CSI5126. Algorithms in bioinformatics

Phylogeny

Marcel Turcotte

uOttawa

School of Electrical Engineering and Computer Science (EECS)
University of Ottawa

Version October 9, 2018

Summary

In this module, we introduce **molecular evolution** concepts. Specifically, we consider building **phylogentic trees**. The general framework is two-step: **large phylogeny problem** and **small phylogeny problem**. We consider the three main approaches: **distance-based**, **character-based**, and **maximum likelihood**.

General objective

- **Explain in your own words** the three main approaches to building phylogenetic trees, with sufficient details so that an actual implementation can be made.

Reading

- Bernhard Haubold and Thomas Wiehe (2006). *Introduction to computational biology: an evolutionary approach*. Birkhäuser Basel. Pages 143-168.

If evolution is true,
why are there still monkeys?

Larry King
(American television and radio host)

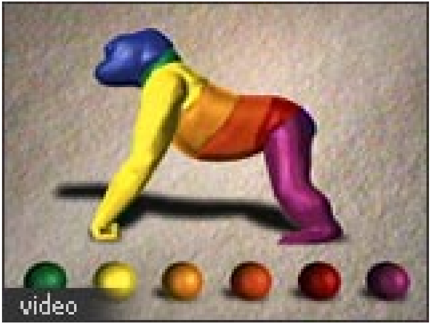
Wellcome - **Tree** of **Life**

`http://www.wellcometreeoflife.org/video/assets/TOL_6min_720p_download.mov`

Evolution - Great Transformations - PBS Nova

- https://www.youtube.com/watch?v=kfHxu8sk6bs (In particular, see the segment 43:13 – 46:47)
- http://www.pbs.org/wgbh/nova/evolution/

Evidences of evolution



Genetic Tool Kit

(www.pbs.org/wgbh/evolution/library/03/4/l_034_04.html)

Evidences of evolution

The Common Genetic Code

(www.pbs.org/wgbh/evolution/library/04/4/l_044_02.html)

(https://www.youtube.com/watch?v=urthr04mqoI)

(http://nsdl.oercommons.org/courses/the-common-genetic-code/view)

See also:

- **Evolution - What Darwin Never Knew - NOVA PBS Documentary**
  `https://www.youtube.com/watch?v=0V27qy6Gfb4`
- **Evolving Ideas: How Do We Know Evolution Happens?**
  (www.pbs.org/wgbh/evolution/library/11/2/e_s_3.html)
- **Evolution library**
  (www.pbs.org/wgbh/evolution/library)
- **Understanding Evolution - Misconceptions about evolution and the mechanisms of evolution - Berkeley**

Definitions

- “The objectives of phylogenetic studies are (1) to **reconstruct** the correct genealogical ties between organisms and (2) to **estimate the time of divergence** between organisms since they last shared a common ancestor.”

Definitions

- "The objectives of phylogenetic studies are (1) to **reconstruct** the correct genealogical ties between organisms and (2) to **estimate the time of divergence** between organisms since they last shared a common ancestor."
- "A phylogenetic tree is a **graph** composed of nodes and branches, in which only one branch connects any two adjacent nodes."

Definitions

- "The objectives of phylogenetic studies are (1) to **reconstruct** the correct genealogical ties between organisms and (2) to **estimate the time of divergence** between organisms since they last shared a common ancestor."
- "A phylogenetic tree is a **graph** composed of nodes and branches, in which only one branch connects any two adjacent nodes."
- "The nodes represents the **taxonomic units**, and the branches define the **relationships** among the units in terms of **descent and ancestry**."

Definitions

- “The objectives of phylogenetic studies are (1) to **reconstruct** the correct genealogical ties between organisms and (2) to **estimate the time of divergence** between organisms since they last shared a common ancestor.”
- “A phylogenetic tree is a **graph** composed of nodes and branches, in which only one branch connects any two adjacent nodes.”
- “The nodes represents the **taxonomic units**, and the branches define the **relationships** among the units in terms of **descent and ancestry**.”
- “The **branch length** usually represents the number of changes that have occurred in that branch.” (or some amount of time)

Definitions

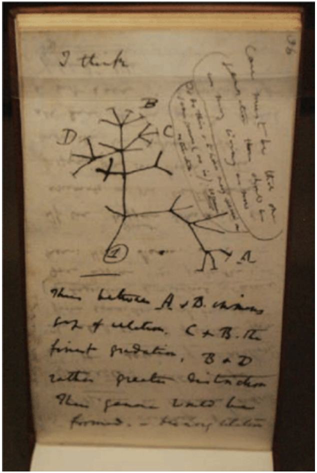
- "The objectives of phylogenetic studies are (1) to **reconstruct** the correct genealogical ties between organisms and (2) to **estimate the time of divergence** between organisms since they last shared a common ancestor."
- "A phylogenetic tree is a **graph** composed of nodes and branches, in which only one branch connects any two adjacent nodes."
- "The nodes represents the **taxonomic units**, and the branches define the **relationships** among the units in terms of **descent and ancestry**."
- "The **branch length** usually represents the number of changes that have occurred in that branch." (or some amount of time)

⇒ Li, W.-H. and Graur, D. (1991) *Fundamentals of Molecular Evolution*. Sinauer.

Terminology

- A **taxon** (plural **taxa**) or **taxonomic unit** is a species or grouping of species.
- Naming the different taxonomic levels: kingdom; phylum; class; order; family; genus; species.

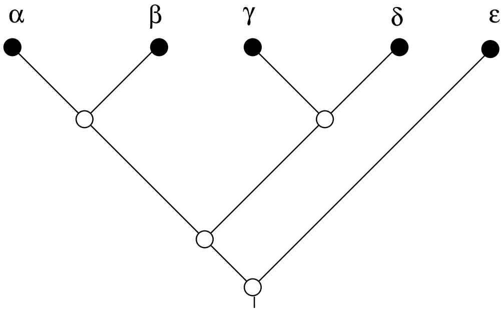
Evolution


I think
B
C
D
1
A
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

"Charles Darwin's famous notebook B containing the first known sketch of an **evolutionary tree**."

A. Rokas (2006) *Genomics and the Tree of Life.* *Science* **313**(5795): 1897–1899.
(DOI: 10.1126/science.1134490)

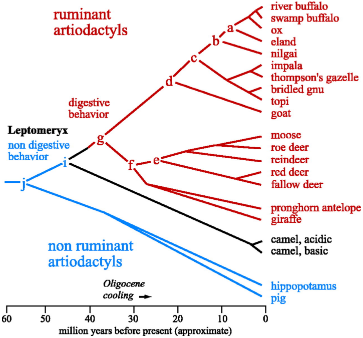
Terminology (cont.)



A **rooted tree** for 5 species. Leaves, $\alpha, \beta, \gamma, \delta$ and ϵ, correspond to **contemporary organisms**, for which data has been collected ($t = 0$). Internal nodes correspond to (inferred) **ancestors** ($t < 0$). **Newick format** of that tree: $(((\alpha, \beta), (\gamma, \delta)), \epsilon)$

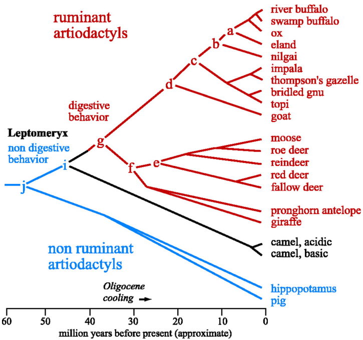
Why?

