One of examples for comparing Character-based Method and Distance-based Method
순서가 있는 형질에 있어서의 계통추론을 위한 두 방법의 비교
참조: Rordford et al. (1986)의 "Phenetics vs. Phylogenetics" 에 잘 설명되어 있음 (웹사이트에 올려져 있음)

- When the character is ordered, we can make phylogenetic tree using the following method
  - ordered: 0→ 1→ 2

Character-based method: Maximum Parsimony (MP)

	1	2	3	4	5
A	1	2	2	1	0
B	1	2	1	0	0
C	0	1	0	1	0
D	0	0	0	0	1
ANC	0	0	0	0	0

Character × Taxon Matri

1) Make distance matrix

$$d(X,Y) = \sum_{i=1}^{n} |V_{X_i} - V_{Y_i}|$$

where $d(X,Y)$ = the distance between taxa X and Y
n = the total number of characters
V_{X_i} = the character state value of taxon X for character i
V_{Y_i} = the character state value of taxon Y for character i

$$d(B,C) = |1-0| + |2-1| + |1-0| + |0-1| + |0-0| = 4$$

	A	B	C	D	ANC
A	—	2	4	7	6
B		—	4	5	4
C			—	3	2
D				—	1
ANC					—

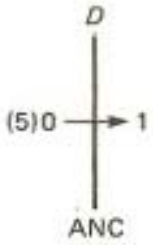
Distance Matrix

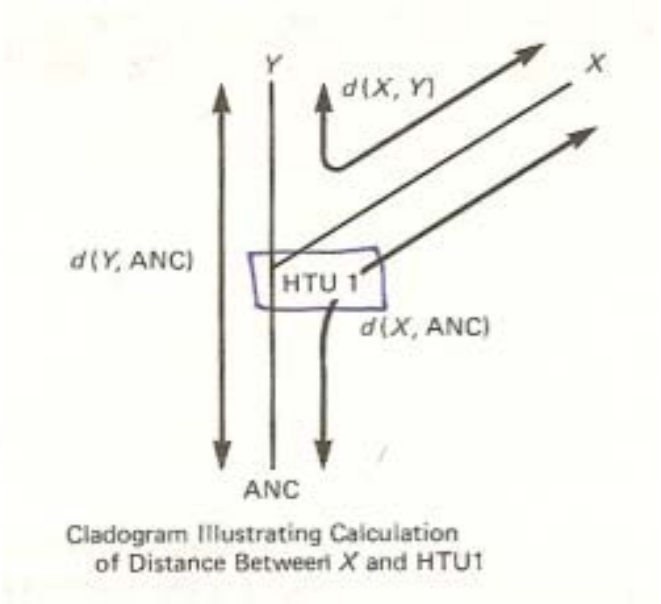
2) Choose minimum distance from ANC

3) Set HTU1 and find minimum-distanced taxa from HTU1

	1	2	3	4	5
A	1	2	2	1	0
B	1	2	1	0	0
C	0	1	0	1	0
D	0	0	0	0	1
ANC	0	0	0	0	0

Character × Taxon Matrix

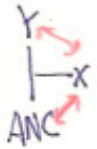




Cladogram Illustrating Calculation of Distance Between *X* and HTU1

$$d(X,\text{HTU1}) = \frac{d(X,Y) + d(X,\text{ANC}) - d(Y,\text{ANC})}{2}$$

where X = an unplaced OTU
HTU1 = the hypothetical ancestor
Y = the first placed OTU
ANC = the original ancestor



For example,

$$d(A,\text{HTU1}) = \frac{d(A,D) + d(A,\text{ANC}) - d(D,\text{ANC})}{2}$$

$$= \frac{7 + 6 - 1}{2} = 6$$

$$d(B,\text{HTU1}) = \frac{d(B,D) + d(B,\text{ANC}) - d(D,\text{ANC})}{2}$$

$$= \frac{5 + 4 - 1}{2} = 4$$

$$d(C,\text{HTU1}) = \frac{d(C,D) + d(C,\text{ANC}) - d(D,\text{ANC})}{2}$$

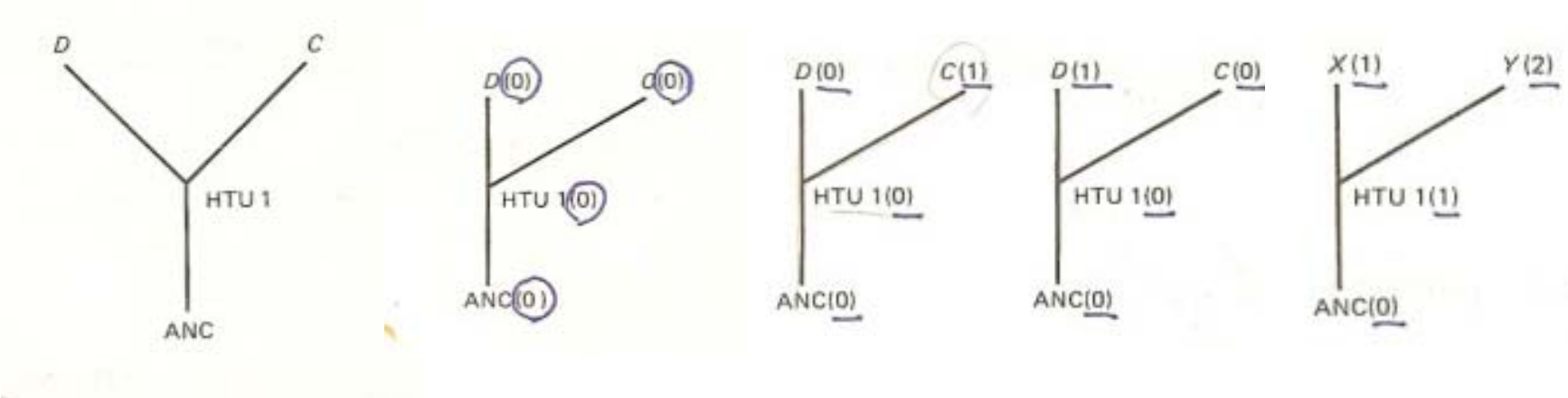
$$= \frac{3 + 2 - 1}{2} = 2$$ shortest

c.f.)

HTU (Hypothetical Taxonomic Unit): 가상분류단위. 계통분석을 할 때 두 OUT들의 분지점에 위치하는 가상적인 분류군을 선정하여 이 분류군의 형질상태를 가정하여 분석에 이용한다.

