



5. Polymorphism, selection and Phylogenetics

- ✂ **Polymorphism in the genomes**
 - Measure of polymorphism
 - Types of polymorphism
- ✂ **Natural and artificial selection: the force shaping the genomes**
 - Neutral tests
 - Domestication selection
- ✂ **Genome-wide association studies (GWAS)**
 - Polymorphism and phenotype
- ✂ **Phylogenetics**
 - Distance method, MP and ML



Polymorphism in the genomes

- Types of polymorphism
 - SNP (point mutation)
 - Copy-number variation
 - Insertion/deletion (indel)
 - Frame-shift
- Measure of polymorphism
 - π (Tajima 1983):
 - θ (Watterson 1975):

Natural and artificial selection



Natural test methods:

- Tajima' s D
- Fay and Wu' s H
- HKA (Hudson-Kreitman-Aguade)
- Fu and Li' s D^* and F^*
- MK (McDonald-Kreitman)
- DAF
- Lewontin-Krakauer test
- Ka/Ks
- Empirical distribution (ranking)
- Coalescent simulation of domestication test
- EHH
- LRR
- iHS
- XP-EHH



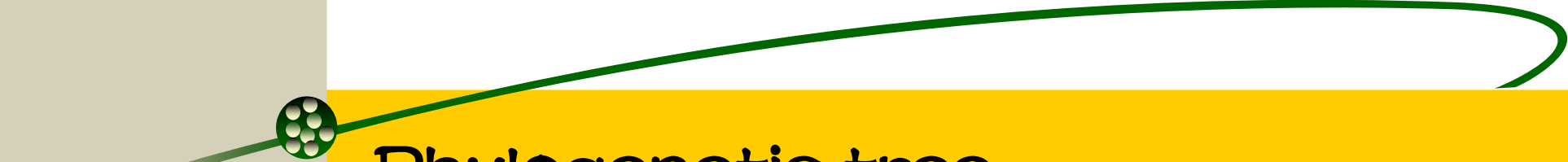
GWAS

- ✿ **A genome-wide association study is defined as any study of genetic variation across the entire genome that is designed to identify genetic associations with observable traits, or the presence or absence of a disease or condition.**
- ✿ **Whole genome information, when combined with phenotype data, offers the potential for increased understanding of basic biological processes affecting human health, improvement in the prediction of disease and patient care, and ultimately the realization of the promise of personalized medicine.**
- ✿ **Rapid advances in understanding the patterns of human genetic variation and maturing high-throughput, cost-effective methods for genotyping are providing powerful research tools for identifying genetic variants that contribute to health and disease.**



Basic Idea A rather vague idea of a study design that involves genotyping cases and controls at a large number ($10^4 \rightarrow 10^6$) of SNP markers spread (in some unspecified way) throughout the genome. Look for associations between the genotypes at each locus and disease status.

0 1 1 1 1 1 0 1 0 2 1 2 2 0 1 0 0 0 1 1	Control
2 0 1 1 1 2 0 0 0 1 0 1 1 0 1 1 0 1 0 0	Control
2 0 1 2 2 0 1 2 1 0 0 1 1 0 1 0 0 1 1 1	Control
1 2 1 1 2 1 1 1 1 0 1 1 1 0 0 2 2 2 0 2	Control
1 1 2 1 0 1 2 1 1 1 1 2 1 2 1 2 1 2 1 1	Case
2 2 1 2 0 1 0 0 0 1 2 2 1 2 1 2 1 0 2 1	Case
0 1 1 0 0 2 1 0 0 2 1 1 1 2 1 1 2 0 1 0	Case
0 1 1 0 0 1 0 2 2 1 1 1 1 2 0 1 2 1 1 2	Case



Phylogenetic tree ——系统发育树

5.1 Introduction

5.2 Distance methods

5.3 Maximum parsimony (MP) and maximum likelihood (ML)



5.1 Introduction

Why do a phylogenetic analysis?

- ✿ important for deciphering relationships in gene function and protein structure and function in different organisms
- ✿ helps to utilize genetic information of a model organism to analyze a second organism
- ✿ helps to sort out gene family relationships
- ✿ valuable tool for tracing the evolutionary history of genes



Evaluating sequence relationships

sequence A ERKSIQDLFQSFTLFERRLLIEF

sequence B ERLSISELIGSLRLYERRLLIEY

sequence C DRKSISDLIGSLRLALLIEF

sequence D DRKDLISSLRKALLIEW

1. Account for all column variations

|A,B and C,D form similar groups based on col. 1

|A,C,D based on col. 3

2. Count differences between sequences

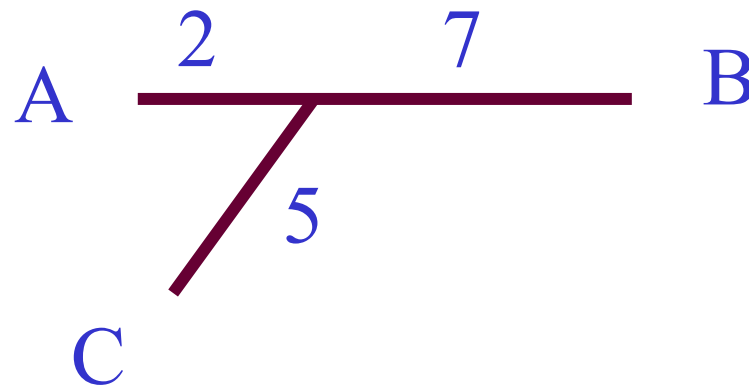
A,B 17/23 similar, 6/23 different

C,D 21/23 similar, 2/23 different

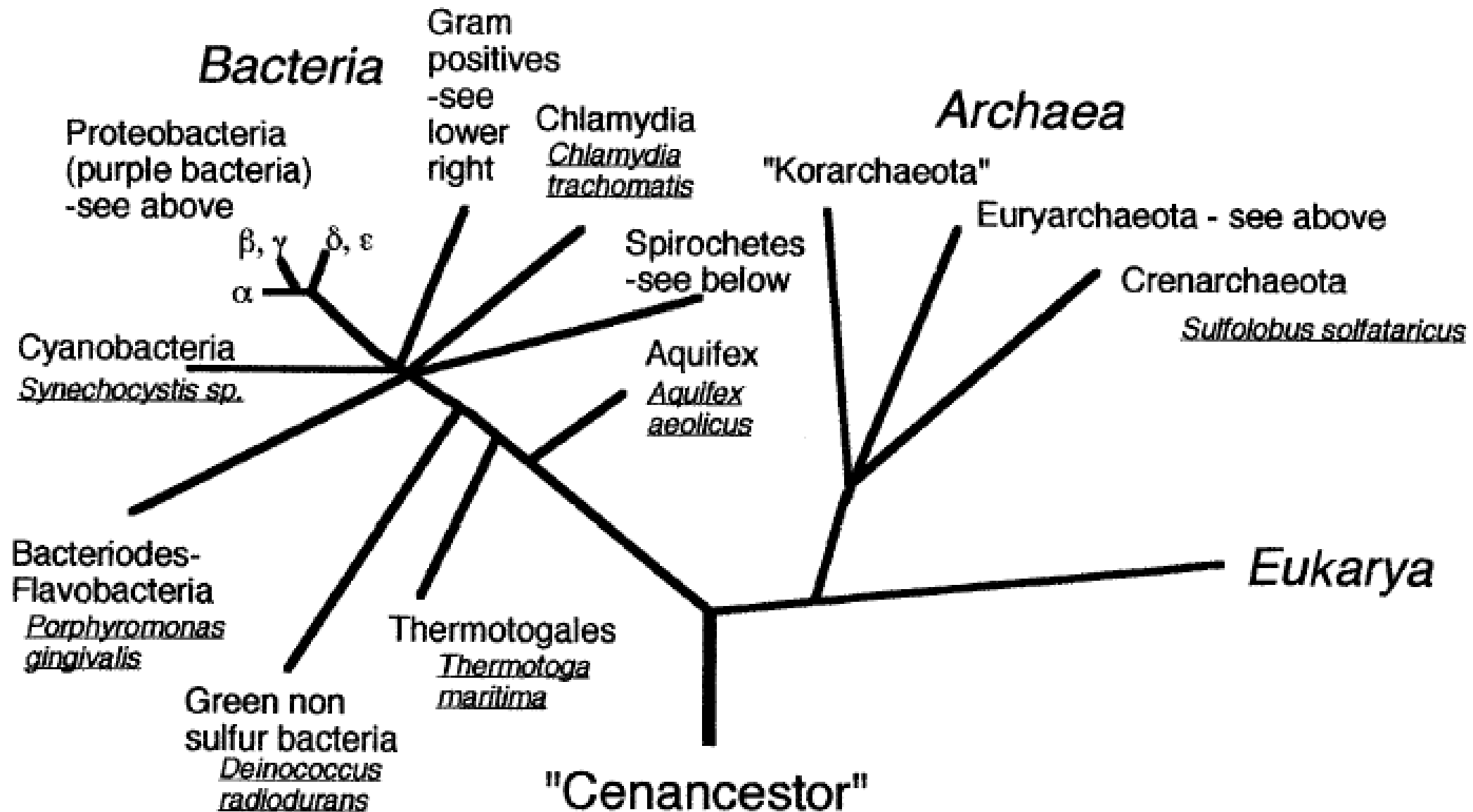
What is a tree?

