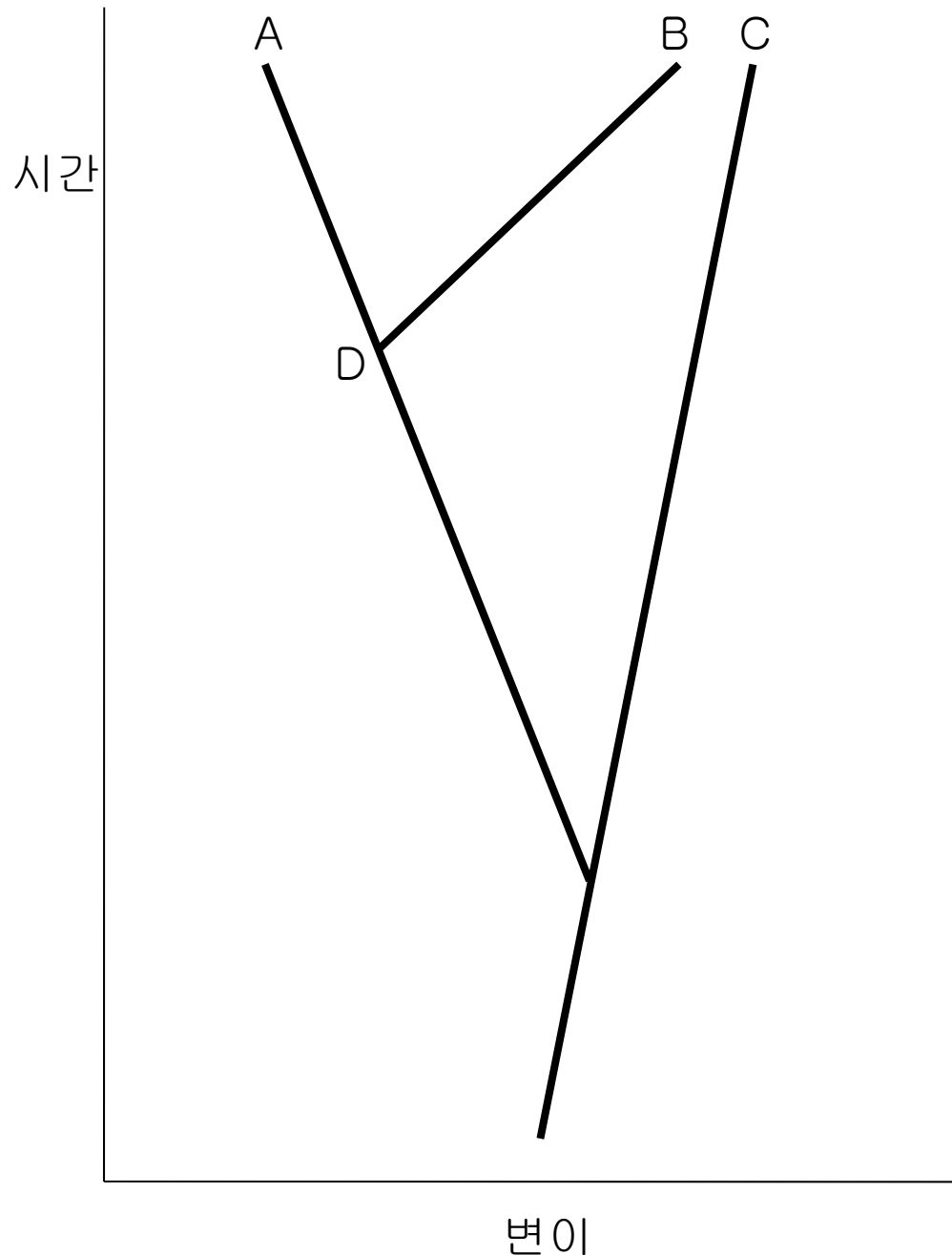


Chap. 4. Distance-Based Method of Phylogenetics

Distance-based method
vs.
Character-based method



- **Phenetic relationship**
(표현론적 관계)
(A(B,C))

→ Phenetic relationship에 의해 계통을 추론하는 법:
Distance-based method

- **Cladistic relationship**
(분지적 관계)
((A,B)C)

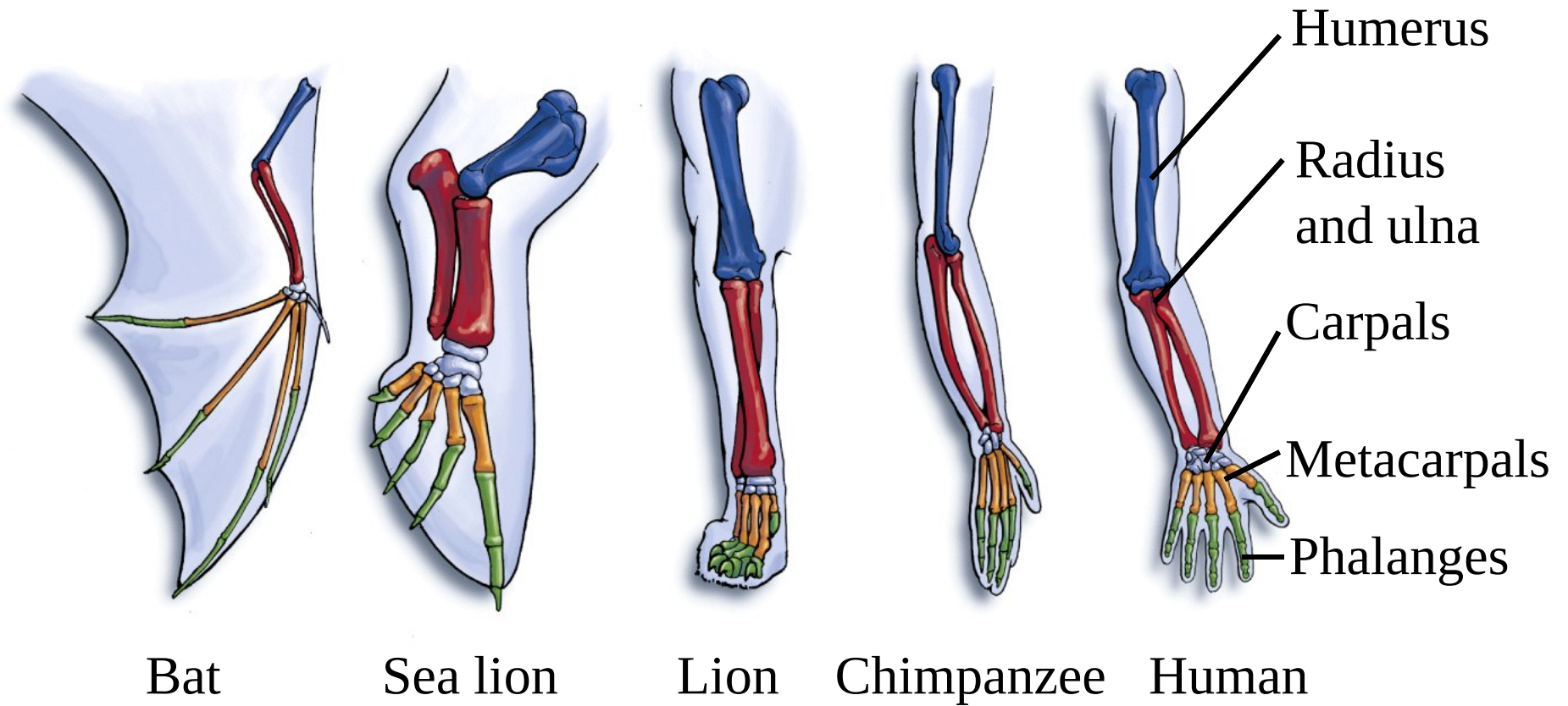
→ Cladistic relationship에 의해 계통을 추론하는 법:
Character-based method

참고) **Chronistic relationship**
(시간적 관계)
(D, (A,B,C))

- 종이하의 분류군 또는 집단간의 관계와 같이 낮은 rank의 분류군에서는 phenetic relationship = cladistic relationship이 될 가능성이 많지만 상위 rank에 있어서는 phenetic relationship은 cladistic relationship과 달라질 가능성이 많다.

Molecular Phylogenetics (분자계통학)의 장점

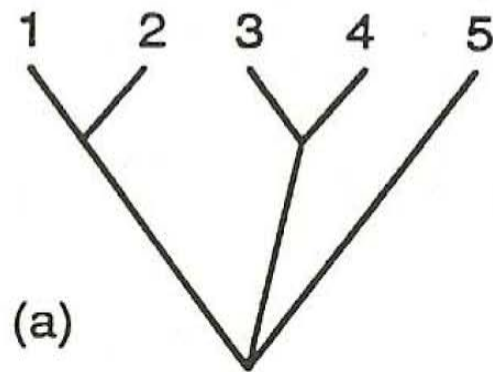
- 닮았나 닮지 않았나 라고 하는 phenotype들은 종종 진화적 계통관계를 잘 못 보여주기도 함. 형태학적 유사성은 언제나 유전적 유사성을 반영하는 것은 아니다.
 - homoplastic characters (불계적 형질) 때문임.
- 형태학적 형질들은 수가 적어 불충분 함.
- 분자적 자료의 비교에 있어서도 homology를 생각하여 비교의 대상이 될 수 있는 것을 비교하여야 함.
 - homologous gene (상동 유전자)들을 비교해야 하며, 또한 sequence alignment 가 필요함.



Homology 상동

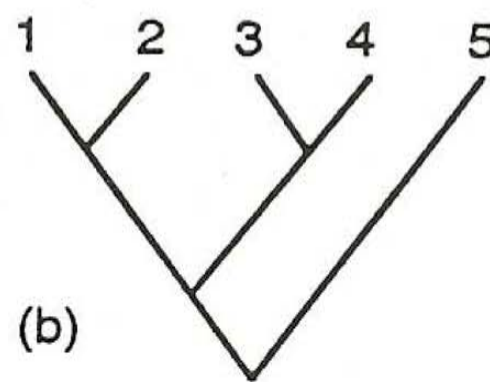
Newick Format

Note: polytomy



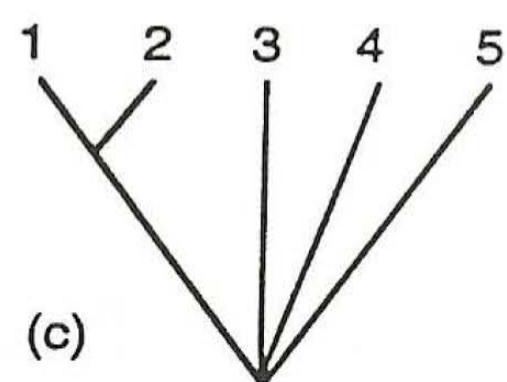
(a)

$((1,2),(3,4),5)$



(b)

$((((1,2),(3,4)),5)$



(c)

$((1,2),3,4,5)$

Three trees and their associated tree descriptions.