OTU (Operational Taxonomic Unit): 분석을 위한 기본 단위. OUT는 종이 될 수 도 있고, 속이 될 수도, 개체가 될 수도 있다. 분류학에 있어서 taxon (분류군)과 비슷한 개념인데, OUT는 수리분류학에서 사용하는 용어이다.

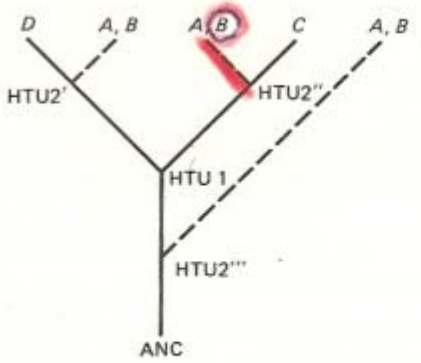
4) Reconstruct character status of HTU1 and add HTU1 in the matrix and distance table



	1	2	3	4	5
A	1	2	2	1	0
B	1	2	1	0	0
C	0	1	0	1	0
D	0	0	0	0	1
ANC	0	0	0	0	0
HTU1	0	0	0	0	0

	A	*B*	*C*	*D*	ANC	HTU1
A	—	2	4	7	6	6
B		—	4	5	4	4
C			—	3	2	2
D				—	1	1
ANC					—	0
HTU1						—

5) Set HTU2 and find minimum distanced taxa again and repeat 3)~4).



	1	2	3	4	5
A	1	2	2	1	0
B	1	2	1	0	0
C	0	1	0	1	0
D	0	0	0	0	1
ANC	0	0	0	0	0

Character × Taxon Matrix

$$d(A,HTU2') = \frac{d(A,D) + d(A,HTU1) - d(D,HTU1)}{2} = 6$$

$$d(B,HTU2') = \frac{d(B,D) + d(B,HTU1) - d(D,HTU1)}{2} = 4$$

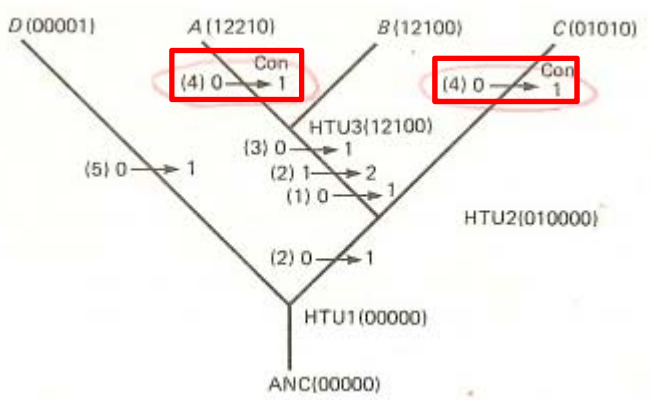
$$d(A,HTU2'') = \frac{d(A,C) + d(A,HTU1) - d(C,HTU1)}{2} = 4$$

$$d(B,HTU2'') = \frac{d(B,C) + d(B,HTU1) - d(C,HTU1)}{2} = 3$$

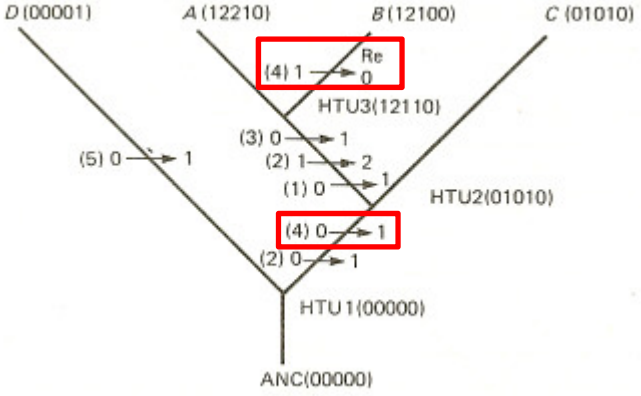
$$d(A,HTU2''') = \frac{d(A,HTU1) + d(A,ANC) - d(ANC,HTU1)}{2} = 6$$

$$d(B,HTU2''') = \frac{d(B,HTU1) + d(B,ANC) - d(ANC,HTU1)}{2} = 4$$

6) Now we got a tree and we may reconstruct character evolution. There are two different optimization method:



DELTRAN (DElayed TRANsformation) optimization : character changes are placed at terminals in the tree



ACCTRAN (ACCelerated TRANsformation) optimization:
character changes are placed at the base rather than terminals.

Distance-based Method: neighbor joining (NJ), UPGMA…

fast and easy

	1	2	3	4	5
A	1	2	2	1	0
B	1	2	1	0	0
C	0	1	0	1	0
D	0	0	0	0	1

1) Calculate coefficient of overall similarity whatever the method is.

2. Calculate the similarity between each pair of taxa by means of the *coefficient of overall similarity* (s_{jk}). The coefficient of overall similarity is variously defined (see Sneath and Sokal, 1973), one of the more commonly used formulas being:

$$s_{jk} = 1 - \frac{\sum_{i=1}^{n} \frac{|x_{ij} - x_{ik}|}{R_i}}{n}$$

where x_{ij} = the ith character state of taxon j
x_{ik} = the ith character state of taxon k
R_i = the range of the ith character, the highest minus the lowest values
n = the number of characters

For example, for the above character × taxon matrix:

$$s_{AB} = 1 - \frac{\frac{0}{1} + \frac{0}{2} + \frac{1}{2} + \frac{1}{1} + \frac{0}{1}}{5} = .7$$

$$s_{AC} = 1 - \frac{\frac{1}{1} + \frac{1}{2} + \frac{2}{2} + \frac{0}{1} + \frac{0}{1}}{5} = .5$$

$$s_{AD} = 1 - \frac{\frac{1}{1} + \frac{2}{2} + \frac{2}{2} + \frac{1}{1} + \frac{1}{1}}{5} = 0$$

$$s_{BC} = 1 - \frac{\frac{1}{1} + \frac{1}{2} + \frac{1}{2} + \frac{1}{1} + \frac{0}{1}}{5} = .4$$

$$s_{BD} = 1 - \frac{\frac{1}{1} + \frac{2}{2} + \frac{1}{2} + \frac{0}{1} + \frac{1}{1}}{5} = .3$$

$$s_{CD} = 1 - \frac{\frac{0}{1} + \frac{1}{2} + \frac{0}{2} + \frac{1}{1} + \frac{1}{1}}{5} = .5$$

2) Choose the most similar pair and link them

	A	B	C	D
A	1	.7	.5	0
B		1	.4	.3
C			1	.5
D				1

3) Combine two taxa and named “P”.