- **Comparative studies**: understanding gene function, adaptation, correlating the appearance of a trait to environmental factors, etc.;
- **Drug design**: designing compounds that are specific to a group of organisms;
- **Bioinformatics**: multiple sequence alignment, protein (secondary) structure prediction, etc.



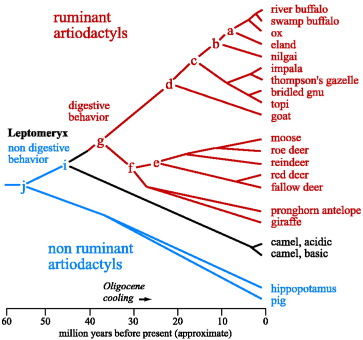
S.A. Benner (2002) *Science* **296**(5569): 864–868.

"One way of testing such hypotheses is to **resurrect the ancestral proteins** and study their behavior in the laboratory.



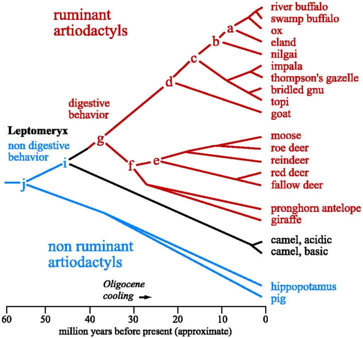
S.A. Benner (2002) *Science* **296**(5569): 864–868.

"One way of testing such hypotheses is to **resurrect the ancestral proteins** and study their behavior in the laboratory. To do this, a DNA molecule encoding **the ancestral protein is synthesized** and expressed in an appropriate host.



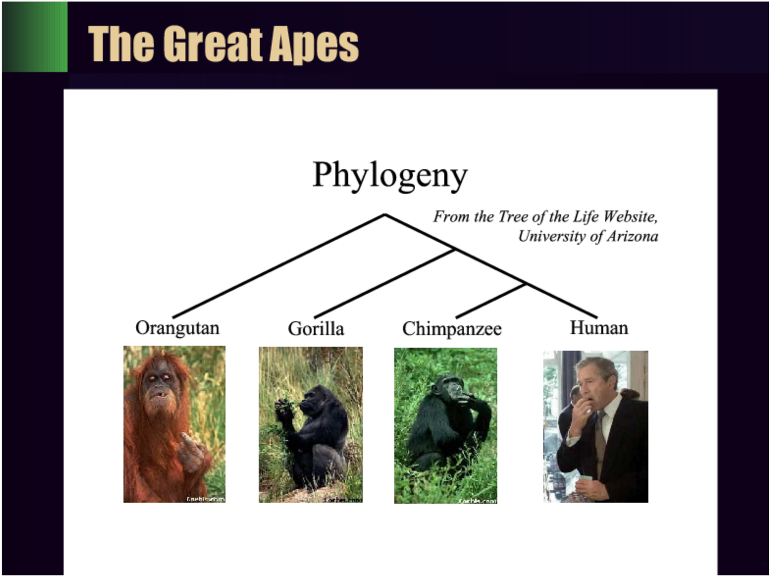
S.A. Benner (2002) *Science* **296**(5569): 864–868.

“One way of testing such hypotheses is to **resurrect the ancestral proteins** and study their behavior in the laboratory. To do this, a DNA molecule encoding **the ancestral protein is synthesized** and expressed in an appropriate host. The ancient protein is then recovered and studied to determine whether its properties are consistent with its inferred ancestral role.



S.A. Benner (2002) *Science* **296**(5569): 864–868.

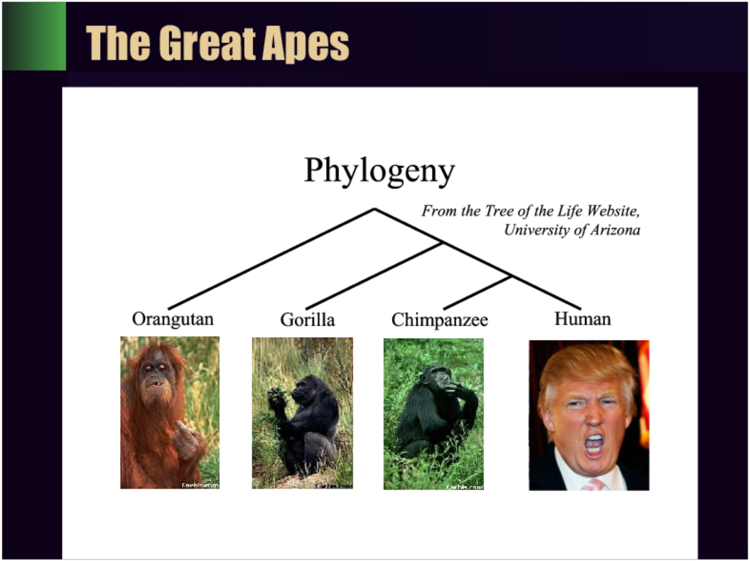
"One way of testing such hypotheses is to **resurrect the ancestral proteins** and study their behavior in the laboratory. To do this, a DNA molecule encoding **the ancestral protein is synthesized** and expressed in an appropriate host. The ancient protein is then recovered and studied to determine whether its properties are consistent with its inferred ancestral role. (...) **digestive ribonuclease emerged near the time when ruminant digestion emerged, in animals in which ruminant digestion developed, at a time where difficult-to-digest grasses emerged, permitting their descendants to exploit a newly available resource emerging at a time of global climatic upheaval.**"



