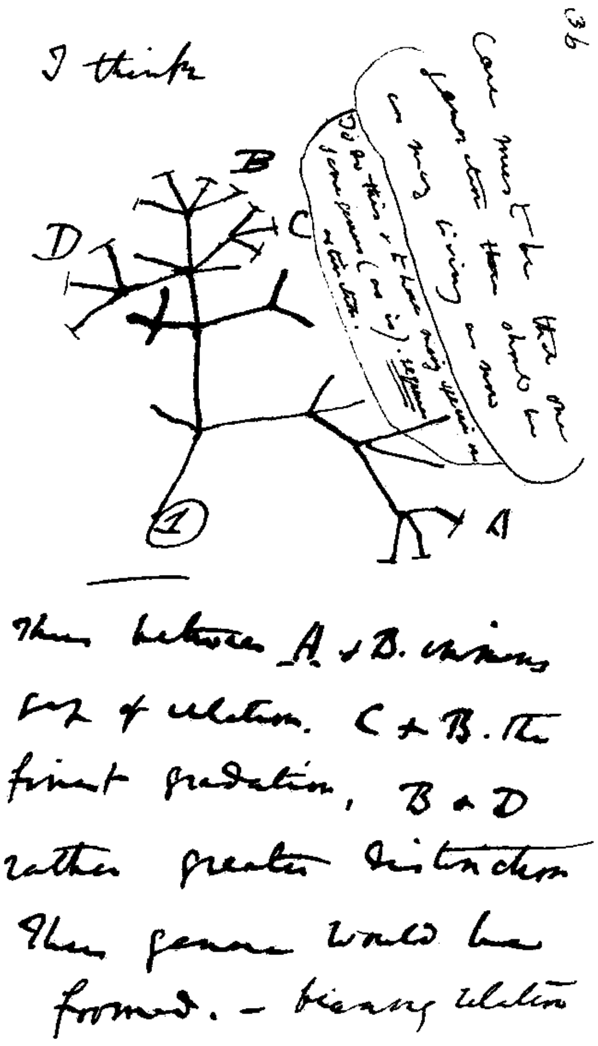
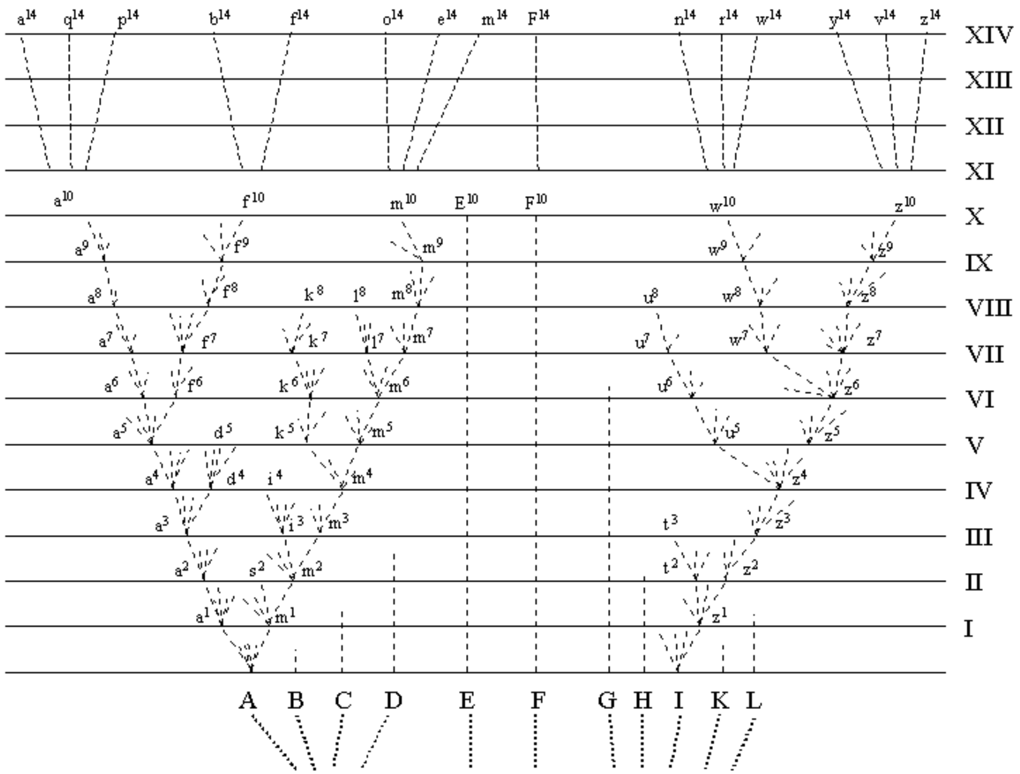
构建系统发生树（进化树）

张朝

2024/11/29

什么是进化树（phylogenetic tree）？


36
I think
Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is) requires extinction.
B
C
D
A
1
Thus between A & B immense gap of relation. C & B the finest gradation, B & D rather greater distinction. Thus genera would be formed. — bearing relation


a14 q14 p14 b14 f14 o14 e14 m14 F14 n14 r14 w14 y14 v14 z14
a10 f10 m10 E10 F10 w10 z10
XIV
XIII
XII
XI
X
IX
VIII
VII
VI
V
IV
III
II
I
A B C D E F G H I K L

Charles Darwin
1837 and 1859

《物种起源》中唯一的插图

分子序列数据的优势

1. 分子序列是可以完全遗传的

2. 对单个分子状态不会有模糊的描述

ATGCs are ATGCs. There is no such thing as half purine and half pyrimidine.

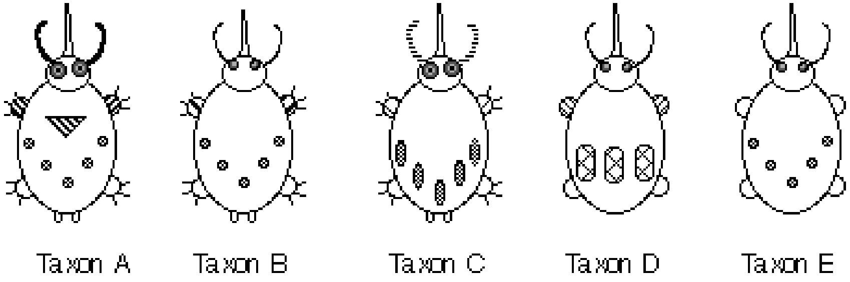
3. 分子数据容易进行量化处理

4. 相对于形态学数据更容易对对同源性进行定义

5.分子数据对于遗传距离更稳健

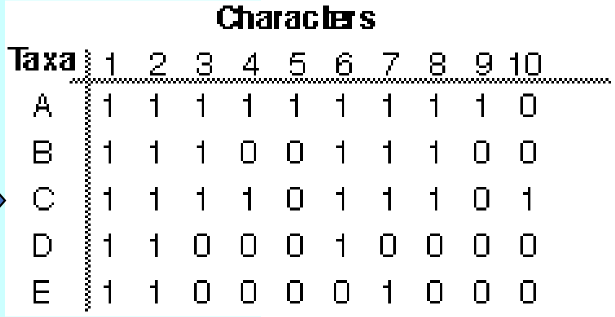
6. 分子数据数量庞大

早期进化树的构建 – 基于表型特征



十个形态学特征(有或无):

1. 尖鼻子
2. 四条腿
3. 脚上有脚指头
4. 大眼睛
5. 背上有三角形的盾状花纹
6. 前腿上条纹
7. 背上有五个点
8. 身体后部有触须
9. 头上触角粗壮
10. 触角上布满了绒毛看上去有条纹感

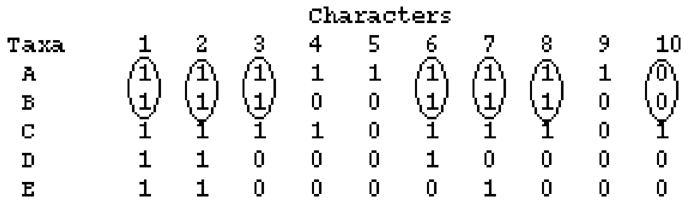


	Characters									
Taxa	1	2	3	4	5	6	7	8	9	10
A	1	1	1	1	1	1	1	1	1	0
B	1	1	1	0	0	1	1	1	0	0
C	1	1	1	1	0	1	1	1	0	1
D	1	1	0	0	0	1	0	0	0	0
E	1	1	0	0	0	0	1	0	0	0

1代表有

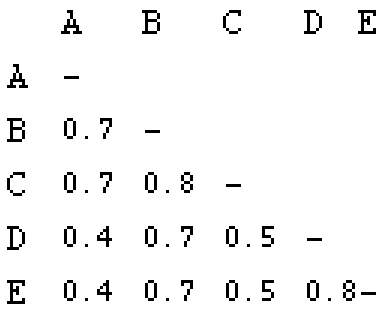
0代表无

基于表型特征的进化树构建



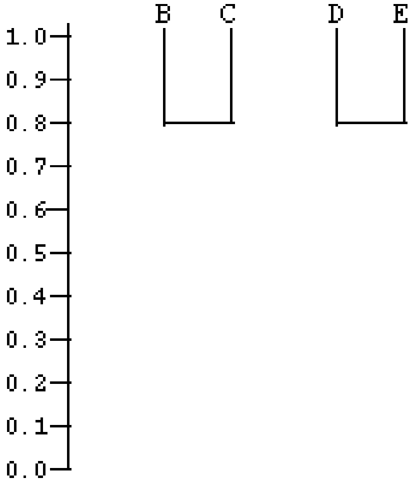
Taxa	Characters 1	2	3	4	5	6	7	8	9	10
A	1	1	1	1	1	1	1	1	1	0
B	1	1	1	0	0	1	1	1	0	0
C	1	1	1	1	0	1	1	1	0	1
D	1	1	0	0	0	1	0	0	0	0
E	1	1	0	0	0	0	1	0	0	0

A和B之间的形态学距离=
7/10 or 0.70

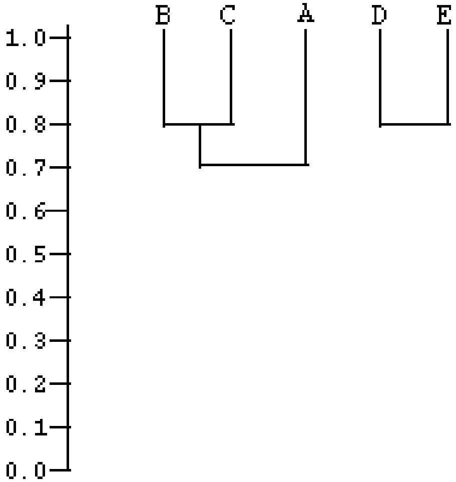


	A	B	C	D	E
A	-				
B	0.7	-			
C	0.7	0.8	-		
D	0.4	0.7	0.5	-	
E	0.4	0.7	0.5	0.8	-

距离矩阵
（Distance matrix）



A, DE与BC之间的关系 (A closer to BC or DE closer to BC)?
A-B = 0.7
A-C = 0.7
A与B、C的平均距离= (0.7 + 0.7) ÷ 2 = 0.7
D-B = 0.7
D-C = 0.5
E-B = 0.7
E-C = 0.5
D，E与B、C的平均距离= (0.7 + 0.5 + 0.7 + 0.5)÷ 4 = 0.6



基于表型特征的进化树构建

ABC 和 DE之间的形态学距离:

A-D = 0.4

B-D = 0.7

C-D = 0.5

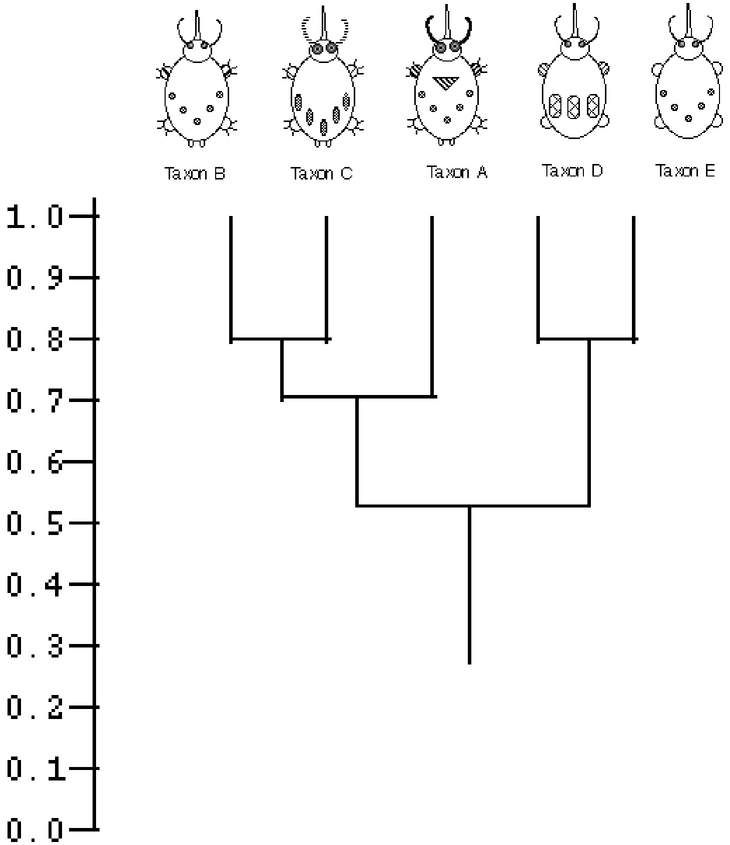
A-E = 0.4

B-E = 0.7

C-E = 0.5

总距离 = 3.20

平均距离 = 3.2 ÷ 6 = 0.53



在分子序列出现之前，构建进化树的方法被称做表型分类法（phenetics）或数量分类学（numerical taxonomy）

分子序列数据的优势

1. 分子序列是可以完全遗传的

2. 对单个分子状态不会有模糊的描述

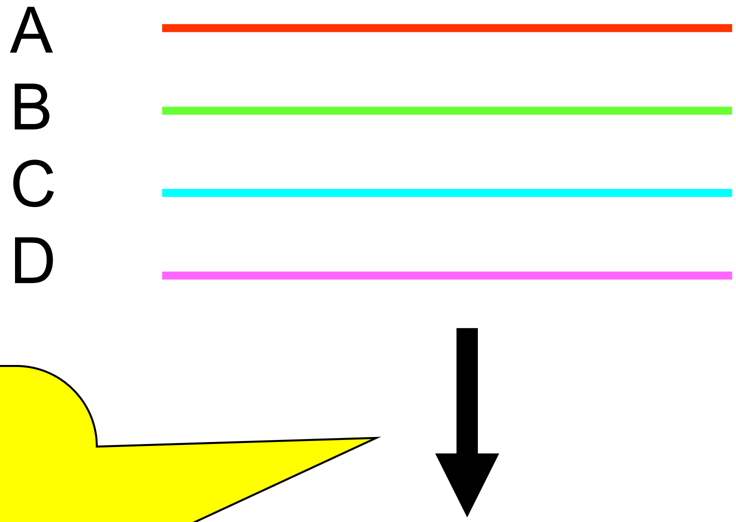
   ATGCs are ATGCs. There is no such thing as half purine and half pyrimidine.

