Bioinformatics 1

Phylogenetic inference: from sequences to trees

Claudia Acquisti

Evolutionary Functional Genomics
Institute for Evolution and Biodiversity, WWU Münster
claudia.acquisti@uni-muenster.de

Computer Lab B, Schlossplatz 2b

Phylogenetic inference [27.11.12-30.12.12]

Stefanie Henze: st.henze@uni-muenster.de

Parijat Tripathi: parijat24@gmail.com

"Where Do We Come From? What Are We? Where Are We Going?" Paul Gauguin, 189

Long-standing questions.....

Where do species come from?

How do they change over time?

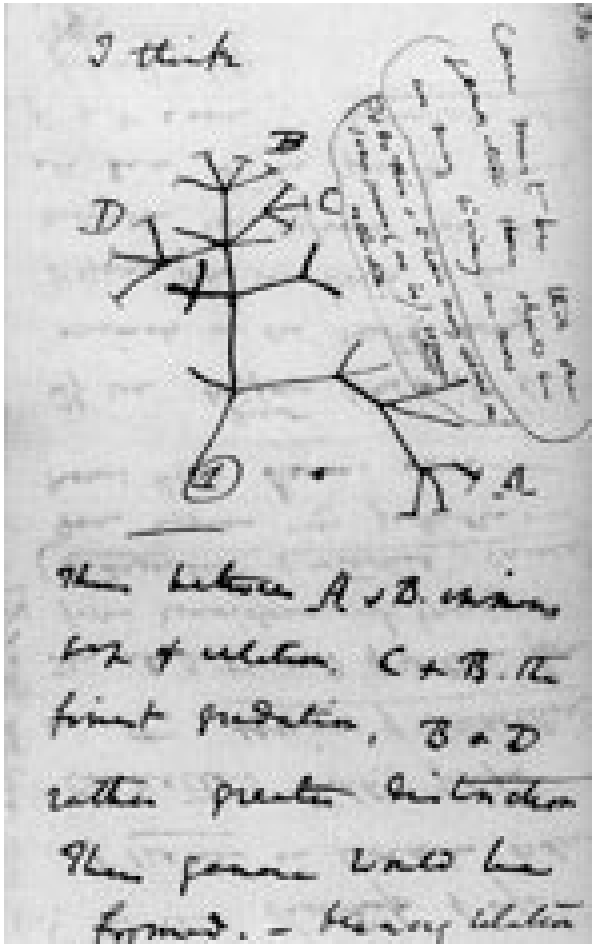
How are their evolutionary histories linked?

