Phylogeny

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

2/28/2017

review

Alignment-based SNP discovery

General procedure

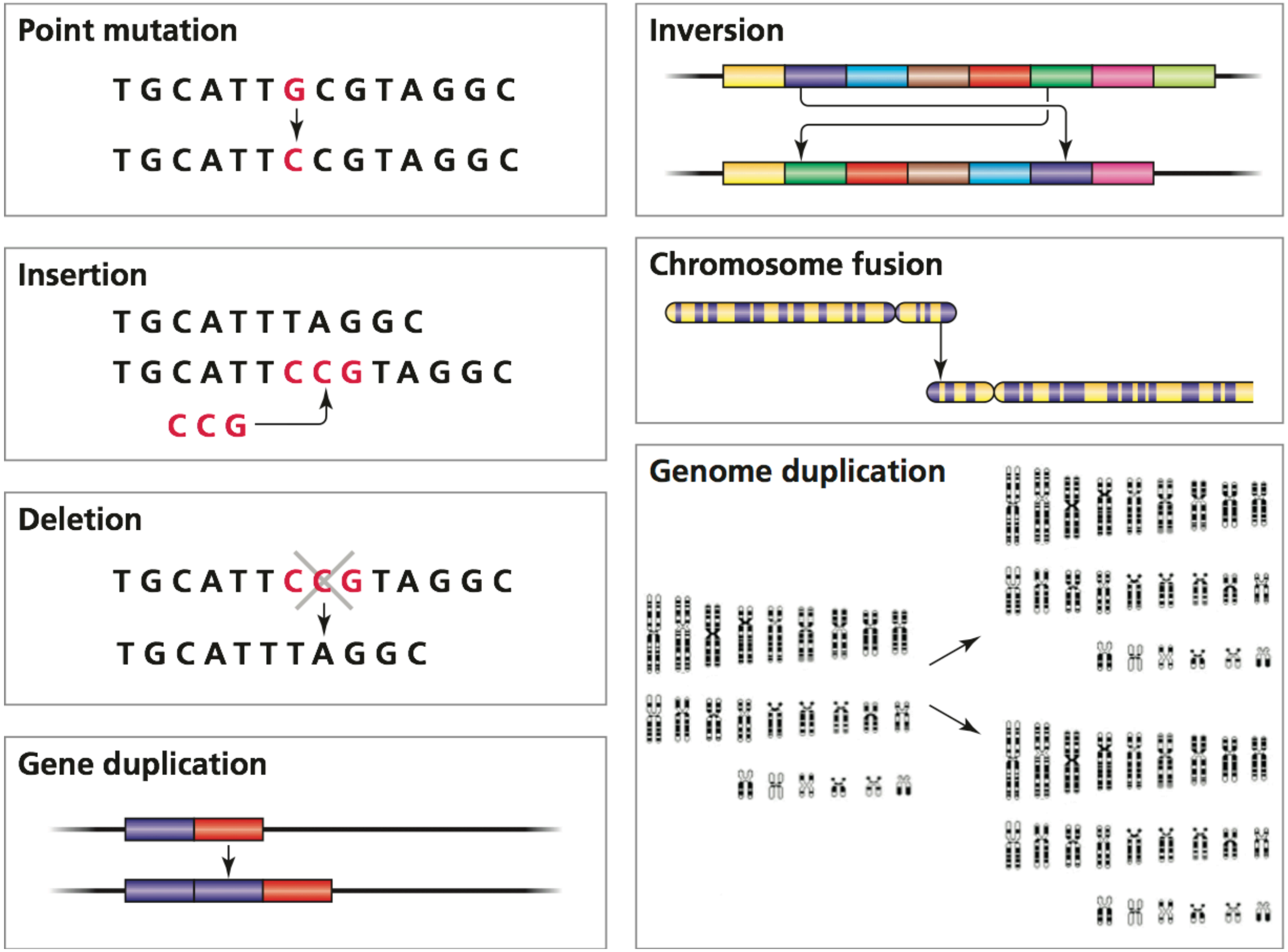
- Reads cleanup (adaptor, quality trimming, e.g., trimmomatic)
- Reads aligned to the reference genome with aligners
  1. BWA, Bowtie (DNA-Seq reads)
  2. GSNAP, Tophat, star (RNA-Seq reads)
- Post-alignment filtering and convert SAM (alignment file) to BAM (samtools or others)
- SNP calling with software packages: Samtools, GATK
- Use population information or some criteria to filter SNP sets

review

Genomic variants (polymorphism)

1. SNP

2. INDEL



3. genomic structural variation

- copy number variation (presence-absence variation)
- other re-arrangements

Outline

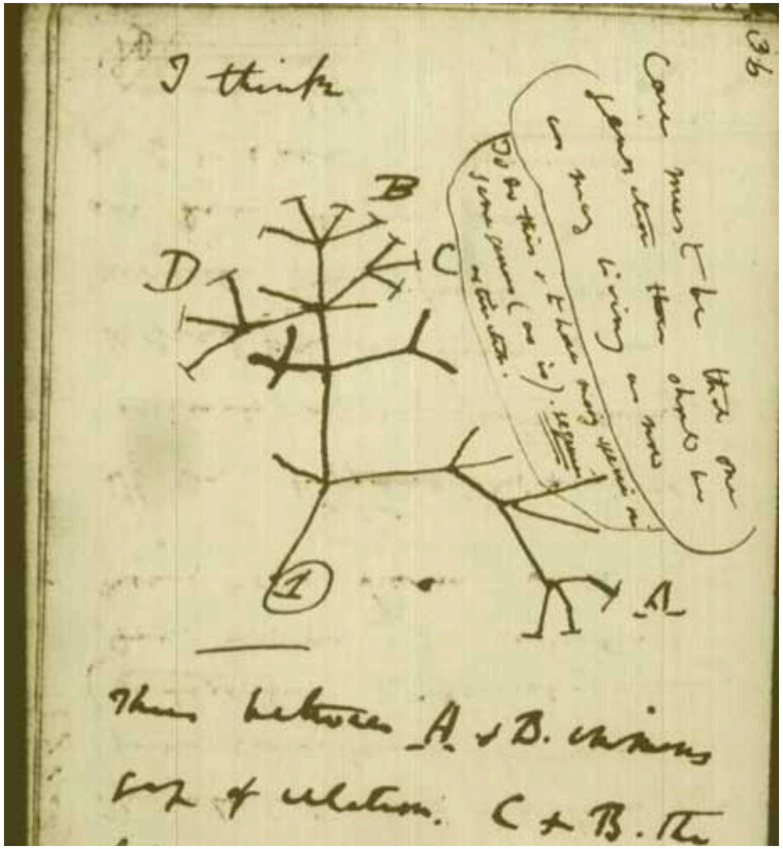
- Background
- Algorithms to build phylogeny
- Interpretation of phylogenic trees

Tree of life

The **tree of life** is used to describe the relationships between organisms. Its use dates back to at least the early 1800s. It was employed by ***Charles Darwin*** to express the concept of the branching divergence of varieties and then species in a process of descent from common ancestors.

Ernst Haeckel coined the term **phylogeny** for the evolutionary relationships of species through time, and went further than Darwin in proposing phylogenic histories of life. The modern development of this idea is called the **phylogenetic tree**.