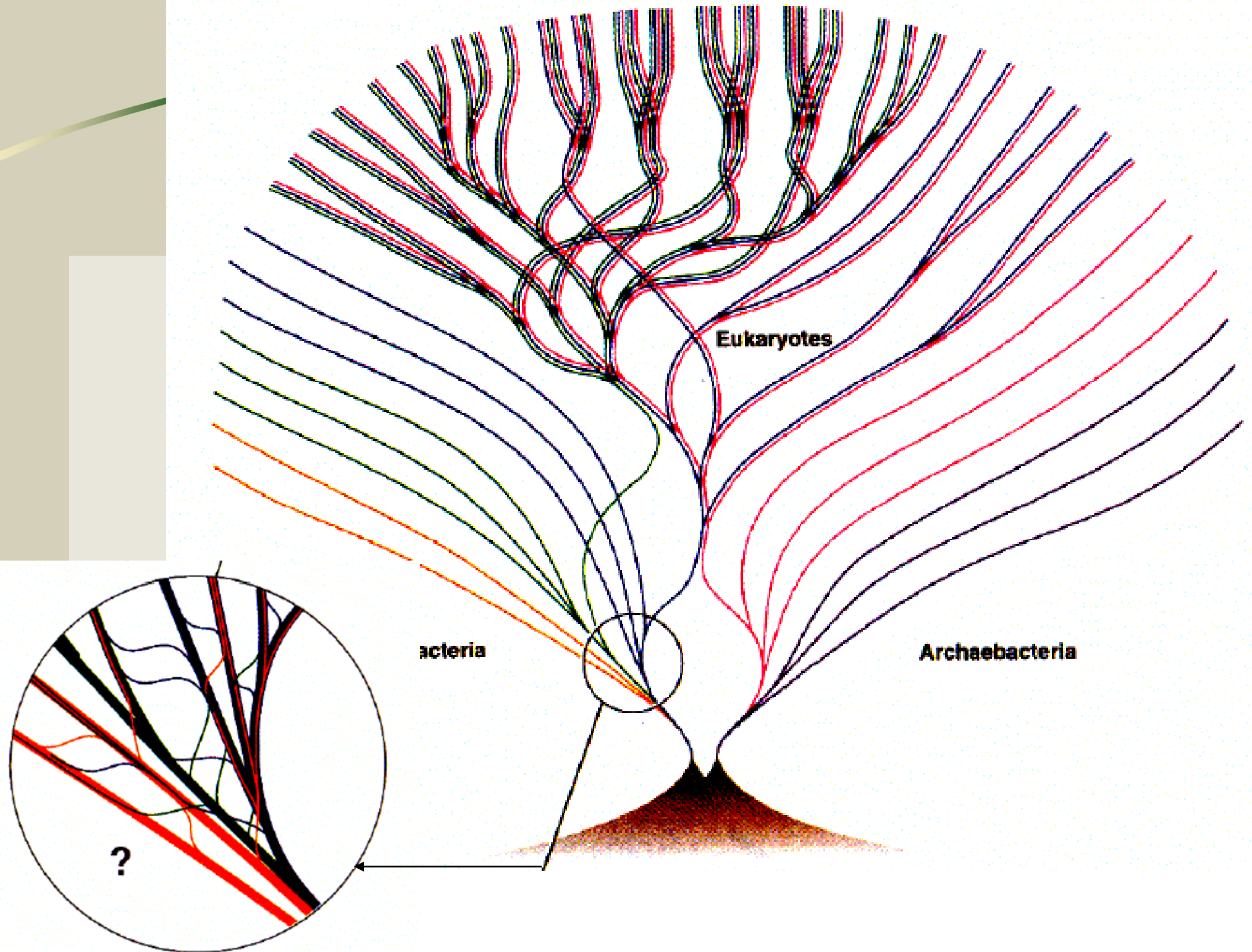
- ✂ a graphical representation of the sequence similarities among a group of nucleic acid or protein sequences
- ✂ for example: no of differences between 3 sequences may be represented by ..

	A	B	C
A		9	7
B			12



Tree of life





Phylogenetic tree (dendrogram)

Nodes: branching points

Branches: lines

Topology: branching pattern

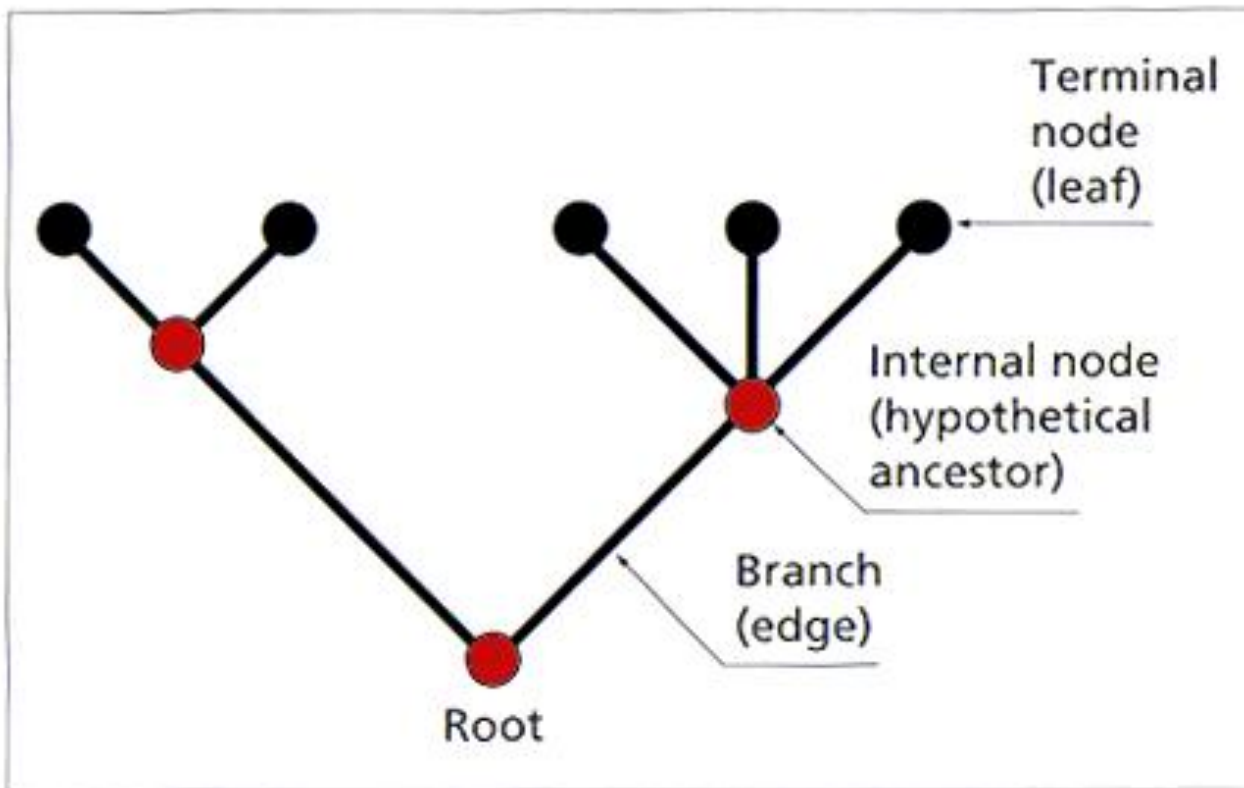
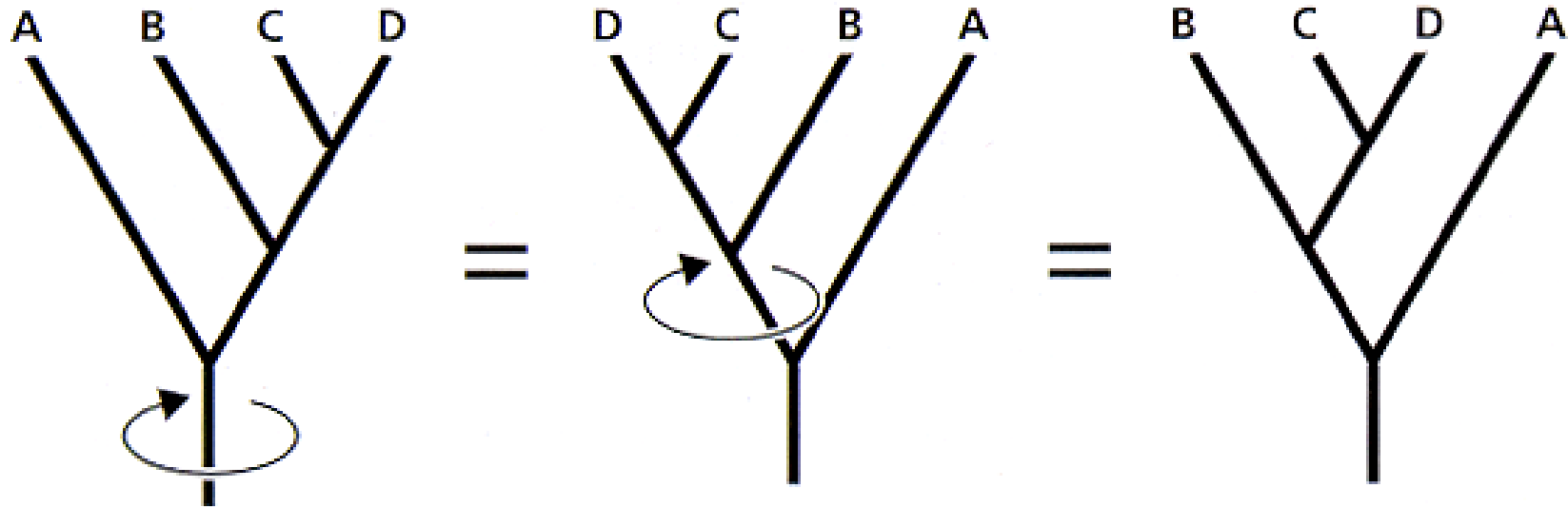


Fig. 2.1 A simple tree and associated terms.

Branches can be rotated at a node, without changing the relationships among the OTU's.



Rooted: unique path from root.
Unrooted: degree of kinship, no evolutionary path.

