Inferencia Filogenética

Types of Data

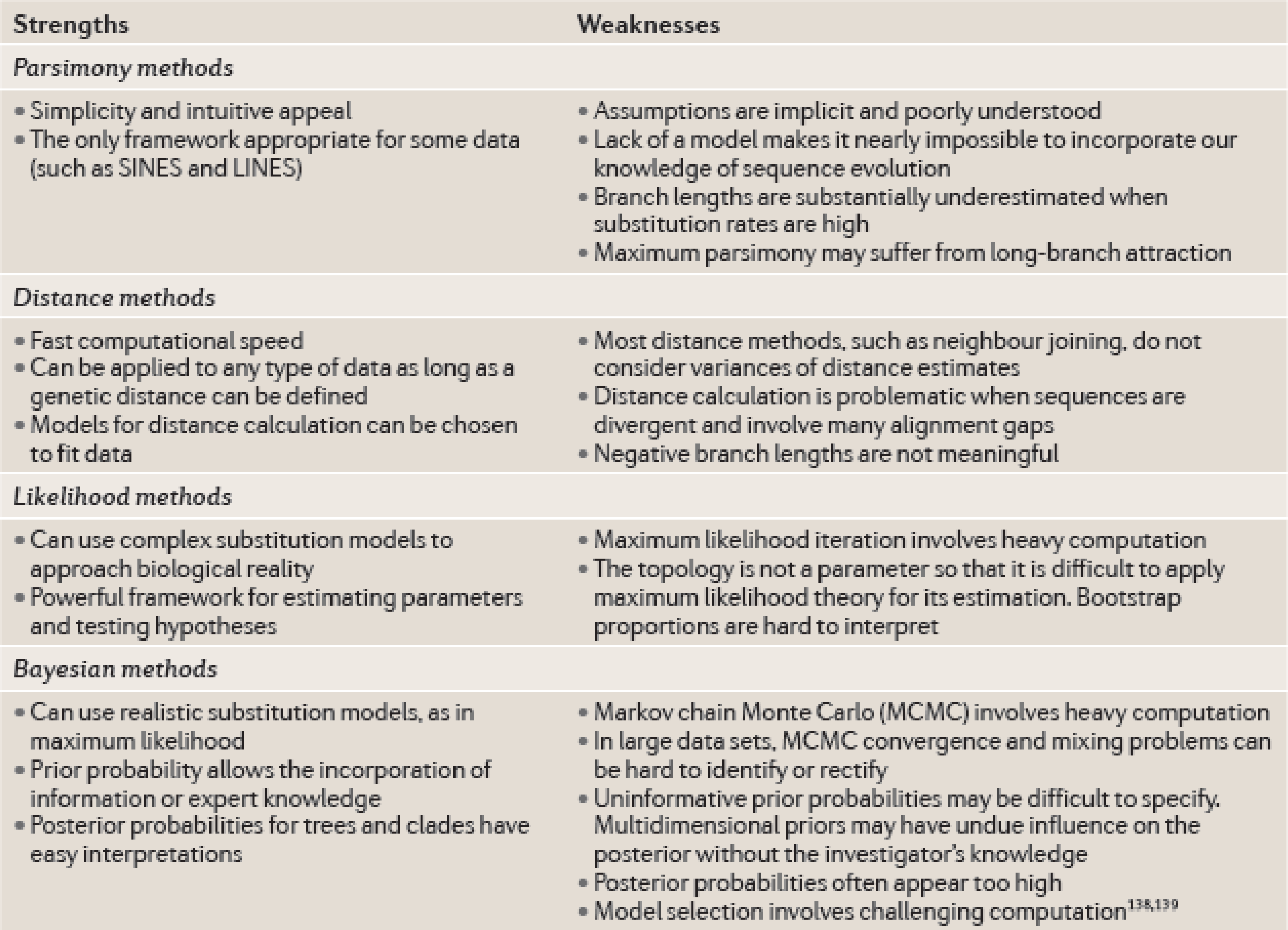
Tree-building method

	Distances	Nucleotide sites
Clustering algorithm	UPGMA neighbour-joining	
Optimality criterion	Minimum evolution	Maximum parsimony Maximum likelihood

Classification of phylogenetic analysis methods

	Distances	Character state
Clustering algorithm	UPGMA neighbour-joining	
Optimality criterion	Minimum evolution	Maximum parsimony Maximum likelihood Bayesian Inference

Summary of strengths and weaknesses



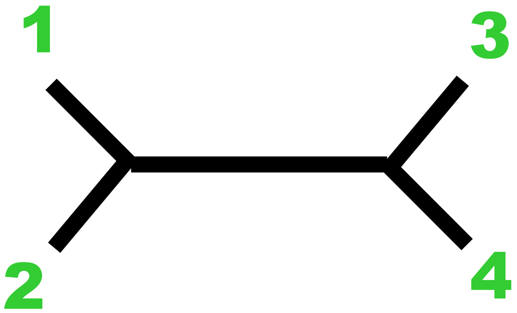
Strengths	Weaknesses
Parsimony methods	
• Simplicity and intuitive appeal • The only framework appropriate for some data (such as SINES and LINES)	• Assumptions are implicit and poorly understood • Lack of a model makes it nearly impossible to incorporate our knowledge of sequence evolution • Branch lengths are substantially underestimated when substitution rates are high • Maximum parsimony may suffer from long-branch attraction
Distance methods	
• Fast computational speed • Can be applied to any type of data as long as a genetic distance can be defined • Models for distance calculation can be chosen to fit data	• Most distance methods, such as neighbour joining, do not consider variances of distance estimates • Distance calculation is problematic when sequences are divergent and involve many alignment gaps • Negative branch lengths are not meaningful
Likelihood methods	
• Can use complex substitution models to approach biological reality • Powerful framework for estimating parameters and testing hypotheses	• Maximum likelihood iteration involves heavy computation • The topology is not a parameter so that it is difficult to apply maximum likelihood theory for its estimation. Bootstrap proportions are hard to interpret
Bayesian methods	
• Can use realistic substitution models, as in maximum likelihood • Prior probability allows the incorporation of information or expert knowledge • Posterior probabilities for trees and clades have easy interpretations	• Markov chain Monte Carlo (MCMC) involves heavy computation • In large data sets, MCMC convergence and mixing problems can be hard to identify or rectify • Uninformative prior probabilities may be difficult to specify. Multidimensional priors may have undue influence on the posterior without the investigator's knowledge • Posterior probabilities often appear too high • Model selection involves challenging computation[138,139]

Yang and Rannala NGR2012

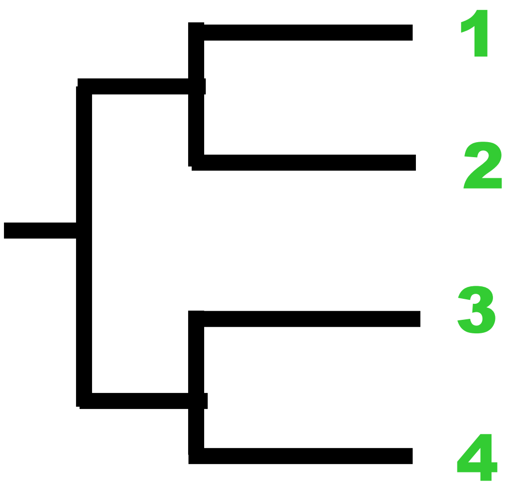
Inferencia Filogenética

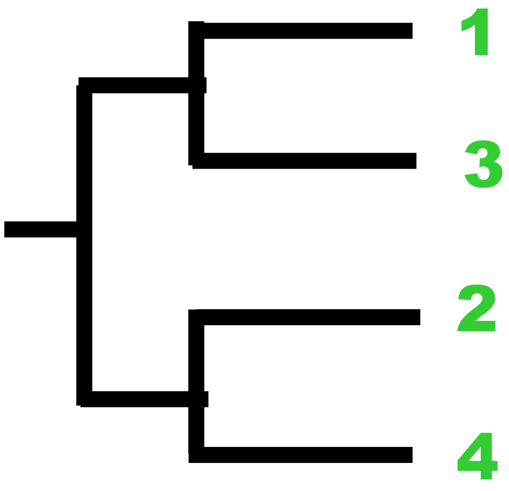
		Types of Data	
		Distances	Nucleotide sites
Tree-building method	Clustering algorithm	UPGMA neighbour-joining	
	Optimality criterion	Minimum evolution	Maximum parsimony Maximum likelihood

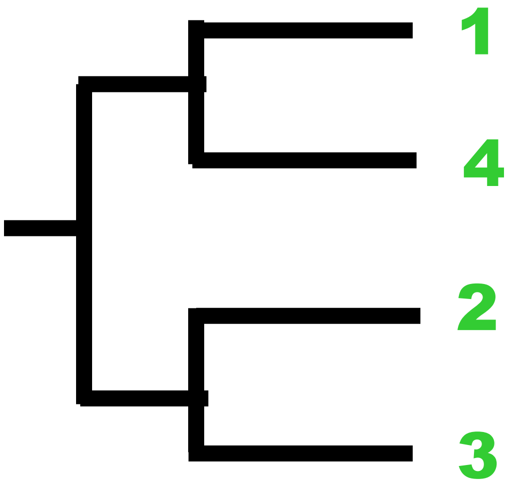
Máxima parsimonia



site	1	2	3	4	5
Species 1-	A	T	A	T	T
Species 2-	A	T	C	G	T
Species 3-	G	C	A	G	T
Species 4-	G	C	C	G	T



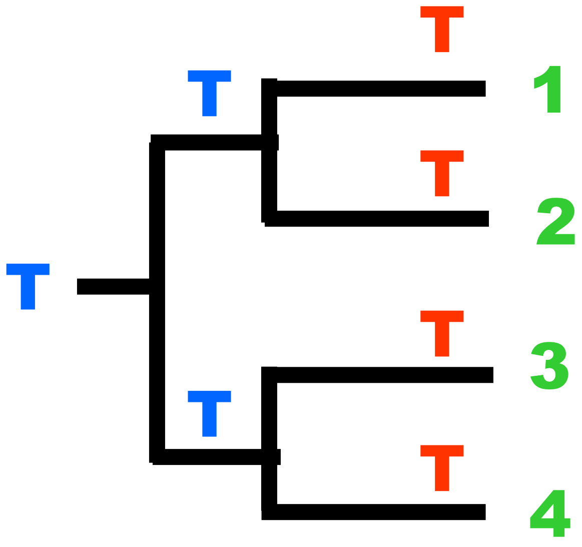




Maxima parsimonia –site 5

site	1	2	3	4	5
Species 1-	A	T	A	T	T
Species 2-	A	T	C	G	T
Species 3-	G	C	A	G	T
Species 4-	G	C	C	G	T

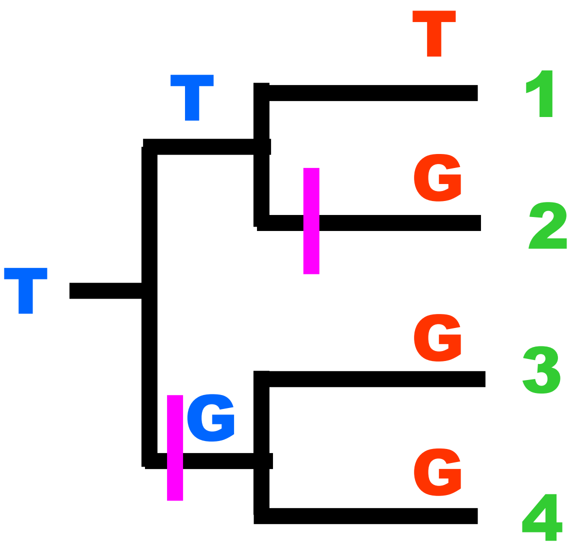
Site 5 (0)



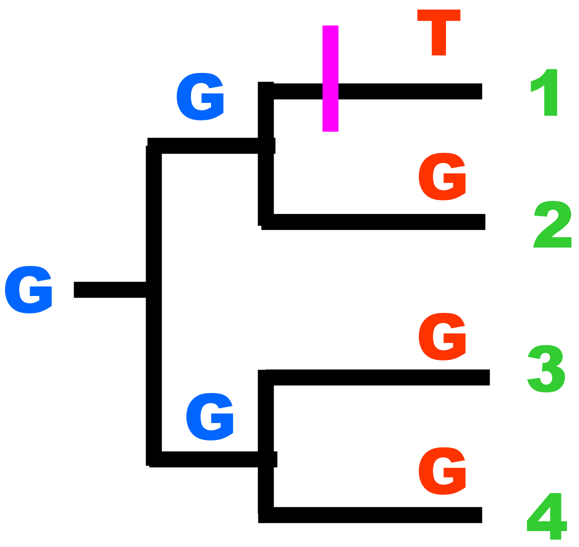
Maxima parsimonia –site 4

site	1	2	3	4	5
Species 1-	A	T	A	T	T
Species 2-	A	T	C	G	T
Species 3-	G	C	A	G	T
Species 4-	G	C	C	G	T

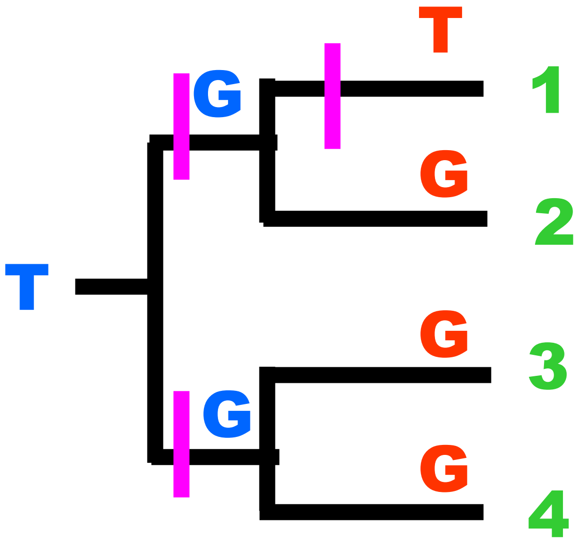
Site 4 (2)



Site 4 (1)



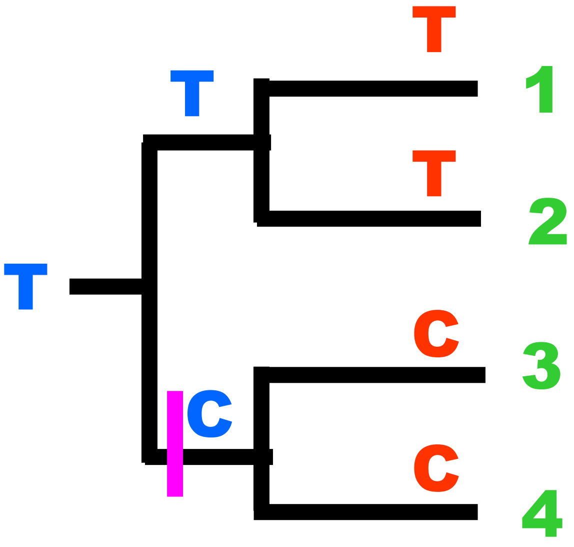
Site 4 (3)



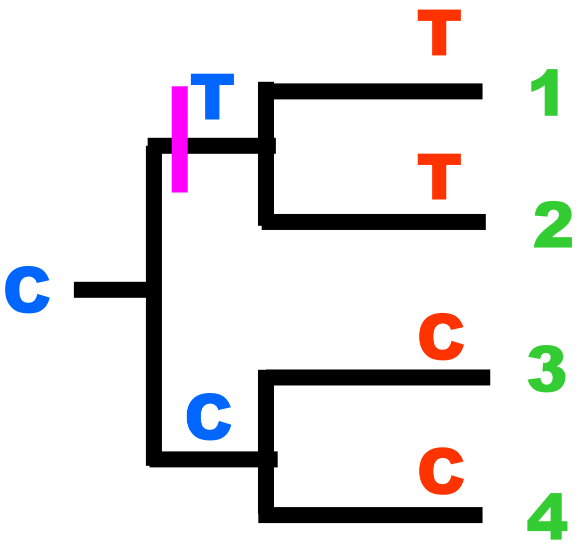
Maxima parsimonia –site 2

site	1	2	3	4	5
Species 1-	A	T	A	T	T
Species 2-	A	T	C	G	T
Species 3-	G	C	A	G	T
Species 4-	G	C	C	G	T

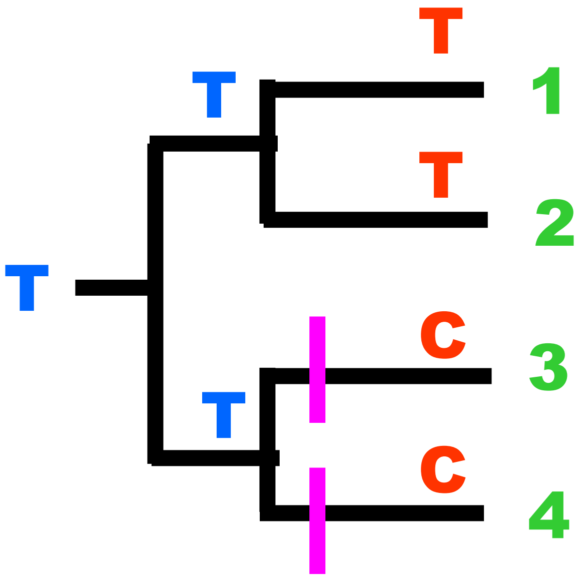
Site 2 (1)



Site 2 (1)



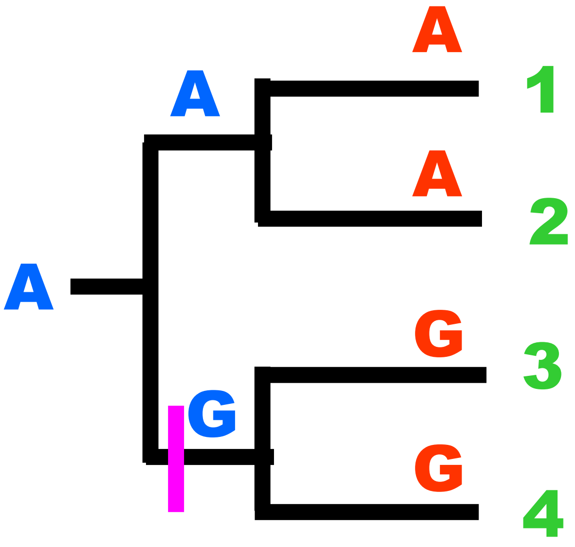
Site 2 (2)



