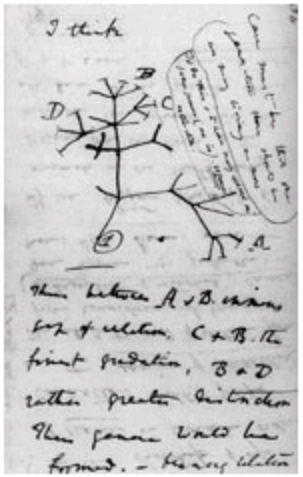
2026년도 한국진화학회 겨울학교

(분자진화학분야)

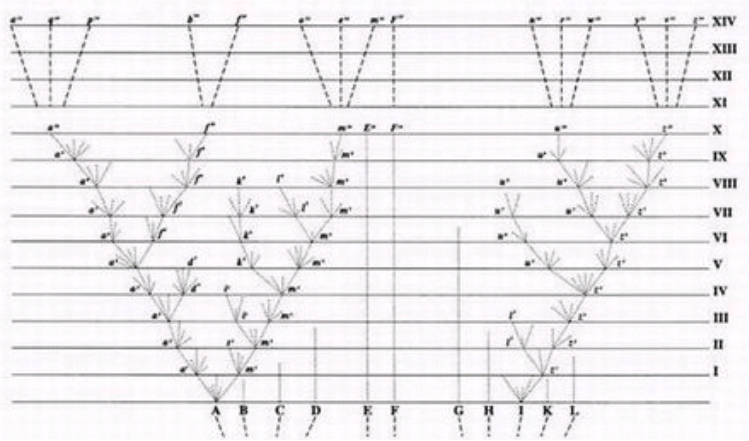
극지연구소
서태건

(ver.2026.01.26)

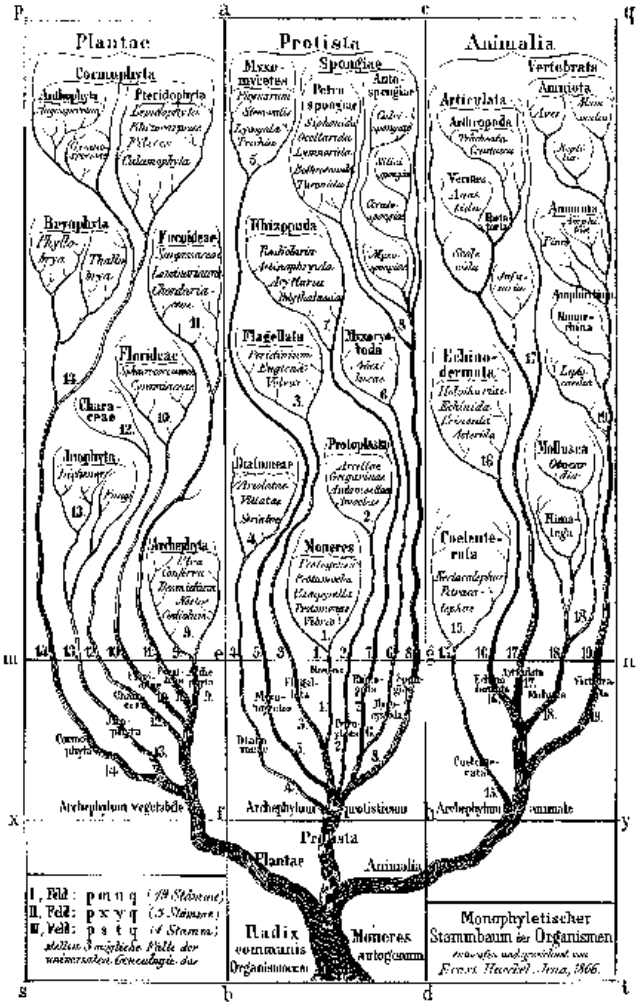
계통수(phylogenetic tree)와 생물의 진화/유연 관계



Darwin's notebook (1837)



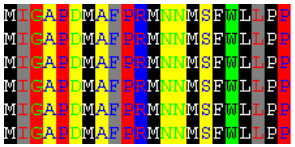
On the Origin of Species by Natural Selection (Darwin, 1859)

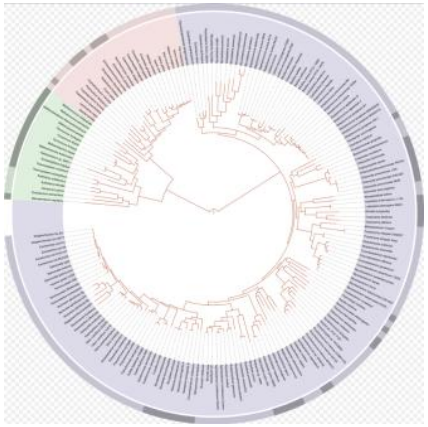


Generelle Morphologie der Organismen (Haeckel 1866)

