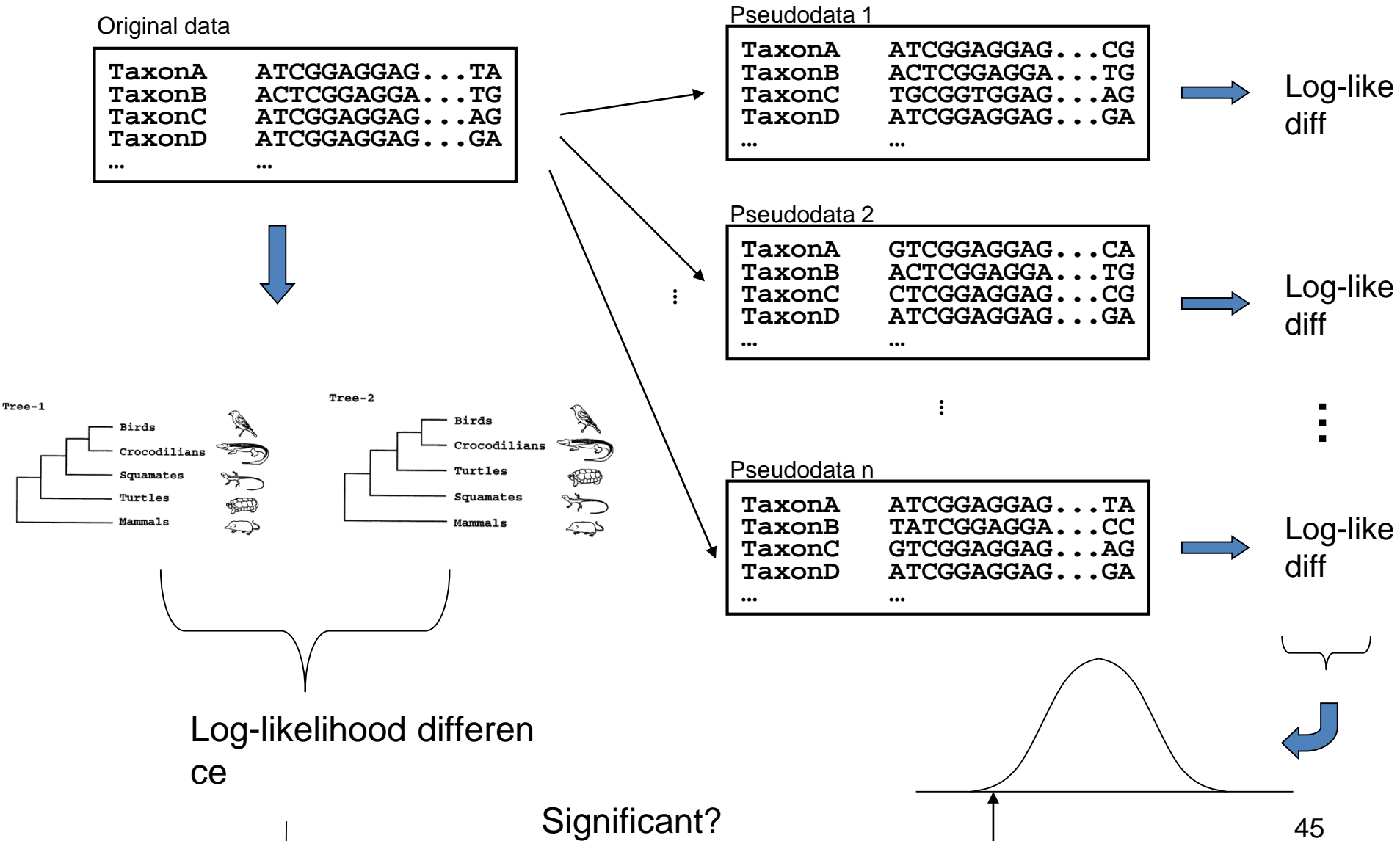
Hierarchy of DNA models



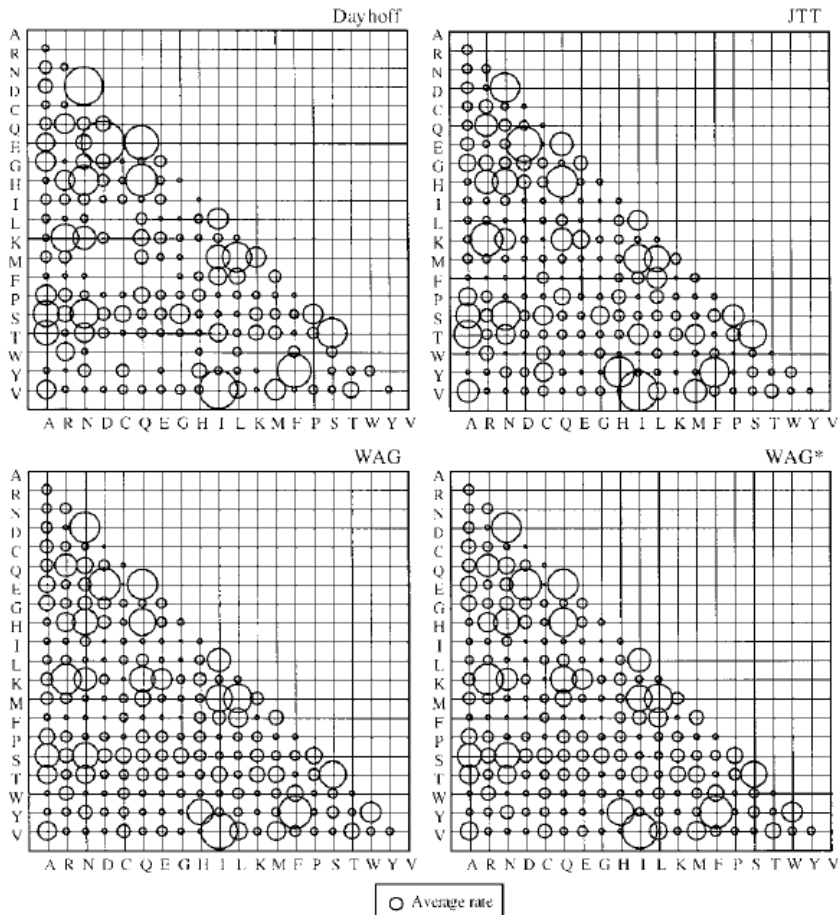
Likelihood ratio test (LRT) 를 적용
할수 없는 일반적인 경우

AIC (Akaike Information Criterion)
BIC (Bayesian Information Criterion)
적용 가능함

Bootstrap method (계통수의 신뢰도 추정)



아미노산 치환 모델



	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	98669	11	40	56	12	34	97	211	8	24	35	21	11	7	126	282	216	0	7	133
Arg	23	99137	13	1	8	93	1	8	80	23	13	370	13	5	52	106	16	21	2	16
Asn	87	14	98198	423	1	40	75	124	180	29	29	256	1	6	21	343	134	2	29	10
Asp	104	1	360	98592	0	51	562	110	29	9	1	57	0	0	7	66	39	0	1	11
Cys	32	10	1	0	99725	0	0	10	10	16	1	0	1	1	10	113	10	1	29	32
Gln	78	100	42	64	0	98754	353	25	203	7	63	123	17	0	78	39	31	0	1	23
Glu	169	1	60	528	0	268	98656	71	15	22	10	66	4	0	25	55	20	0	6	24
Gly	207	4	56	58	4	11	40	99351	4	0	6	21	2	6	17	161	18	0	0	35
His	20	96	211	40	9	227	21	9	99132	3	37	21	1	19	47	24	13	3	37	28
Ile	57	26	31	11	15	7	30	1	3	98727	217	37	49	77	6	17	111	0	11	568
Leu	36	6	14	0	0	28	6	6	15	95	99465	15	77	62	16	12	19	5	9	113
Lys	23	189	128	34	0	58	41	24	9	17	15	99251	36	0	17	67	79	0	4	7
Met	61	36	2	2	1	42	15	15	1	121	439	191	98764	36	8	42	59	0	1	163
Phe	16	6	6	0	1	0	0	14	16	71	133	0	14	99457	6	32	8	8	207	8
Pro	215	42	17	6	6	58	25	31	31	4	27	27	2	4	99260	168	45	0	0	31
Ser	350	62	196	44	53	21	39	204	12	9	15	76	9	18	122	98415	316	8	10	20
Thr	323	11	93	31	5	20	17	27	8	71	28	110	15	5	40	38415	98699	0	13	102
Trp	1	86	10	0	3	1	1	3	10	1	41	1	1	32	1	54	1	99733	19	2
Tyr	21	3	38	1	32	1	11	0	42	14	24	11	0	275	1	23	24	6	99453	18
Val	178	10	6	8	16	13	18	47	15	323	148	8	38	5	24	21	91	0	8	99020

π 0.087 0.041 0.040 0.047 0.033 0.038 0.050 0.089 0.034 0.037 0.085 0.080 0.015 0.040 0.051 0.070 0.058 0.010 0.030 0.065

Table 1: Transition probability($\times 10^5$) of the Dayhoff model during a time interval of one substitution per 100 amino acids (1 PAM).

아미노산 치환의 상대적인
속도를 데이터로부터 경험적
으로 얻음

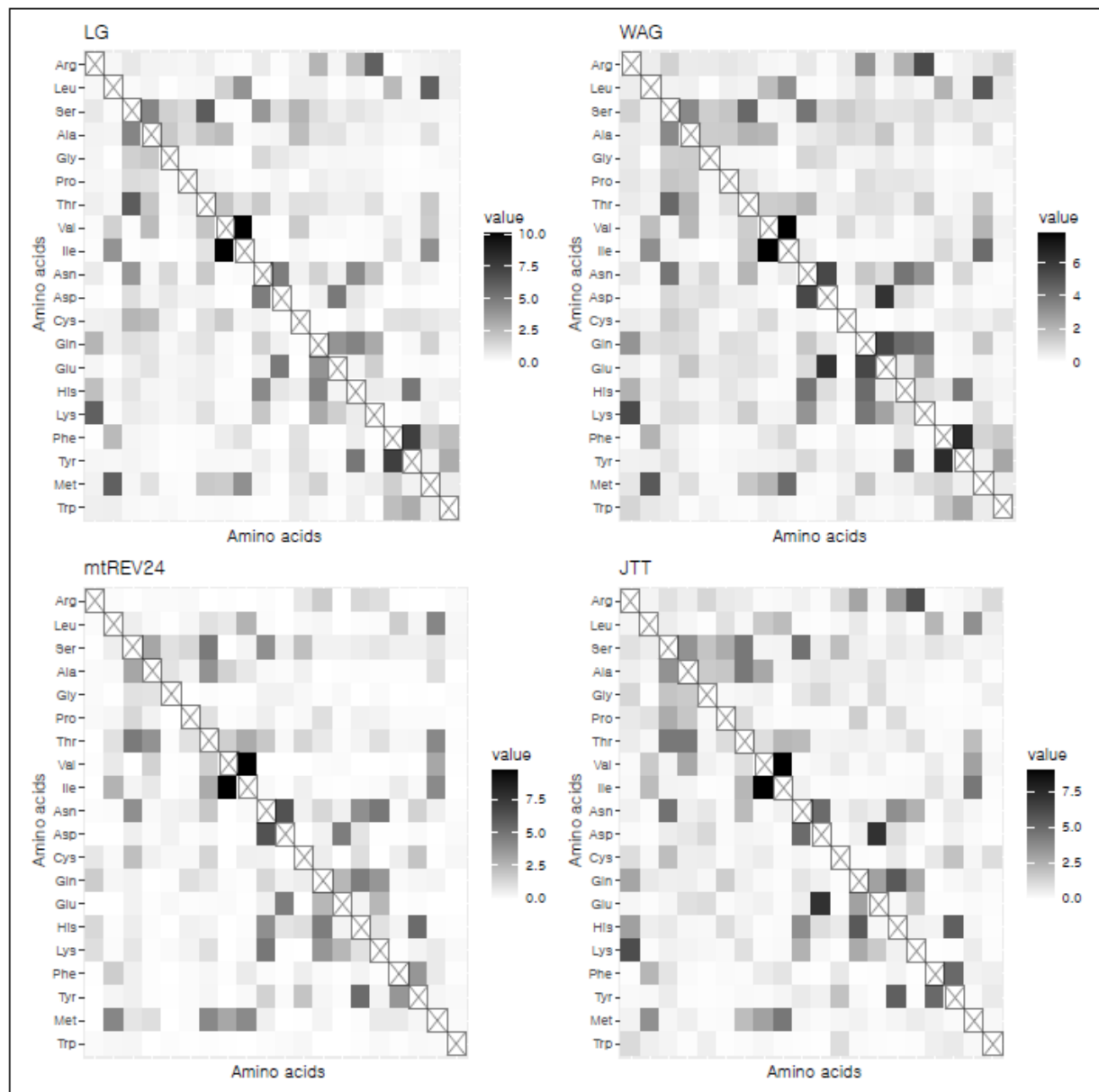
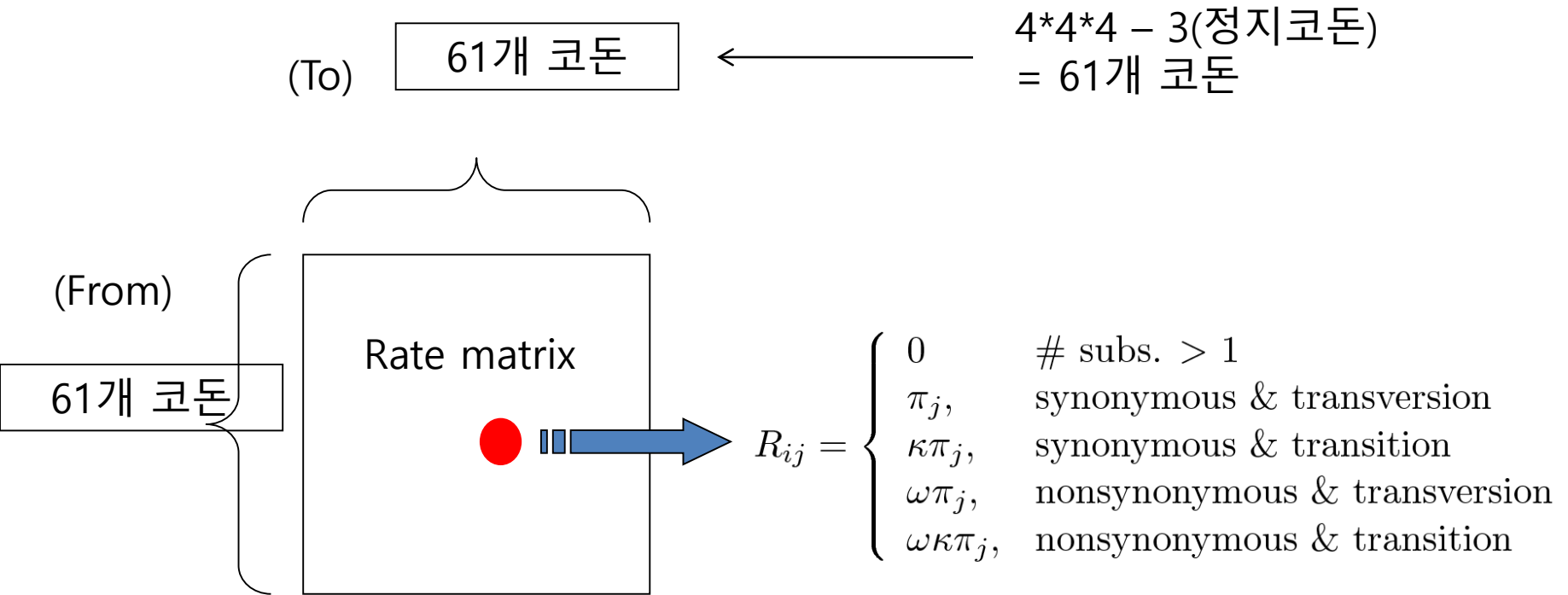