The Great Apes

Phylogeny

From the Tree of the Life Website, University of Arizona



The Great Apes

Phylogeny

From the Tree of the Life Website,
University of Arizona

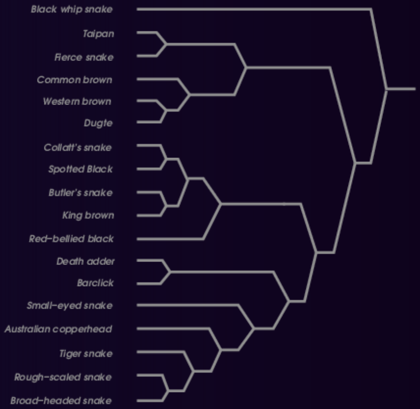
Orangutan

Gorilla

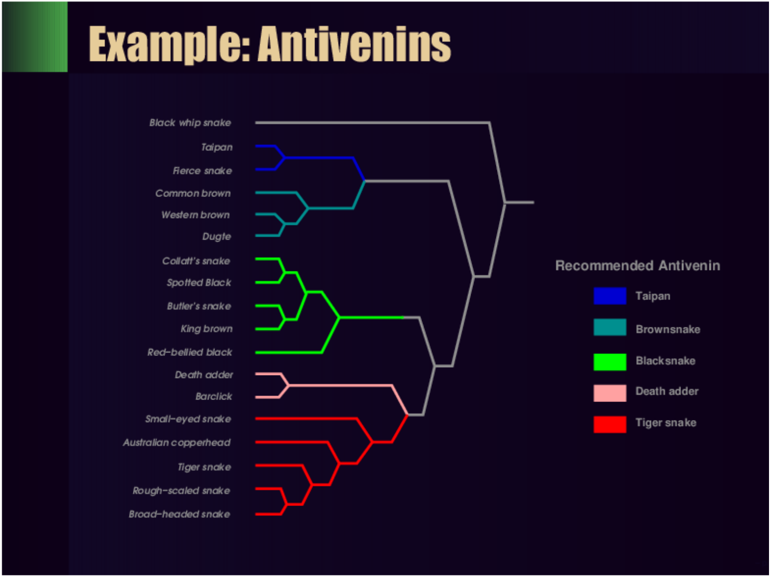
Chimpanzee

Human

Example: Antivenins



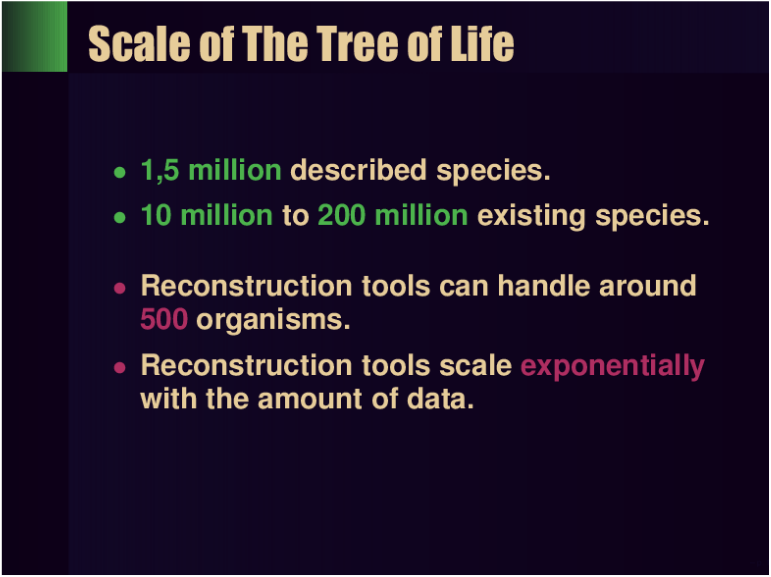
Venomous
Australian
Snakes



Example: Antivenins

Recommended Antivenin

- 🟦 Taipan
- 🟦 Brownsnake
- 🟩 Blacksnake
- 🟥 Death adder
- 🟥 Tiger snake



Scale of The Tree of Life

- 1,5 million described species.
- 10 million to 200 million existing species.

- Reconstruction tools can handle around 500 organisms.
- Reconstruction tools scale exponentially with the amount of data.

Credits for the last four slides

Bernard Moret

1980–2006: University of New Mexico (www.cs.unm.edu/∼moret)

2006–: École Polytechnique Fédérale de Lausanne (people.epfl.ch/bernard.moret)

See also: Tree of Life Web Project (www.tolweb.org).

Summary of the applications

- Study **ancestor-descendant** relationships
  (Evolutionary biology, adaption, genetic drift, selection, speciation, etc.)
- **Paleogenomics**: inferring ancestral genomic information from extinct species
  (Comparing Chimpanzee, Neanderthal and Human DNA)
- Origins of **epidemics**
  (Comparing, at the molecular level, various virus strains)
- **Drug design**: specifically targeting groups of organisms
  (Efficient enumeration of phylogenetically informative substrings)
- **Linguistics**
  (Languages tree divergence times)

Caveats

⇒ Felsenstein, J. (2004) *Inferring phylogenies*. Sinauer.

Caveats

⇒ Felsenstein, J. (2004) *Inferring phylogenies*. Sinauer.

- “Phylogenies, or evolutionary trees, are the basic structures necessary to think clearly about differences between species, and to analyze those differences statistically.”

Caveats

⇒ Felsenstein, J. (2004) *Inferring phylogenies*. Sinauer.

- "Phylogenies, or evolutionary trees, are the basic structures necessary to think clearly about differences between species, and to analyze those differences statistically."
- "I estimated that there are about **3,000 papers** on methods for inferring phylogenies."

Caveats

⇒ Felsenstein, J. (2004) *Inferring phylogenies*. Sinauer.

- “Phylogenies, or evolutionary trees, are the basic structures necessary to think clearly about differences between species, and to analyze those differences statistically.”
- “I estimated that there are about **3,000 papers** on methods for inferring phylogenies.”
- “The field of inferring phylogenies has been wracked by **outrageously excessive controversy** (...) there have been many biologists who strove to bring back the field to normality (...).”

Caveats

⇒ Felsenstein, J. (2004) *Inferring phylogenies*. Sinauer.

- "Phylogenies, or evolutionary trees, are the basic structures necessary to think clearly about differences between species, and to analyze those differences statistically."
- "I estimated that there are about **3,000 papers** on methods for inferring phylogenies."
- "The field of inferring phylogenies has been wracked by **outrageously excessive controversy** (...) there have been many biologists who strove to bring back the field to normality (...)."

Joe Felsenstein is the author of a software package called PHYLIP, which is one the most widely used software system for phylogenetic studies.

What's **the data**? (1/2)

	Characters					
Species	**1**	**2**	**3**	**4**	**5**	**6**
α	1	0	0	1	1	0
β	0	0	1	0	0	0
γ	1	1	0	0	0	0
δ	1	1	0	1	1	1
ϵ	0	0	1	1	1	0

Here, the 0s and 1s are indicating the **presence** or **absence** of a **character** (has feathers?, lays eggs?, curved beak?, flies?, ...).

A character is a measurable feature having well-defined mutually exclusive states.

What's **the data**? (1/2) (cont.)

- Based on anatomical and behavioural characters, the **panda** was classified as a **raccoon** (1870). Recently, 1985, the panda was re-classified as a **bear** when an analysis based on molecular data was done.

⇒ Images from www.wikipedia.org

What's **the data**? (1/2)

Platypus (*Ornithorhynchus anatinus*)

What's **the data**? (1/2)

Platypus (*Ornithorhynchus anatinus*) “the only **mammals that lay eggs** instead of giving birth to live young”,

What's **the data**? (1/2)

Platypus (*Ornithorhynchus anatinus*) “the only **mammals that lay eggs** instead of giving birth to live young”, “[i]t is the sole representative of its family (Ornithorhynchidae) and genus (Ornithorhynchus)”,

What's **the data**? (1/2)

Platypus (*Ornithorhynchus anatinus*) "the only **mammals that lay eggs** instead of giving birth to live young", "[i]t is the sole representative of its family (Ornithorhynchidae) and genus (Ornithorhynchus)", "[t]he platypus is considered to be one of the strangest specimens of the animal kingdom:

What's **the data**? (1/2)

Platypus (*Ornithorhynchus anatinus*) "the only **mammals that lay eggs** instead of giving birth to live young", "[i]t is the sole representative of its family (Ornithorhynchidae) and genus (Ornithorhynchus)", "[t]he platypus is considered to be one of the strangest specimens of the animal kingdom: a **venomous**,

What's **the data**? (1/2)

Platypus (*Ornithorhynchus anatinus*) "the only **mammals that lay eggs** instead of giving birth to live young", "[i]t is the sole representative of its family (Ornithorhynchidae) and genus (Ornithorhynchus)", "[t]he platypus is considered to be one of the strangest specimens of the animal kingdom: a **venomous**, **egg-laying**,

What's **the data**? (1/2)

Platypus (*Ornithorhynchus anatinus*) "the only **mammals that lay eggs** instead of giving birth to live young", "[i]t is the sole representative of its family (Ornithorhynchidae) and genus (Ornithorhynchus)", "[t]he platypus is considered to be one of the strangest specimens of the animal kingdom: a **venomous**, **egg-laying**, **duck-billed**

What's **the data?** (1/2)

Platypus (*Ornithorhynchus anatinus*) "the only **mammals that lay eggs** instead of giving birth to live young", "[i]t is the sole representative of its family (Ornithorhynchidae) and genus (Ornithorhynchus)", "[t]he platypus is considered to be one of the strangest specimens of the animal kingdom: a **venomous**, **egg-laying**, **duck-billed mammal**".