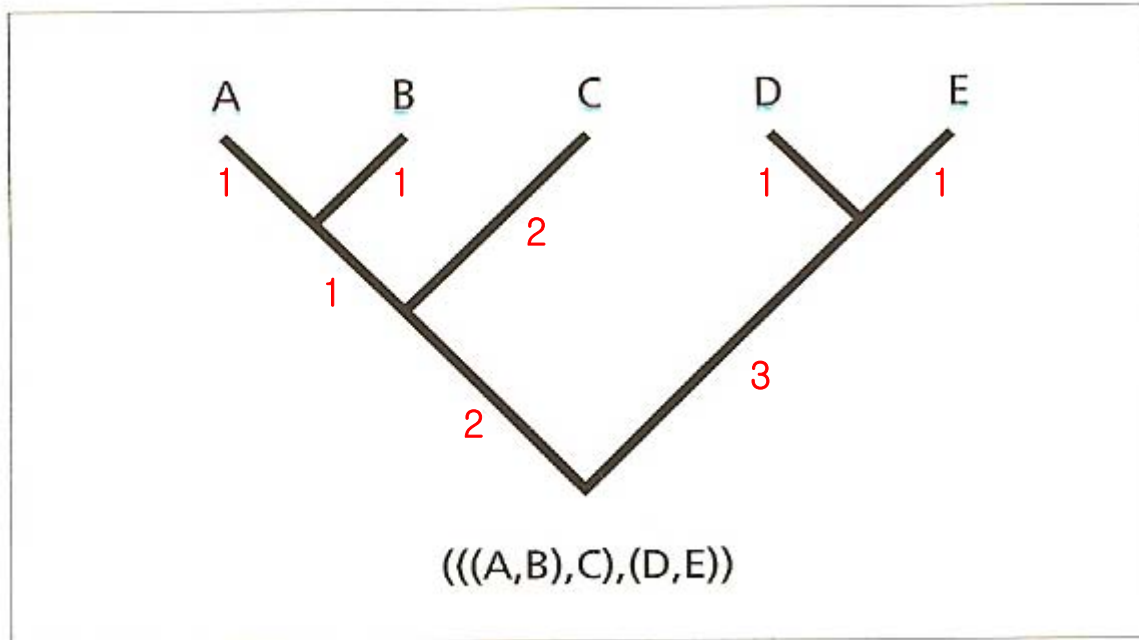


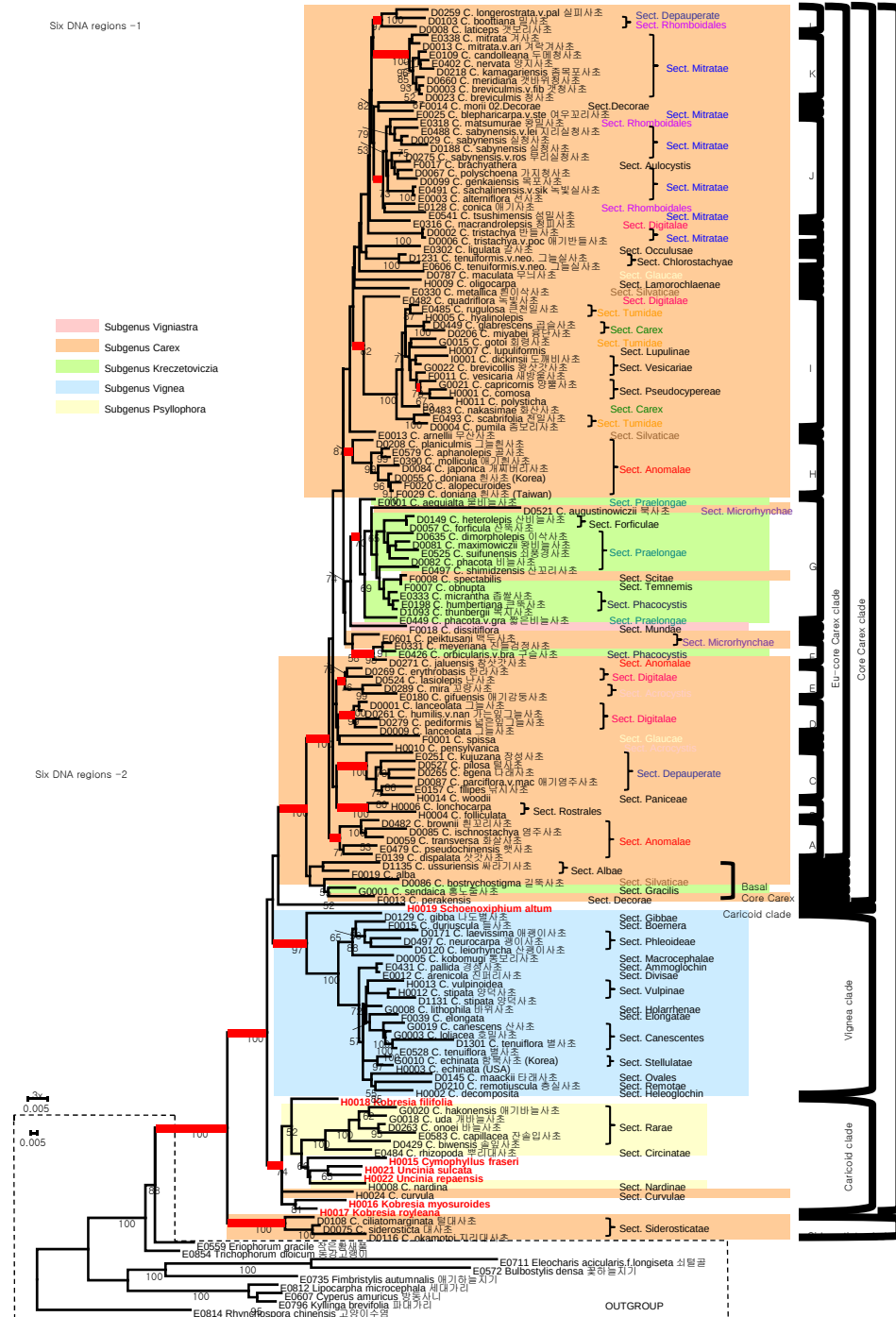
Fig. 2.4 A tree and its shorthand representation using nested parentheses.

Newick format

Tree lengths can be included in the newick format!

→ (((A:1,B:1):1,C:2):2,(D:1,E:1):3)

Six DNA regions -2

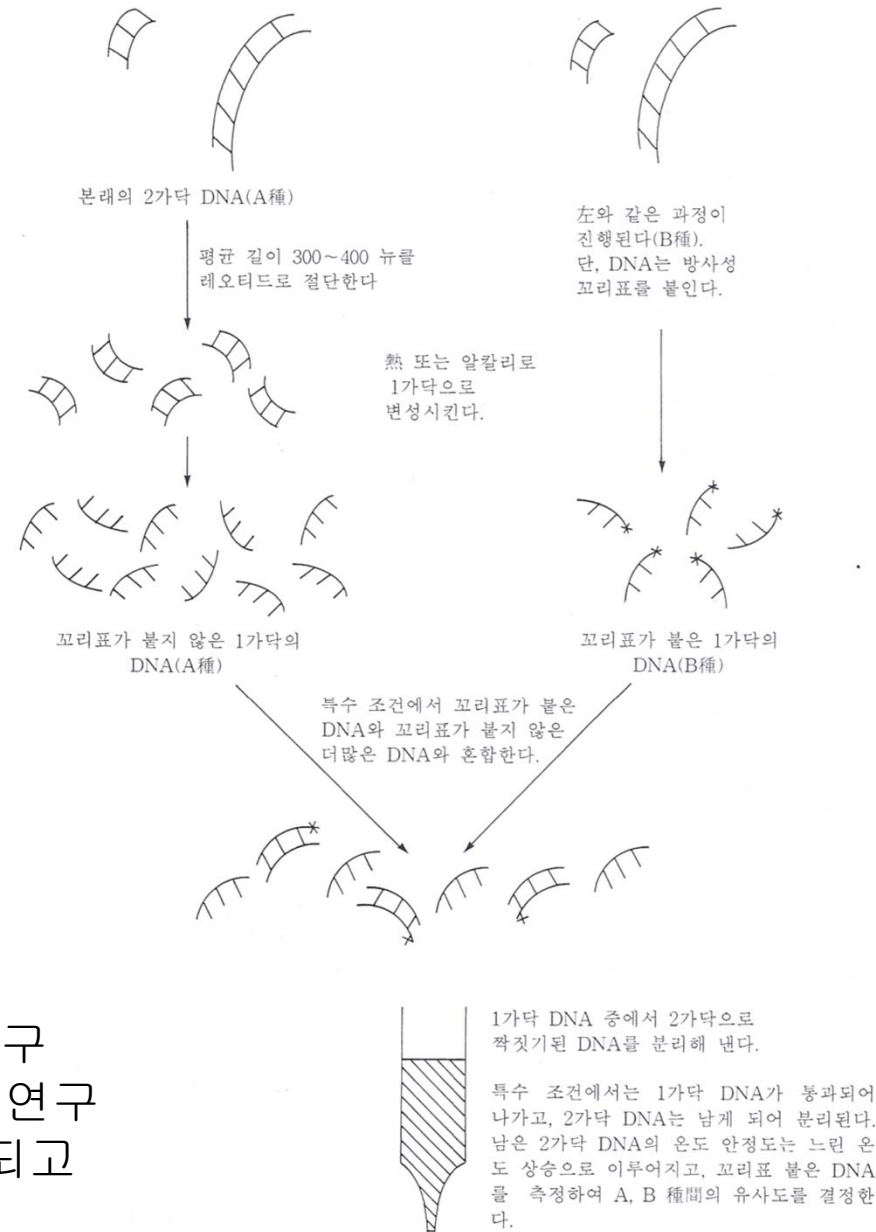


History of Molecular Phylogenetics

- 1) DNA–DNA hybridization
- 2) Restriction Fragment Length Polymorphism (RFLP)
(Protein sequencing, Immunological method)
- 3) DNA sequencing
 - i) Single gene comparison
 - ii) Multiple gene comparison
 - iii) Comparative genomics

계통학에서 사용되는 분자적 방법

1) DNA-DNA hybridization



- 생물간의 전체 유전체의 비교.
- DNA 분석 초기 동물군에서 활발히 연구
ex) 독수리, 콘돌, 황새간의 유연관계 연구
- 현재는 재현성이 부족하여 거의 사용되고 있지 않음

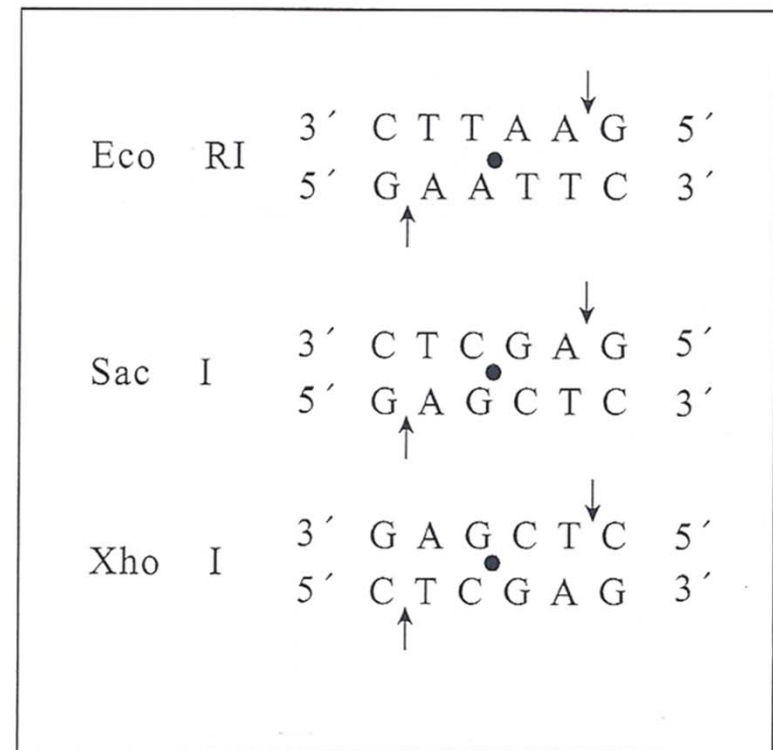
2) RFLP (restriction fragment length polymorphism)

그림 5-47 제한효소의 세 가지 예.

각 제한효소는 DNA 분자의 염기서열에서 특정한 인식부위(recognition sites)를 찾아서 그 지점에서만 작용이 제한된다. 각종 제한효소는 상품으로 나와있고 목록도 작성되어 있다.

↑ ↓ : 인식 부위

● (점) : 대칭축(axis of symmetry)



Enzyme	Source	Recognition Sequence	Cut
<u>EcoRI</u>	<u>Escherichia coli</u>	5'GAATTC	5'---G/AATTC---3'
<u>EcoRII</u>	<u>Escherichia coli</u>	5'CCWGG	5'---/CCWGG---3'
<u>BamHI</u>	<u>Bacillus amyloliquefaciens</u>	5'GGATCC	5'---G/GATCC---3'
<u>HindIII</u>	<u>Haemophilus influenzae</u>	5'AAGCTT	5'---A/AGCTT---3'
<u>TaqI</u>	<u>Thermus aquaticus</u>	5'TCGA	5'---T/CGA---3'
<u>NotI</u>	<u>Nocardia otitidis</u>	5'GCGGCCGC	5'---GC/GGCCGC---3'
<u>HinfI</u>	<u>Haemophilus influenzae</u>	5'GANTC	5'---G/ANTC---3'
<u>Sau3A</u>	<u>Staphylococcus aureus</u>	5'GATC	5'---/GATC---3'
<u>PvuII*</u>	<u>Proteus vulgaris</u>	5'CAGCTG	5'---CAG/CTG---3'
<u>SmaI*</u>	<u>Serratia marcescens</u>	5'CCCGGG	5'---CCC/GGG---3'

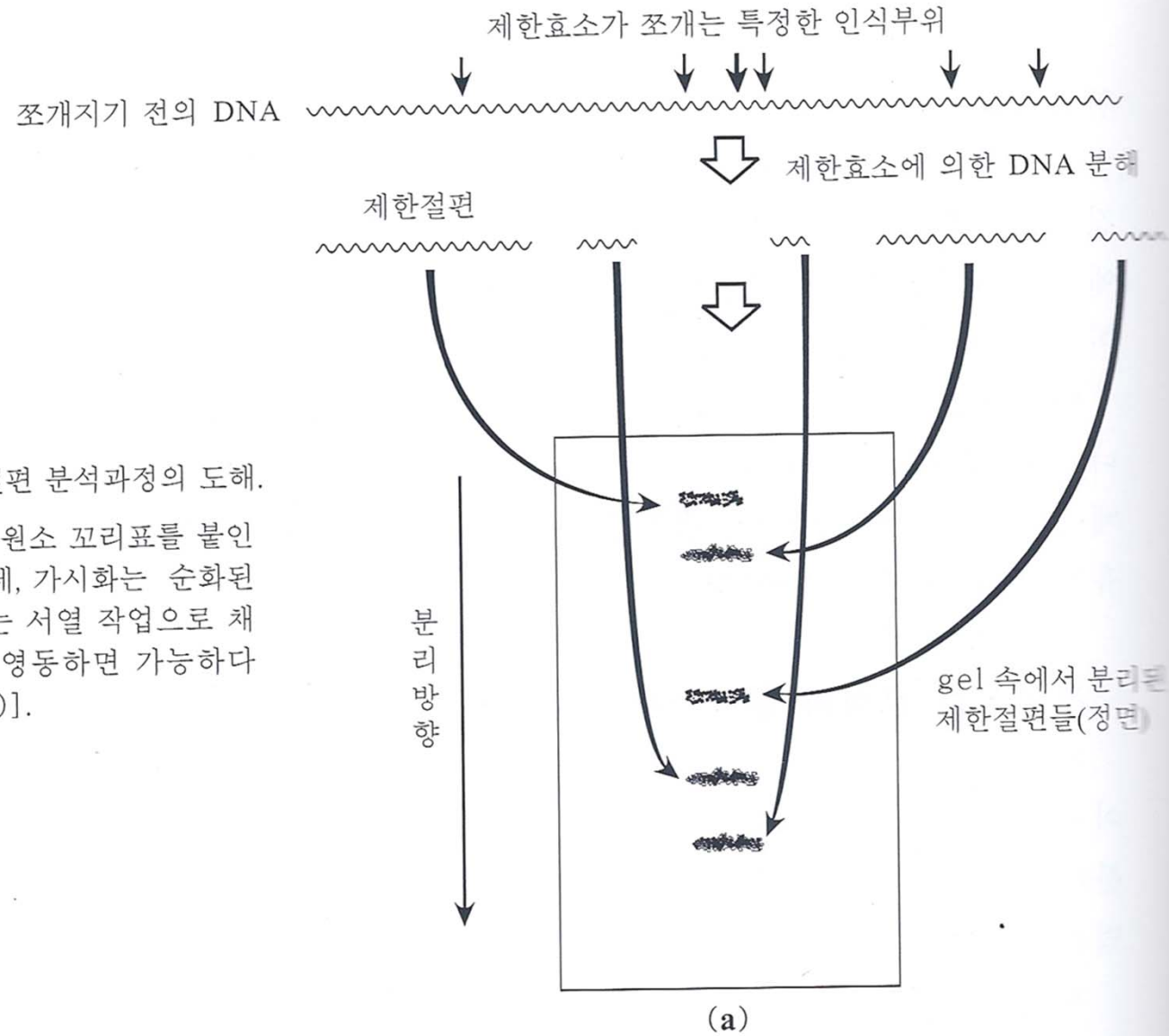
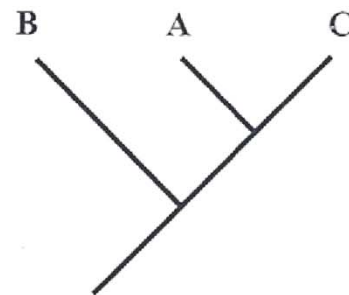
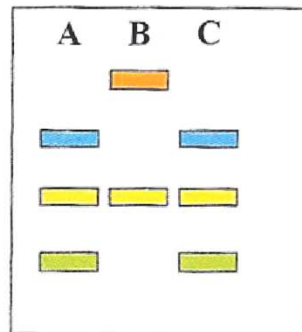
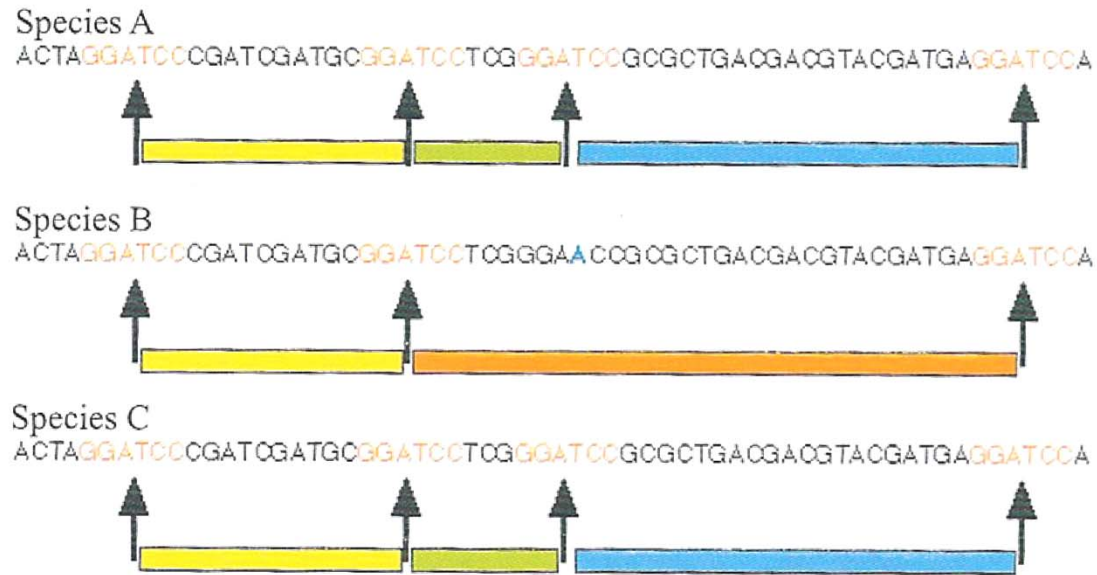