그림 2. 아미노산 모형의 $s_{a_i a_j}$ 값들. $s_{a_i a_j}$ 값이 크면 해당 아미노산 쌍은 치환이 빈번하게 일어남을 의미한다. 임의로 LG(Le and Gascuel 2008), WAG(Whelan and Goldman 2001), mtREV24(Adachi et al. 1996), JTT(Jones et al. 1992) 네 종류 아미노산 치환 모형을 선택했다. 가로축 아미노산 순서 (좌에서 우)는 세로축 아미노산 순서(위에서 아래)와 같다.

코돈치환 모델



(e.g., Goldman & Yang 1994)

	CGT	CGC	CGA	CGG	AGA	AGG
CGT	$-$	$\kappa\pi_{CGC}$	π_{CGA}	π_{CGG}	0	0
CGC	$\kappa\pi_{CGT}$	$-$	π_{CGA}	π_{CGG}	0	0
CGA	π_{CGT}	π_{CGC}	$-$	$\kappa\pi_{CGG}$	π_{AGA}	0
CGG	π_{CGT}	π_{CGC}	$\kappa\pi_{CGA}$	$-$	0	π_{AGG}
AGA	0	0	π_{CGA}	0	$-$	$\kappa\pi_{AGG}$
AGG	0	0	0	π_{CGG}	$\kappa\pi_{AGA}$	$-$

(아르기닌을 코딩하는 코돈끼리의 치환률 예시)

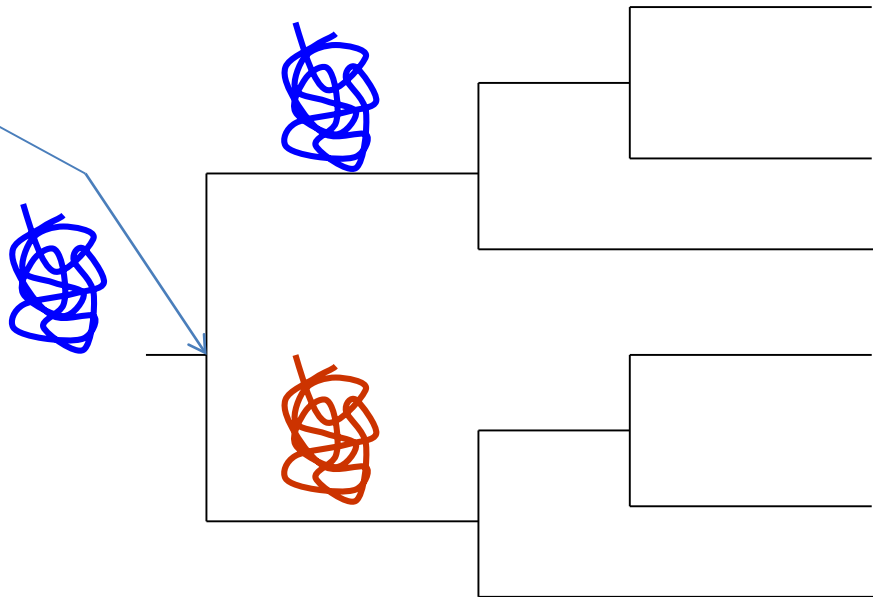
코돈치환 모델을 이용한 단백질의 분자진화 연구

Gene duplication
이 일어나면 한쪽
의 기능이 자유롭게
변할수 있다

단백질 3
차구조

기능이 달라진 단백
질은 어떤 과정에 의
해서?

시간



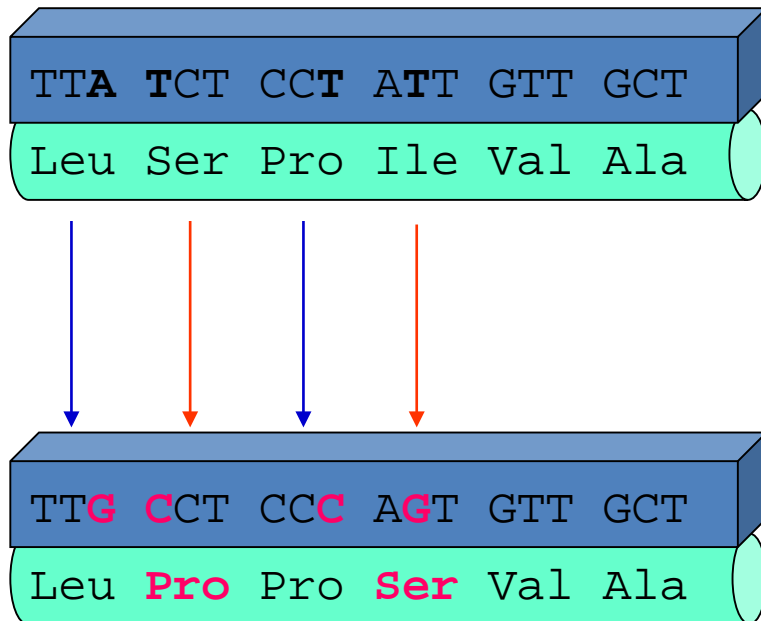
- 단백질 코딩 영역의 DNA 염기치환의 분류

- Synonymous 염기치환 → 아미노산이 변하지 않음

ex. TT**A** (Leu) → TT**G** (Leu)

- Nonsynonymous 염기치환 → 아미노산이 변함

ex. **T**CT(Ser) → **C**CT(Pro)



→
Synonymous substitution

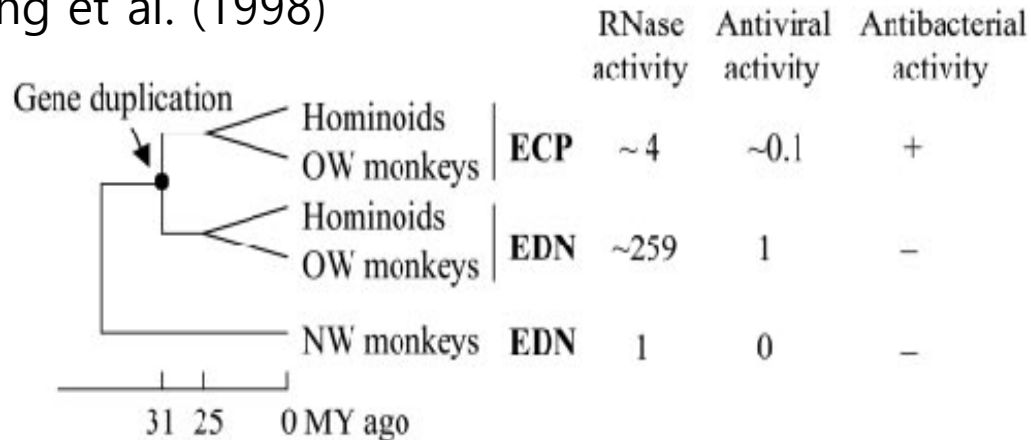
→
Nonsynonymous substitution

- Synonymous, Nonsynonymous 치환을 구분함으로써 얻을 수 있는 정보

- Synonymous 치환
 - 자연선택의 영향을 받지 않음
- Nonsynonymous 치환
 - 자연선택의 영향을 받음
 - 단백질 기능향상 → 치환속도 증가
 - 단백질 기능저하 → 치환속도 감소
- Nonsynonymous/synonymous 치환 속도의 비율 ($= \omega$, dn/ds)
 - > 1 : **positive (diversifying) selection**
 - = 1 : **neutral evolution**
 - < 1 : **purifying selection**

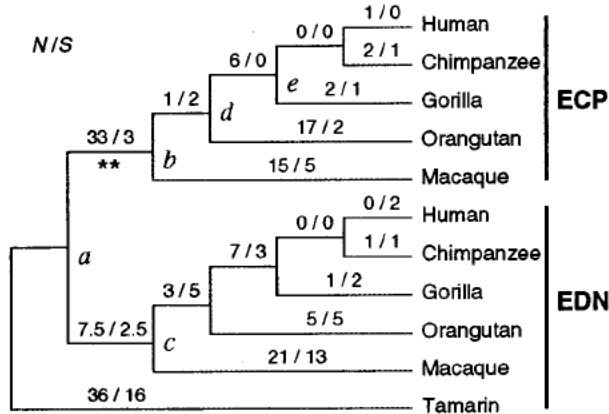
- Synonymous/Nonsynonymous 염기치환을 이용한 연구의 예

Zhang et al. (1998)

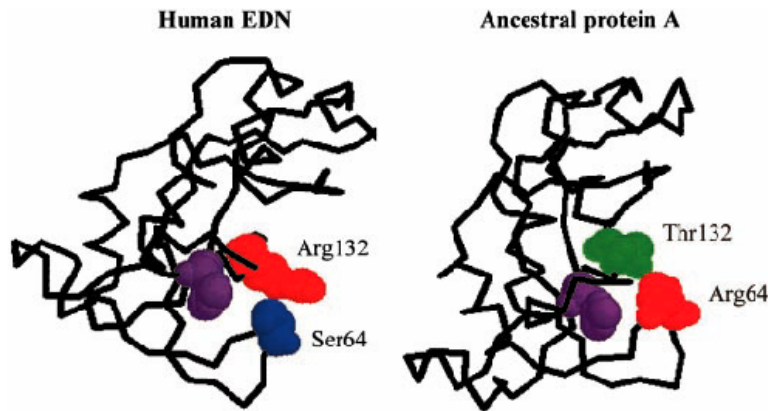


- NW (New World) monkeys : 중남미에 분포
- OW(Old World) monkeys : 아시아,아프리카에 분포
- EDN: eosinophil-derived neurotoxin
- ECP : eosinophil cationic protein
- EDN은 본래 antibacterial activity를 가지고 있지 않음. ECP는 가짐
- Hominoids/OW monkey는 EDN의 활성이 크다