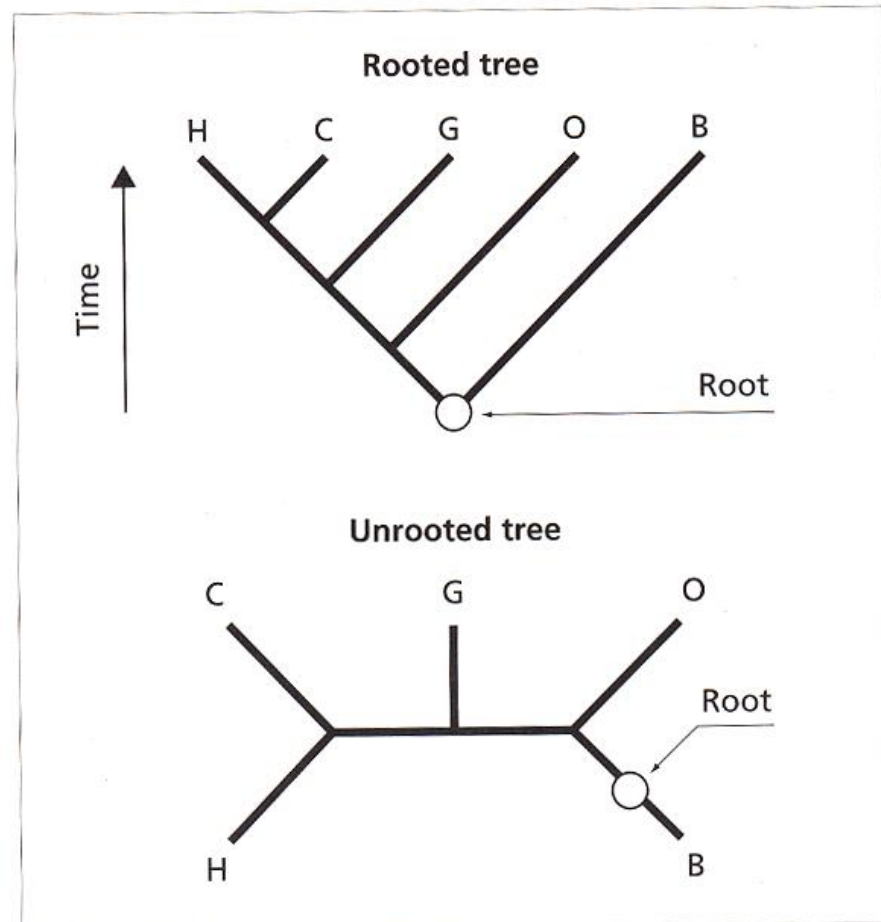


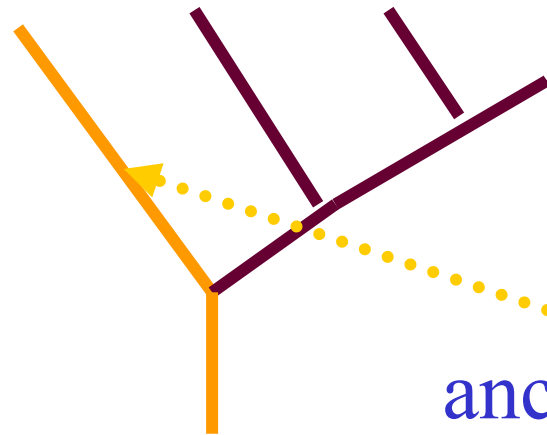
Fig. 2.6 Rooted and unrooted trees for human (H), chimp (C), gorilla (G), orang-utan (O), and gibbon (B). The rooted tree (top) corresponds to the unrooted tree below.

Assumptions about rate of change in branches of tree

- may assume constant rate of change throughout tree - then branch length is proportional to no. of changes and we can easily root the tree using a simple algorithm OR
- may have variable rate and then root can be any branch



no ancestor sequence



ancestor sequence



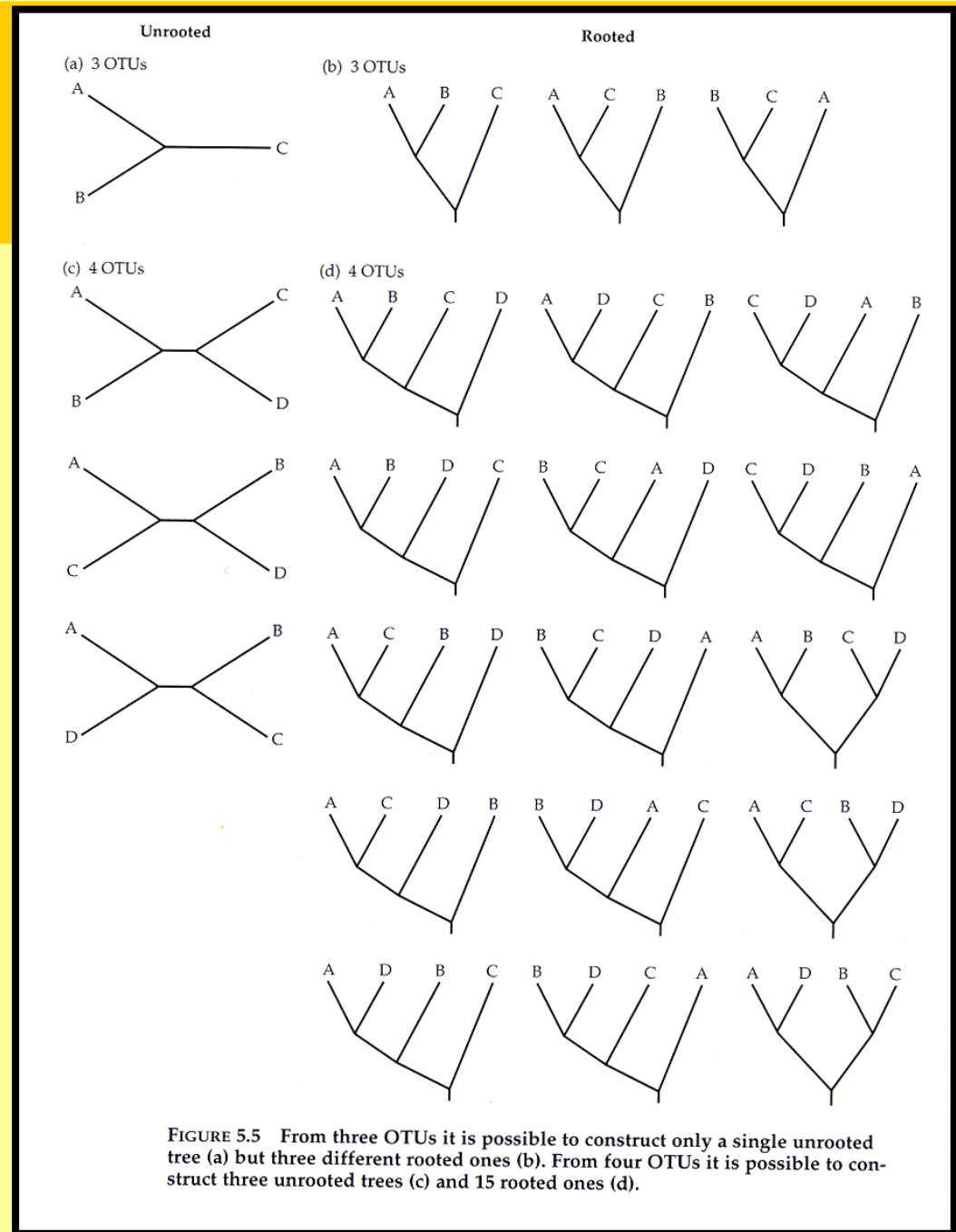
Tree reconstruction as optimization

- ✱ Number of possible trees is 2^S . Considering branch lengths makes the problem harder. Almost as many unrooted trees.
- ✱ Want the tree that maximizes some quality score. Score based on either
 - Characters (e.g. sites in a sequence) directly, or
 - Distances between character sets (e.g. alignment scores)
- ✱ Another NP-hard optimization problem...

Number of possible phylogenetic trees

**3 OTU's: 1 unrooted tree
3 rooted trees**

**4 OTU's: 3 unrooted trees
15 rooted trees.**



实用分类单位(operational
taxonomic units, OTU)

TABLE 5.1 Numbers of possible rooted and unrooted trees for up to 20 OTUs

Number of OTUs	Number of rooted trees	Number of unrooted trees
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
16	6,190,283,353,629,375	213,458,046,676,875
17	191,898,783,962,510,625	6,190,283,353,629,375
18	6,332,659,870,762,850,625	191,898,783,962,510,625
19	221,643,095,476,699,771,875	6,332,659,870,762,850,625
20	8,200,794,532,637,891,559,375	221,643,095,476,699,771,875

Data from Felsenstein (1978b).