UPGMA

	A	B	C	D
A	1	.7	.5	0
B		1	.4	.3
C			1	.5
D				1

	P	C	D
P	1		
C		1	.5
D			1

$$s_{Pm} = \frac{N_j s_{jm} + N_k s_{km}}{N_j + N_k}$$

where
- P = new taxon formed from j and k
- m = all remaining taxa
- s_{jm} = similarity coefficient between taxon j and taxon m (from original similarity matrix)
- N_j = the number of original taxa contained in the jth taxon (by creation of new taxa)
- s_{km} and N_k = as above, but for taxon k

4) Calculate similarity between P and others

$$s_{PC} = \frac{(1).5 + (1).4}{1 + 1} = \frac{.9}{2} = .45$$

$$s_{PD} = \frac{(1)0 + (1).3}{1 + 1} = \frac{.3}{2} = .15$$

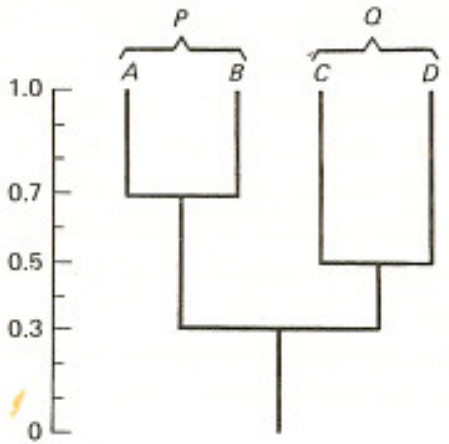
	P	C	D
P	1	.45	.15
C		1	.50
D			1

6) Repeat 2), 3) 4).

	P	Q
P	1	
Q		1

iii. Calculate s_{QP}

$$s_{QP} = \frac{(1).45 + (1).15}{1 + 1} = \frac{.6}{2} = .3$$



※ 같은 matrix를 갖고, character-based method (parsimony)와 distance-based method (UPGMA)로 계통수를 작성하였는데, 같은 matrix임에도 불구하고 다른 두 방법으로부터 도출된 계통수는 서로 다르다.

→ 분석의 성질을 잘 파악하고 이에 맞는 분석방법의 선택이 필요

Consensus Tree (종합수)

- 같은 가능성의 다른 위상을 갖는 여러 개의 계통수를 갖고 있을 때 이들의 정보를 종합할 필요가 있다. 여러 가지 다른 위상의 계통수를 종합한 계통수를 consensus tree (종합수) 라고 한다. 특히 parsimony 분석에 있어서 분석의 결과 흔히 여러 개의 equally parsimonious tree들이 생성되는데, consensus tree를 만듦으로서 전체 분석을 종합할 수 있다.

- Strict consensus tree: tree들의 topology들이 서로 conflict (상충) 하는 경우 polytomy (다분)를 만들어 줌.
- 50% majority rule consensus tree: 서로 다른 topology들 중 하나를 선택함에 있어서 다수결의 원리를 따른다.

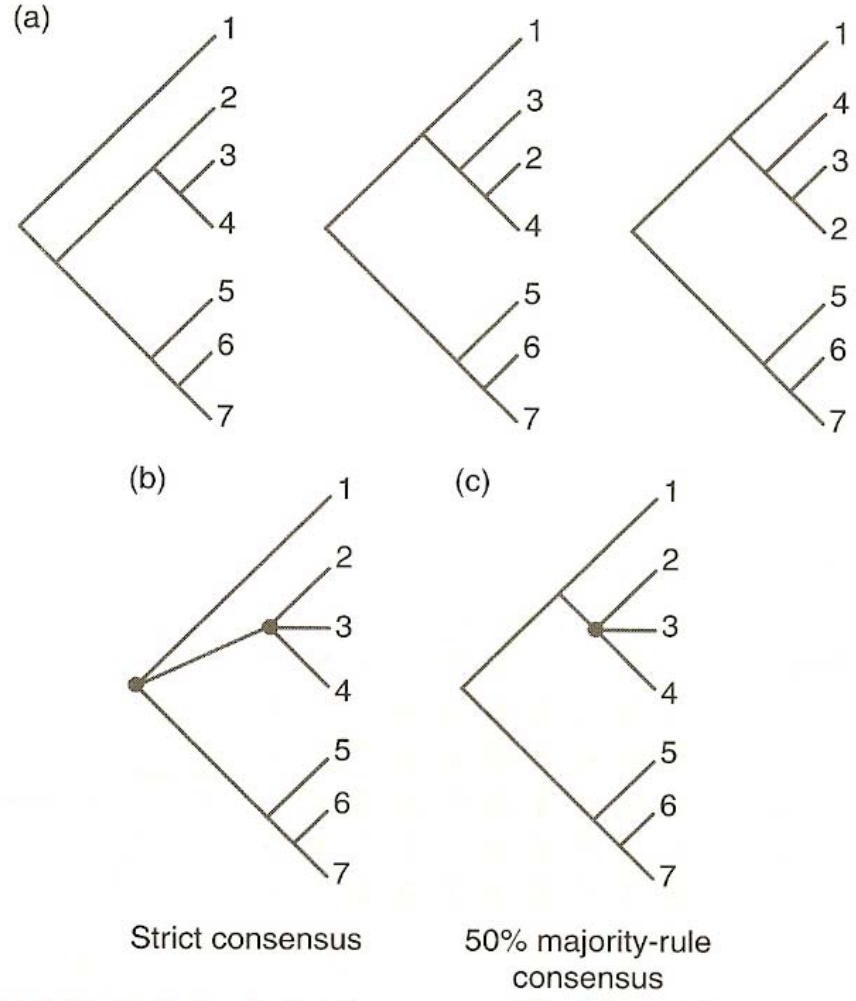
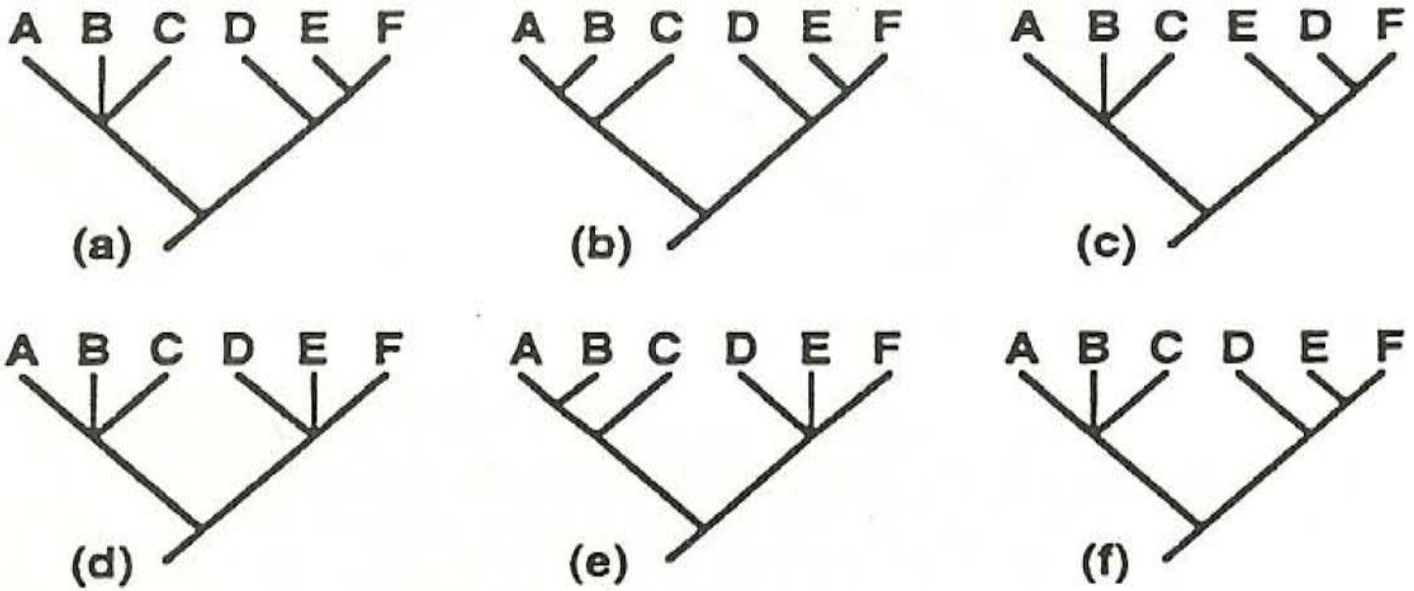