(commons.wikimedia.org/wiki/Image:Ornithorhynchidae-00.jpg)

Genome analysis of the platypus reveals unique signatures of evolution. Nature (2008) vol. 453 (7192) pp. 175-183

What's **the data**? (1/2)

- The **thylacine** (*Thylacinus cynocephalus*) is a now extinct (wolf-like) carnivorous marsupial.

(commons.wikimedia.org/wiki/Image:Thylacinus.jpg)

Hard to resolve relationships using **morphology** and **behaviour** alone

1. Similar characteristics can evolve independently in distantly related organisms — **convergent-evolution**;
2. It is often difficult to find characteristics that are **common** to all the organisms under study.

What's **the data?** (2/2)

	Characters					
Species	**1**	**2**	**3**	**4**	**5**	**6**
α	A	G	A	C	G	G
β	C	G	T	G	A	G
γ	A	C	A	G	A	G
δ	A	C	A	C	G	A
ϵ	C	G	T	C	G	G

Nowadays, biologists rely on molecular sequence data, in particular DNA or RNA sequences, which allows the comparison of a broader range of species. **What characters, other than molecular sequence data, would allow to compare *E. coli*, yeast, clam shell and human?**

Genes or species trees

- Herein, a **molecular sequence alignment** (DNA, RNA or proteins) is used as input. Each column (site) of this alignment represents a character.

Genes or **species** trees

- Herein, a **molecular sequence alignment** (DNA, RNA or proteins) is used as input. Each column (site) of this alignment represents a character.
- The **taxonomic units** (nodes of the tree) can represent genes, species or populations; not all at once obviously.

Genes or **species** trees

- Herein, a **molecular sequence alignment** (DNA, RNA or proteins) is used as input. Each column (site) of this alignment represents a character.
- The **taxonomic units** (nodes of the tree) can represent genes, species or populations; not all at once obviously.
- A **gene tree** represents the evolutionary history of a single gene; e.g. the evolution of the globin family (with its numerous gene duplication events).

Genes or **species** trees

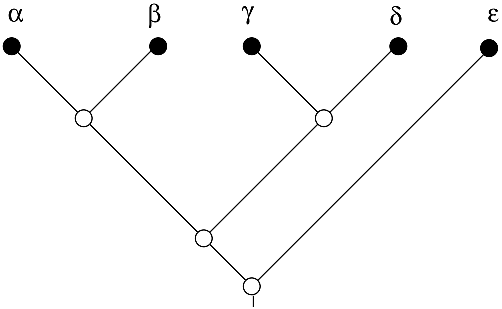
- Herein, a **molecular sequence alignment** (DNA, RNA or proteins) is used as input. Each column (site) of this alignment represents a character.
- The **taxonomic units** (nodes of the tree) can represent genes, species or populations; not all at once obviously.
- A **gene tree** represents the evolutionary history of a single gene; e.g. the evolution of the globin family (with its numerous gene duplication events).
- A **species tree** will generally be built using multiple genes (say 100).

Genes or species trees

- Herein, a **molecular sequence alignment** (DNA, RNA or proteins) is used as input. Each column (site) of this alignment represents a character.
- The **taxonomic units** (nodes of the tree) can represent genes, species or populations; not all at once obviously.
- A **gene tree** represents the evolutionary history of a single gene; e.g. the evolution of the globin family (with its numerous gene duplication events).
- A **species tree** will generally be built using multiple genes (say 100).
- Since our main interest is to study the methods, **we will limit our discussion to species trees**.

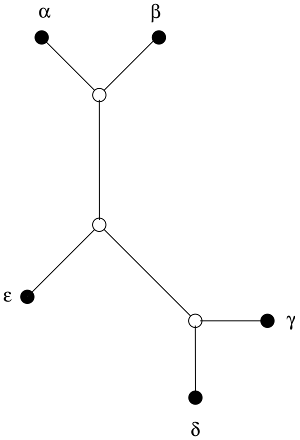
Rooted tree

A **rooted tree** not only gives the relationships between the taxonomic units, it also indicates the direction of evolution (time). Such trees can be scaled or unscaled.



Unrooted tree

An **unrooted tree** specifies the relationships between species.



Rooting the tree

Biologists generally **prefer rooted trees**.

- Under the **molecular clock assumption**, the root of the tree would be located at equal distance from all the leaves (contemporary organisms);
- The **outgroup method** consists of including into the analysis an organism that is known to have branched off earlier than the taxa under study (for which paleontological evidences exist, for instance), the root will be placed along the edge connecting the outgroup to the ancestor of the ingroup (taxa under study).

Molecular clock theory

- Proposed by Emile Zuckerkandl and Linus Pauling, 1962.

⇒ [5, pages 453–455]

Molecular clock theory

- Proposed by Emile Zuckerkandl and Linus Pauling, 1962.
- **Accepted mutations occur at a constant rate.**

⇒ [5, pages 453–455]

Molecular clock theory

- Proposed by Emile Zuckerkandl and Linus Pauling, 1962.
- **Accepted mutations occur at a constant rate.**
- The number of accepted mutations is proportional to the length of the time interval.

⇒ [5, pages 453–455]

Molecular clock theory

- Proposed by Emile Zuckerkandl and Linus Pauling, 1962.
- **Accepted mutations occur at a constant rate.**
- The number of accepted mutations is proportional to the length of the time interval.
- Once the "**clock**" has been calibrated (using fossil evidences, for instance) the unknown length of some time interval can be deduced from the number of accepted mutations.

⇒ [5, pages 453–455]

Molecular clock theory

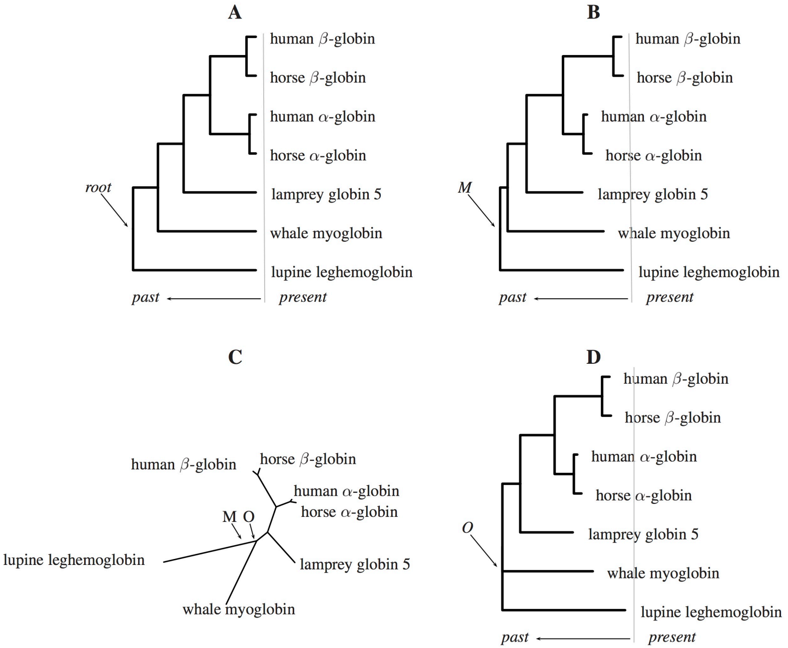
- Proposed by Emile Zuckerkandl and Linus Pauling, 1962.
- **Accepted mutations occur at a constant rate.**
- The number of accepted mutations is proportional to the length of the time interval.
- Once the "**clock**" has been calibrated (using fossil evidences, for instance) the unknown length of some time interval can be deduced from the number of accepted mutations.
- **Note**: different proteins have different clocks (hemoglobin ticks faster than cytochrome c).