TYPES OF TREES

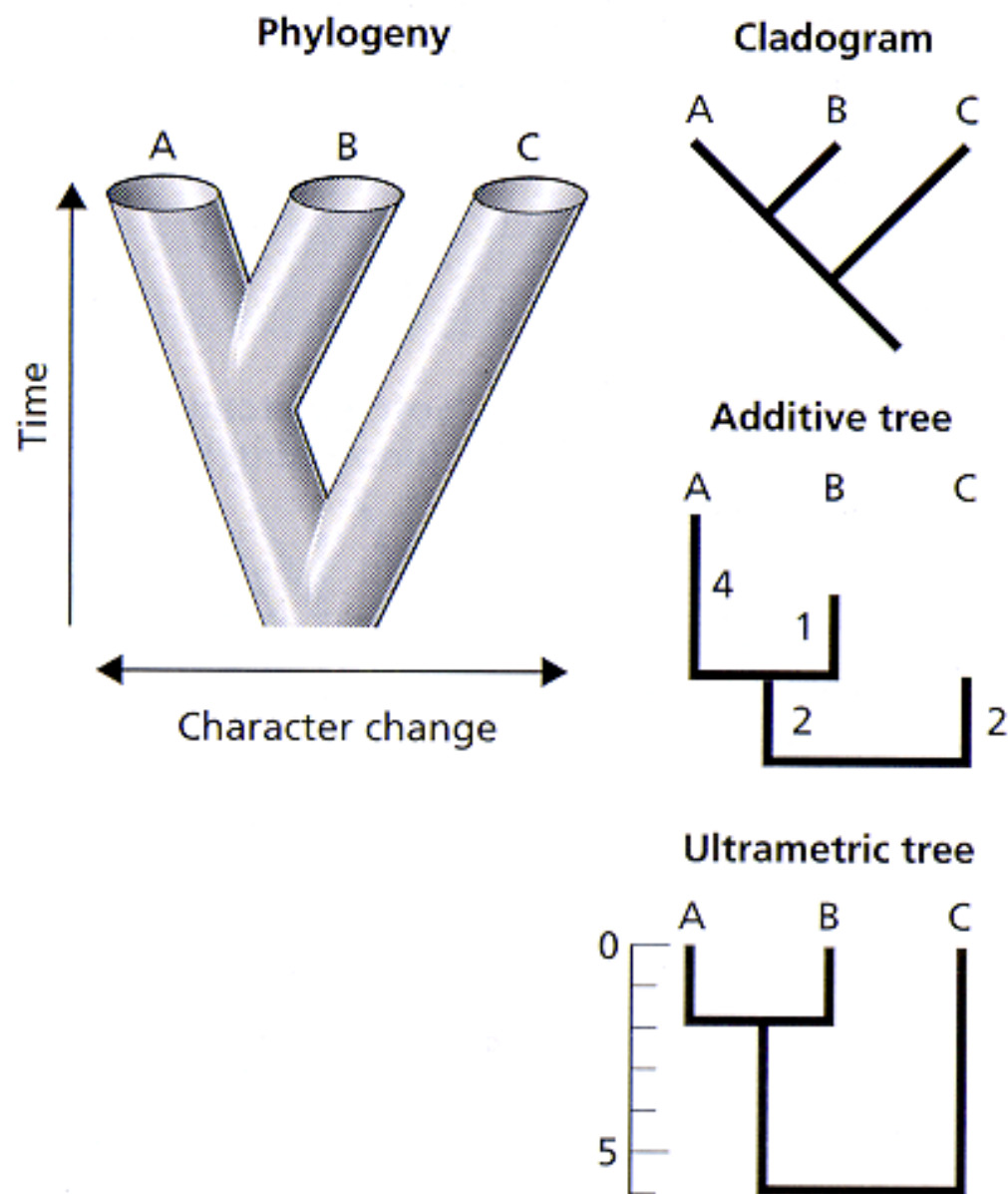


Fig. 2.5 A phylogeny and the three basic kinds of tree used to depict that phylogeny. The cladogram represents relative recency of common ancestry; the additive tree depicts the amount of evolutionary change that has occurred along the different branches, and the ultrametric tree depicts times of divergence.

Newick (shorthand) format

- text based representation of relationships.

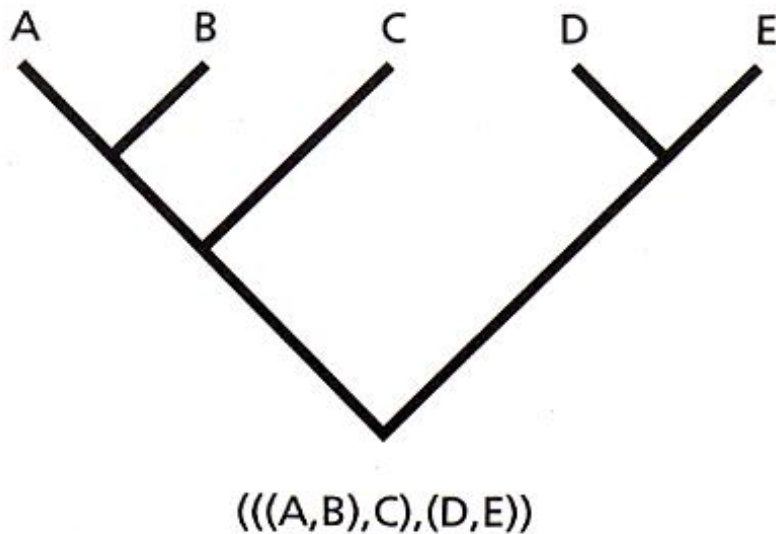
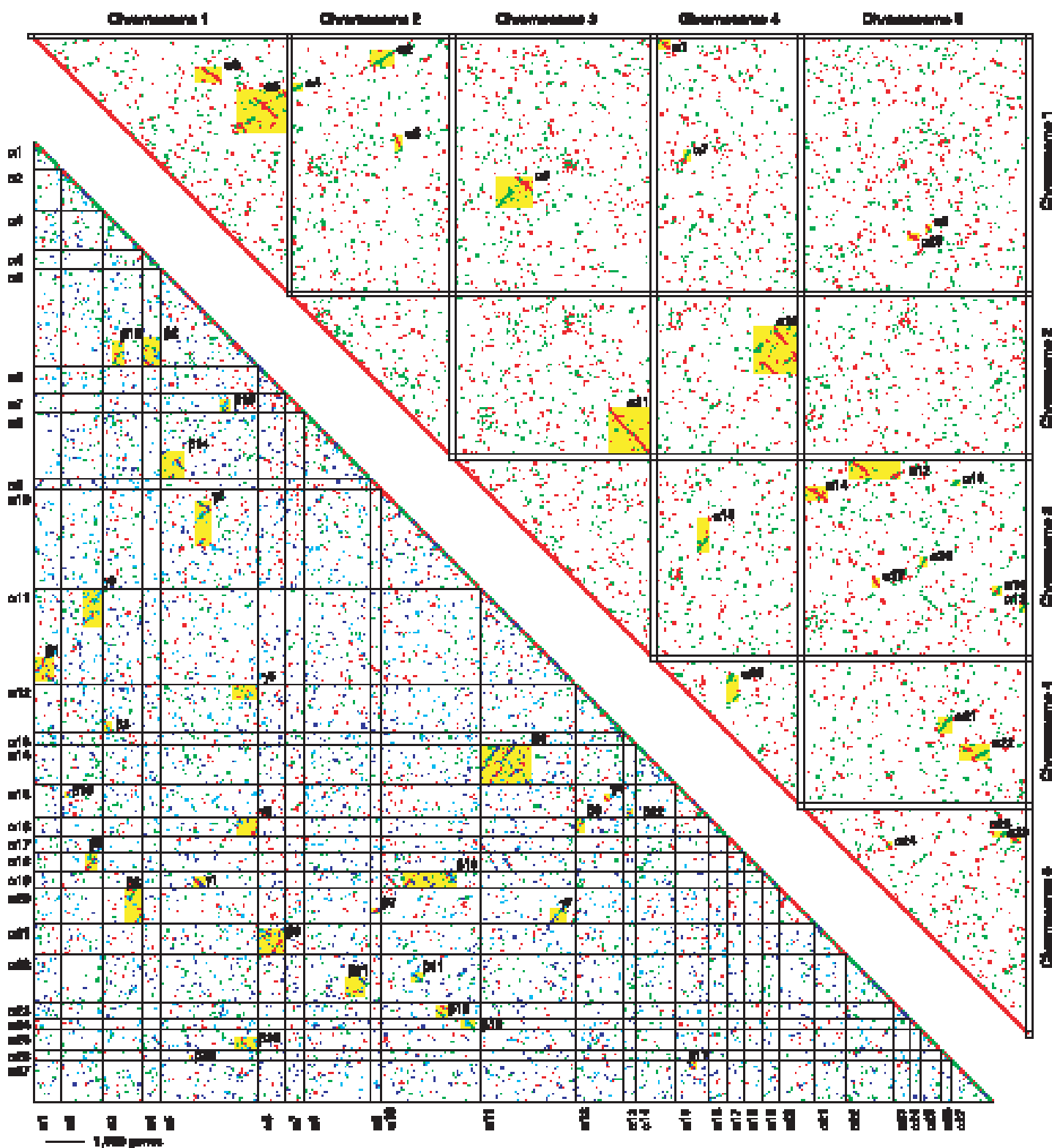


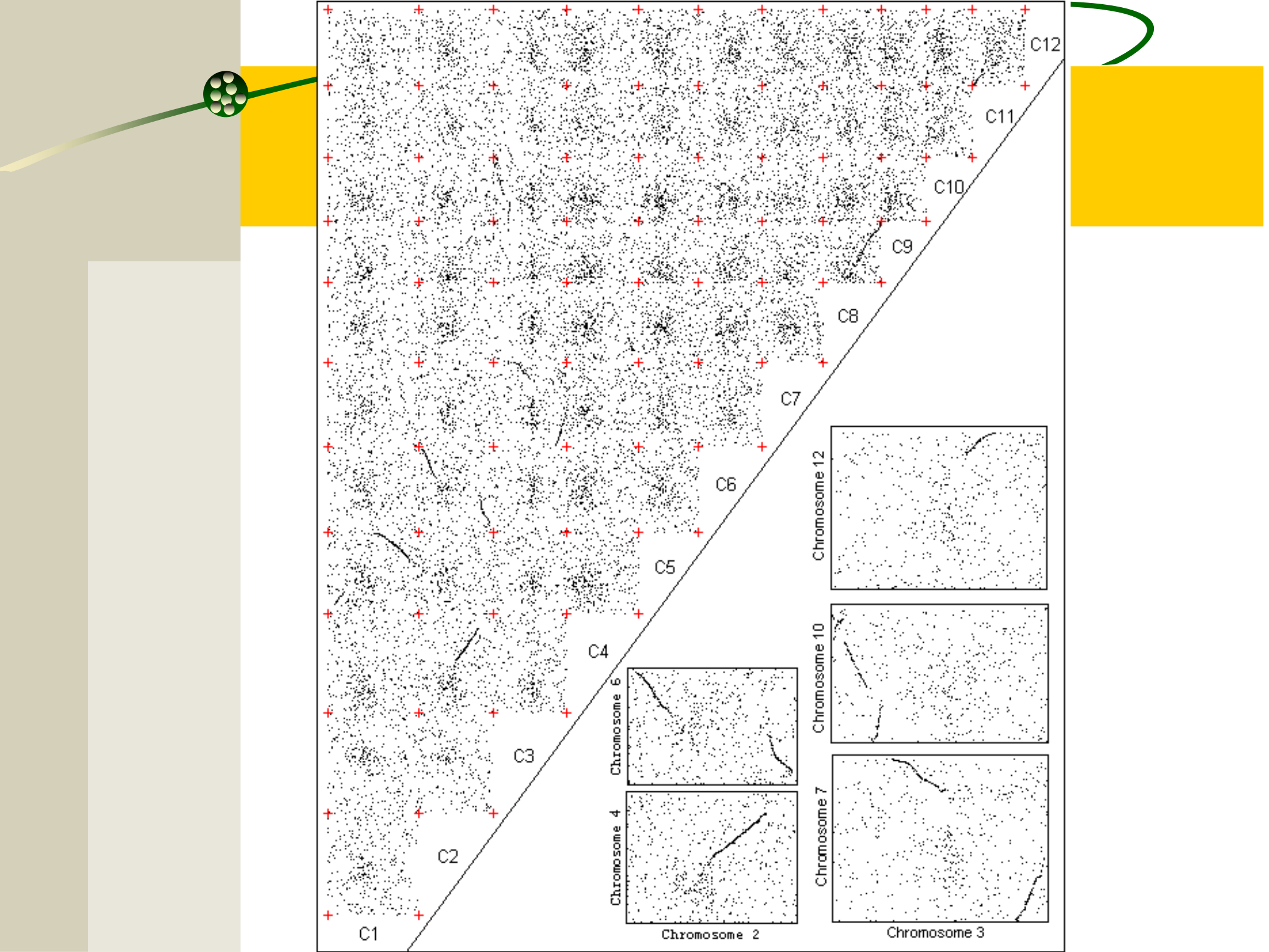
Fig. 2.4 A tree and its shorthand representation using nested parentheses.