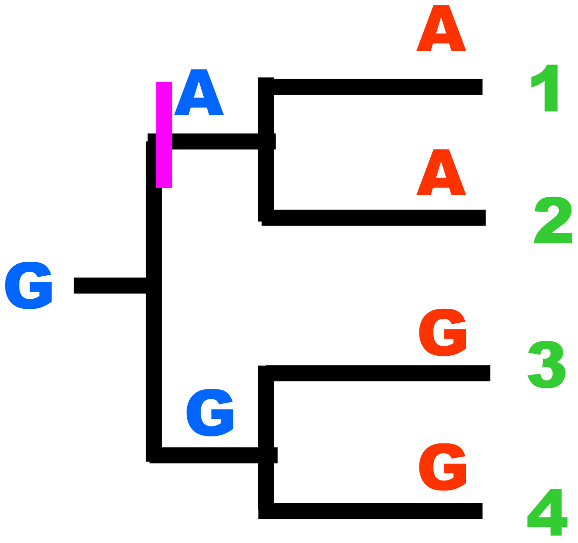
Maxima parsimonia –site 1

site	1	2	3	4	5
Species 1-	A	T	A	T	T
Species 2-	A	T	C	G	T
Species 3-	G	C	A	G	T
Species 4-	G	C	C	G	T

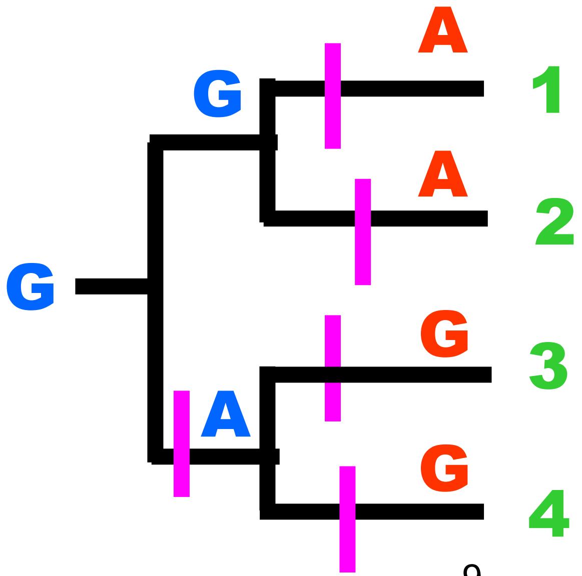
Site 1 (1)



Site 1 (1)



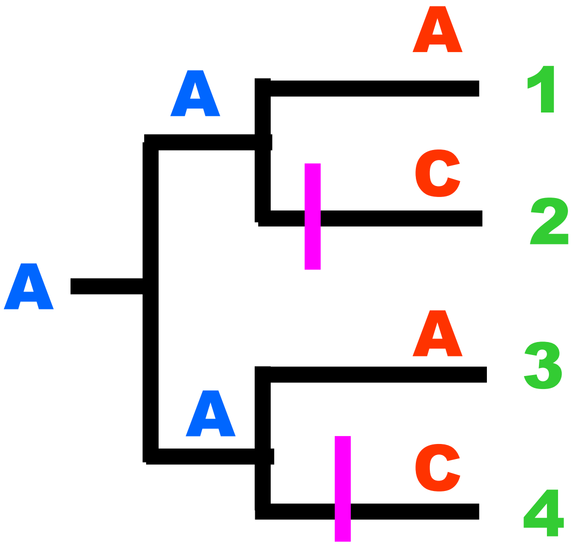
Site 1 (5)



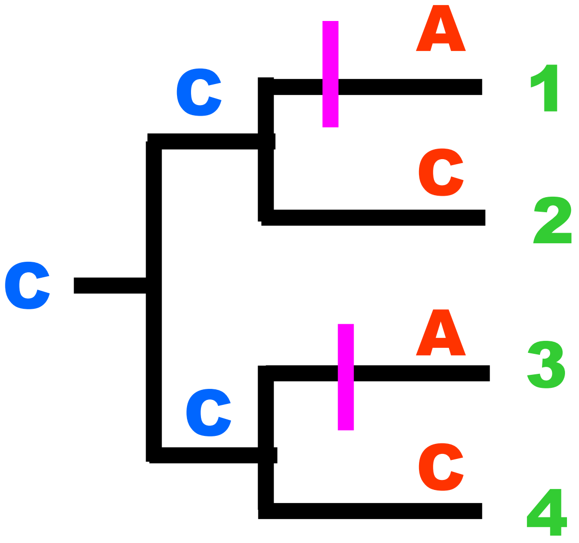
Maxima parsimonia –site 3

site	1	2	3	4	5
Species 1-	A	T	A	T	T
Species 2-	A	T	C	G	T
Species 3-	G	C	A	G	T
Species 4-	G	C	C	G	T

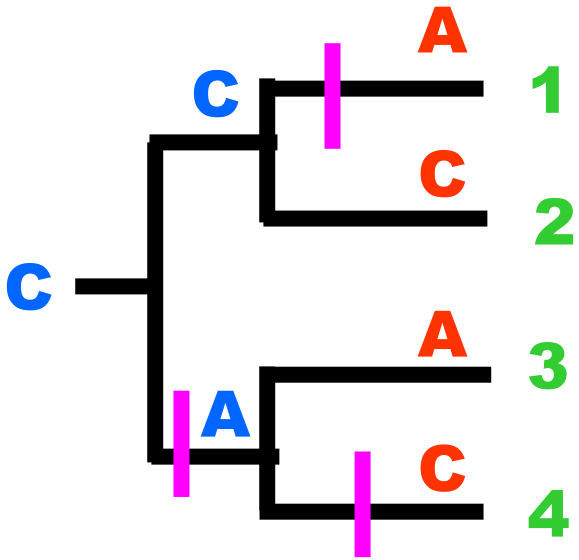
Site 3 (2)



Site 3 (2)



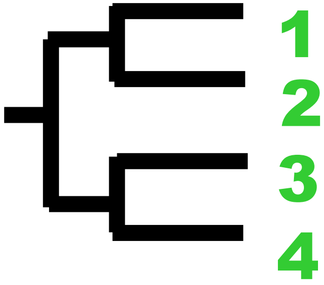
Site 3 (3)

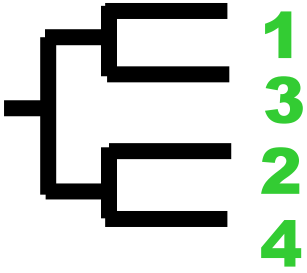


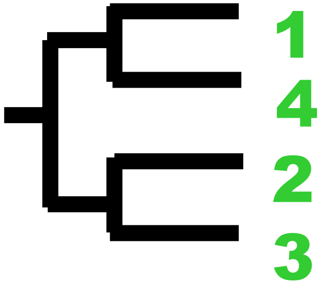
Maxima parsimonia

Total tree lengh ---> L=Σli from i=1 to k (nucleotide number)

site	1	2	3	4	5
Species 1-	A	T	A	T	T
Species 2-	A	T	C	G	T
Species 3-	G	C	A	G	T
Species 4-	G	C	C	G	T



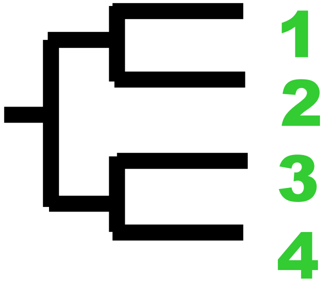




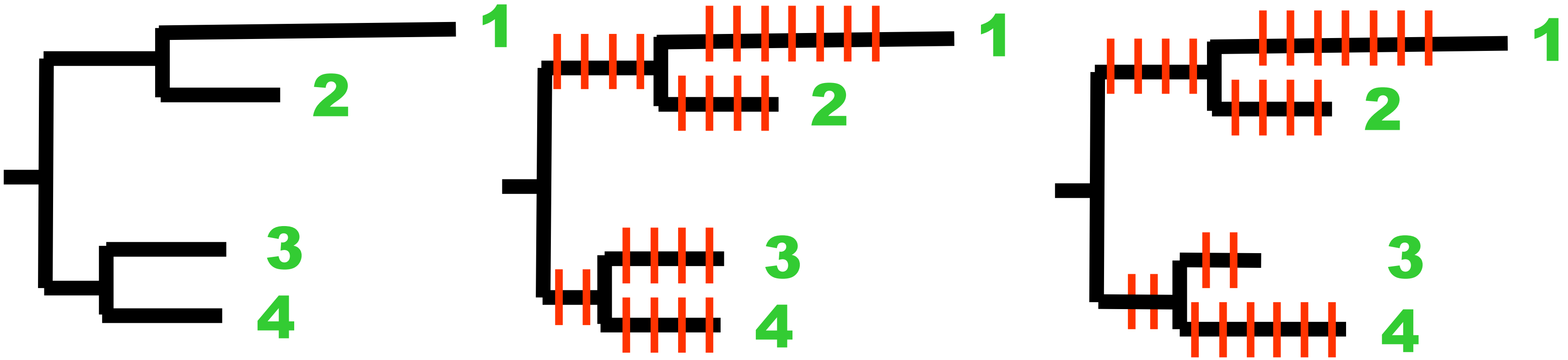
Sites	1	2	3	4	5	total
Steps	1	1	2	1	0	5
Steps	2	2	1	1	0	6
Steps	2	2	2	1	0	7

Branch lenght

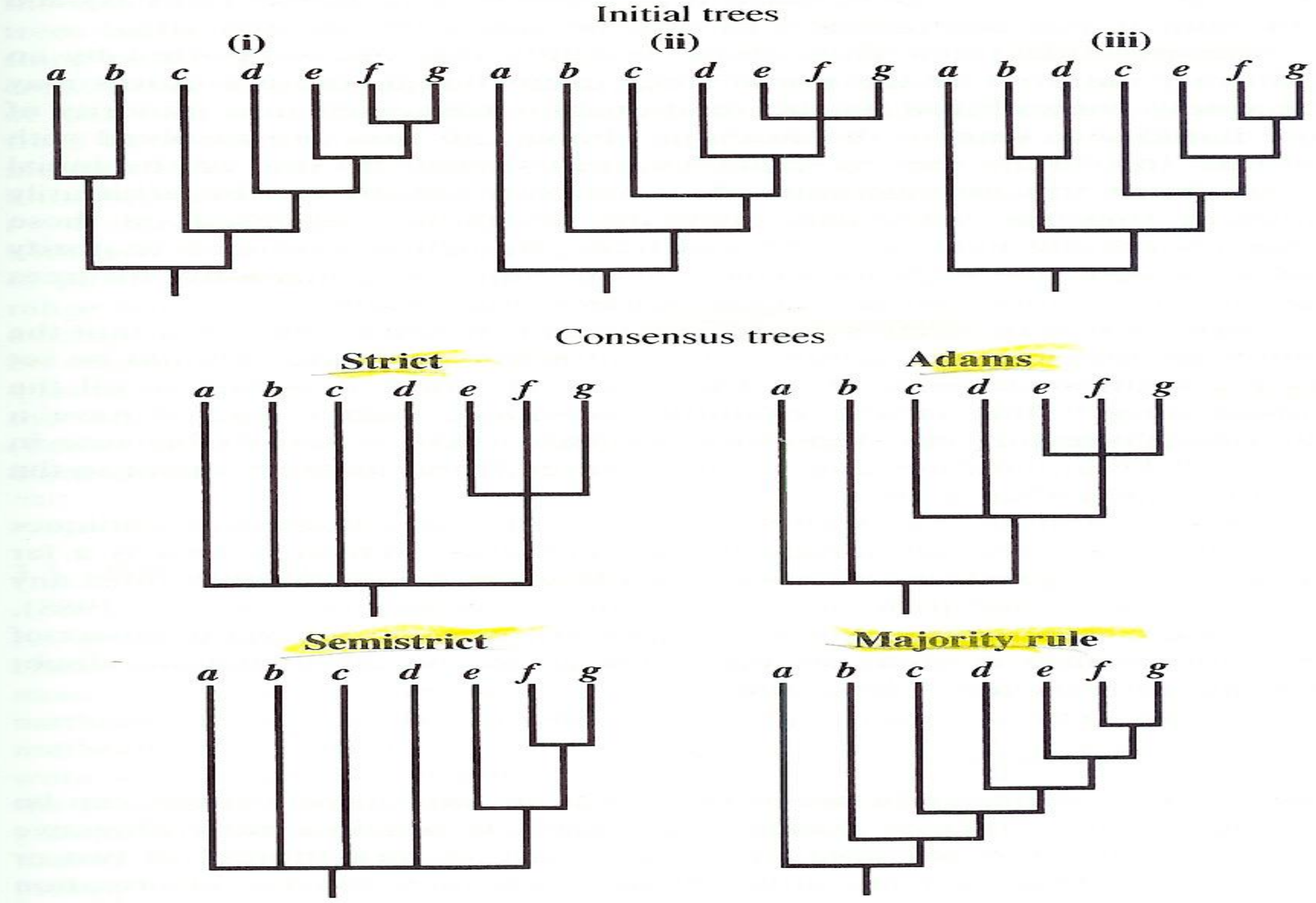
Total tree lengh ---> L=Σli from i=1 to k (nucleotide number)



Sites	1	2	3	4	5	....	N	total
Steps	1	1	2	1	0	....		25



Árvores de Consenso



Árvores de Consenso

Strict – all groups that occur on all trees

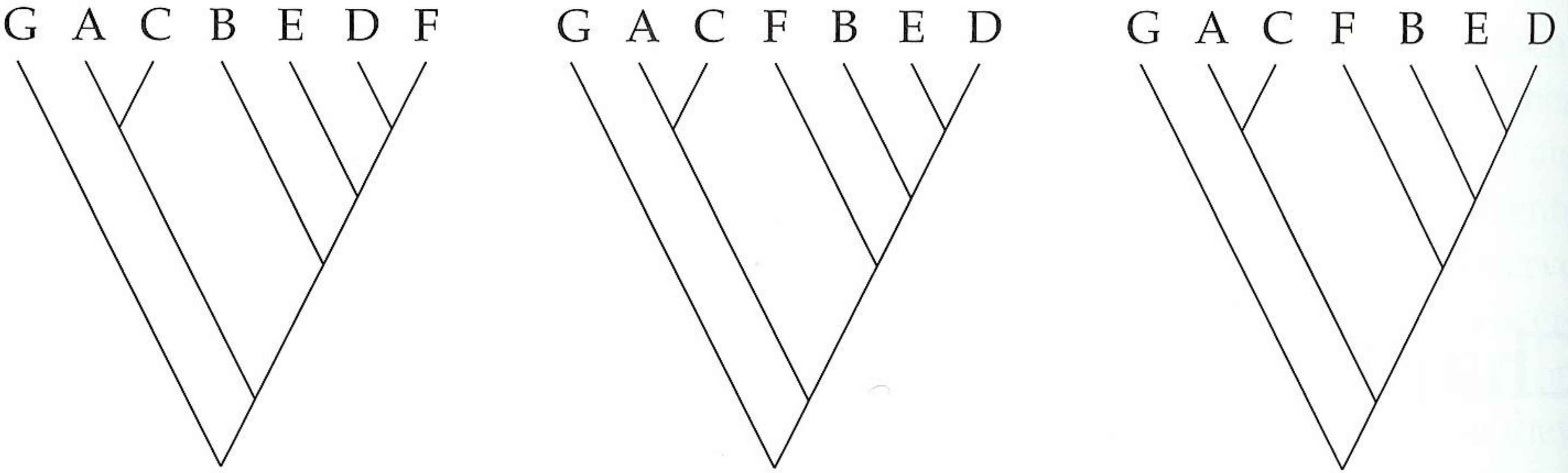
Majority Rule – all groups that occur on 50-100% of the trees
M100=strict

Semistrict – features that are resolved in all the initial trees
or are resolved in some of the initial trees and not contradicted
In the others

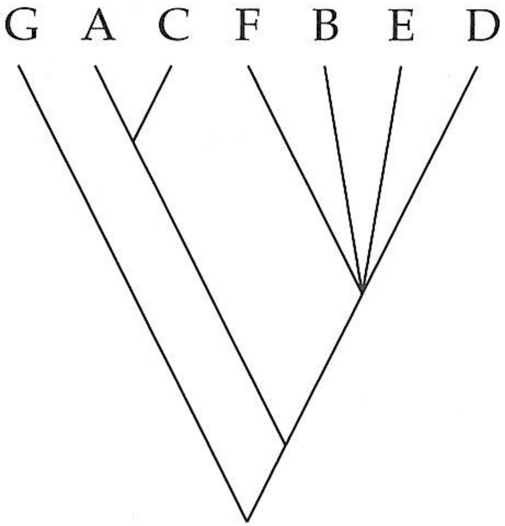
Adams - Commom ancestor of a group of taxa in a consensus
tree should be set at the furthest distance from the origin at
which it occurs in all the initial trees