분자계통수의 추정이 다른 연구에 미치는 영향은 매우 크다

염기/아미노산 서열 데이터

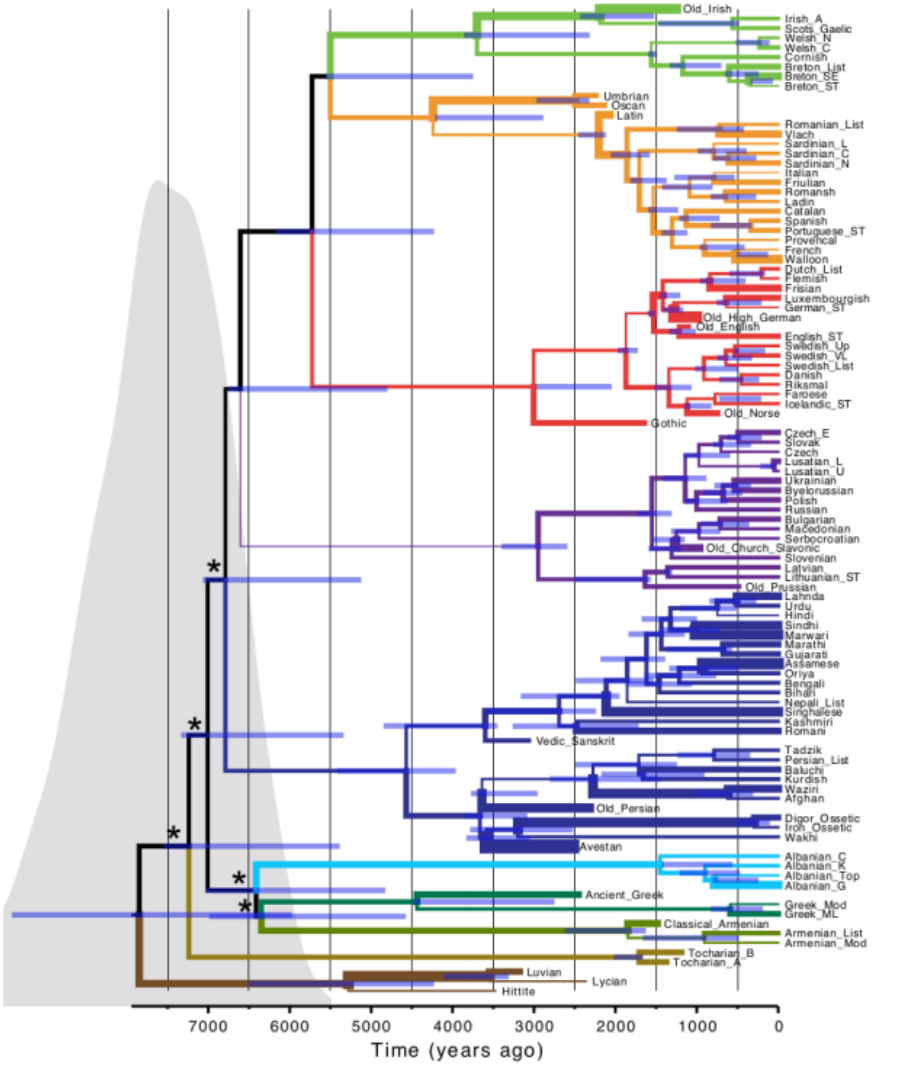




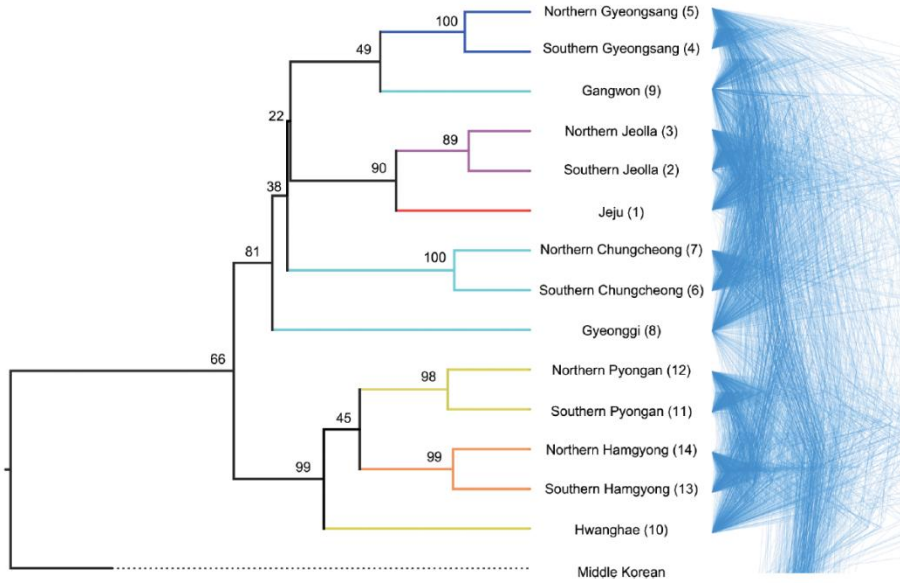
통계학적 모델/분석방법

- 종 분화 과정 규명
- 분기연대 추정
- 적응 진화 기작 연구
- 단백질 구조 연구
- 계통지리학 연구
- 바이러스 백신 개발에 응용
- ⋮

계통수 추정은 무생물 (언어)의 진화과정 규명에도 이용된다

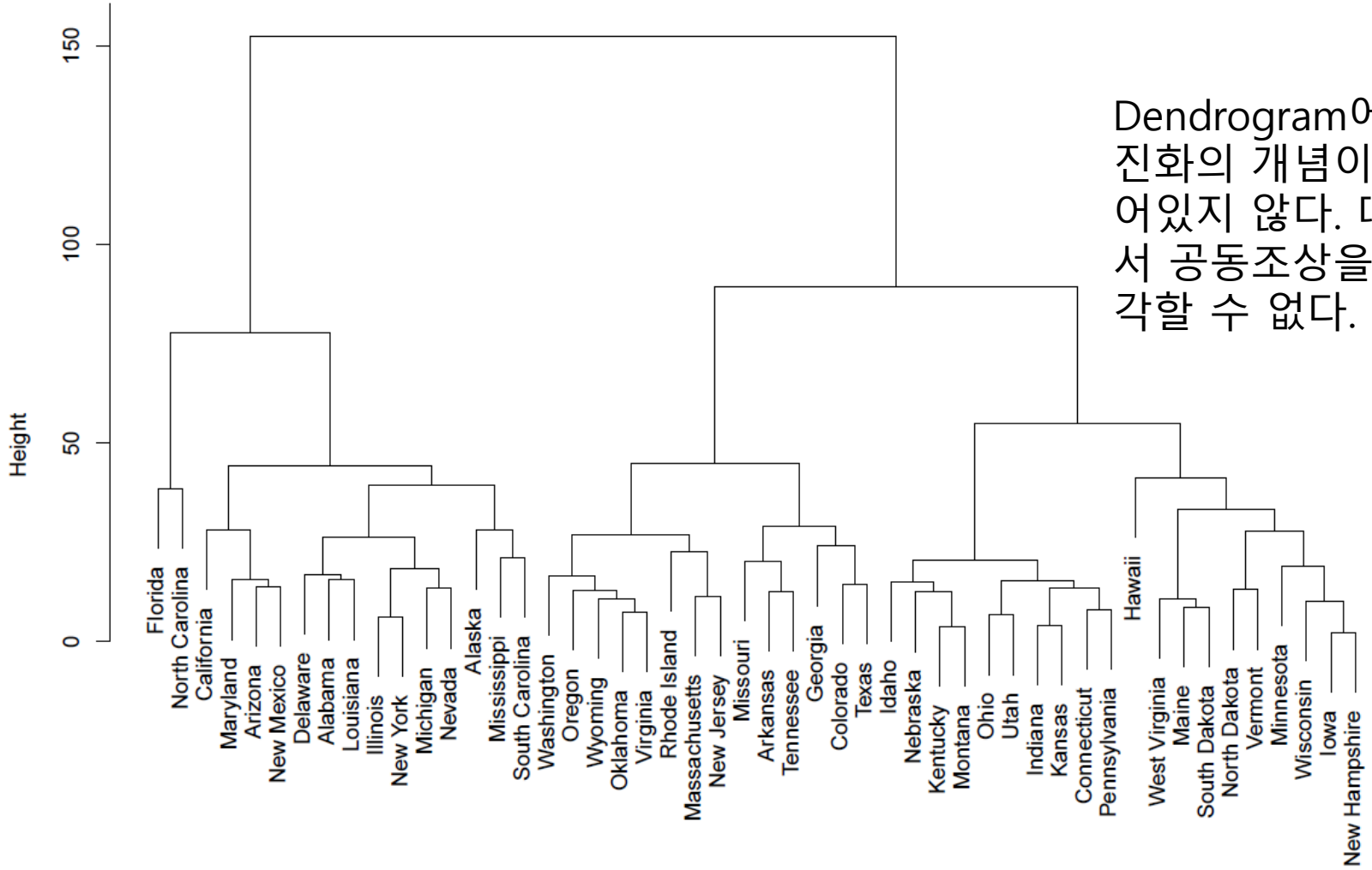


(Bouckaert et al. 2012)



(Lee 2015)

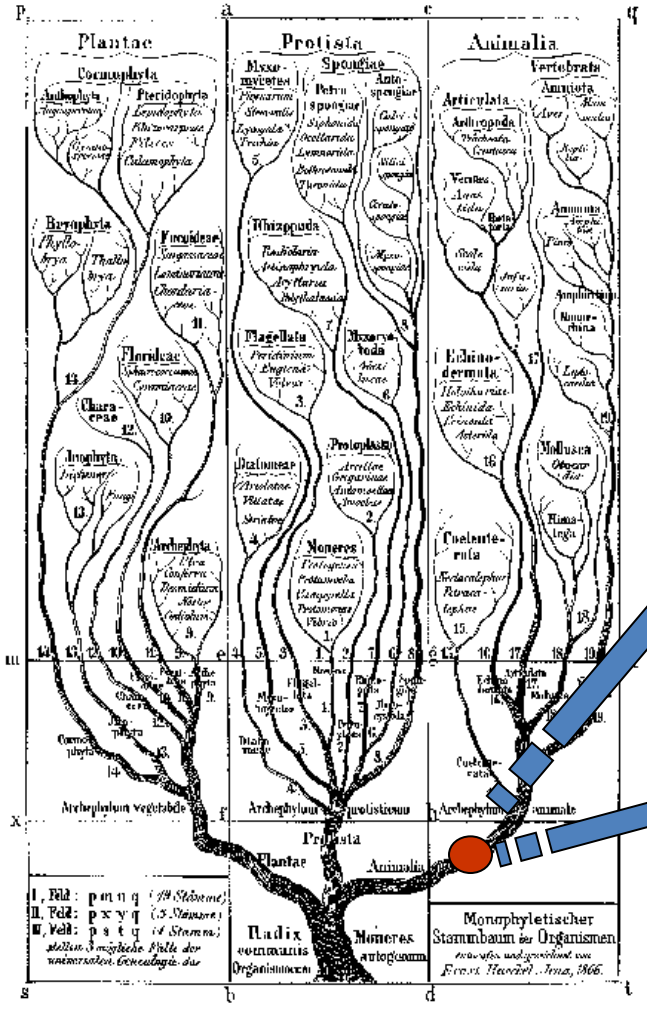
Phylogeny 와 dendrogram(clustering)은 다르다



Dendrogram에는 진화의 개념이 들어있지 않다. 따라서 공동조상을 생각할 수 없다.

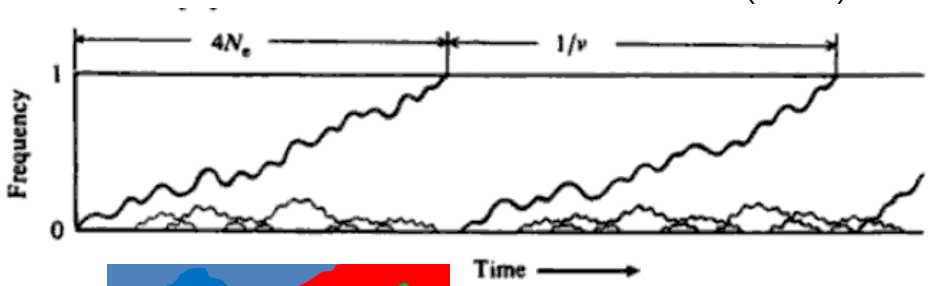
범죄발생 유사도에 의해 미국50개 주를 clustering 한 결과

계통수와 집단유전학과의 관계

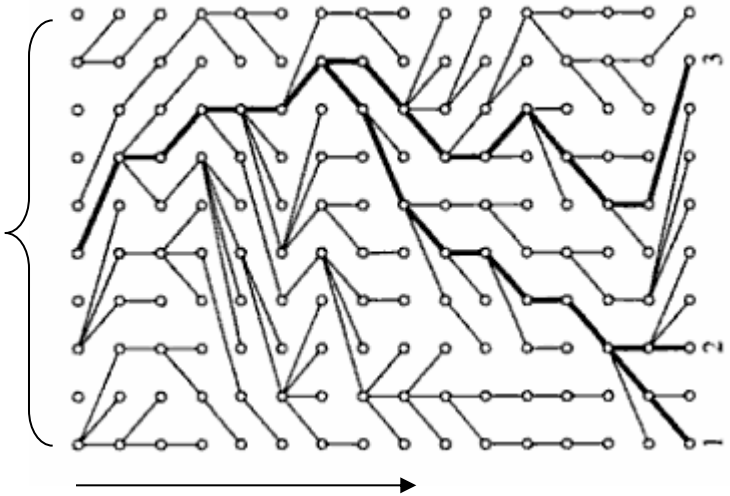


Haeckel (1866)

Kimura (1985)

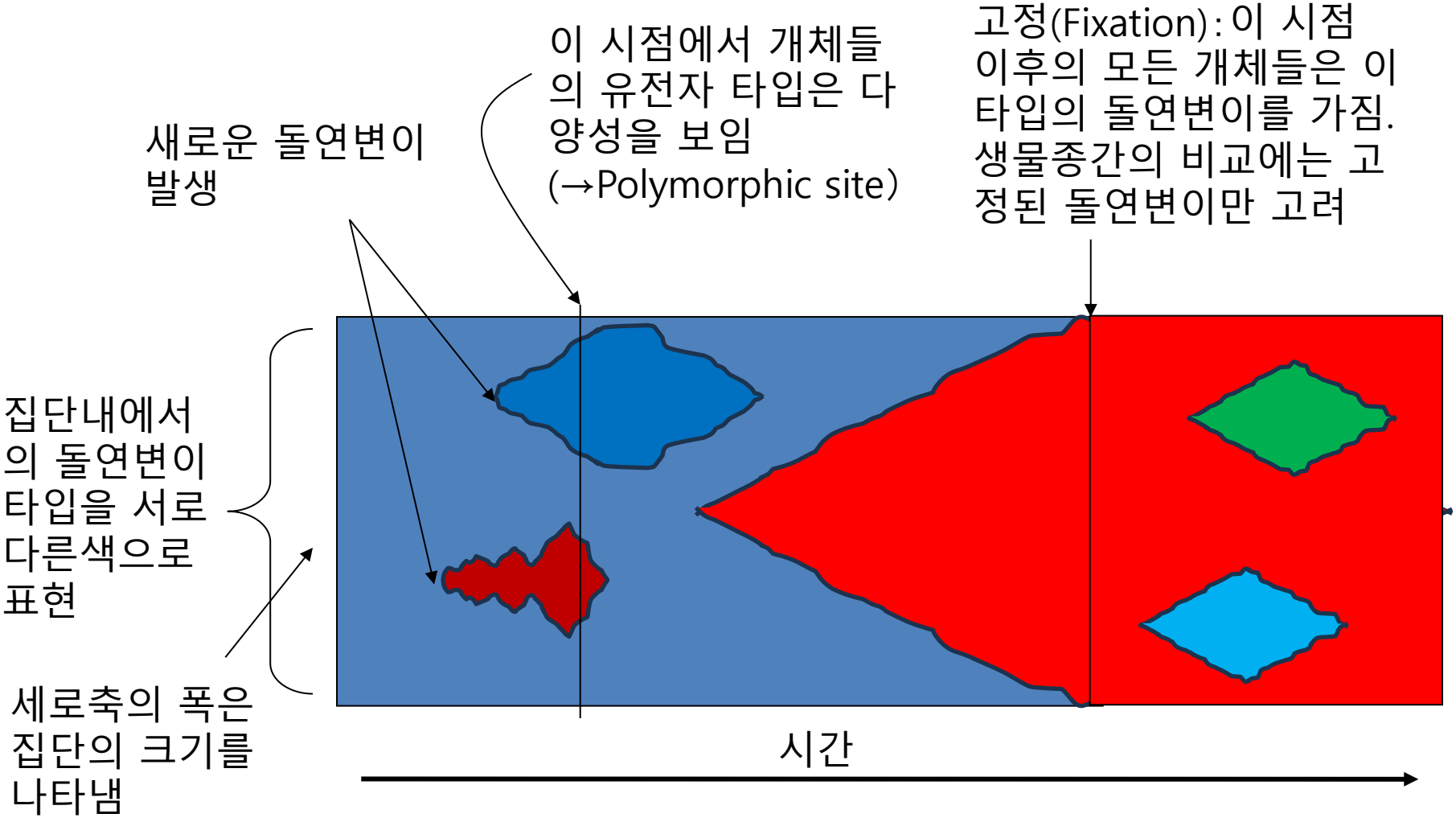


$\theta(=4N\mu)$



Coalescent theory, Hein et al. (2005)