Tree of life: an ancient metaphor

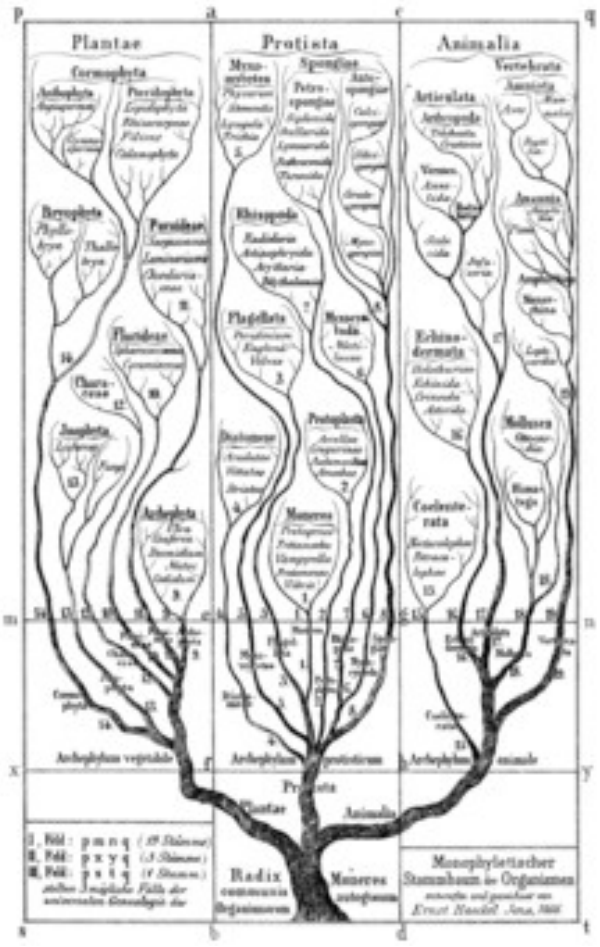
Urartu Helmet 1300 BC

Gustav Klimt, Tree of life 1911

First attempts of phylogenetic inference

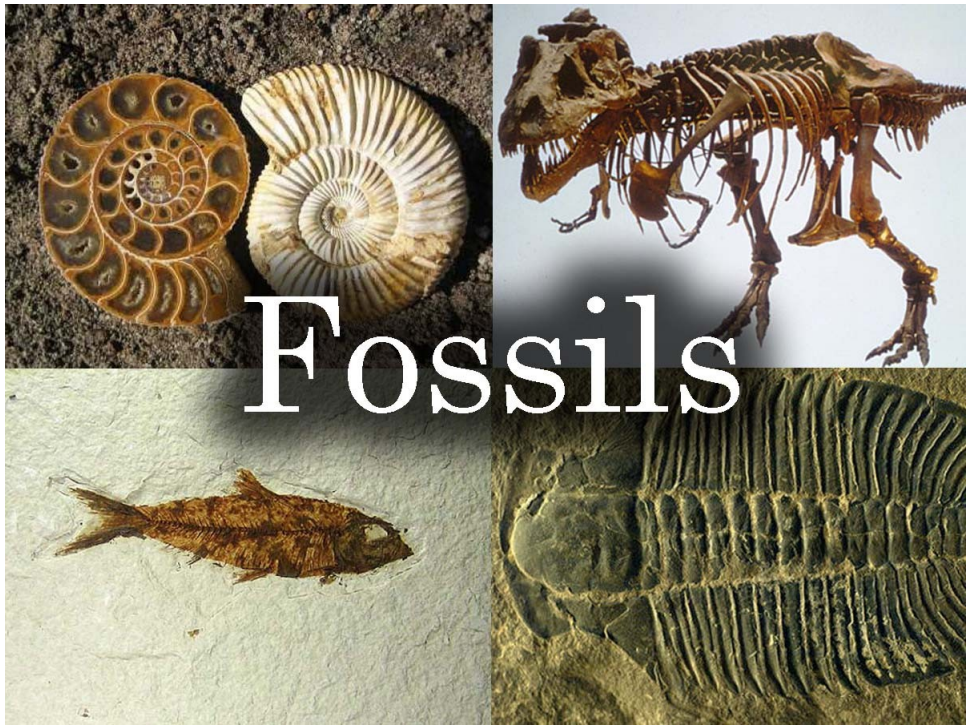


C. Darwin 1837



E. Haeckel 1866

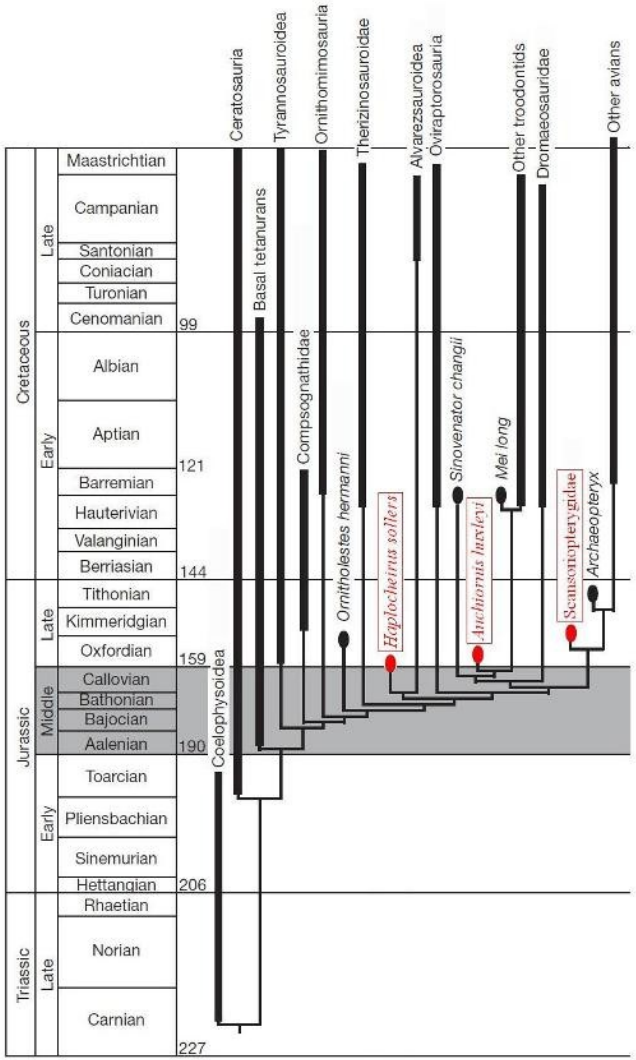
Tackling the problem the classical way



Fragmentary
Incomplete

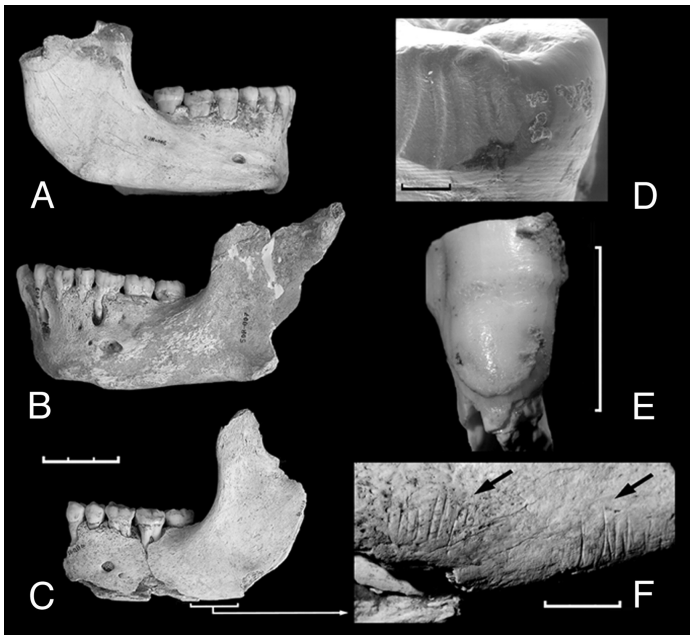
Success stories: transitional fossils

Archaeopteryx
1880, Berlin specimen

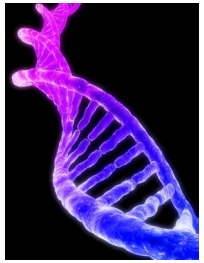


Tackling the problem the classical way

Comparative Morphology and Physiology

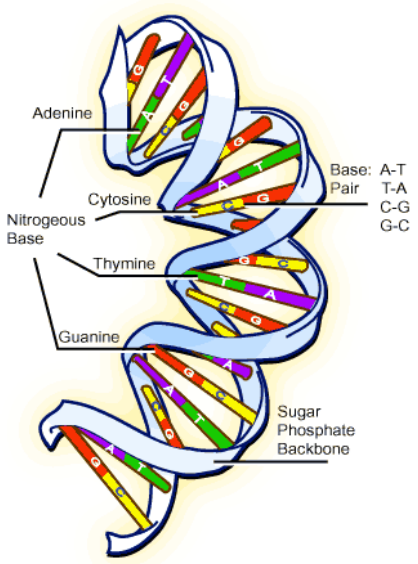