3. 分子数据容易进行量化处理

4. 相对于形态学数据更容易对对同源性进行定义

5.分子数据对于遗传距离更稳健

6. 分子数据数量庞大



Compute the pairwise alignments for all against all (6 pairwise alignments) the similarities are stored in a table

	A	B	C	D
A				
B	11			
C	3	1		
D	2	2	10	

Alignments can be easy or difficult (序列比对可以容易也可以很难)

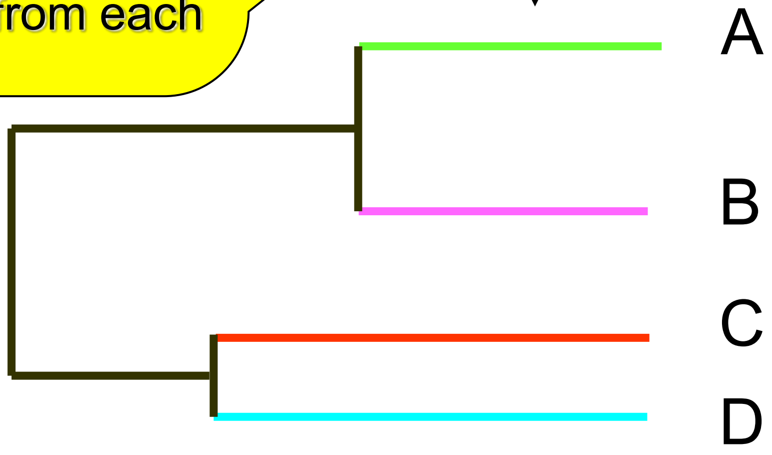
```
GCGGCCCA TCAGGTAGTT GGTGG
GCGGCCCA TCAGGTAGTT GGTGG
GCGTTCCA TCAGCTGGTT GGTGG
GCGTCCCA TCAGCTAGTT GGTGG
GCGGCGCA TTAGCTAGTT GGTGA
***...** *.**.*.*** ****.
```

```
TTGACATG CCGGGG---A AACCG
T-GACATG CCGGTG--GT AAGCC
TTGGCATG -CTAGG---A ACGCG
TTGACATG -CTAGGGAAC ACGCG
TTGACATC -CTCTG---A ACGCG
* *.***.  *...*   . *..*.
```

Cluster the sequences to create a tree (guide tree):

- Represents the order in which pairs of sequences are to be aligned
- Similar sequences are neighbors in the tree
- Distant sequences are distant from each other in the tree

	A	B	C	D
A				
B	11			
C	3	1		
D	2	2	10	



碱基替代模型

- **Substitution models for nucleotide sequence**

Jukes-Cantor **(JC, nst=1):** Equal base frequencies, all substitutions equally likely (Jukes and Cantor 1969)

Felsenstein 1981 **(F81, nst=1):** Variable base frequencies, all substitutions equally likely (Felsenstein 1981)

Kimura 2-parameter **(K80, nst=2):** Equal base frequencies, variable transition and transversion frequencies (Kimura 1980)

Hasegawa-Kishino-Yano **(HKY, nst=2):** Variable base frequencies, variable transition and transversion frequencies (Hasegawa et. al. 1985)

Tamura-Nei **(TrN):** Variable base frequencies, equal transversion frequencies, variable transition frequencies (Tamura Nei 1993)

Kimura 3-parameter **(K3P):** Variable base frequencies, equal transition frequencies, variable transversion frequencies (Kimura 1981)

Transition Model **(TIM):** Variable base frequencies, variable transitions, transversions equal

Transversion Model **(TVM):** Variable base frequencies, variable transversions, transitions equal

Symmetrical Model **(SYM):** Equal base frequencies, symmetrical substitution matrix (A to T = T to A)

General Time Reversible **(GTR, nst=6):** Variable base frequencies, symmetrical substitution matrix (Lanave et al. 1984, Tavare 1986, Rodriguez et. al. 1990)

Kimura 2-Parameter Distance (K2P)

- Two rates of substitutions, for transitions and transversions

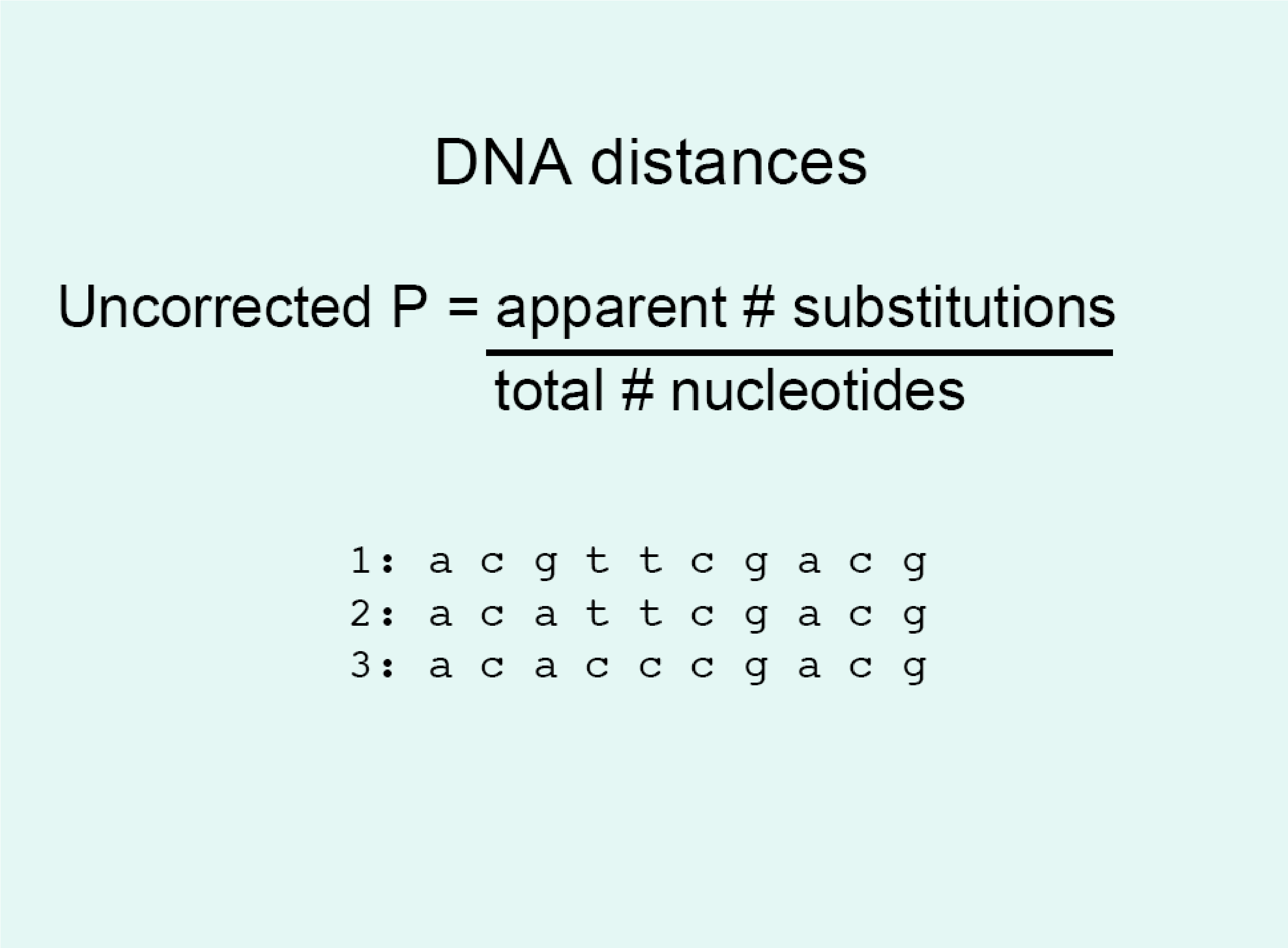
	a	c	g	t
a	-			
c	**c=>a**	-		
g	**g=>a**	**g=>c**	-	
t	**t=>a**	**t=>c**	**t=>g**	-

Generalized Time Reversable Distance

- Six different substitution rates
- Backward rates are same as forward rates

	a	c	g	t
a	-			
c	**1**	-		
g	**2**	**3**	-	
t	**4**	**5**	**6**	-

Why do we need models?
（我们为什么需要模型）

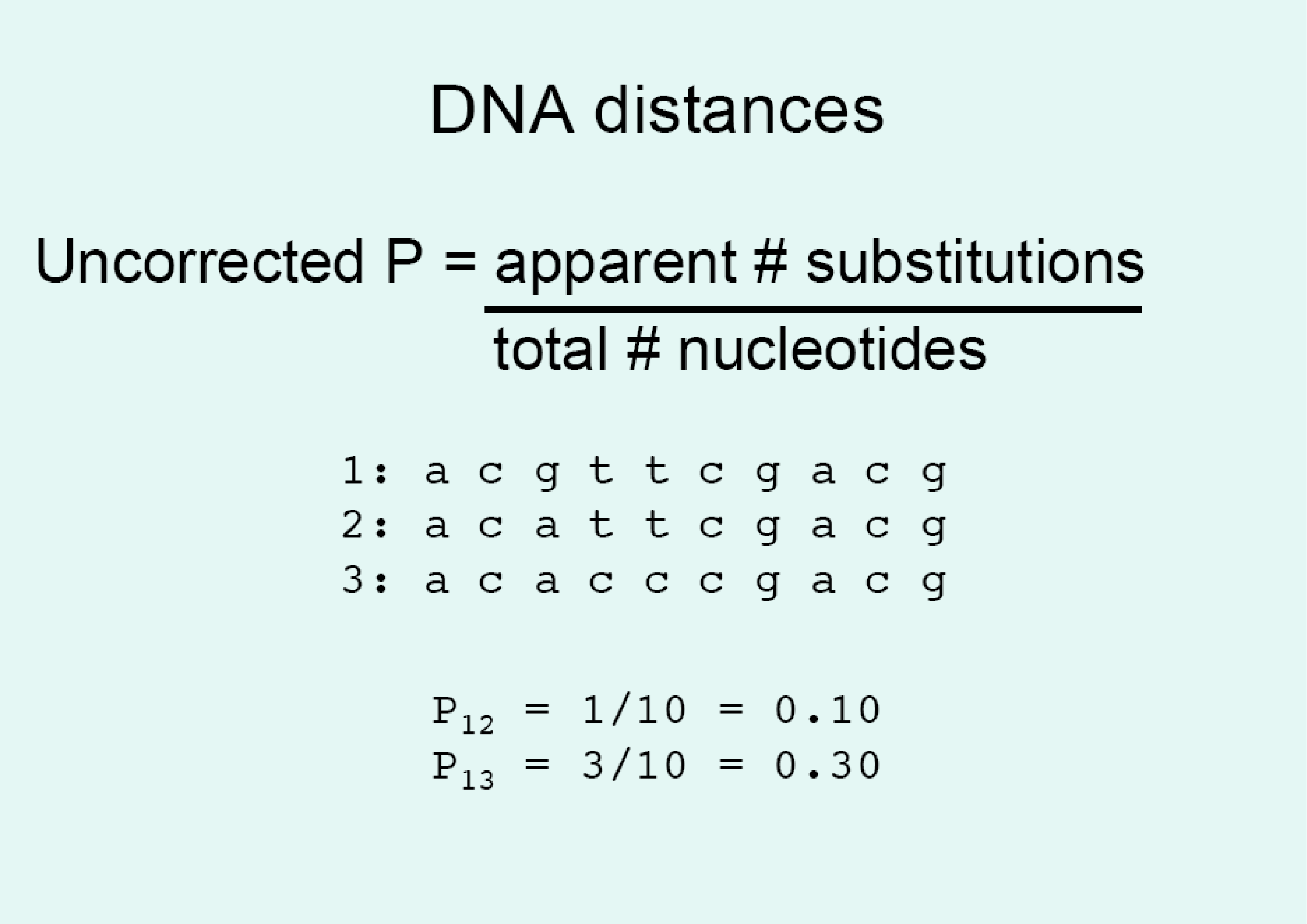


DNA distances

$$\text{Uncorrected P} = \frac{\text{apparent \# substitutions}}{\text{total \# nucleotides}}$$

```
1: a c g t t c g a c g
2: a c a t t c g a c g
3: a c a c c c g a c g
```

Why do we need models?
（我们为什么需要模型）



DNA distances

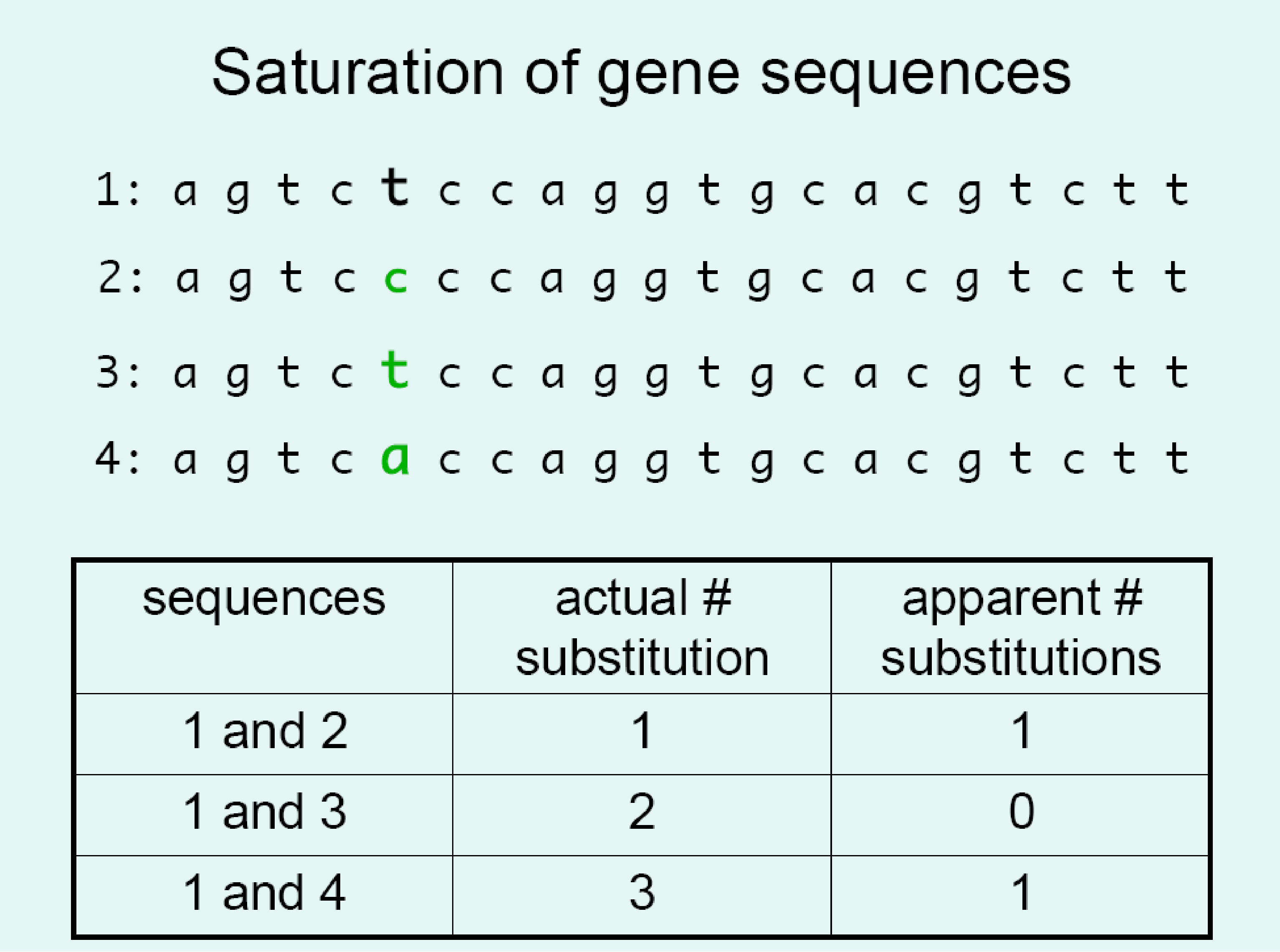
$$\text{Uncorrected P} = \frac{\text{apparent \# substitutions}}{\text{total \# nucleotides}}$$

```
1: a c g t t c g a c g
2: a c a t t c g a c g
3: a c a c c c g a c g
```