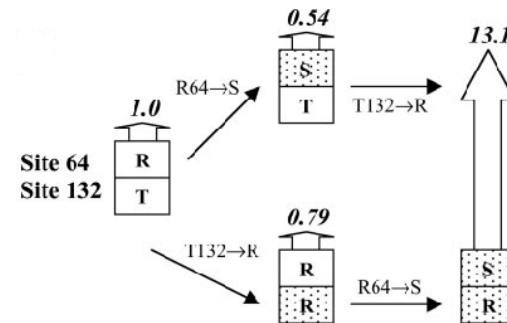
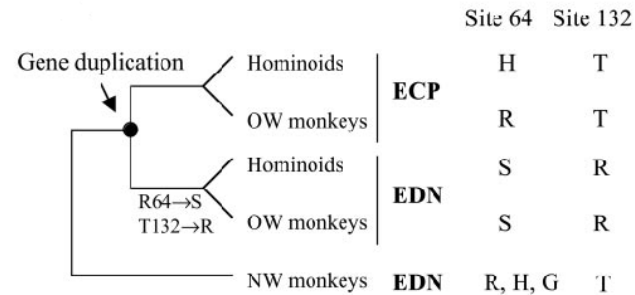
Zhang & Rosenberg (2002)



Gene duplication 직후에 Nonsynonymous 염기치환 (N)이 synonymous 염기치환(S)에 비하여 매우 왕성하게 일어났음을 알수 있다 -> 이 nonsynonymous 치환이 단백질의 기능과 밀접하게 관련이 있을것이라 추측됨



조상형의 단백질을 추정 → EDN의 기능에 있어서 64번째, 132번째 아미노산 치환이 중요



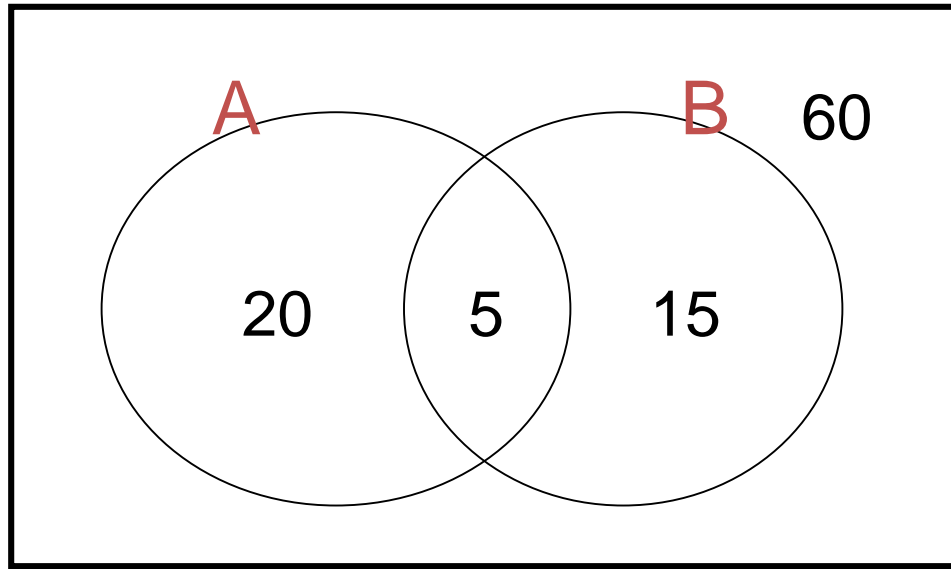
EDN의 기능향상과 관련하여 단백질을 실제로 합성하여 증명

모형	장점	단점
(1) DNA 치환 모형	계산속도가 빠름	비현실적인 (암묵적인) 가정 <ul style="list-style-type: none"> ● 정지코돈의 존재 가정 ● 정지코돈의 치환 가정
(2) 아미노산 치환 모형	경험적으로 얻은 치환 정보 반영 동의치환 포화에 영향을 받지 않음	동의치환 정보 손실
(3) 코돈 치환 모형	자연선택 검출 가능	계산속도 느림

표 1. 세 그룹 모형의 장단점

베이지안 분자계통수 추정

조건부 확률



$$P(A) = (20 + 5) / 100 = 1/4$$

$$P(B) = (15 + 5) / 100 = 1/5$$

$$P(A \cap B) = 5 / 100 = 1/20$$

$$P(B | A) = 5 / (20 + 5) = 1/5$$

$$P(A | B) = 5 / (15 + 5) = 1/4$$

$$P(A \cap B) = P(A | B)P(B)$$

$$P(A \cap B) = P(B | A)P(A)$$

$$\begin{cases} P(B | A) = P(B) = 1/5 \\ P(A | B) = P(A) = 1/4 \end{cases}$$

A와 B는 독립 $\Rightarrow P(A \cap B) = P(A)P(B)$

Bayes 정리

D: data θ : parameter



Tomas Bayes (1703-1761)

$$P(\theta | D) = \frac{P(\theta, D)}{P(D)}$$

$$= \frac{P(D | \theta)P(\theta)}{P(D)}$$

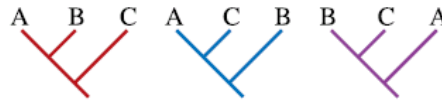
$$= \frac{P(D | \theta)P(\theta)}{\sum_{\theta} P(D | \theta)P(\theta)} \quad \text{or} \quad \frac{P(D | \theta)P(\theta)}{\int_{\theta} P(D | \theta)P(\theta)}$$

데이터가 주어진 상태에서 파라미터의 확률 밀도 함수

$P(D|\theta)$: Non-Bayesian(Frequentist, classical) framework에서는 θ : 고정, D: 랜덤

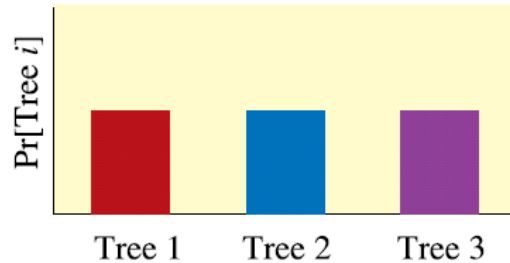
$P(\theta|D)$: Bayesian framework에서는 D: 고정, θ : 랜덤 → 미지의 파라미터에 대해서 확률분포를 생각하는것이 가능

분자 계통수의 베이지안 추정법



①

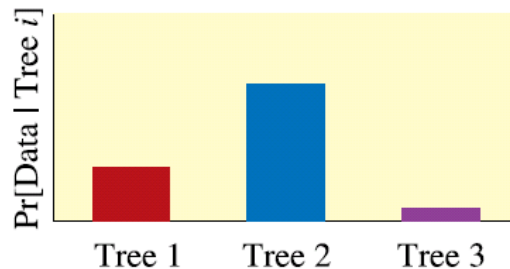
The **prior probability** of a tree represents the probability of the tree before the observations have been made. Typically, all trees are considered equally probable, a priori. However, other information can be used to give some trees more prior probability (e.g., the taxonomy of the group).



(1) 사전확률 (事前確率):
테이타를 관측하기전에 사
전에 가지고 있던 정보에
의한 확률(정보가 없는 경
우 균등한 확률을 생각하는
것이 무난함)

②

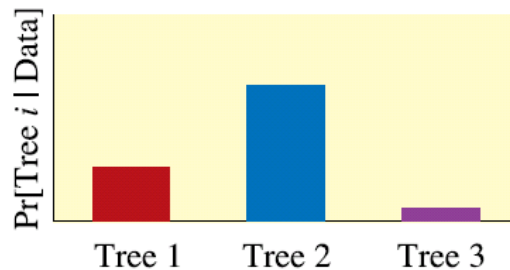
The **likelihood** is proportional to the probability of the observations (often an alignment of DNA sequences) conditional on the tree. This probability requires making specific assumptions about the processes generating the observations.



(2) likelihood: 테이타가
계통수에 얼마만큼 잘 적합
한가 나타내는 수치

③

The **posterior probability** of a tree is the probability of the tree conditional on the observations. It is obtained by combining the prior and likelihood for each tree using Bayes' formula.



(3) 사후확률(事後確率):
사전확률과 likelihood를 곱
해서 얻은 확률

Fig. 1. The main components of a Bayesian analysis.

$$P(\text{Tree} | \text{Data}) = \frac{P(\text{Data} | \text{Tree})P(\text{Tree})}{P(\text{Data})}$$

계통수의 사후확률 계산의 간단한 예 *

* *branch length*, 모델 파라미터 등의 사후분포도 생각해야 하므로, 실제의 계산은 이보다 훨씬 더 복잡함

Tree1, Tree2, Tree3의 계통수에 대하여 계산한 log-likelihood score가 각각 -10.0, -11.0, -12.0 일 경우, 베이즈 정리를 이용하여 사후확률을 계산해보자

각 계통수의 사전확률을 1/3이라고 가정하면 **Tree1**의 사후확률은...

$$\begin{aligned} P(T_1 | D) &= \frac{P(T_1, D)}{P(D)} = \frac{P(D | T_1)P(T_1)}{P(D)} \\ &= \frac{P(D | T_1)P(T_1)}{P(D | T_1)P(T_1) + P(D | T_2)P(T_2) + P(D | T_3)P(T_3)} \\ &= \frac{e^{-10.0} \times 1/3}{e^{-10.0} \times 1/3 + e^{-11.0} \times 1/3 + e^{-12.0} \times 1/3} \\ &\approx 0.6652 \end{aligned}$$

(참고) 사전확률이 각각 1/2, 1/4, 1/4 일 경우
Tree1의 사후확률을 계산하면 ?

계통수의 사후확률 계산의 간단한 예(2) *

* *branch length, 모델 파라미터 등의 사후분포도 생각해야 하므로, 실제의 계산은 이보다 훨씬 더 복잡함*

Tree1, Tree2, Tree3의 계통수에 대하여 계산한 log-likelihood score가 각각 -30.0, -33.0, -36.0 일 경우, 베이즈 정리를 이용하여 사후확률을 계산해보자 (5페이지의 예에 비하여 염기배열의 길이가 3배)

각 계통수의 사전확률을 1/3이라고 가정하면 **Tree1**의 사후확률은...

$$\begin{aligned} P(T_1 | D) &= \frac{P(T_1, D)}{P(D)} = \frac{P(D | T_1)P(T_1)}{P(D)} \\ &= \frac{P(D | T_1)P(T_1)}{P(D | T_1)P(T_1) + P(D | T_2)P(T_2) + P(D | T_3)P(T_3)} \\ &= \frac{e^{-30.0} \times 1/3}{e^{-30.0} \times 1/3 + e^{-33.0} \times 1/3 + e^{-36.0} \times 1/3} \\ &\approx 0.95033 \end{aligned}$$

(참고) 사전확률이 각각 1/2, 1/4, 1/4 일 경우 Tree1의 사후확률을 계산하면 ?

중요포인트: 데이터의 수가 많으면 사전확률은 사후 확률에 거의 영향을 미치지 않음. 사전확률선택의 주관성에 대한 걱정 불필요.