그림 5-48(a) 제한절편 분석과정의 도해.
 절편들은 방사성 동위원소 꼬리표를 붙인
 교잡으로 가시화되는데, 가시화는 순화된
 다량의 DNA(반복되는 서열 작업으로 채
 취한)를 모아서 전기영동하면 가능하다
 [출처: Quicke(1996)].

Restriction Site Analysis

The restriction enzyme BamHI cuts DNA everywhere it finds the sequence **GGATTC**



(b)

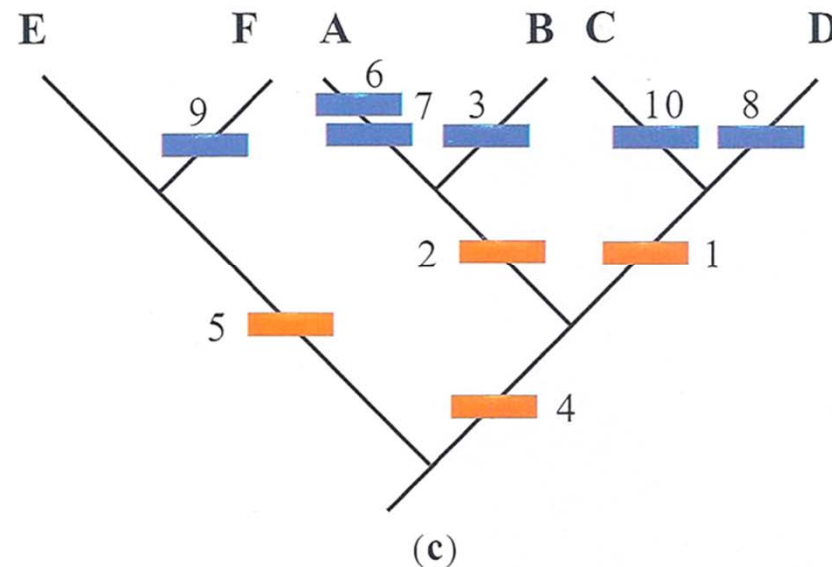
[출처: <http://www.life.uiuc.edu/plantbio/260/rcstsite.jpg>]

그림 5-48(c)

제한절편의 분석과정의 도해.

DNA 제한절편 분석으로 분류군 간의 유연관계를 밝히는 과정을 요약한 그림이다. 제한효소가 DNA의 인식부위를 만나면 절단하여 절편을 만든다. 이들을 gel 속에서 전기영동하여 절편들(침전선으로 나타남)의 분리패턴을 비교하면, 각 종에서 얻은 DNA의 인식부위 유무를 matrix로 나타낸다. 이를 토대로 A종, B종, C종, D종, E종, F종 간의 유연관계를 밝힌 것이다.

		Characters(Presence/absence)									
		1	2	3	4	5	6	7	8	9	10
A		0	1	0	1	0	1	1	0	0	0
B		0	1	1	1	0	0	0	0	0	0
C		1	0	0	1	0	0	0	0	0	1
D		1	0	0	1	0	0	0	1	0	0
E		0	0	0	0	1	0	0	0	0	0
F		0	0	0	0	1	0	0	0	1	0



[출처: <http://www.life.uiuc.edu/plantbio/260/MolSyst.html>]

- 식물에서는 주로 cpDNA의 RFLP 연구가 1980년대말~90년대 초반까지 활발히 이루어짐
- RFLP의 특징은 매우 긴 구간에 대하여 전반적으로 작은 부분의 sampling에 의하여 공유형질을 조사할 수 있다는 것으로 효율적으로 전체의 유연관계를 판단할 수 있다.
- 이에 반하여 DNA sequencing 방법은 상대적으로 작은 구간에 대하여 모든 DNA 염기서열을 비교하는 것이다. 즉 정밀한 분석은 되지만 가끔 전체를 대변하지 못할 때도 있다 → 유전자별로 분석한 결과가 틀린 경우도 있음.

(Protein sequencing, Immunological method)

3) DNA sequencing

- i) 단일 DNA 구간에 의한 자료
- ii) 복수의 DNA 구간에 의한 자료
- iii) Comparative genomics

- 계통연구를 위한 DNA 염기서열 분석의 장점:

- i) 객관적인 형질상태(character status): A/C/G/T
- ii) 계통상 의미 있는 형질의 수(informative character)가 무수히 많다.
- iii) 기존의 데이터에 새로운 데이터의 첨가가 용이하다.
- iv) 매우 적은 양의 시료에서도 데이터를 얻을 수 있다.
- v) 실험 결과가 광범위한 분류군에 이용될 수 있다.
 - 진화속도가 느린 유전자: 전체 피자식물의 계통
 - 진화속도가 빠른 유전자 또는 유전자들 사이의 염기서열 (intergenic spacer) 등: 속 또는 종들의 구분

- Basic process to reconstruct phylogenetic tree based on sequencing method

Sequencing → search → filtering → alignments → phylogenetic analyses

Phylogenetic Tree

- Node 절
- Terminal node 말단절
- Internal node 내부절
- Inferred ancestor (추정) 조상
- Newick format: $((I, II), (III, IV)), V)$
- Multifuracting vs. bifuracting



- Scaled tree vs. unscaled tree

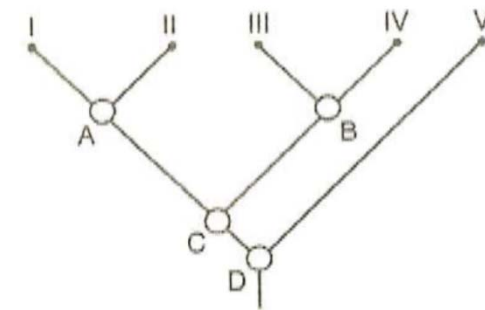
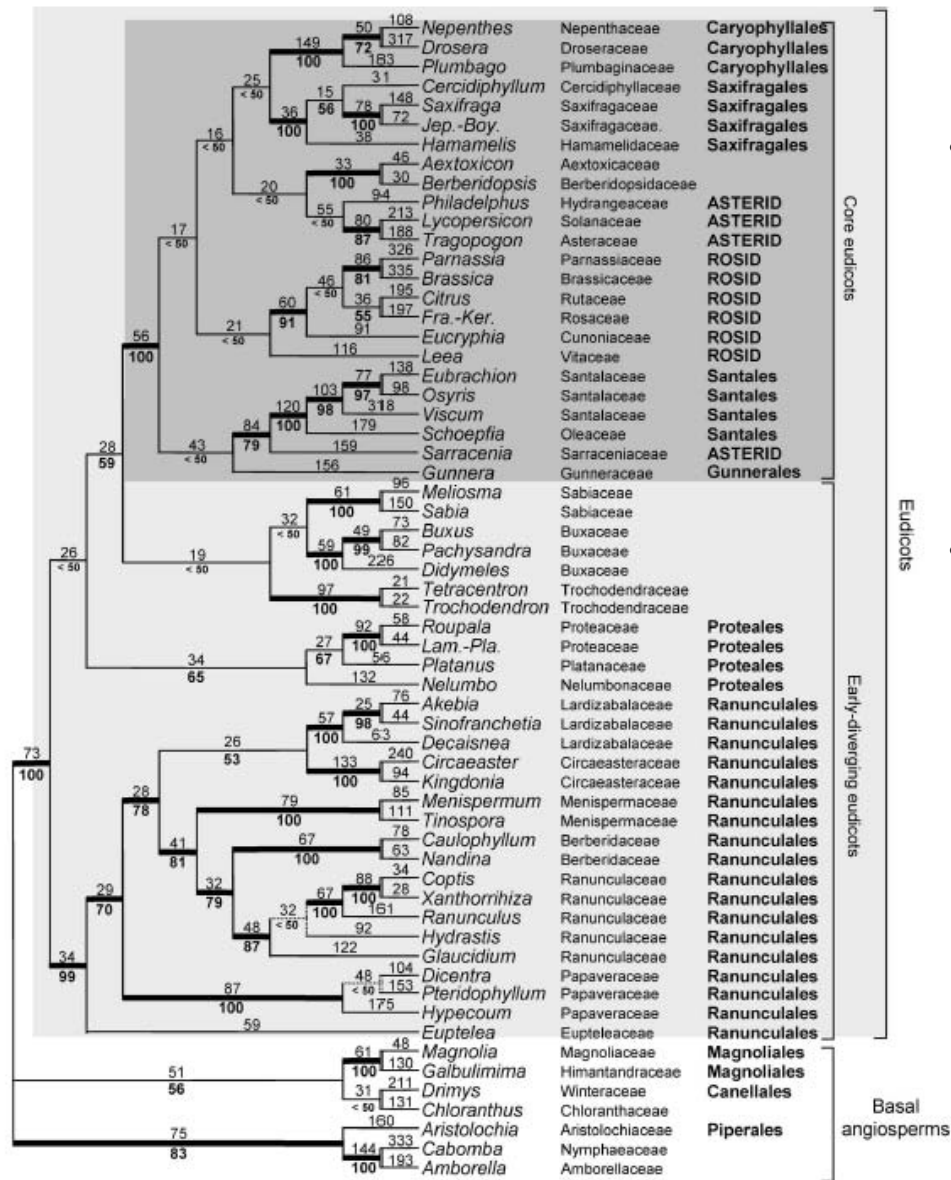


FIGURE 4.1 A phylogenetic tree illustrating the evolutionary relationships among five species (I, II, III, IV, and V). Filled circles represent terminal nodes while open circles correspond to internal nodes. Inferred common ancestors to the species for which sequence data are available are labeled with letters (A, B, C, and D). The root of the tree corresponds to D.

- 위 그림의 circle에 해당하는 분류군을 분자계통학과 진화학에서는 **MRCA** (most recent common ancestor)라고 부른다.



- 왼쪽 그림과 같이 모든 분류군이 동일 선상에 놓여 있고, tree는 이들간의 “관계” 만을 보여주고 있을 때 이를 **unscaled tree** 라고 한다. 이때 Y축에서의 분류군간의 거리와, X축에서의 node 간의 거리는 무의미하다.
- 분류군의 분지양상만 고려하여 표시한 수지도 (dendrogram)인 **cladogram (분지도; 분계도)**이 여기에 속한다.

Fig. 4. One of three most parsimonious trees based on the combination of *atpB*, *rbcL*, 18S rDNA, and 26S rDNA sequences. Numbers along branches, dotted lines, and bold lines are as in Fig. 2. Abbreviations for several genera are as follows: *Jep.*, *Jepsonia*; *Boy.*, *Boykinia*; *Fra.*, *Fragaria*; *Ker.*, *Kerria*; *Lam.*, *Lambertia*; *Pla.*, *Placospermum*.