$P_{12} = 1/10 = 0.10$

$P_{13} = 3/10 = 0.30$

Why do we need models?
（我们为什么需要模型）



Saturation of gene sequences

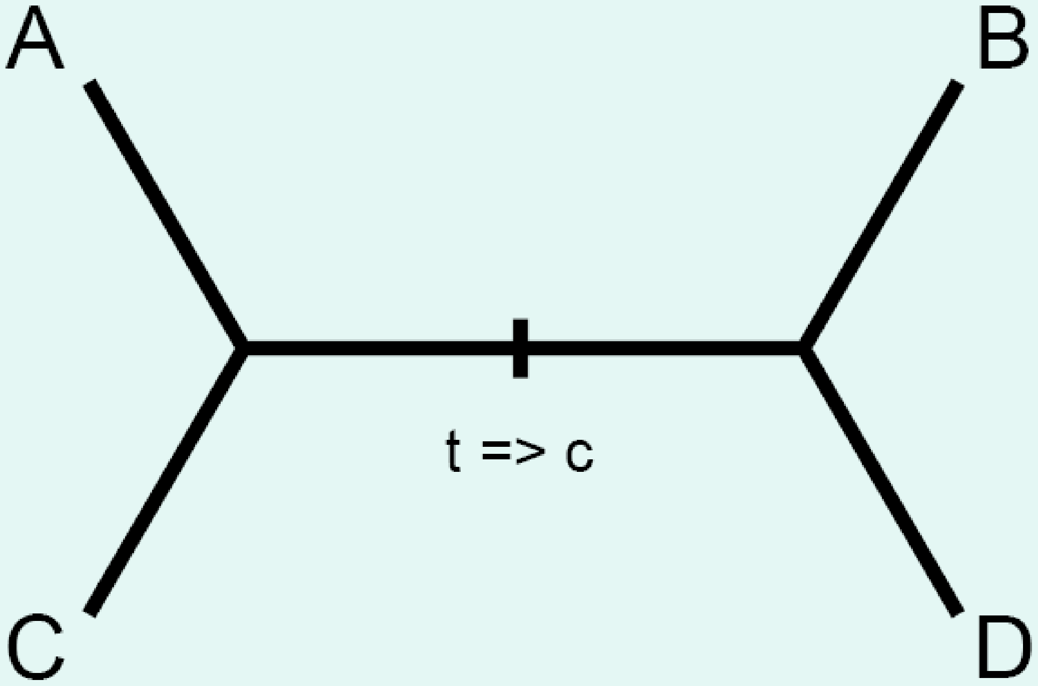
1: a g t c t c c a g g t g c a c g t c t t

2: a g t c c c c a g g t g c a c g t c t t

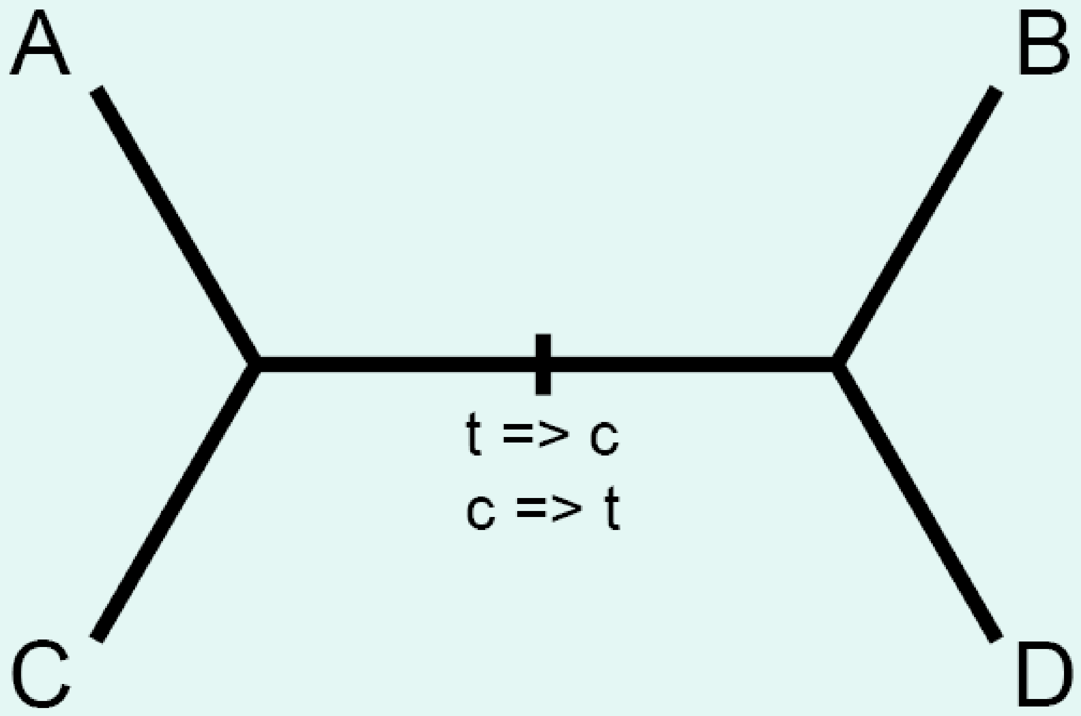
3: a g t c t c c a g g t g c a c g t c t t

4: a g t c a c c a g g t g c a c g t c t t

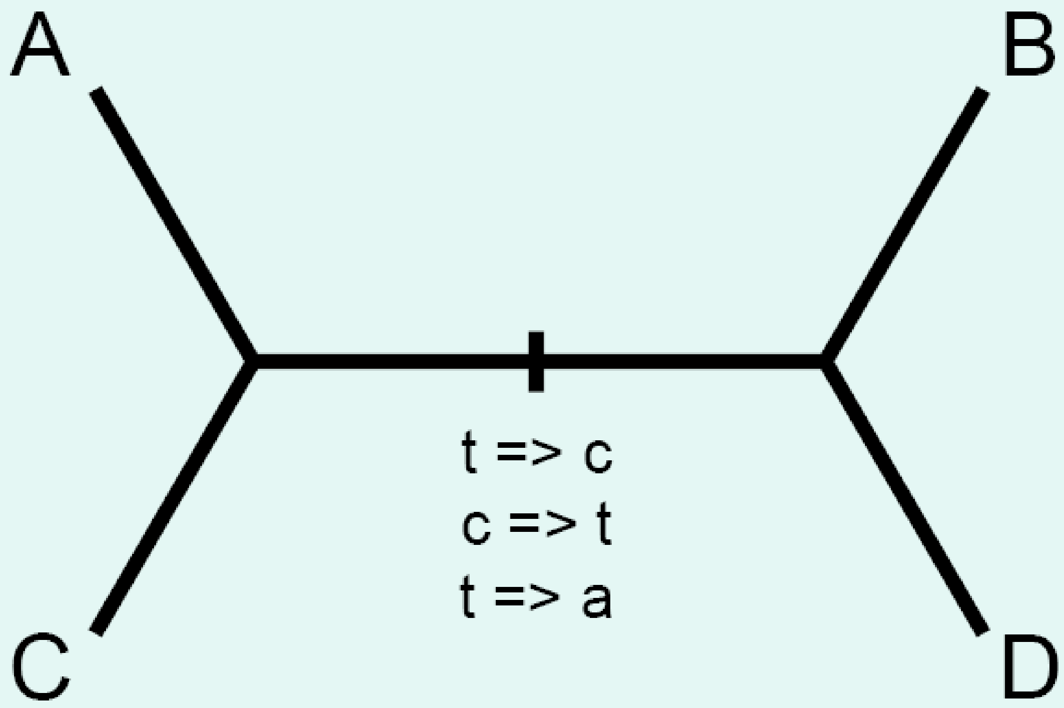
sequences	actual # substitution	apparent # substitutions
1 and 2	1	1
1 and 3	2	0
1 and 4	3	1



Distance	actual # substitution	apparent # substitutions
A or C to B or D	1	1



Distance	actual # substitution	apparent # substitutions
A or C to B or D	1	1
A or C to B or D	2	0



Distance	actual # substitution	apparent # substitutions
A or C to B or D	1	1
A or C to B or D	2	0
A or C to B or D	3	1

Substitution rates

- All rates equal (essentially the Jukes-Cantor model)
- Transitions have different rate than transversions
- Two different kinds of transversions (this gives us three rates)
- Maximally, all types of substitutions are allowed to have their own rate
- But note that substitution rates are not independent of base frequencies in the data!

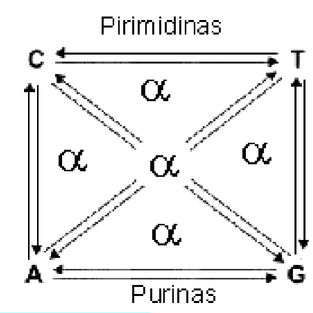
Base Frequencies: equal or variable?

（碱基频率：到底是均等还是可变？）

- Jukes-Cantor **(JC, nst=1):** Equal base frequencies, all substitutions equally likely

To different data set, JC model uses the same fixed substitution rate (alpha is a constant, approximately equal to zero).

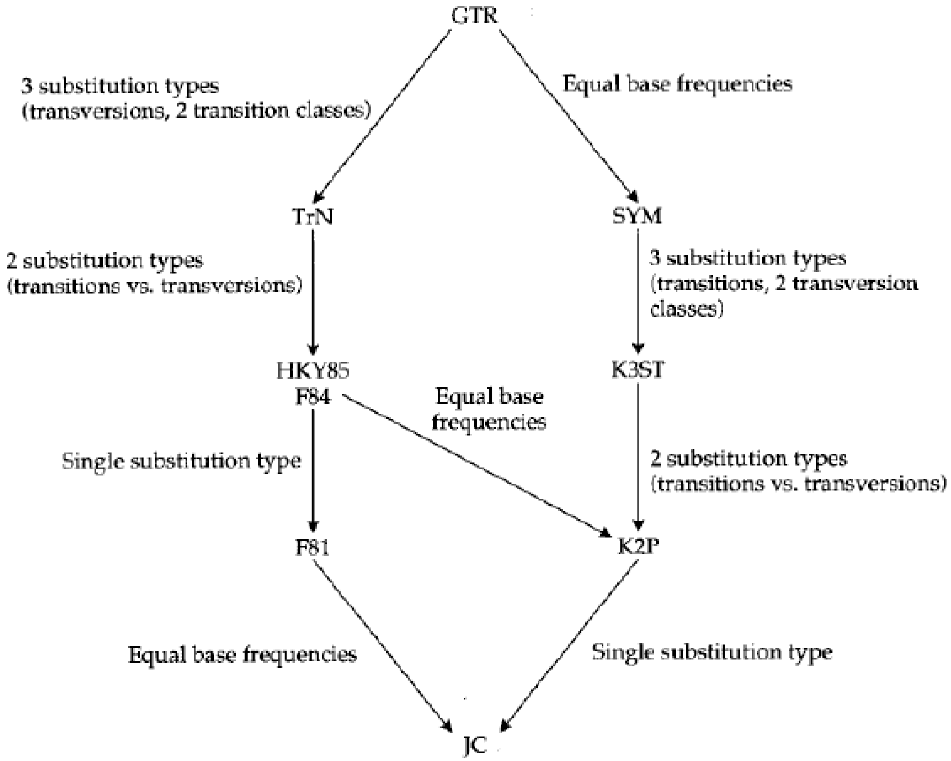
- Felsenstein 1981 **(F81, nst=1):** Variable base frequencies, all substitutions equally likely



To different data set, F81 model uses the different estimated substitution rate, although transition and transverstion rate is the same in F81 (alpha estimated from data set).

Variable base frequencies are more realistic （可变更符合实际）.

Substitution Rate Models


GTR
3 substitution types (transversions, 2 transition classes)
Equal base frequencies
TrN
SYM
2 substitution types (transitions vs. transversions)
3 substitution types (transitions, 2 transversion classes)
HKY85
F84
Equal base frequencies
K3ST
Single substitution type
2 substitution types (transitions vs. transversions)
F81
K2P
Equal base frequencies
Single substitution type
JC

- GTR = general time reversible
- JC = Jukes Cantor
- Relax each assumption to move to the model below

Take home message

- 序列比对是构建进化树的基础

- 基于序列的相似性和碱基替代模型计算距离

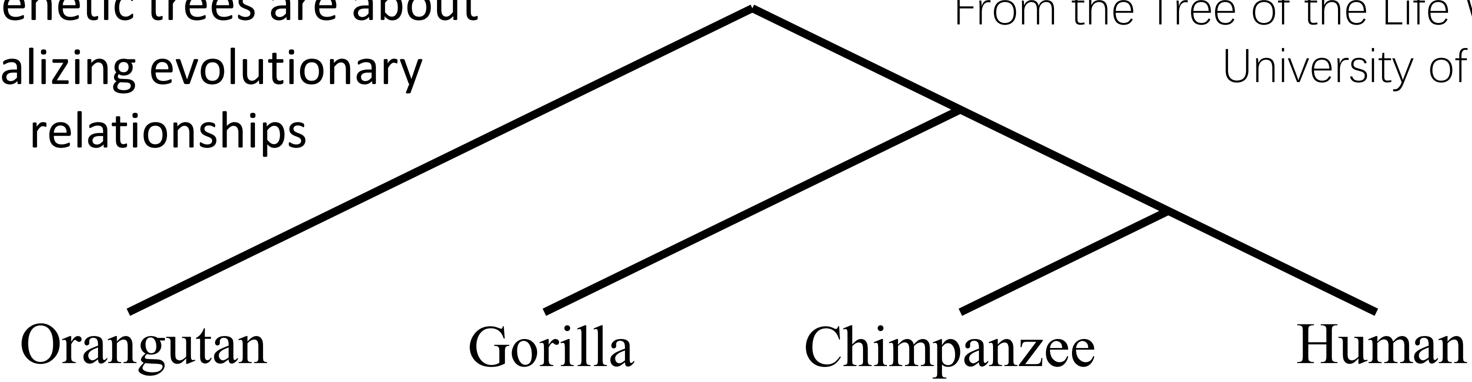
- 基于距离构建进化树

- 进化树构建

Phylogeny

Phylogenetic trees are about visualizing evolutionary relationships

From the Tree of the Life Website,
University of Arizona



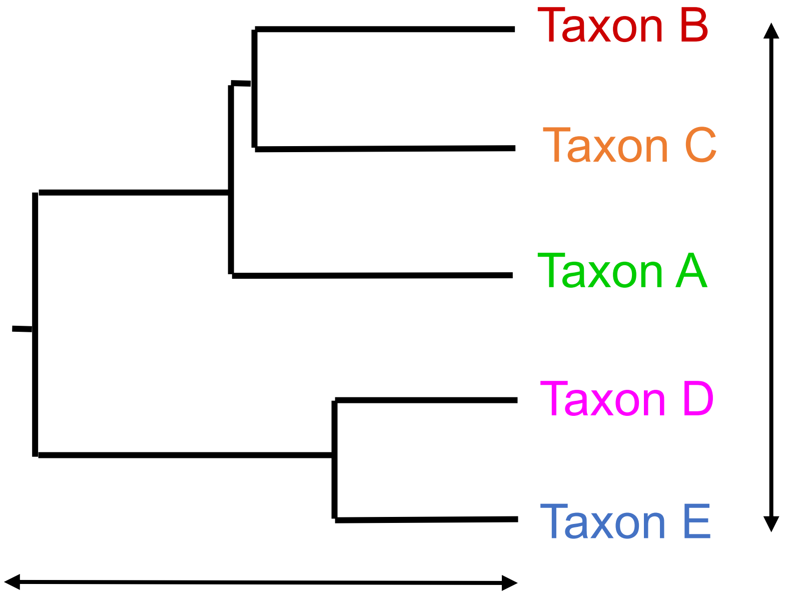
Orangutan

Gorilla

Chimpanzee

Human

Phylogenetic trees diagram the *evolutionary relationships* between the taxa



No meaning to the spacing between the taxa, or to the order in which they appear from top to bottom.