- wikipedia


36
I think
B
C
D
1
A
Thus between A & B. immense
gap of relation. C & B. the

A page from Darwin's Notebook B showing his sketch of the tree of life

theguardian.com

Evolution and Phylogeny

- Evolution is a process of change. At the molecular level, evolution is a process of mutation with selection.

- Molecular evolution is the study of changes in genes and proteins throughout different braches of the tree of life.

- Phylogeny is **the inference** of evolutionary relationships, providing **hypotheses** of past biological events.

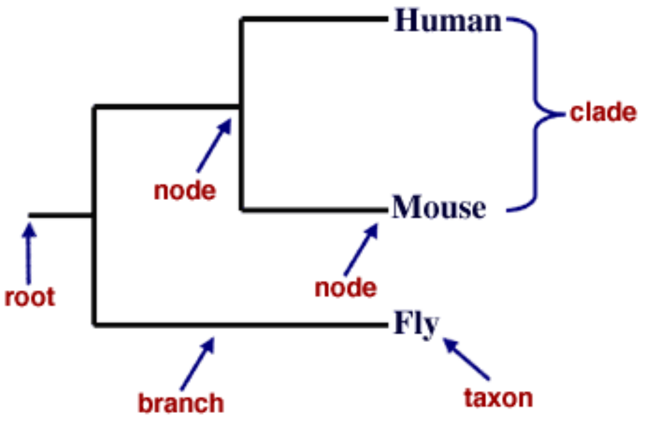
Applications of phylogenic trees (examples)

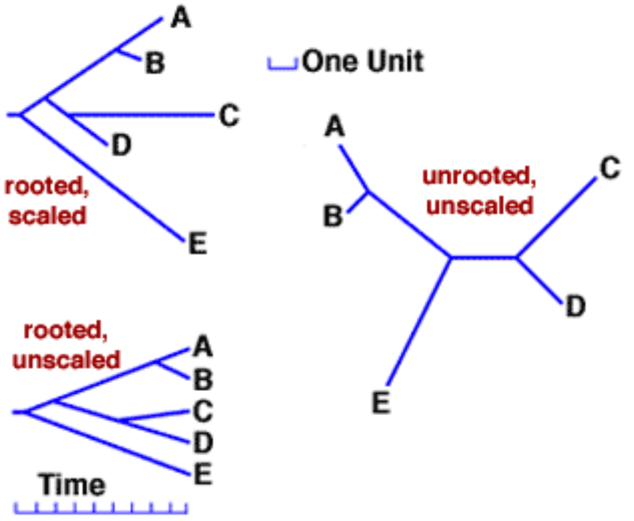
1. to represent the relationships among species/isolates/ varieties/genotypes/lines
2. to describe relationships among homologs in a gene family
3. to infer the evolutionary and epidemiological dynamics of pathogens
4. to classify metagenomic sequences

Nowadays, every biologist needs to know something about phylogenetic inference!

Tree components

- A phylogeny is a tree containing **nodes** that are connected by **braches**. The pattern of branching determines the tree's **topology**.
- **Root** represents a common ancestor of all taxa shown in the tree.
- **Rooted trees** are thus directional, since all taxa evolved from the root.
- **Unrooted trees** illustrate relationships only.





www.talkorigins.org/faqs/comdesc/phylo.html

An unrooted tree

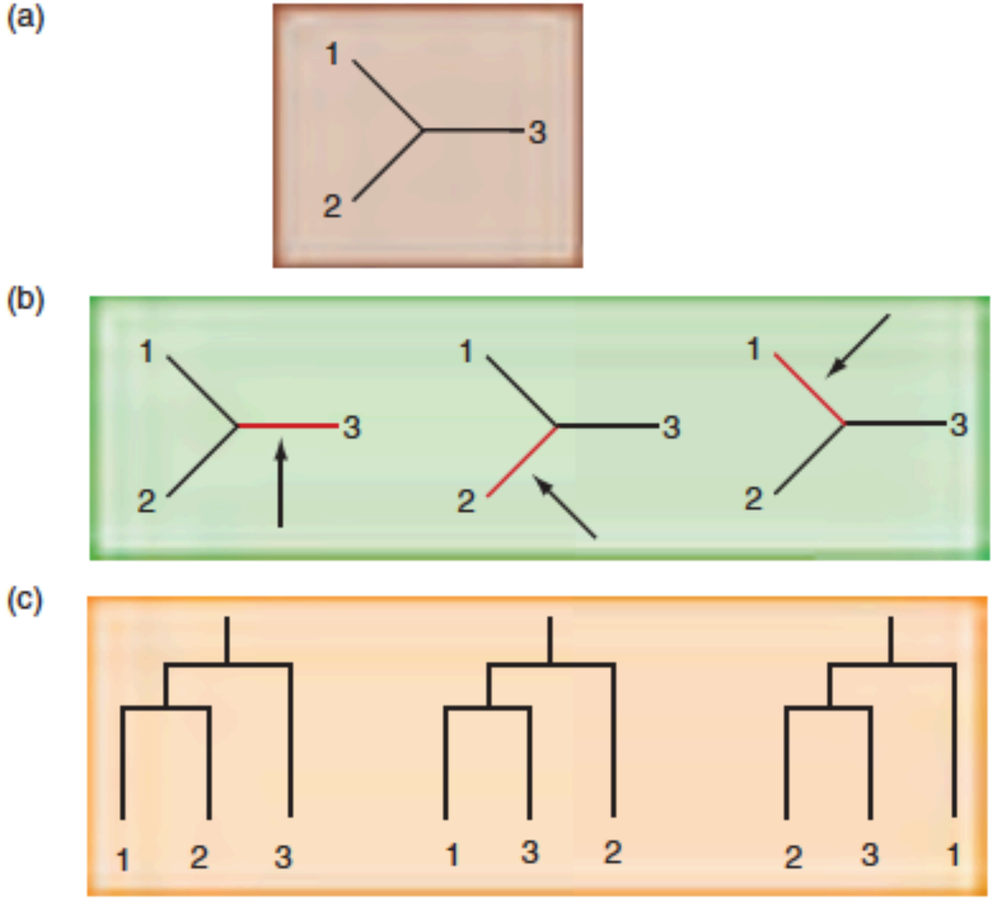
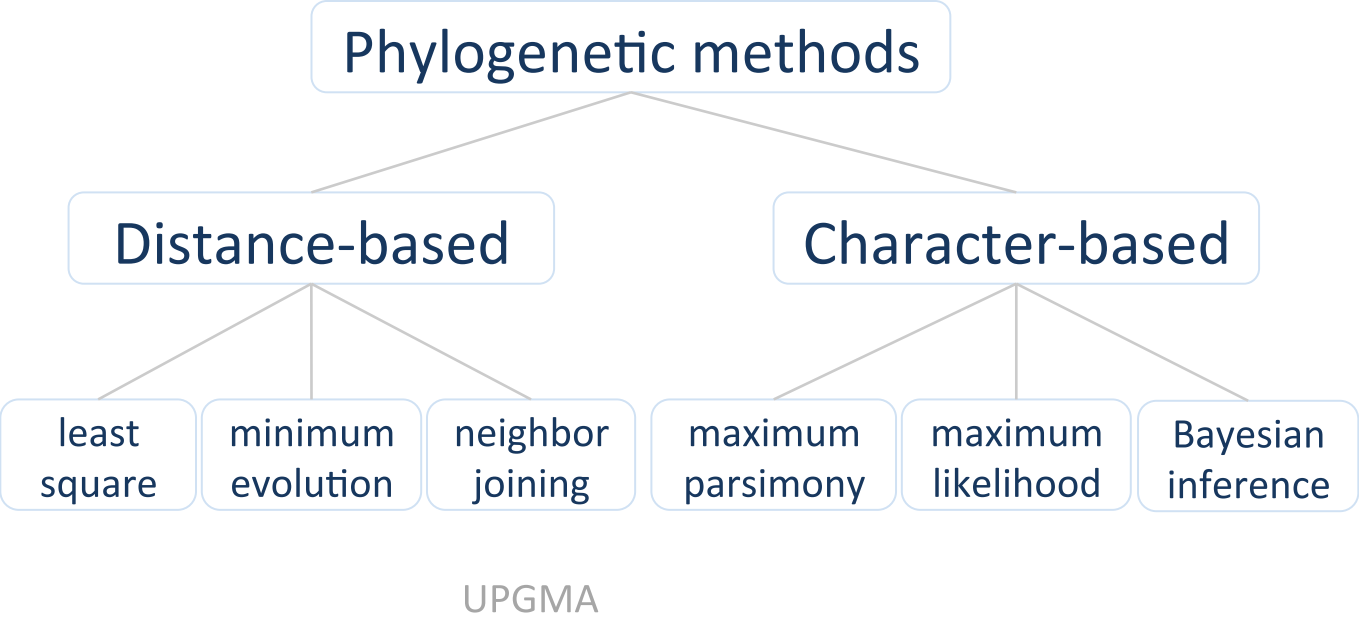