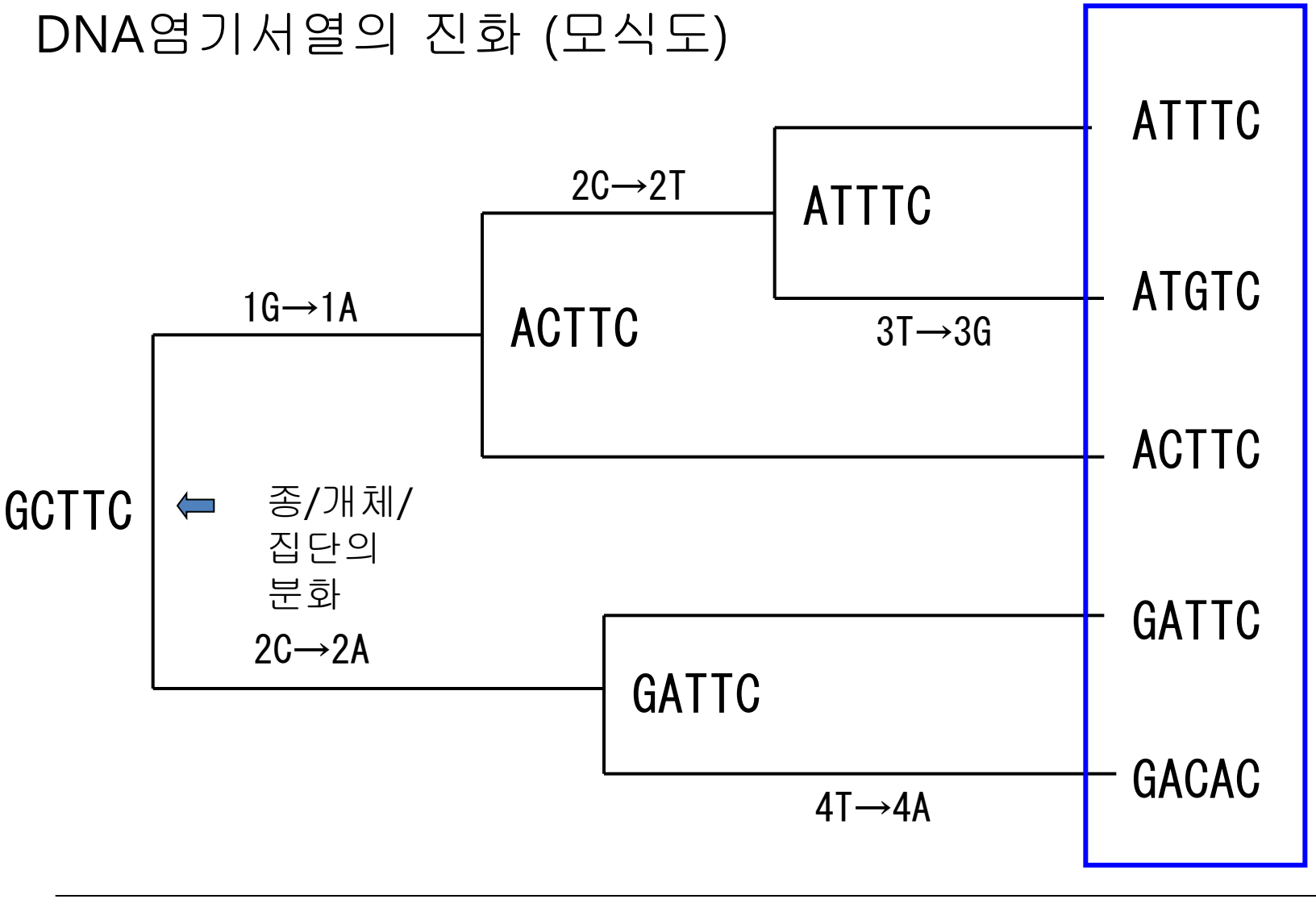
돌연변이의 발생과 집단내에서의 고정



집단의 다양성을 나타내는 척도 $\theta(=4N\mu)$

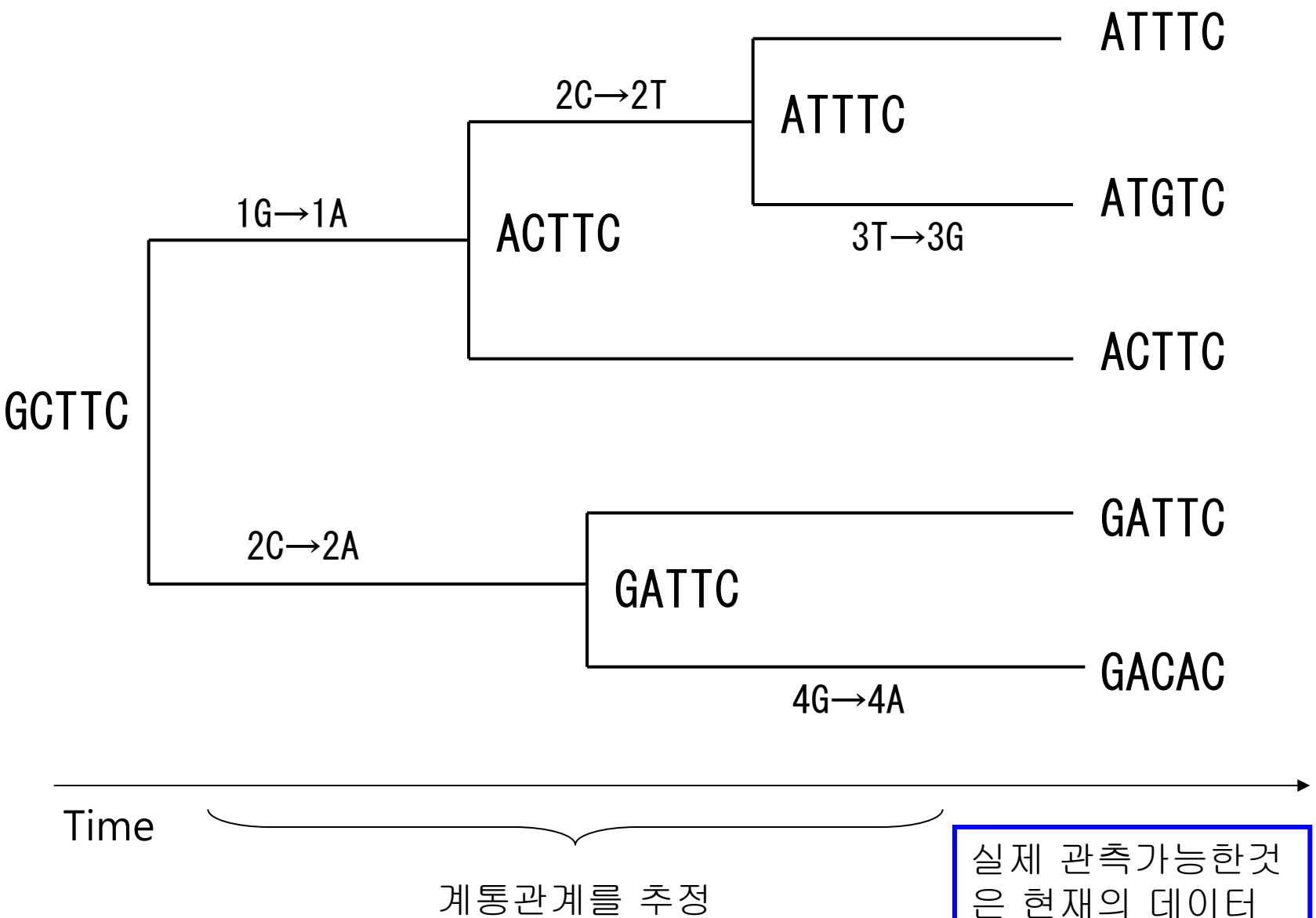
N: effective population size (집단의 크기)
μ: mutation rate per generation (세대당 돌연변이율)

DNA염기서열의 진화 (모식도)

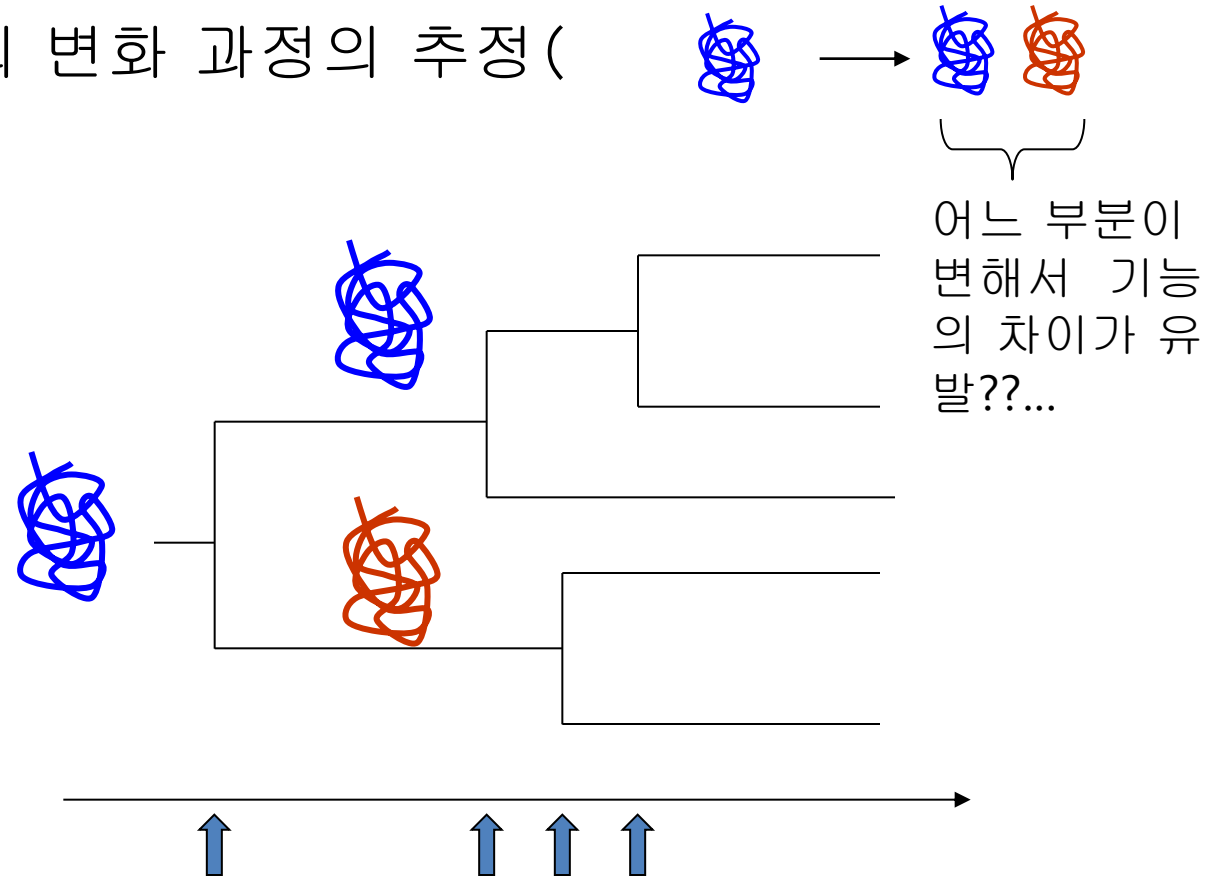


실제 관측가능한것은 현재의 데이터

DNA염기서열의 진화 (모식도)



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



분자계통수로 추정하는 생명의 기원 (HIV의 예)

원인모를 새病 잇따라 엣病

美의료계 非常

사망률 38%의 「면역결핍증」… 輸血감염 추정

젊은 여성들에 「毒性중독증」 발생 79명 사망

한때 자취감췄던 나병·말라리아·뇌염도 늘어

의학 텔렉스

(동아일보 1983년 3월 17일 기사)