This dimension either can have no scale (for ‘cladograms’), can be proportional to genetic distance or amount of change (for ‘phylograms’ or ‘additive trees’), or can be proportional to time (for ‘ultrametric trees’ or true evolutionary trees).

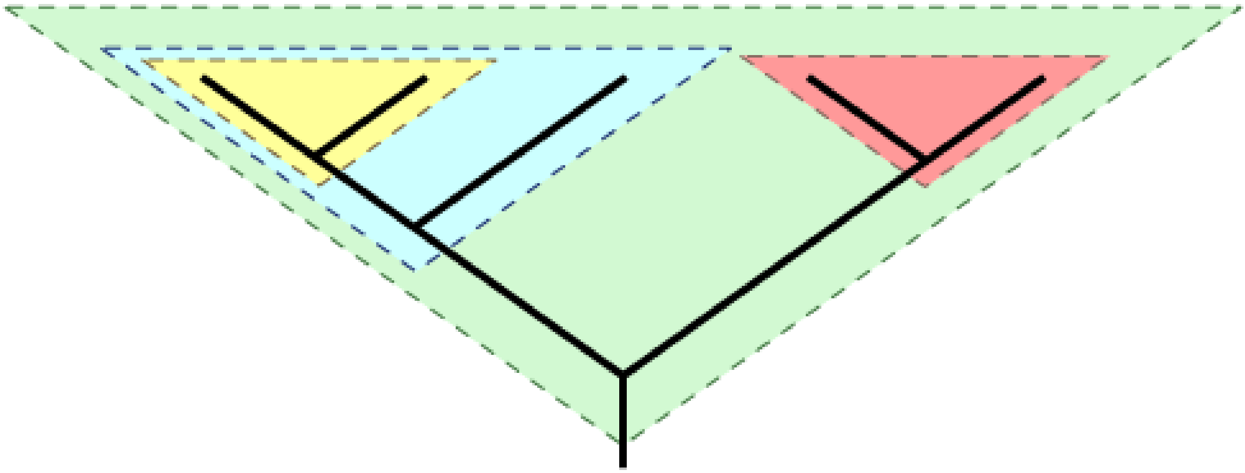
((A,(B,C)),(D,E)) = The above phylogeny as nested parentheses

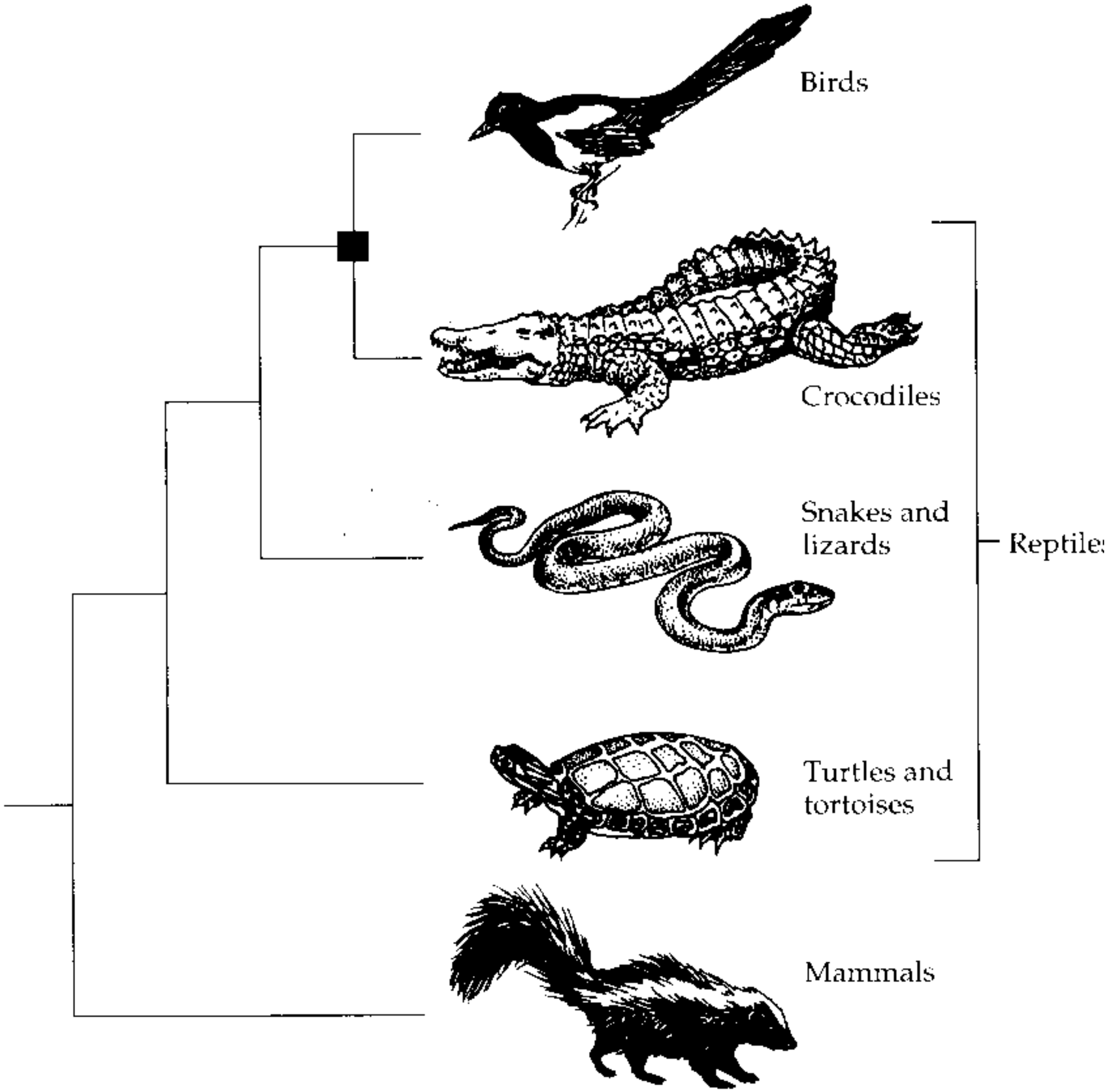
These say that B and C are more closely related to each other than either is to A, and that A, B, and C form a clade that is a sister group to the clade composed of D and E. If the tree has a time scale, then D and E are the most closely related.

Clades

- Evolutionary trees depict clades.
- A clade is a group of organisms that includes an ancestor and *all* descendents of that ancestor. You can think of a clade as a branch on the tree of life.
- clades指的是一个由一个共同祖先及其所有后代组成的生物群组

Each of these highlighted areas is a clade:

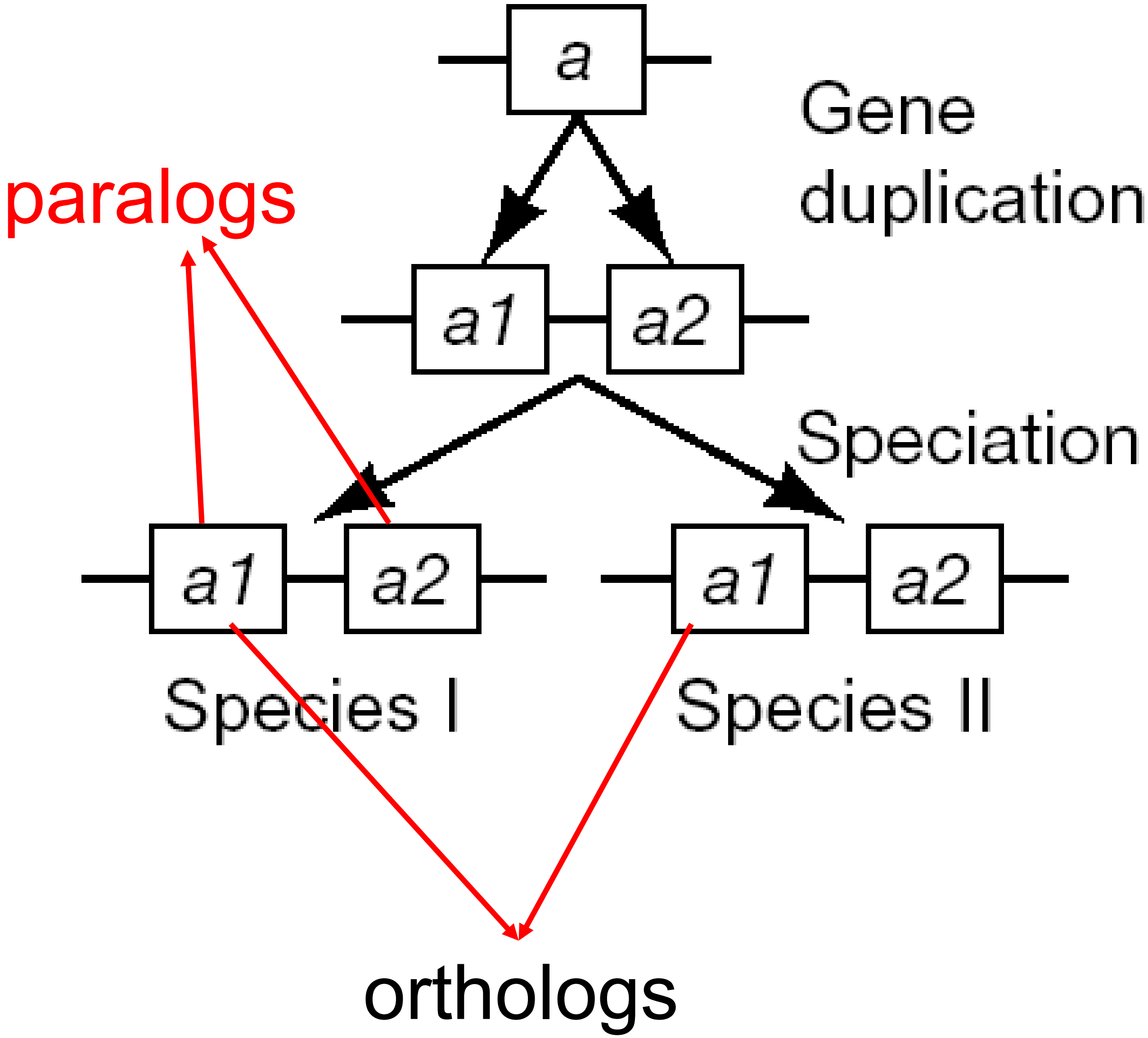




Molecular Evolution - Li

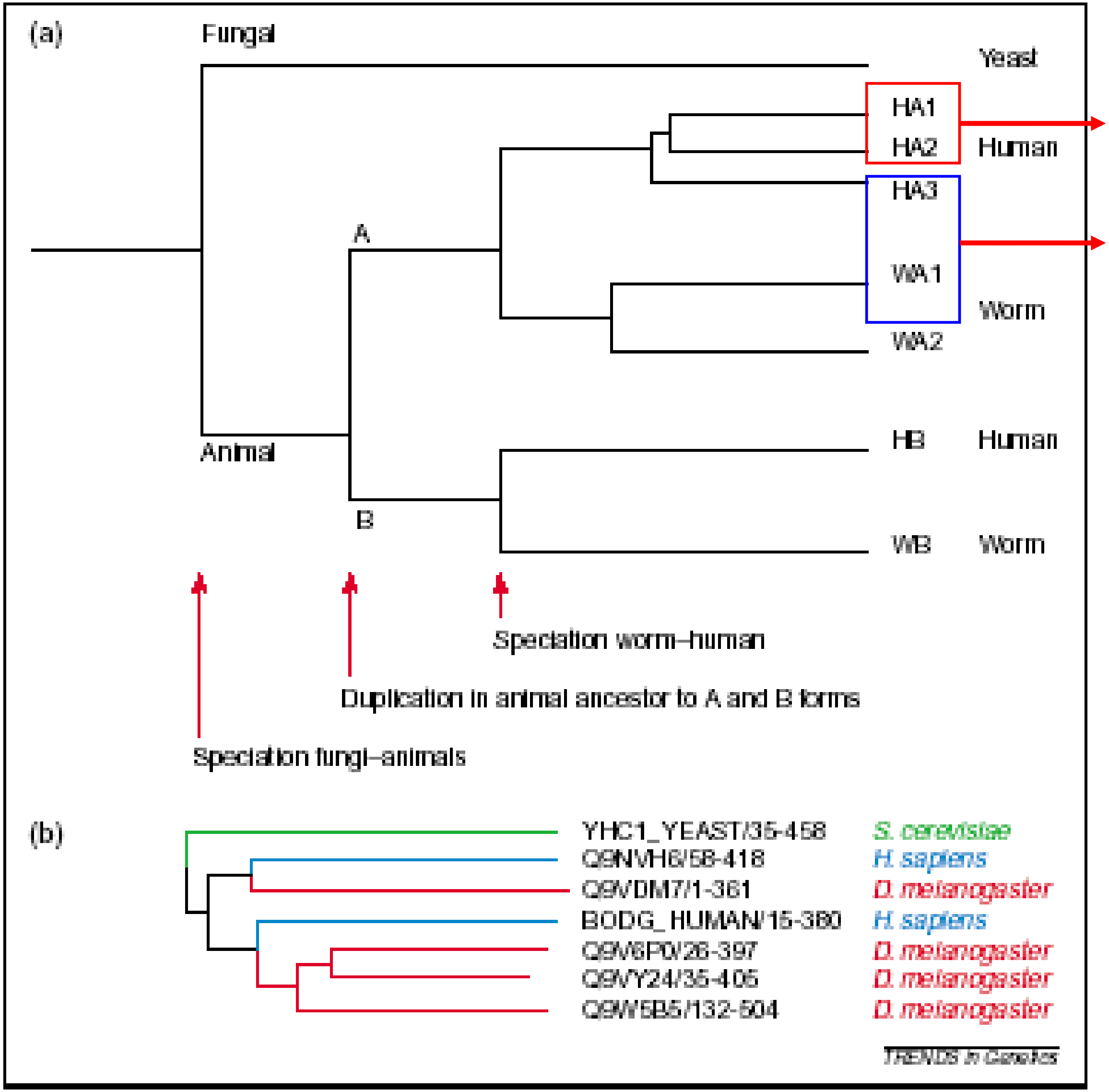
Terminology

- Homologue： Orthologue Paralogue
- 同源基因（Homologue）：同源基因是指在不同物种中，由于共**同的祖先而具有相似序列和功能的基因**。这些基因可能在进化过程中经历了复制、突变和自然选择，但仍然保留了一些原始的特征。
- 直系同源基因（Orthologue）：直系同源基因是指在**不同物种**中，由于**物种分化**（speciation）而产生的基因。这些基因在物种分化之前存在于共同的祖先中，随着物种的分化，这些基因在不同物种中独立进化。
- 旁系同源基因（Paralogue）：旁系同源基因是指在**同一个物种**中，由于**基因复制**事件而产生的基因。这些基因原本是同一个基因，但由于复制，它们在基因组中有了两个或更多的副本。
- 其他术语：Monophyletic， Kingdom



paralogs

orthologs



paralogs

orthologs

Erik L.L. Sonnhammer Orthology,paralogy and proposed classification for paralog subtypes