⇒ [5, pages 453–455]

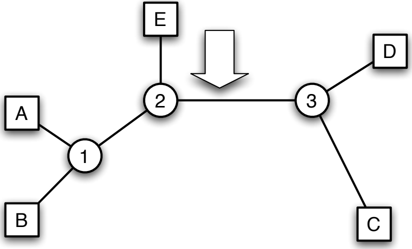
Molecular clock theory

- Proposed by Emile Zuckerkandl and Linus Pauling, 1962.
- **Accepted mutations occur at a constant rate.**
- The number of accepted mutations is proportional to the length of the time interval.
- Once the "**clock**" has been calibrated (using fossil evidences, for instance) the unknown length of some time interval can be deduced from the number of accepted mutations.
- **Note**: different proteins have different clocks (hemoglobin ticks faster than cytochrome c).
- "A great deal of ink (and blood) has been spilt over the molecular clock (...)"

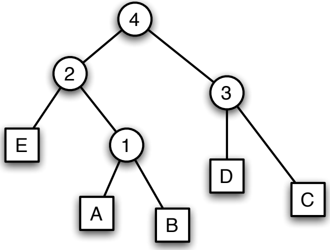
⇒ [5, pages 453–455]


A
human β-globin
horse β-globin
human α-globin
horse α-globin
lamprey globin 5
whale myoglobin
lupine leghemoglobin
root
past
present
B
human β-globin
horse β-globin
human α-globin
horse α-globin
lamprey globin 5
whale myoglobin
lupine leghemoglobin
M
past
present
C
human β-globin
horse β-globin
human α-globin
horse α-globin
M
O
lupine leghemoglobin
lamprey globin 5
whale myoglobin
D
human β-globin
horse β-globin
human α-globin
horse α-globin
lamprey globin 5
O
whale myoglobin
lupine leghemoglobin
past
present

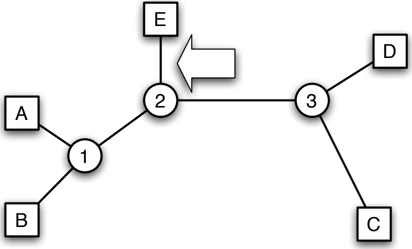
Rooting a tree: molecular clock



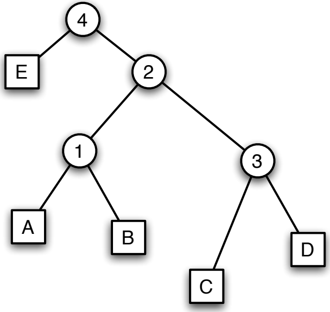
Rooting a tree: molecular clock (cont.)



Rooting a tree: outgroup

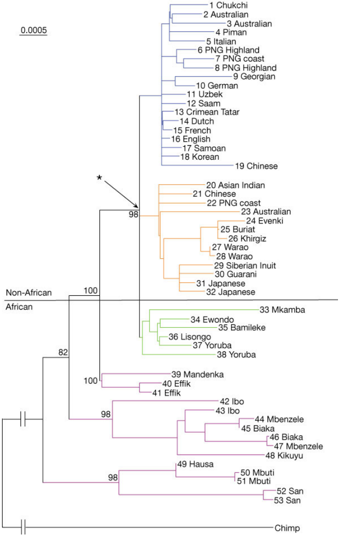


Rooting a tree: outgroup (cont.)



Hypothetically, E is the outgroup — e.g. chimpanzee while the ingroup consists of human populations.

Rooting a tree: outgroup (cont.)

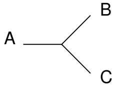


Neighbour-joining phylogram based on complete mtDNA genome sequences.
Source: Max Ingman, Henrik Kaessmann, Svante Pääbo and Ulf Gyllensten (2000) Nature 408, 708-713

General paradigm

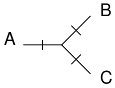
1. **Enumerate** trees;
2. **Select** the “best” tree.

Sequential addition strategy



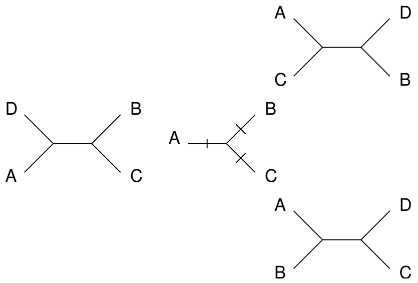
Given three species, there is a single unrooted tree.

Sequential addition strategy



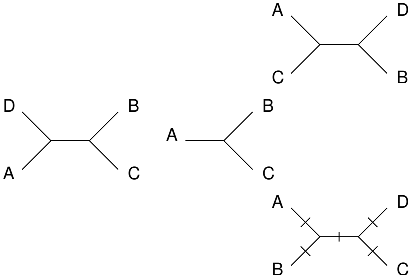
Each branch can serve as an insertion point, adding a new branch off the middle of any existing branch.

Sequential addition strategy



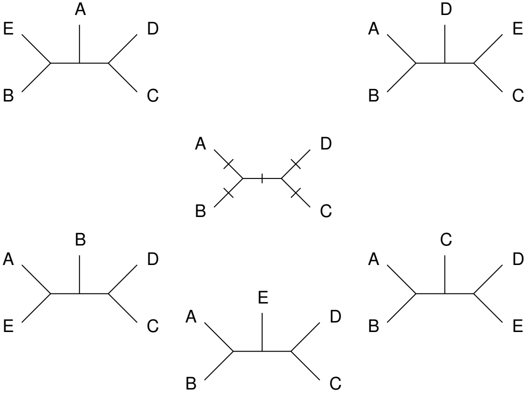
Therefore producing 3 four species unrooted trees.

Sequential addition strategy



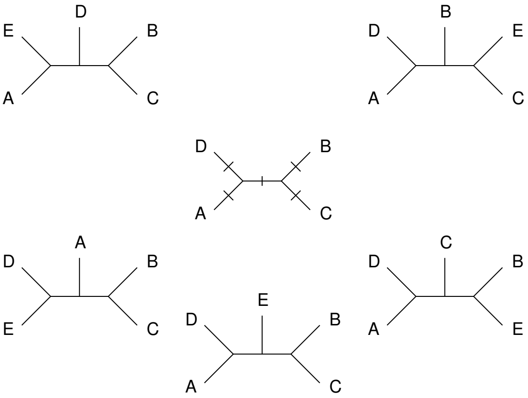
The same process is applied to all 3 four species trees.

Sequential addition strategy



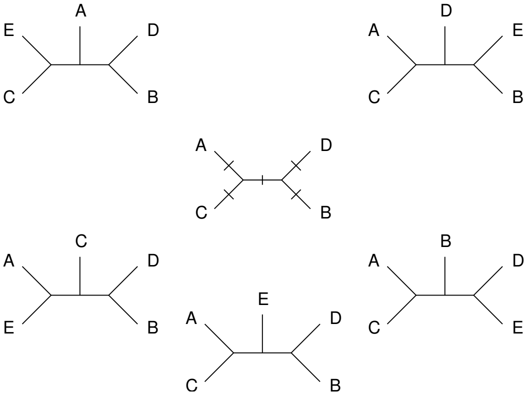
A four species unrooted tree has 5 edges, thus leading to 5 new unrooted trees.