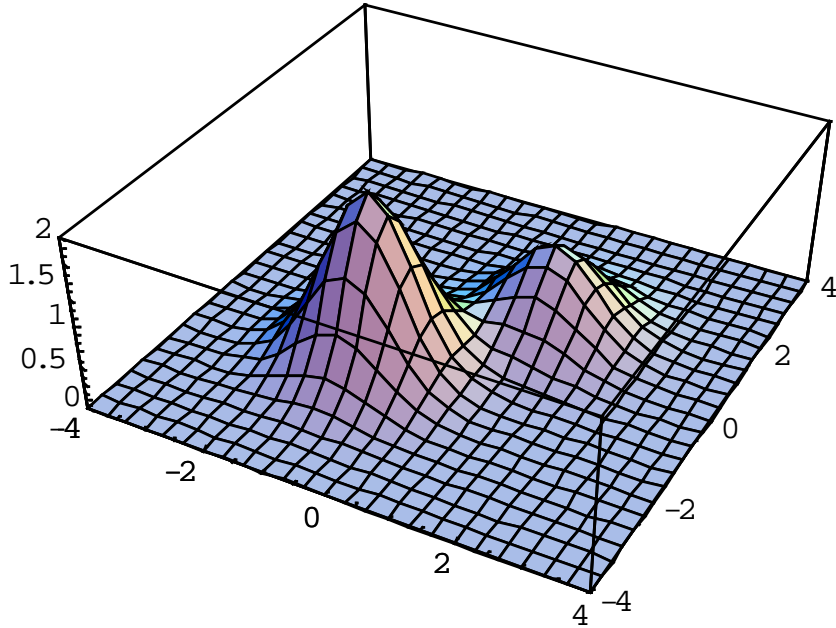
MCMC (Markov chain Monte Carlo)

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)} \quad (D: \text{data}, \theta: \text{parameter})$$
$$= \frac{P(D | \theta)P(\theta)}{\sum_{\theta} P(D | \theta)P(\theta)} \quad \text{or} \quad \frac{P(D | \theta)P(\theta)}{\int_{\theta} P(D | \theta)P(\theta)}$$

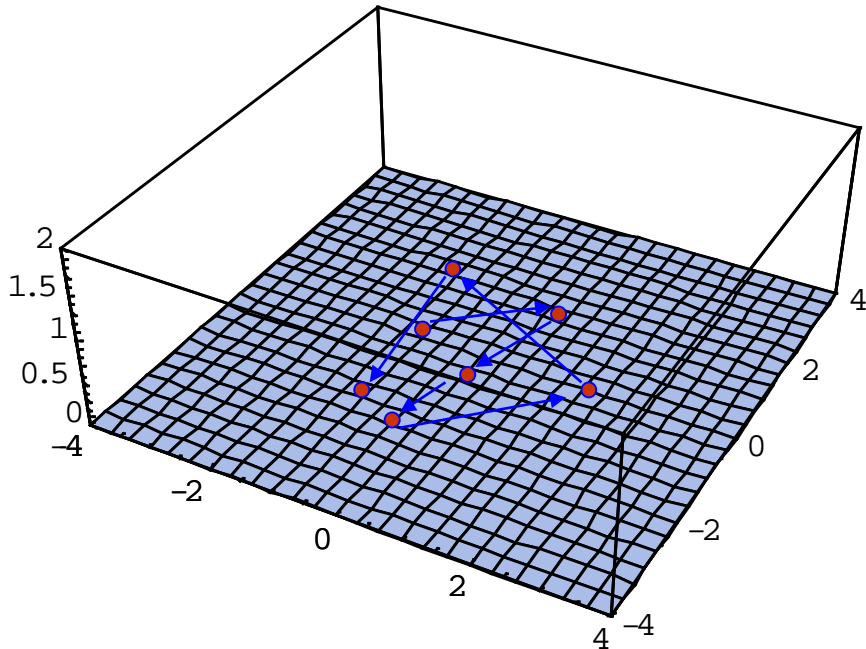
- (1) $P(D)$ 를 직접 계산하는것은 어려움 (고차원의 파라미터의 경우 다중적분에 상당한 계산시간이 소요됨)
- (2) $P(D)$ 는 파라미터 θ 에 의존하지 않는 확률이므로 직접 계산하지 않아도 사후확률을 구할수 있음 -> MCMC 이용

MCMC = Metropolis–Hastings algorithm

MCMC (Markov chain Monte Carlo)의 원리



(1) 이러한 형태의 사후분포로부터 샘플링을 하려고 할 때...

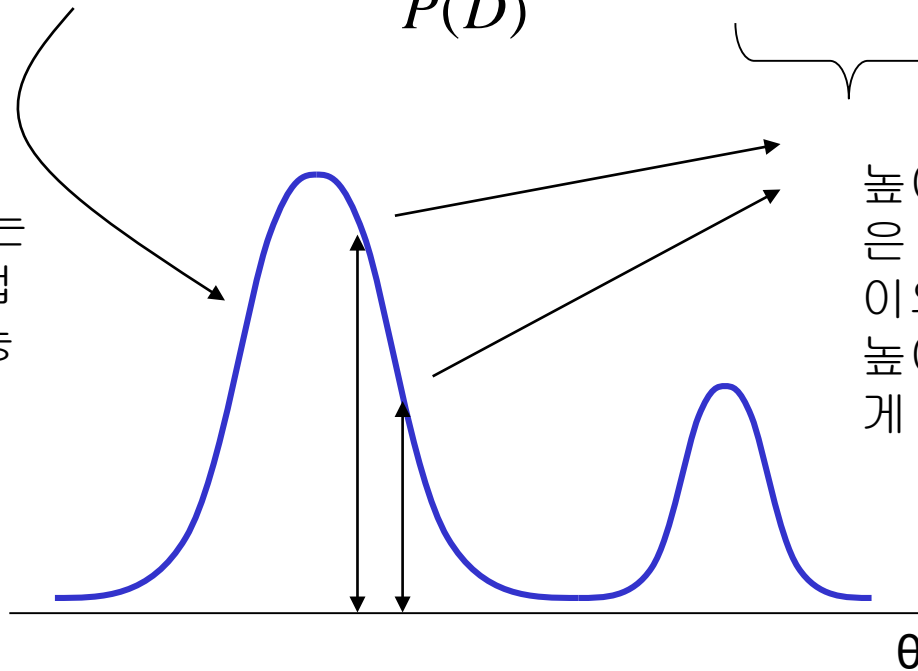


(2) 샘플링될 데이터 포인트를 랜덤하게 움직인다 (점의 존재 가능성이 사후분포의 높이에 비례하도록...). 랜덤하게 움직이면서 정기적으로 샘플링함.

MCMC(Markov chain Monte Carlo)의 원리

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)} \propto P(D | \theta)P(\theta)$$

$P(D)$ 를 모르는
상태에서 직접
계산은 불가능

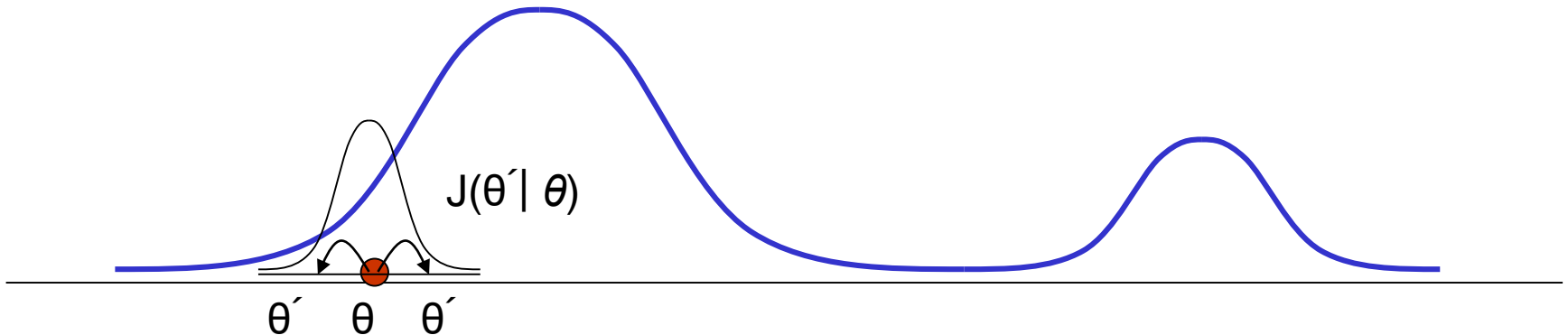


높이의 정확한 값
은 모르지만 두 높
이의 비 (상대적인
높이)는 비교적 쉽
게 계산 할수 있음

D: data θ : parameter

MCMC(Markov chain Monte Carlo)의 원리

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)} \propto P(D | \theta)P(\theta)$$



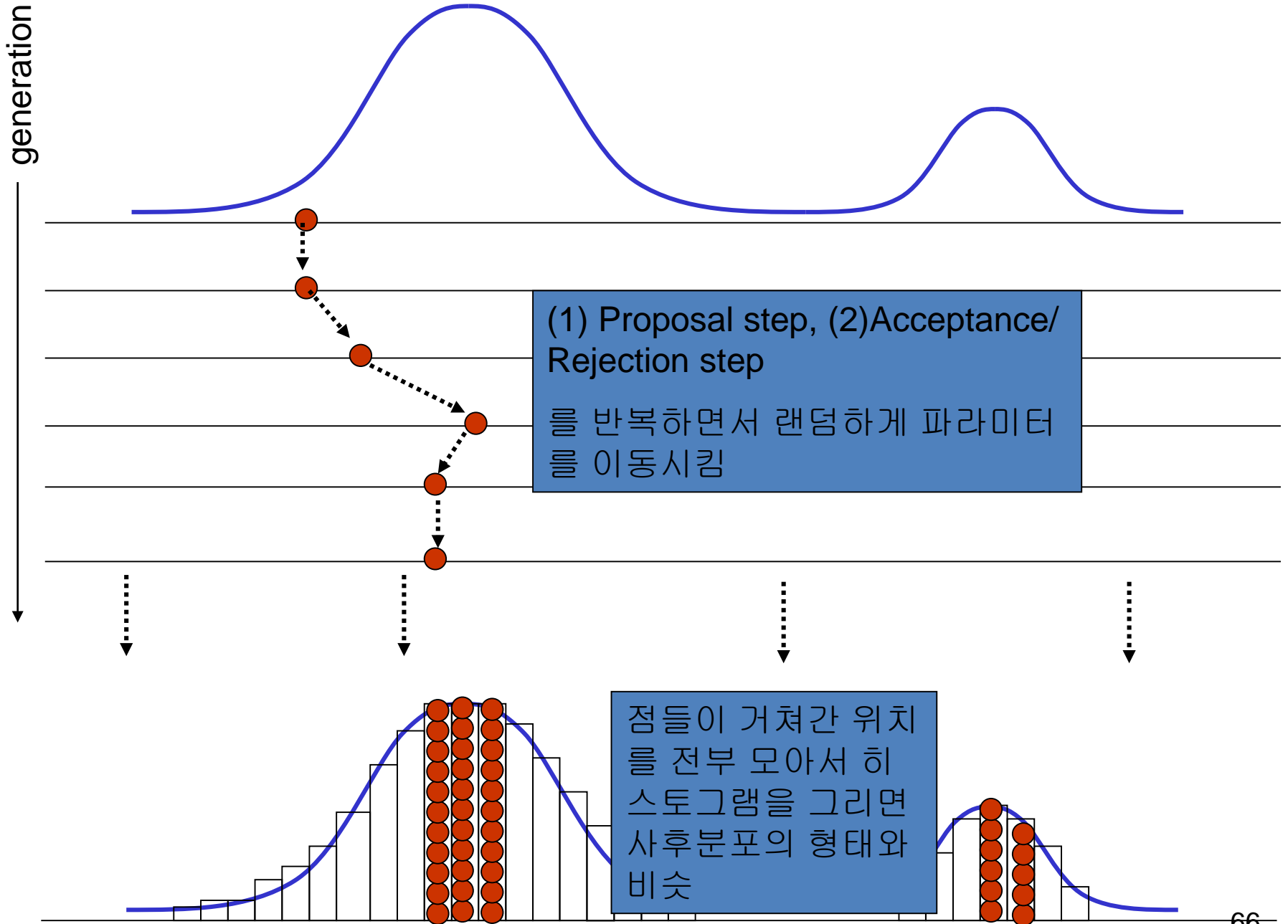
(1) Proposal step (현재의 파라미터의 다음 위치를 결정)

(2) Acceptance/rejection step (다음 위치를 채택/기각 할지 결정)