FIGURE 7.10 For three operational taxonomic units (such as three aligned protein sequences 1–3), there is (a) one possible unrooted tree. (b) Any of these edges may be used to select a root (see arrows), from which (c) three corresponding rooted trees are possible.

Approaches to construct phylogenic trees



Phylogenetic methods

Distance-based

Character-based

least square

minimum evolution

neighbor joining

maximum parsimony

maximum likelihood

Bayesian inference

UPGMA

Distance-based method

- Distance calculation to build a distance matrix

A **distance matrix** is a table that indicates pairwise **dissimilarity**.

	A	B	C	D
A	0	3	4	6
B	3	0	5	7
C	4	5	0	6
D	6	7	6	0

	B	C	D
A	3	4	6
B		5	7
C			6

example from Barbara Holland

Distance between DNA sequences

- Distance (e.g., percentage of difference) between DNA sequences

A	ATTTGCGGTA
B	ATCTGCGATA
C	ATTGCCGTTT
D	TTCGCTGTTT

	A	B	C	D
A	0	0.2	0.4	0.7
B	0.2	0	0.5	0.6
C	0.4	0.5	0	0.3
D	0.7	0.6	0.3	0

example from Barbara Holland

e.g., DNA sequences

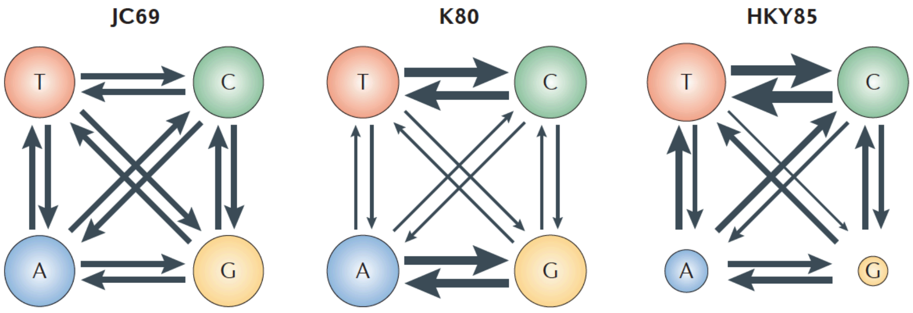
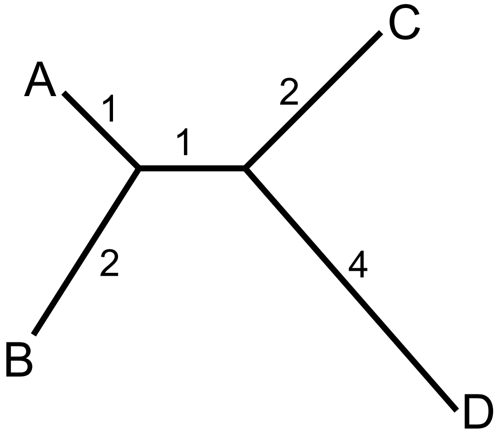


Figure 1 | **Markov models of nucleotide substitution.** The thickness of the arrows indicates the substitution rates of the four nucleotides (T, C, A and G), and the sizes of the circles represent the nucleotide frequencies when the substitution process is in equilibrium. Note that both JC69 and K80 predict equal proportions of the four nucleotides.

From distance matrix to a tree

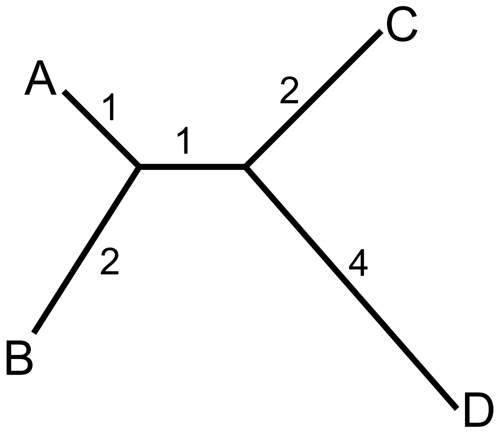
	B	C	D
A	3	4	6
B		5	7
C			6



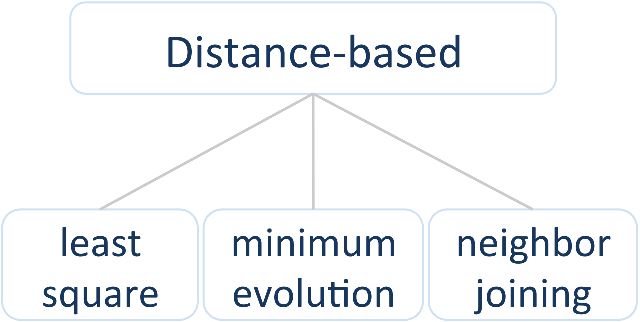
Perfect match

	B	C	D
A	3	4	7
B		4	7
C			6

?



No perfect matches



Distance-based

least square

minimum evolution

neighbor joining

example from Barbara Holland

Distance-based method – least squares

- **least squares**

	B	C	D
A	3	4	7
B		4	7
C			6

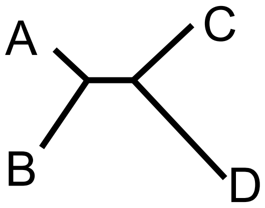
1) Given the distance matrix data, a tree is assumed and a new distance matrix is determined based on the tree

2) Determine Q values

3) Repeat 1 and 2 for all the possible trees

4) The tree with the smallest Q score is the least squares estimate of the true tree.