Phylogenomics Tree

Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events
 John E. Bowers, Brad A. Chapman, Junkang Rong & Andrew H. Paterson

Nature, 2003, VOL 422, 27





5.2 Distance methods

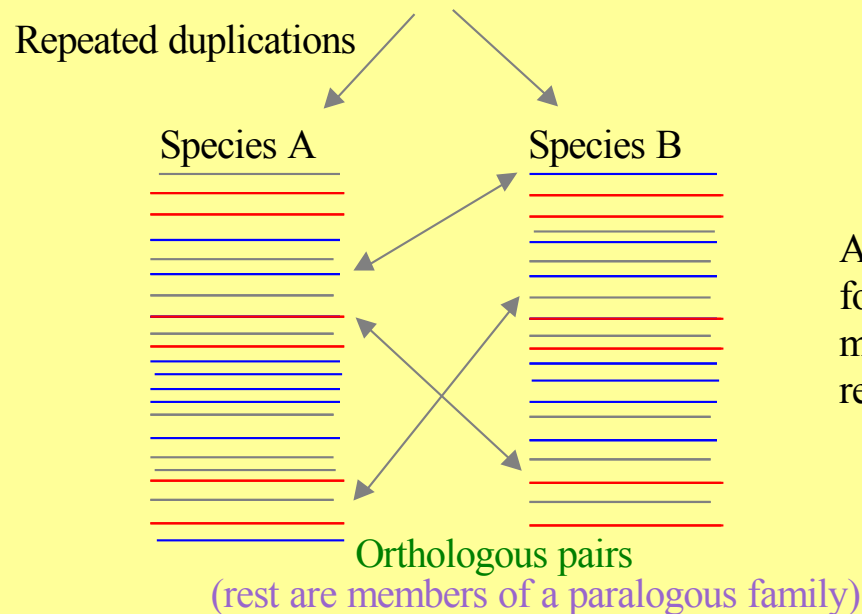
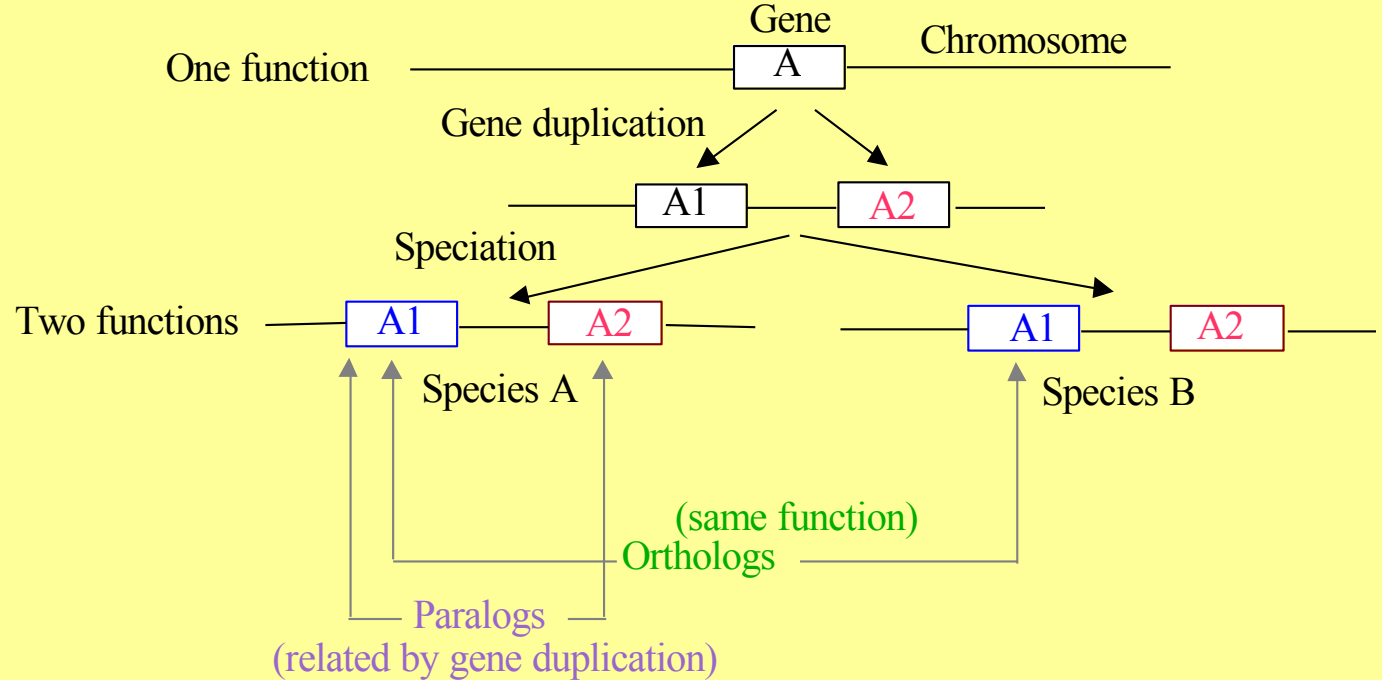


Basic idea:

- Employs the number of changes between each pair in a group of sequences to produce a phylogenetic tree of the group;
- For phylogenetic analysis, the distance score between two sequences is used. This score between two sequences is the number of mismatched positions in the alignment or the number of sequence positions that must be changed to generate the other sequence. Gaps may be ignored in these calculations or treated like substitutions.
- When a scoring or substitution matrix is used, the calculation is slightly more complicated, but the principle is the same.

ORIGIN OF GENE FAMILIES

Theory of gene duplication



All- by- all sequence analysis followed by single- linkage or maximal linkage clustering reveals orthologs and paralogs.

Homoplasy: Sequence similarity NOT due to shared ancestry, such as convergent or parallel evolution; horizontal transmission

祖先序列

A
C
T
G
A
A
C
G
T
A
A
C
G
C

A
C
T
G
A→C→T
A
C→G
G
T→A
A
A→C→T
C
G
C

A
C→A
T
G
A
A
C→A
G
T→A
A
A→T
C
G
C→T→C

单一替换 (single substitution)

多重替换 (multiple substitutions)

同义替换 (coincidental substitutions)

平行替换 (parallel substitutions)

趋同替换 (convergent substitution)

反转替换 (back substitution)

遗传模型和序列距离

Jukes-Cantor单参数模型 (1969)

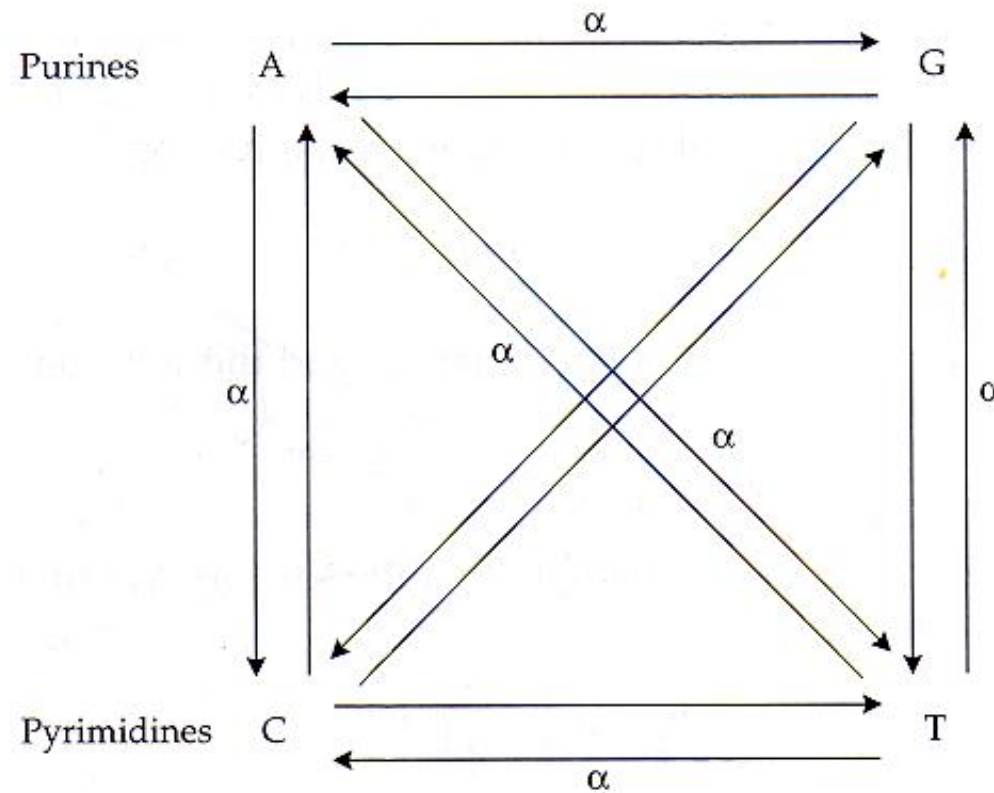


FIGURE 3.1 One-parameter model of nucleotide substitution. The rate of substitution in each direction is α .