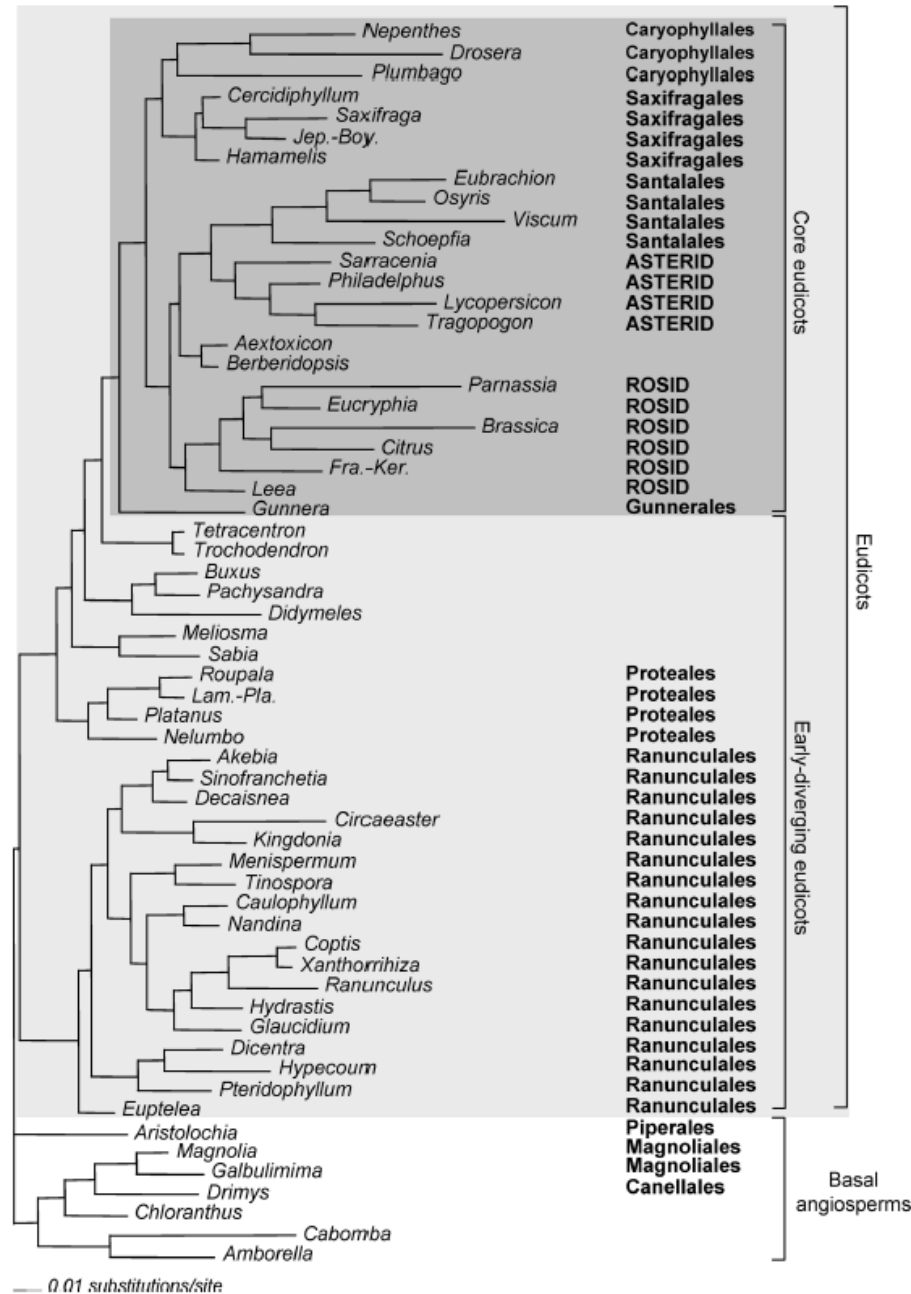


Fig. 5. Maximum likelihood tree generated from analysis of four-gene data set.

- Scaled tree에 있어서는 각 분류군이 동일직선상에 놓여있지 않다. 즉 각각의 branch의 길이들이 모두 틀리고, 이 때 X축에서의 node 간의 거리(가지의 길이)는 형질의 변화 양에 해당하고, Y축에서의 분류군 간의 거리는 unscaled tree와 마찬가지로 무의미하다.
- 분류군의 분지양상과 함께 변화한 형질의 수를 반영하는 계통도를 **phylogram** 이라고 한다.

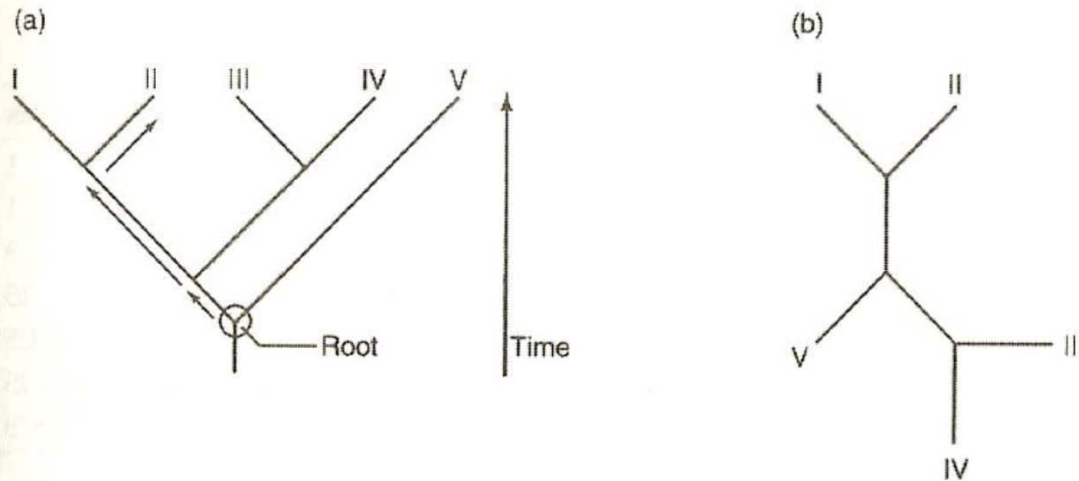


FIGURE 4.2 (a) Rooted and (b) unrooted trees. Arrows indicate a unique path leading from the root to species D in the rooted tree. No inferences regarding the direction in which evolution occurred can be made from the unrooted tree.

Rooted vs. Unrooted Tree

- 모든 계통수의 도출법은 공통적으로 먼저 **unrooted tree** (tree network)가 만들어진 후 root를 설정하여 진화의 방향을 표시하면 **rooted tree**가 된다. Root의 설정은 outgroup을 지정함으로써 이루어 지는데, outgroup이란 연구 대상의 분류군과 공동 조상을 가장 먼저 이루는 군임 (진화의 역사에 있어서 공동조상으로부터 연구 대상과 가장 빨리 분지된 군).

Rooted tree: 궁극적으로 모든 대상 종들이 뻗어 나왔다고 생각되는 하나의 root를 설정한 계통수.

→ root 설정 후 비로소 조상-후손관계가 성립된다.

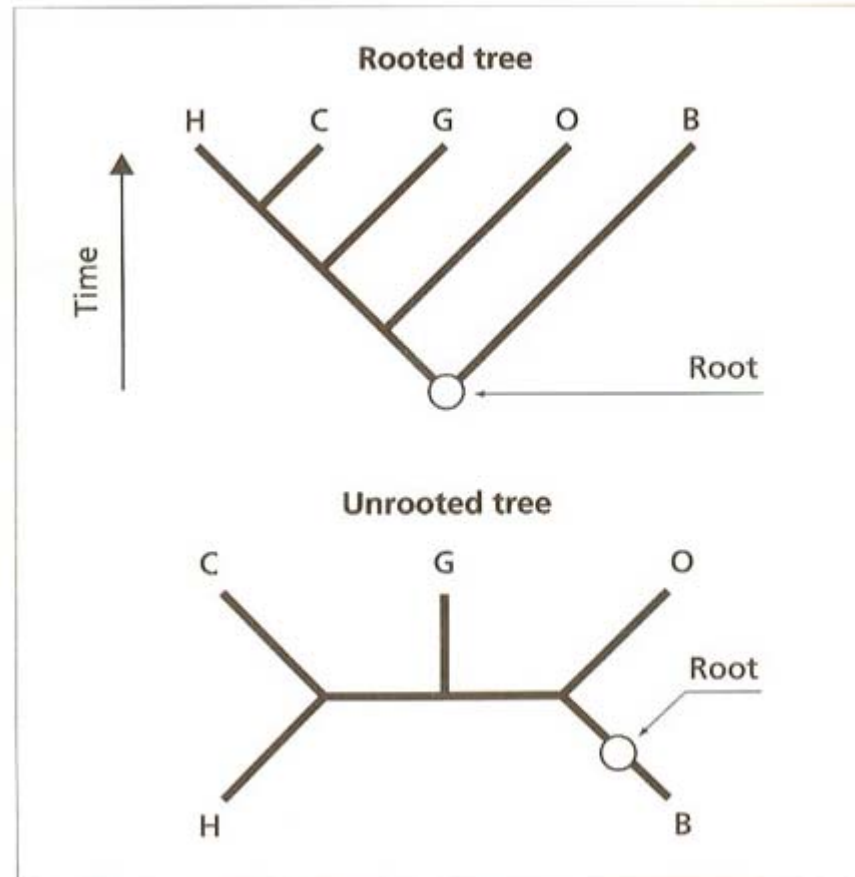


Fig. 2.6 Rooted and unrooted trees for human (H), chimp (C), gorilla (G), orang-utan (O), and gibbon (B). The rooted tree (top) corresponds to the unrooted tree below.

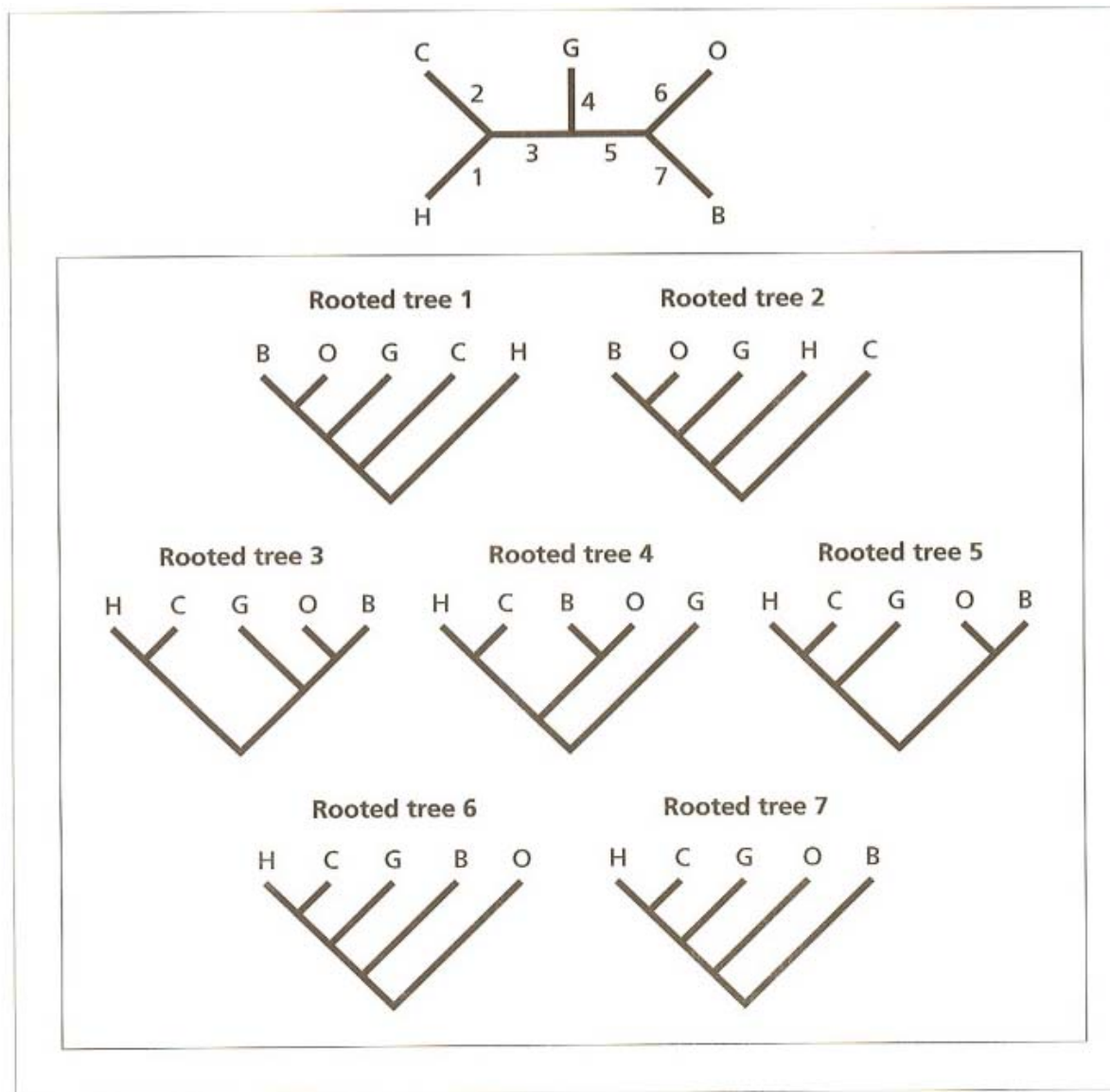
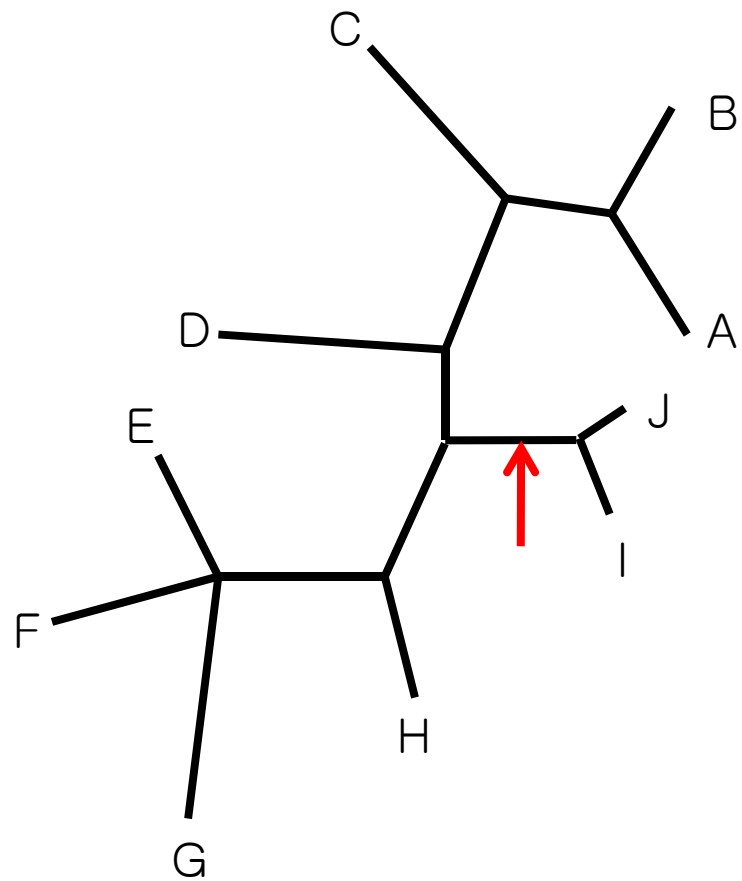
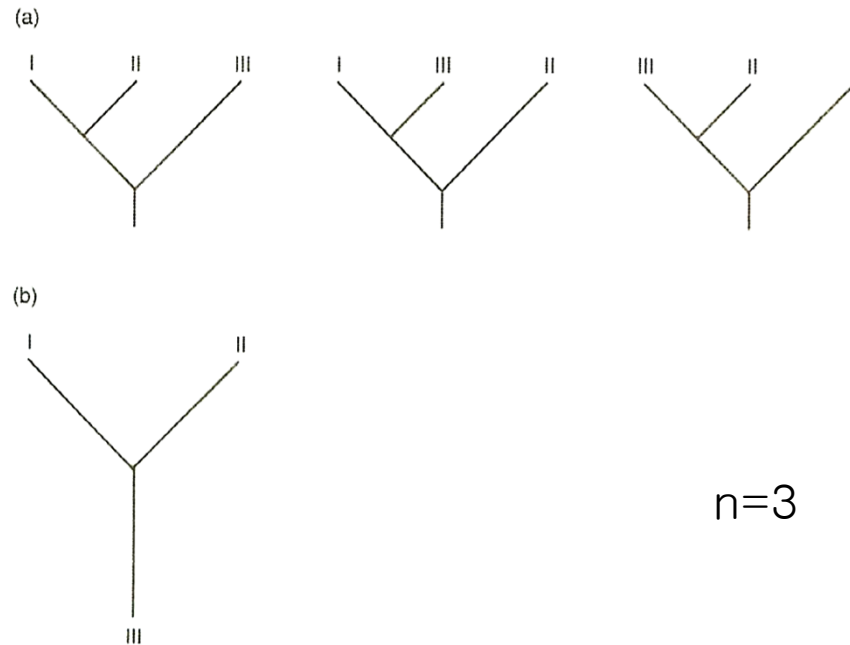