FIGURE 5.6 *Three trees (a) inferred from a single data set can be summarized in a single consensus tree using either (b) strict consensus criteria or (c) 50% majority-rule consensus criteria.*



Three rival trees (a,b,c) and their strict (d), semistrict(e) and 50% majority rule (f) consensus trees.

- semistrict consensus tree: 분해된 topology를 갖는 tree와 polytomy를 비교 시에는 "정보"를 갖는 분해된 topology를 따른다.

Tree confidence
1) Bootstrapping
2) Jackknifing
3) Decay analysis

Tree confidence (계통수내의 절들의 신뢰도): Tree의 clade들이 얼마나 정확한가를 보여주는 수치들

- 계통분석을 하여 하나의 tree를 얻어내는 것이 분석의 끝이 아니라 얻어진 tree내부의 각각의 clade가 얼만큼 신뢰할 수 있는 clade 인가를 인식하여야 한다. 신뢰도가 낮은 clade는 아무런 의미가 없다.
- 일반적으로 clade의 신뢰도는 bootstrap 법으로 계산한다.
- Bootstrap법은 통계학적 기법으로 원자료를 일정 법칙에 따라 변형시킨 pseudo-matrix를 만들어서 이를 분석하여 원자료에 의한 결과와 같은 clade를 형성하는지 확인하는 것이다. 이 때 반복적으로 또 다른 pseudo-matrix를 만들어 이들을 분석하여 특정 clade가 전체 pseudo-matrix들 중 몇%에서 여전히 존재하는지를 나타낸 것이다.

- Pseudo-matrix를 만드는 방법은 예를 들어 10개의 형질이 있다면 형질 번호인 10번까지의 숫자를 무작위로 추첨하여 (반복추첨 가능) 추첨된 10개의 번호에 대한 형질들을 모아 matrix를 만듬