Kimura两参数模型 (Motoo Kimura, 木村资生, 1980) :

转换 (transition, α): 嘌呤间或嘧啶间的替换;

颠换 (transversion, β): 嘌呤和嘧啶间的替换。

在大多数DNA片段中, 转换出现的频率高于颠换。

同义与非同义替换

在蛋白质编码基因中，仍为同义密码子的核苷酸替换称为同义或沉默替换（synonymous/silent substitution），而导致非同义密码子的替换，称为非同义或氨基酸更换替换（nonsynonymous/amino acid replacement substitution）。另外，导致形成终止密码子的突变，称为无义突变（nonsense mutation）。

序列距离----遗传距离

Generally, genetic distance is thought of as related to the time and requires a genetic model specifying the processes such as mutation and drift causing the populations to diverge.

Geometric distance: Euclidean distance

Jukes和Cantor(1969)提出了DNA序列距离 K (最早为氨基酸序列引入)计算公式:

$$K = \frac{3}{4} \ln\left(\frac{3}{4q-1}\right) \approx 2\mu t$$

其中 q 为经过 t 世代祖先序列的趋异变化后同源DNA序列中具有相同碱基的概率, μ 为碱基替换频率。



$$q_t = \frac{1}{4} + \frac{3}{4} \left(1 - \frac{8\mu}{3}\right)^t$$

Kimura在其两参数模型下证实，由于趋异变化，由转换造成差异(I型变化)或由颠换造成差异(II型变化)的碱基，随时间而变化。如果 $k = \alpha + 2\beta$ 是单位时间碱基替换的总频率，则

$$K = -\frac{1}{2} \ln[(1 - 2p_I - p_{II})\sqrt{1 - 2P_{II}}] \approx 2kt$$

分歧度

(sequence divergence)

DNA序列距离 K 又可称为DNA序列间的分歧度，即序列间相异性的一个指标。蛋白质序列的分歧度分为两序列同义变化的分歧度(K_S)和非同义变化的分歧度(K_A)。根据Jukes-Cantor单参数模型和Kimura两参数模型等遗传模型，可以分别计算得到两序列的分歧度(或称为蛋白质序列间的距离)。

A simple example

A. Sequences

sequence A	A C G C G T T G G G C G A T G G C A A C
sequence B	A C G C G T T G G G C G A C G G T A A T
sequence C	A C G C A T T G A A T G A T G A T A A T
sequence D	A C A C A T T G A G T G A T A A T A A T

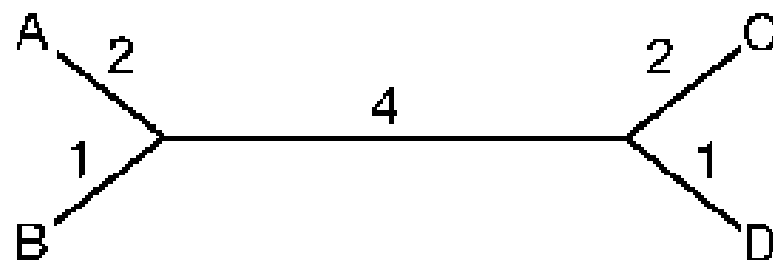
B. Distances between sequences, the number of steps required to change one sequence into the other.

n_{AB}	3
n_{AC}	7
n_{AD}	8
n_{BC}	6
n_{BD}	7
n_{CD}	3

C. Distance table

	A	B	C	D
A	–	3	7	8
B	–	–	6	7
C	–	–	–	3
D	–	–	–	–

D. The assumed phylogenetic tree for the sequences A-D showing branch lengths. The sum of the branch lengths between any two sequences on the trees has the same value as the distance between the sequences.



UPGMA

- ✱ Unweighted pair group method using arithmetic mean (平均连接聚类法).
- ✱ A form of agglomerative clustering (successive pairing of closest nodes)
- ✱ Method:
 - Let initial organisms be leaves of the tree.
 - Iteratively add a parent to the closest pair of existing nodes
 - Define the distance from an internal node to be the mean distance to its children.

实例：线粒体DNA序列

5个线粒体序列的差异核苷酸数(对角线下)和Jukes-Cantor距离(对角线上)

	人 类 (hu)	黑 猩 猩 (ch)	大 猩 猩 (go)	猩猩(or)	长 臂 猿 (gi)
人 类(hu)	–	0.015	0.045	0.143	0.198
黑 猩 猩 (ch)	1	–	0.030	0.126	0.179
大 猩 猩 (go)	3	2	–	0.092	0.179
猩猩(or)	9	8	6	–	0.179
长 臂 猿 (gi)	12	11	11	11	–

最近的距离是人类和黑猩猩之间的，将它们合并为一个类。其它序列与这个新类之间的距离就是该序列到新类各成员间的平均距离：

$$d_{(hu-ch),go} = \frac{1}{2}(d_{hu,go} + d_{ch,go}) = 0.037$$

$$d_{(hu-ch),or} = \frac{1}{2}(d_{hu,or} + d_{ch,or}) = 0.135$$

$$d_{(hu-ch),gi} = \frac{1}{2}(d_{hu,gi} + d_{ch,gi}) = 0.189$$

	(hu-ch)	go	or	gi
hu-ch		0.037	0.135	0.189
go				0.179
or				0.179
gi				



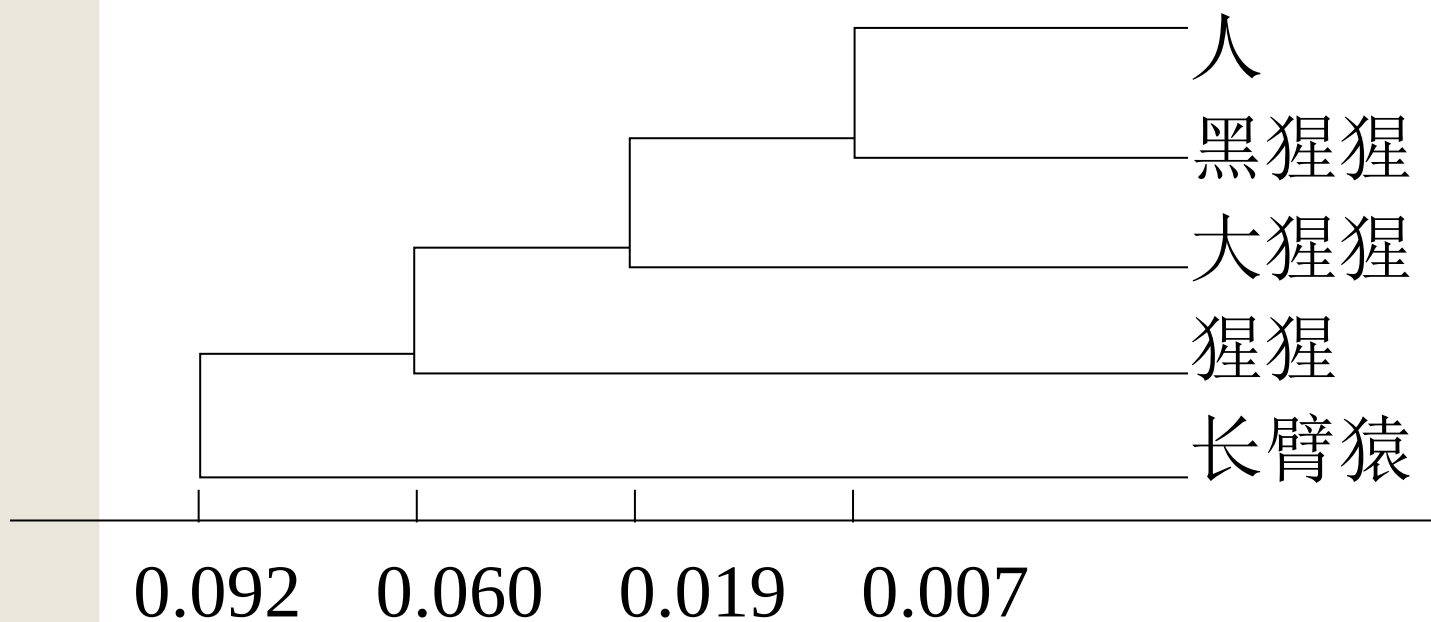
其中人类—黑猩猩（**hu-ch**）与大猩猩（**go**）之间的距离最小。将它们合并为一类。新距离为：

$$d_{(hu-ch-go),gi} = \frac{1}{3}(d_{hu,gi} + d_{ch,gi} + d_{go,gi}) = 0.185$$

$$d_{(hu-ch-go),or} = \frac{1}{3}(d_{hu,or} + d_{ch,or} + d_{gp,pr}) = 0.121$$

	(hu-ch-go)	or	gi
(hu-ch-go)		0.121	0.185
or			0.179
gi			

平均连接聚类法系统树



Neighbor Joining Method (邻接法)

✂ Another agglomerative clustering method

✂ Method:

- Let A be active set, initially all leaves (organisms)
Let $r(i)$ be the average distance from i to all other leaves
- Define $D(i,j) = d(i,j) - (r(i) + r(j))$
- Pick i,j with smallest $D(i,j)$. Create parent k of i and j .
- Set $d(i,k) = \frac{1}{2} [d(i,j) + r(i) - r(j)]$, and $d(j,k) = d(i,j) - d(i,k)$
- For all other $m \in A$, set $d(k,m) = \frac{1}{2} [d(i,m) + d(j,m) - d(i,j)]$
- Replace i and j in A with k .
- Repeat until $|A| = 1$.