- HIV (Human Immunodeficiency Virus)가 AIDS를 유발

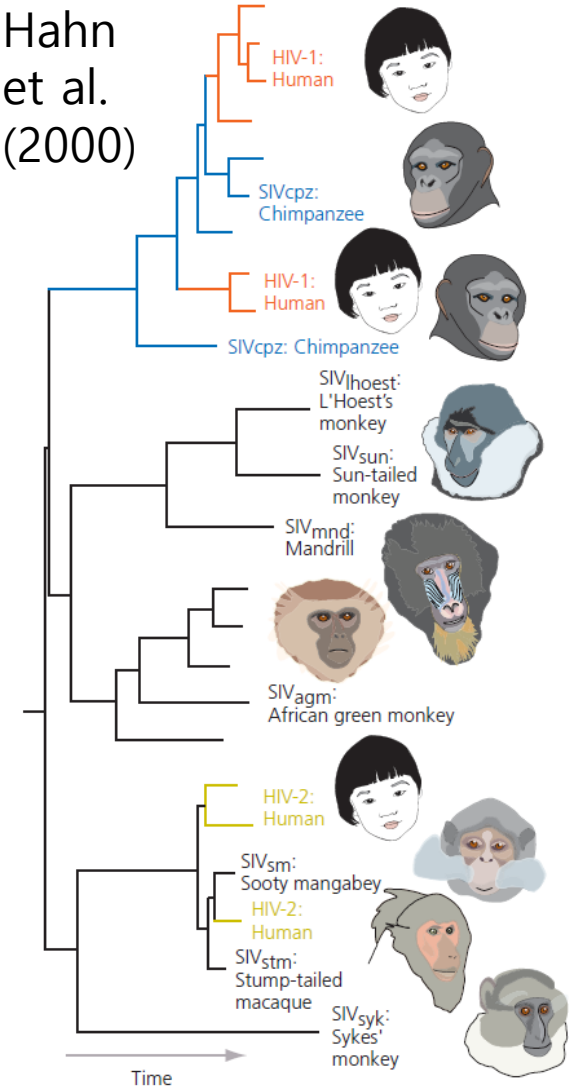
- 20세기 이전에 AIDS (후천성 면역결핍증)는 지구상에 존재하지 않았던 질병

- 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) : 바이러스의 숙주의 영역을 넘어서 다른 숙주를 감염시키게 되는 현상

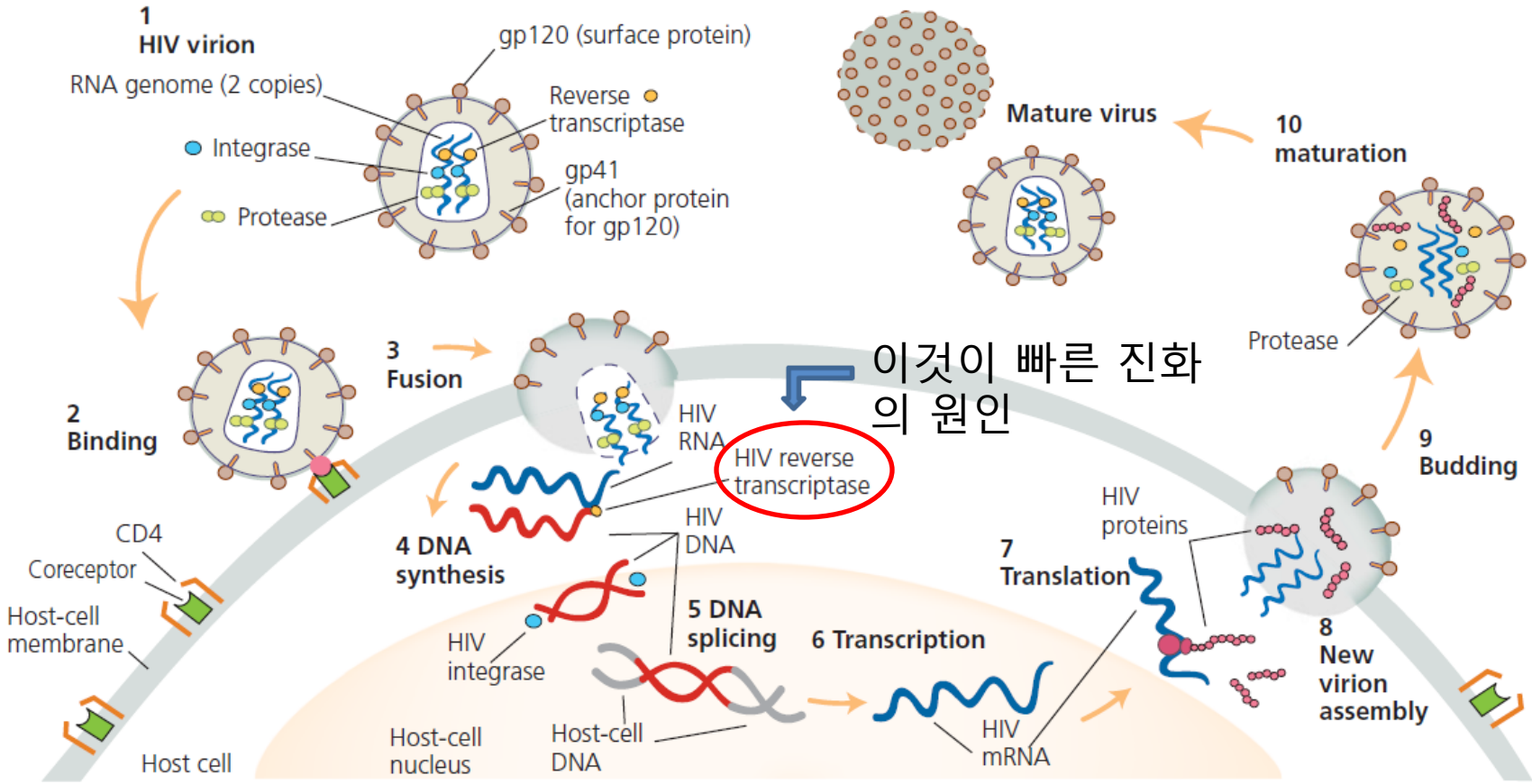
FIV (Feline ~) → 고양이과를 감염시킴

SIV (Simian ~) → 원숭이 부류를 감염시킴

↘ **HIV → 인간에게 감염시킬수 있는 능력획득**

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

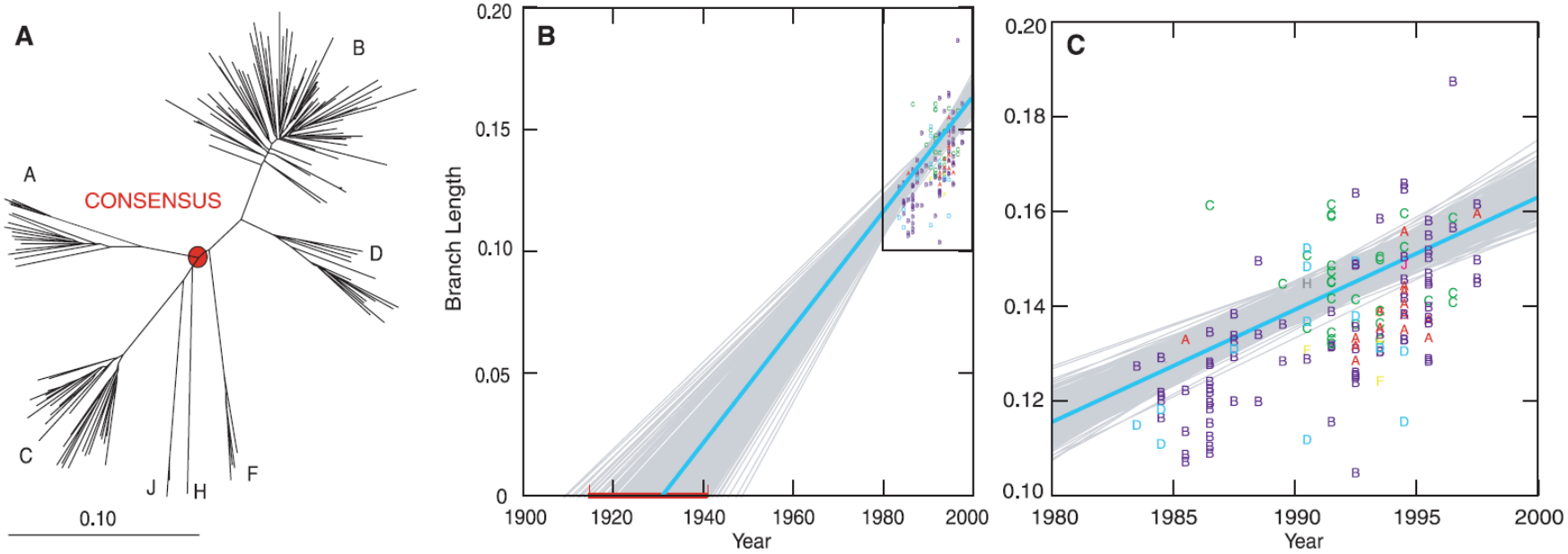
HIV는 왜 빨리 진화하는가?



HIV의 life cycle (Herron & Freeman 2014)

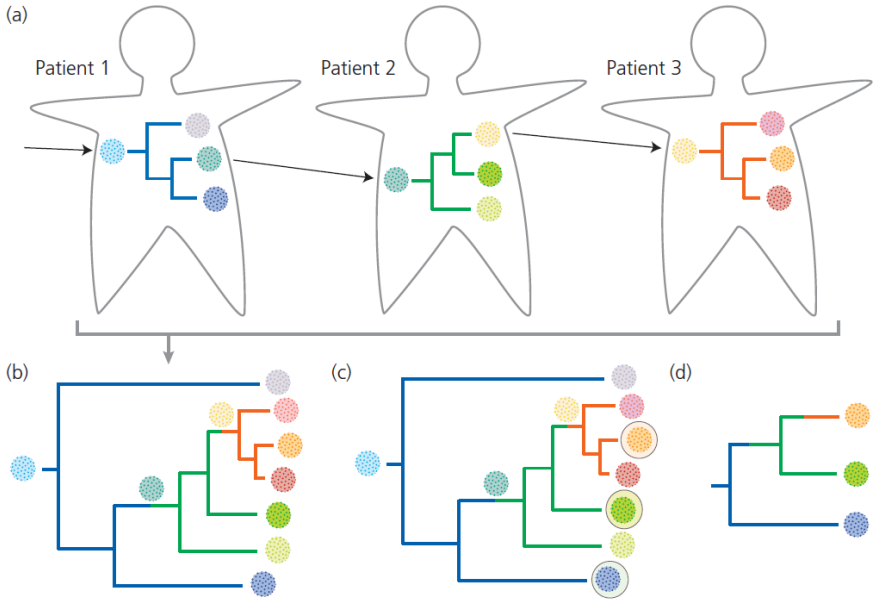
HIV-1 의 최초 등장 시기 추정

(Korber et al. 2000)