1



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

2

$$Q=\sum_{i=1}^{s}\sum_{i=1}^{s}(\hat{d}_{ij}-d_{ij})^2$$

3

$Q_1, Q_2, Q_3, \ldots$

4

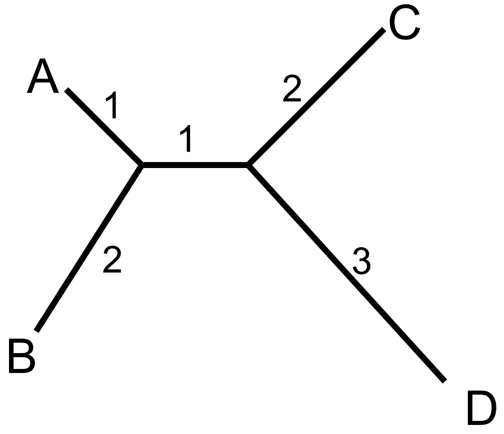
The tree with the smallest *Q*

Distance-based method – least squares

- **least squares**

d_{ij}

	B	C	D
A	3	4	7
B		4	7
C			6



$d_{ij}(hat)$

	B	C	D
A	3	4	5
B		5	6
C			5

$$Q = \sum_{i=1}^{s}\sum_{i=1}^{s}(\hat{d}_{ij} - d_{ij})^2$$

1) Given the distance matrix data, a tree is assumed and a new distance matrix is determined based on the tree

2) Determine Q values

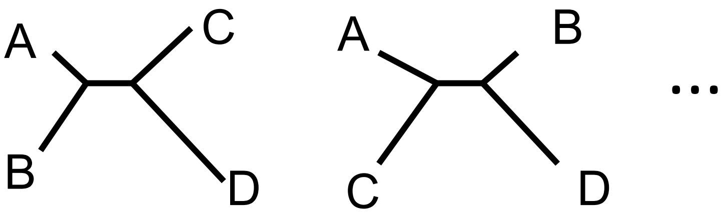
$(3\text{-}3)^2 + (4\text{-}4)^2 + (5\text{-}7)^2 +$
$(5\text{-}4)^2 + (6\text{-}7)^2 +$
$(5\text{-}6)^2 = 7$

Distance-based method – least squares

- **least squares**

d_{ij}

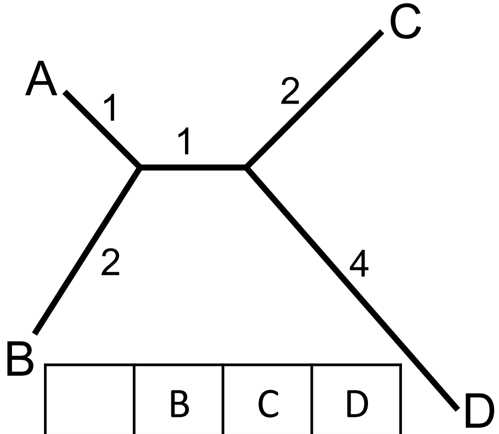
	B	C	D
A	3	4	7
B		4	7
C			6



tree 1

tree 2

$$Q = \sum_{i=1}^{s}\sum_{i=1}^{s}(\hat{d}_{ij} - d_{ij})^2$$



	B	C	D
A	3	4	6
B		5	7
C			6

Q = 2

1) Given the distance matrix data, a tree is assumed and a new distance matrix is determined based on the tree

2) Determine Q values

3) Repeat 1 and 2 for all the possible trees

4) The tree with the smallest Q score is the least squares estimate of the true tree.

Distance-based method –minimum evolution

- **minimum evolution**

uses the tree length (which is the sum of branch lengths) instead of Q for tree selection. Under the minimum evolution criterion, shorter trees are more likely to be correct than longer trees are.

sum of d_{ij}(hat)

	B	C	D
A	3	4	6
B		5	7
C			6

	B	C	D
A	3	4	5
B		5	6
C			5

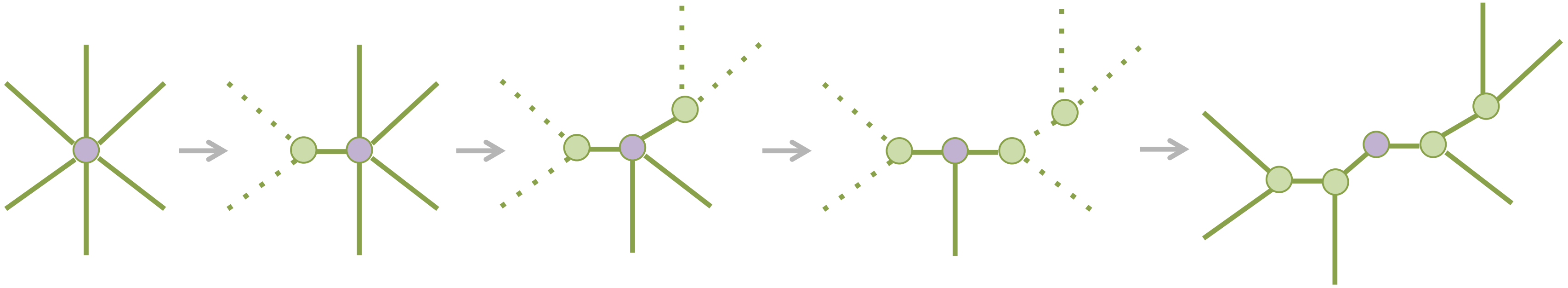
3 + 4 + 6 + 5 + 7 + 6 = 31 < 3 + 4 + 5 + 5 + 6 + 5 = 28

Distance-based method – neighbor joining

- **neighbor joining** (the most popular algorithm)

- Find a pair of leaves that are close to each other but far from other leaves

The algorithm operates by starting with a star tree and successively choosing a pair of taxa to join together (based on the taxon distances), until a fully resolved tree is obtained.



neighbor joining (NJ) procedure - I

Distance matrix

	A	B	C	D
A	0	2	4	6
B		0	3	7
C			0	10
D				0

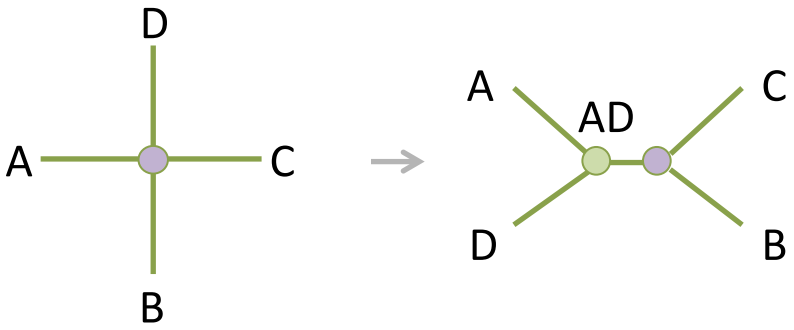
$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