TRENDS in Genetics Vol.18 No.12 December 2002

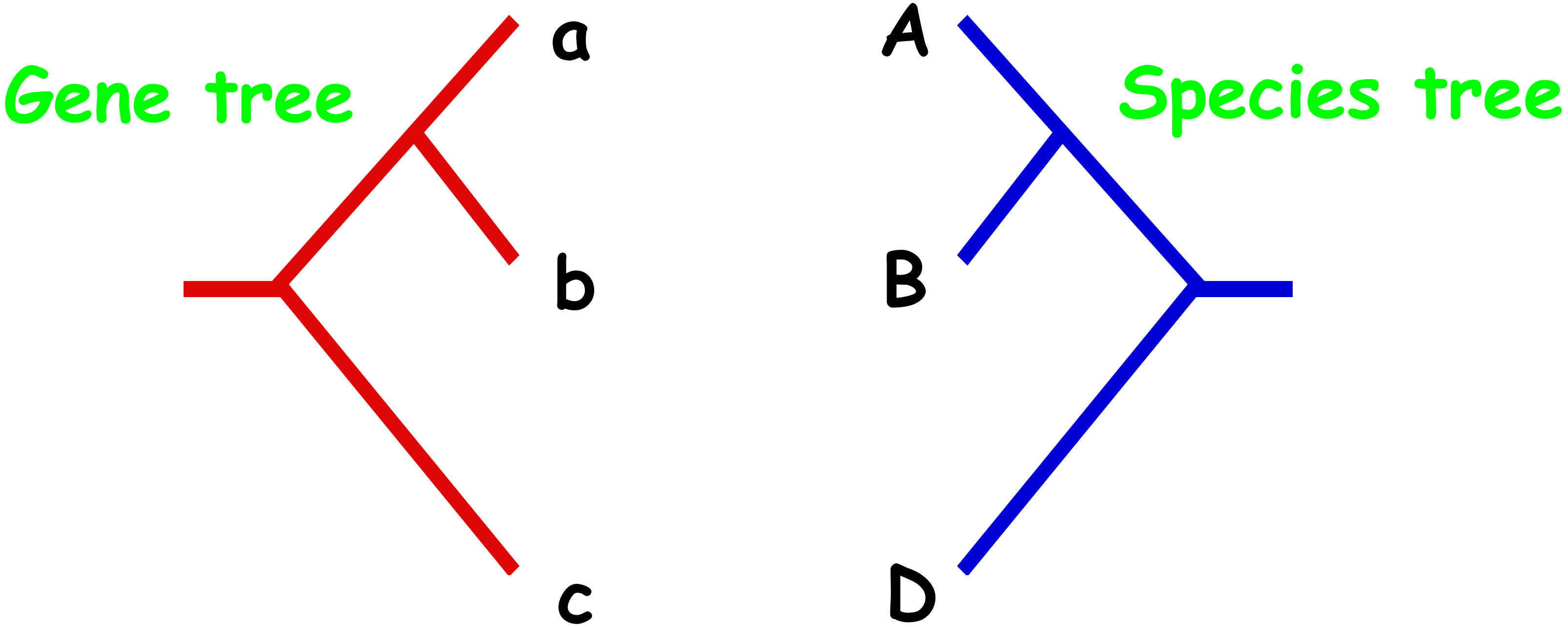
http://tig.trends.com 0168-9525/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved.

Trees

- Diagram consisting of branches and nodes
- **Species tree** (how are my species related?)
  - contains only one representative from each species.
  - all nodes indicate speciation events
- **Gene tree** (how are my genes related?)
  - normally contains a number of genes from a single species
  - nodes relate either to speciation or gene duplication events

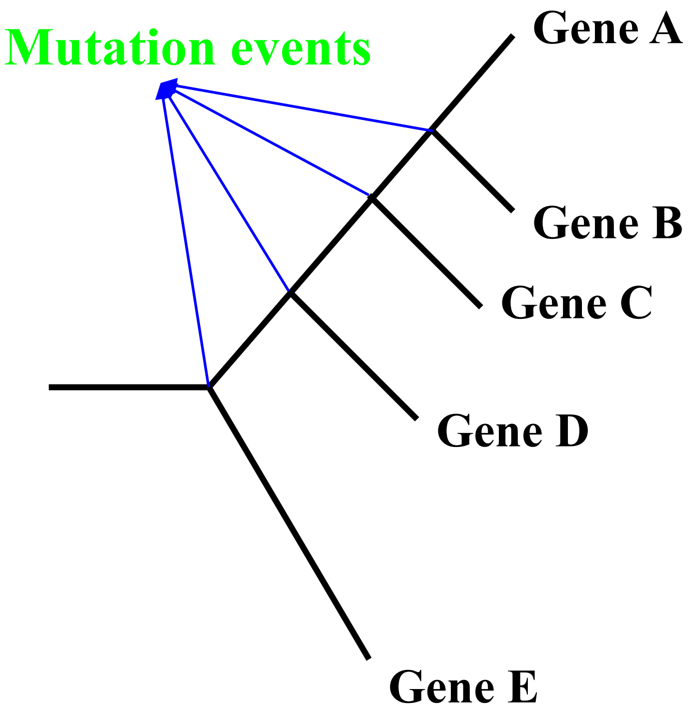
基于单个同源基因差异构建的演化树应称为基因树（gene tree），由多个基因构建的能代表多个物种演化关系的树，则称为物种树（species tree）

Gene tree, species tree

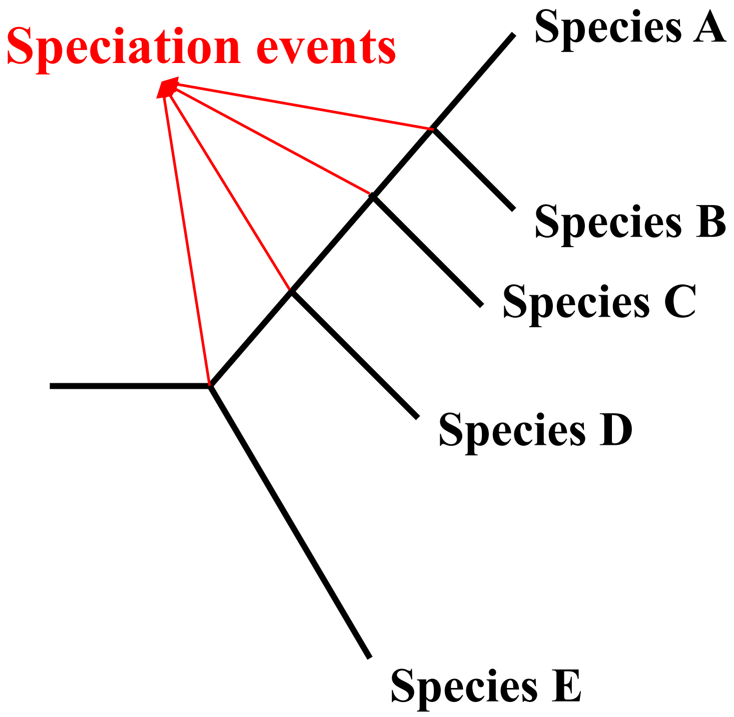


We often assume that gene trees give us species trees

Gene tree - Species tree



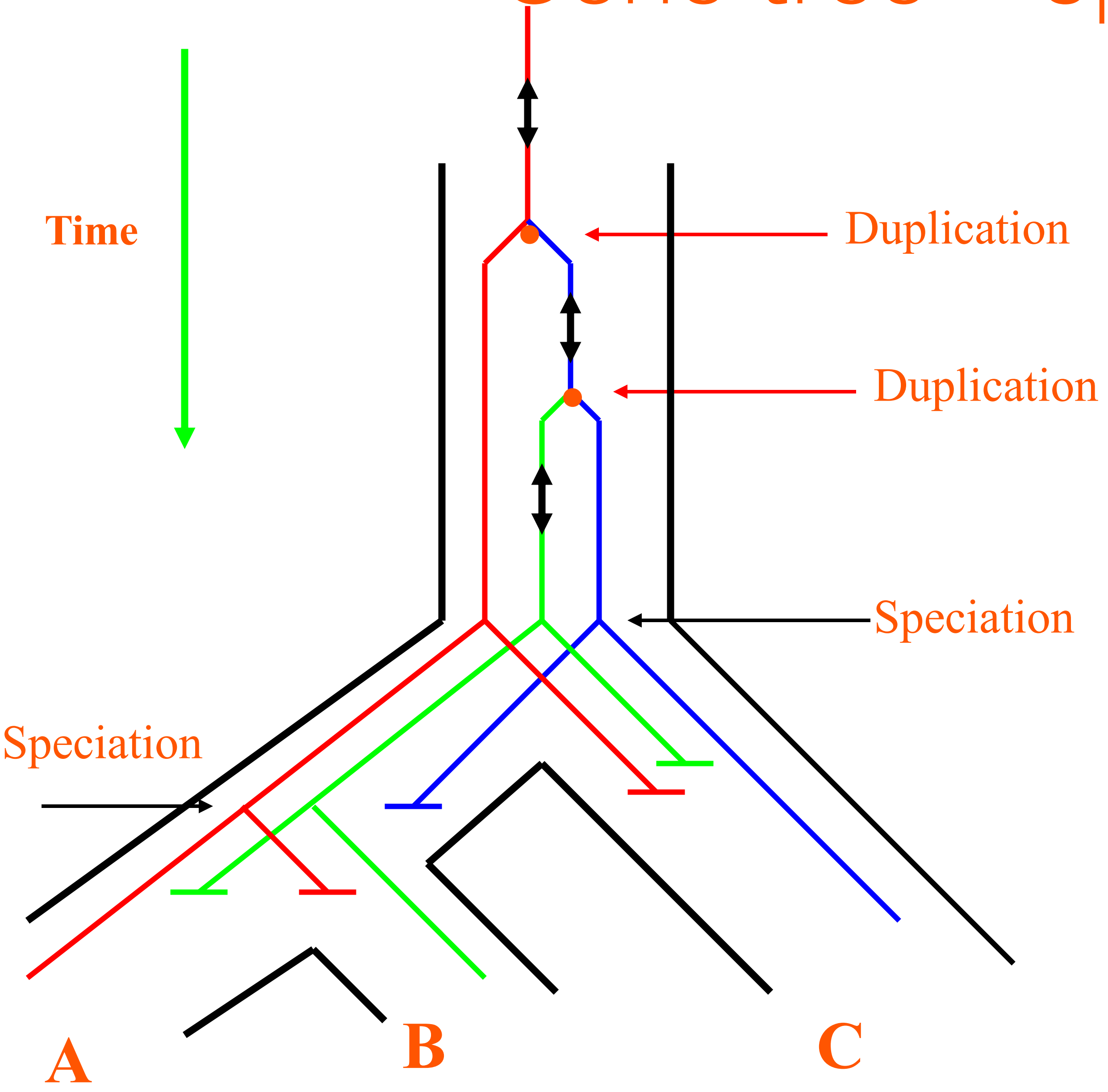
Gene tree

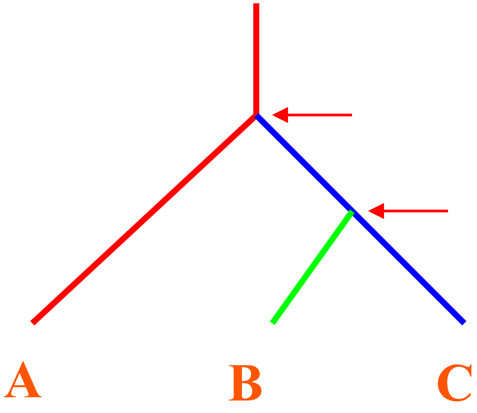


Species tree

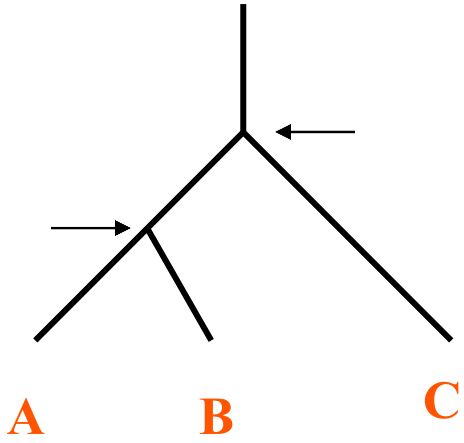
The two events - mutation and speciation- are not expected to occur at the same time. So gene trees cannot represent species tree.

Gene tree - Species tree





Gene tree



Species tree

进化树构建的基本方法

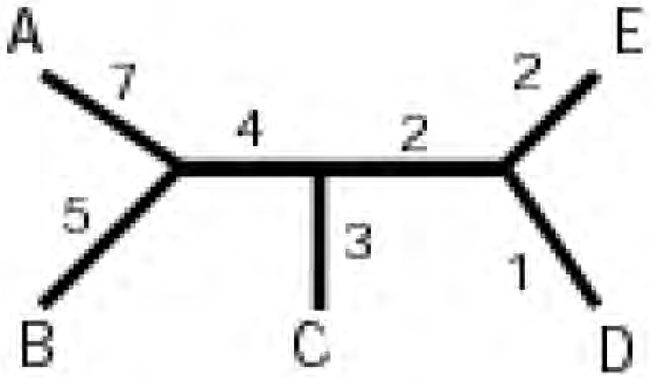
- 距离法 (Distance)

- 穷举法
  - 最大简约法 (maximum parsimony, MP)
  - 最大似然法 (maximum likelihood, ML)

- 贝叶斯法 (Bayes)

Distance Methods

	A	B	C	D	E
A	0				
B	12	0			
C	14	12	0		
D	14	12	6	0	
E	15	13	7	3	0



A tree exactly fitting the matrix does not always exist.