Other methods

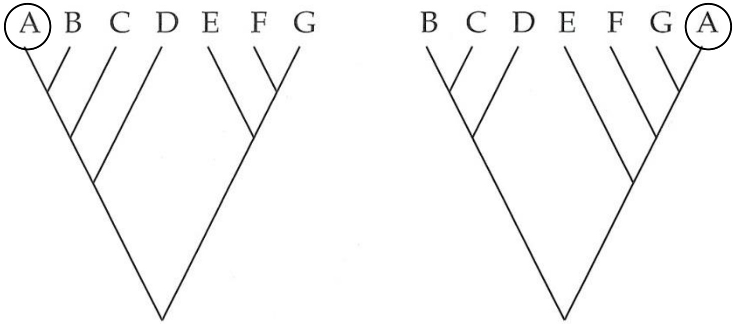
Árvores de Consenso - strict



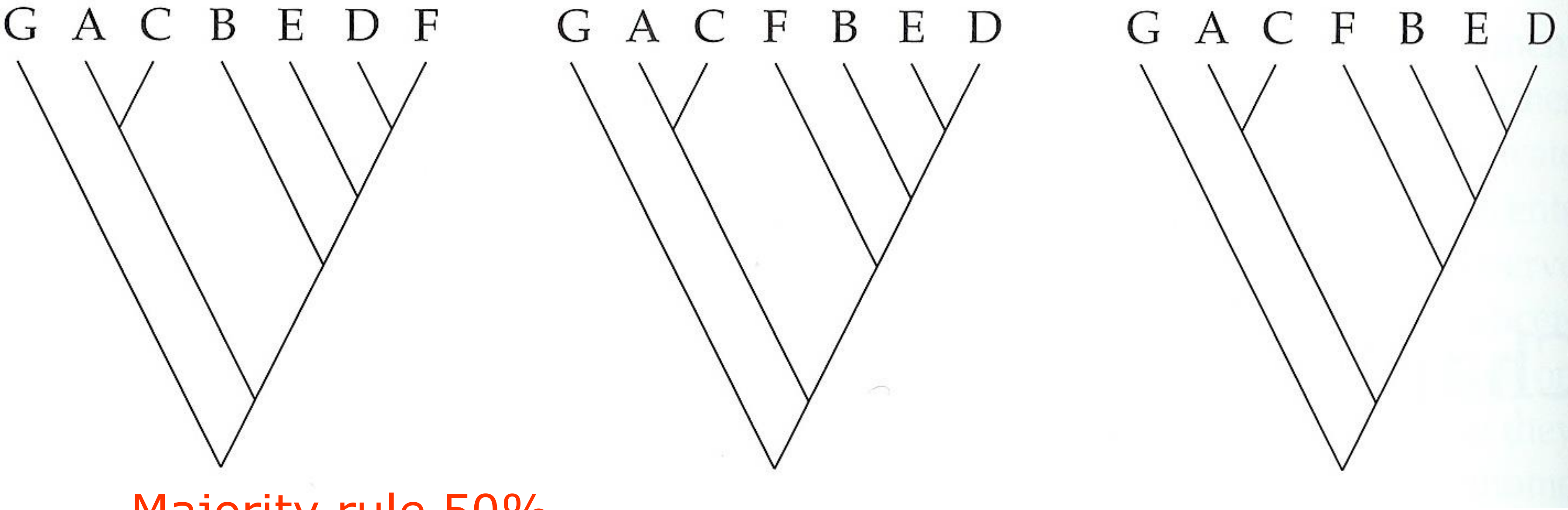
strict



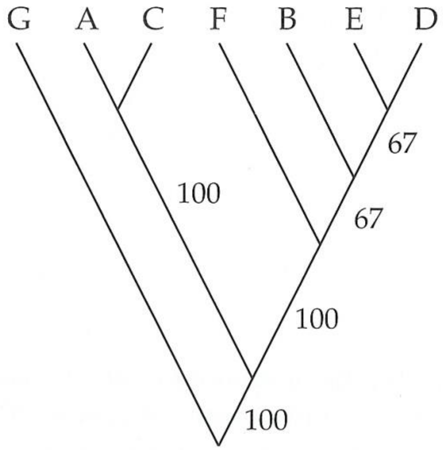
problem



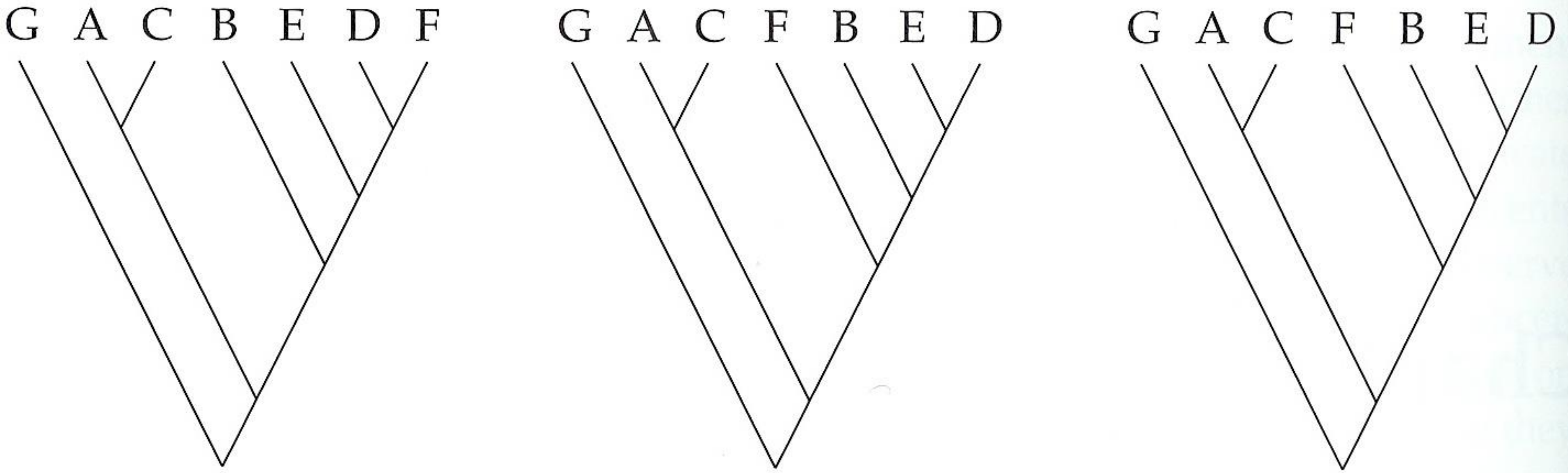
Árvores de Consenso – Majority rule



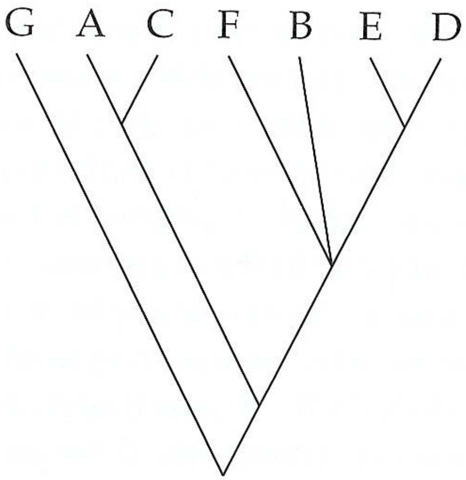
Majority rule 50%



Árvores de Consenso – Adams

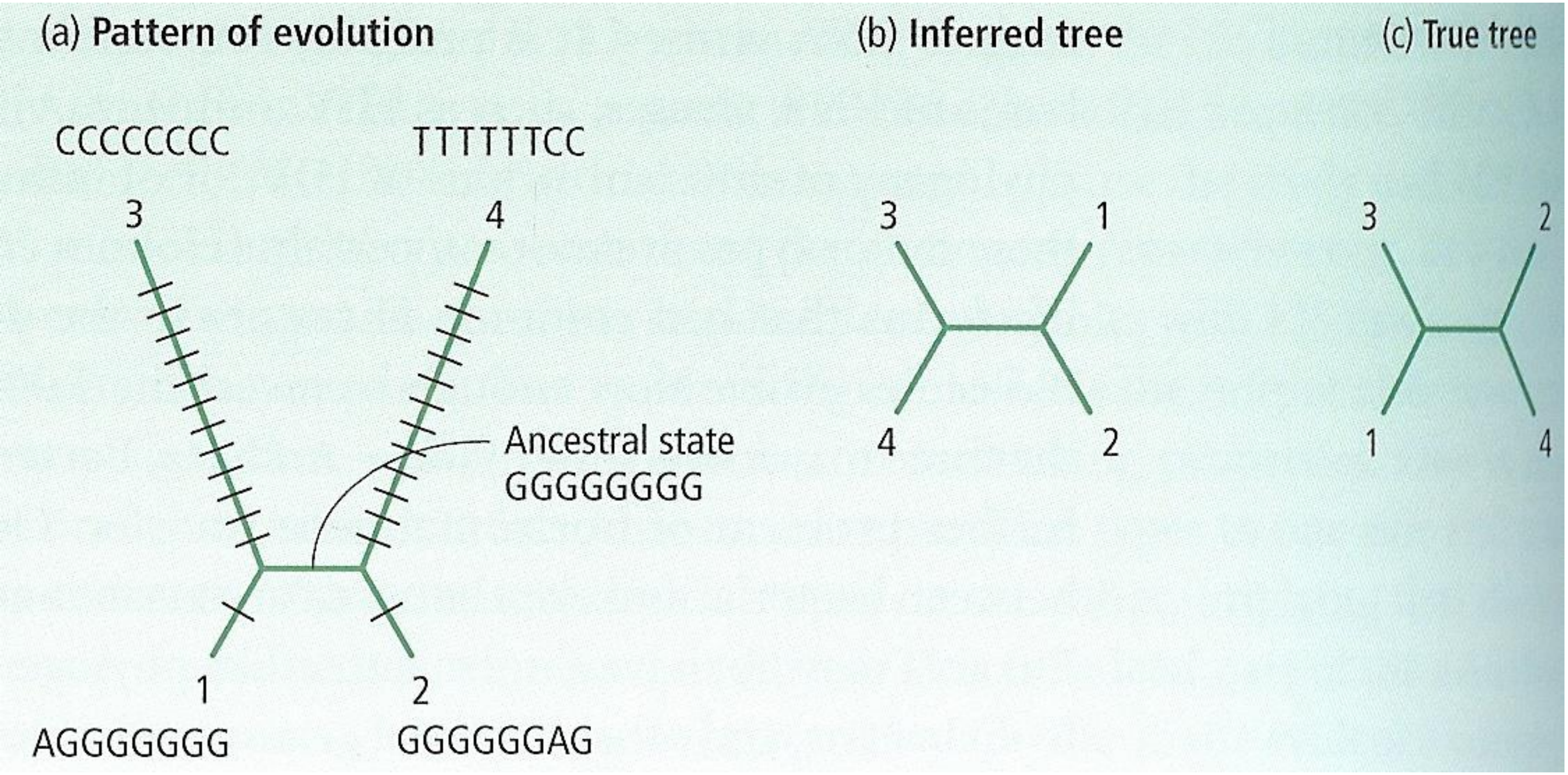


Adams



problem

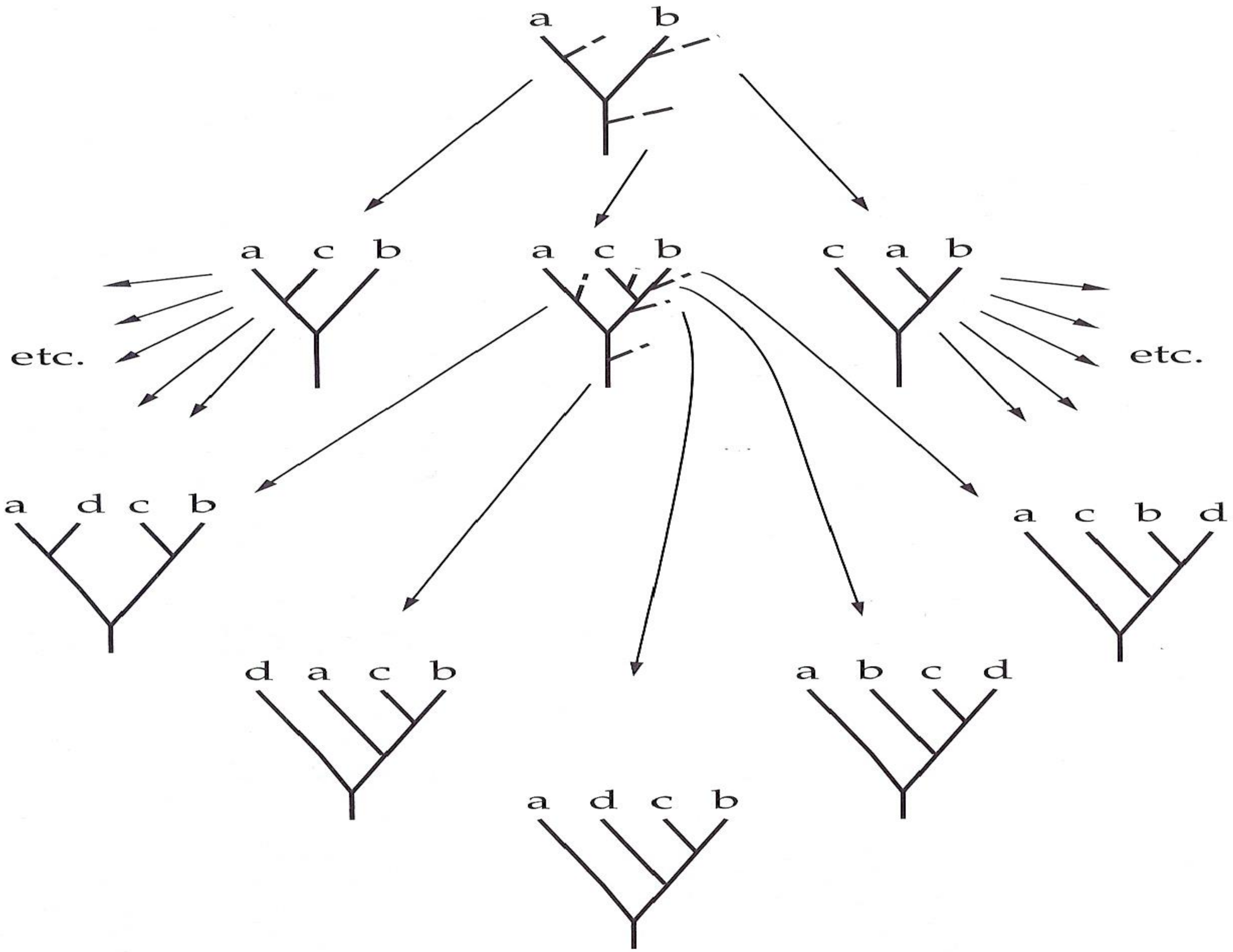
Long branch attraction



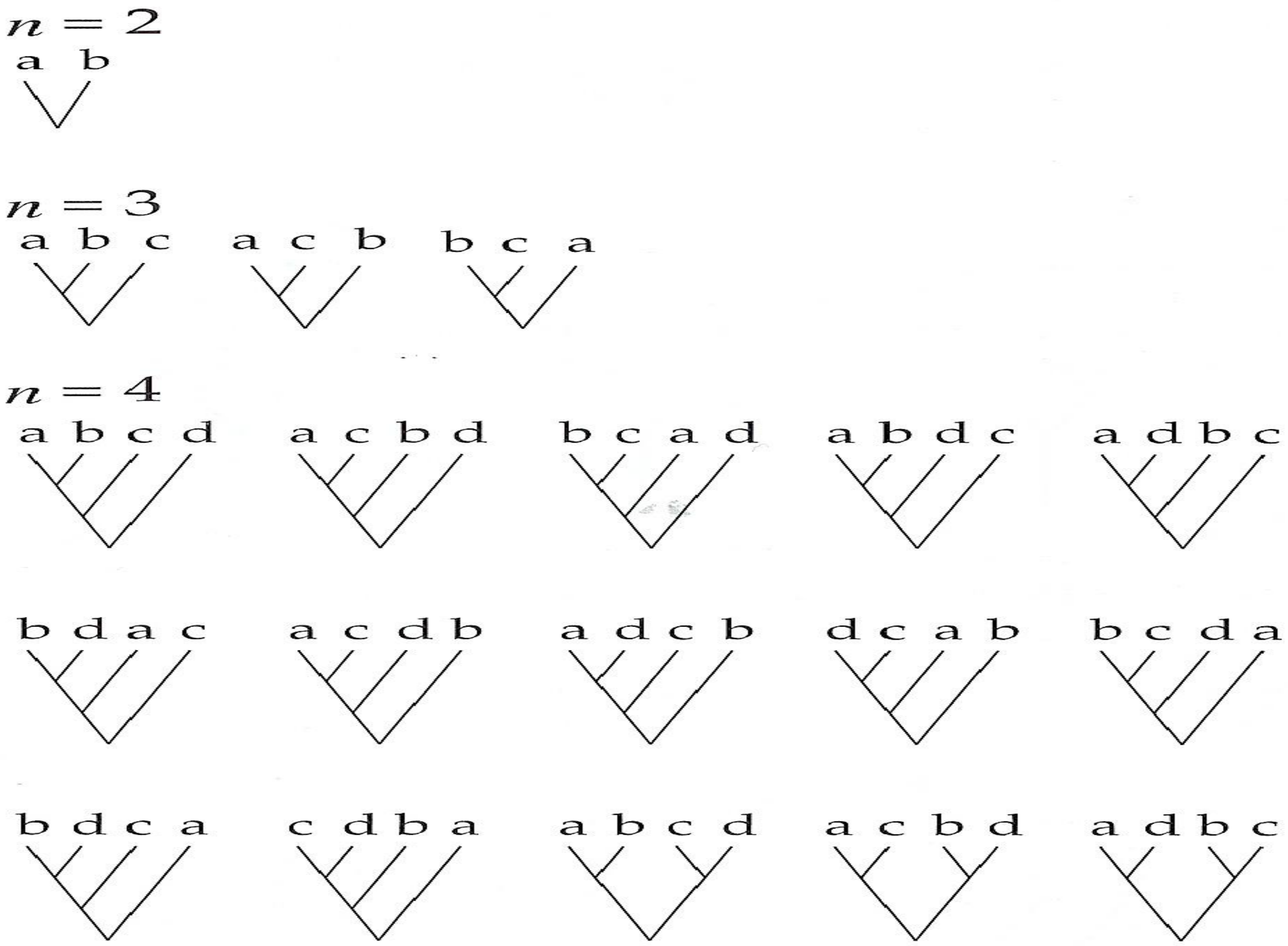
Bremer support or Decay Index for parsimony

If the most parsimonious tree that had the group ABC had 138 changes of state, and the most parsimonious tree that lacked that group had 143 changes the Decay index for that group is 143-138=5

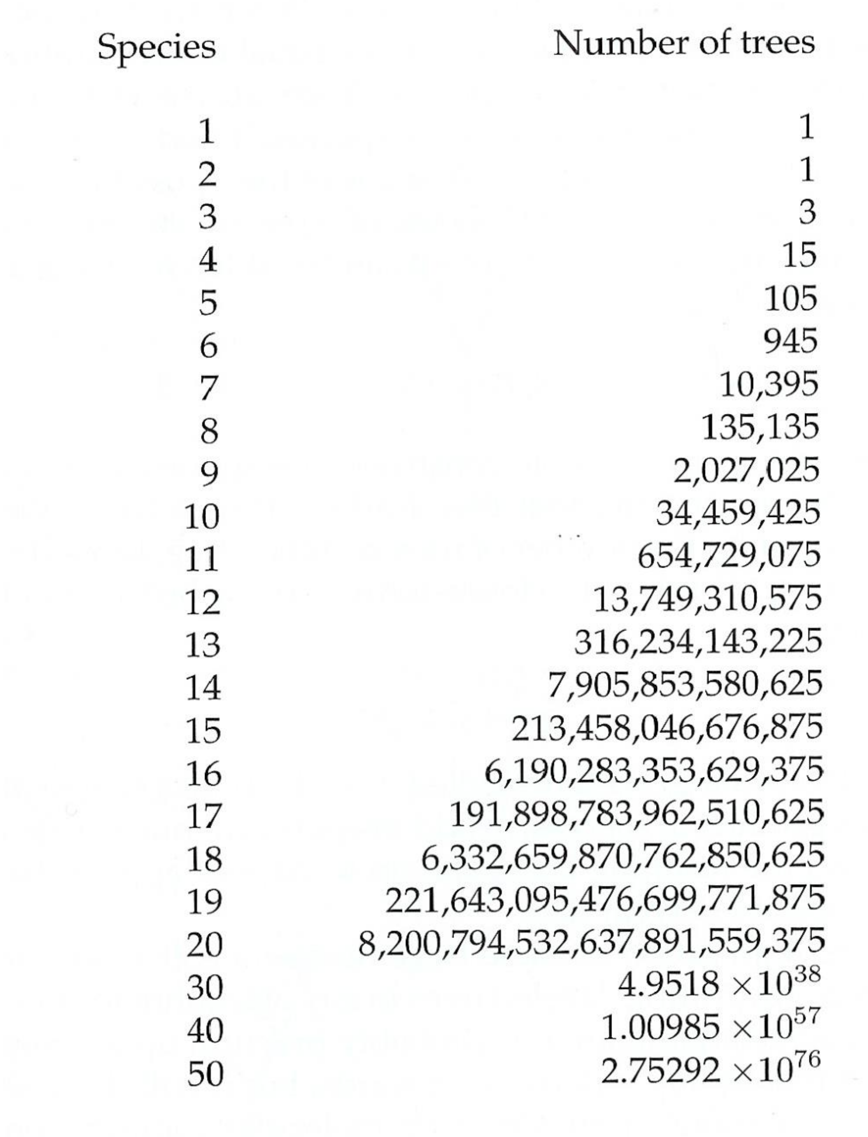
How many tree are there?