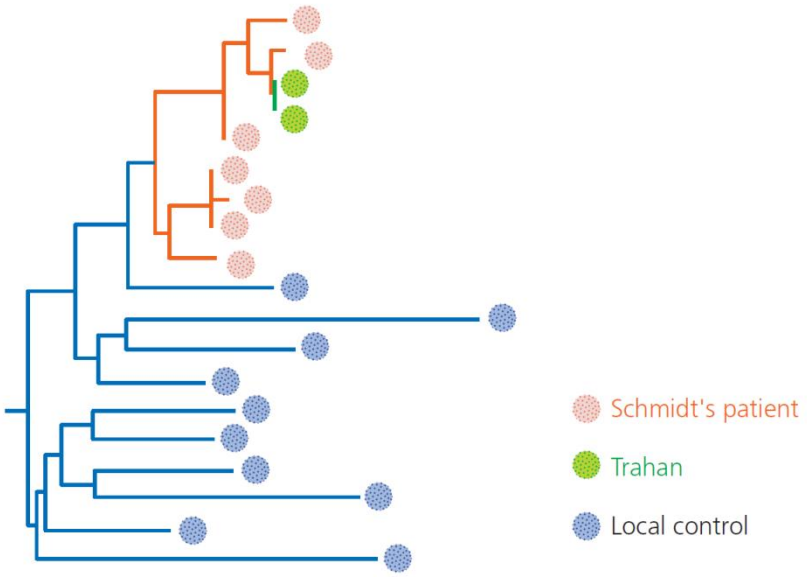


- 매년 DNA 치환량을 측정하여 회귀직선을 구하면 공동조상의 연대를 추정할수 있다 ← 분자시계 (molecular clock) 개념 이용
- HIV가 최초 등장한 시기는 대략 1931년 (1915~1941)
- 이 결과는 OPV (Oral Polio Vaccine) AIDS 가설을 반증하는 증거의 하나가 됨

법의학분야에서도 사용되는 분자계통수



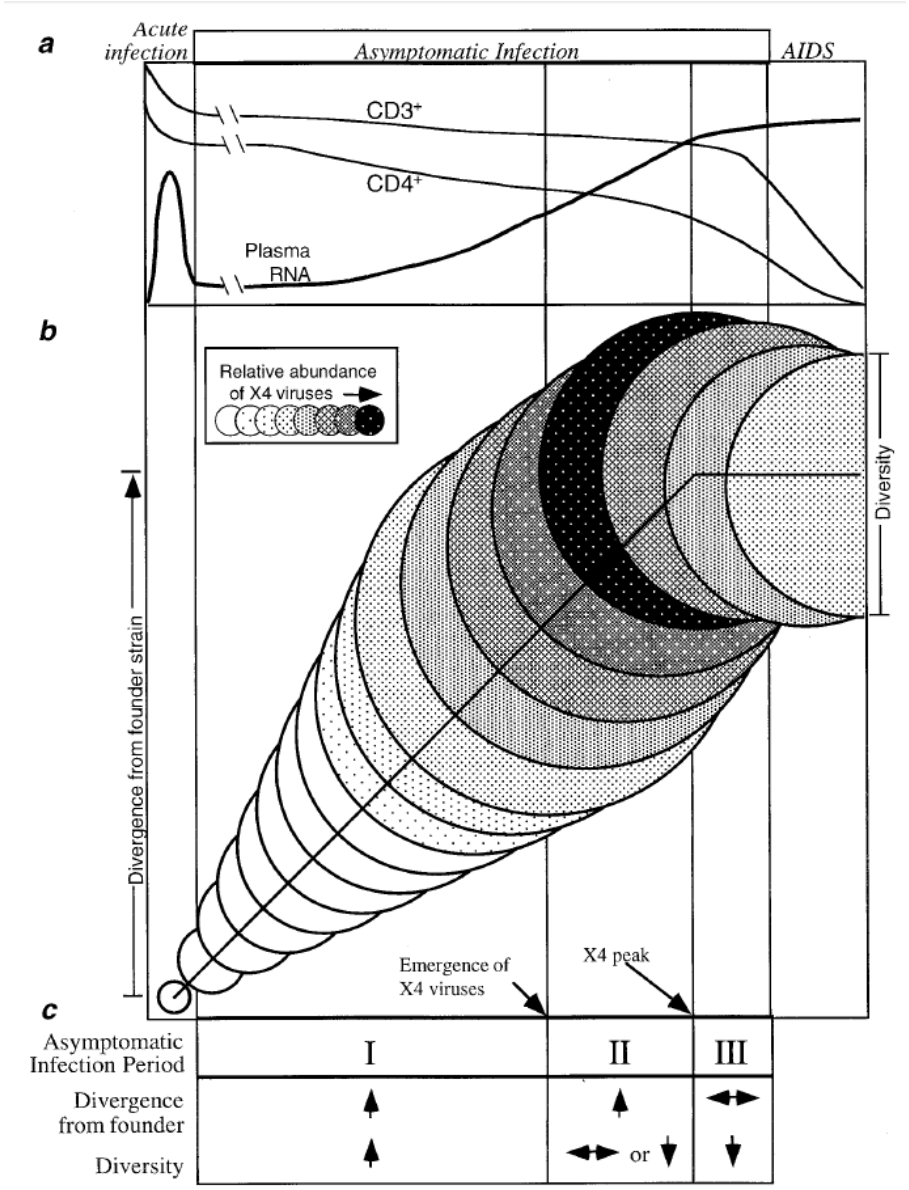
환자에서 환자로 HIV가 전이되는 양상과 각각의 분자계통수

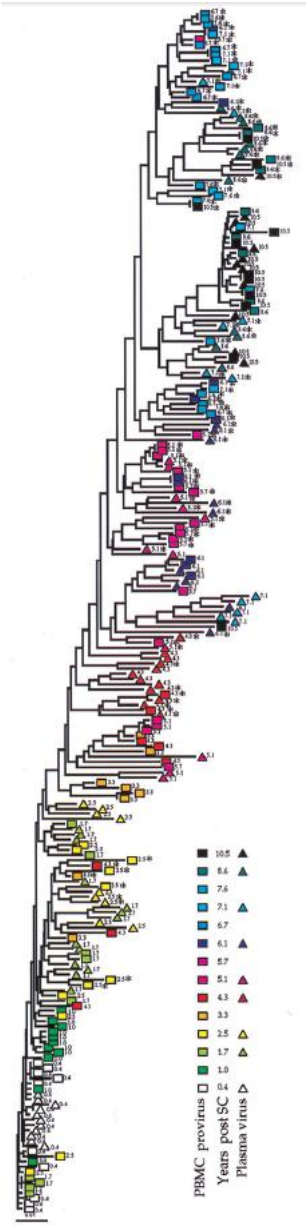


미국 Louisiana 주의 내과의사 Schmidt가 내연관계의 간호사 Trahan에게 고의로 HIV를 주사했다고 의심되는 사건 → 징역 50년 유죄판결 확정 (2000년)

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

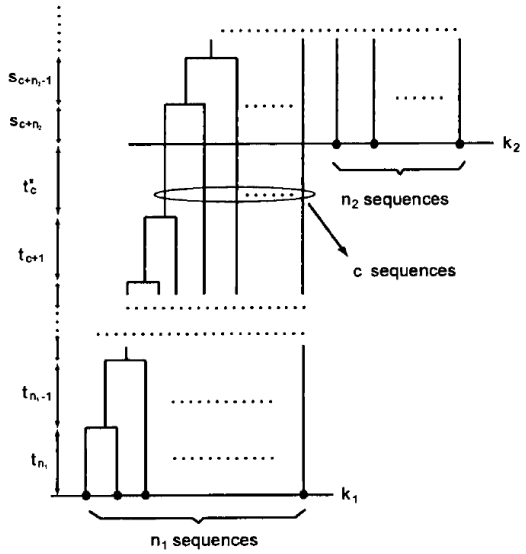
환자의 체내에서 HIV가 진화하는 방식





(Shankarappa et al 1999)

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



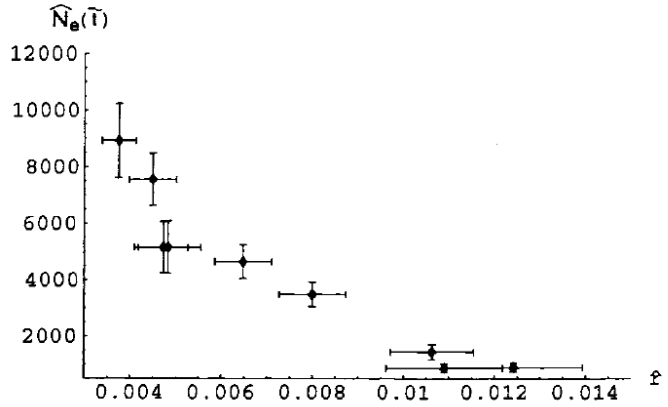
$$L_1 = \left[\prod_{i=n_1}^{c+1} p(t_i|e(t_i),\ N_e)\right] \times p(t_c^*,\ s_{c+n_2}|e(t_c^*),\ k_2,\ N_e)$$
$$\times \left[\prod_{i=c+n_{2-1}}^{2} p(s_i|e(s_i),\ N_e)\right]$$

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

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

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

좌측의 계통수가 얻어지는 가능성을 likelihood 함수로 표현



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

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

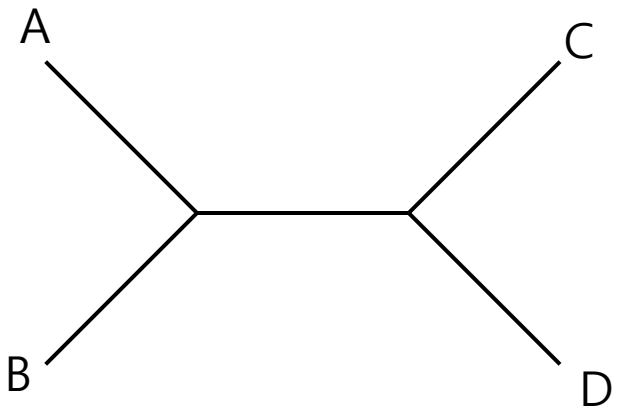
계통수 추정법의 종류

- 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 × likelihood

분자계통수에 관한 기본 사항

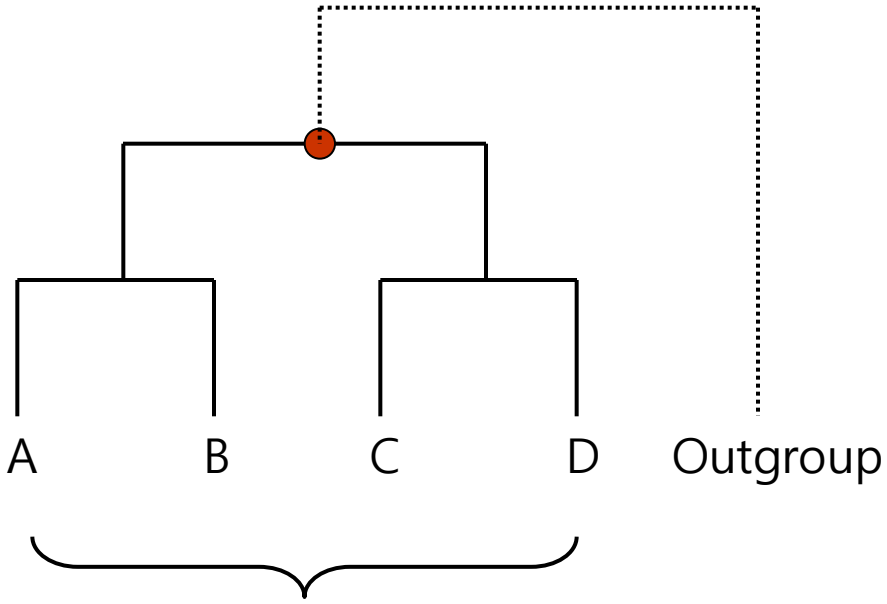
Phylogeny, phylogenetic tree (계통수)