반복하면
서 점을
움직임

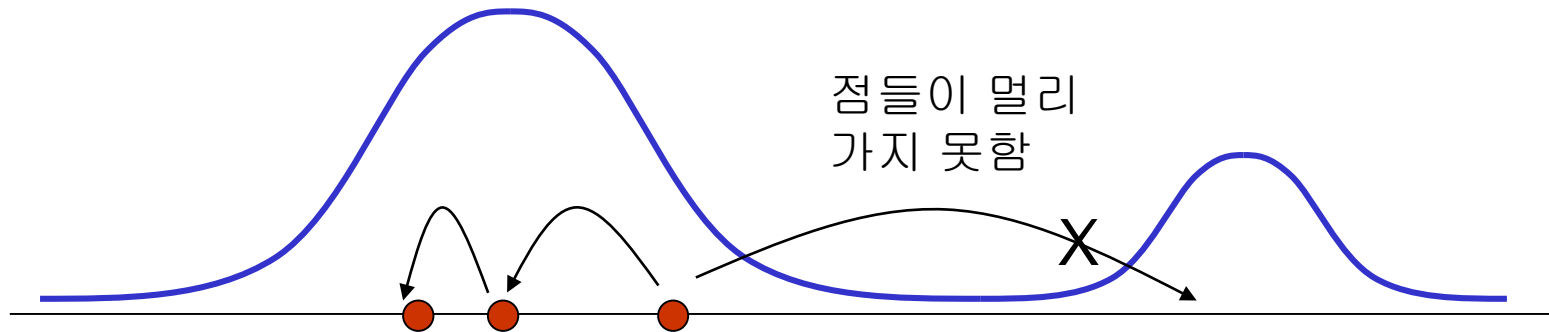
Accept proposed state with probability $\min\left(1, \frac{P(\theta' | D)J(\theta | \theta')}{P(\theta | D)J(\theta' | \theta)}\right)$

generation

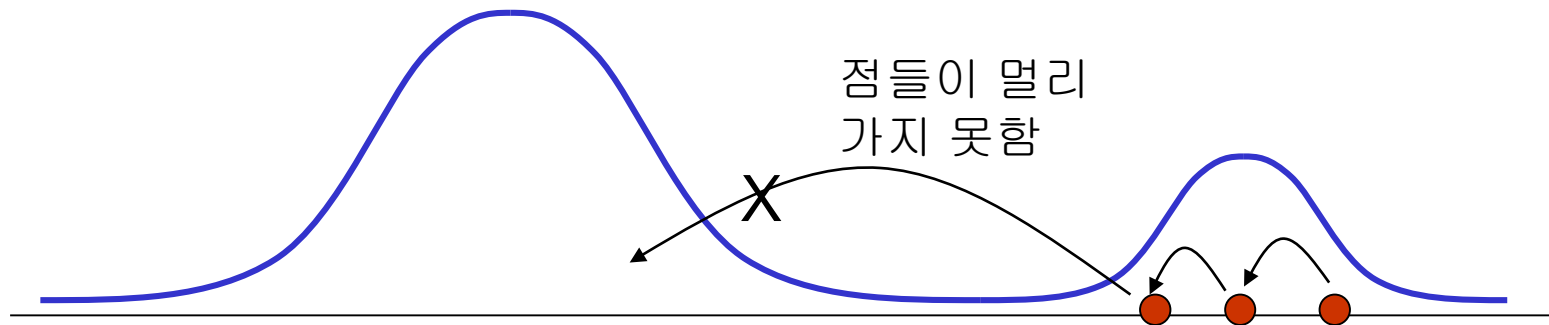


Proposal step에서 θ' 와 θ 사이의 거리가 중요

거리가 짧은 경우



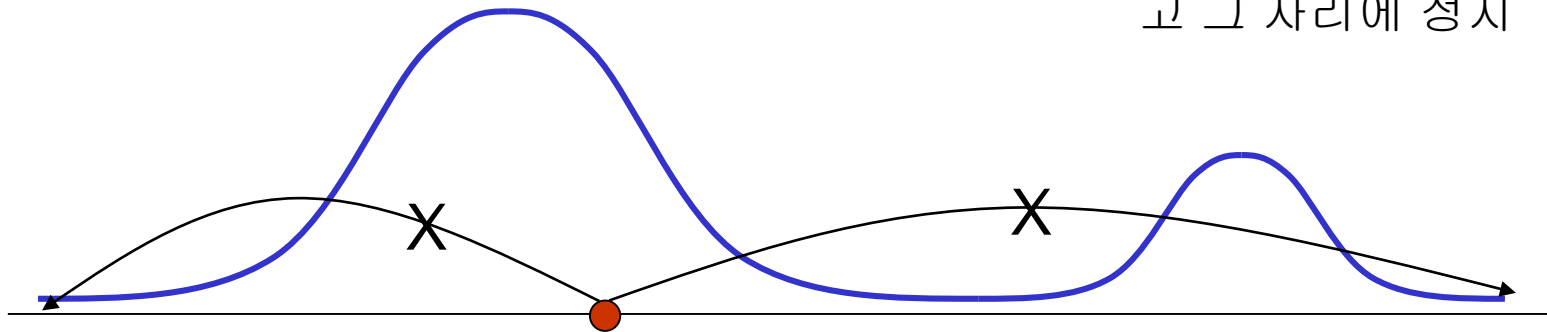
서로다른 MCMC
실행



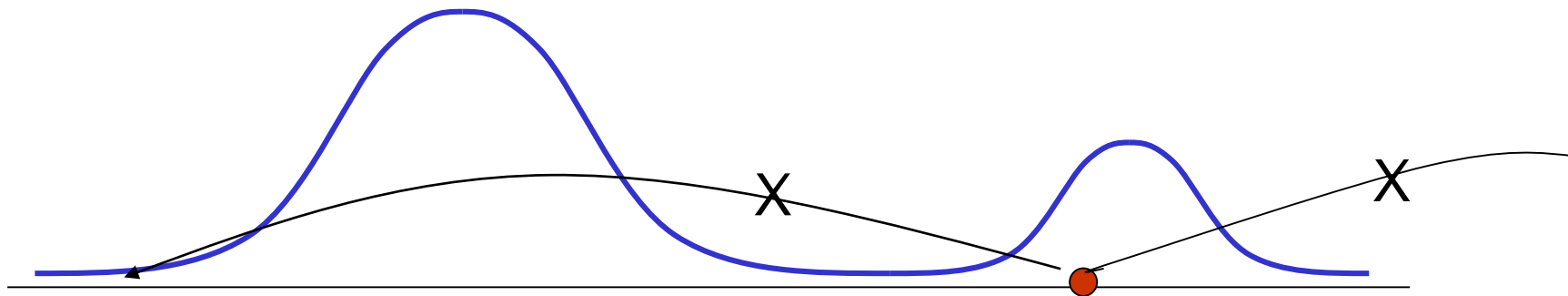
Proposal step에서 θ' 와 θ 사이의 거리가 중요

거리가 먼 경우

Acceptance rate이 낮다.
→ 점들이 움직이지 않고
그 자리에 정지



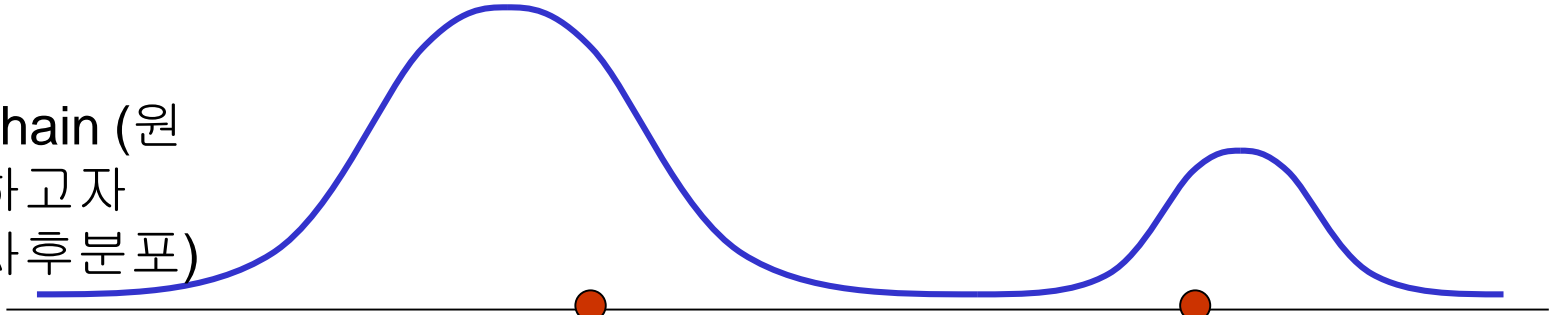
서로다른 MCMC
실행



Cold chain & Hot chain

(θ' 와 θ 사이의 거리를 정하기 힘들때 유용한 수단)

Cold chain (원래 구하고자 하는 사후분포)



두종류의 Markov chain을 동시에 실행하면서 정기적으로 점을 교환함

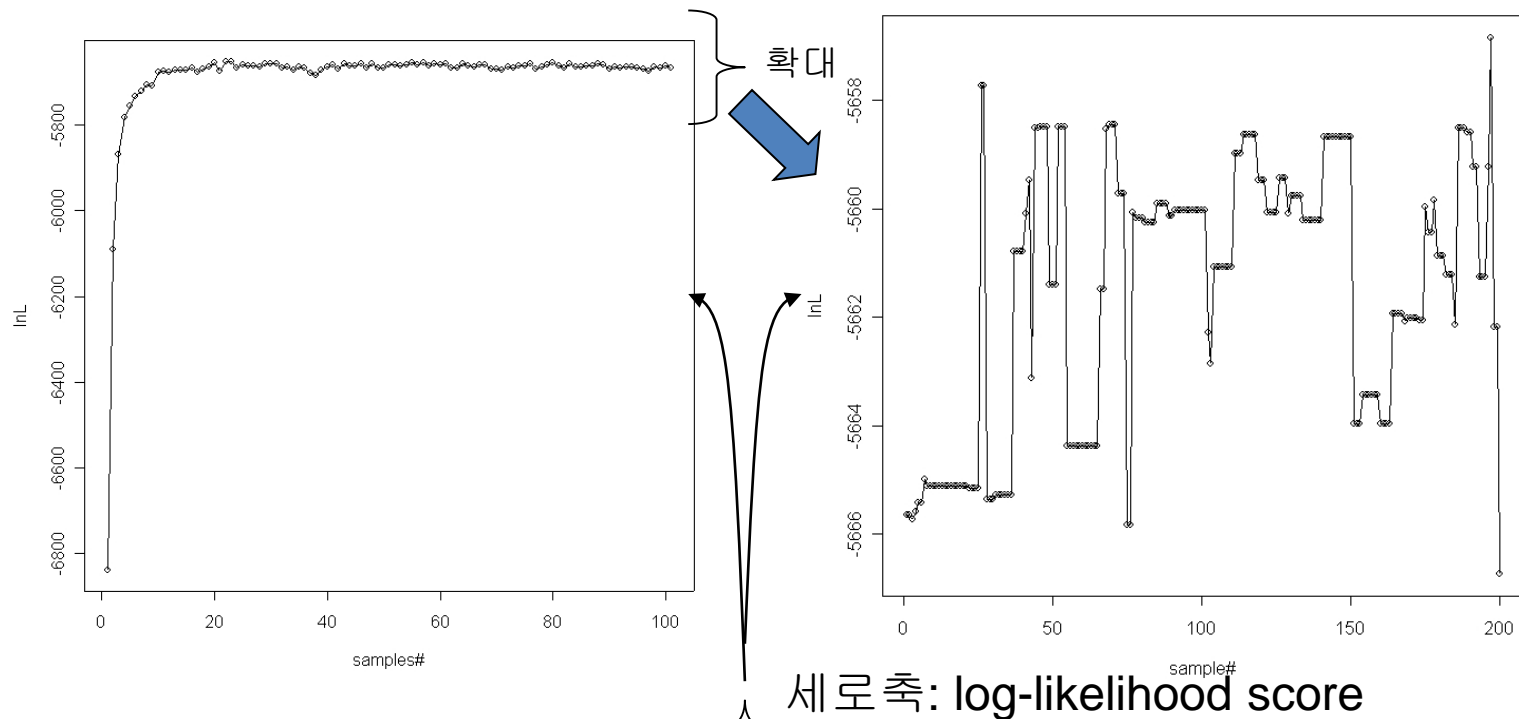
Hot chain (원래 구하고자 하는 사후분포의 높이를 전체적으로 낮춤)



Hot chain: 열을 가해서 녹아내린 모습을 연상하면 이해하기 쉬움. θ' 와 θ 의 거리에 비교적 영향을 덜 받고 점들이 잘 움직임

Number of burn-in : MCMC의 결과가 초기치의 영향을 받지 않게 하기 위해 초기의 샘플을 무시한다. 무시하는 **generation**의 수