Fig. 2.7 The seven rooted trees that can be derived from an unrooted tree for five sequences. Each rooted tree 1–7 corresponds to placing the root on the corresponding numbered branch of the unrooted tree. (Sequence labels as for Fig. 2.6.)

Question: Draw a ROOTED cladogram based on the following unrooted tree with the arrowed rooting position.





n=3

FIGURE 4.3 All possible (a) rooted and (b) unrooted trees when only three species are considered.

- n 개의 종이 있을 때 가능한 Rooted tree (N_R) 와 unrooted tree (N_U) 의 수

$$N_R = (2n - 3)! / 2^{n-2} (n - 2)!$$

$$N_U = (2n - 5)! / 2^{n-3} (n - 3)!$$

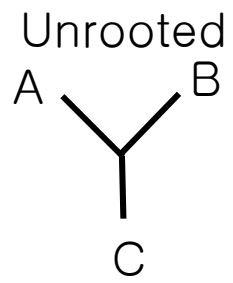
- 135 human mtochondrial DNA sequences data에서 가능한 tree의 수는 2.113×10^{267} !!!!!

TABLE 4.1 Number of possible rooted and unrooted trees that can describe the possible relationships among fairly small numbers of data sets.

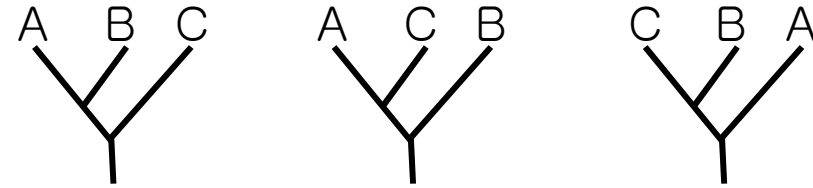
Number of Data Sets	Number of Rooted Trees	Number of Unrooted Trees
2	1	1
3	3	1
4	15	3
5	105	15
10	34,459,425	2,027,025
15	213,458,046,767,875	7,905,853,580,625
20	8,200,794,532,637,891,559,375	221,643,095,475,699,771,875

Find the best tree in HERE

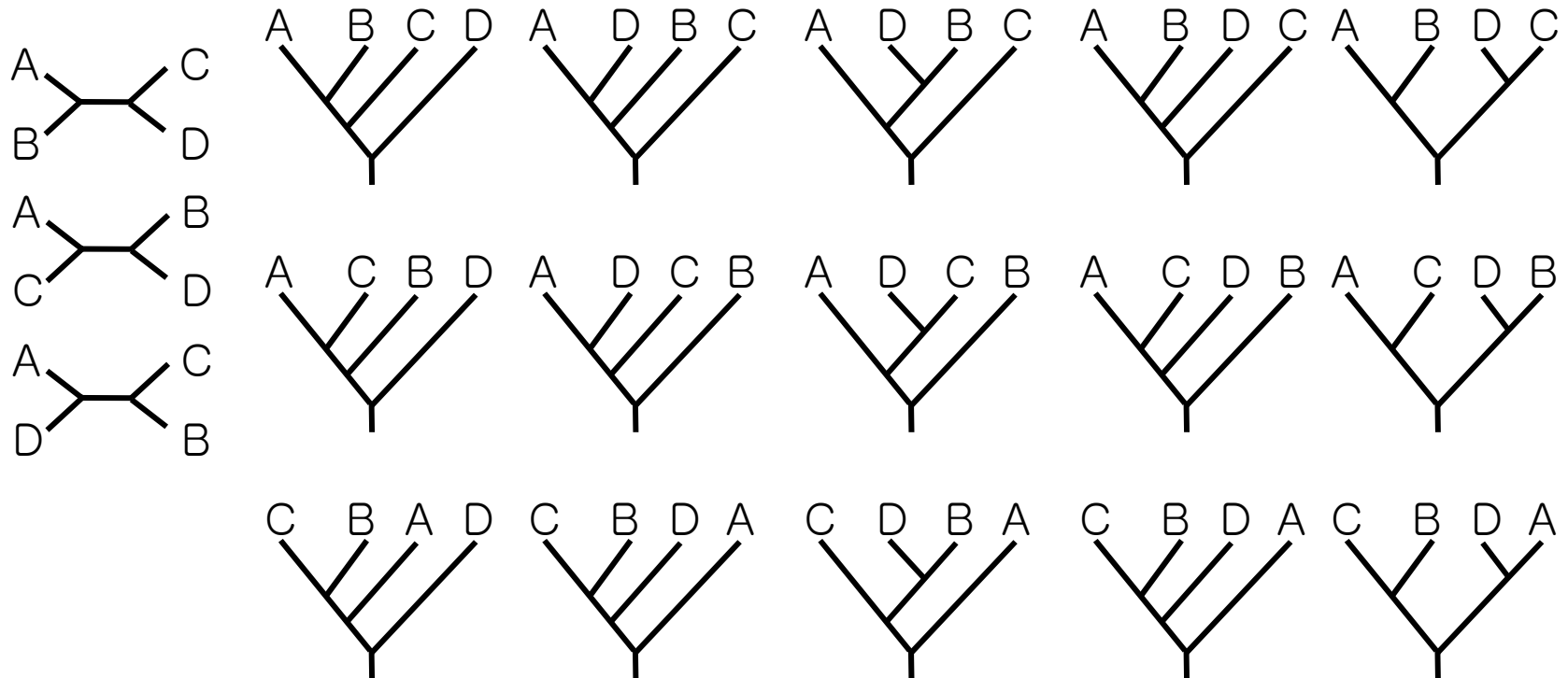
N=3



Rooted



N=4



N=n

...

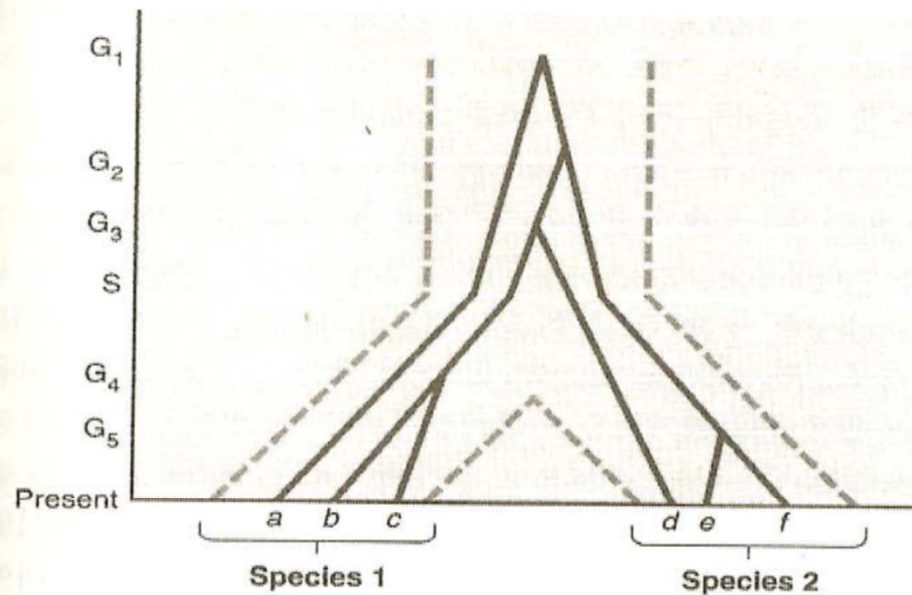


FIGURE 4.4 Individuals may actually appear to be more closely related to members of a species other than their own when only one gene is considered. Gene divergence events (G_1 through G_5) often occur before as well as after speciation events (S). The evolutionary history of gene divergence resulting in the six alleles denoted a through f is shown in solid lines; speciation (i.e., population splitting) is shown by broken lines. Individual d would actually appear to be more closely related to individuals in species 1 if only this locus were considered even though it is a member of species 2.

– Species tree가 언제나 gene tree와 일치하는 것은 아님. → homologous 한 유전자를 비교해야 함.

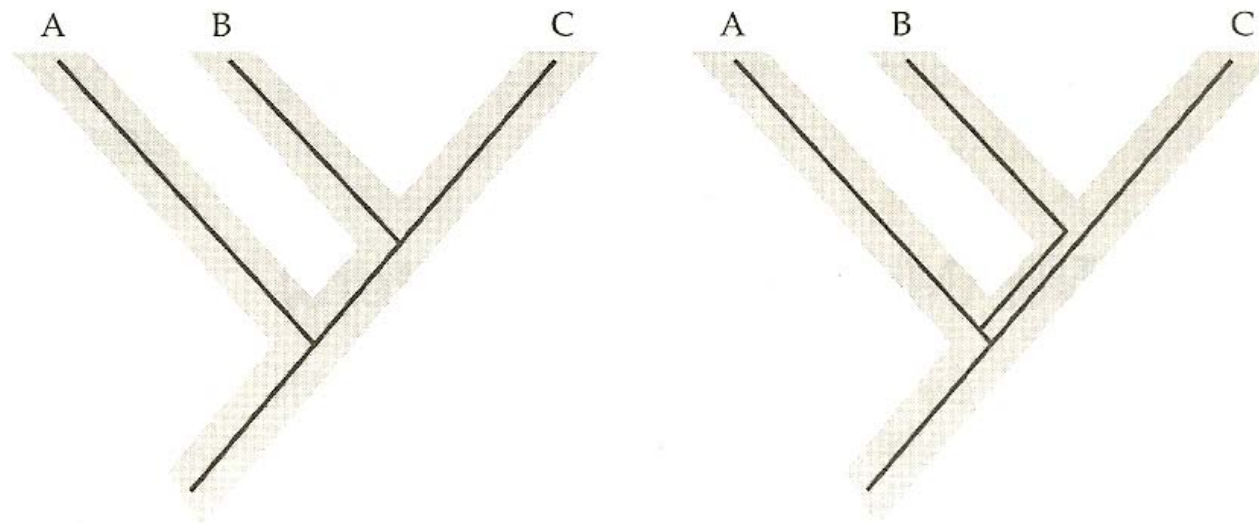


그림 5.3 유전자계통수와 종계통수의 비교. 유전자 계통수는 진한 선으로 종계통수는 회색음영처리 하였다. 좌측의 계통수는 유전자가 종과 동일한 역사를 갖는다. 오른쪽 계통수는 다형현상이 종 B와 C의 계열에 나타난다. 두 유전자복제중의 하나는 A종과 더 근연이다. 이러한 유전자의 선택은 종 계통수에 대한 잘못된 추론을 야기 할 것이다. (After Avise 1994.)