Sequential addition strategy



There will be 15 five species unrooted trees.

Sequential addition strategy



Number of trees

# Species	# rooted trees	# unrooted trees
5	105	15
10	34,459,425	2,027,025
15	213,458,046,676,875	7,905,853,580,625
20	8,200,794,532,637,891,559,375	221,643,095,476,699,771,875

It can be shown that the number of rooted and unrooted trees for a given n (number of species) are as follows.

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

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

Methods

- **Distance-based**
  a distance is a measure of the overall differences/similarities between two objects
- **Character-based**
  a character is a characteristic that has well-defined, limited number of states
- **Maximum likelihood**
  Finds a tree such that the likelihood of the data given the tree structure is maximum

"While disputes between the champions of the two approaches [character-based and distance-based] have often been surprisingly intense, it is fair to say that both approaches are widely used and work well with most data sets." [2, page 85]

I. **Distance-based** reconstruction

- Let D_{ij} be the pairwise distance between two species; for instance, measured by comparing, in some ways, **sequence data** from the two species.

I. **Distance-based** reconstruction

- Let D_{ij} be the pairwise distance between two species; for instance, measured by comparing, in some ways, **sequence data** from the two species.
- Let d_{ij} be the distance between i and j in some **tree**; the sum of the length of all the branches on the unique path between i and j.

I. **Distance-based** reconstruction

- Let D_{ij} be the pairwise distance between two species; for instance, measured by comparing, in some ways, **sequence data** from the two species.
- Let d_{ij} be the distance between i and j in some **tree**; the sum of the length of all the branches on the unique path between i and j.
- Distance-based methods seek to find a tree (**topology** + **branch length**) such the D_{ij} and the d_{ij}, for all i and j, are in **good agreement**.

I. **Distance-based** reconstruction

- Let D_{ij} be the pairwise distance between two species; for instance, measured by comparing, in some ways, **sequence data** from the two species.
- Let d_{ij} be the distance between i and j in some **tree**; the sum of the length of all the branches on the unique path between i and j.
- Distance-based methods seek to find a tree (**topology + branch length**) such the D_{ij} and the d_{ij}, for all i and j, are in **good agreement**.
- For instance, find a tree minimizing

$$Q = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}(D_{ij} - d_{ij})^2 \qquad (1)$$

Minimum least squares approach.

I. **Distance-based** reconstruction

If the tree topology was given, the problem would simply be to find the length of the branches minimizing Equation 1, which is a quadratic function.

I. **Distance-based** reconstruction

If the tree topology was given, the problem would simply be to find the length of the branches minimizing Equation 1, which is a quadratic function.

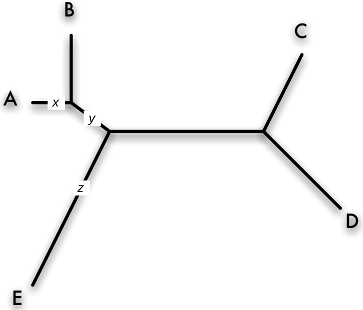
The d_{ij} need to be expressed as the sum of the length of the branches, x, on the unique path between i and j, $\mathcal{P}_{ij}$.

$$Q = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}(D_{ij} - \sum_{x \in \mathcal{P}_{ij}} x)^2 \tag{2}$$

and

$$d_{ij} = \sum_{x \in \mathcal{P}_{ij}} x$$

I. **Distance-based** reconstruction



$$d_{AE} = x + y + z$$

I. Distance-based reconstruction

This involves solving the set of linear equations obtained by taking the derivative of Q with respect to the branch lengths and equating those to 0.

$$\frac{dQ}{dx} = -2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(D_{ij} - \sum_{x \in \mathcal{P}_{ij}} x) = 0$$

I. **Distance-based** reconstruction

This involves solving the set of linear equations obtained by taking the derivative of Q with respect to the branch lengths and equating those to 0.

$$\frac{dQ}{dx} = -2 \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(D_{ij} - \sum_{x \in \mathcal{P}_{ij}} x) = 0$$

Methods for finding exact solutions for the weighted and unweighted least squares branch length equations have been proposed that run in polynomial time (order 3 or less), and iterative/heuristic methods have also been proposed, which in practice converge rapidly towards the the correct lengths.

I. Distance-based reconstruction

- What's next?

I. **Distance-based** reconstruction

- What's next?
- Well, we assumed that the topology was known!

I. **Distance-based** reconstruction

- What's next?
- Well, we assumed that the topology was known!
- In principle, it would be possible to enumerate all tree topologies, solve the branch length equations for each one, and select the topology that minimizes the least squares.

I. Distance-based reconstruction

- What's next?
- Well, we assumed that the topology was known!
- In principle, it would be possible to enumerate all tree topologies, solve the branch length equations for each one, and select the topology that minimizes the least squares.
- This is not practical since the number of topologies grows rapidly w.r.t. the number of species.
  [The same approaches as for parsimony methods for enumerating tree topologies can be applied. However, under certain assumptions, a simple algorithm, UPGMA, performs adequately.]

I. **Distance-based** reconstruction

Given n species and m characters.

1. Compute a distance matrix between all ($\Theta(n^2)$) pairs, this matrix is symmetrical;
2. Generate the topology of the tree;
3. Estimate the length of the branches.

Distance-based reconstruction

- For now, let's assume a simple distance measure, e.g. **Hamming distance** or **fractional alignment difference** ($\frac{D}{L}$) where D is the number of sites that differ (excluding indel containing sites), L is the number of sites
- More realistic models will be presented together with the Maximum Likelihood approach

Distance-based reconstruction

```
a GTGCTGCACGG CTCAGTATA GCATTTACCC TTCCATCTTC AGATCCTGAA
b ACGCTGCACGG CTCAGTGCG GTGCTTACCC TCCCATCTTC AGATCCTGAA
g GTGCTGCACGG CTCGGCGCA GCATTTACCC TCCCATCTTC AGATCCTATC
d GTATCACACGA CTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCCTAAA
e GTATCACATAG CTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCTAAAA
```

Build a multiple sequence alignment and **select columns**.

⇒ [2, page 87]

Distance-based reconstruction

Species	α	β	γ	δ	ϵ
α	0	9	8	12	15
β	9	0	11	15	18
γ	8	11	0	10	13
δ	12	15	10	0	5
ϵ	15	18	13	5	0

The above matrix has been filled by computing the Hamming distance for all pairs of sequences.

⇒ [2, page 87]

UPGMA

Unweighted Pair Group Method using Arithmetic averages[*].

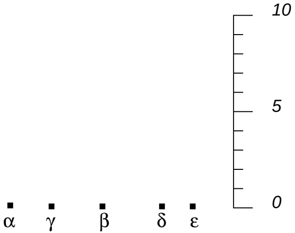
```
{ Initialization }
Assign each species $i$ to its own cluster $C_i$
Define one leaf of $T$ for each species, place it at height zero.

{ Iterations }
Find the pair of clusters $i$ and $j$ which minimises $d_{ij}$.
Define a new cluster $C_k = C_i \bigcup C_j$.
Calculate $d_{kl}$ for all $l$.
Create the parent node $k$ of $i$ and $j$ at height $d_{ij}/2$ in $T$.
Add $k$ to the current list of clusters and remove $i$ and $j$.

{ Termination }
Stop when the list of clusters contains only one entry.
```

⇒ Early 1960s, simple and intuitive.