Rooted, Labeled and bifurcated



Rooted, Labeled and bifurcating



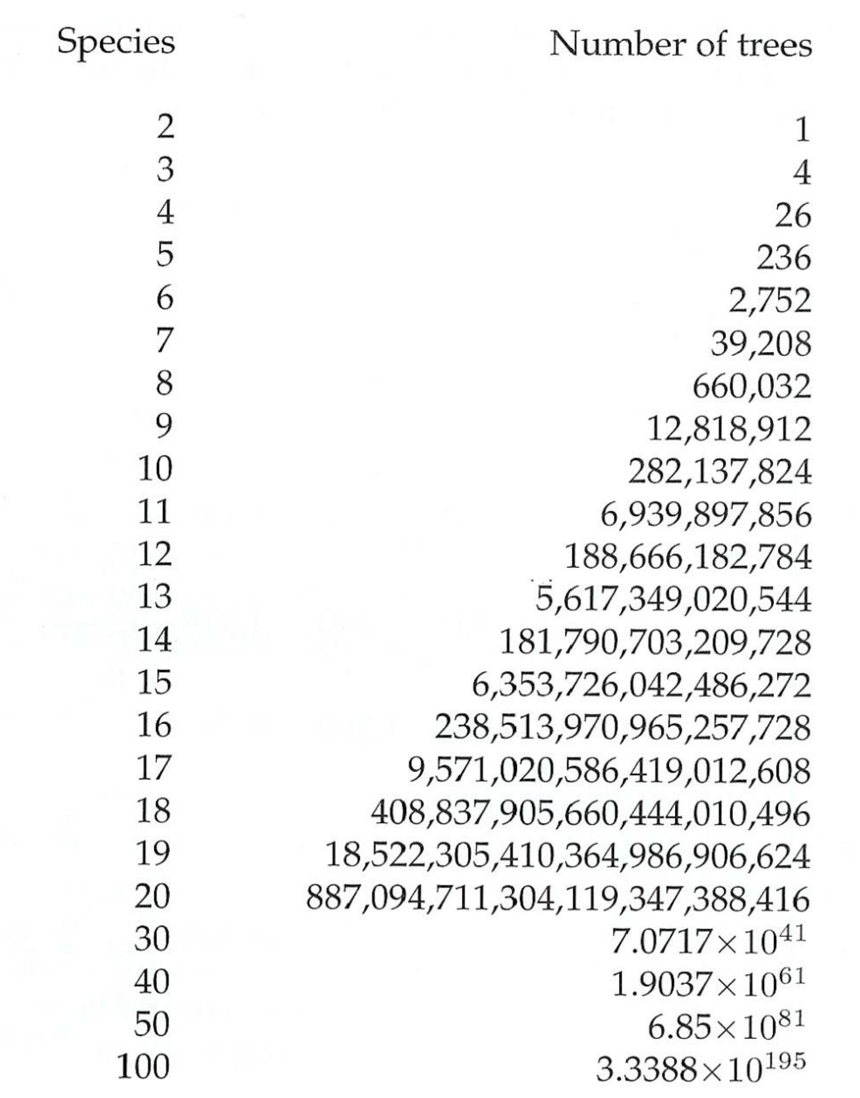
Species	Number of trees
1	1
2	1
3	3
4	15
5	105
6	945
7	10,395
8	135,135
9	2,027,025
10	34,459,425
11	654,729,075
12	13,749,310,575
13	316,234,143,225
14	7,905,853,580,625
15	213,458,046,676,875
16	6,190,283,353,629,375
17	191,898,783,962,510,625
18	6,332,659,870,762,850,625
19	221,643,095,476,699,771,875
20	8,200,794,532,637,891,559,375
30	4.9518×10^{38}
40	1.00985×10^{57}
50	2.75292×10^{76}

$$\frac{(2n-3)!}{2^{n-2}(n-2)!}$$

Sometimes called

$$(2n-3)!!$$

Rooted, Labeled and multifurcating



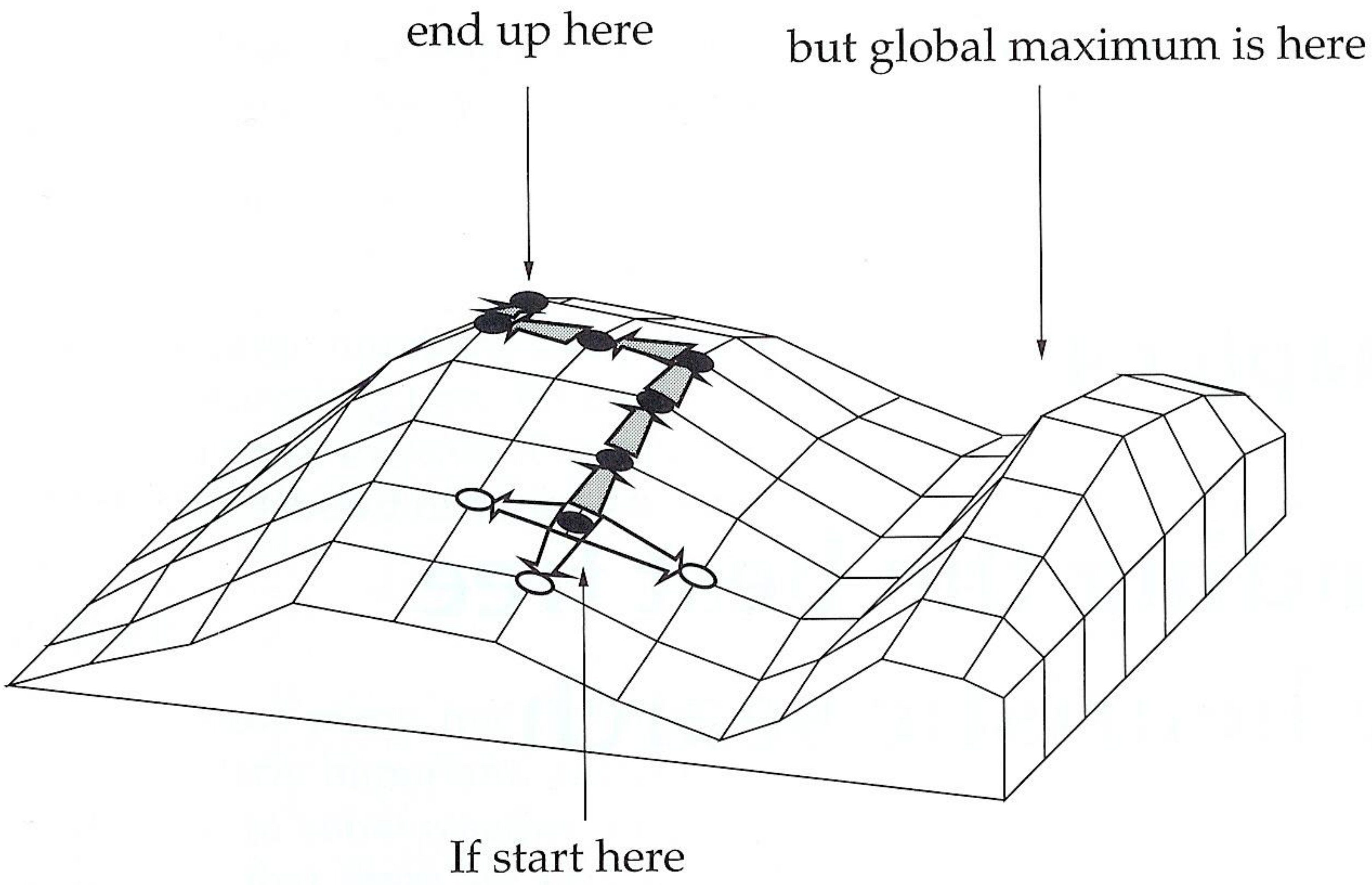
Species	Number of trees
2	1
3	4
4	26
5	236
6	2,752
7	39,208
8	660,032
9	12,818,912
10	282,137,824
11	6,939,897,856
12	188,666,182,784
13	5,617,349,020,544
14	181,790,703,209,728
15	6,353,726,042,486,272
16	238,513,970,965,257,728
17	9,571,020,586,419,012,608
18	408,837,905,660,444,010,496
19	18,522,305,410,364,986,906,624
20	887,094,711,304,119,347,388,416
30	7.0717×10^{41}
40	1.9037×10^{61}
50	6.85×10^{81}
100	3.3388×10^{195}

Exhaustive Search

Computer algorithms to calculate trees from distance matrices are straightforward sequential cluster methods, and consequently fast. But for parsimony and maximum likelihood the calculations are more tedious because all possible alternatives need to be considered to find the best of all solutions and for maximum likelihood it is even worse because for each tree a search is involved for the maximum likelihood value (Kuhner & Felsenstein 1994).

Some shortcuts to the exhaustive search have been invented in order to save computation time, namely the branch-and-bound technique and the heuristic approach. The branch-and-bound method is an exact algorithm like the exhaustive search, but instead of analysing all of the possible trees, it starts by evaluating a random tree, then follows several alternative paths by successive incorporation of taxa, abandoning a certain path every time the score obtained is higher than the score of the random tree with all the taxa.

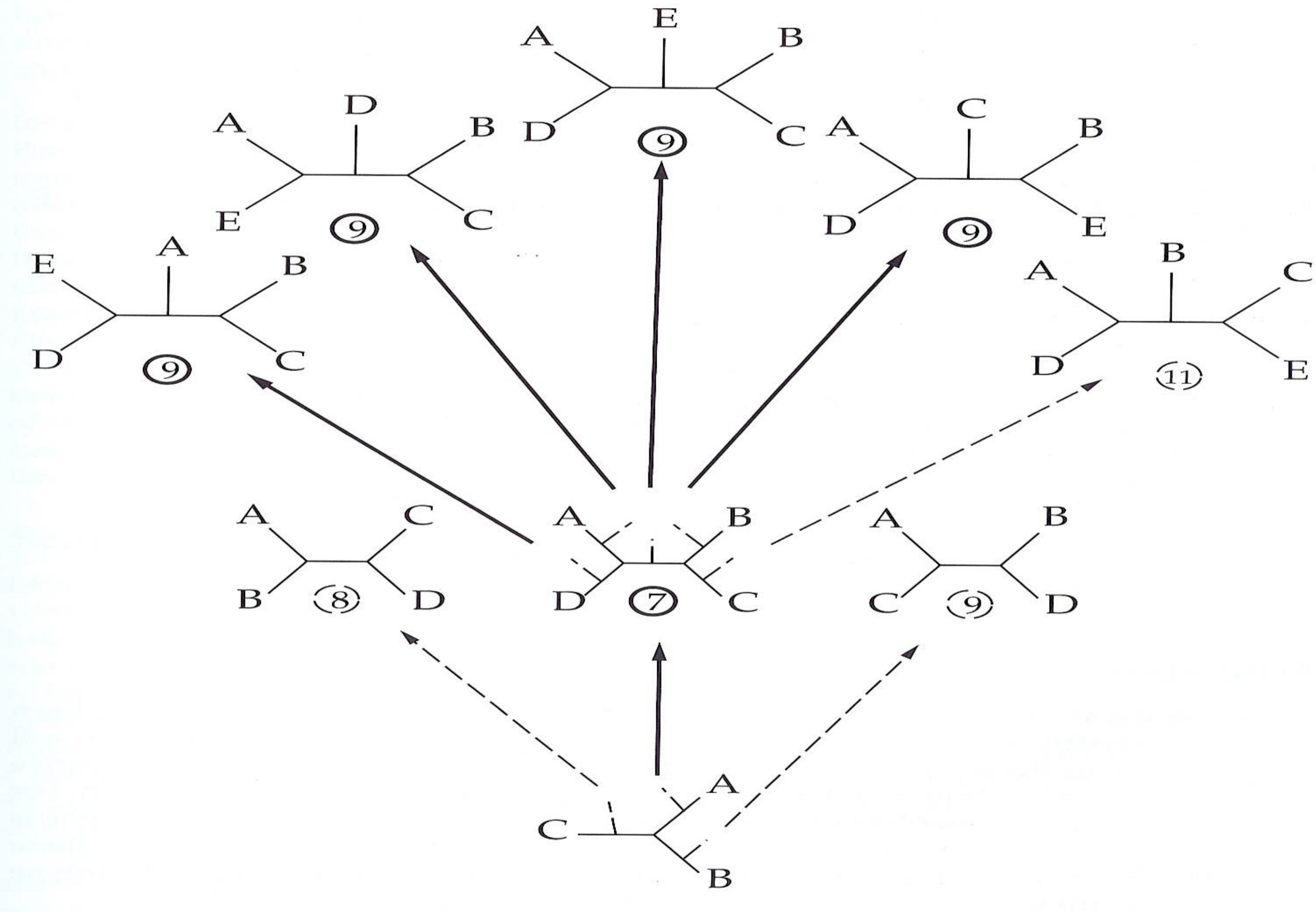
Search



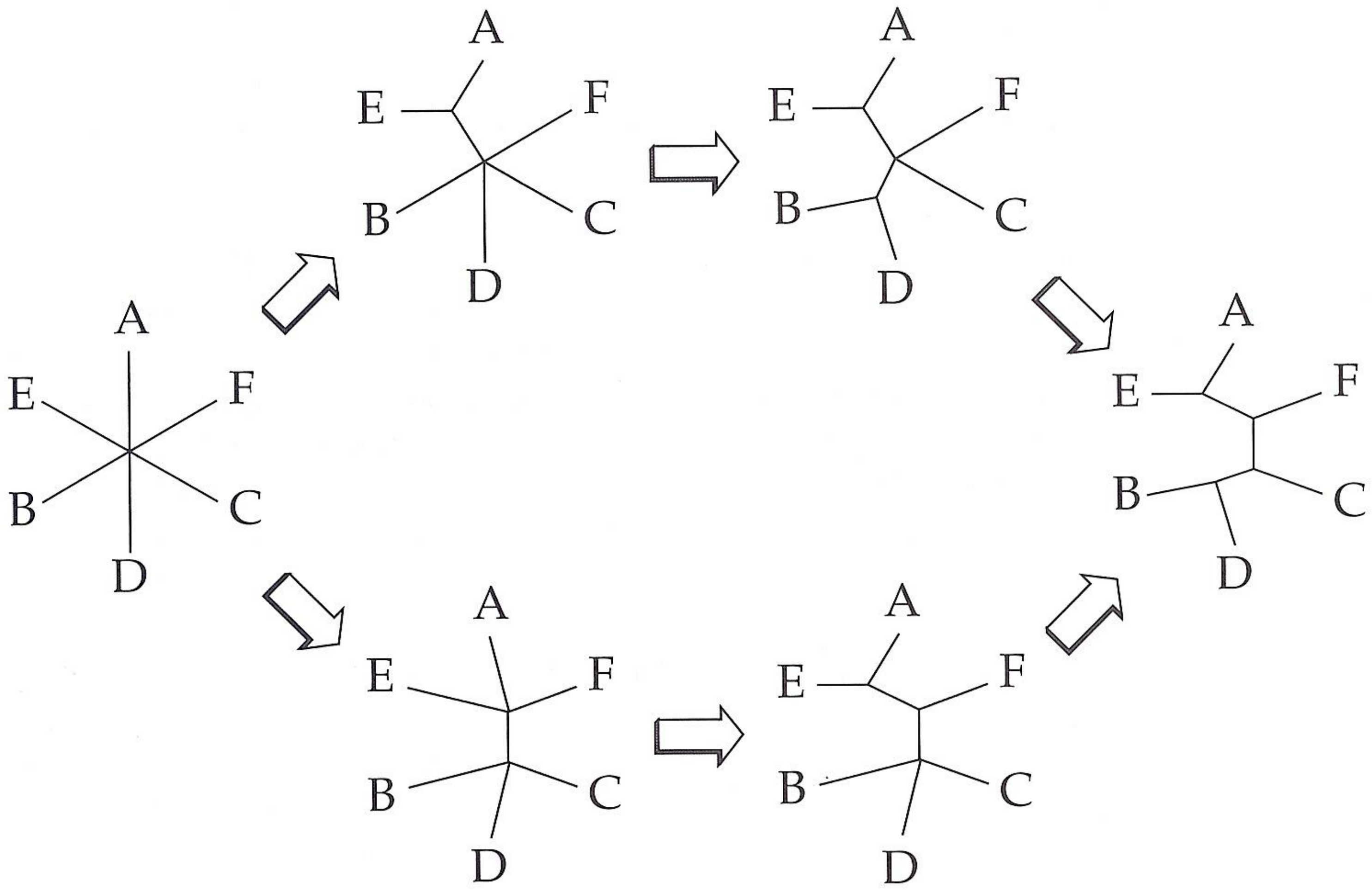
Heuristic Search

The heuristic approach is used when the data set is too large, and consequently too time consuming, to analyse with an exact algorithm, but it sacrifices the guarantee of finding the best of all trees. Three techniques have been used, the stepwise addition, the star decomposition and the branch swapping.
The stepwise addition functions by successive addiction of taxa to a growing tree. In each step the resulting trees are evaluated, and only the best ones are kept for the next step. The star decomposition begins with all taxa connected in a star-like way and by successive pairwise clustering with evaluation, only the optimal trees of each step are saved, leading to the final tree.

Sequential Addition



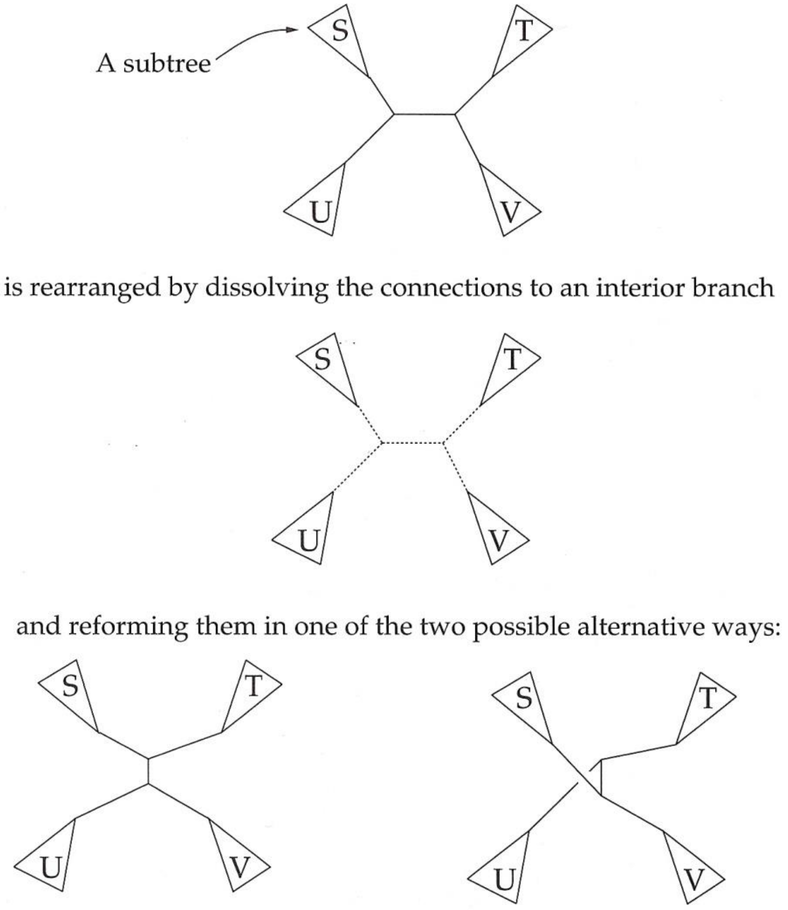
Star decomposition



Search

Both of these techniques usually find local optimum trees, but not necessarily the global optimal tree, unless the number of taxa is small or the data very simple. The branch swapping method tries to increase the chance of finding the global optimum by performing sets of predefining rearrangements of the tree branches with the respective evaluation. If the branch swapping is not only made on the best trees of each step but also on the suboptimal trees it increases the chance that the final result is actually the global optimal tree (Swofford *et al.* 1996).