Unrooted tree



`((A,B),C,D);` 혹은 `((C,D),A,B);`

Rooted tree



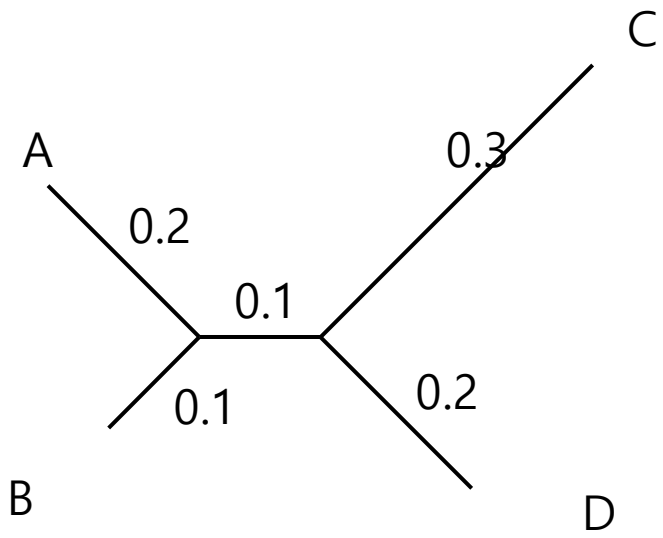
`(((A,B),(C,D)),Outgroup);`

* 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

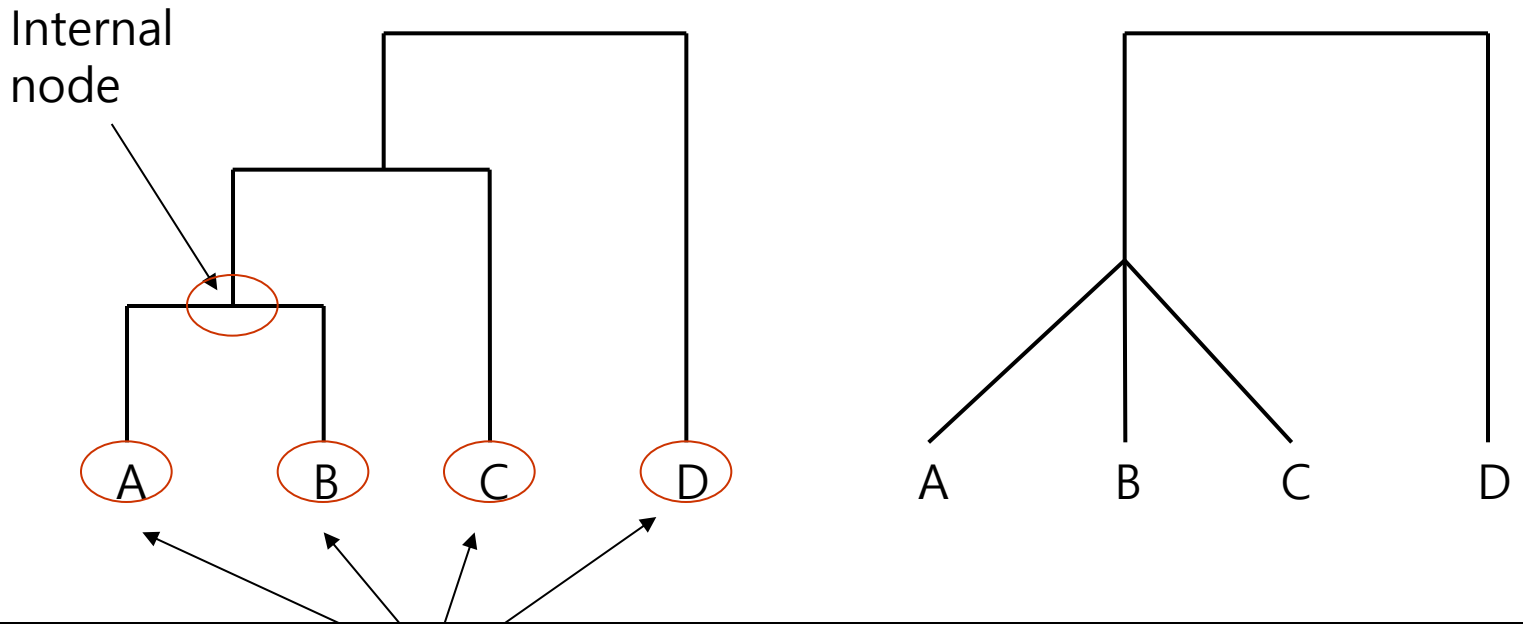


`((A:0.2,B:0.1):0.1,C:0.3,D:0.2);` 혹은
`((C:0.3,D:0.2):0.1,A:0.2,B:0.1);` 로 계통수를 표시함