Homologous 상동

Orthologous genes: originated from speciation

Paralogous genes: originated from gene duplications

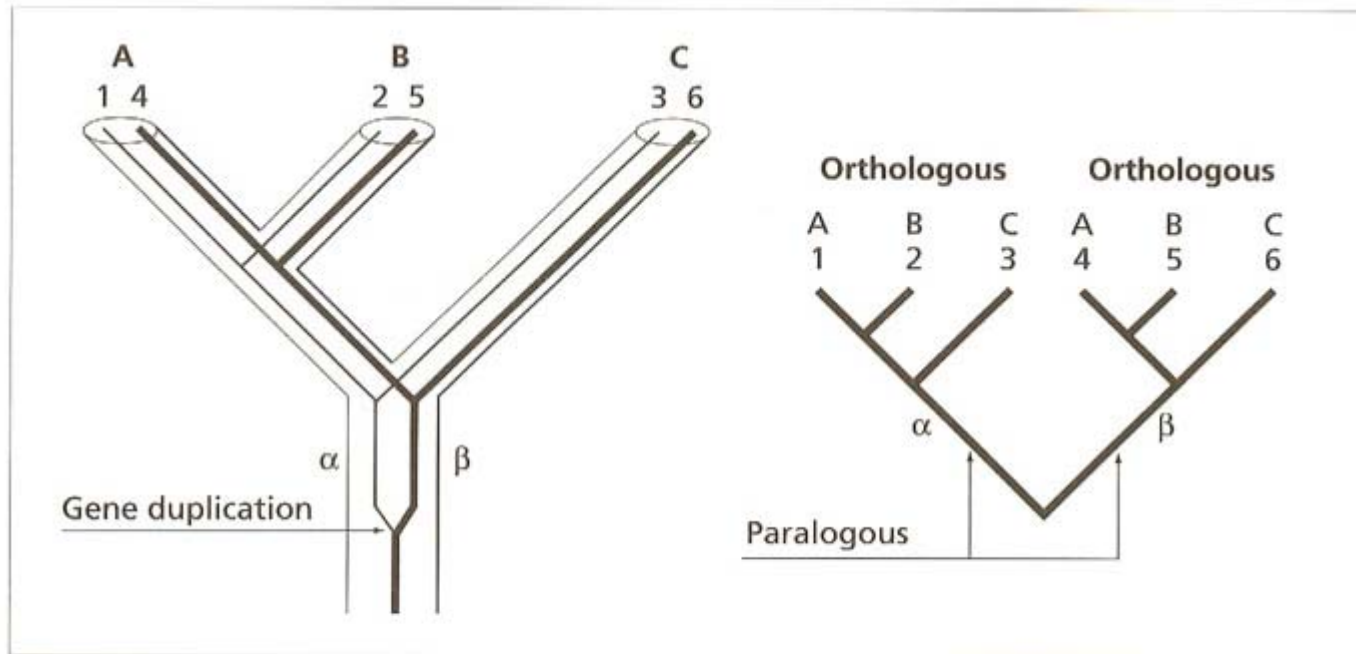


Fig. 2.22 Phylogeny for three species A–C and six genes that stem from a gene duplication resulting in two paralogous clades of genes, α and β . The α genes 1–3 are orthologous with each other, as are the β genes 4–6; however each α gene is paralogous with each β gene as they are separated by a gene duplication event, not a speciation event.

Distance Matrix Method: 가장 간단한 계통수 제작법

Distance based method의 대표적 예

UPGMA:

Unweighted-pair-group method with arithmetic mean (산술평균 비가중 쌍별 그룹법)

	10	20	30	40	50
A:	GTGCTGCACGG	CTCAGTATA	GCATTTACCC	TCCCATCTTC	AGATCCTGAA
B:	ACGCTGCACGG	CTCAGTGCG	GCGCTTACCC	TCCCATCTTC	AGATCCTGAA
C:	GTGCTGCACGG	CTCGGCGCA	GCATTTACCC	TCCCATCTTC	AGATCCTATC
D:	GTATCACACGA	CTCAGCGCA	GCATTTGCCC	TCCCGTCTTC	AGATCCTAAA
E:	GTATCACATAG	CTCAGCGCA	GCATTTGCCC	TCCCGTCTTC	AGATCTAAAA

FIGURE 4.5 A five-way alignment of homologous DNA sequences.

Pairwise Distance Matrix

Species	A	B	C
B	d_{AB}	-	-
C	d_{AC}	d_{BC}	-
D	d_{AD}	d_{BD}	d_{CD}

TABLE 4.2 A pairwise distance matrix that summarizes the number of nonmatching nucleotides between all possible pairs of sequences shown in Figure 4.5.

Species	A	B	C	D
B	9	-	-	-
C	8	11	-	-
D	12	15	10	-
E	15	18	13	5

- 가장 작은 거리를 나타내는 것을 선택하여 하나의 group으로 묶음.
- 이것과 나머지 것들과의 거리를 구함.

→ D 와 E 가 가장 작은 거리를 나타냄: (D, E)를 우선적으로 묶음

→ Distance between (DE) and X:

$$d_{(DE)X} = \frac{1}{2} (d_{DX} + d_{EX})$$

$$\text{ex) } d_{(DE)A} = \frac{1}{2} (d_{DA} + d_{EA}) = \frac{1}{2} (12 + 15) = 13.5$$

TABLE 4.3 The distance matrix that results when species D and E of Table 4.2 are combined and considered as a single group.

Species	A	B	C
B	9	-	-
C	8	11	-
DE	13.5	16.5	11.5

- The smallest number is 8:
(A, C)

TABLE 4.4 The distance matrix that results when species A and C of Table 4.3 are combined and considered as a single group.

Species	B	AC
AC	10	—
DE	16.5	12.5

Final Newick format: (((A, C)B, (D, E))

Estimation of Branch Lengths

– Cladogram

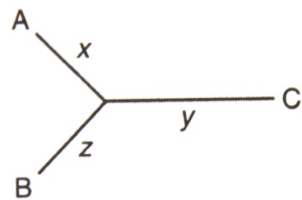


FIGURE 4.8 The simplest tree whose branch lengths might have some meaningful information. Each of the three branches on this tree can be represented as a single variable (x , y , and z).

$$d_{AC} = x + y$$

$$d_{AB} = x + z$$

$$d_{BC} = y + z$$

$$x = (d_{AB} + d_{AC} - d_{BC})/2$$

$$y = (d_{AC} + d_{BC} - d_{AB})/2$$

$$z = (d_{AB} + d_{BC} - d_{AC})/2$$

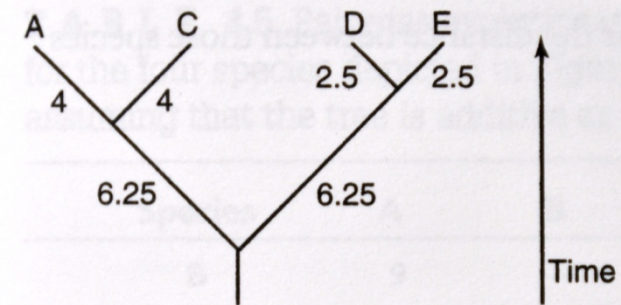


FIGURE 4.7 A scaled tree showing the branch lengths separating four of the species depicted in Figure 4.6. Branch lengths are shown next to each branch. Branches are also drawn to scale to reflect the amount of differences between all species. If evolution has occurred at a constant rate for all of these lineages, then branch length also corresponds to divergence time.

IF the molecular clock is present!

	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

Uninformative character: 1, 2, 6

Variable character: 2, 3, 4, 5, 6

Informative character: 4, 5

Parsimony method: using informative characters

Distance method: using all characters

Transformed Distance Method (Farris, 1977)

- UPGMA 단점: 모든 계통의 진화속도가 일정하다고 가정.
- TDM 은 각 lineage들의 진화속도를 다르게 가정하여 계산할 수 있음.

PHYLOGENY AND DIVERSIFICATION OF B-FUNCTION MADS-BOX GENES IN ANGIOSPERMS: EVOLUTIONARY AND FUNCTIONAL IMPLICATIONS OF A 260-MILLION-YEAR-OLD DUPLICATION¹

SANGTAE KIM,^{2,6} MI-JEONG YOO,² VICTOR A. ALBERT,³
JAMES S. FARRIS,⁴ PAMELA S. SOLTIS,⁵ AND DOUGLAS E. SOLTIS^{2,6}

²Department of Botany, University of Florida, Gainesville, Florida 32611 USA; ³The Natural History Museums and Botanical Garden, University of Oslo, P.O. Box 1172 Blindern, NO-0318 Oslo, Norway; ⁴Molekylärsystematiska laboratoriet, Naturhistoriska riksmuseet, Box 50007, SE-104 05 Stockholm, Sweden; and ⁵Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611 USA

B-function MADS-box genes play crucial roles in floral development in model angiosperms. We reconstructed the structural and functional implications of B-function gene phylogeny in the earliest extant flowering plants based on analyses that include 25 new *AP3* and *PI* sequences representing critical lineages of the basalmost angiosperms: *Amborella*, *Nuphar* (Nymphaeaceae), and *Illicium* (Austrobaileyales). The ancestral size of exon 5 in *PI*-homologues is 42 bp, typical of exon 5 in other plant MADS-box genes. This 42-bp length is found in *PI*-homologues from *Amborella* and Nymphaeaceae, successive sisters to all other angiosperms. Following these basalmost branches, a deletion occurred in exon 5, yielding a length of 30 bp, a condition that unites all other angiosperms. Several shared amino acid strings, including a prominent “DEAER” motif, are present in the *AP3*- and *PI*-homologues of *Amborella*. These may be ancestral motifs that were present before the duplication that yielded the *AP3* and *PI* lineages and subsequently were modified after the divergence of *Amborella*. Other structural features were identified, including a motif that unites the previously described TM6 clade and a deletion in *AP3*-homologues that unites all Magnoliales. Phylogenetic analyses of *AP3*- and *PI*-homologues yielded gene trees that generally track organismal phylogeny as inferred by multigene data sets. With both *AP3* and *PI* amino acid sequences, *Amborella* and Nymphaeaceae are sister to all other angiosperms. Using nonparametric rate smoothing (NPRS), we estimated that the duplication that produced the *AP3* and *PI* lineages occurred approximately 260 mya (231–290). This places the duplication after the split between extant gymnosperms and angiosperms, but well before the oldest angiosperm fossils. A striking similarity in the multimer-signalling C domains of the *Amborella* proteins suggests the potential for the formation of unique transcription-factor complexes. The earliest angiosperms may have been biochemically flexible in their B function and “tinkered” with floral organ identity.

Key words: *Amborella*; *AP3*; B-class; basal angiosperms; MADS; *PI*.

in Figure 4.5 that have accumulated along each lineage during each stage of its evolution, then the pairwise evolutionary distances are shown in Table 4.5 (an abbreviated version of Table 4.2). In this situation, D can be used as an external reference to transform the distances that separate the other species using this equation:

$$d'_{ij} = (d_{ij} - d_{iD} - d_{jD})/2 + \bar{d}_D$$

where d'_{ij} is the transformed distance between species i and j , and $\bar{d}_D$ is the average distance between the outgroup and all ingroups (in this case: $d_D = (d_{AD} + d_{BD} + d_{CD})/3 = 37/3$). The term $\bar{d}_D$ is introduced simply to ensure that all transformed distance

TABLE 4.5 Pairwise evolutionary distances for the four species depicted in Figure 4.7, assuming that the tree is additive as it should be.

Species	A	B	C
B	9	—	—
C	8	11	—
D	12	15	10

TABLE 4.6 Transformed distance matrix for the three ingroups of Table 4.5 when D is used as an outgroup.

Species	A	B
B	10/3	—
C	16/3	16/3

d'_{ij} : 종 i 와 j 간의 transformed distance

d_D : outgroup과 모든 ingroup 분류군들간의 평균거리

in this case: $d_D = (d_{AD} + d_{BD} + d_{CD})/3 = (12 + 15 + 10)/3 = 37/3$

$$d'_{AB} = (d_{AB} - d_{AD} - d_{BD})/2 + 37/3 = (9 - 12 - 15)/2 + 37/3 = -9 + (37/3) = 10/3$$

Neighbor's Relation Method 근친접근법

- UPGMA의 또다른 변형으로 모든 분지들의 길이가 최소화되도록 계통수를 작성. Neighbor와의 관계에 중점을 둔 방법
- Neighbor (근친): 오직 하나의 내부절에 의해 분리된 한쌍의 분류군들을 뜻함
- 4개의 종으로 만들 수 있는 모든 쌍의 배열 중 어떤 배열이 사점조건을 만족하는지 파악해 냄.
- 4개의 분류군을 정하고 $d_{AB}+d_{CD}$, $d_{AC}+d_{BD}$, $d_{AD}+d_{BC}$ 를 계산하여 합이 가장 적은 두개의 쌍에 1점, 나머지는 0점 부여
- 모든 조합을 반복
- 가장 높은 누적점수를 갖는 쌍이 근친이 됨
- 근친이 된 그룹을 인식하여 이를 하나의 점으로 인식.
- 이를 반복해 나감..

$$d_{AC} + d_{BD} = d_{AD} + d_{BC} = a + b + c + d + 2e = d_{AB} + d_{CD} + 2e$$

A,B가 붙고, C,D가 붙은 사점 조건은 다음을 만족.

$$d_{AB} + d_{CD} < d_{AC} + d_{BD}$$

and

$$d_{AB} + d_{CD} < d_{AD} + d_{BC}$$

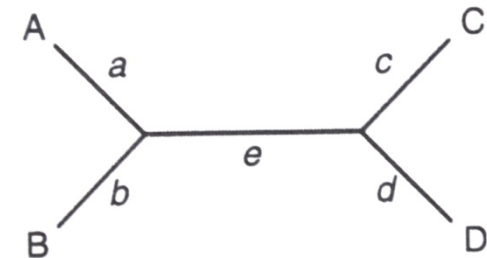


FIGURE 4.10 A generic phylogenetic tree with four species (A, B, C, and D) and each branch uniquely labeled (a, b, c, d, and e).

Neighbor Joining (NJ) Method (근친 연결법; Saitou and Nei, 1987)

- 모든 종을 한 개의 중앙 결절에서 파생시키는 별모양의 계통수로부터 시작하여 tree에 있어서 전체 branch 길이를 최소화 하는 방법임

$$S_{12} = (1/(2(N-2))(\Sigma(d_{1k} + d_{2k}) + (1/2)d_{12} + (1/N-2)(\Sigma d_{ij}))$$

S12: any pair of species can take positions 1 and 2

N: number of species

k: accepted outgroup

Dij: distance between species

→ 가장 많이 사용되는 방법

Maximum Likelihood Approaches (최대가능도법)

- Make a tree maximizing sum of substitution rate in each site.
- 최근 가장 많이 사용되는 방법
- Character based method 이후에 다시 설명

Chap. 5. Character-based Method of Phylogenetics

Parsimony: 최소가정법. 진화는 항상 그 변화 단계의 수를 최소화 하는 방향으로 일어난다는 가정하에 수행하는 계통도의 재구성. 단어의 어원은 극도로 절약하는 구두쇠를 의미하고, 단어 자체는 구두쇠의, 경제적인 등의 의미를 갖는다.