Nearest-neighbor interchange NNI

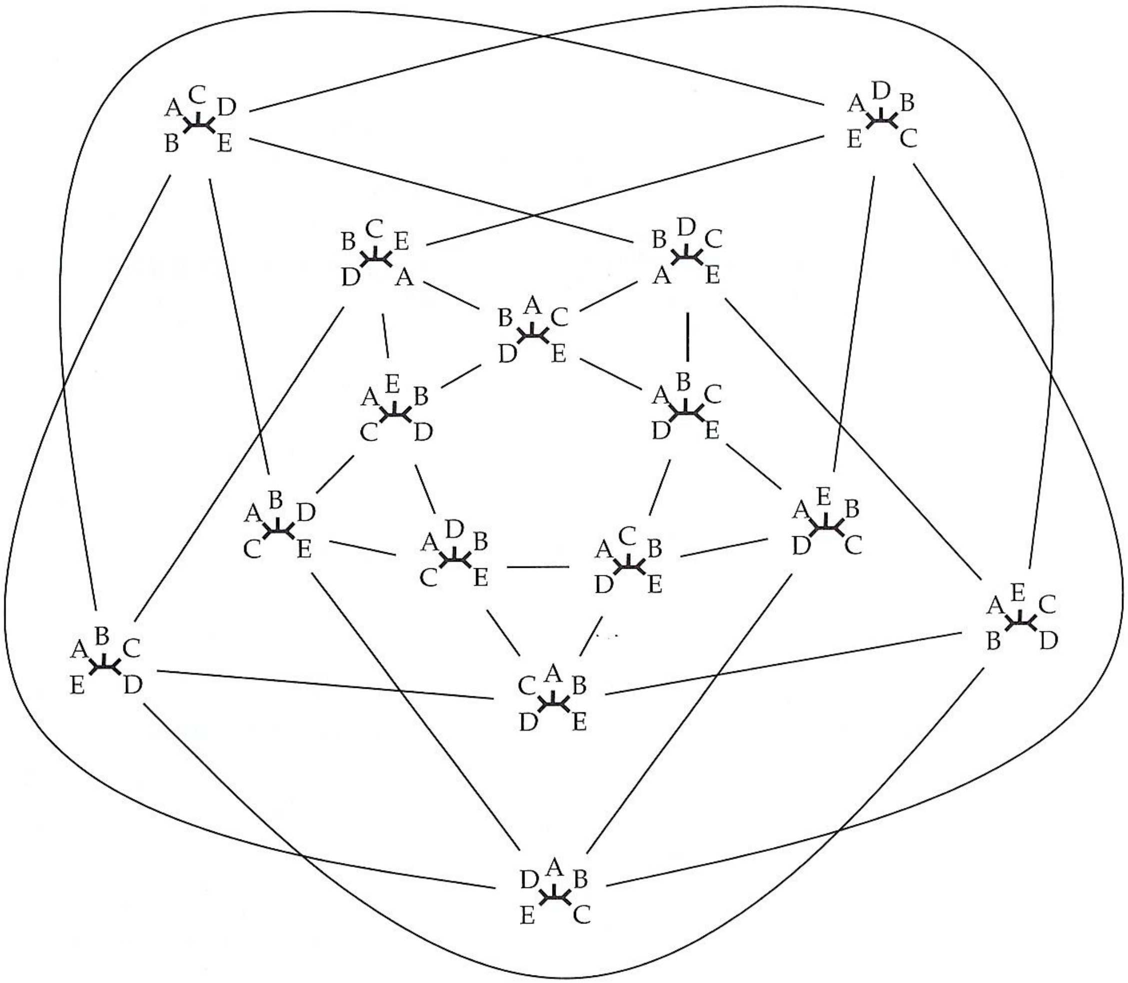


is rearranged by dissolving the connections to an interior branch

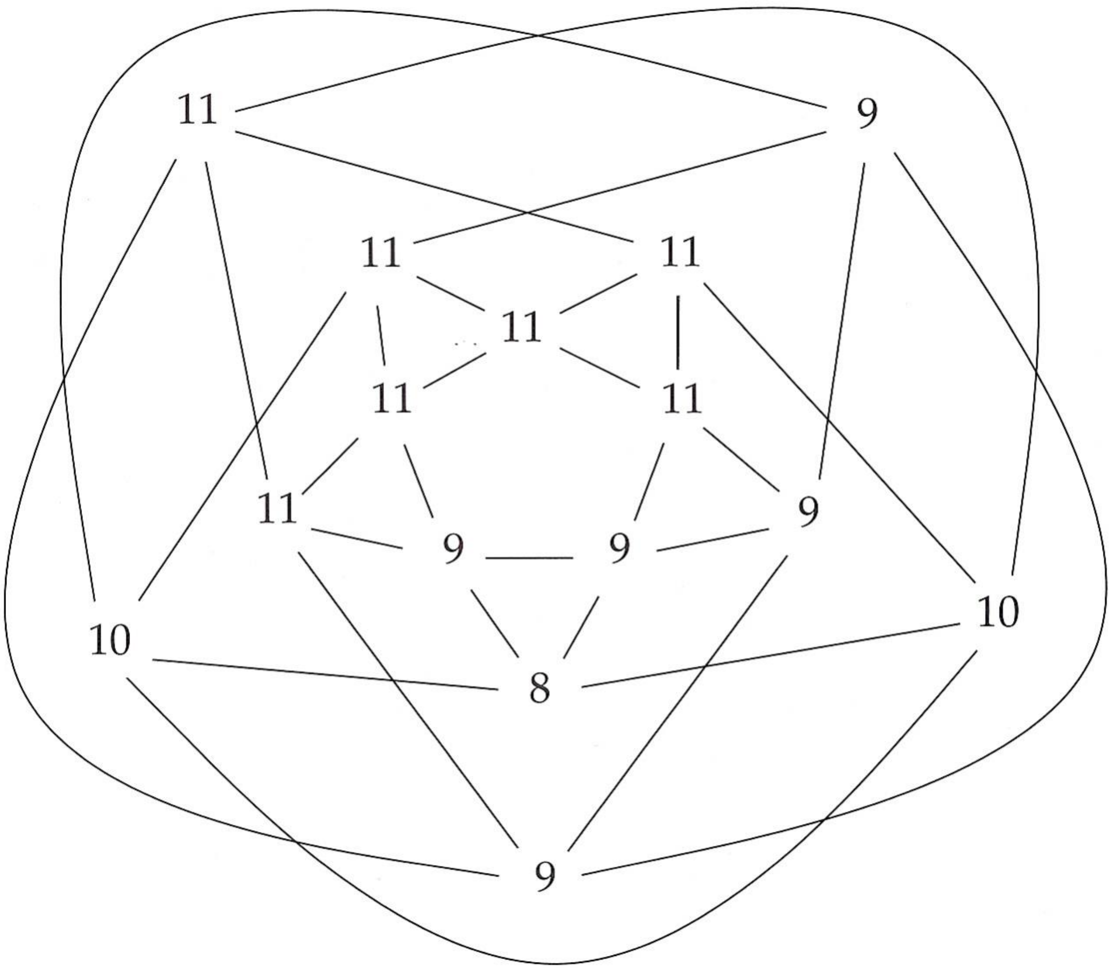
and reforming them in one of the two possible alternative ways:

Figure 4.2: The process of nearest-neighbor interchange. An interior branch is dissolved and the four subtrees connected to it are isolated. These then can be reconnected in two other ways.

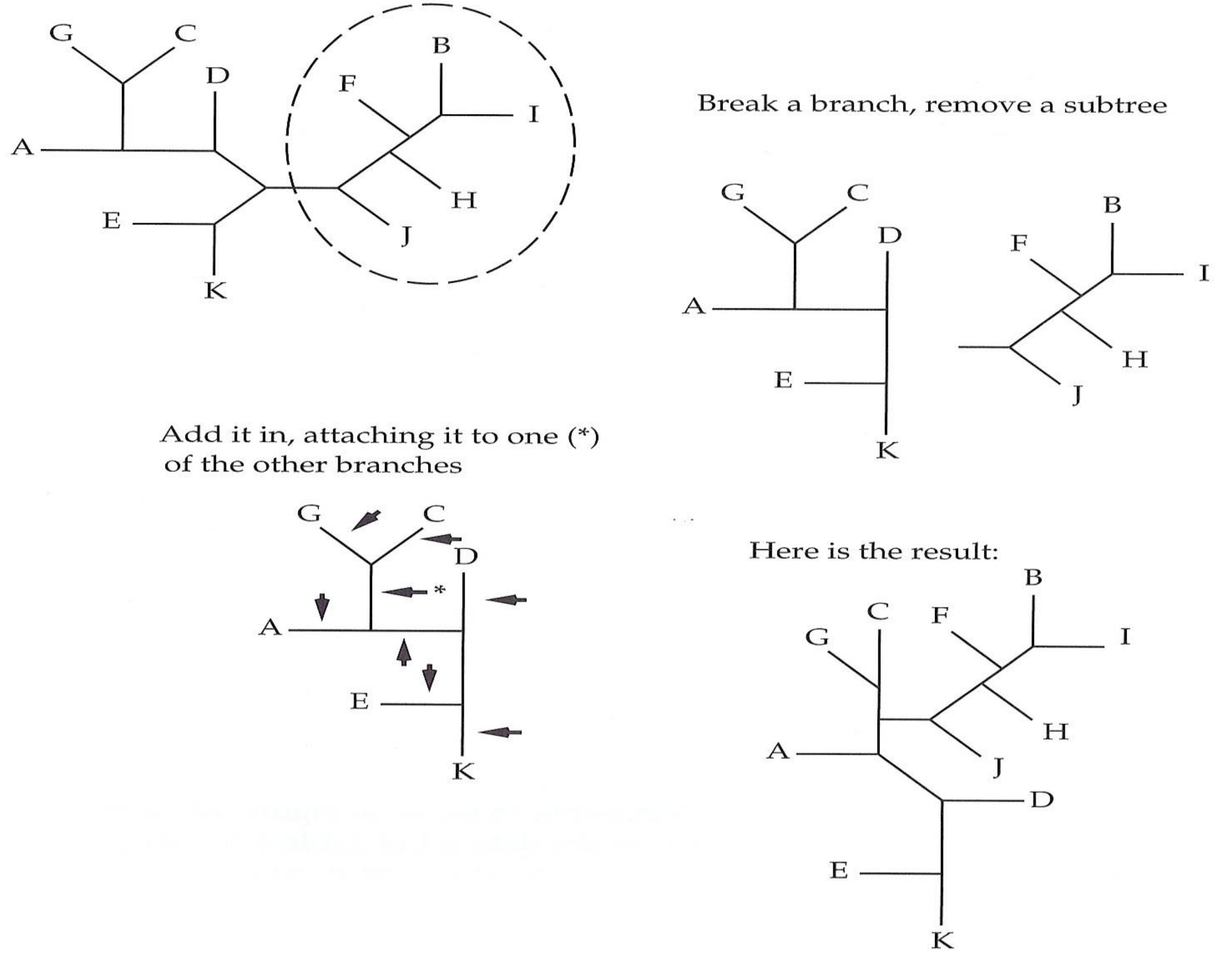
NNI – number of changes of state



NNI – number of changes of state



Subtree prunning and regrafting - SPR

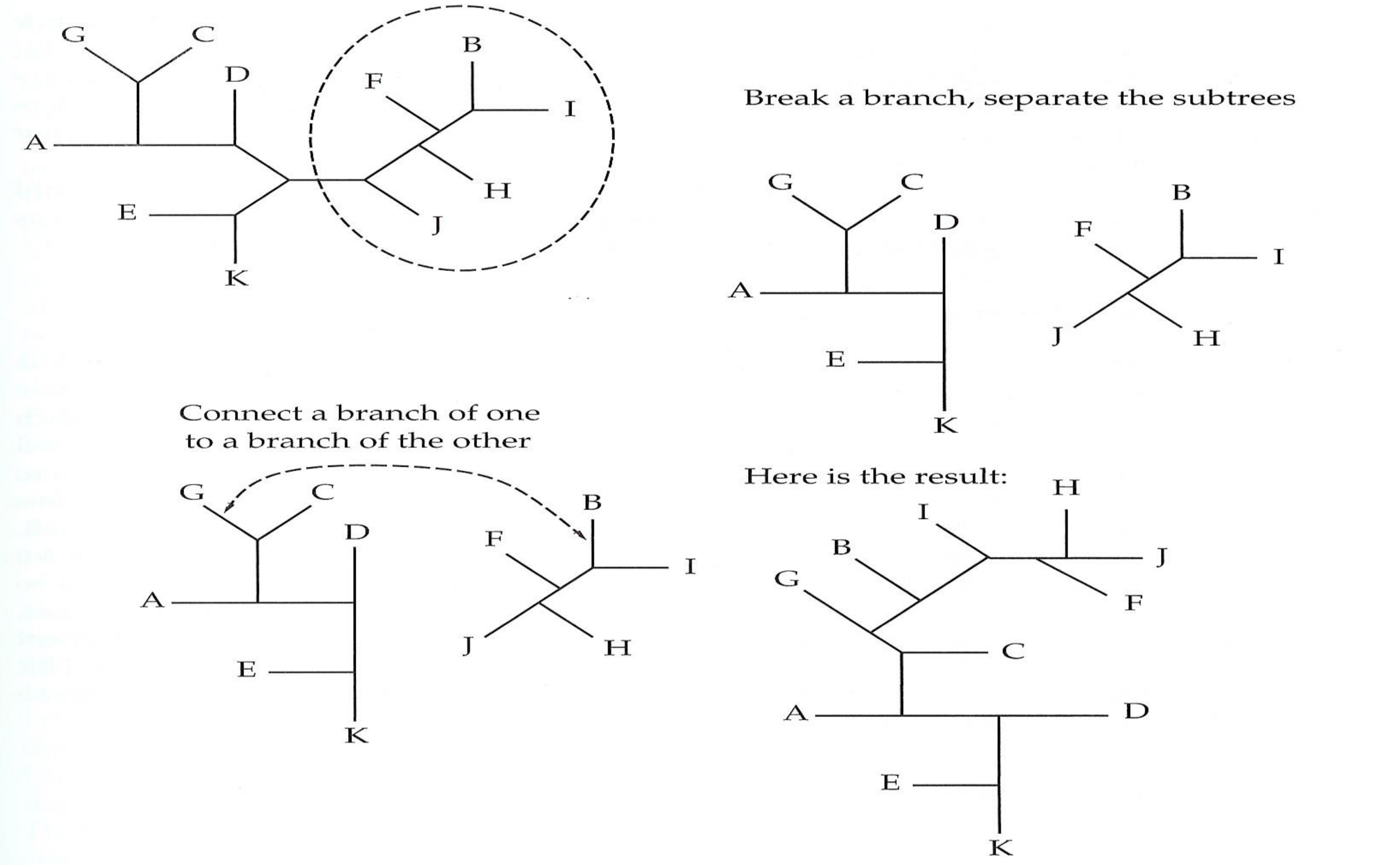


Break a branch, remove a subtree

Add it in, attaching it to one (*) of the other branches

Here is the result:

Tree Bisection and Reconnection - TBR



Break a branch, separate the subtrees

Connect a branch of one to a branch of the other

Here is the result:

Branch and Bound

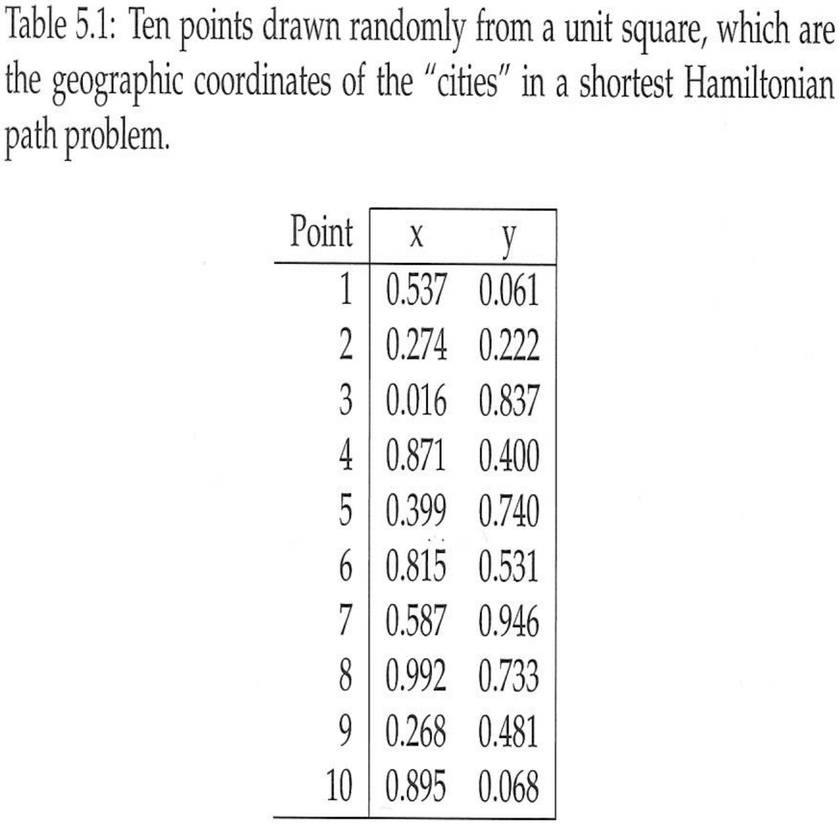
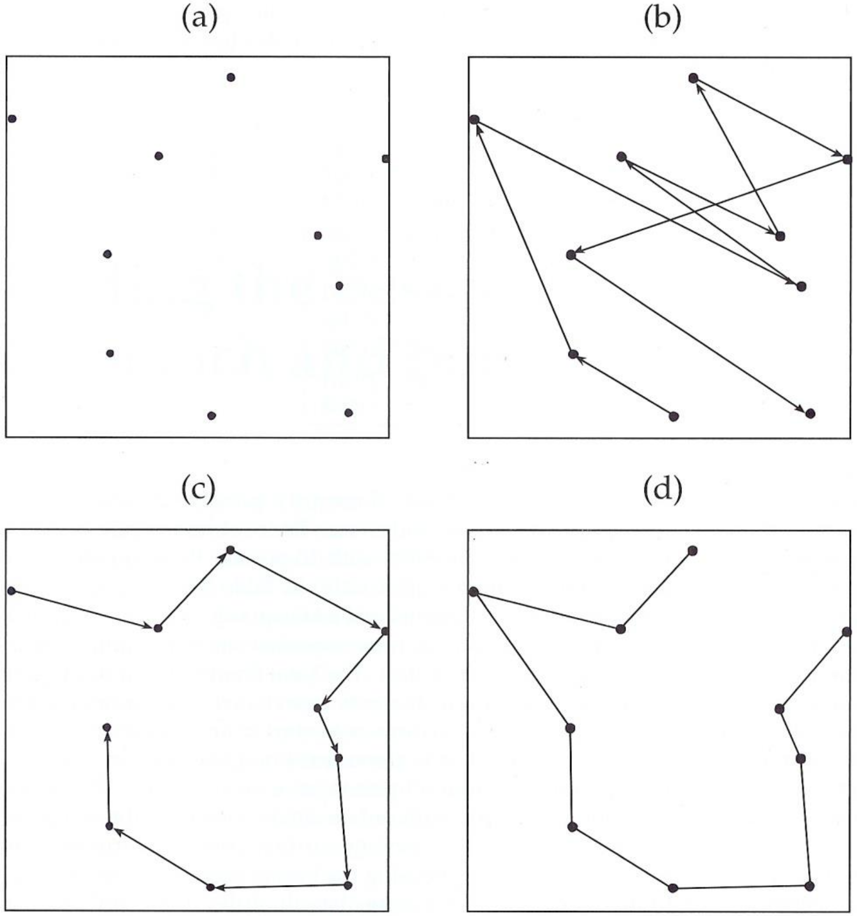
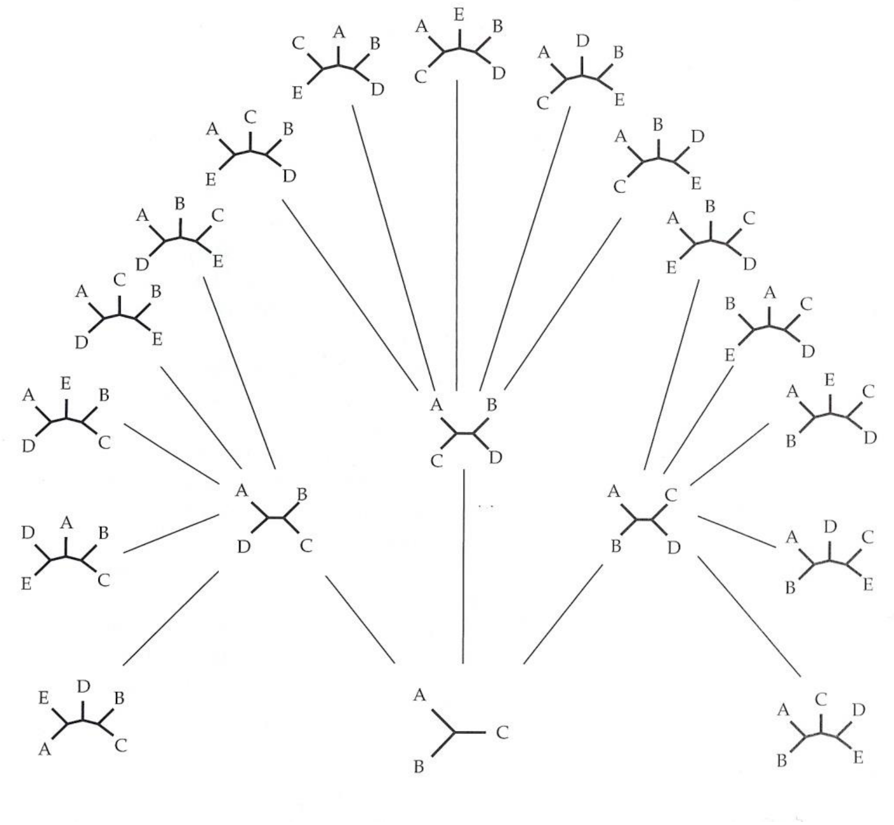


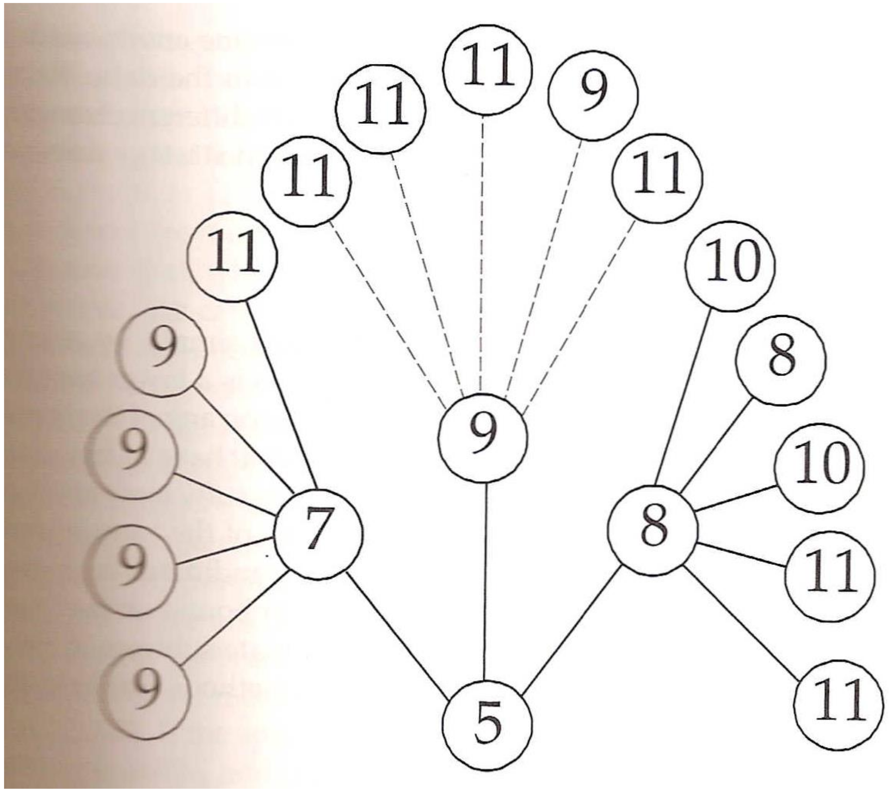
Table 5.1: Ten points drawn randomly from a unit square, which are the geographic coordinates of the "cities" in a shortest Hamiltonian path problem.

Point	x	y
1	0.537	0.061
2	0.274	0.222
3	0.016	0.837
4	0.871	0.400
5	0.399	0.740
6	0.815	0.531
7	0.587	0.946
8	0.992	0.733
9	0.268	0.481
10	0.895	0.068



Branch and Bound Search





Quartets

Another heuristic tree search procedure for maximum likelihood trees has recently been introduced (Strimmer & von Haeseler 1996). The method applies maximum-likelihood reconstruction to all possible quartets that can be formed from n sequences. These trees serve as starting points for the reconstruction of a set of optimal trees with all sequences. Improved versions of the original algorithm show high accuracy in returning the true tree without compromising speed or requiring more computer memory (Strimmer *et al.* 1997).

Inferencia Filogenética- Distâncias

		Types of Data	
		Distances	Nucleotide sites
Tree-building method	Clustering algorithm	UPGMA neighbour-joining	
	Optimality criterion	Minimum evolution	Maximum parsimony Maximum likelihood

Métodos de Distancia

site 1 2 3 4 5
Species 1- A T A T T
Species 2- A T C G T
Species 3- G C A G T
Species 4- G C C G T

	1	2	3	4
1				
2	2			
3	3	3		
4	4	2	1	

Métodos de Distancia II

Minimize total tree lengh ---> L=Σei

from i=1 to 2n-3 where n is the number of sequences

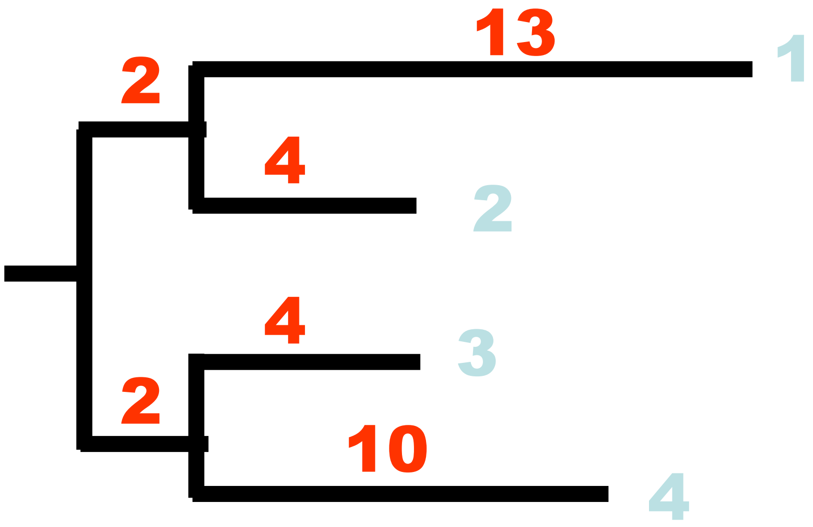
(2n-3) is the number of branches

ei is the branch lengh

	1	2	3	4
1				
2	2			
3	3	3		
4	4	2	1	

Métodos de Distancia III

	1	2	3	4
1				
2	17			
3	21	12		
4	27	18	14	



Clustering Algorithm - UPGMA

1 – Find i and j that have the smallest distance D_{ij}.

2 – Create a new group (ij), which has $n_{(ij)}=n_i+n_j$ members

3 – Connect i and j on the tree to a new node
(which corresponds to the new goup ij).
Give the two branches connecting i to (ij) and j to (ij) each Lenght $D_{ij}/2$.

4 - Compute the distance between the new group and all the other groups (except for i and j) by using:

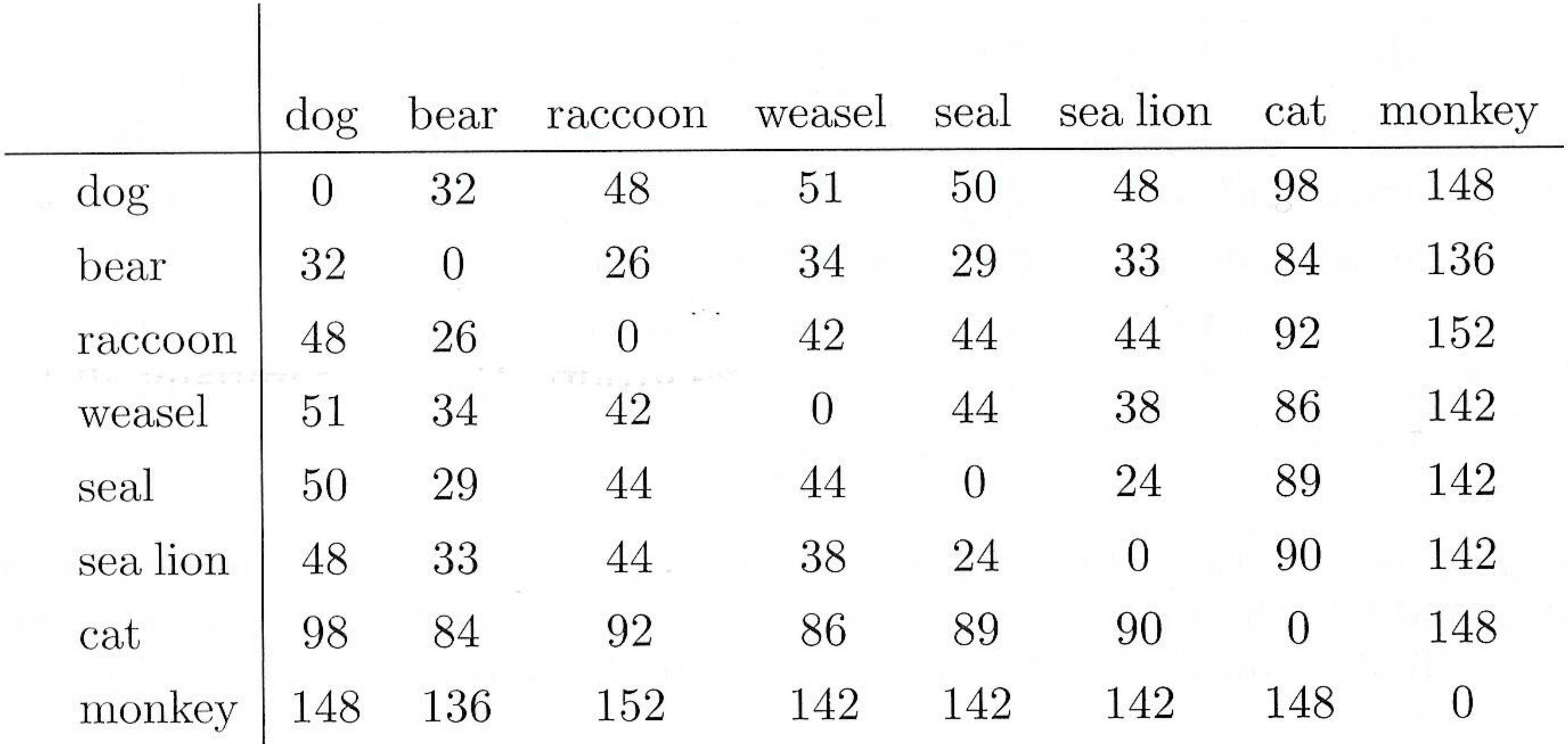
$$D_{(ij),k} = (n_i/(n_i+n_j))\, D_{ik}+(n_j/(n_i+n_j))\, D_{jk}$$

UPGMA

5- Delete the columns and rows of the data matrix that Correspond to groups i and j and add a column and a row for group (ij)

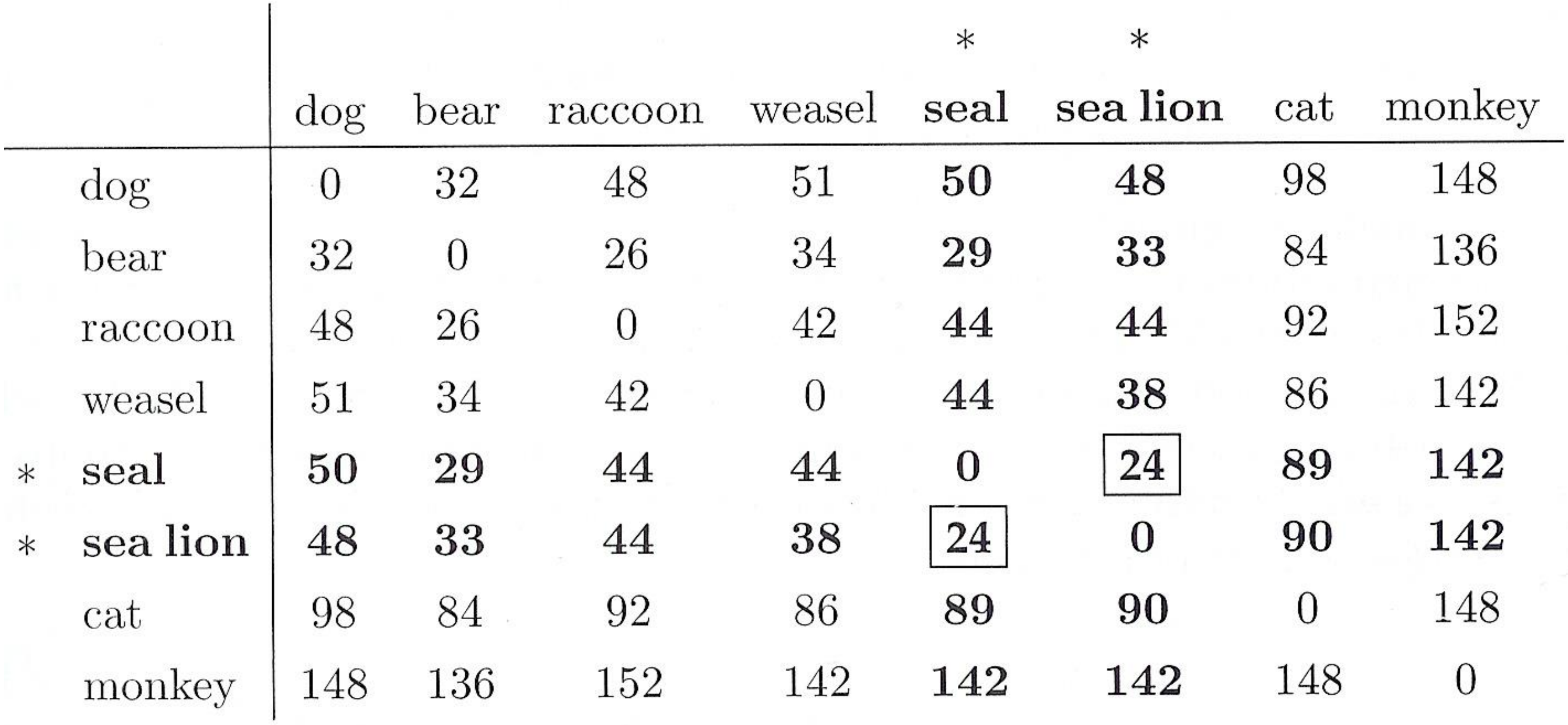
6 – If there is only one item in the data matrix ,stop. Otherwise return to 1