邻接法的一般步骤

- ① 计算第*i*终端节点(即分类单位*i*)的净分歧度 r_i

$$r_i = \sum_{k=1}^N d_{ik}$$

其中*N*为终端节点数, d_{ik} 为节点*i*和节点*k*之间的距离, 有 $d_{ik}=d_{ki}$

- ② 计算并确定最小速率校正距离 (rate-corrected distance) M_{ij} :

$$M_{ij} = d_{ij} - \frac{r_i + r_j}{N - 2}$$



③定义一个新节点 u ， u 节点由节点 i 和 j 组合而成。
节点 u 与节点 i 和 j 的距离为：

$$S_{iu} = \frac{d_{ij}}{2} + \frac{r_i + r_j}{2(N-2)} \quad S_{ju} = d_{ij} - S_{iu}$$

节点 u 与系统树其它节点 k 的距离为：

$$d_{ku} = \frac{d_{ik} + d_{jk} - d_{ij}}{2}$$

④从距离矩阵中删除列节点 i 和 j 的距离， N 值(总节点数)减去1

⑤如果尚余2个以上终端节点，返回到步骤①继续计算，直至系统树完全建成。

同样以线粒体DNA为例

邻接法计算线粒体序列(图5.4)的距离 d_{ij} (上对角线部分)和 M_{ij} (下对角线部分)

		hu	ch	go	or	gi	净分歧度
		j=1	j=2	j=3	j=4	j=5	r_i
hu	i=1	0.000	0.015	0.045	0.143	0.198	0.401
ch	i=2	-0.235	0.000	0.030	0.126	0.179	0.350
go	i=3	-0.204	-0.202	0.000	0.092	0.179	0.346
or	i=4	-0.171	-0.171	-0.203	0.000	0.179	0.540
gi	i=5	-0.181	-0.183	-0.181	-0.246	0.000	0.735



第一步，or和gi之间的 M_{ij} 值最小，则它们用节点1取代，进入第2步，则新节点(节点1)到这二个节点的距离为：

$$d_{gi, \text{节点1}} = d_{or, gi} - d_{or, \text{节点1}} = 0.122$$

$$d_{or, \text{节点1}} = \frac{1}{2} d_{or, gi} + \frac{r_{or} - r_{gi}}{6} = 0.057$$



		hu	ch	go	节点1	
		j=1	j=2	j=3	j=4	r_i
hu	i=1	0.000	0.015	0.045	0.081	0.141
ch	i=2	-0.110	0.000	0.030	0.063	0.108
go	i=3	-0.086	-0.084	0.000	0.046	0.121
节点1	i=4	-0.085	-0.086	-0.110	0.000	0.190



		go	节点1	节点2	
		j=1	j=2	j=3	r_i
go	i=1	0.000	0.046	0.030	0.076
节点1	i=2	-0.141	0.000	0.065	0.111
节点2	i=3	-0.141	-0.141	0.000	0.095



		go	节点3
		j=1	j=2
go	i=1	0.000	0.005
节点3	i=2		0.000

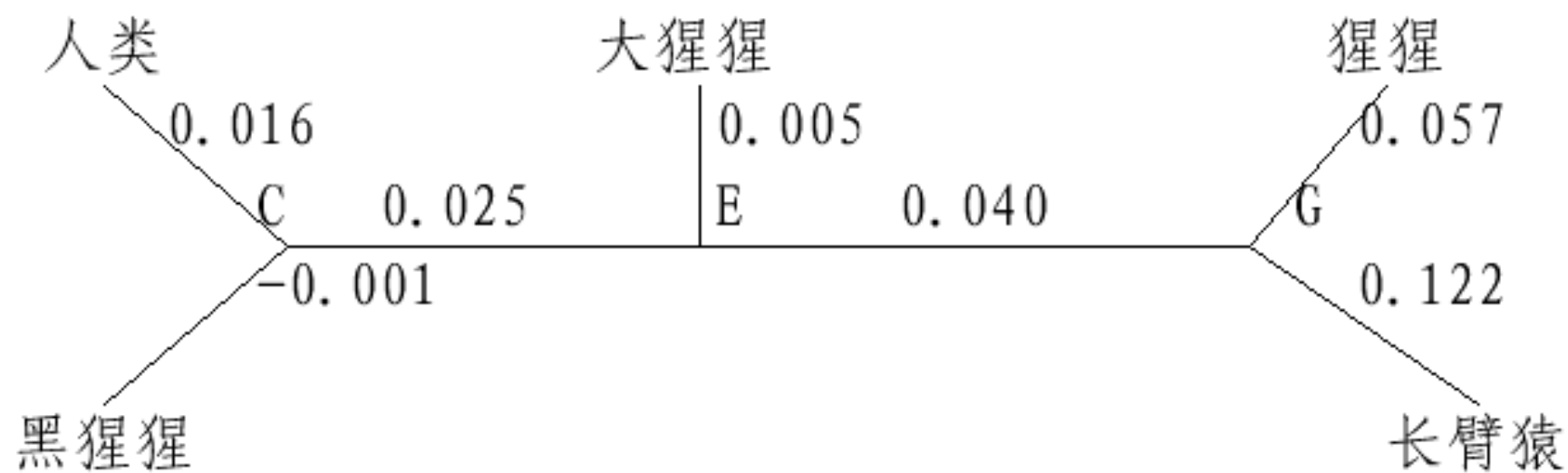


图 5.10 根据线粒体序列 (图 5.4) 构建邻接法无根系统树



5.3 Maximum parsimony method (简约法)

Parsimony:

General scientific criterion for choosing among competing hypotheses that states that we should accept the hypothesis that explains the data most simply and efficiently.

Maximum parsimony method of phylogeny reconstruction:

The optimum reconstruction of ancestral character states is the one which requires the fewest mutations in the phylogenetic tree to account for contemporary character states.

First step : Identify all of the informative sites

Sequence	Site								
	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T
					*		*		*

Invariant: all OTU's possess the same character state at the site.

Any invariant site is **uninformative**.

对于DNA序列，信息位点是指那些至少存在2个不同的碱基且每个不同碱基至少出现两次的位点。只有一个碱基且只在一个序列中出现的位点不属于信息位点，因为那种独特的碱基位点是由于在直接通向它所在序列的分枝上发生单个碱基变更所引起的。这种碱基变更可与任何拓扑结构相容。

Sequence	Site								
	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T
					*		*		*

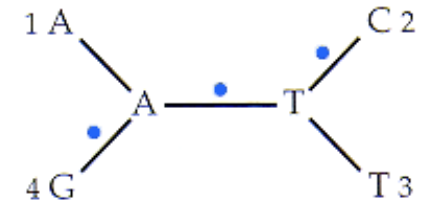
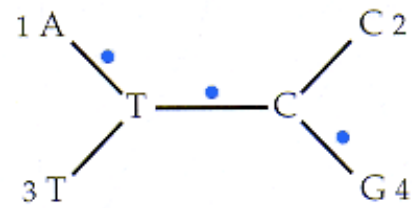
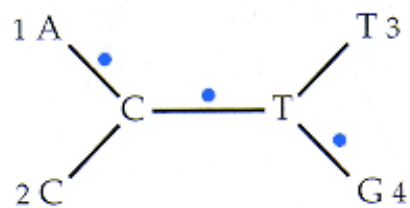
Two types of variable sites:

Informative: favors a subset of trees over other possible trees.

Uninformative: a character that contains no grouping information relevant to a cladistic problem (i.e. autapomorphies).

Sequence	Site								
	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T
					*		*		*

(b)

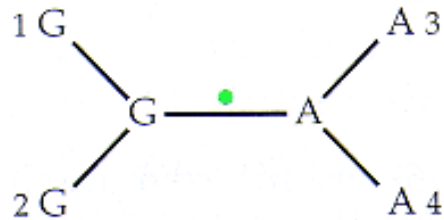


Uninformative: each tree 3 steps

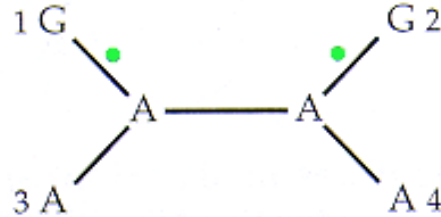
2nd step: Calculate the minimum number of substitutions at each informative site

Sequence	Site								
	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T
					*		*		*

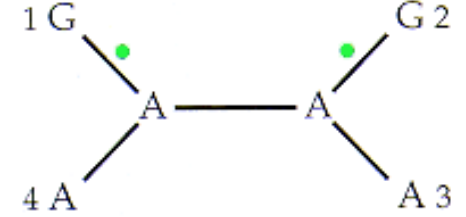
(c)



1 step



2 steps

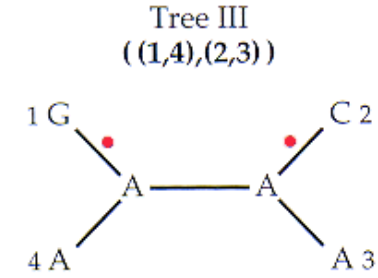
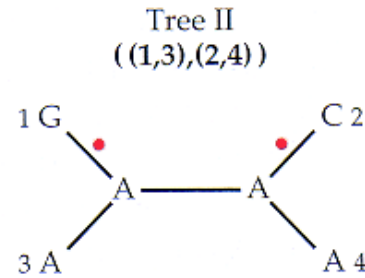
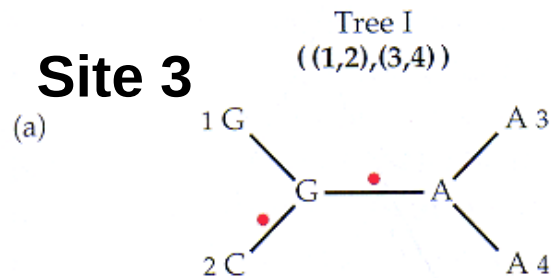


2 steps

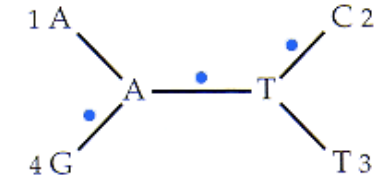
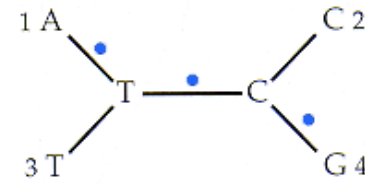
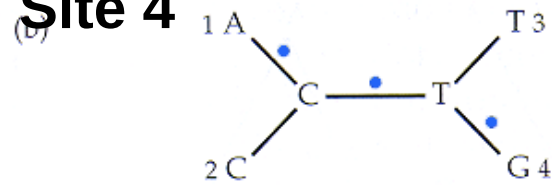
Informative: favors tree 1 over other 2 trees.

Final step: Sum the number of changes over all informative sites for each possible tree and choose the tree associated with the smallest number of changes.