[*] See [6, page 166]

Distance-based reconstruction

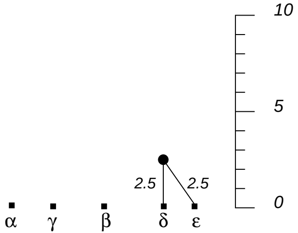
Species	α	β	γ	δ	ϵ
α	0	9	8	12	15
β	9	0	11	15	18
γ	8	11	0	10	13
δ	12	15	10	0	5
ϵ	15	18	13	5	0



⇒ Assign each species to its own cluster, place it at height 0.

Distance-based reconstruction

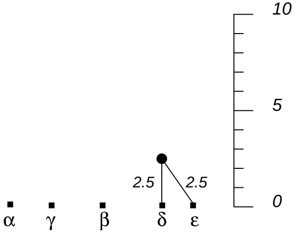
Species	α	β	γ	δ	ϵ
α	0	9	8	12	15
β	9	0	11	15	18
γ	8	11	0	10	13
δ	12	15	10	0	5
ϵ	15	18	13	5	0



⇒ Find the pair of clusters which minimises $d_{i,j}$, define a new cluster $C_k = C_i \bigcup C_j$, create the parent node k at height $d_{i,j}/2$.

Distance-based reconstruction

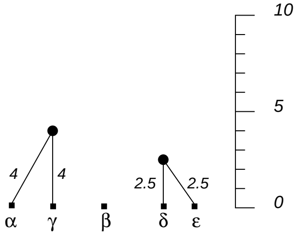
Species	α	β	γ	(δ, ϵ)
α	0	9	8	13.5
β	9	0	11	16.5
γ	8	11	0	11.5
(δ, ϵ)	13.5	16.5	11.5	0



$\Rightarrow$ Calculate $d_{k,l}$ for all l, where $d_{i,j} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}$.

Distance-based reconstruction

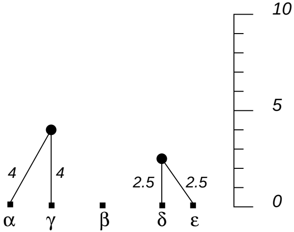
Species	α	β	γ	(δ, ϵ)
α	0	9	8	13.5
β	9	0	11	16.5
γ	8	11	0	11.5
(δ, ϵ)	13.5	16.5	11.5	0



$\Rightarrow$ Find the pair of clusters which minimises $d_{i,j}$, define a new cluster $C_k = C_i \bigcup C_j$, create the parent node k at height $d_{i,j}/2$.

Distance-based reconstruction

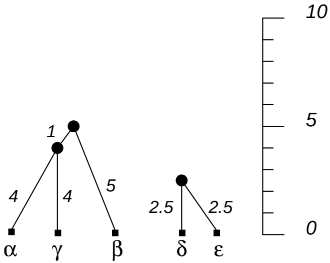
Species	(α, γ)	β	(δ, ϵ)
(α, γ)	0	10	12.5
β	10	0	16.5
(δ, ϵ)	12.5	16.5	0



$\Rightarrow$ Calculate $d_{k,l}$ for all l, where $d_{i,j} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}$.

Distance-based reconstruction

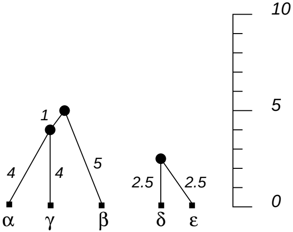
Species	(α, γ)	β	(δ, ϵ)
(α, γ)	0	10	12.5
β	10	0	16.5
(δ, ϵ)	12.5	16.5	0



⇒ Find the pair of clusters which minimises $d_{i,j}$, define a new cluster $C_k = C_i \bigcup C_j$, create the parent node k at height $d_{i,j}/2$.

Distance-based reconstruction

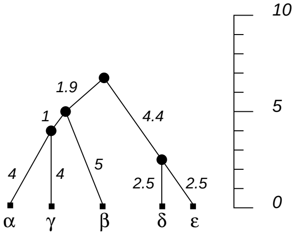
Species	$((\alpha,\gamma),\beta)$	(δ,ϵ)
$((\alpha,\gamma),\beta)$	0	$13.8\overline{3}$
(δ,ϵ)	$13.8\overline{3}$	0



$\Rightarrow$ Calculate $d_{k,l}$ for all l, where $d_{i,j} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}$.

Distance-based reconstruction

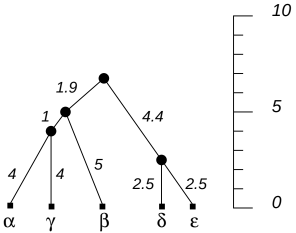
Species	$((\alpha,\gamma),\beta)$	(δ,ϵ)
$((\alpha,\gamma),\beta)$	0	$13.8\overline{3}$
(δ,ϵ)	$13.8\overline{3}$	0



$\Rightarrow$ Find the pair of clusters which minimises $d_{i,j}$, define a new cluster $C_k = C_i \bigcup C_j$, create the parent node k at height $d_{i,j}/2$.

Distance-based reconstruction

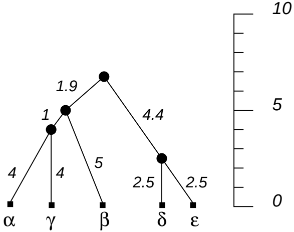
Species	$(((\alpha, \gamma), \beta), (\delta, \epsilon))$
$(((\alpha, \gamma), \beta), (\delta, \epsilon))$	0



⇒ UPGMA produces **ultrametric** trees, a tree such that the distance from any internal node (including the root) to its descendant leaves is the same. Thus, UPGMA assumes that evolution proceeds at the same rate in all the lineages, this is called the **molecular clock hypothesis**.
(An assumption that is often violated)

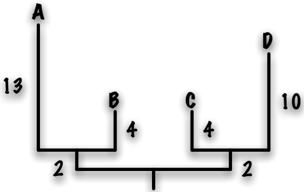
Distance-based reconstruction

Species	α	β	γ	δ	ϵ
α	0	9	8	12	15
β	9	0	11	15	18
γ	8	11	0	10	13
δ	12	15	10	0	5
ϵ	15	18	13	5	0



⇒ Consider $d_T(\alpha, \beta)$ and $d_{\alpha,\beta}$. Is $d_T(i, j) = d_{i,j}$ for all i, j?