Q_{AB} = (4-2)*2 - (2 + 4 + 6) - (2 + 3 + 7) = -20
Q_{AC} = (4-2)*4 - (2+ 4 + 6) - (4 + 3 + 9) = -20
Q_{AD} = (4-2)*6 - (2 + 4 + 6) - (6 + 7 + 9) = -22
Q_{BC} = (4-2)*3 - (2 + 3 + 7) - (4 + 3 + 9) = -20
Q_{BD} = (4-2)*7 - (2 + 3 + 7) - (6 + 7 + 9) = -20
Q_{CD} = (4-2)*9 - (4 + 3 + 6) - (6 + 7 + 6) = -14

Q matrix

	A	B	C	D
A	0	-20	-20	**-22**
B		0	-20	-20
C			0	-14
D				0

1. Based on the current distance matrix calculate the matrix Q.
2. Find the pair of distinct taxa *i* and *j* for which $Q_{i,j}$ has the lowest value. These taxa are joined to a newly node (e.g., AB), which is connected to the central node.



neighbor joining (NJ) procedure - II

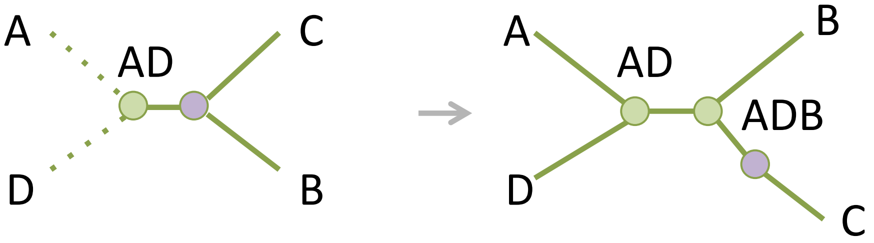
Updated distance matrix

$$d(u,k) = \frac{1}{2}[d(f,k) + d(g,k) - d(f,g)]$$

	AD	B	C
AD	0		
B		0	
C			0

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

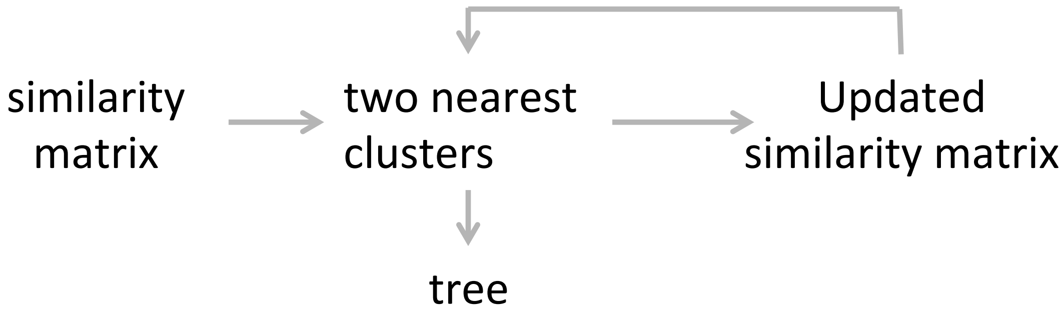
1. Based on the current distance matrix calculate the matrix Q.
2. Find the pair of distinct taxa *i* and *j* for which $Q_{i,j}$ has its lowest value. These taxa are joined to a newly node (AD), which is connected to the central node.
3. Calculate the distance from each of the taxa in the pair to this new node.
4. Calculate the distance from each of the taxa outside of this pair to the new node.
5. Repeat until resolving all



UPGMA (unweighted pair group method with arithmetic mean)

- UPGMA is a simple agglomerative (bottom-up) hierarchical clustering method, constructing a rooted tree.

- Based on a similarity matrix, at each step, the nearest two clusters are combined into a higher-level cluster.



similarity matrix

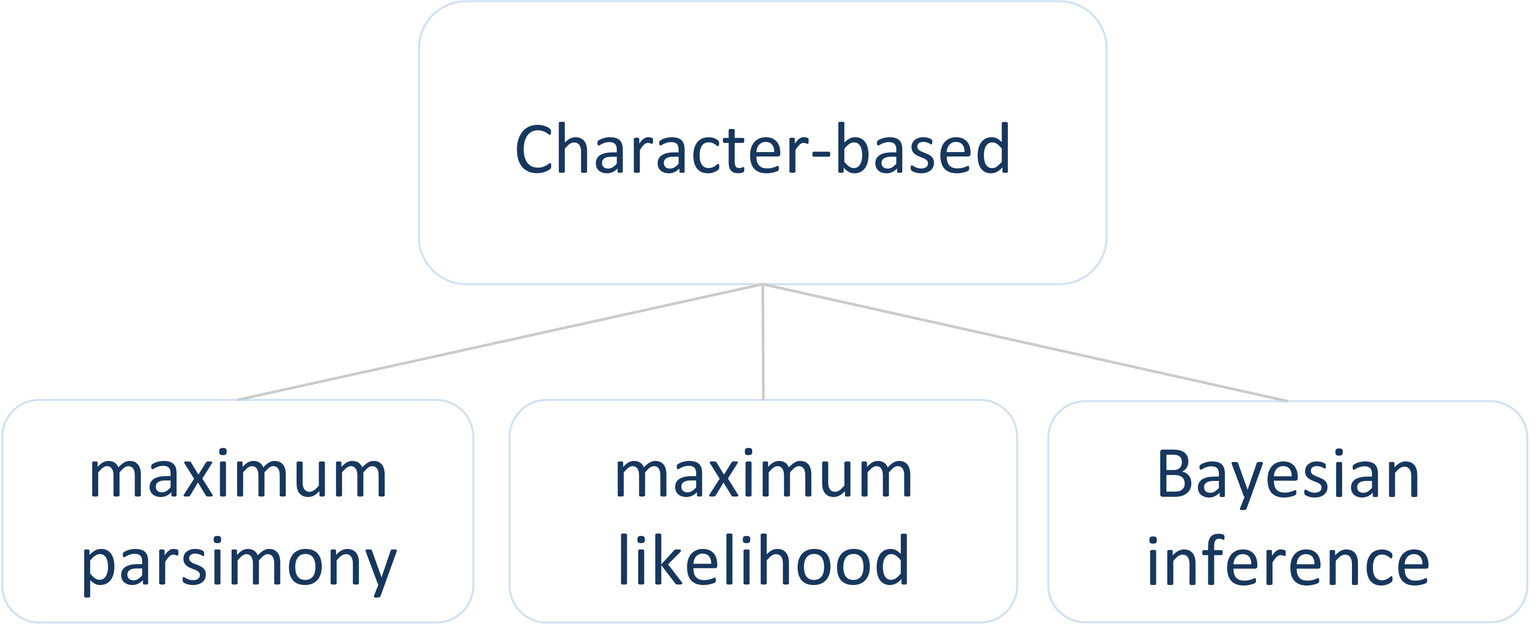
two nearest clusters

Updated similarity matrix

tree

- UPGMA assumption: the distances from the root to every branch tip are equal.

Character-based methods



Character-based

maximum parsimony

maximum likelihood

Bayesian inference

character-based method - maximum parsimony

The **maximum parsimony** method minimizes the number of changes on a phylogenetic tree by assigning character states to interior nodes on the tree.