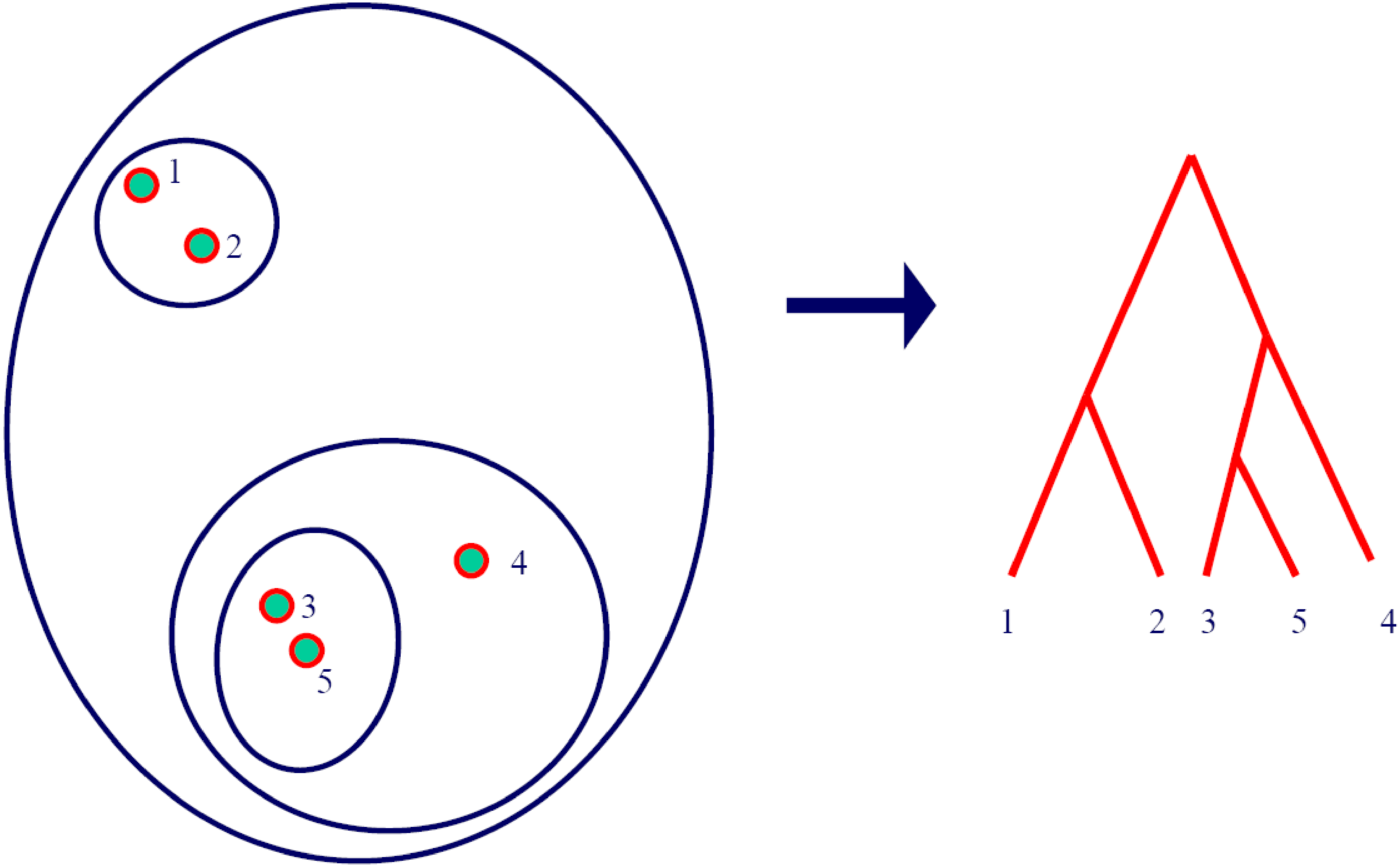
Distance methods

- Normally fast and simple
- e.g. UPGMA, Neighbour Joining, Minimum Evolution

UPGMA

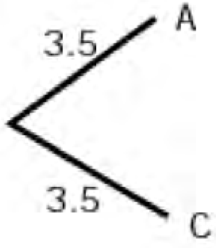
- UPGMA (Unweighted group method with arithmetic mean)
  - Sequential clustering algorithm
  - Start with things most similar
    - Build a composite OTU
  - Distances to this OTU are computed as arithmetic means
  - From new group of OTUs, pick pair with highest similarity etc.
- Average-linkage clustering

UPGMA: Visually



UPGMA: example

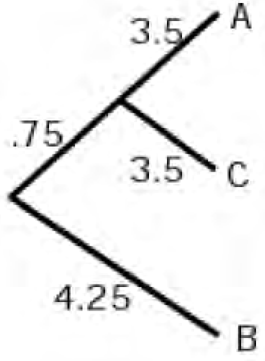
	A	B	C	D
A	0			
B	8	0		
C	7	9	0	
D	12	14	11	0



$$M_{B(AC)} = (M_{BA} + M_{BC})/2 = (8+9)/2=8.5$$
$$M_{D(AC)} = (M_{DA} + M_{DC})/2= (12+11)/2=11.5$$

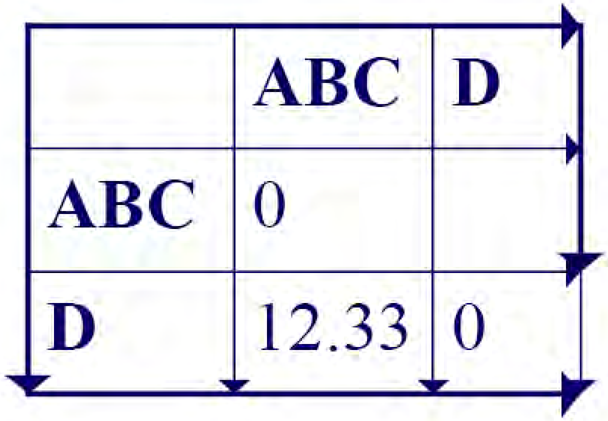
UPGMA: example

	AC	B	D
AC	0		
B	8.5	0	
D	11.5	14	0

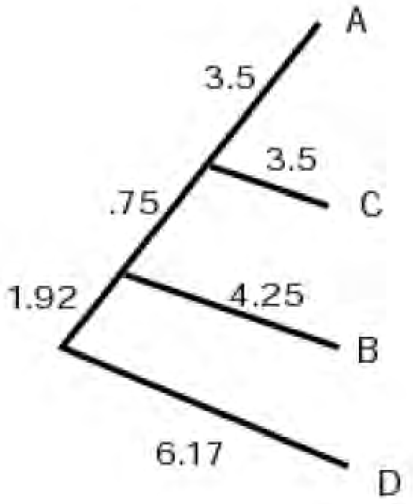


$$M_{(ABC)D} = (M_{AD} + M_{BD} + M_{CD})/3 = (12+14+11)/3$$

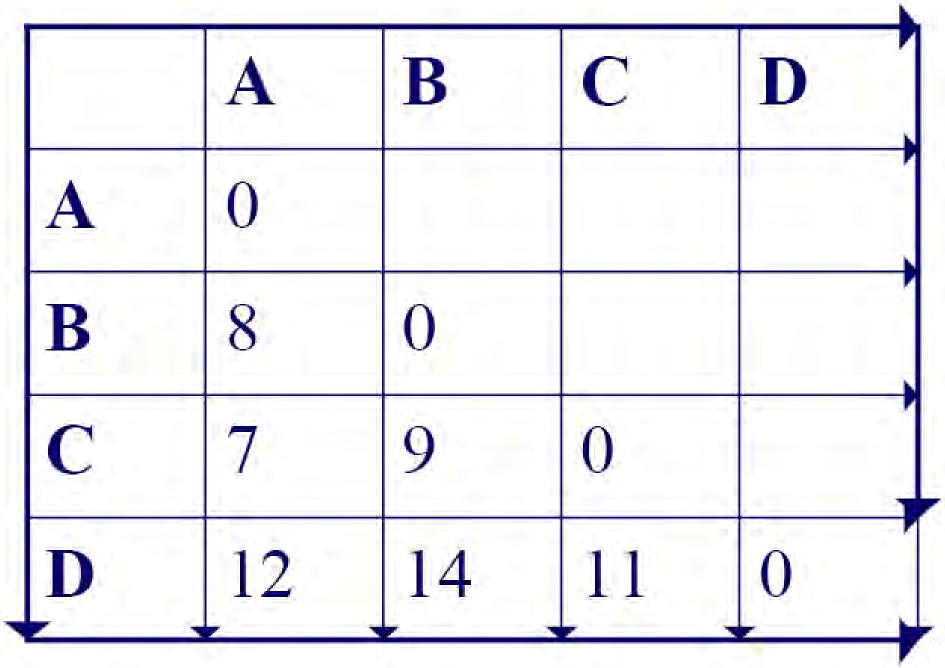
UPGMA: example



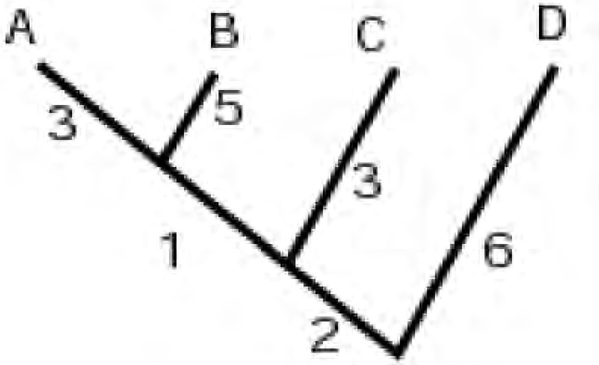
	ABC	D
ABC	0	
D	12.33	0



UPGMA weaknesses



	A	B	C	D
A	0			
B	8	0		
C	7	9	0	
D	12	14	11	0



In fact, exact fitting tree exists !

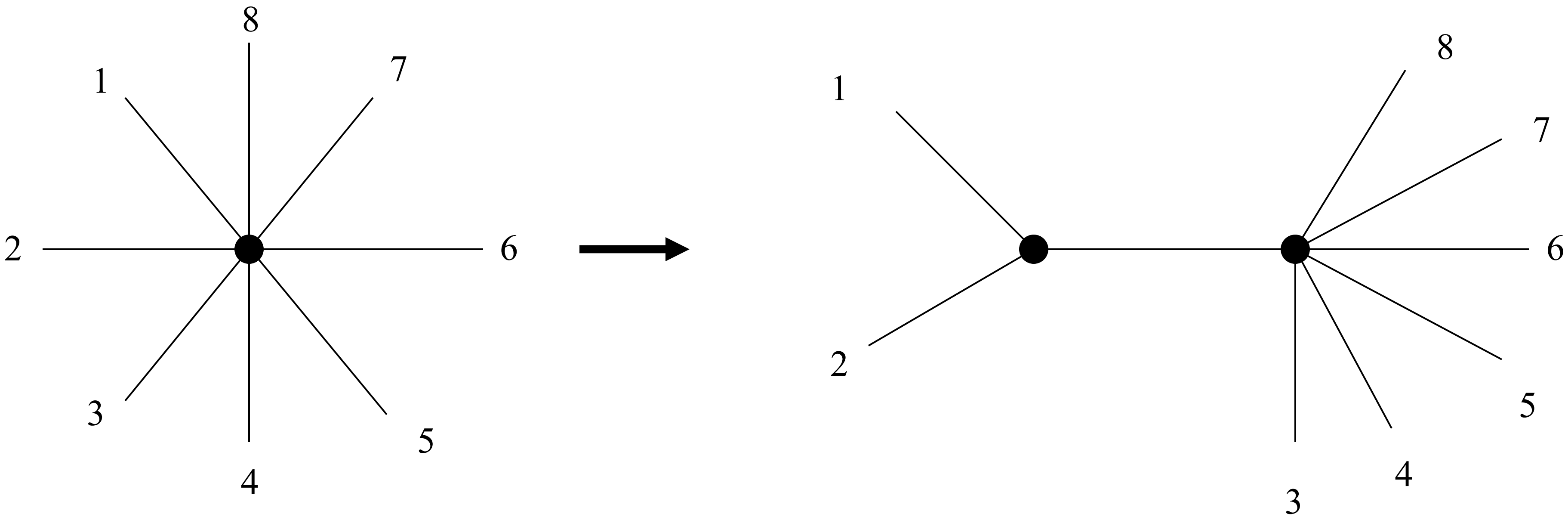
UPGMA weaknesses

- UPGMA assumes that the rates of evolution are the same among different lineages
- In general, should not use this method for phylogenetic tree reconstruction (unless believe assumption)
- Produces a rooted tree

Neighbor Joining

- Most widely-used distance based method for phylogenetic reconstruction
- UPGMA illustrated that it is not enough to just pick closest neighbors
- Idea here: take into account averaged distances to other leaves as well
- Produces an unrooted tree

Neighbor Joining (NJ)



Start off with star tree; pull out pairs at a time

NJ Algorithm

Step 1: Let $u_i = \sum_k \frac{M_{ik}}{n-2}$

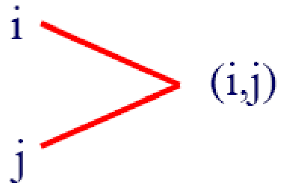
- (Almost) “average” distance to other nodes

Step 2: Choose i and j for which $M_{ij} - u_i - u_j$ is smallest

- Look for nodes that are close to each other, and far from everything else
- Turns out minimizing this is minimizing sum of branch lengths

NJ Algorithm

Step 3: Define a new cluster (*i*, *j*), with a corresponding node in the tree



Distance from *i* and *j* to node (*i*,*j*):

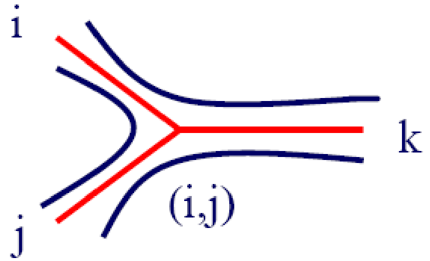
$$d_{i,\ (i,j)} = 0.5(M_{ij} + u_i - u_j)$$
$$d_{j,\ (i,j)} = 0.5(M_{ij} + u_j - u_i)$$

Default: split distance but if on average one is further away, make it longer

NJ Algorithm

Step 4: Compute distance between new cluster and all other clusters:

$$M_{(ij)k} = \frac{M_{ik} + M_{jk} - M_{ij}}{2}$$



Step 5: Delete *i* and *j* from matrix and replace by *(i, j)*

Step 6: Continue until only 2 leaves remain

NJ Performance

- Works well in practice
- If there is a tree that fits the matrix, it will find it
- Can sometimes get trees with negative length edges (!)

Maximum Parsimony

- Check each topology
- Count the minimum number of changes required to explain the data
- Choose the tree with the smallest number of changes

Table 6.3. *Example of phylogenetic analysis to find the correct unrooted tree from four aligned sequences by the maximum parsimony method*

Taxa	Sequence position (sites) and character								
	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G

Adapted from Li and Graur 1991.