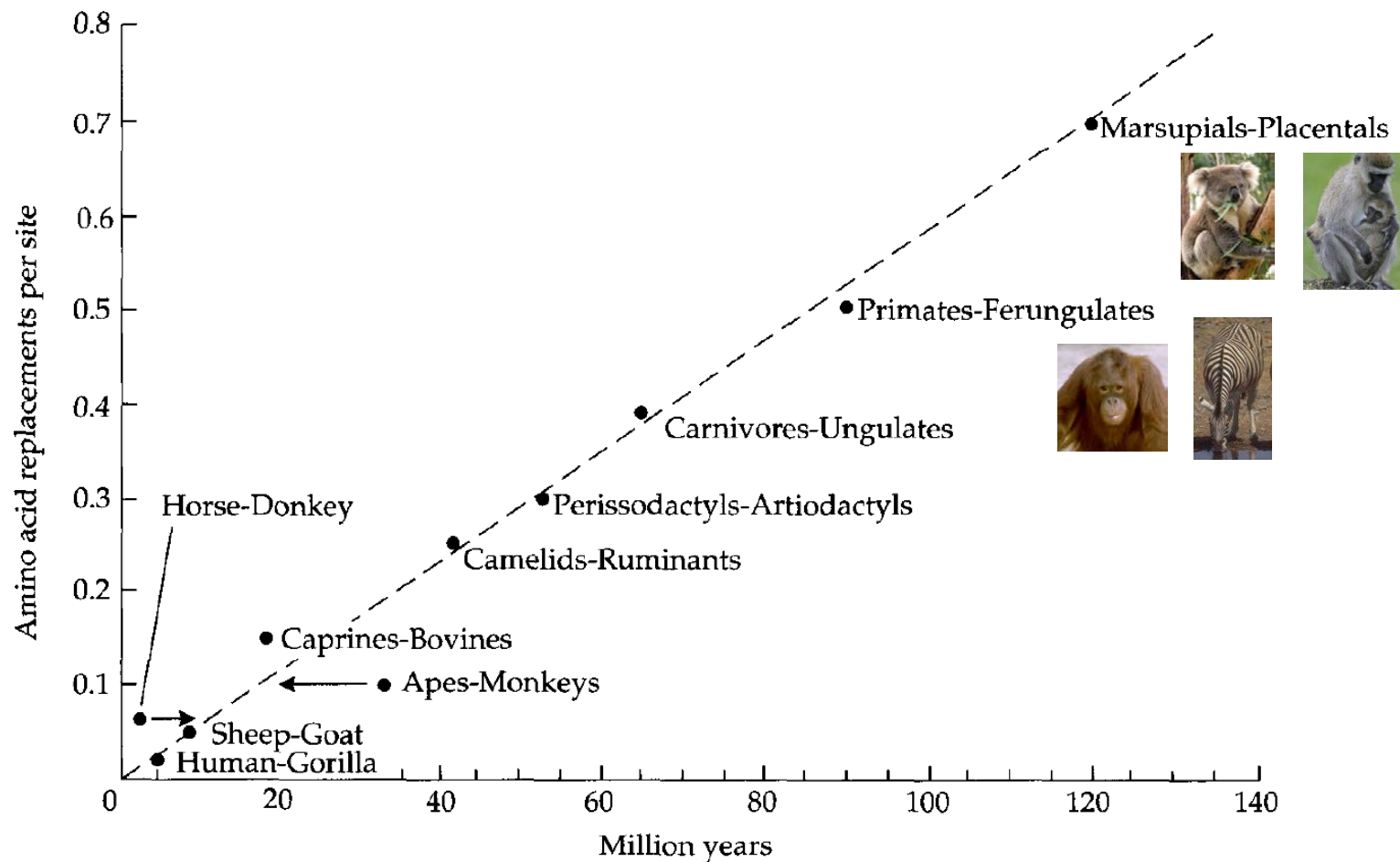
Number of interval : MCMC로 부터 샘플링되는 점들 사이의 상관관계를 줄이기 위해서 일정 **generation**간격으로 점들을 샘플링. 그 간격의 수



$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)} \propto \underbrace{P(D | \theta)P(\theta)}_{\text{세로축: log-likelihood score}}$$

분자계통수를 이용한 분기연대 추정

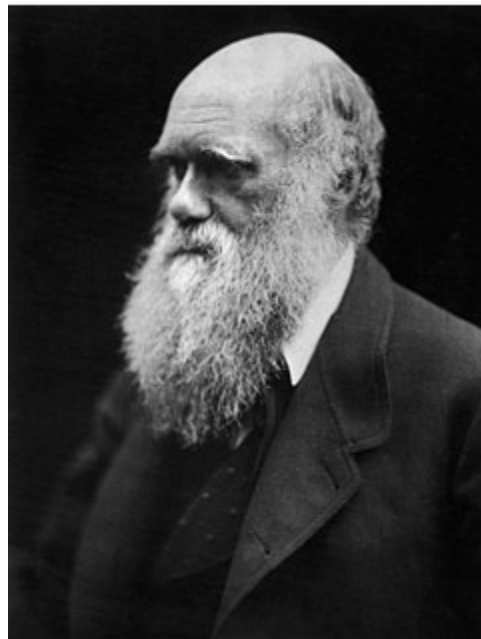
분자진화의 중립설 (neutral theory of molecular evolution) 과 분자시계 (molecular clock)



헤모글로빈 α , β , 시토크롬 c , 피브리노겐타이드A 아미노산 치환 속도
(Langley and Fitch 1974; Graur & Li 2000)



라마르크(용불용설)
(1744~1829)



다윈 (종의기원)
(1809~1882)



기무라 (분자진화의
중립설)
(1924~1994)

돌연변이의 분류 및 고정(fixation)

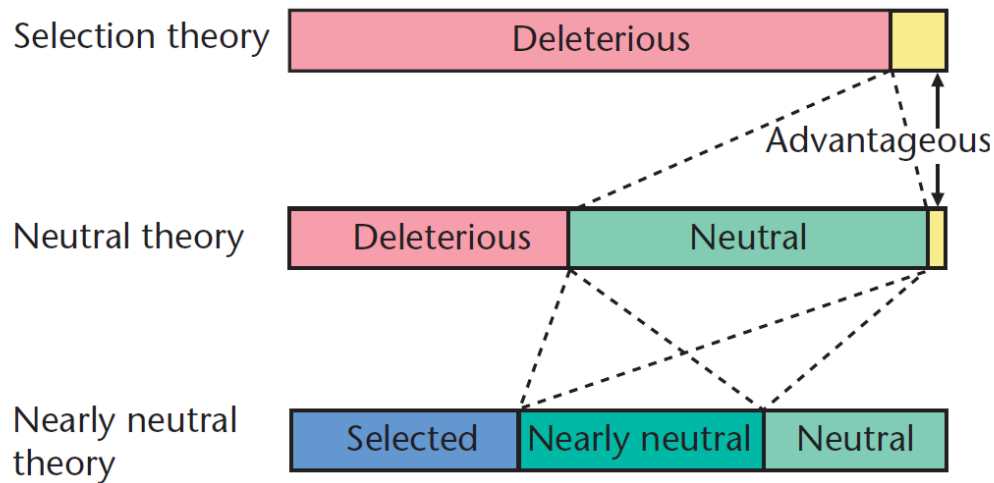


Figure 1 The classification of new mutants under the selection, neutral and nearly neutral theories. Note that while most selected mutants are deleterious, the group also includes advantageous mutants.

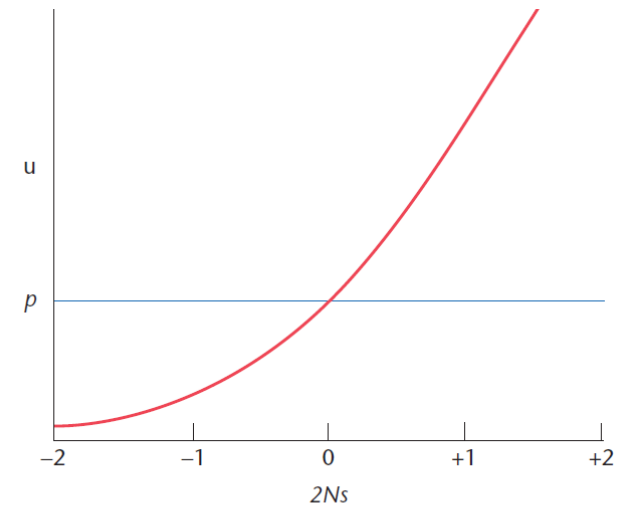
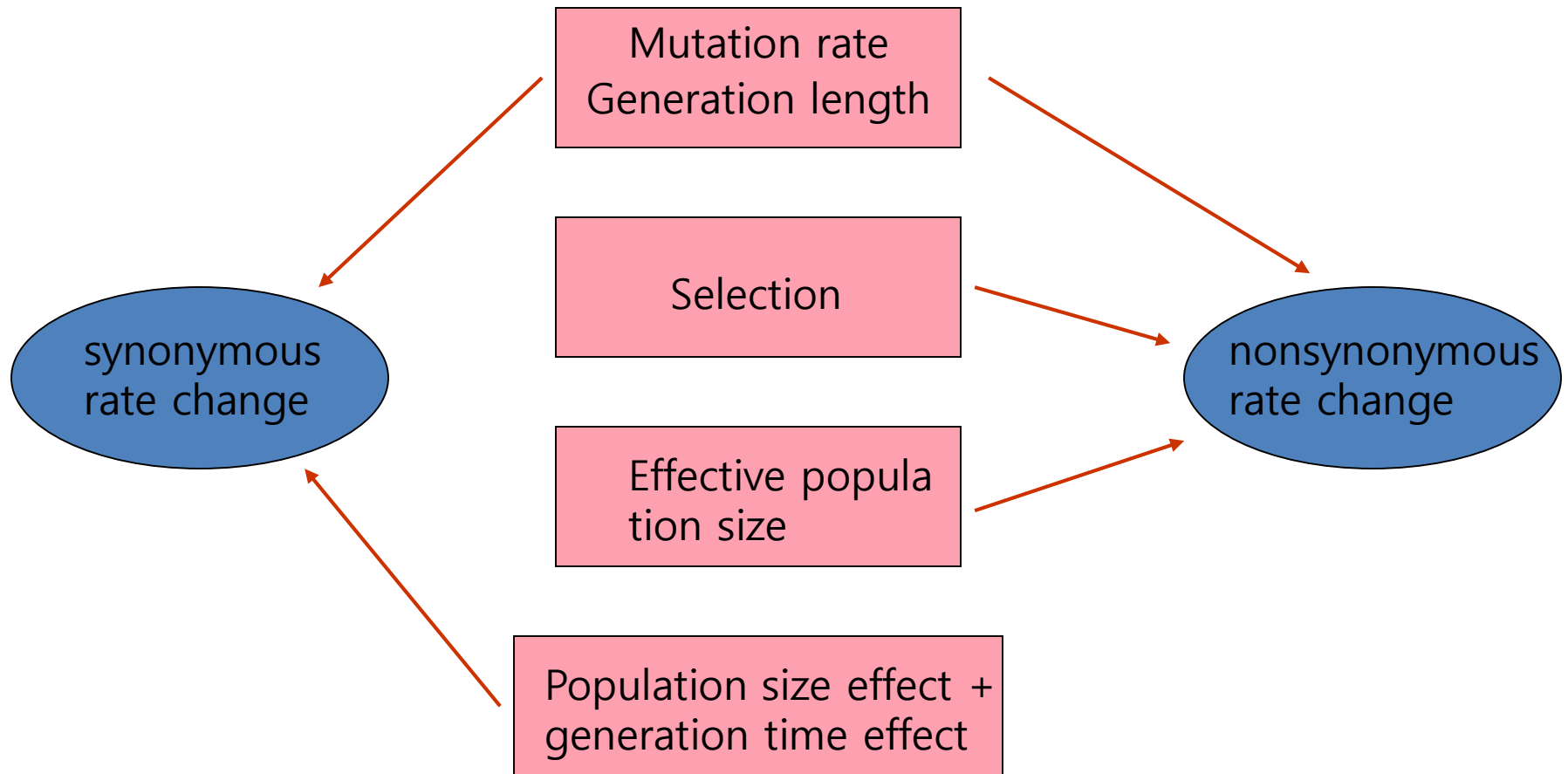


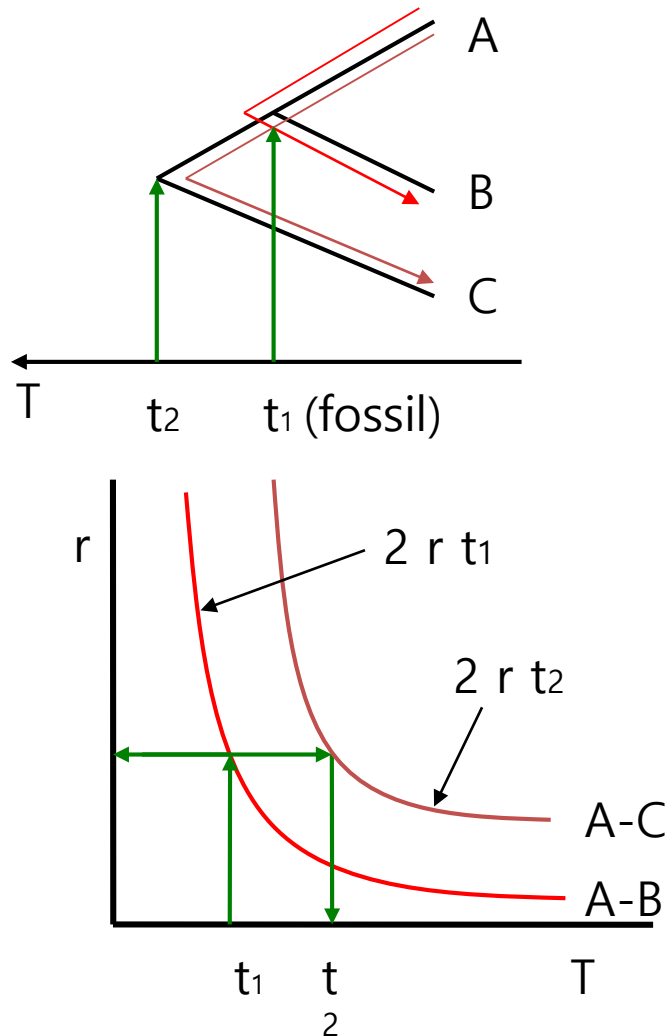
Figure 2 Fixation probability of a mutant in a finite population as a function of $2Ns$. p is the initial frequency of the mutant. The region of $2Ns < 0$ is that of slightly deleterious mutations.