The **character length** is the minimum number of changes required for that site, whereas the **tree score** is the sum of character lengths over all sites. The **maximum parsimony tree** is the tree that minimizes the tree score.

```
           (1)  GGA          ACA (3)
1 GGA            \1          1/          Tree score
2 GGG             \    2    /
3 ACA             GGG --- ACG            Tree I:   4
4 ACG             /         \
                 /0          0\
           (2)  GGG          ACG (4)

           (1)  GGA          GGG (2)
                 \1          1/
                  \    1    /
                  GCA --- GCG            Tree II:  5
                  /         \
                 /1          1\
           (3)  ACA          ACG (4)

           (1)  GGA          GGG (2)
                 \2          1/
                  \    0    /
                  GCG --- GCG            Tree III: 6
                  /         \
                 /1          2\
           (4)  ACG          ACA (3)
```

http://www.icp.ucl.ac.be/~opperd/private/parsimony.html

maximum likelihood

- **maximum likelihood**

The maximum likelihood estimates (MLEs) of parameters are the parameter values that maximize the likelihood.

From a statistical point of view, the tree is a model, whereas branch lengths on the given tree and substitution parameters are parameters in the model.

the probability or likelihood of observing the data given the model: P(Data|Model) or P(D|M)

1. optimization of branch lengths to calculate the tree likelihood, P(D|M), for each candidate tree

2. a search in the tree space for the tree with the maximum likelihood.

Bayesian inference methods

- **Bayesian inference methods**

$$P(T,\theta|D) = \frac{P(T,\theta)P(D|T,\theta)}{P(D)}$$

where P(T,θ) is the **prior probability** for tree T and parameter θ, P(D|T,θ) is the likelihood or probability of the data given the tree and parameter, and P(T,θ|D) is the **posterior probability**. The denominator P(D) is a normalizing constant, as its role is to ensure that P(T,θ|D) sums over the trees and integrates over the parameters to one.

the posterior probability of a tree is simply the probability that the tree is correct

Strengths and weaknesses

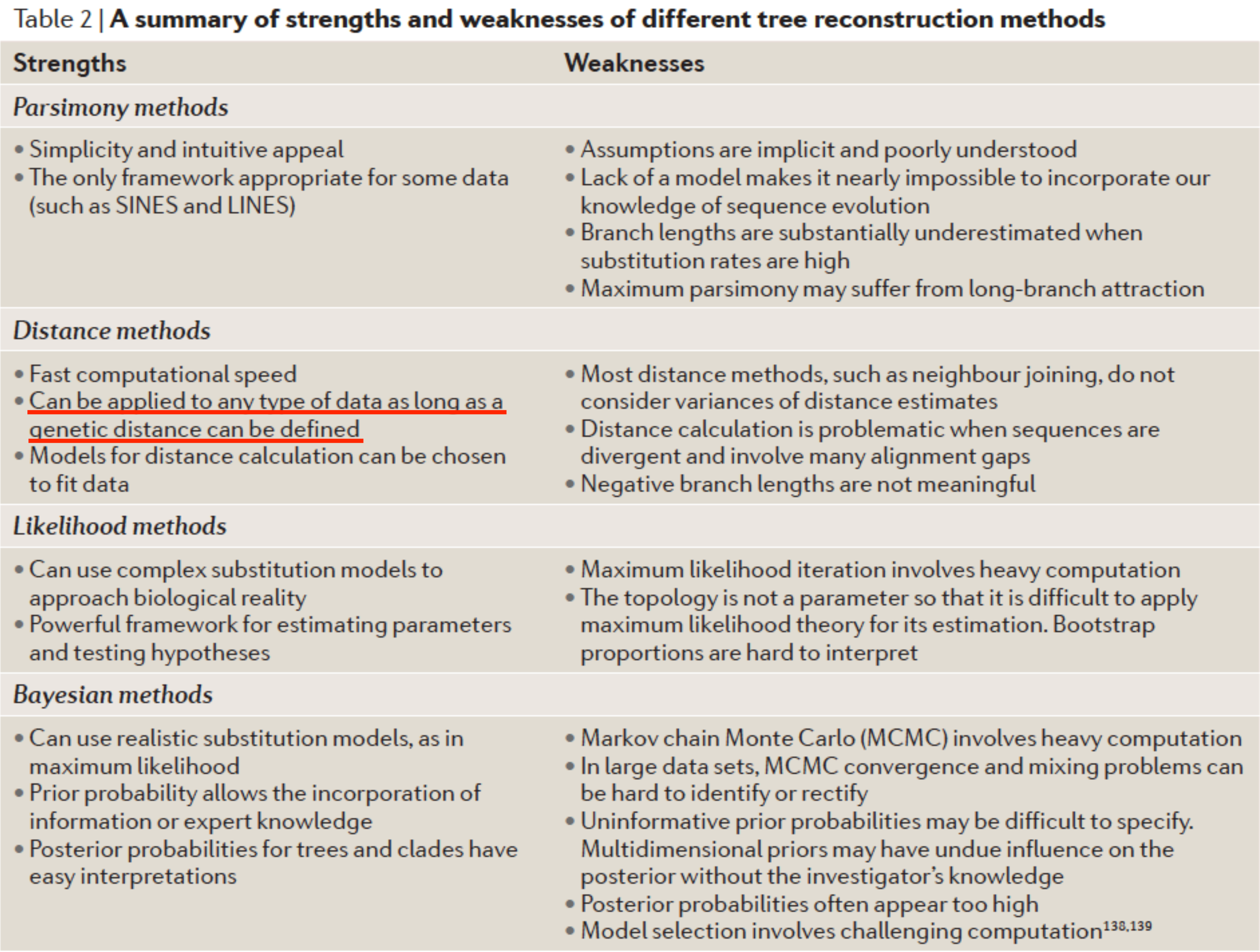
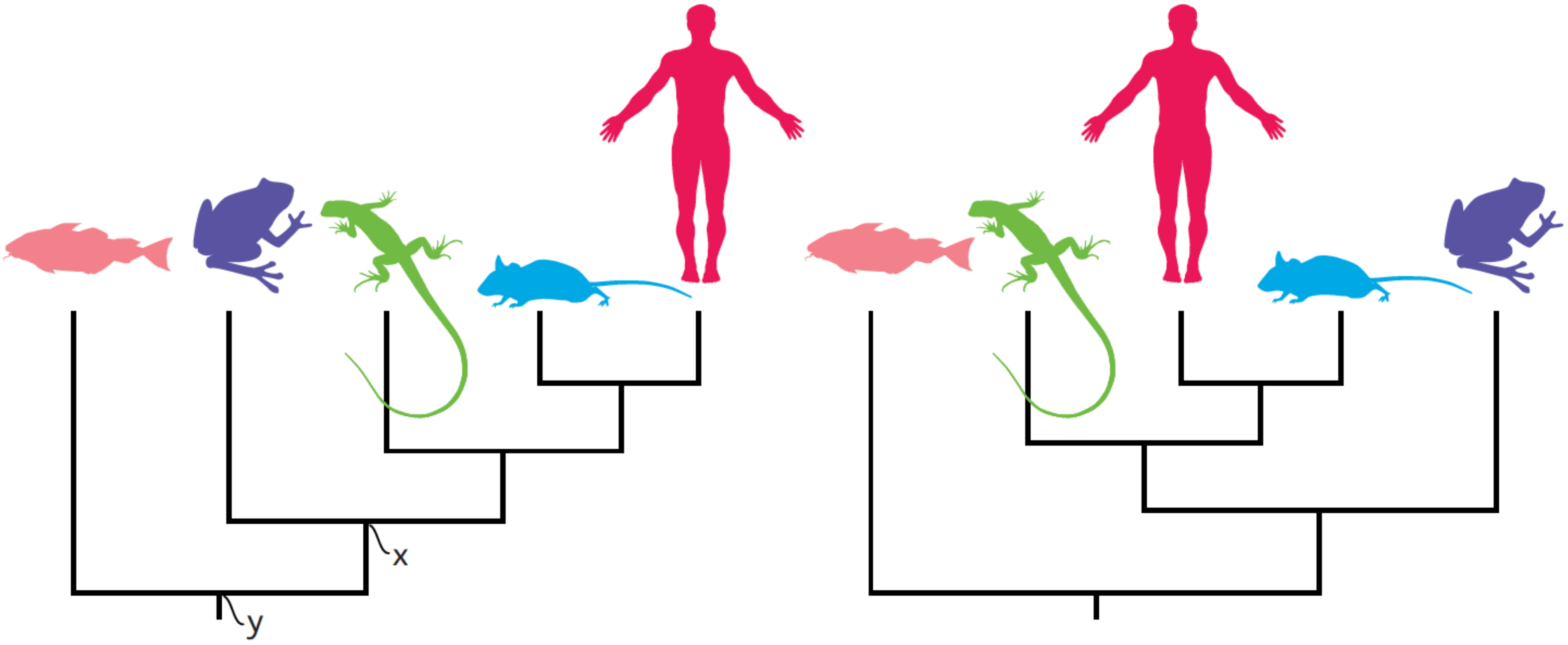