- 계통수의 형태

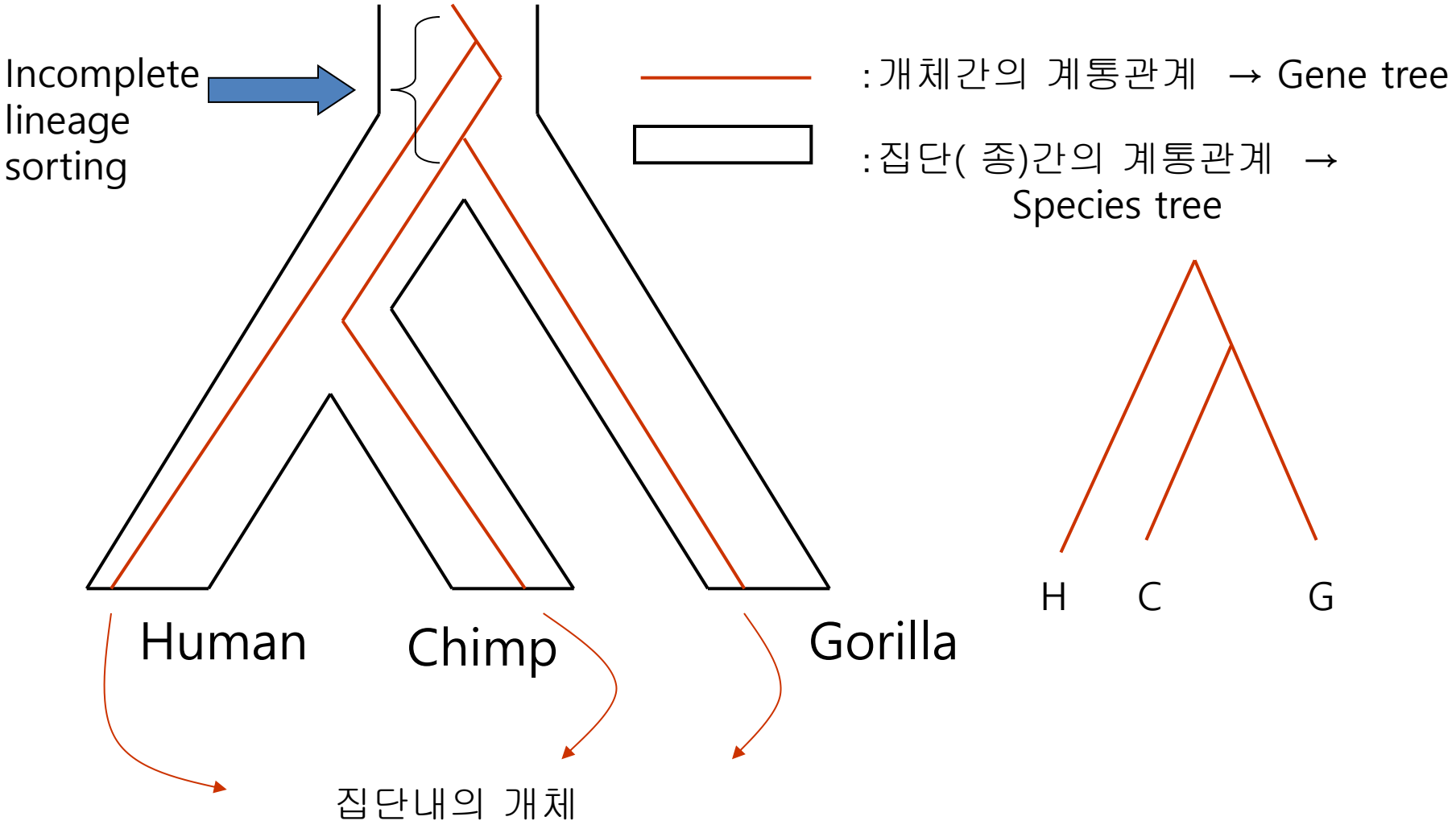
Bifurcating tree

Multifurcating tree

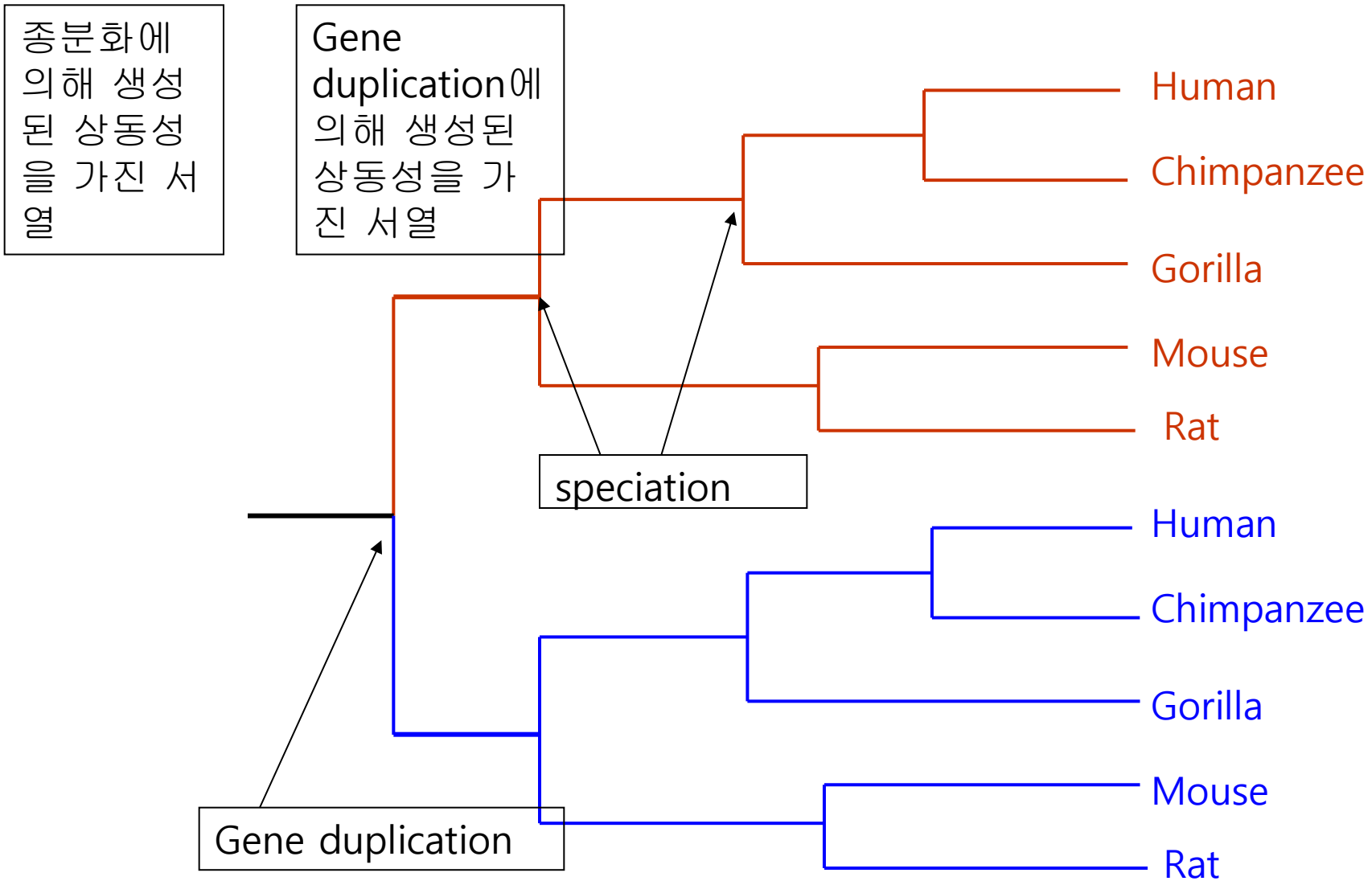


OTU (Operational Taxonomic Unit) , taxon (복수형명사:taxa), terminal node, tip

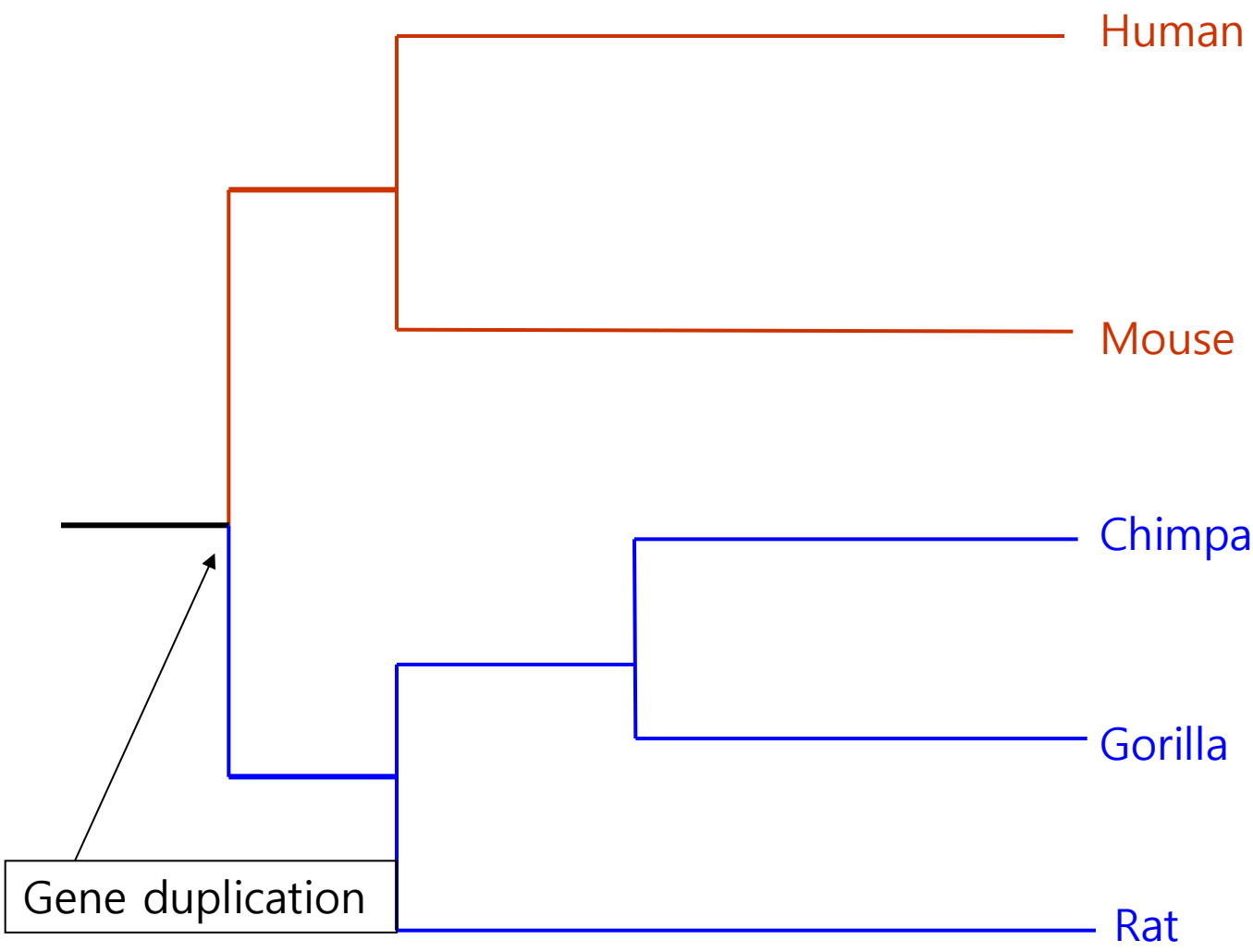
- Gene tree와 species tree는 다른 경우가 있음에 주의