(Ohta 2008)

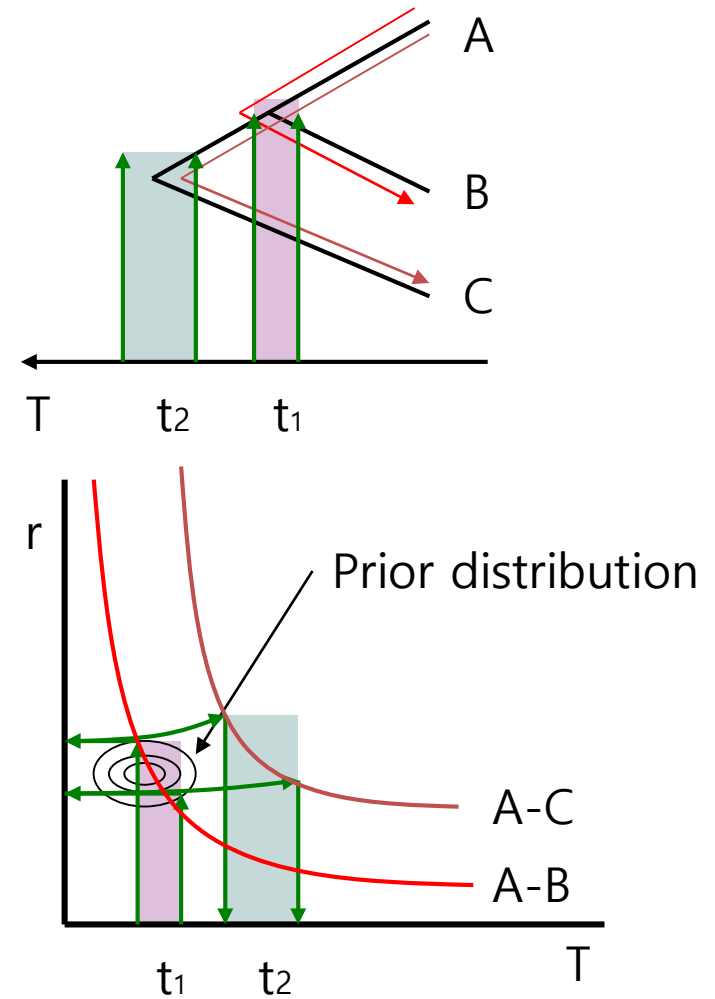
진화속도 변화의 요인



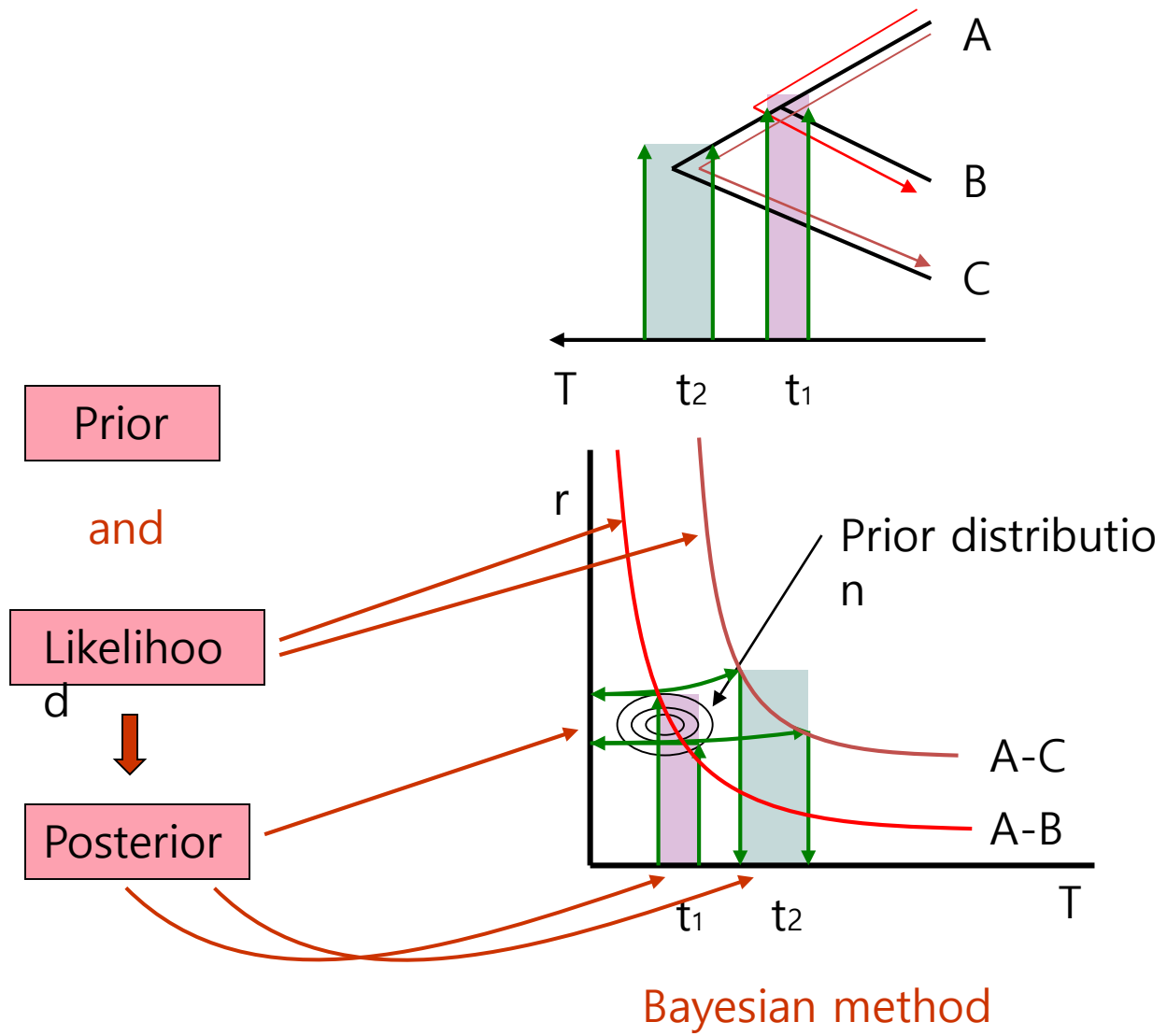
Molecular clock vs. Bayesian method



진화속도가 계통수 전체에 있어서 일정하다고 가정

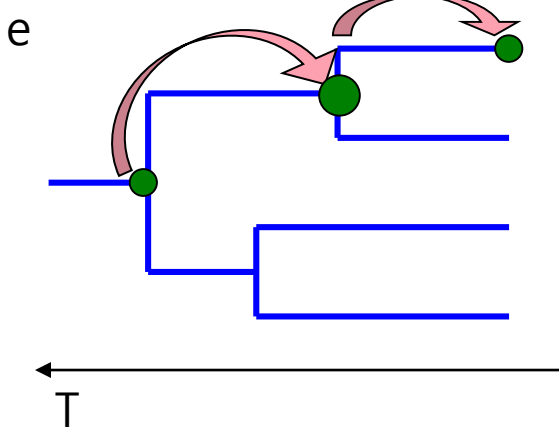


진화속도가 변할수 있다고 가정



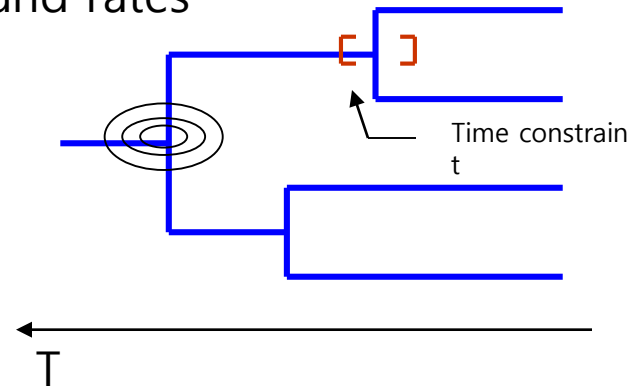
베이지안 분기연대 추정 (autocorrelated rate change model)

(1) Model of rate change



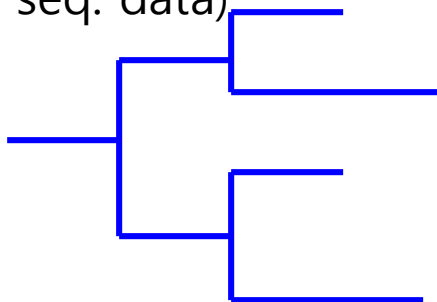
and

(2) Prior distribution of time and rates

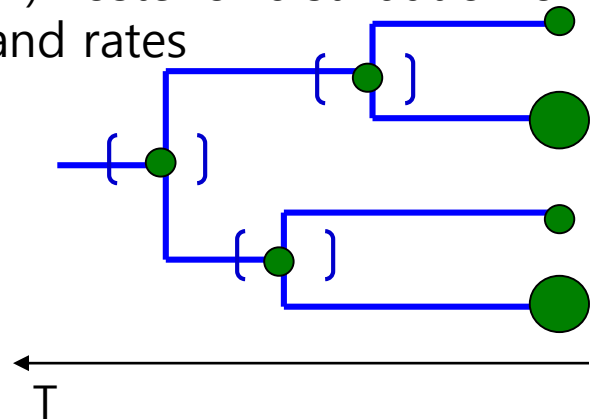


(3) Estimated branch lengths
(from seq. data)

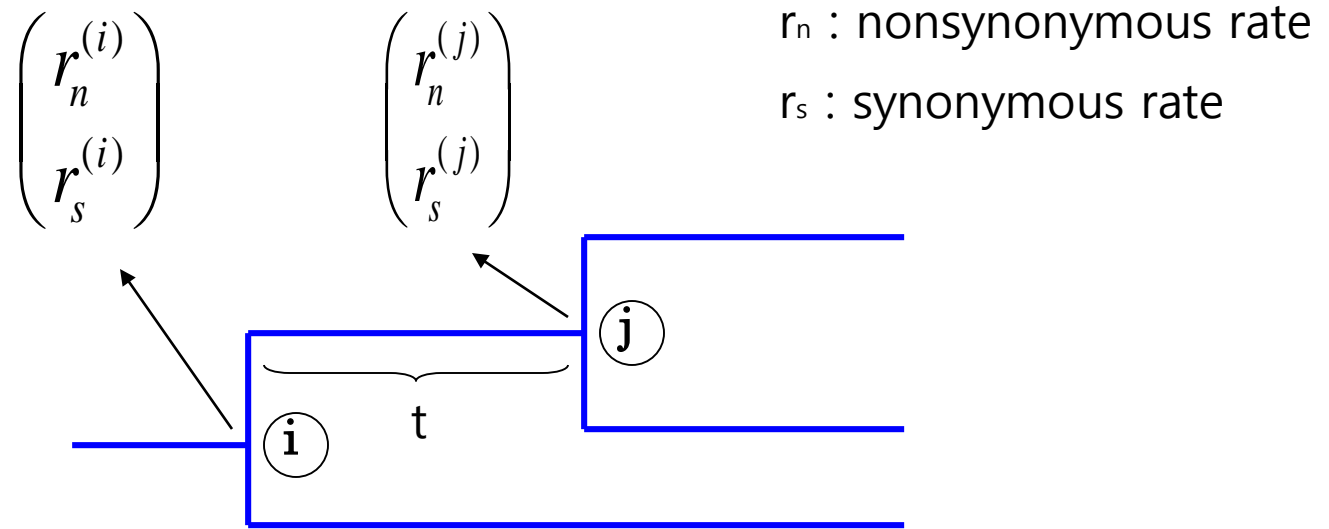
and



(4) Posterior distribution of time and rates



- 진화속도 변화 모델 : {log-normal 분포, 코돈모델}의 예시



$$\begin{pmatrix} \log r_n^{(j)} \\ \log r_s^{(j)} \end{pmatrix} \sim N \left(\begin{pmatrix} \log r_n^{(i)} \\ \log r_s^{(i)} \end{pmatrix}, \begin{pmatrix} \nu_n t & 0 \\ 0 & \nu_s t \end{pmatrix} \right)$$