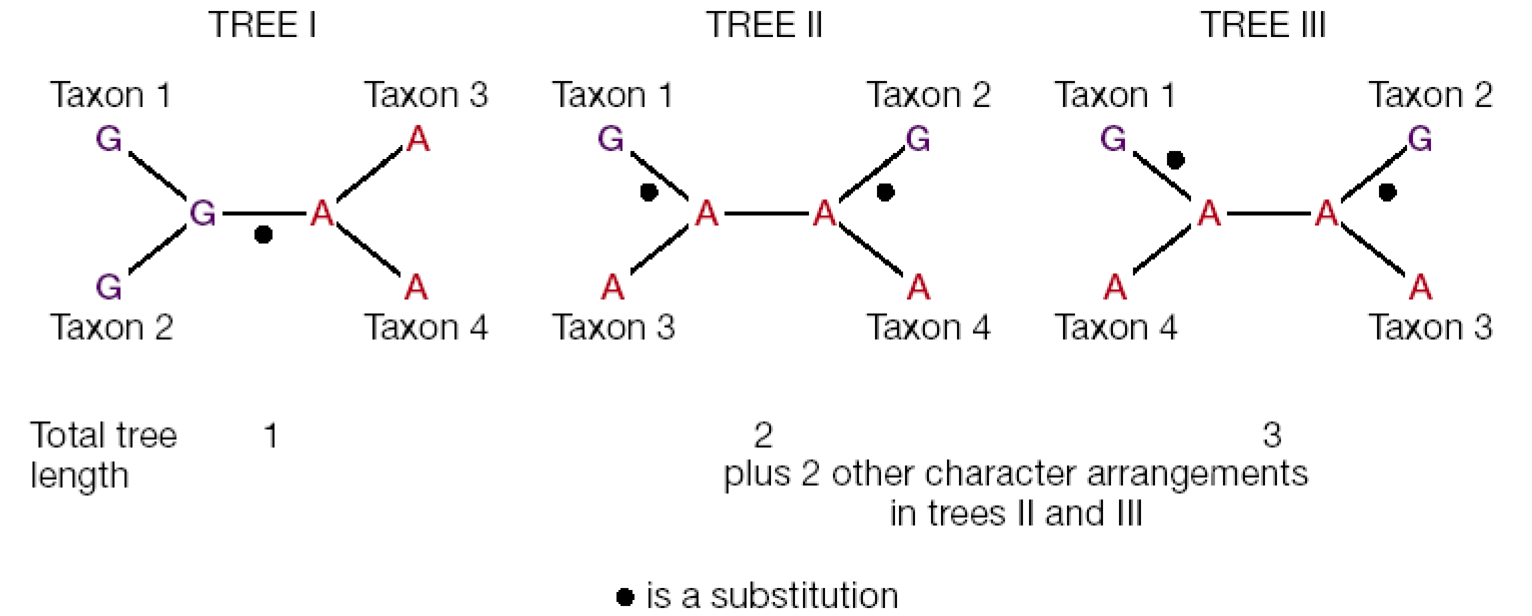
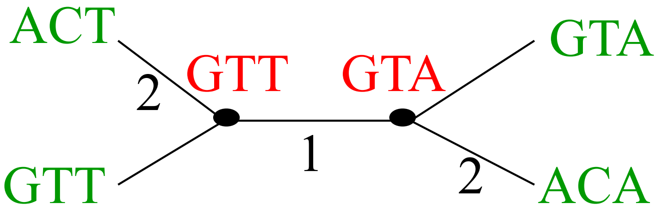
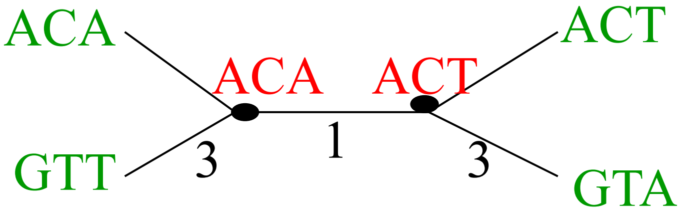


Figure 6.6. Example of phylogenetic analysis based on sequence position 5 in Table 6.3, using the maximum parsimony method. (Redrawn, with permission, from Li and Graur 1991 [copyright Sinauer Associates].)

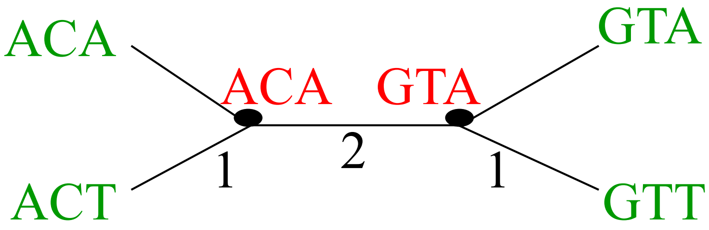
Maximum Parsimony



MP score = 5



MP score = 7



MP score = 4

Optimal MP tree

Maximum Parsimony: Limitations

With only a few sequences, becomes computationally intractable (“NP-hard”)

# of rooted trees = $\frac{(2n-3)!}{2^{n-2}(n-2)!}$

# of unrooted trees = $\frac{(2n-5)!}{2^{n-3}(n-3)!}$

Number of possible trees (Felsenstein 1978)

#of species	**#rooted trees**	**#unrooted trees**
2	1	1
3	3	1
4	15	3
5	105	15
10	3.44×10^{7}	2.03×10^{6}
15	2.13×10^{14}	7.91×10^{12}
20	8.20×10^{21}	2.21×10^{20}

Maximum Likelihood

- Given a probabilistic model for nucleotide (or protein) substitution (e.g., Jukes & Cantor), pick the tree that has highest probability of generating observed data
  - I.e., Given data D and model M, find tree T such that $Pr(D|T, M)$ is maximized
- Models gives values $p_{ij}(t)$, the probability of going from nucleotide i to j in time t

Maximum Likelihood

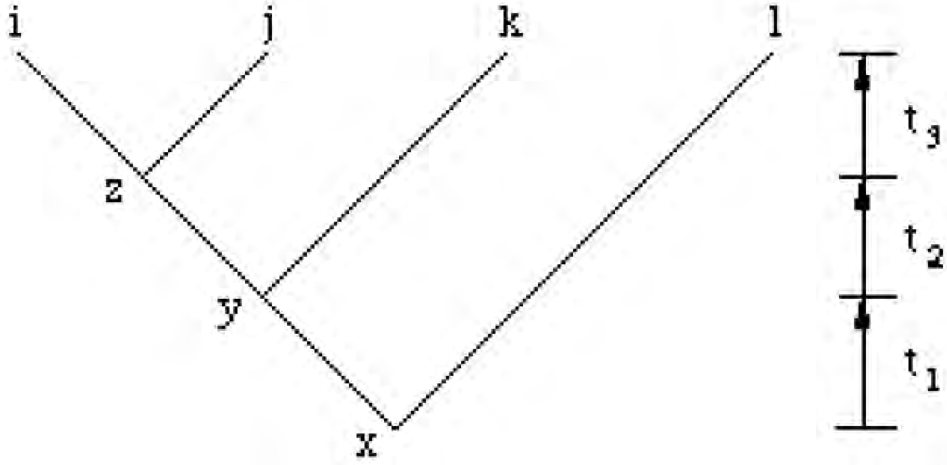
- Makes 2 independence assumptions
  - Different sites evolve independently
  - Diverged sequences (or species) evolve independently after diverging
- If D_i is data for *i*th site

$$Pr(D|T, M) = \prod_i Pr(D_i|T, M)$$

Maximum Likelihood

How to calculate $Pr(D_i|T,M)$?

$p_{xy}(t)$ ~ prob of going from x to y in time t



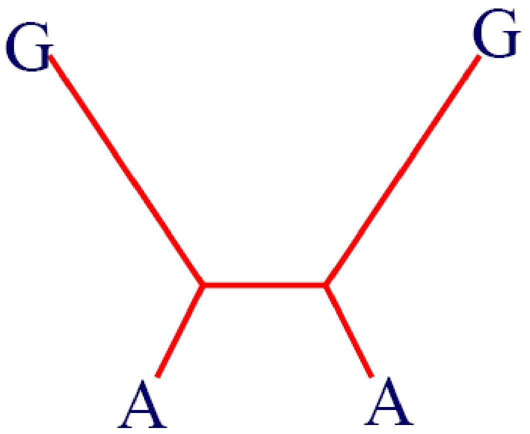
$$Pr(i,j,k,l|T,M) =$$
$$\sum_x \sum_y \sum_z pr(x)(p_{xl} \cdot (t_1+t_2+t_3) \cdot p_{xy}(t_1)$$
$$\cdot p_{yk}(t_2+t_3) \cdot p_{yz}(t_2) \cdot p_{zi}(t_3) \cdot p_{zj}(t_3))$$

Maximum Likelihood

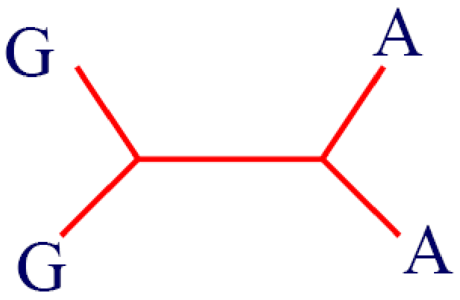
- Given tree topology and branch lengths, can efficiently calculate $Pr(D|T, M)$ using dynamic programming
  - I.e., don't have to enumerate over all internal states
- Finding best maximum likelihood tree is expensive
  - Must consider all topologies
  - Find best edge lengths for each topology
    - Idea: use some search procedure, e.g., EM, to optimize these lengths

Long Branches Attraction

Parsimony analysis implicitly assumes that rate of change along branches are similar



Real tree: two long branches
where G has turned to A independently



Inferred tree

Comparison of Methods

Neighbor-joining	**Maximum parsimony**	**Maximum likelihood**
Very fast	Slow	***Very*** slow
Easily trapped in local optima	Assumptions fail when evolution is rapid	Highly dependent on assumed evolution model
Good for generating tentative tree, or choosing among multiple trees	Best option when tractable (<30 taxa, strong conservation)	Good for very small data sets and for testing trees built using other methods

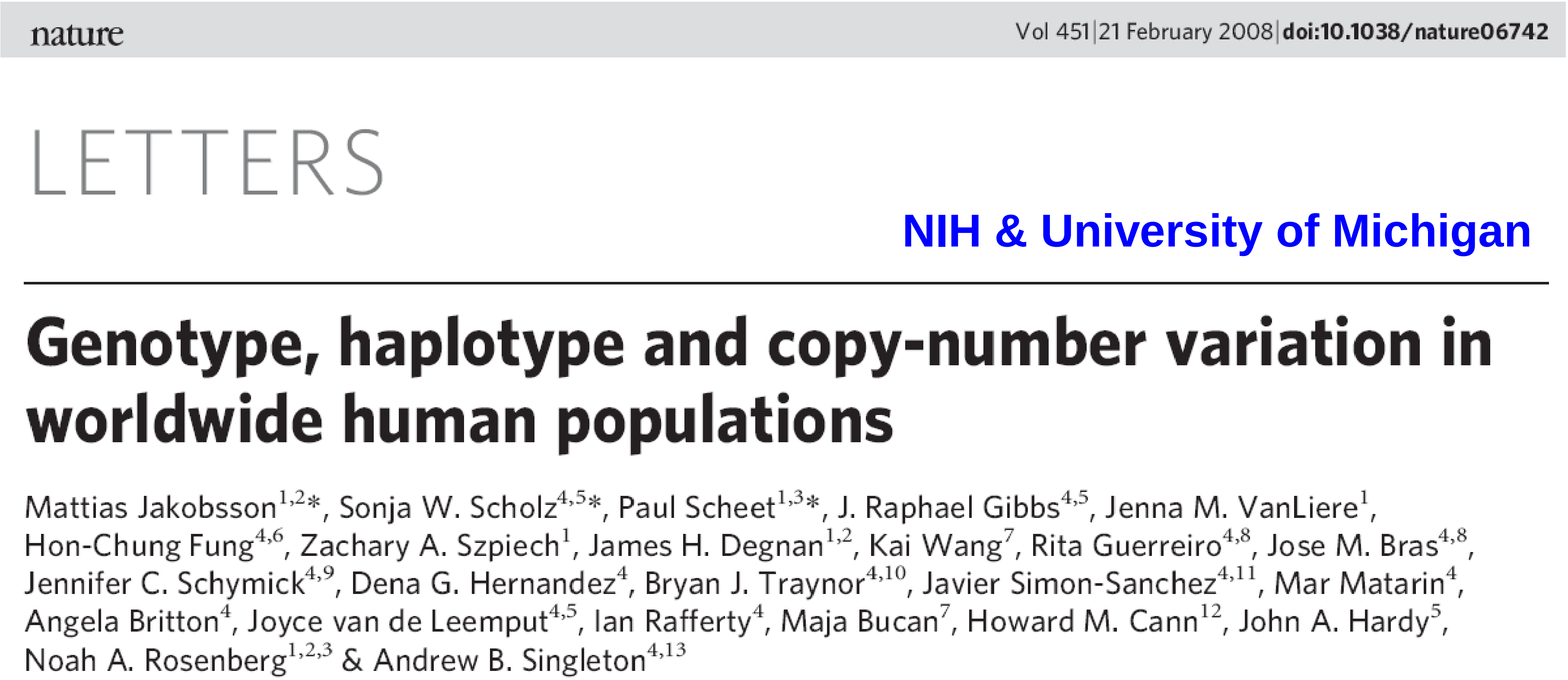
How confident am I that my tree is correct?

Bootstrapping: how dependent is the tree on the dataset

1. Randomly choose *n* objects from your dataset of *n*, *with replacement*
2. Rebuild the tree based on the subset of the data
3. Repeat 1,000 – 10,000 times
4. How often are the same children joined?

Jackknifing: how dependent is the tree on the dataset

1. Randomly choose *k* objects from your dataset of *n*, *without replacement*
2. Rebuild the tree based on the subset of the data
3. Repeat 1,000 – 10,000 times
4. How often are the same children joined?



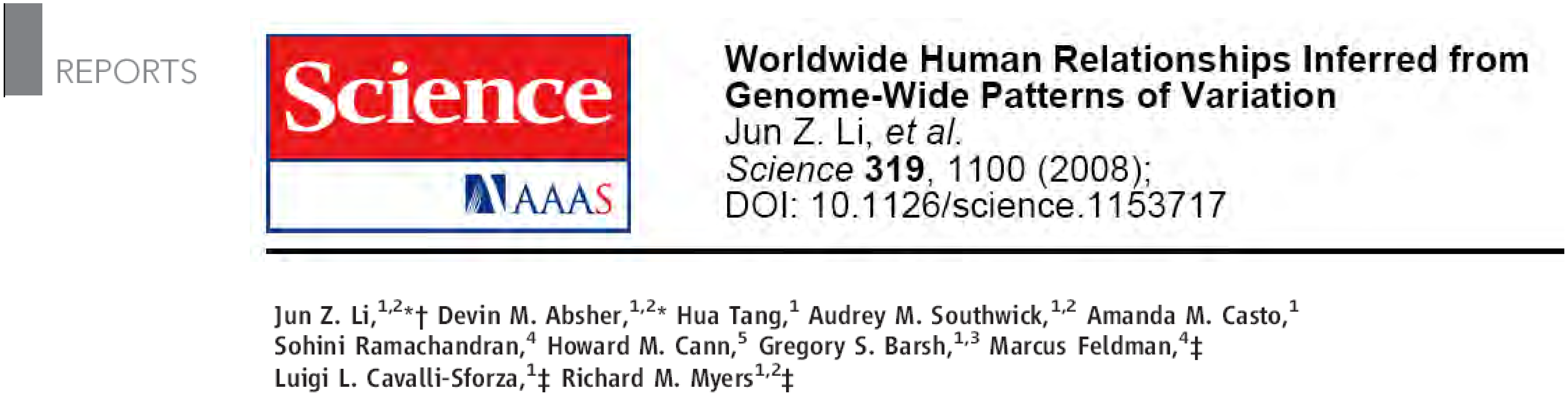
nature Vol 451|21 February 2008|doi:10.1038/nature06742