- 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 sites만 고려됨

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

Tree 1

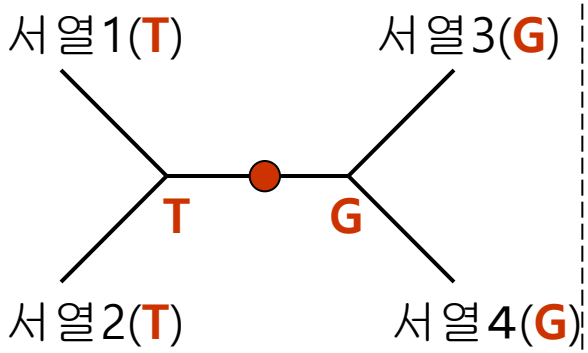
MP tree

Tree 2

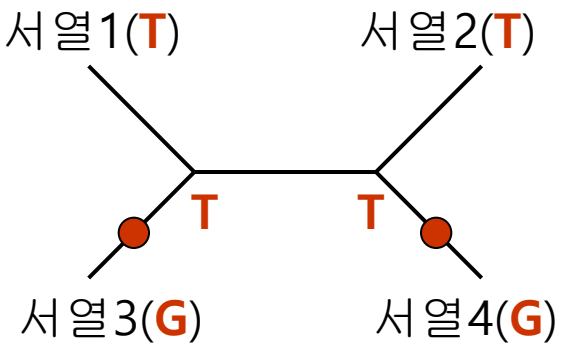
MP tree

Tree 3

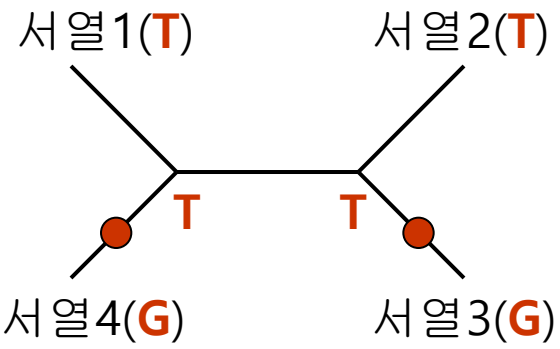
4th site



Score = 1

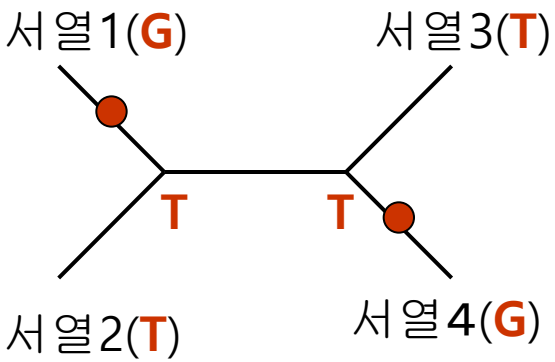


Score = 2

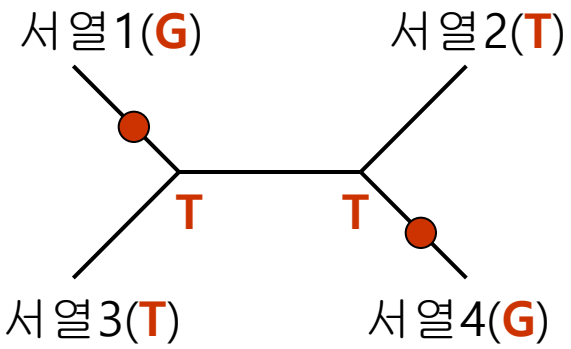


Score = 2

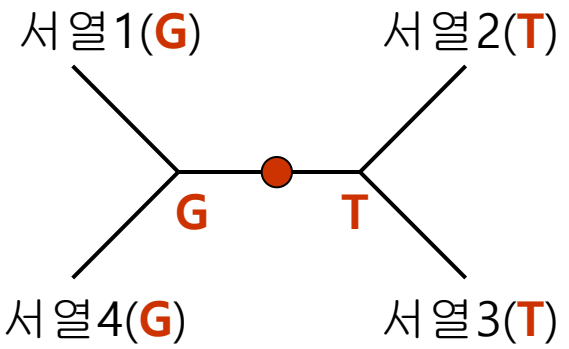
6th site



Score = 2



Score = 2



Score = 1

Score 합: 3

Score 합: 4

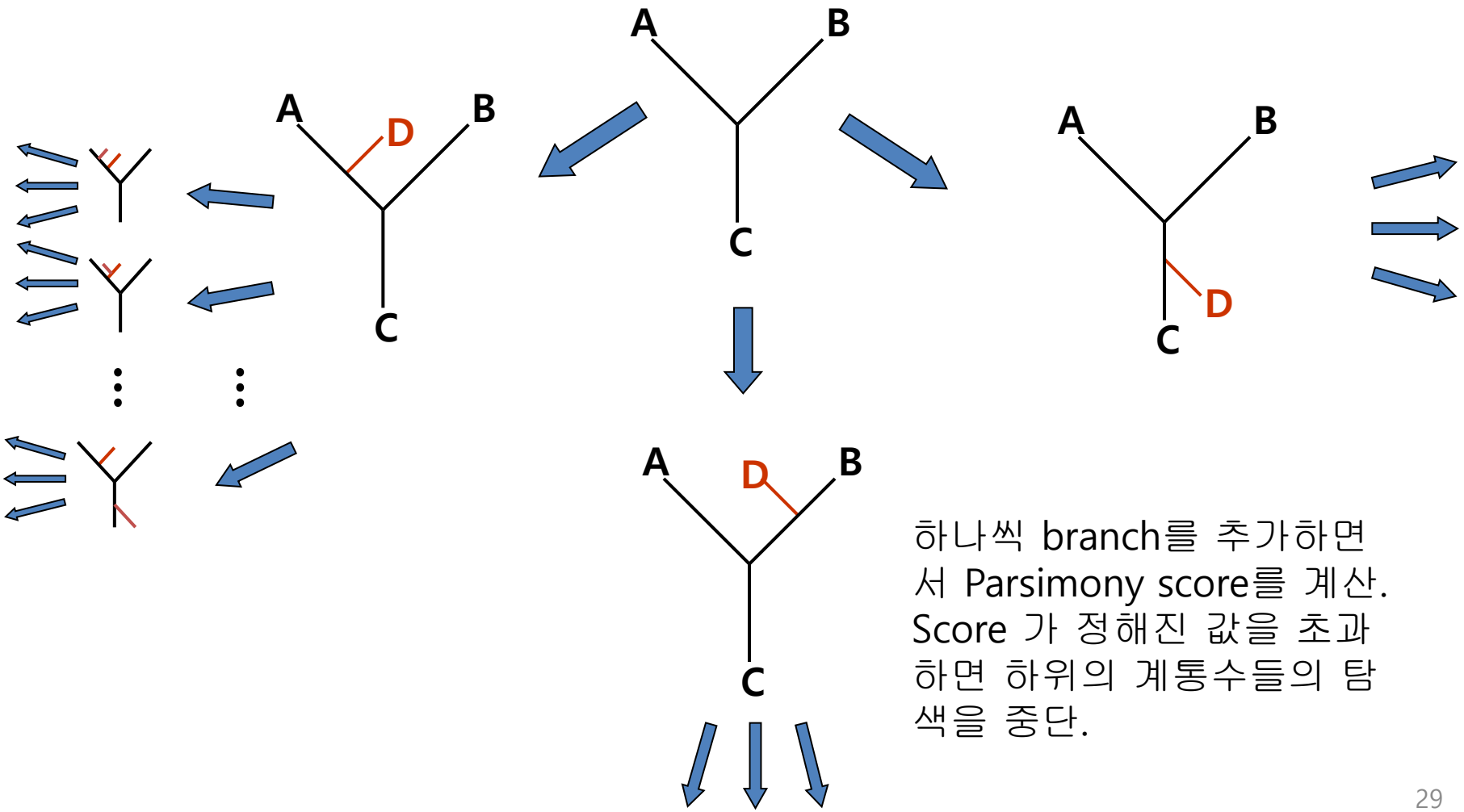
Score 합: 3

- 염기서열의 수가 증가함에 따라 가능한 계통수의 수는 급격히 증가

  → 모든 계통수에 대하여 parsimony score를 계산/비교하는 것은 불가능

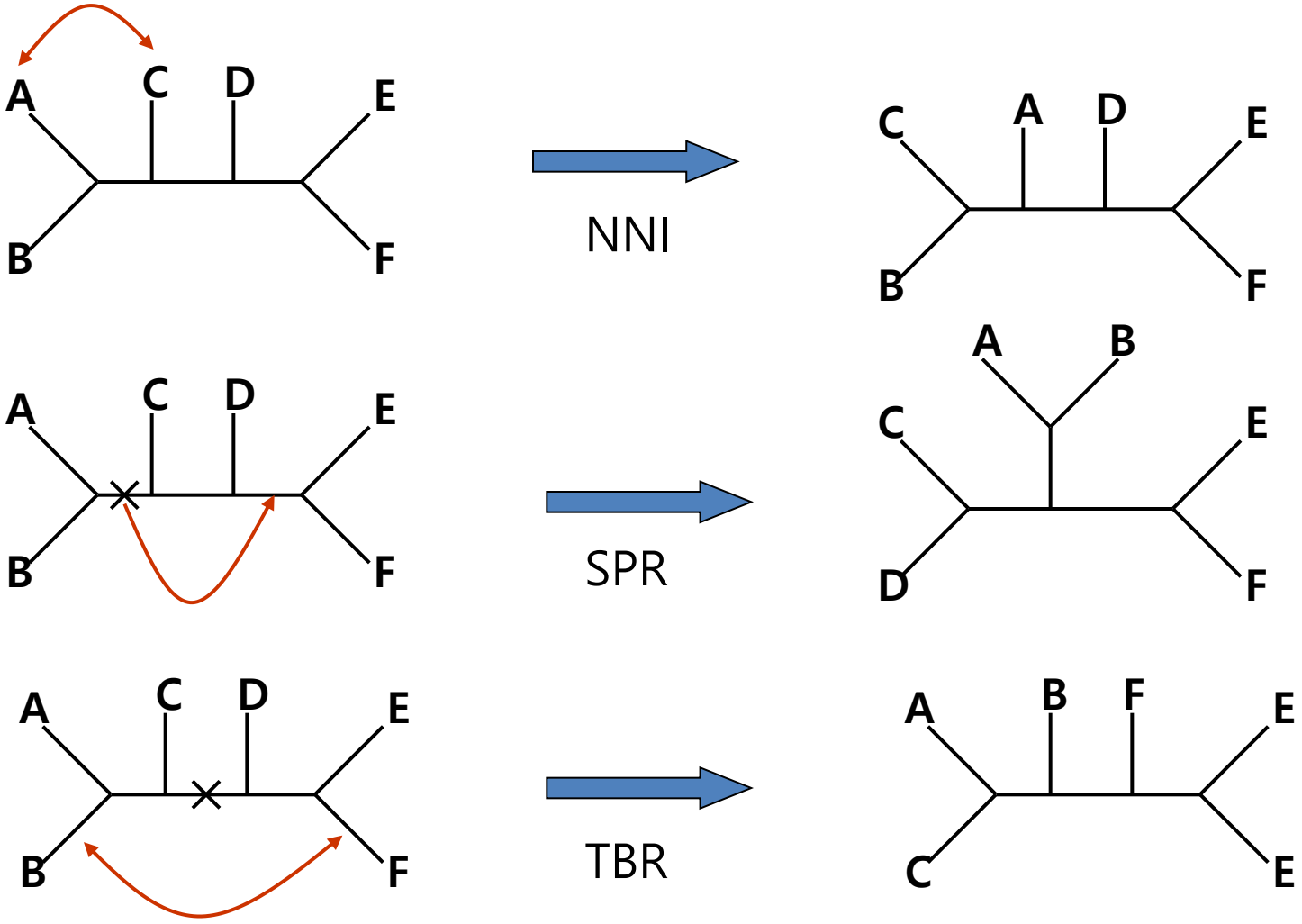
Num of OTU	Num of rooted tree	Num of unrooted tree
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025
11	654,729,075	34,459,425
12	13,749,310,575	654,729,075
13	316,234,143,225	13,749,310,575
14	7,905,853,580,625	316,234,143,225
15	213,458,046,676,875	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 。

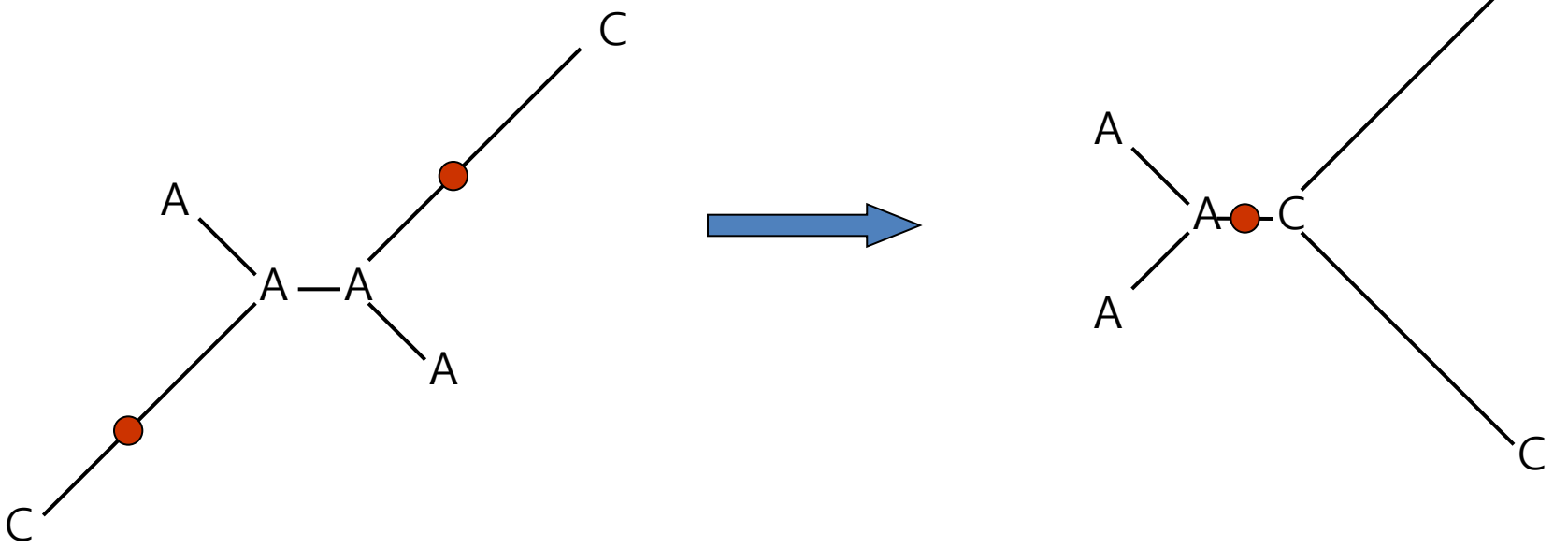


하나씩 branch를 추가하면서 Parsimony score를 계산. Score 가 정해진 값을 초과하면 하위의 계통수들의 탐색을 중단.

- 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

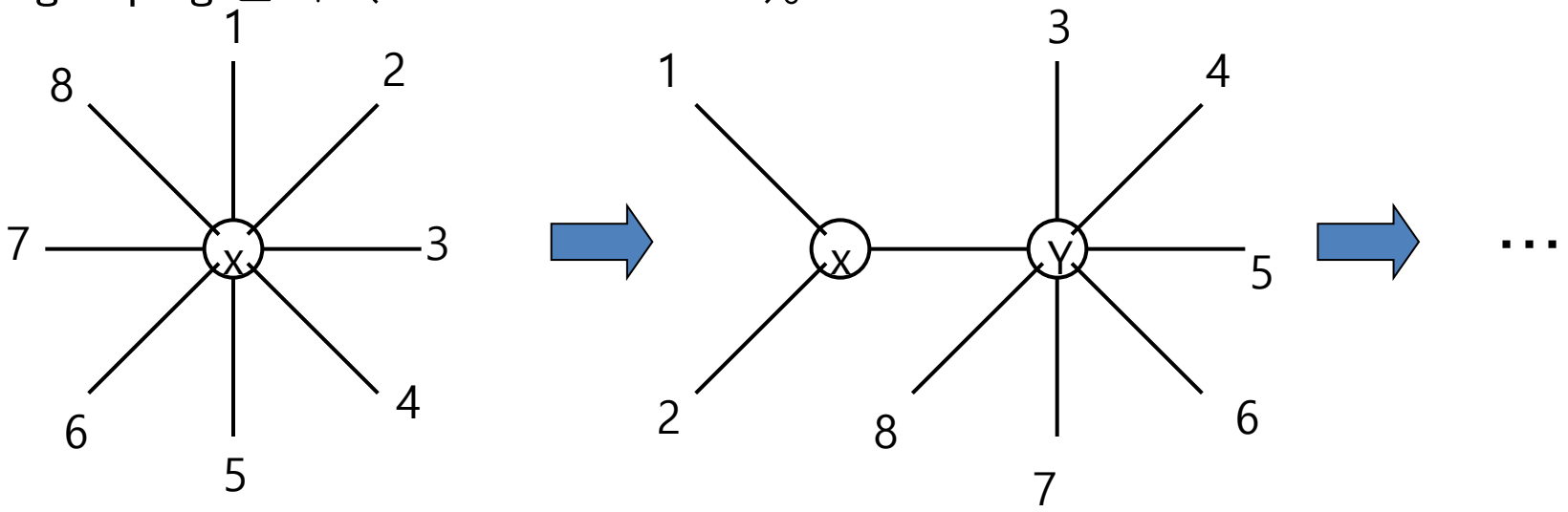


Homoplasy가 존재하는 데이타
(공동조상에서 유래한것이 아니지만 상동성을 가짐)

추정된 계통수

Neighbor-Joining (NJ) method

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



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

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

d_{ij} :서열i와j사이의 진화적인 거리. 사이트당 염기치환수를 나타냄