하나의 자료가 있고, 이 자료에 대하여 여러 가지 형태의 계통수를 가정할 수 있다. 이 계통수 상에서 형질 하나하나가 어떻게 변화해 가는지를 추정할 때 node에서 묶어지는 형질의 수가 많아지면 많아질 수록 변화하는 단계 (step) 수가 줄어든다. **가장 최소의 step 수를 갖는 계통수**가 바로 parsimony 방법에 의해 선택되어지는 가장 좋은 계통수이다.

Campbell Biology Chap. 26 – PARSIMONY Method



Species I



Species II



Species III

Three phylogenetic hypotheses:

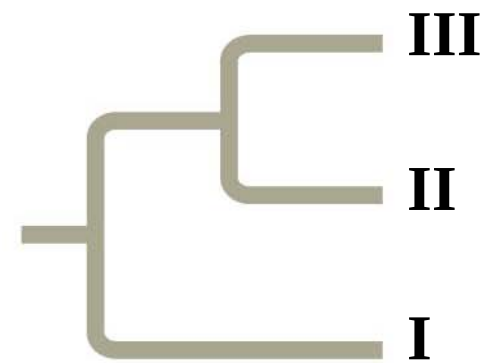
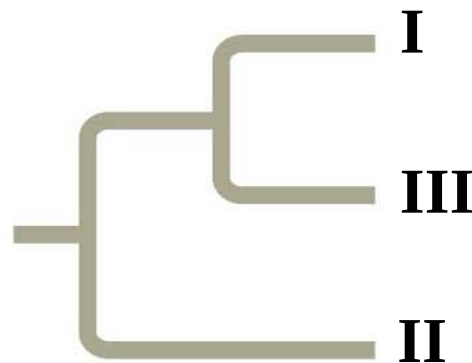
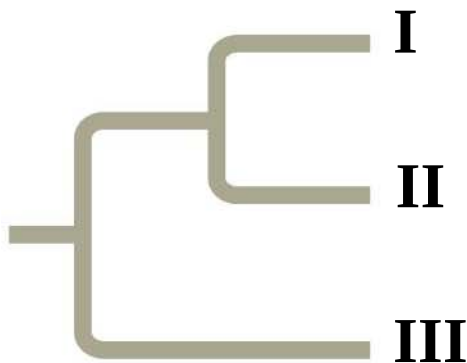


Fig. 26-15-2

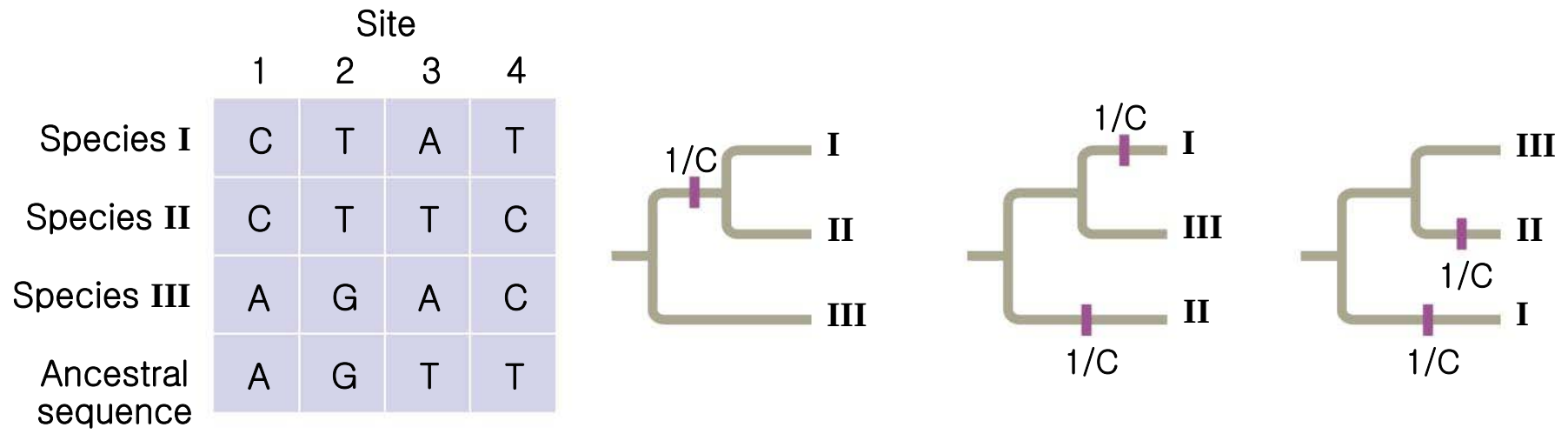


Fig. 26-15-3

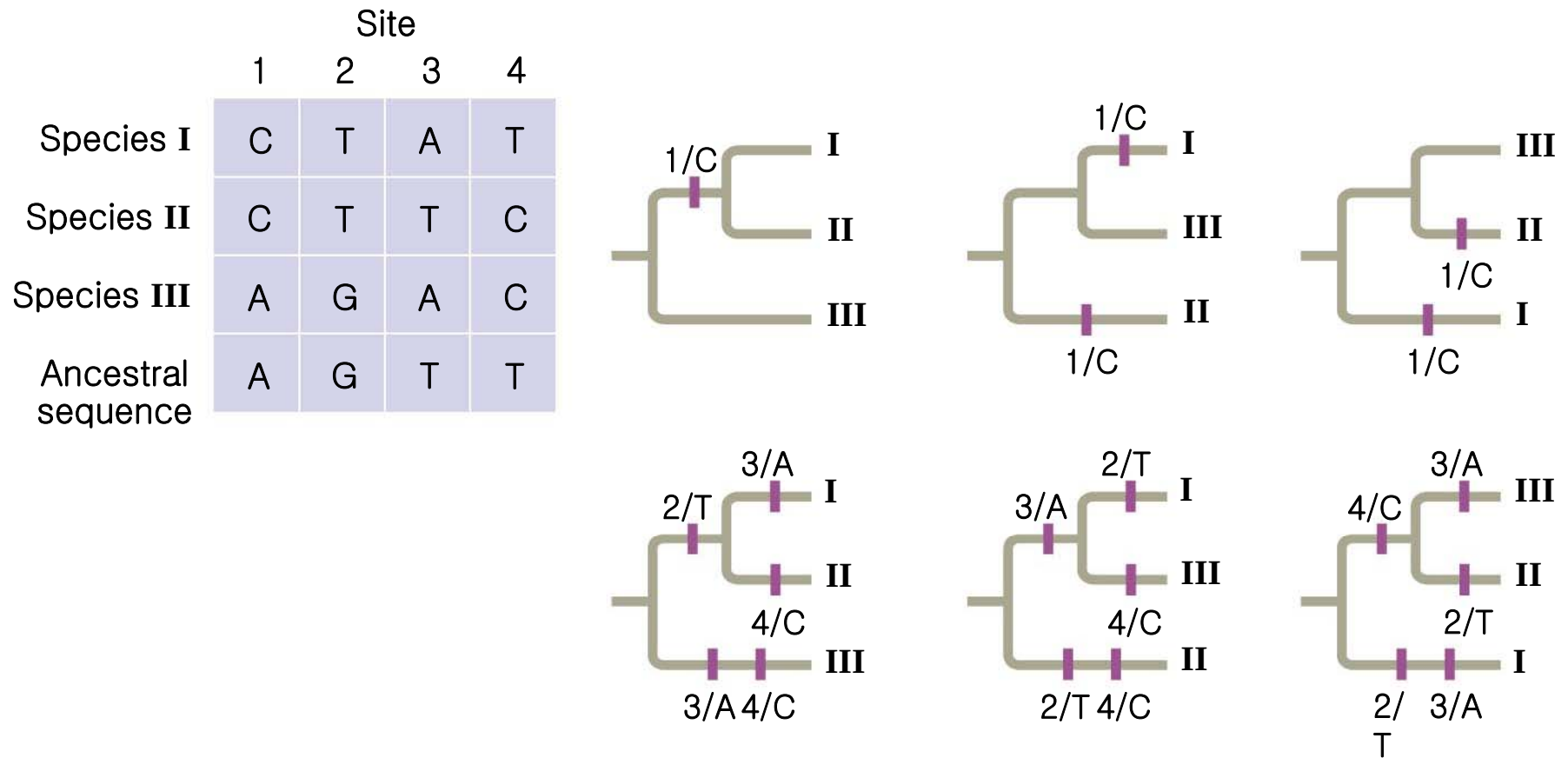
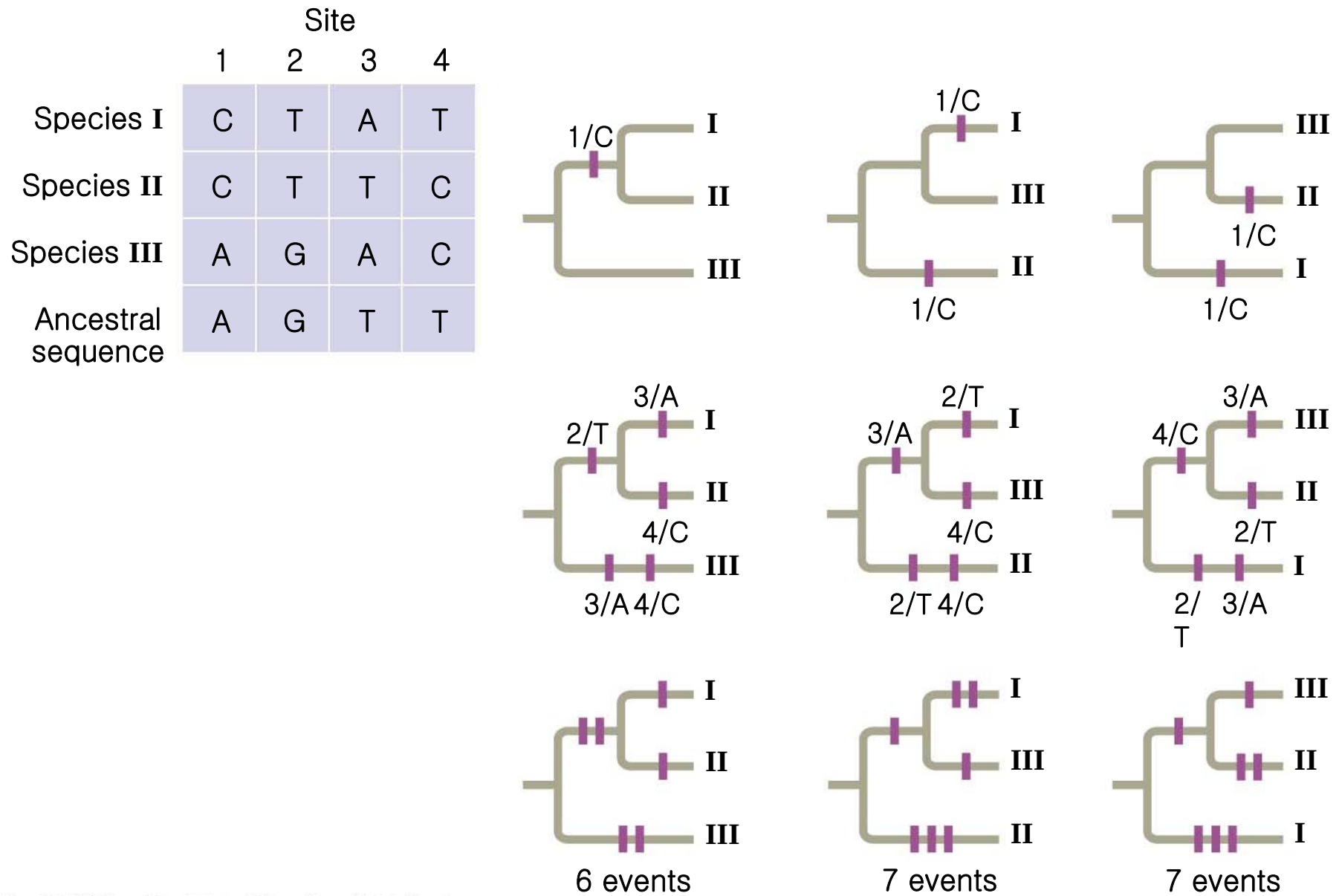


Fig. 26-15-4



STRATEGIES FOR FASTER SEARCH

1) Exhaustive search 완전한 검색

Table 2

The number of distinct, unrooted, bifurcating trees as a function of the number of taxa

Number of taxa	Number of trees
10	2×10^6
22	3×10^{23}
50	3×10^{74}
100	2×10^{182}
1,000	$2 \times 10^{2,860}$
10,000	$8 \times 10^{38,658}$
100,000	$1 \times 10^{486,663}$
1,000,000	$1 \times 10^{5,866,723}$
10,000,000	$5 \times 10^{68,667,340}$

전체 가능한 모든 tree topology들에 대하여 모두 step 수를 계산하여 가장 좋은 tree를 구해 낸다. 분류군이 10개만 넘어가도 거의 불가능한 계산이 되어 버려 실질적으로는 적용하기 힘들다.

2) Branch-and-bound method 가지-경계법

Exhaustive search 보다는 효율적이지만 여전히 20 분류군 이하에서 사용하는 방법이다.