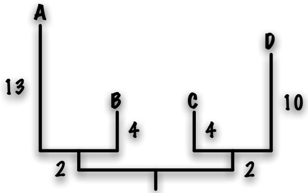
UPGMA — **failure of the molecular clock hypothesis**



“True tree”

Felsenstein 2004, page 167

UPGMA — failure of the molecular clock hypothesis



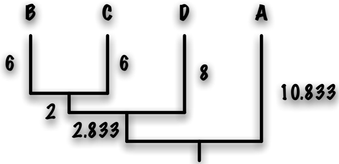
$$\Rightarrow \begin{matrix} 0 & 17 & 21 & 27 \\ 17 & 0 & 12 & 18 \\ 21 & 12 & 0 & 14 \\ 27 & 18 & 14 & 0 \end{matrix}$$

“True tree”

Felsenstein 2004, page 167

UPGMA — failure of the molecular clock hypothesis

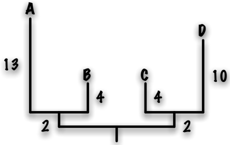
$$\begin{matrix} 0 & 17 & 21 & 27 \\ 17 & 0 & 12 & 18 \\ 21 & 12 & 0 & 14 \\ 27 & 18 & 14 & 0 \end{matrix} \Rightarrow$$



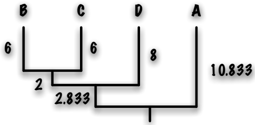
Reconstructed tree

Felsenstein 2004, page 167

UPGMA — failure of the molecular clock hypothesis



$$\Rightarrow \begin{matrix} 0 & 17 & 21 & 27 \\ 17 & 0 & 12 & 18 \\ 21 & 12 & 0 & 14 \\ 27 & 18 & 14 & 0 \end{matrix} \Rightarrow$$



UPGMA joins B and C together since both are evolving slowly (short branch attraction).

Distance-based tree reconstruction

- Given a tree T, a distance matrix $d_{i,j}$ is *additive* if $d_T(i,j) = d_{i,j}$, and *nonadditive* otherwise;
- UPGMA has no guarantee to produce "additive" trees;
- UPGMA produces *ultrametric* trees, i.e. a tree such that the distance from the root to any leaf is the same;
- UPGMA assumes that evolution proceeds at a fixed constant rate, the so called molecular clock hypothesis;
- Other distance-based methods, such as the *neighbour-joining* algorithm, do not assume the existence of a molecular clock; good approximation of the least squares methods, fast (works well with hundreds of species).

Distance **matrix**

- The distance matrix needs to be a **metric**.
  - **Symmetric:** $D_{i,j} = D_{j,i}$ and $D_{i,i} = 0$;
  - **Triangle inequality:** $D_{i,j} + D_{j,k} \geq D_{i,k}$.

Distance **matrix**

- The distance matrix is **additive** iff there exist a phylogeny such that:
  - $d_{i,j} \geq 0$;
  - $\forall i, j, D_{i,j} = d_{i,j}$.

Here $d_{i,j}$ is the sum of all the weigths along the path from i to j in the tree $\mathbf{T}$.

Additive Tree Reconstruction

If **D** is an **additive** matrix, the corresponding additive tree **T** is unique and it can be reconstructed in $\mathcal{O}(n^2)$, where is n is the number of species.

- Wing-Kin Sung (2010) Algorithms in Bioinformatics: A Practical Introduction. Chapman & Hall/CRC. Chapter 7. QH 324.2 .S86 2010

Neighbour-Joining algorithm

```
1. Let $Z = \{\{1\}, \{2\}, \ldots, \{n\}\}$ be the initial set of clusters;
2. For all $\{i\}, \{j\} \in Z$, set $d(\{i\}, \{j\}) = D_{i,j}$;
3. for $i = 2$ to $n$ do
   3.1 For all cluster $A \in Z$, set $u_A = \frac{1}{n-2} \sum_{D \in Z} d(D, A)$;
   3.2 Find $A, B \in Z$ minimizing $d(A, B) - u_a - u_B$;
   3.3 Let $C$ be the new cluster formed by connecting $A$ and $B$
       to a new root $r$. Let $(r, r_A)$ be $\frac{1}{2}d(A, B) + \frac{1}{2}(u_A - u_B)$
       and $(r, r_B)$ be $\frac{1}{2}d(A, B) + \frac{1}{2}(u_B - u_A)$;
   3.4 Set $Z = Z \bigcup \{C\} \setminus \{A, B\}$;
   3.5 For all $D = Z \setminus \{C\}$, let
       $d(D, C) = d(C, D) = \frac{1}{2}(d(A, D) + d(B, D) - d(A, B))$;
4. end for
```

- Wing-Kin Sung (2010), Page 189.

Summary

- Distance-based methods are taking as **input** a **multiple sequence alignment** (n sequences by m columns). A **distance** is calculated for each pairwise alignment in order to fill an $n \times n$ (distance) matrix. The **distance matrix** is used to **infer the topology** of a tree, as well as the length of its branches.

Summary

- Distance-based methods are taking as **input** a **multiple sequence alignment** (n sequences by m columns). A **distance** is calculated for each pairwise alignment in order to fill an $n \times n$ (distance) matrix. The **distance matrix** is used to **infer the topology** of a tree, as well as the length of its branches.
- **Neighbour-joining** is the most widely used distance-based approach. It produces un-rooted, additive (but not necessarily ultrametric) trees. It is an iterative algorithm that requires $\mathcal{O}(n^3)$ time.

Summary

- Distance-based methods are taking as **input** a **multiple sequence alignment** (n sequences by m columns). A **distance** is calculated for each pairwise alignment in order to fill an $n \times n$ (distance) matrix. The **distance matrix** is used to **infer the topology** of a tree, as well as the length of its branches.
- **Neighbour-joining** is the most widely used distance-based approach. It produces un-rooted, additive (but not necessarily ultrametric) trees. It is an iterative algorithm that requires $\mathcal{O}(n^3)$ time.
- **Distance-based methods are fast!**

Remarks

- "However, the real problems arise when the data are not additive. Then, one has to find a tree whose distances *best approximate* the given data."

Remarks

- "However, the real problems arise when the data are not additive. Then, one has to find a tree whose distances *best approximate* the given data."
- "(...) the problem [distance-based tree reconstruction] remains open as there is no approach that both leads to a provably efficient algorithm and that follows a completely accepted definition of a *good approximation*." [5, page 448]

Remarks

- “However, the real problems arise when the data are not additive. Then, one has to find a tree whose distances *best approximate* the given data.”
- “(...) the problem [distance-based tree reconstruction] remains open as there is no approach that both leads to a provably efficient algorithm and that follows a completely accepted definition of a *good approximation*.” [5, page 448]
- **Information is lost when reducing a pairwise alignment to a single number!** In particular, the reconstructed tree says nothing about the “characters” of the internal nodes (ancestors) and the evolutionary events that occurred along its branches.

Remarks

- "However, the real problems arise when the data are not additive. Then, one has to find a tree whose distances *best approximate* the given data."
- "(...) the problem [distance-based tree reconstruction] remains open as there is no approach that both leads to a provably efficient algorithm and that follows a completely accepted definition of a *good approximation*." [5, page 448]
- **Information is lost when reducing a pairwise alignment to a single number!** In particular, the reconstructed tree says nothing about the "characters" of the internal nodes (ancestors) and the evolutionary events that occurred along its branches.
- Hence the need for **alternatives**.

References

W.-H. Li and D. Graur.
Fundamentals of Molecular Evolution.
Sinauer, 1991.

D. E. Krane and M. L. Raymer.
Fundamental Concepts of Bioinformatics.
Benjamin Cummings, 2003.

J. Felsenstein.
Inferring Phylogenies.
Sinauer, 2004.

N. C. Jones and P. A. Pevzner.
An introduction to bioinformatics algorithm.
MIT Press, 2004.

References (cont.)

- D. Gusfield.
  Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.
  Cambridge University Press, 1997.

- R. Durbin, S. Eddy, A. Krogh, and G. Mitchison.
  Biological Sequence Analysis.
  Cambridge University Press, 1998.

- Radu Mihaescu, Dan Levy, and Lior Pachter.
  Why Neighbor-Joining Works.
  Algorithmica, 54(1):1–24, May 2009.

Pensez-y!

L’impression de ces notes n’est probablement pas nécessaire!