UPGMA – exemplo numérico



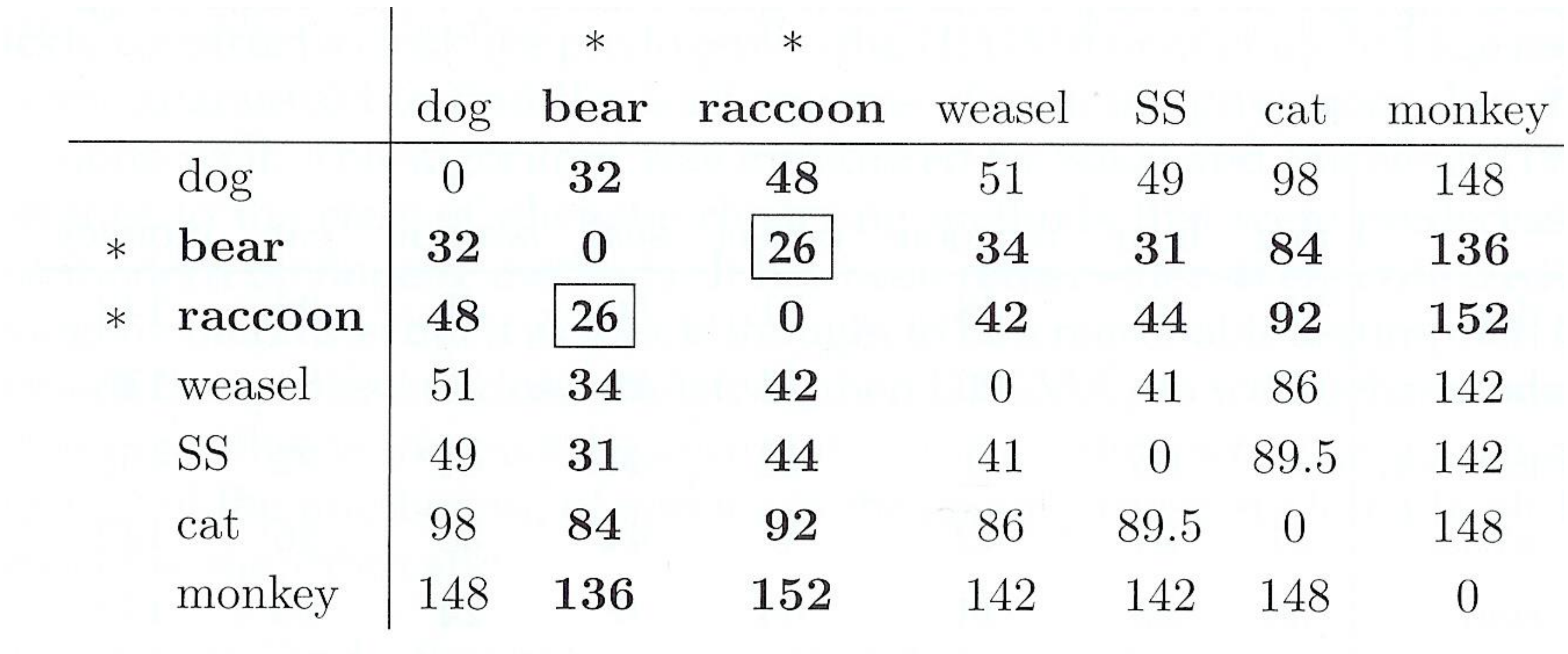
	dog	bear	raccoon	weasel	seal	sea lion	cat	monkey
dog	0	32	48	51	50	48	98	148
bear	32	0	26	34	29	33	84	136
raccoon	48	26	0	42	44	44	92	152
weasel	51	34	42	0	44	38	86	142
seal	50	29	44	44	0	24	89	142
sea lion	48	33	44	38	24	0	90	142
cat	98	84	92	86	89	90	0	148
monkey	148	136	152	142	142	142	148	0

UPGMA



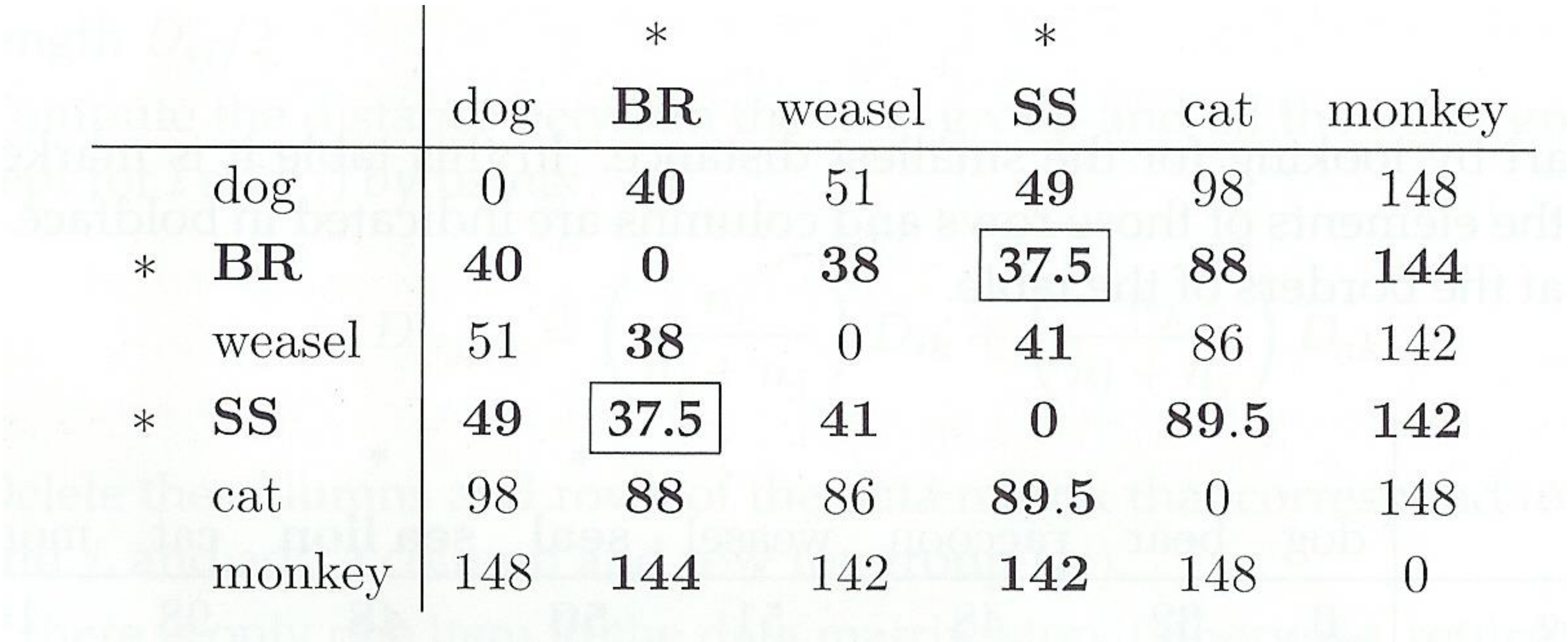
						*	*		
		dog	bear	raccoon	weasel	**seal**	**sea lion**	cat	monkey
	dog	0	32	48	51	**50**	**48**	98	148
	bear	32	0	26	34	**29**	**33**	84	136
	raccoon	48	26	0	42	**44**	**44**	92	152
	weasel	51	34	42	0	**44**	**38**	86	142
*	**seal**	**50**	**29**	**44**	**44**	**0**	$\boxed{\mathbf{24}}$	**89**	**142**
*	**sea lion**	**48**	**33**	**44**	**38**	$\boxed{\mathbf{24}}$	**0**	**90**	**142**
	cat	98	84	92	86	**89**	**90**	0	148
	monkey	148	136	152	142	**142**	**142**	148	0

UPGMA



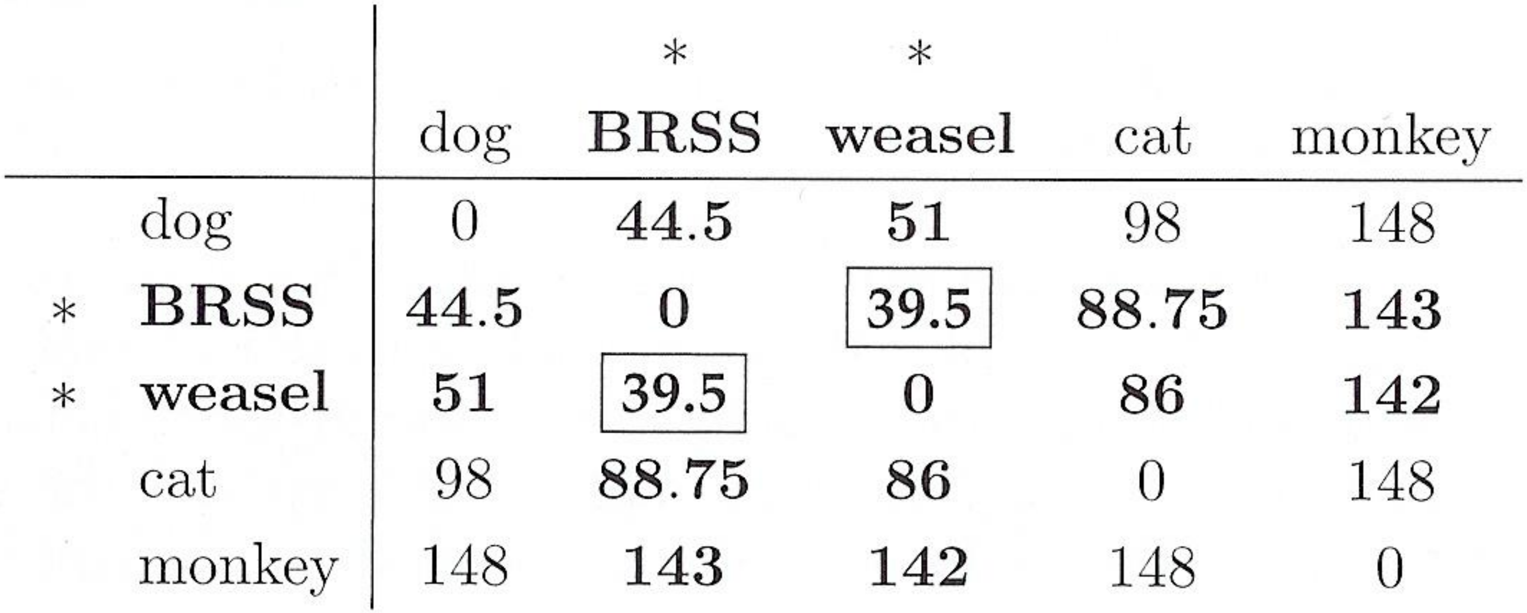
			*	*				
		dog	**bear**	**raccoon**	weasel	SS	cat	monkey
	dog	0	**32**	**48**	51	49	98	148
*	**bear**	**32**	**0**	$\boxed{26}$	**34**	**31**	**84**	**136**
*	**raccoon**	**48**	$\boxed{26}$	**0**	**42**	**44**	**92**	**152**
	weasel	51	**34**	**42**	0	41	86	142
	SS	49	**31**	**44**	41	0	89.5	142
	cat	98	**84**	**92**	86	89.5	0	148
	monkey	148	**136**	**152**	142	142	148	0

UPGMA



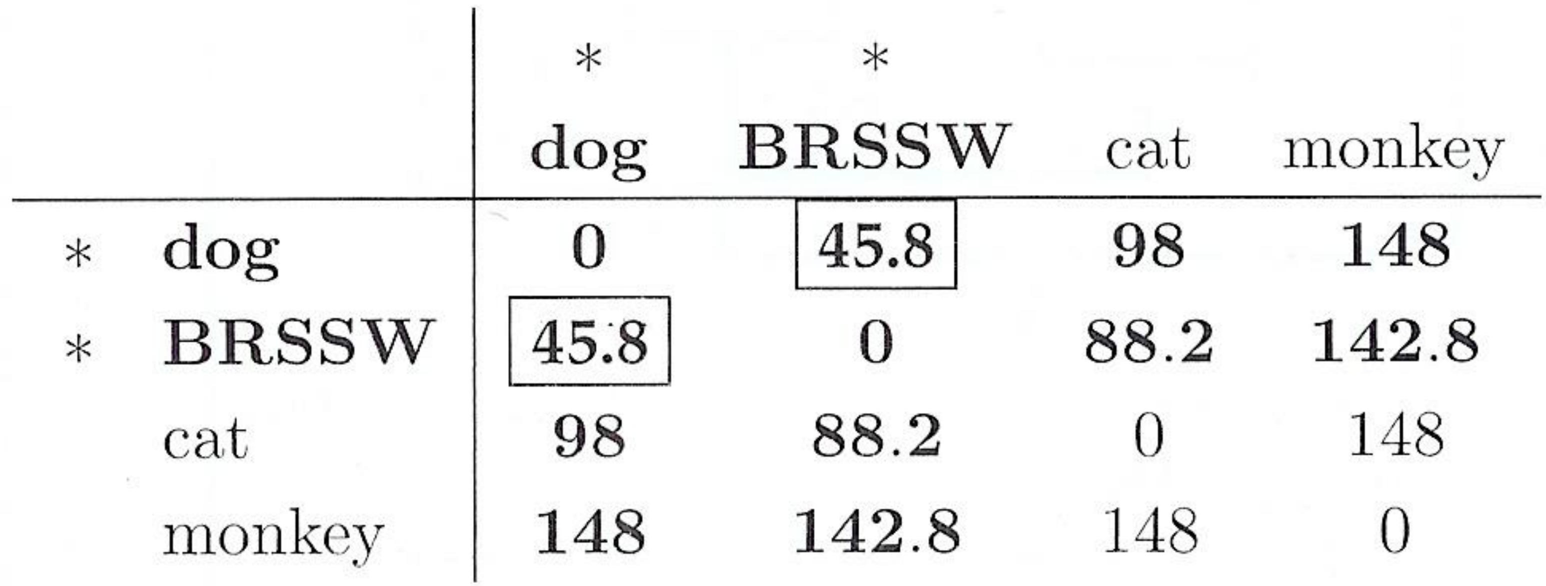
			*		*		
		dog	**BR**	weasel	**SS**	cat	monkey
	dog	0	**40**	51	**49**	98	148
*	**BR**	**40**	**0**	**38**	$\boxed{\mathbf{37.5}}$	**88**	**144**
	weasel	51	**38**	0	**41**	86	142
*	**SS**	**49**	$\boxed{\mathbf{37.5}}$	**41**	**0**	**89.5**	**142**
	cat	98	**88**	86	**89.5**	0	148
	monkey	148	**144**	142	**142**	148	0

UPGMA



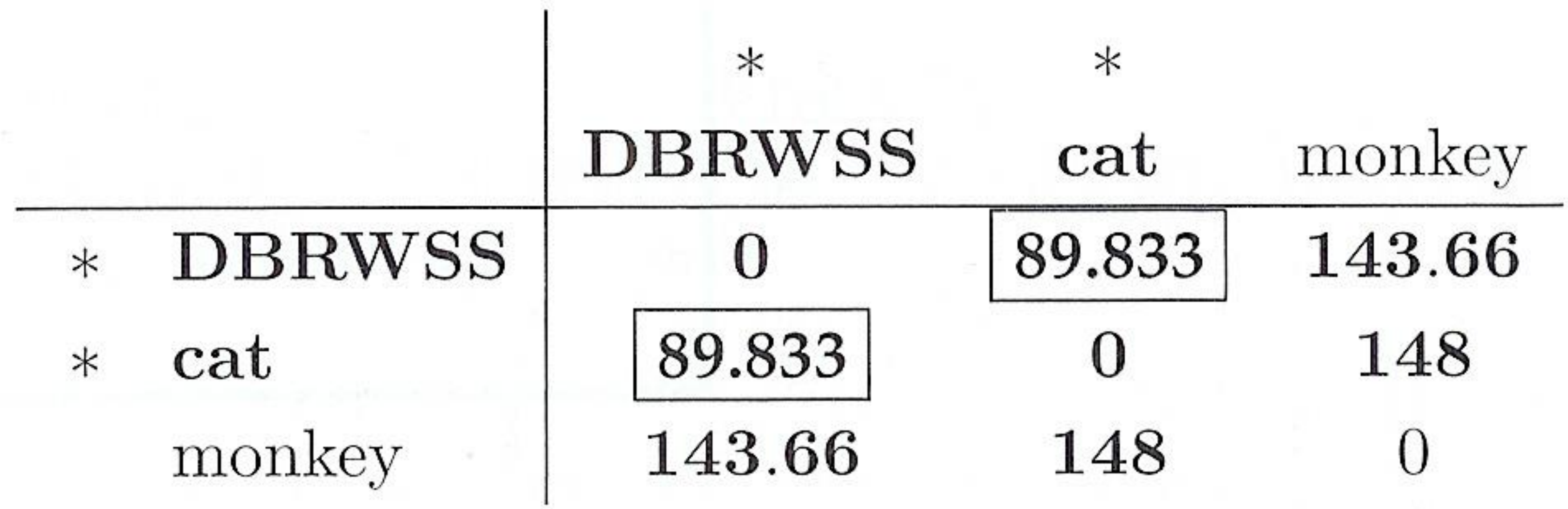
			*	*		
		dog	BRSS	weasel	cat	monkey
	dog	0	44.5	51	98	148
*	BRSS	44.5	0	$\boxed{39.5}$	88.75	143
*	weasel	51	$\boxed{39.5}$	0	86	142
	cat	98	88.75	86	0	148
	monkey	148	143	142	148	0

UPGMA



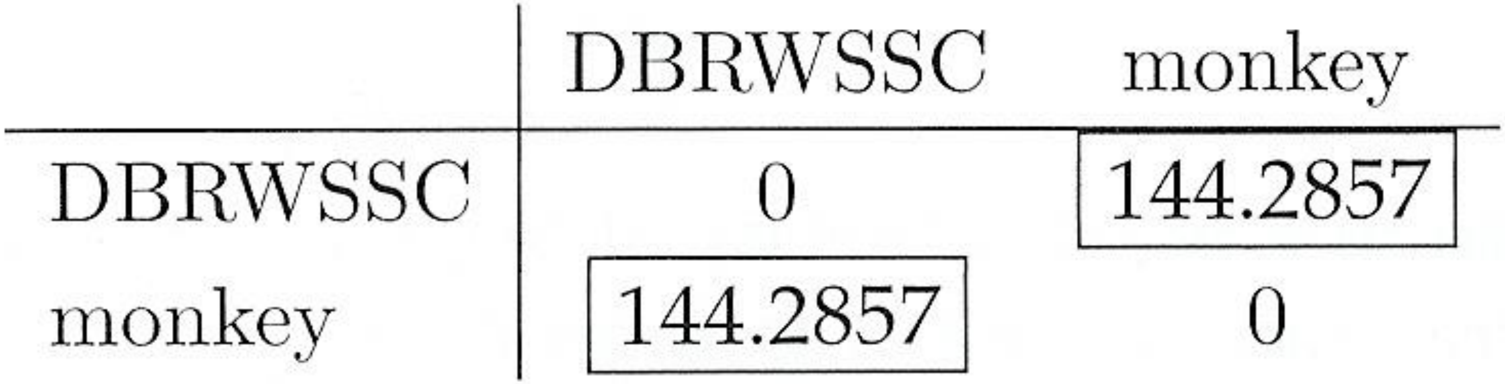
		*	*		
		dog	BRSSW	cat	monkey
*	dog	0	$\boxed{45.8}$	98	148
*	BRSSW	$\boxed{45.8}$	0	88.2	142.8
	cat	98	88.2	0	148
	monkey	148	142.8	148	0