Calculation of 20 taxa:
about 10^{21} trees

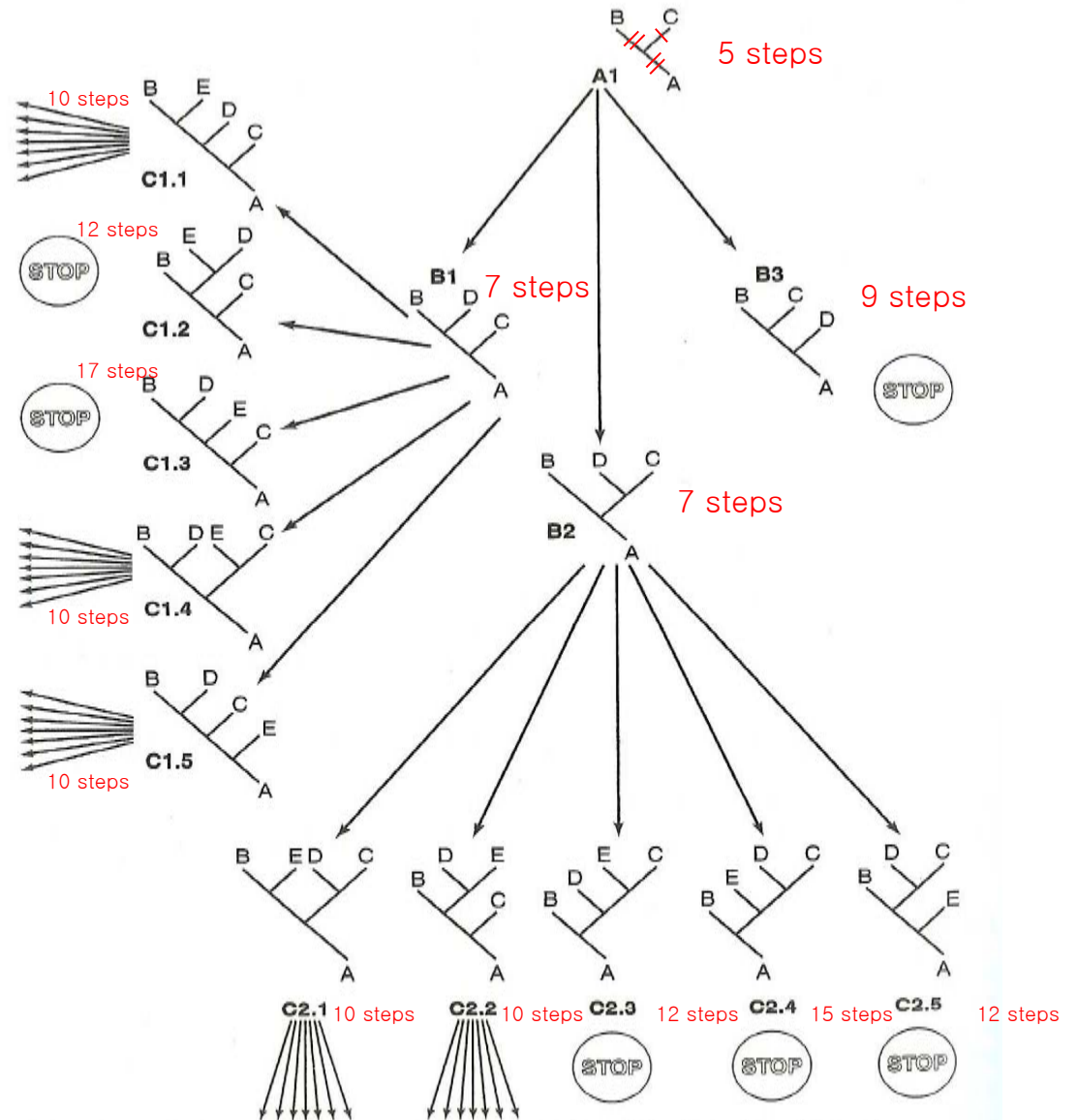


FIGURE 5.4 A diagrammatic representation of the branch and bound approach for maximum parsimony analysis. Only one tree can describe the relationship between three sequences (A, B, and C) and it is used as a starting point (tree A1). A fourth sequence (D) from the data set can be added to tree 1 in only three different ways to give rise to trees B1, B2, and B3. If the length of one of those trees is greater than that of the upper bound, then that tree is discarded and no additional branches are added to it in subsequent steps. An additional (fifth) branch is then added to all remaining trees and the process is repeated until a branch has been added for all

3) Heuristic search 발견적 방법

- 임의의 계통수를 구성 후 이를 시작점으로 보다 tree를 조금씩 변화시켜 더욱 짧은 계통수를 찾아나감.
- 실질적으로 가장 많이 쓰이는 방법으로 가장 효율적이다.
- 하지만 정밀한 조사를 하는 것이 아님으로 오류가 있을 수가 있다. 그러므로 분석을 한번만 하는 것이 아니고 여러가지 다른 최초 계통수를 사용하여 다수의 분석을 수행한 후 가장 적은 step 수를 갖는 계통수를 찾아낸다.

여러가지 Tree를 약간씩 변화시키는 방법으로는 branch swapping (분지교환) 방법이 가장 많이 쓰인다.

Branch swapping (분지교환): 임의의 tree를 구성하고 일정수의 분류군을 포함하는 가지를 잘라 다른 가지에 붙인 후 step 수를 조사한다. 이를 반복하여 가장 적은 step수의 tree를 찾아나간다.

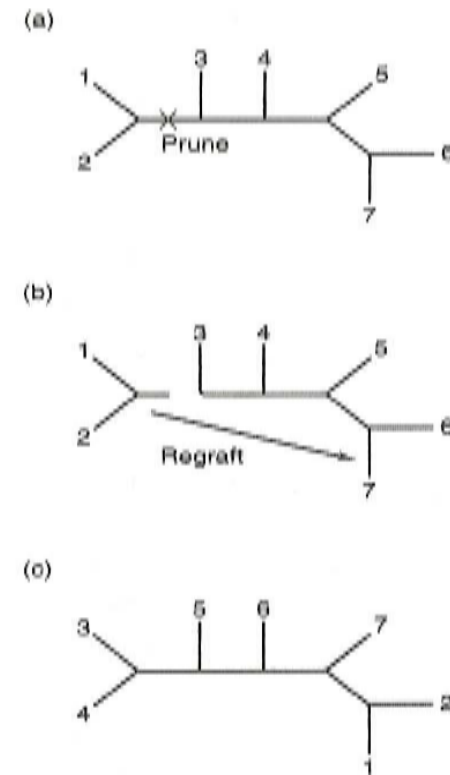


FIGURE 5.5 An example of branch swapping for an unrooted tree that describes the relationship between seven sequences labeled 1 through 7. The branch containing sequences 1 and 2 is randomly chosen for pruning and then randomly placed between terminal nodes 6 and 7 to generate a new tree with a topology that is similar to the original.

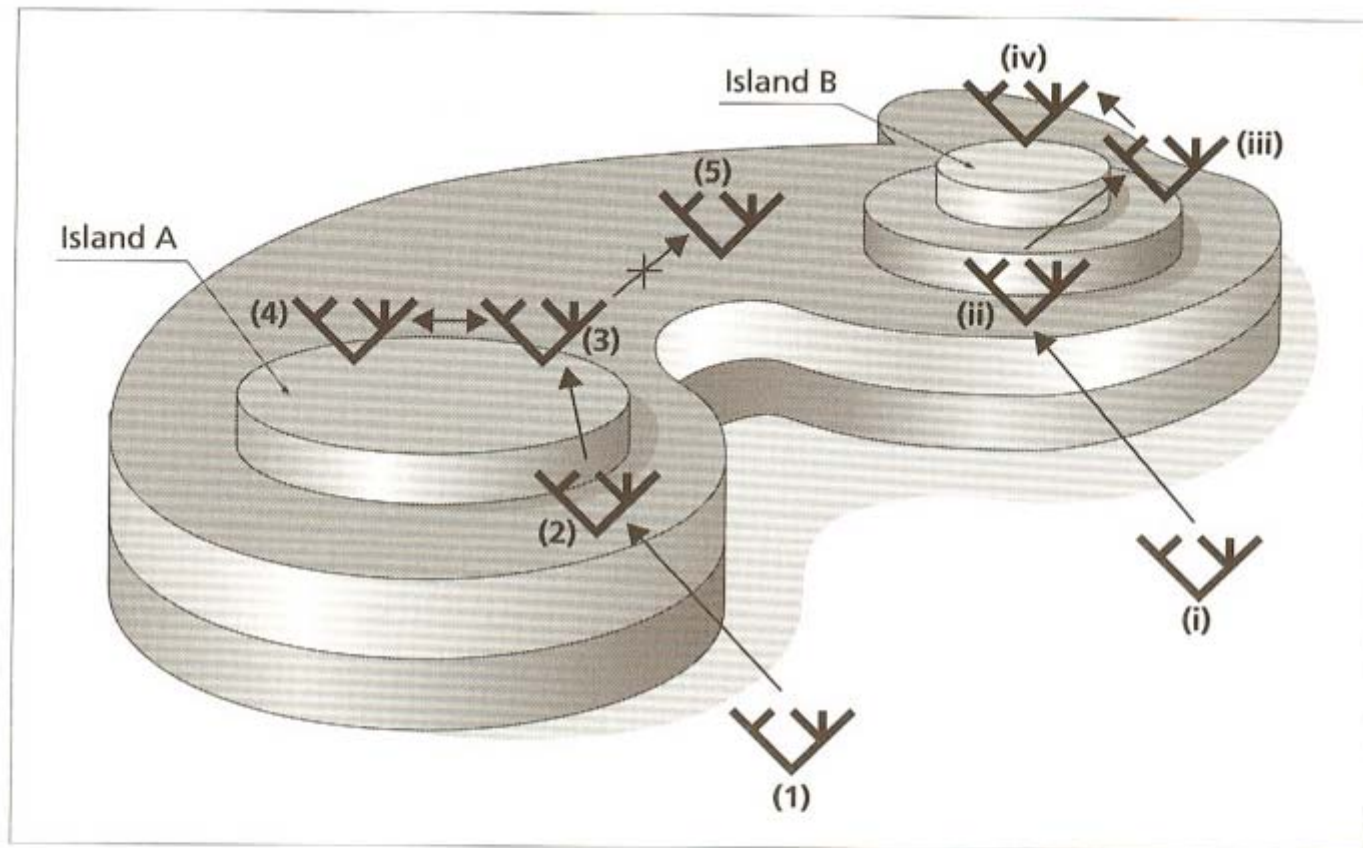


Fig. 6.4 Landscapes and the problem of islands of trees (locally optimal sets of trees). A hill climbing algorithm that started from tree 1 would succeed in finding trees 3 and 4 (which comprise island a), but would fail to discover that tree iv (island b) was even better, because to get to that tree it would have to cross a plateau of trees that were worse than trees 3 and 4. However, a search starting from tree i would succeed in finding the best tree. If the set of possible trees contains more than one island then heuristic methods may land on a suboptimal island, and the optimal island will not be discovered. After Maddison (1991).

참조: 4) Star Decomposition method (성상해체법)

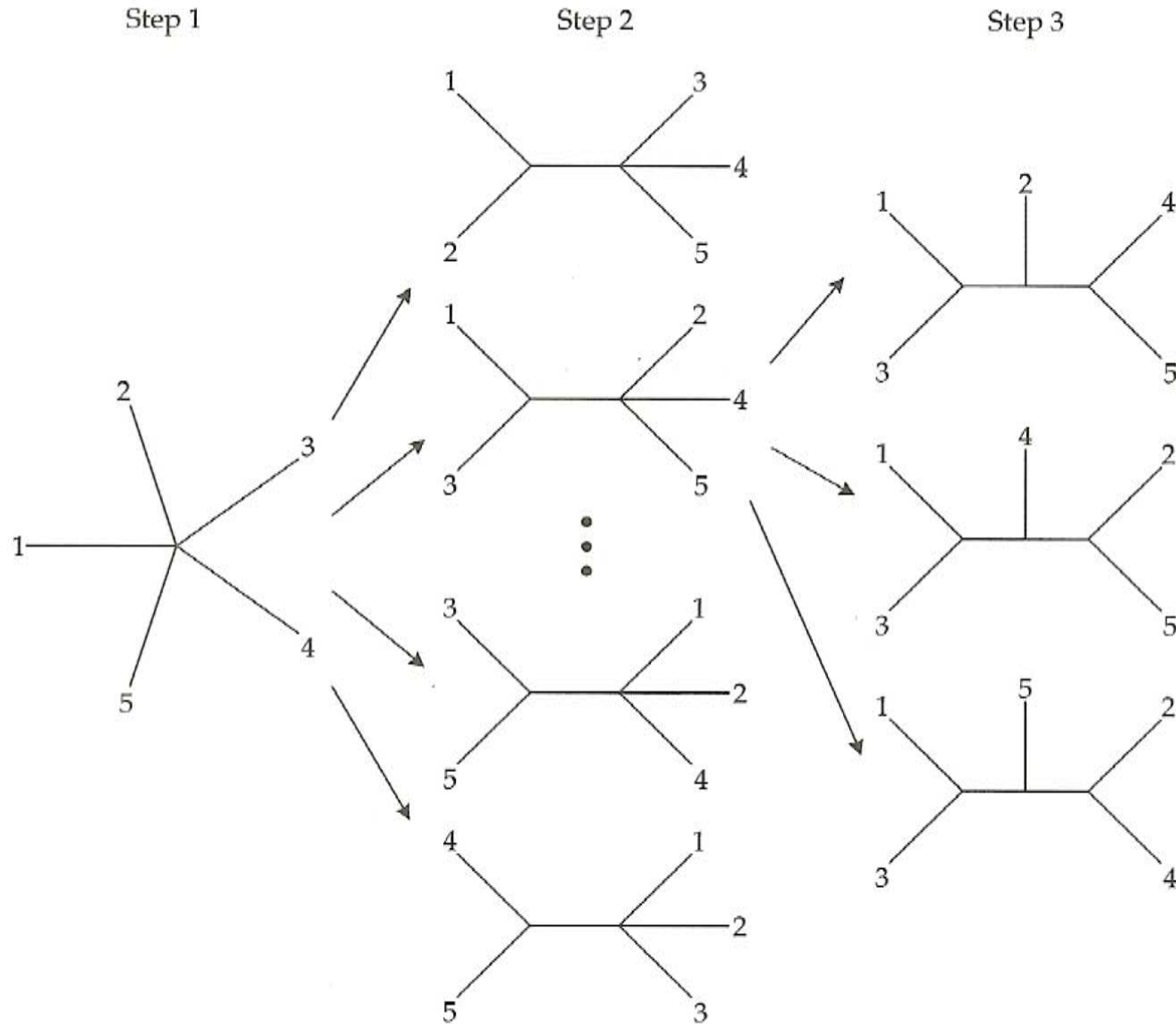


Figure 25 Heuristic tree selection using star decomposition method. At each step, the optimality criterion evaluated for each possible joining of a pair of lin-

edges leading away from the central node. The best tree found during each step becomes the starting point for the next step.