Table 2 | **A summary of strengths and weaknesses of different tree reconstruction methods**

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]

Which tree is correct?

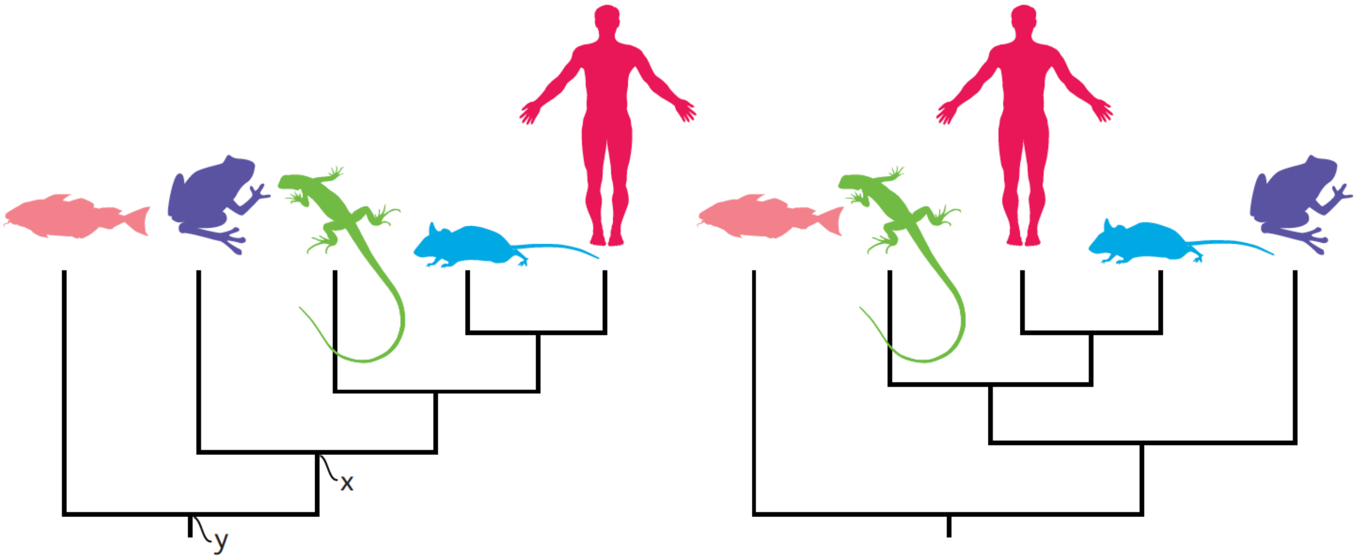
"just as beginning students in geography need to be taught how to read maps, so beginning students in biology should be taught how to read trees and to understand what trees communicate." - Robert O'Hara



The more recently species share a common ancestor, the more closely related they are.

Baum et al., 2005, Science. 310: 979-978

Interpretation of a tree



The correct way to read a tree is as a set of hierarchically nested groups, known as ***clades***.

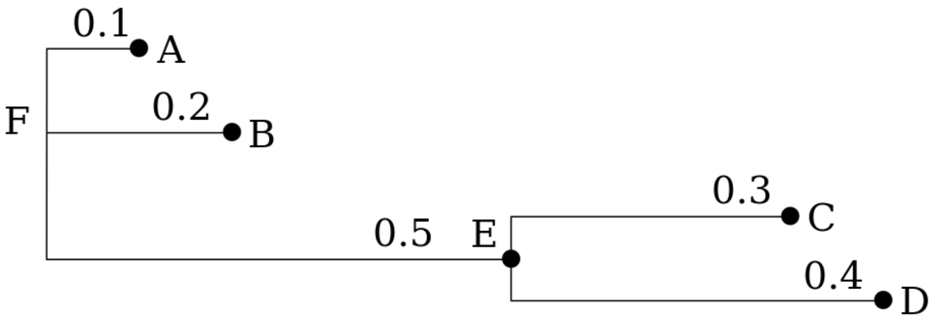
Flat file formats for phylogenetic trees

- **Newick**: Newick files are simply text files that consist of one or more tree descriptions in the Newick notation.

- **Nexus**: Nexus is widely used in phylogenetics and can contain trees in Newick notation and furthermore also information about taxa and phylogenetic data sets such as sequence alignments.