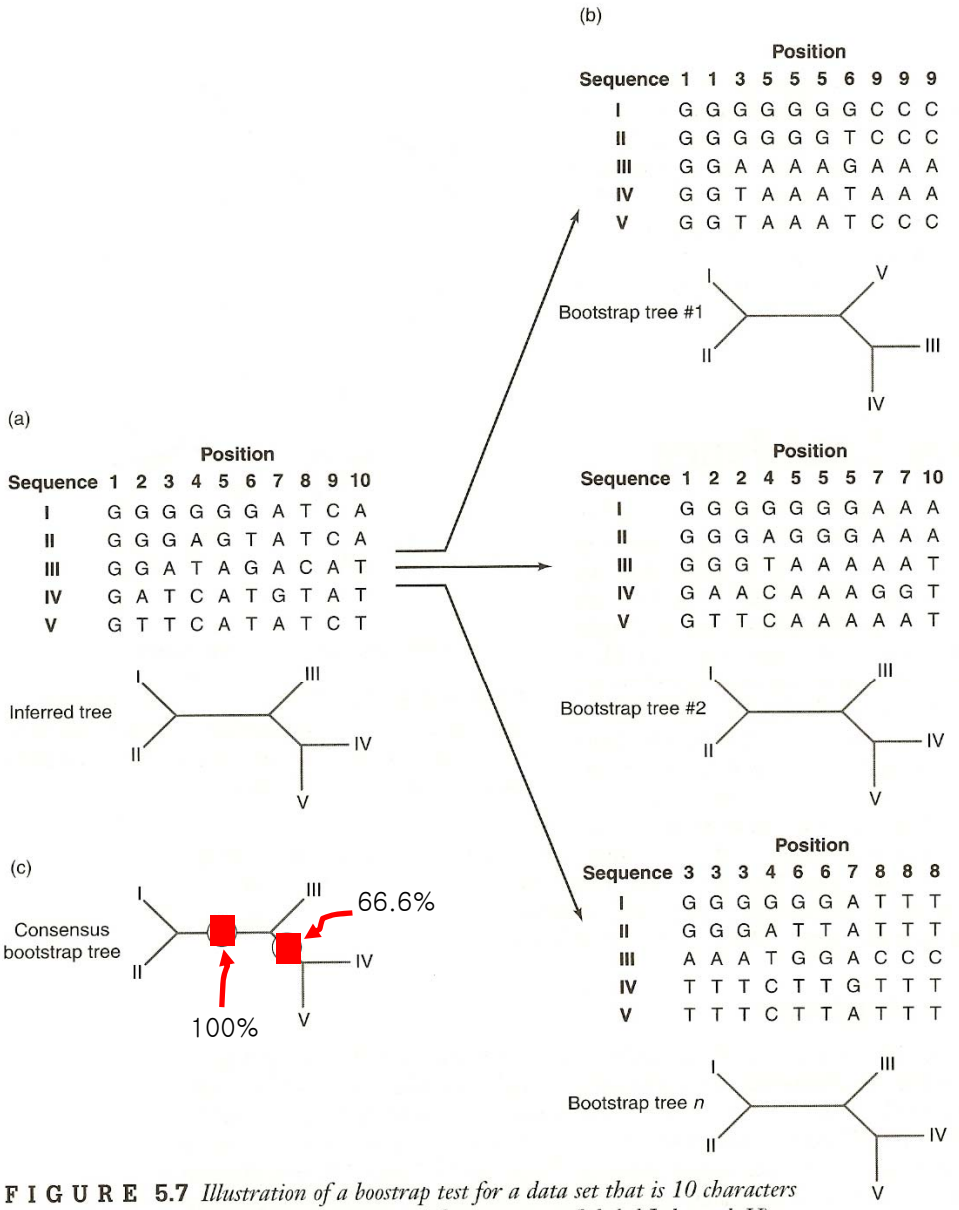


FIGURE 5.7 *Illustration of a boostrap test for a data set that is 10 characters long (positions are labeled 1 through 10) and five sequences (labeled I through V) deep. (a) The original data set and its most parsimonious tree is also shown. A random number generator chooses 10 times from the original 10 positions and creates the three resampled data sets with their corresponding trees (b). A consensus tree (c) for the three resampled data sets is shown with circled values indicating the fraction of bootstrapped trees that support the clustering of sequences I and II and sequences IV and V.*

Tree confidence
1) Bootstrapping
2) Jackknifing
3) Decay analysis

Jackknifing 분석은 bootstrapping 과 동일한 개념으로 수행되는데, pseudo-matrix 대신 전체 matrix 의 일부 (예를 들어 80%)만을 취한 sub-matrix를 만드는 것이다.

Original matrix

	1	2	3	4	5	6
A	T	C	A	C	C	T
B	T	A	A	C	C	G
C	T	A	A	T	C	G
D	T	A	A	T	A	G
E	T	A	T	T	A	G

Jackknifing

Submatrix 1

	1	2	3	4
A	T	C	A	C
B	T	A	A	C
C	T	A	A	T
D	T	A	A	T
E	T	A	T	T

Submatrix 2

	1	3	4	6
A	T	A	C	T
B	T	A	C	G
C	T	A	T	G
D	T	A	T	G
E	T	T	T	G

Submatrix 3 …

	3	4	5	6
A	A	C	C	T
B	A	C	C	G
C	A	T	C	G
D	A	T	A	G
E	T	T	A	G

…

↓ ↓ ↓

Tree 1 Tree 2 Tree 3

각각을 분석 후 original matrix에 의한 결과와 비교하여 여전히 지지되는 clade들의 %를 구함

Tree confidence
1) Bootstrapping
2) Jackknifing
3) Decay analysis

Decay analysis

D1

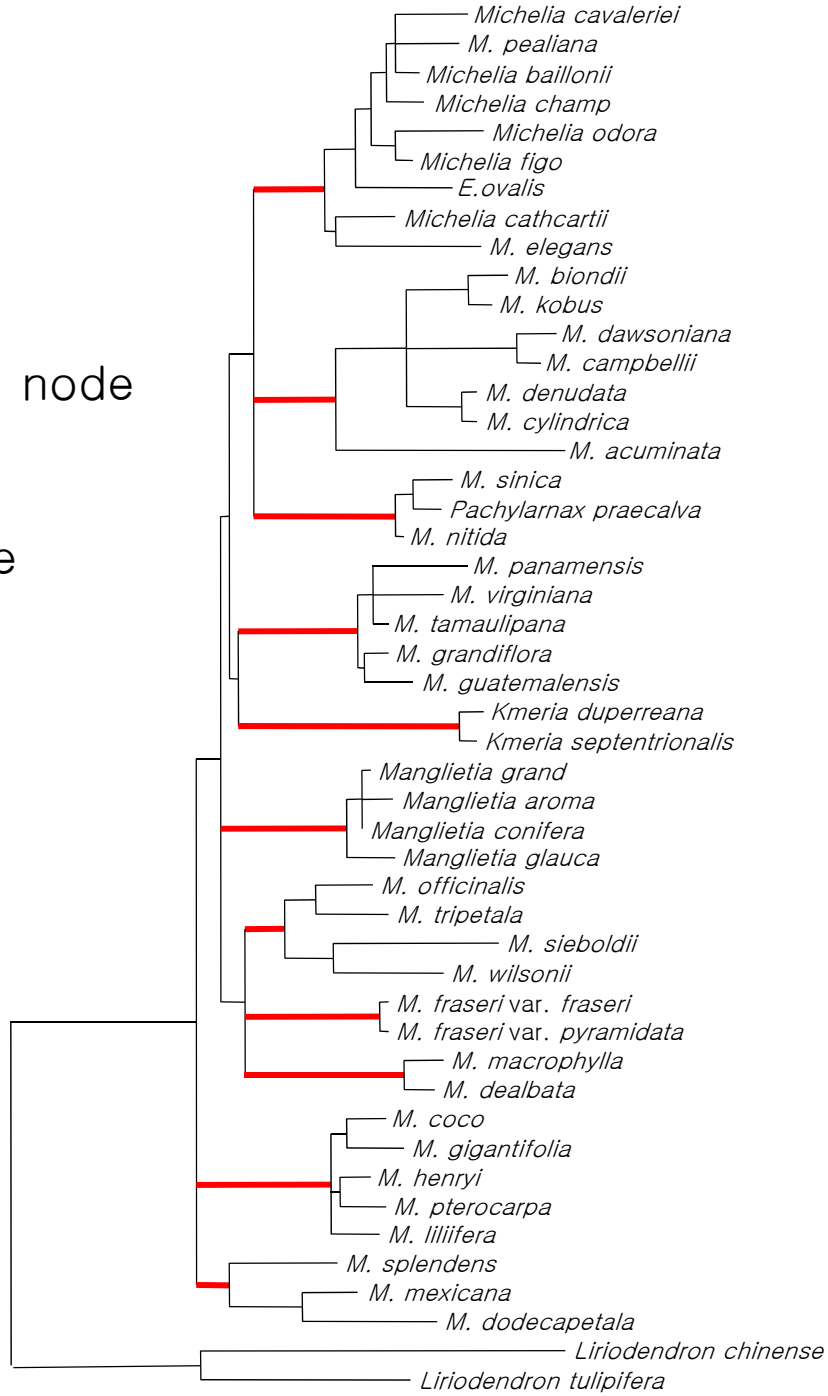
↑

Recognize “collepsed” node

↑

256 steps: 1 tree
257 steps: 5 trees ⟶ Strict consensus tree
258 steps: 83 trees ⟶ … D2
259 steps: 345 trees ⟶ … D3
… …