- Posterior distribution (사후분포) of parameters

(MCMC 알고리즘을 이용하여 수치적으로 계산)

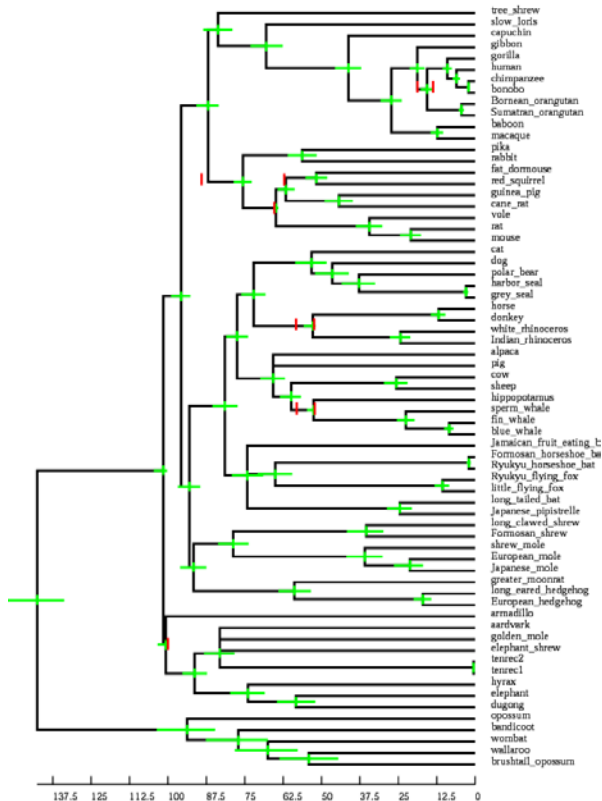
Divergence times
 Evolutionary rates
 Rate variation
 Sequence data

$$P(\mathbf{T}, \mathbf{r}_n, \mathbf{r}_s, \mathbf{v}_n, \mathbf{v}_s | \mathbf{X}) = \frac{P(\mathbf{X} | \mathbf{T}, \mathbf{r}_n, \mathbf{r}_s, \mathbf{v}_n, \mathbf{v}_s) P(\mathbf{T}, \mathbf{r}_n, \mathbf{r}_s, \mathbf{v}_n, \mathbf{v}_s)}{P(\mathbf{X})}$$

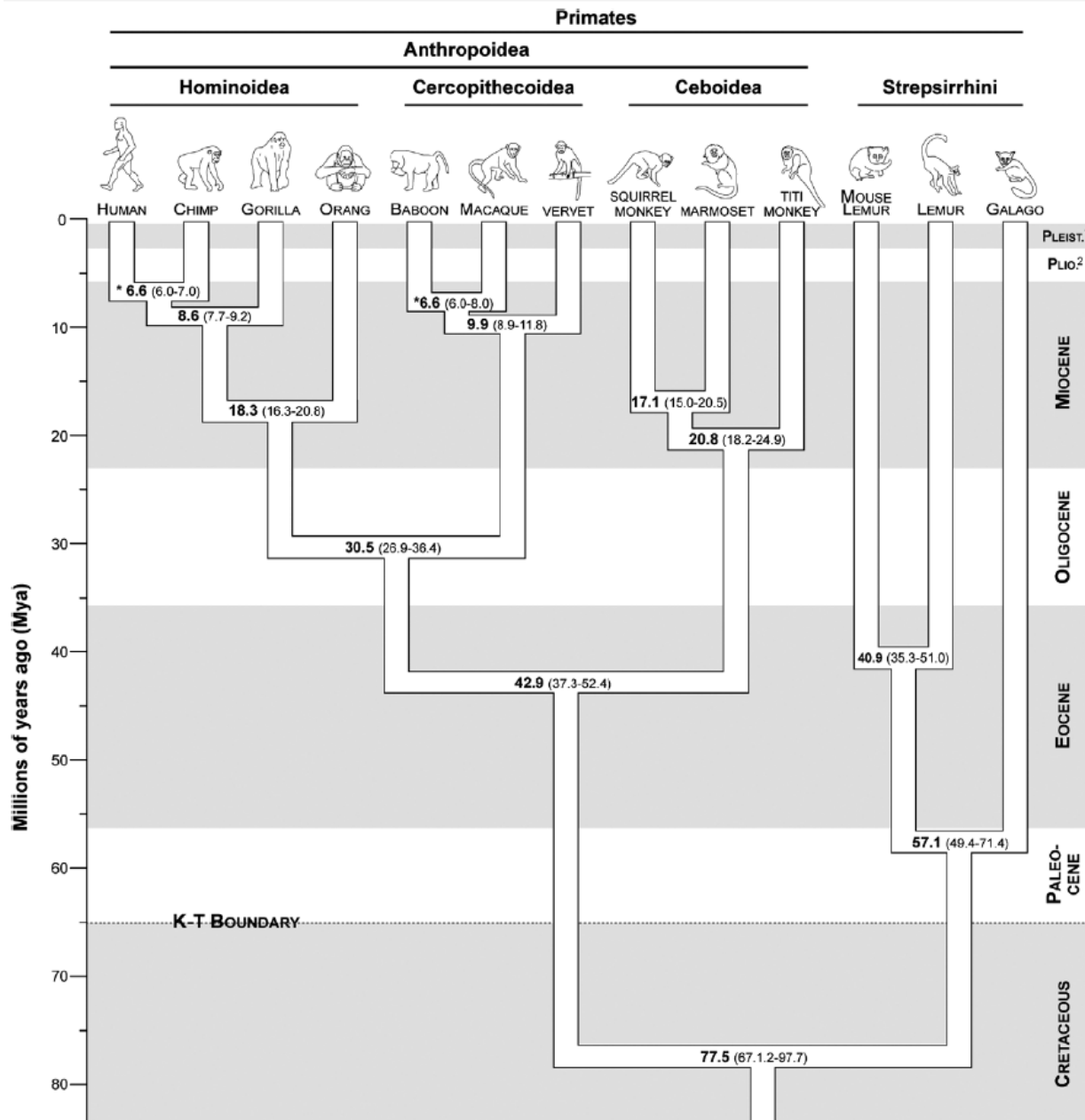
$$= \frac{1}{P(\mathbf{X})} \underbrace{P(\mathbf{X} | \mathbf{T}, \mathbf{r}_n, \mathbf{r}_s)}_{\text{Phylogenetic likelihood calculation (Codon model)}} \underbrace{P(\mathbf{r}_n, \mathbf{r}_s | \mathbf{T}, \mathbf{v}_n, \mathbf{v}_s)}_{\text{Lognormal distribution}} \underbrace{P(\mathbf{T}) P(\mathbf{v}_n) P(\mathbf{v}_s) P(\mathbf{r}_n) P(\mathbf{r}_s)}_{\text{Prior distribution}}$$

Phylogenetic likelihood calculation (Codon model)
 Lognormal distribution
 Prior distribution
 \mathbf{T} : Gamma + modified Dirichlet
 $\mathbf{v}_n, \mathbf{v}_s$: exponential
 $\mathbf{r}_n, \mathbf{r}_s$: exponential

미토콘드리아 유전자를 이용한 포유류의 분기연대 추정 (Seo et al. 2008)

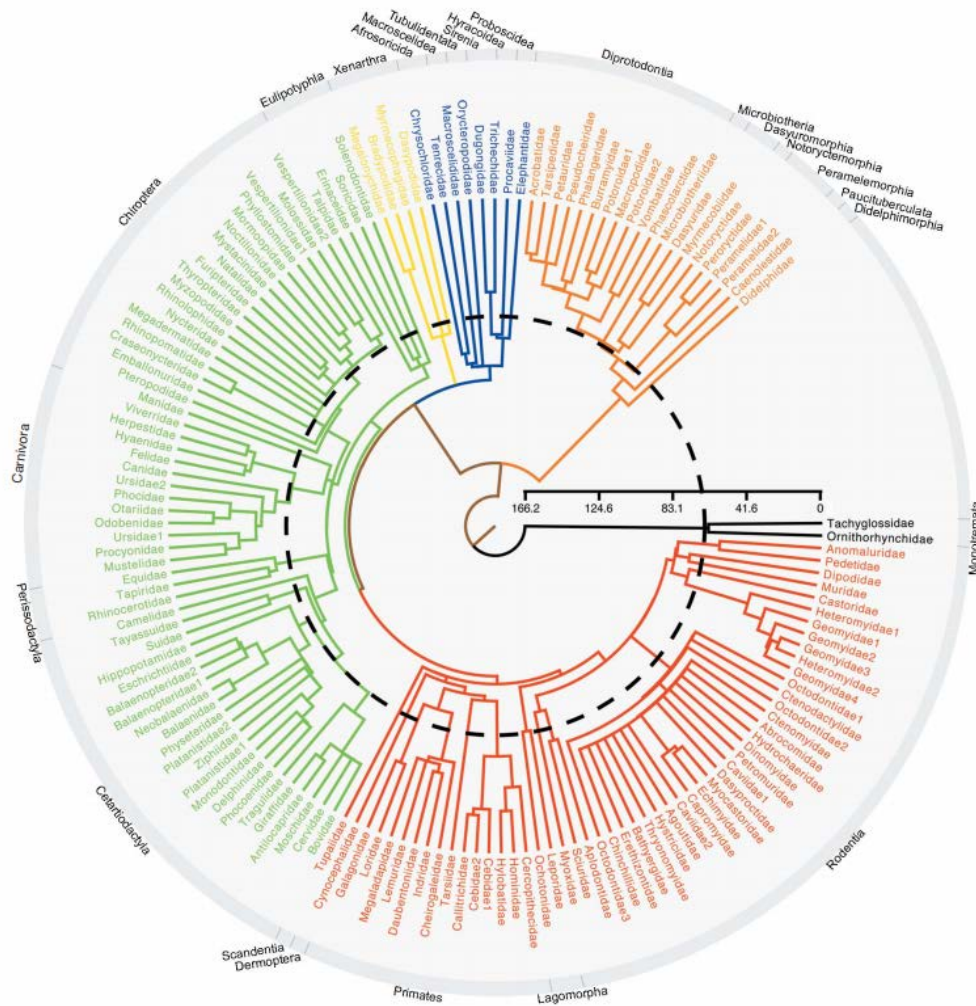


Branching	Mt-Codon ^(a)	Mt-proteins ^(b)	12 Nuclear + 2 Mt ^(c)	19 Nuclear + 3 Mt ^(d)
Base of Afrotheria	91.3 ± 2.0	79.9 ± 2.9	82.8 ± 2.7	79.9 ± 3.0
Euarchontoglires/Laurasiatheria	95.7 ± 1.6	—	91.2 ± 1.8	94.0 ± 3.4
Euarchonta/Glires	87.0 ± 1.8	89.0 ± 1.9	81.6 ± 1.8	87.3 ± 3.2
Base of Euarchonta	83.7 ± 2.4	—	78.4 ± 2.2	86.0 ± 3.1
Base of Primates	68.2 ± 2.7	—	73.1 ± 2.7	77.1 ± 3.3
Patyrrhini/Catarrhini	41.3 ± 2.2	—	37.5 ± 3.1	—
Hominoidea/Cercopithecoidea	27.3 ± 1.7	34.6 ± 1.6	25.5 ± 2.7	—
Human/gibbon	18.9 ± 1.3	21.7 ± 1.0	15.6 ± 2.1	—
Human/chimpanzee	6.13 ± 0.60	7.4 ± 0.7	—	—
Base of Lagomorpha (rabbit/pika)	56.3 ± 2.4	—	50.5 ± 3.2	50.9 ± 4.0
Mouse/rate	21.0 ± 1.8	16.2 ± 1.4	16.0 ± 1.9	16.3 ± 2.2
Base of Eulipotyphla	91.6 ± 2.2	61.0 ± 3.1	75.3 ± 3.1	75.9 ± 2.3
Base of Chiroptera	74.1 ± 2.6	65.2 ± 2.9	74.9 ± 3.0	65.3 ± 1.4
Base of Carnivora	53.3 ± 2.6	49.0 ± 2.7	56.8 ± 3.2	55.1 ± 2.5
Base of Cetartiodactyla	65.7 ± 2.1	64.1 ± 2.3	67.3 ± 2.7	63.8 ± 0.8
White rhinoceros/Indian rhinoceros	24.3 ± 1.7	26.1 ± 2.3	—	—
Opossum/wallaroo	93.8 ± 4.8	99.5 ± 5.4	—	—



영장류의 분기연대 추정 결과 (Steiper & Young 2006)

포유류의 분기연대 추정 결과



K-T boundary 이전에도 포유동물은 상당히 진화가 진행되었음을 알수 있다