LETTERS

NIH & University of Michigan

Genotype, haplotype and copy-number variation in worldwide human populations

Mattias Jakobsson[1,2]*, Sonja W. Scholz[4,5]*, Paul Scheet[1,3]*, J. Raphael Gibbs[4,5], Jenna M. VanLiere[1], Hon-Chung Fung[4,6], Zachary A. Szpiech[1], James H. Degnan[1,2], Kai Wang[7], Rita Guerreiro[4,8], Jose M. Bras[4,8], Jennifer C. Schymick[4,9], Dena G. Hernandez[4], Bryan J. Traynor[4,10], Javier Simon-Sanchez[4,11], Mar Matarin[4], Angela Britton[4], Joyce van de Leemput[4,5], Ian Rafferty[4], Maja Bucan[7], Howard M. Cann[12], John A. Hardy[5], Noah A. Rosenberg[1,2,3] & Andrew B. Singleton[4,13]



REPORTS

Science AAAS

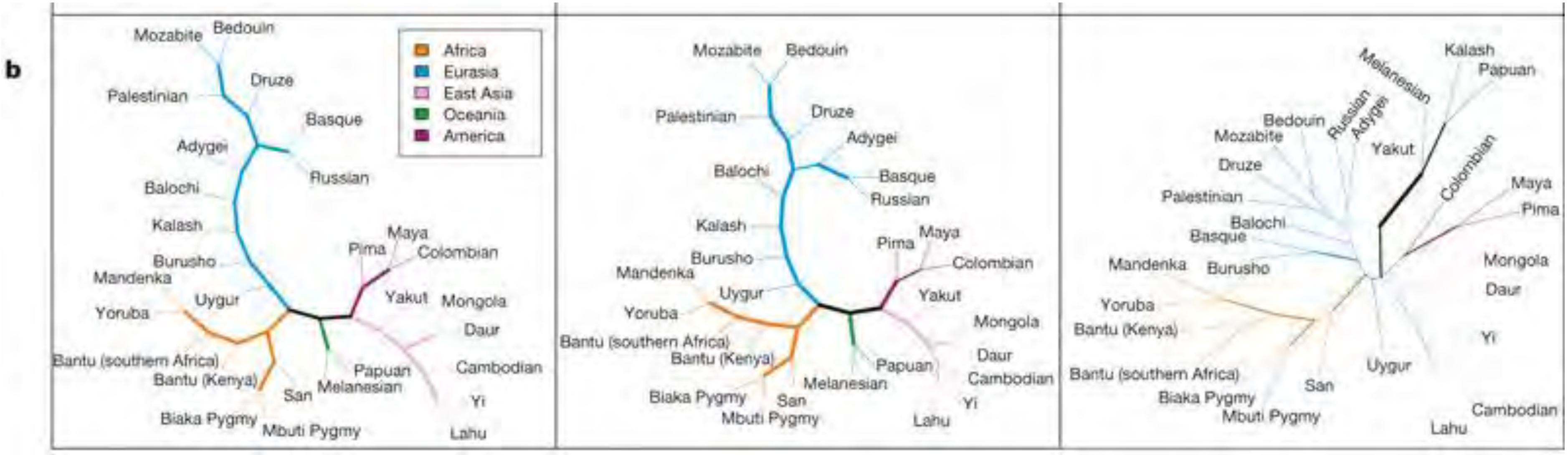
Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation

Jun Z. Li, *et al.*
Science **319**, 1100 (2008);
DOI: 10.1126/science.1153717

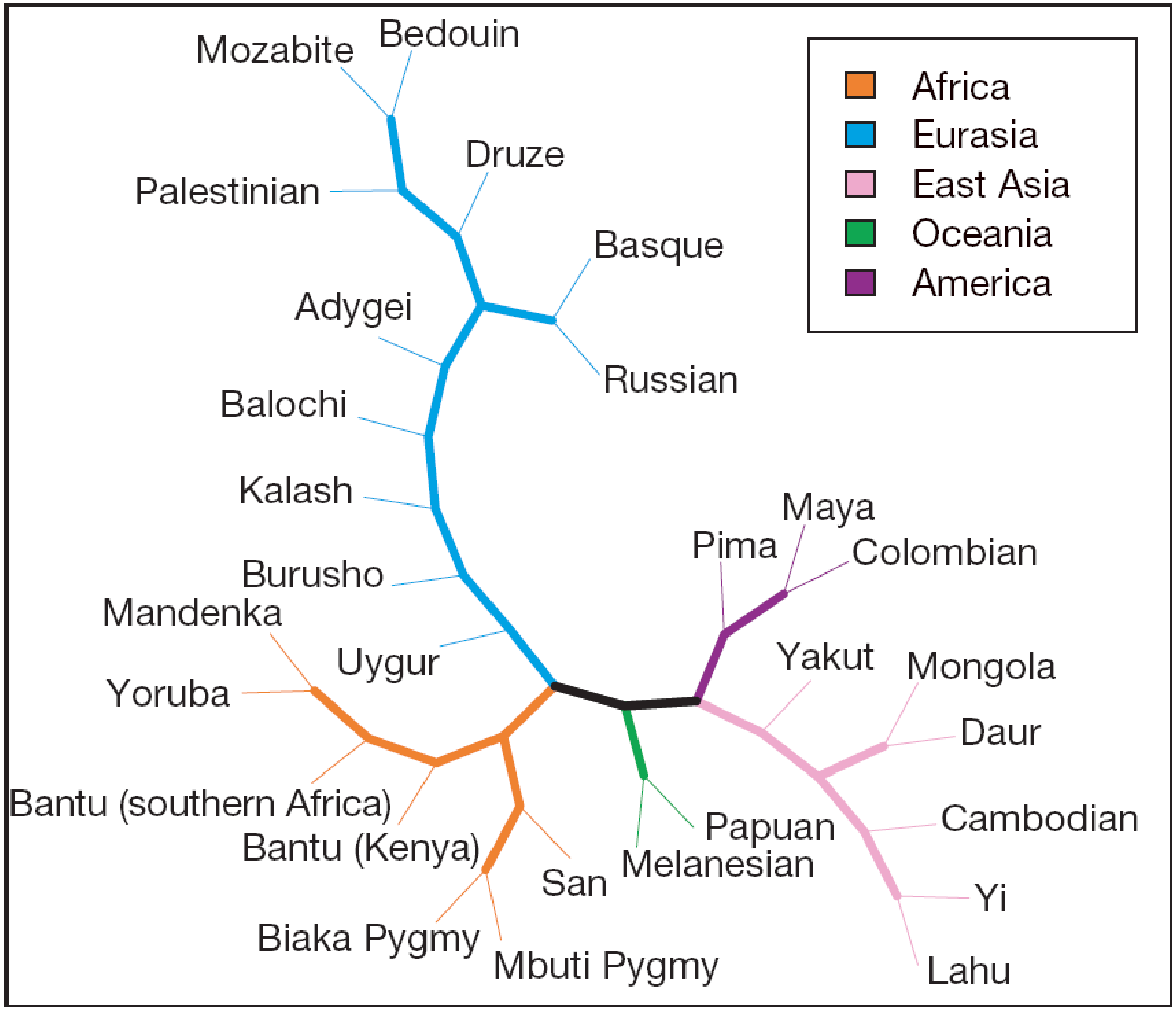
Jun Z. Li,[1,2]*† Devin M. Absher,[1,2]* Hua Tang,[1] Audrey M. Southwick,[1,2] Amanda M. Casto,[1]
Sohini Ramachandran,[4] Howard M. Cann,[5] Gregory S. Barsh,[1,3] Marcus Feldman,[4]‡
Luigi L. Cavalli-Sforza,[1]‡ Richard M. Myers[1,2]‡

Stanford University

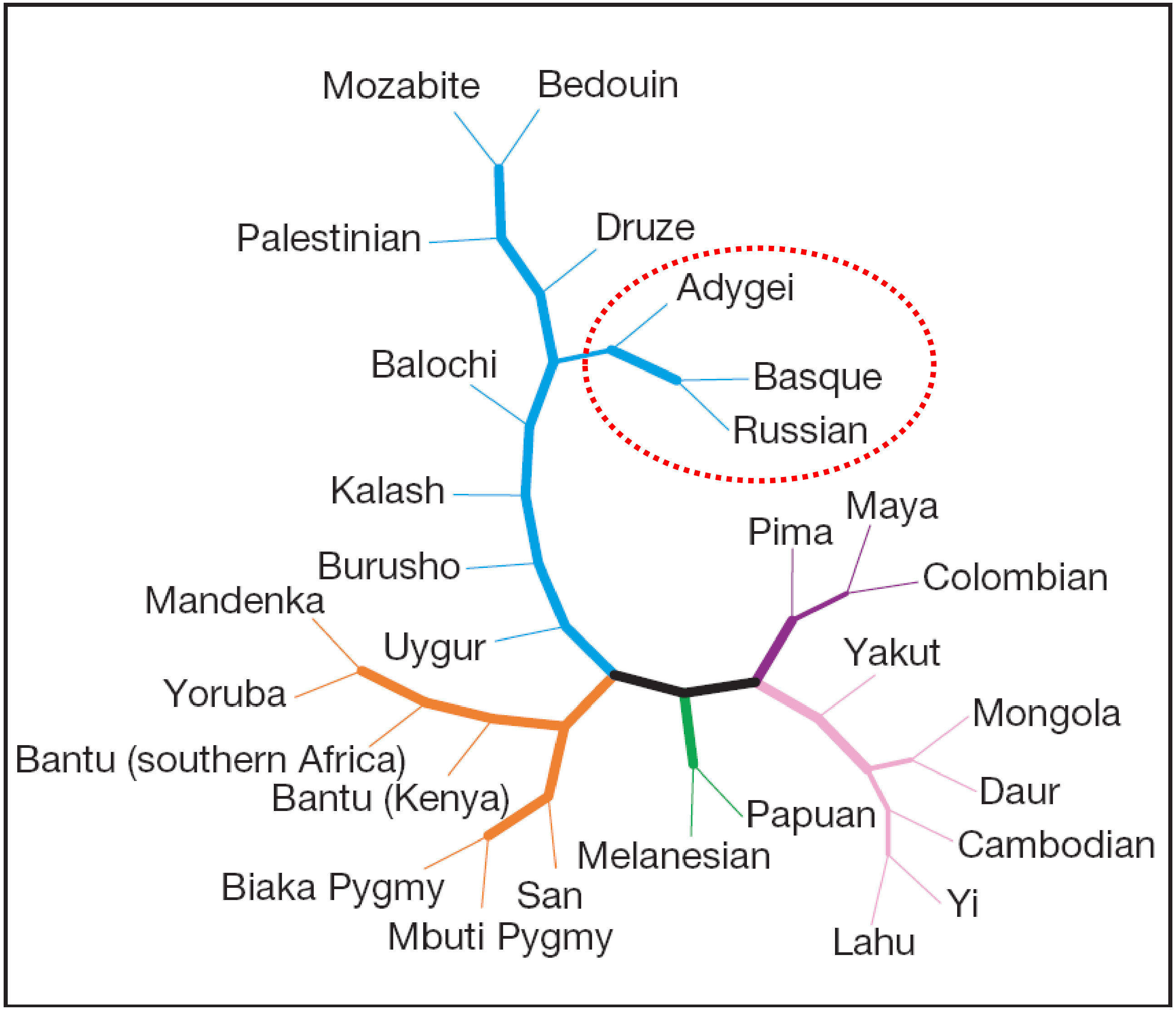
Neighbour-joining trees of population relationships



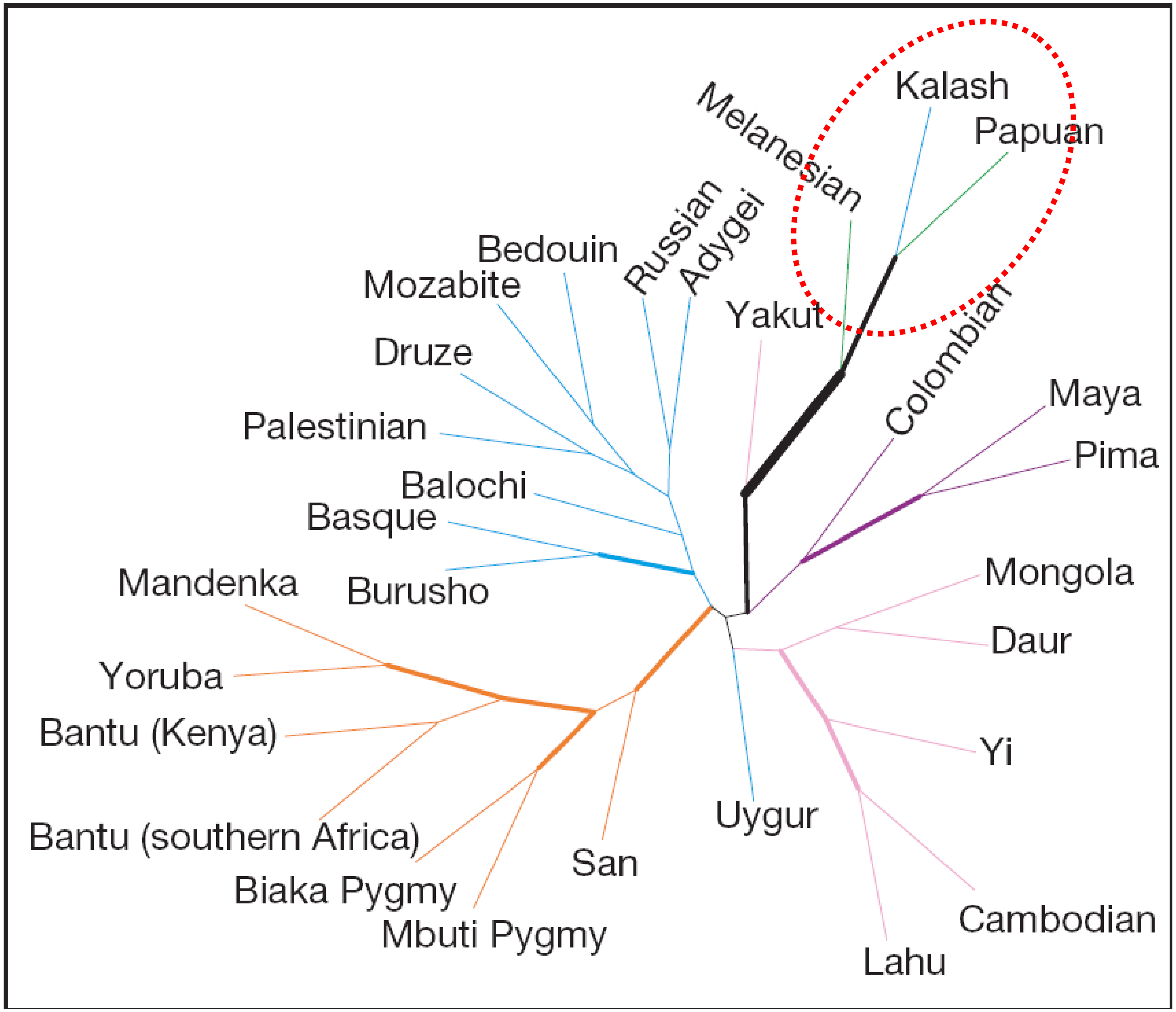
NJ tree based on SNP genotypes



NJ tree based on SNP haplotypes



NJ tree based on CNVs


Kalash
Papuan
Melanesian
Russian
Adygei
Bedouin
Mozabite
Yakut
Colombian
Druze
Maya
Pima
Palestinian
Balochi
Basque
Mandenka
Burusho
Mongola
Daur
Yoruba
Bantu (Kenya)
Yi
Uygur
Bantu (southern Africa)
San
Biaka Pygmy
Mbuti Pygmy
Cambodian
Lahu