Parsimony 분석 시 256 step의 tree 가 1개 이고, 257 step의 tree가 5개, 258 step의 tree가 83개… 등으로 나타날 때, 257 step의 tree 5개에 대한 strict consensus tree를 만들어 보면 가장 좋은 tree인 256 step의 tree 에 비하여 node들이 polytomy를 형성할 때 (collepse 됨) 이를 D1 으로 정의한다. 마찬가지로 D2, D3… 등으로 clade들의 신뢰도를 나타내게 된다.



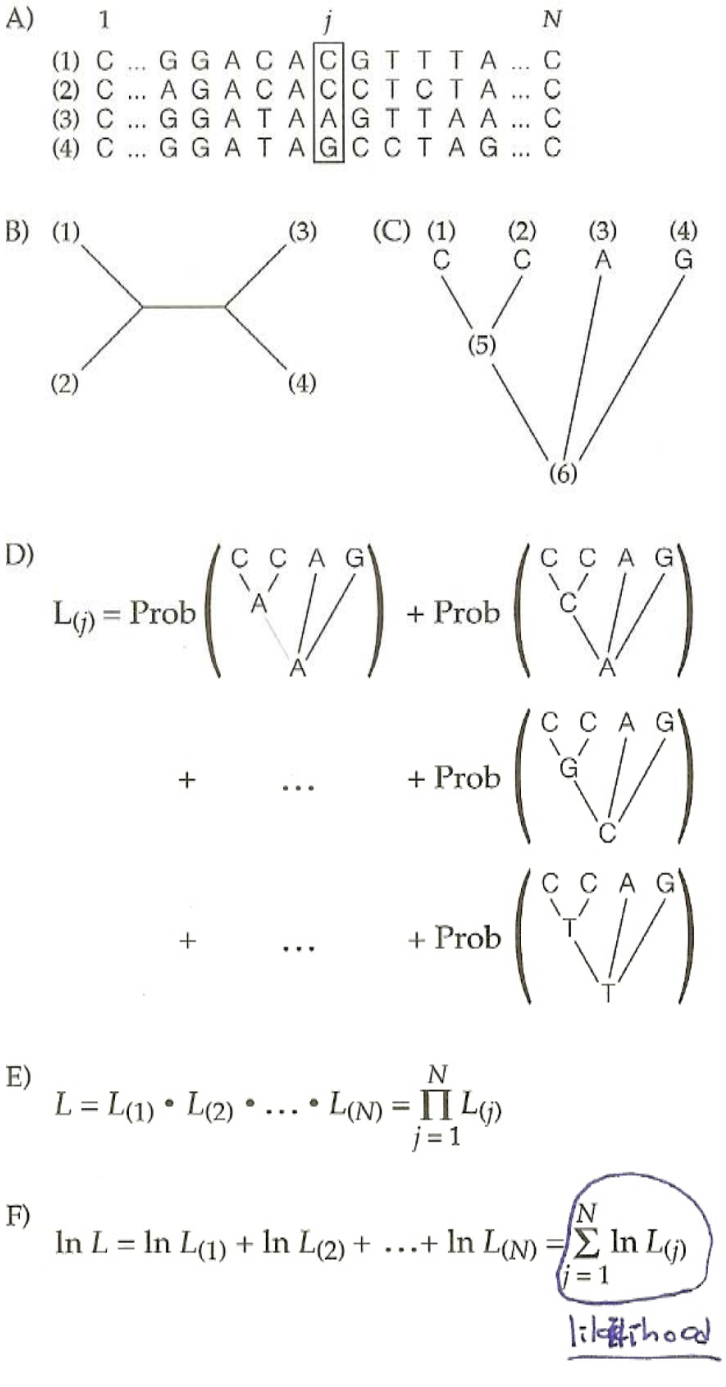


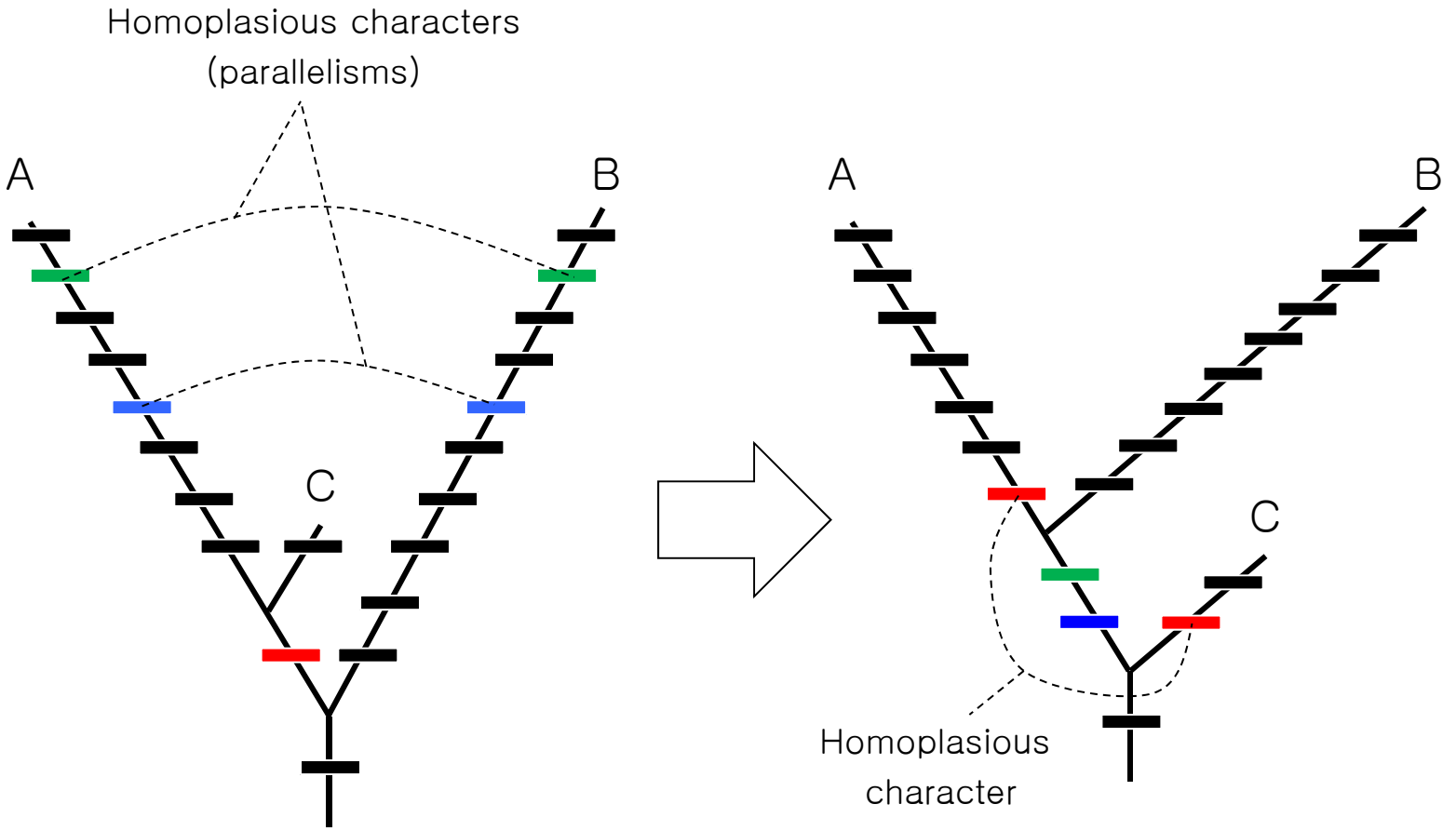
Figure 10 Overview of the calculation of the likelihood of a tree. (A) Hypothetical sequence alignment. (B) An unrooted tree for the four taxa whose sequences appear in (A). (C) Tree after rooting at an arbitrary internal node. (D) The likelihood for a particular site is the sum of the probabilities of every possible reconstruction of ancestral states given some model of base substitution. (E) The likelihood of the full tree is the product of the likelihoods at each site. (F) The likelihood is usually evaluated by summing the log of the likelihoods at each site, and reported as the log likelihood of the full tree.

Maximum likelihood method:

- 통계학에 기반을 둔 방법
- site 당 염기가 진화할 수 있는 모든 확률의 합을 계산하고 이를 모든 site에서 계산하여 합친 값을 최대화한 하나의 가능성을 찾아내는 방법
- long branch attraction을 줄여줄 수 있는 방법으로 제시됨.

Long Branch Attraction

긴가지 친화현상 → 김상태 2010 Basal-most angiosperm review 참조



True Phylogeny:
21 steps

Tree Generated by Parsimony Analysis:
20 steps