UPGMA



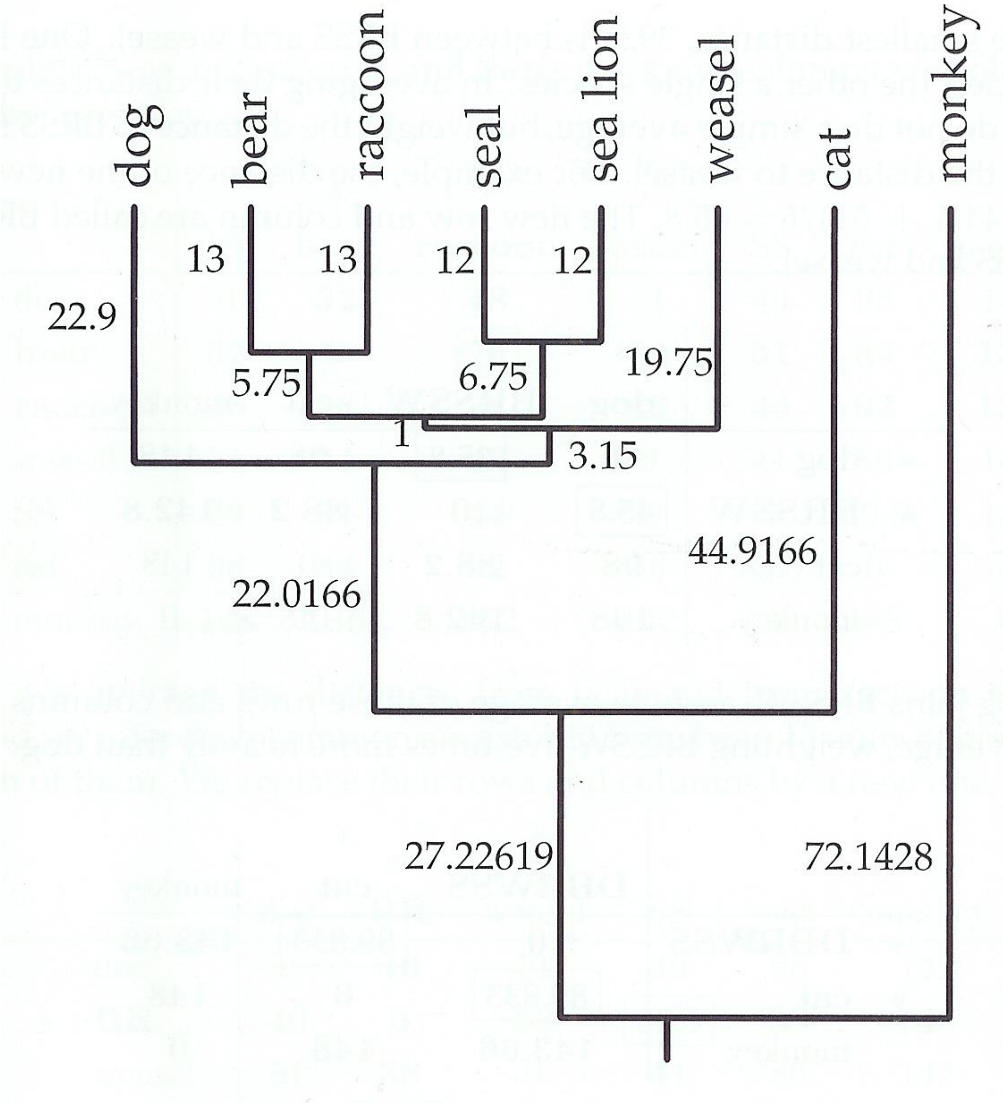
		*	*	
		DBRWSS	cat	monkey
*	DBRWSS	0	$\boxed{89.833}$	143.66
*	cat	$\boxed{89.833}$	0	148
	monkey	143.66	148	0

UPGMA



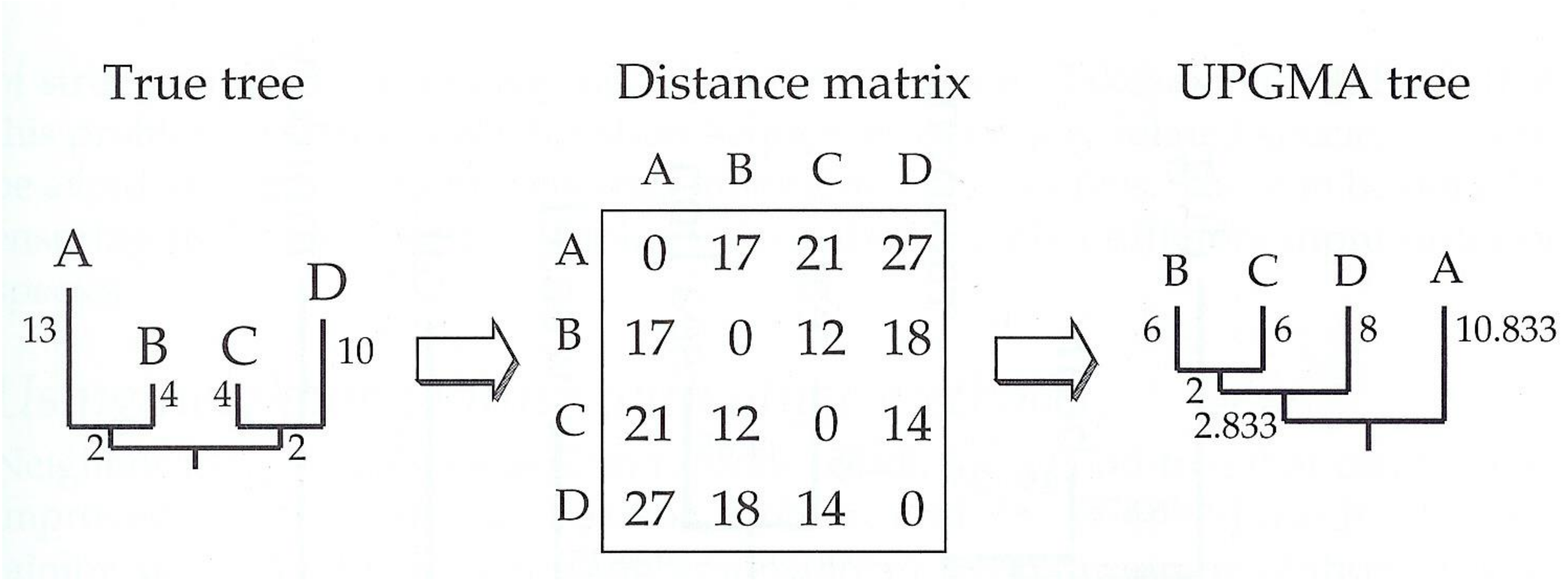
	DBRWSSC	monkey
DBRWSSC	0	144.2857
monkey	144.2857	0

UPGMA



dog
bear
raccoon
seal
sea lion
weasel
cat
monkey
13
13
12
12
22.9
19.75
5.75
6.75
1
3.15
44.9166
22.0166
27.22619
72.1428

Similarity vs evolutionary relationship



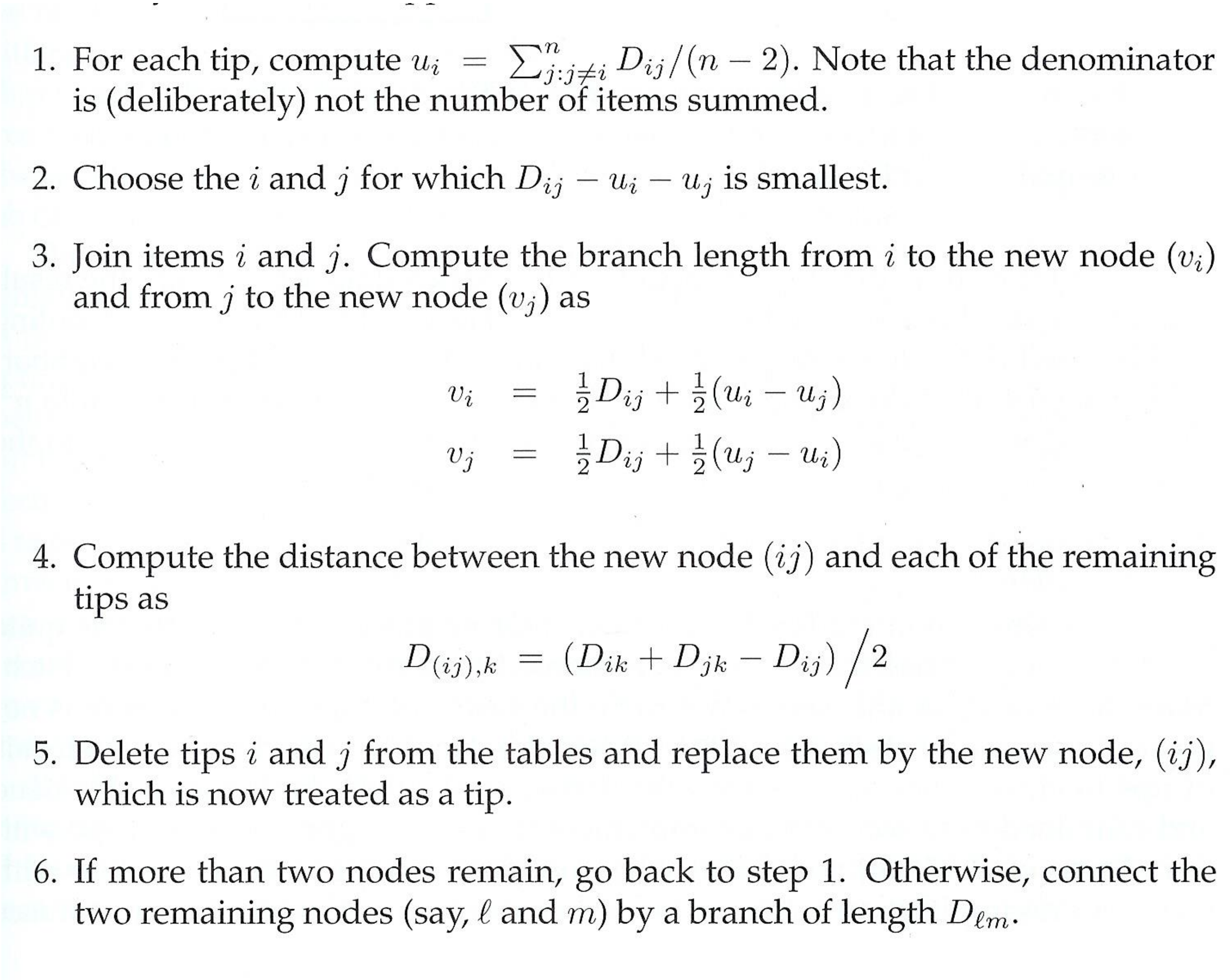
True tree

Distance matrix

	A	B	C	D
A	0	17	21	27
B	17	0	12	18
C	21	12	0	14
D	27	18	14	0

UPGMA tree

NJ algorithm



1. For each tip, compute $u_i = \sum_{j:j\neq i}^{n} D_{ij}/(n-2)$. Note that the denominator is (deliberately) not the number of items summed.

2. Choose the i and j for which $D_{ij} - u_i - u_j$ is smallest.

3. Join items i and j. Compute the branch length from i to the new node (v_i) and from j to the new node (v_j) as

$$\begin{aligned} v_i &= \tfrac{1}{2}D_{ij} + \tfrac{1}{2}(u_i - u_j) \\ v_j &= \tfrac{1}{2}D_{ij} + \tfrac{1}{2}(u_j - u_i) \end{aligned}$$

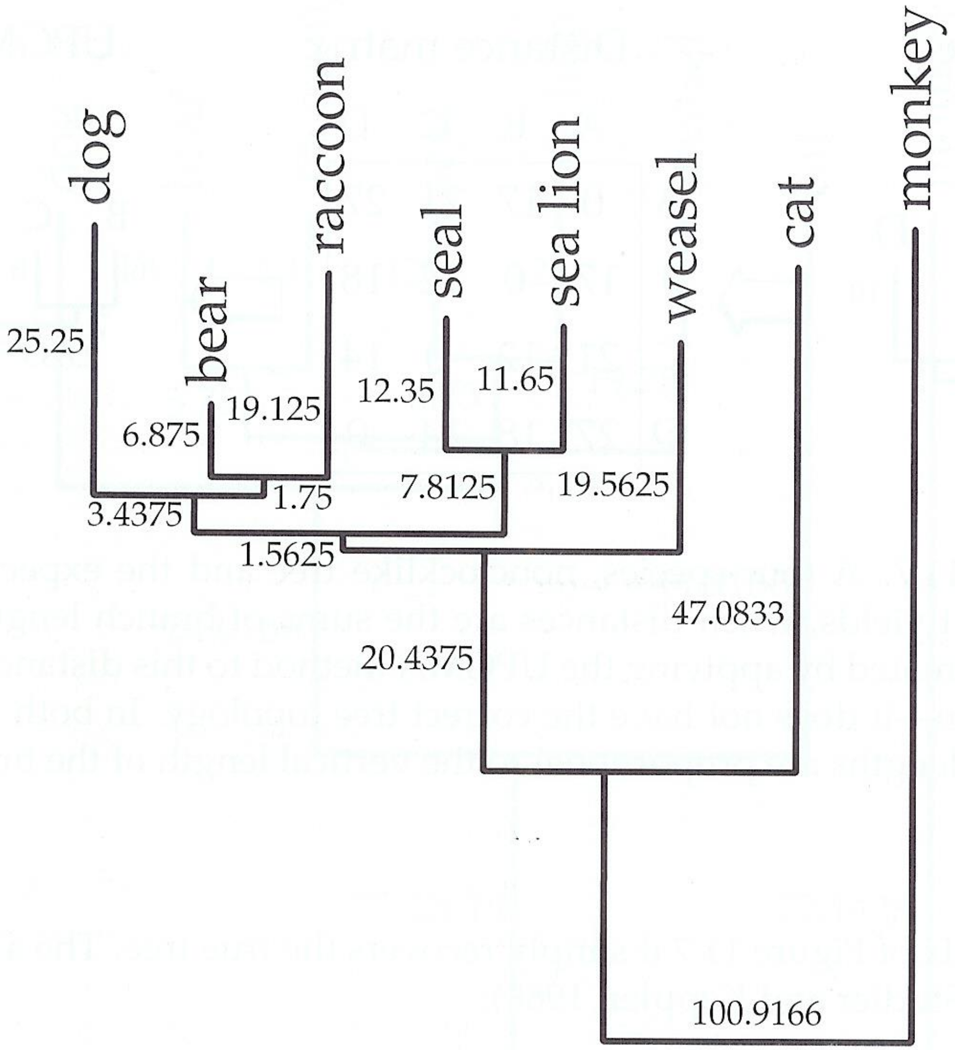
4. Compute the distance between the new node (ij) and each of the remaining tips as

$$D_{(ij),k} = (D_{ik} + D_{jk} - D_{ij}) \Big/ 2$$

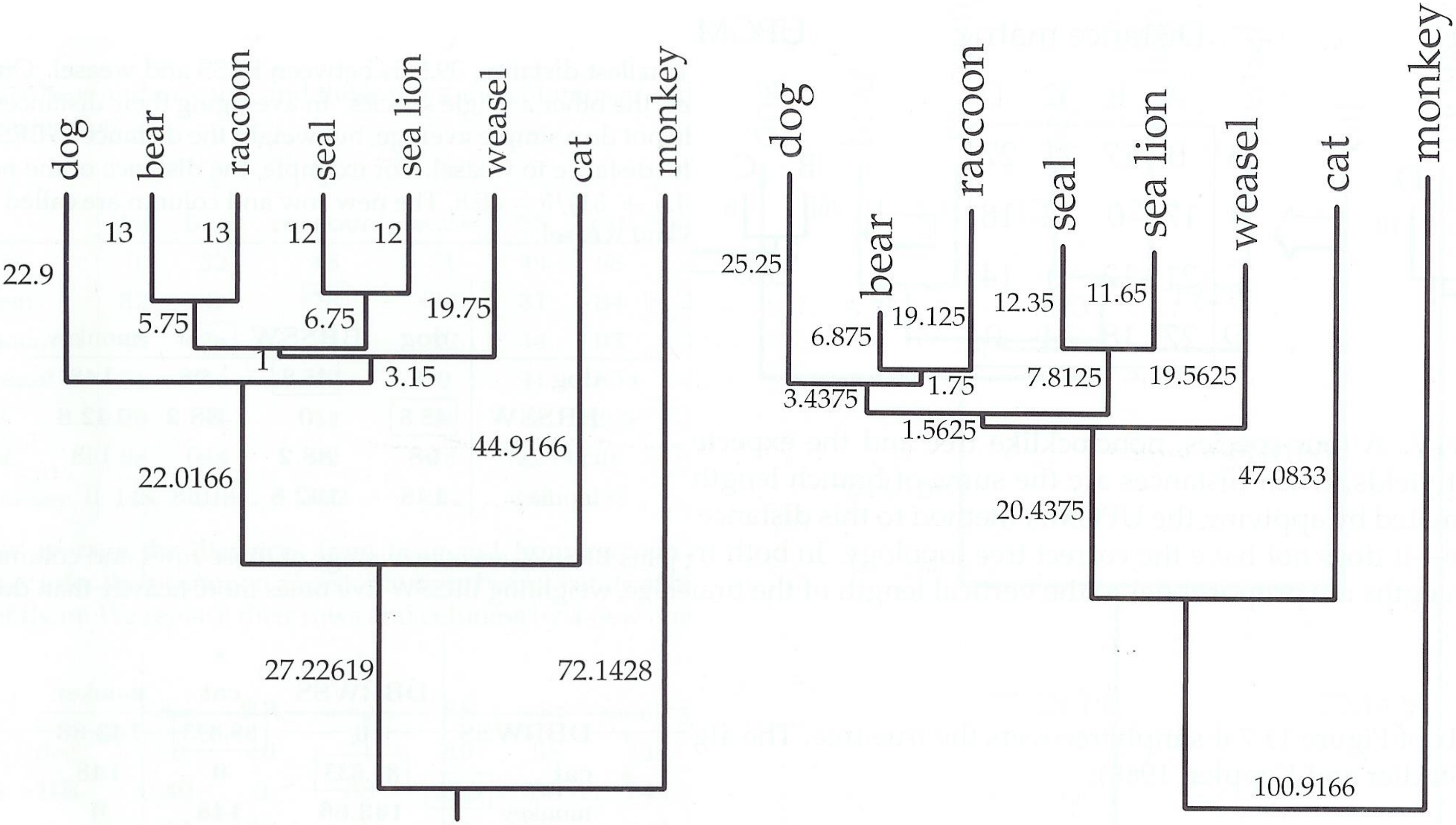
5. Delete tips i and j from the tables and replace them by the new node, (ij), which is now treated as a tip.

6. If more than two nodes remain, go back to step 1. Otherwise, connect the two remaining nodes (say, ℓ and m) by a branch of length $D_{\ell m}$.

NJ tree


dog
bear
raccoon
seal
sea lion
weasel
cat
monkey
25.25
6.875
19.125
12.35
11.65
3.4375
1.75
7.8125
19.5625
1.5625
47.0833
20.4375
100.9166

UPGMA vs NJ


dog
bear
raccoon
seal
sea lion
weasel
cat
monkey
13
13
12
12
22.9
5.75
6.75
19.75
1
3.15
44.9166
22.0166
27.22619
72.1428
dog
raccoon
seal
sea lion
weasel
cat
monkey
bear
25.25
19.125
6.875
12.35
11.65
3.4375
1.75
7.8125
19.5625
1.5625
47.0833
20.4375
100.9166