- **phyloXML**: phyloXML is a XML format for the analysis, exchange, and storage of phylogenetic trees (or networks) and associated data. It allows to store much more information about the tree nodes.

an example of the Newick format



Newick file:	Notes:
`(,,(,))`	no nodes are named
`(A,B,(C,D))`	leaf nodes are named
`(A,B,(C,D)E)F`	all nodes are named
`(:0.1,:0.2,(:0.3,:0.4):0.5)`	all but root node have a distance to parent
`(:0.1,:0.2,(:0.3,:0.4):0.5):0.0`	all have a distance to parent
`(A:0.1,B:0.2,(C:0.3,D:0.4):0.5)`	distances and leaf names (popular)
`(A:0.1,B:0.2,(C:0.3,D:0.4)E:0.5)F`	distances and all names
`((B:0.2,(C:0.3,D:0.4)E:0.5)F:0.1)A`	a tree rooted on a leaf node (rare)

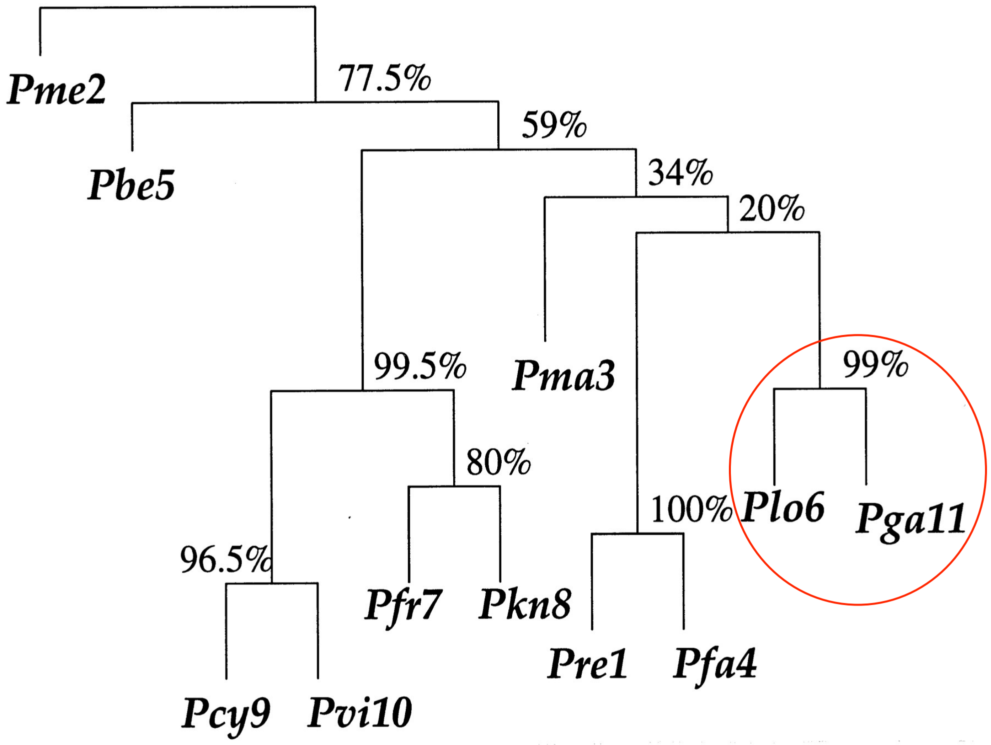
Outgroup rooting

- Many methods (NJ) construct unrooted tree. An outgroup can be introduced to identify the “root”. Although the inferred tree for all species is still unrooted, the root is believed to be located along the branch that leads to the outgroup so that the tree for the ingroup species is rooted. This strategy is called outgroup rooting.

- A good outgroup needs to satisfy:

1. not a member of the ingroup
2. close related to the ingroup

Tree evaluation: Bootstrap analysis



Boostrapping measures how consistently the data support given taxon bipartitions (Hedges, 1992).

Plo6 and Pga11 are grouped together in 99% bootstrap replicates.

* B = 200 bootstrap replications.

Bradley Efron et al. PNAS 1996;93:13429

An R package - ape

Analysis of Phylogenetics and Evolution ("ape") is an R software package for use in molecular evolution and phylogenetics.

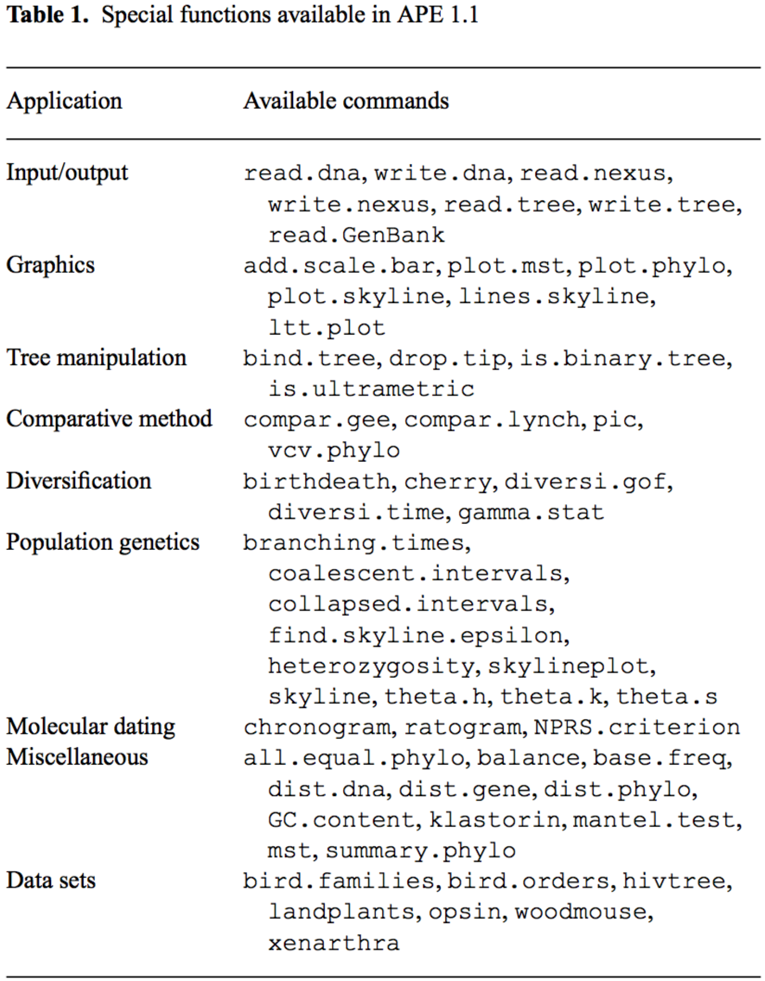


Table 1. Special functions available in APE 1.1

Application	Available commands
Input/output	read.dna, write.dna, read.nexus, write.nexus, read.tree, write.tree, read.GenBank
Graphics	add.scale.bar, plot.mst, plot.phylo, plot.skyline, lines.skyline, ltt.plot
Tree manipulation	bind.tree, drop.tip, is.binary.tree, is.ultrametric
Comparative method	compar.gee, compar.lynch, pic, vcv.phylo
Diversification	birthdeath, cherry, diversi.gof, diversi.time, gamma.stat
Population genetics	branching.times, coalescent.intervals, collapsed.intervals, find.skyline.epsilon, heterozygosity, skylineplot, skyline, theta.h, theta.k, theta.s
Molecular dating	chronogram, ratogram, NPRS.criterion
Miscellaneous	all.equal.phylo, balance, base.freq, dist.dna, dist.gene, dist.phylo, GC.content, klastorin, mantel.test, mst, summary.phylo
Data sets	bird.families, bird.orders, hivtree, landplants, opsin, woodmouse, xenarthra

citations

Yang *et al*., 2012, Molecular phylogenetics: principles and practice, Nature Reviews Genetics, 13: 303-314

Paradis *et al*., 2004, APE: Analyses of Phylogenetics and Evolution in R language, Bioinformatics, 20 (2): 289-290

Baum *et al*. The tree-thinking challenge, 2005, Science, 310:979-980