Long-branch attraction 이란 계통수 형성에 있어서 terminal node들이 매우 긴 tree들은 논리적으로 정확한 이론에 의한 계통수 작성법에도 불구하고 간혹 틀린 계통관계를 산출해 낼 수 있다는 것이다. 이것은 DNA의 형질상태가 4개 밖에 되지 않아서 가지가 길어지면 계통적 연관관계가 없는 데도 불구하고 같은 형질상태를 나타낼 수 있기 때문이다.

Who can minimize long-branch attraction?

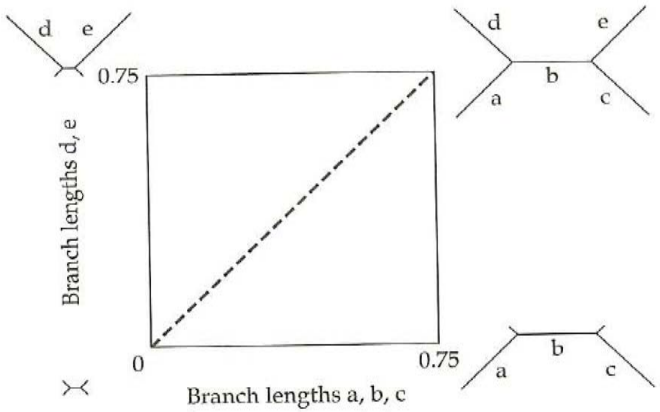


Figure 4 The complete parameter space for the four-taxa, two-rates problem originally outlined by Felsenstein (1978a), extended to four character states (e.g., DNA sequence data). In this problem, two opposing peripheral branches (d and e) co-vary; their length (measured in terms of proportion of differences) is plotted along the vertical axis. The remaining three branches (the other two peripheral branches and the central branch) also co-vary; their length is plotted along the horizontal axis. Because there are only four states, the maximum expected divergence for two sequences separated by an infinitely long time is 0.75. The dotted line along the diagonal represents equal rates of change along all lineages. The four trees shown represent relative branch lengths for trees near the respective corners of the graph. Extensive simulations have evaluated the performance of most major methods of phylogenetic analysis throughout this parameter space (see Huelsenbeck and Hillis, 1993; Huelsenbeck, 1995a; and Figure 5).