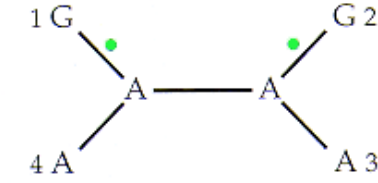
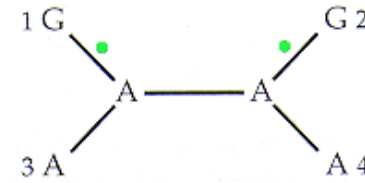
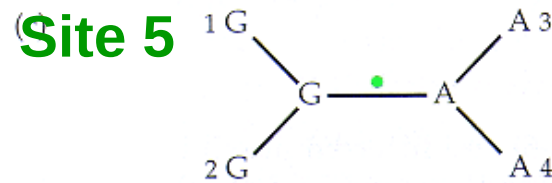
Site 3



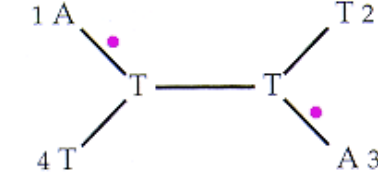
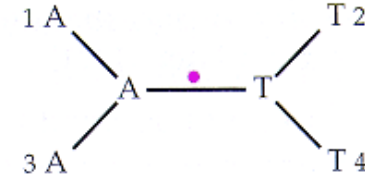
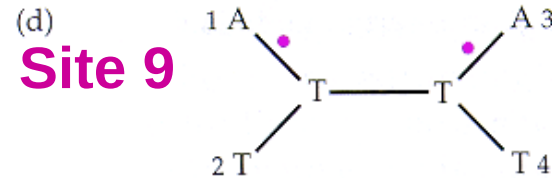
Site 4



Site 5



Site 9



3 steps

3 steps

4 steps



Parsimony Search Methods

Exhaustive search method: searches all possible fully resolved topologies and guarantees that all of the minimum length cladograms will be found. (not a practical option, time consuming)

Branch and bound methods: begins with a cladogram. The length of starting cladogram is retained as an upper bound for use during subsequent cladogram construction. As soon as a length of part of the tree exceeds the upperbound, the cladogram is abandoned. If equal length, cladogram is saved as an optimal topology. If length is less, it is substituted for the original as the optimal upperbound. (good option for fewer than 20 taxa, time consuming)

Heuristic methods: approximate or “hill climbing technique”
Begin with a cladogram, add taxa and swap branches until a shorter length cladogram is found. Procedure can be replicated many times to increase chance of finding minimum length cladogram.



Different types of parsimony analyses:

Unweighted parsimony: all character state changes are given equal weight in the step matrix.

Weighted parsimony: different weights assigned to different character state changes.

Transversion parsimony: transitions are completely ignored in the analysis, only transversions are considered.

Maximum Likelihood Estimate

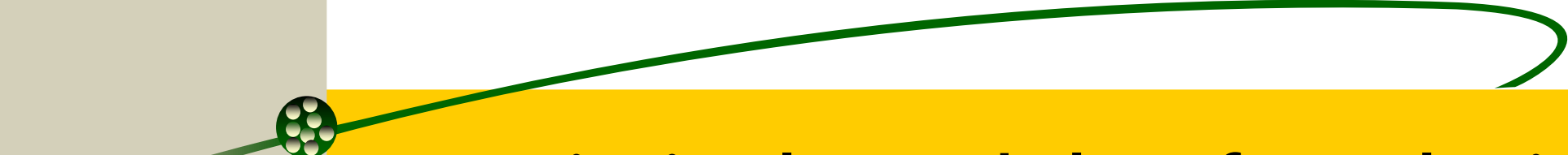
- ✱ Consider Bayes rule:
$$P(\text{model}|\text{data}) = P(\text{data}|\text{model}) * P(\text{model}) / P(\text{data})$$
- ✱ $P(\text{data})$ is constant for any problem. If priors over model are uniform, the model that gives the highest likelihood for the data is the best.
- ✱ If we have a parameterized statistical model of evolution, then we could find most likely set of parameters given a set of observations
- ✱ Now we have a numeric optimization problem to score even a single tree!

ML algorithms (1)

- ✱ This method uses probability calculations to find a tree that best accounts for the variation in a set of sequences.
- ✱ The method is similar to the maximum parsimony method in that the analysis is performed on each column of a multiple sequence alignment. All possible trees are considered. Hence, the method is only feasible for a small number of sequences.
- ✱ The method can be used to explore relationships among more diverse sequences, conditions that are not well handled by maximum parsimony methods.

ML algorithms (2)

- ✦ Given a tree topology, optimize score (likelihood) by setting lengths and parameters (e.g. evolution rates) for each edge
- ✦ Heuristic search through possible tree topologies.
 - Greedy approaches (or beam search)
 - Branch swapping
 - Start with MP as approximate solution?



Statistical models of evolution

- ✶ Various models of observable sequences, based on phylogenetic tree + randomness
 - **IID models:** If characters (e.g. sites in an alignment) are independently and identically distributed, we can consider just a single site.
 - **Markov models:** If probability of transitioning among character values depends only on current state, we have a Markov chain from ancestors to descendents
 - Discrete (transition matrix)
 - Continuous time (exponentiated instantaneous matrix $\ast t$)

More statistical models

✶ Lots of 4 state (DNA) models

- Jukes-Cantor: all off-diagonal transition probabilities equal
- More freedom = more parameters (e.g. Kimura 2 param)

✶ Cavender-Felsenstein (CF) model

- Simplest possible Markov model.
- Two states (0/1). Fixed root with specified state distribution
- Each edge specifies mutation probability ≤ 0.5
- Each site evolves i.i.d. according to a Markov process
- Equivalent to a model where each edge gets an expected number of changes that is Poisson distributed.



Still more statistical models

✧ Mixture models

- Not a single Markov process, but a mixture (e.g. drawing at random from two distributions of characters)
- Non-iid: characters drawn from different classes of distributions (e.g. different positions in a codon)
- Classes with differently parameterized Poisson processes mixed based on prior distributions of parameters drawn from gamma distributions are widely used.

✧ More complex models do not necessarily give better trees given limited data size!



Statistical performances

- ✱ Identifiability: a set of trees are identifiable under a model if the distributions of characters are disjoint for different tree topologies (with trivial exceptions, like 0 branch lengths)
- ✱ Consistency: A method is consistent if it will recover the true tree with arbitrarily high probability given enough data (i.e. a long enough sequence)



Identifiability

- ✂ Under iid Markovian models, trees are identifiable
- ✂ If a particular mixture is known and identical over all possible trees, then the trees are identifiable
- ✂ Unclear for more complex mixture models (like the Poisson / gamma distribution)



Consistency

- ✱ Maximum parsimony is not always statistically consistent, even in very simple models like CF
- ✱ However, in simulation studies parsimony often performs well, so its consistency under biologically realistic conditions may be good
- ✱ For many problems, maximum likelihood estimators are provably consistent.



MP versus ML

- ✱ MP is much faster, since scoring a tree is linear in tree size. Scoring an ML tree requires optimizing the edge weights (which may have multiple parameters per edge).
- ✱ In some circumstances, MP and ML give identical answers
 - When there are few mutations
 - In a model where sites are independent, but not necessarily identically distributed, and there is no assumption of scaling across the sites

16S ribosomal trees

- ✂ Pioneered by Woese
 - Archaeal kingdom hypothesis came from 16S tree!
- ✂ Exist in all organisms, highly conserved
- ✂ Suitable for very broad phylogenies
- ✂ Sequenced in tens of thousands of organisms
- ✂ Not appropriate for close relatives...



Whole genome trees

- ✧ Characteristics over entire genomes
 - Presence/absence of homologous genes
 - Chromosomal rearrangements (inversions, translocations)
- ✧ Good for distant taxa
- ✧ Mostly practical for microbes, since relatively few completely sequenced genomes of others yet.



Multiple hit correction

- ✂ A site can change and then change back to where it was during the course of evolution.
- ✂ Poisson correction: assume that changes after time t are independent of ones before t . $D(i,j) = -\frac{1}{2} \log [1 - h(i,j)]$
- ✂ Other models are possible as well...



Ultrametric trees

- ✂ In addition to the additive tree criterion, require that all distances to the root be equal
- ✂ Idea is that differences in sequence should be proportional to the evolutionary time since divergence
- ✂ Assumes that DNA evolves at a constant rate across lineages.

The Molecular Clock

- ✿ Are mutation rates constant? If so, sequence distance = evolutionary time distance
- ✿ Why not?
 - Selective pressure. E.g. in coding regions, mutations that disrupt function will occur less often than neutral mutations
 - Punctuated equilibrium. E.g. causative factors in mutation rates (e.g. UV light exposure) vary over evolutionary time
 - Random changes in mutation rate (can model these!)
- ✿ Apparently not true in general, but perhaps...
 - Over relatively short intervals? In certain regions?



Comparing distance to state methods

- ✖ Distance methods are suitable for continuous characters. For sequences, we need to define a distance function.
- ✖ MP doesn't provide branch lengths.
- ✖ Speed: distance < MP << ML



Tree merging

✧ Consensus trees

- Multiple trees constructed from same data to evaluate robustness

✧ Supertrees

- Trees constructed on overlapping sets of taxa and then combined to form larger trees without considering across-subtree interactions

Consensus trees

- ✦ When many different trees have similar scores, look to see where they are identical
- ✦ Various consensus criteria
 - Unanimous, majority, 2/3, score-weighted voting
- ✦ Bootstrapping
 - Resample characters (sequences) from inferred tree
 - Compute new tree from resampled data
 - Report on consensus between original and resampled data.



Supertrees

- ✂ Compute trees from overlapping taxa
- ✂ Only way to create very large trees
- ✂ Each tree needs significant overlap (and consensus) with at least one other tree.
- ✂ Warnow's DCM algorithm



When to use which...

- ✂ Continuous characters, lots of data, computational constraints: Neighbor Joining
- ✂ Discrete characters, moderate data, no homoplasy: maximum parsimony
- ✂ Discrete characters, limited sequence lengths, some homoplasy: maximum likelihood
- ✂ Many taxa: supertrees
- ✂ Complete genomes: whole genome methods