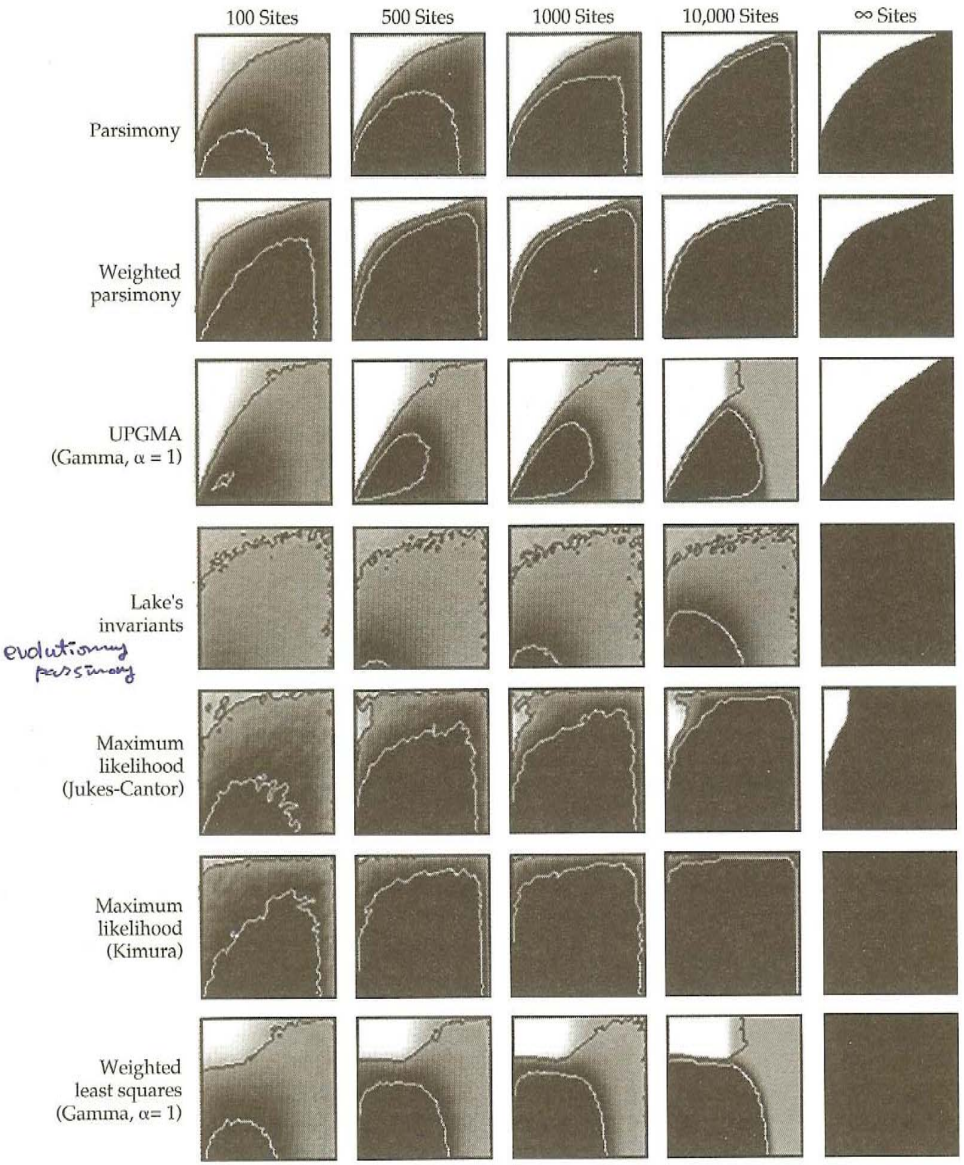
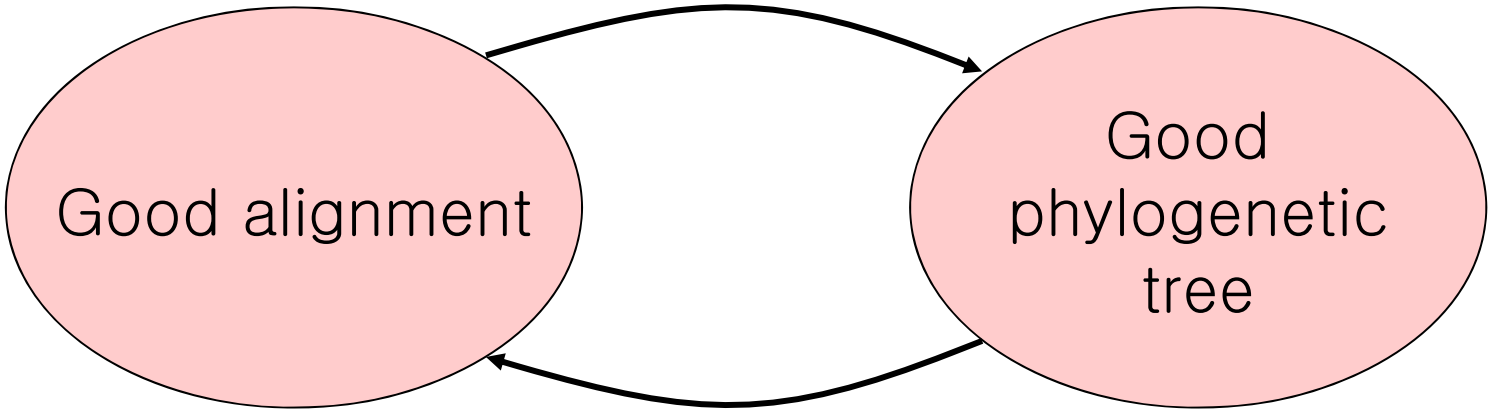


Figure 5 Graphs of the parameter space described in Figure 4, showing the performance of seven phylogenetic methods (rows) for five different sequence lengths (columns). Performance is indicated by shading, from black (method correct 100% of the time) to white (method correct 0% of the time). The white lines represent 95%-correct contours; the black lines represent 33%-correct contours. The model for the simulation was a Kimura two-parameter model with a 5:1 transition:transversion ratio. See Huelsenbeck (1995a) for details of the simulation and the performance of other methods (as well as simulations under other models).

지금까지 많은 학자들이 long-branch attraction 을 줄일 수 있는 계통분석법을 개발하려고 노력하였고, 각각의 방법을 computer simulation 에 의해 평가해 본 결과 Kimura's two parameter를 탑재한 maximum likelihood 법이 그 중 가장 전반적으로 long-branch attraction을 줄여줄 수 있는 방법으로 제시되고 있다

- How can we get good alignment and good phylogenetic tree?



지금까지 우리는 자료를 정렬 (alignment) 하는 법과 정렬된 자료를 이용하여 계통수를 작성하는 법을 공부하였다. 그런데, 자료를 보다 정밀하게 잘 정렬하려면 분류군간의 계통관계를 알아야 하고, 좋은 계통수를 작성하려면 좋은 alignment를 이용해야 하는 순환논리에 직면하게 된다. 이를 해결하려면 최초 정렬 후 분석 후 이 결과를 이용하여 다시 자료를 정렬하고 정렬된 결과를 이용하여 계통수를 다시 그리는 순환 반복을 수행할 때 더 좋은 계통수를 얻을 수 있게 된다.

- Tree of Life Project(ToL): 전 세계 생물 종 간의 계통관계를 집약하여 정리하는 국제 콘소시엄(http://tolweb.org/tree/)

- KToL: Korean Tree of Life

http://www.youtube.com/watch?v=H6IrUUDboZo