High complexity
Big controversies



The molecular approach

Comparison of the blueprint of life between species



1. (Almost) all species use DNA as genetic material

2. Mathematical modeling of sequence change

3. Large amount of phylogenetic information

DNA damage

Copying errors

MUTATIONS: heritable changes to the genome, essential for evolution.

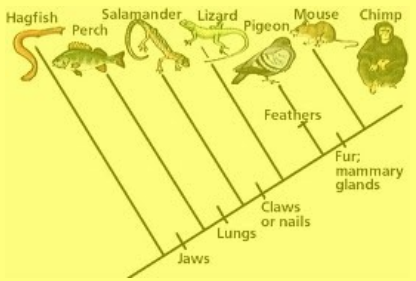
LARGE SCALE → SMALL SCALE

Chromosomal rearrangements

Point mutations

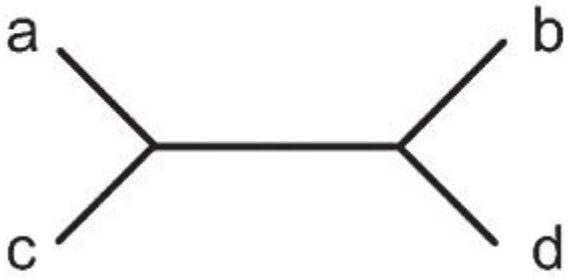
POLYMORPHISMS

?

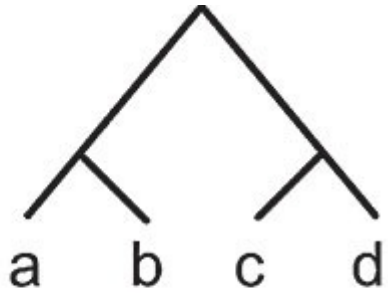


SUBSTITUTIONS

Tree topologies



Unrooted



Rooted

How many trees can we build?

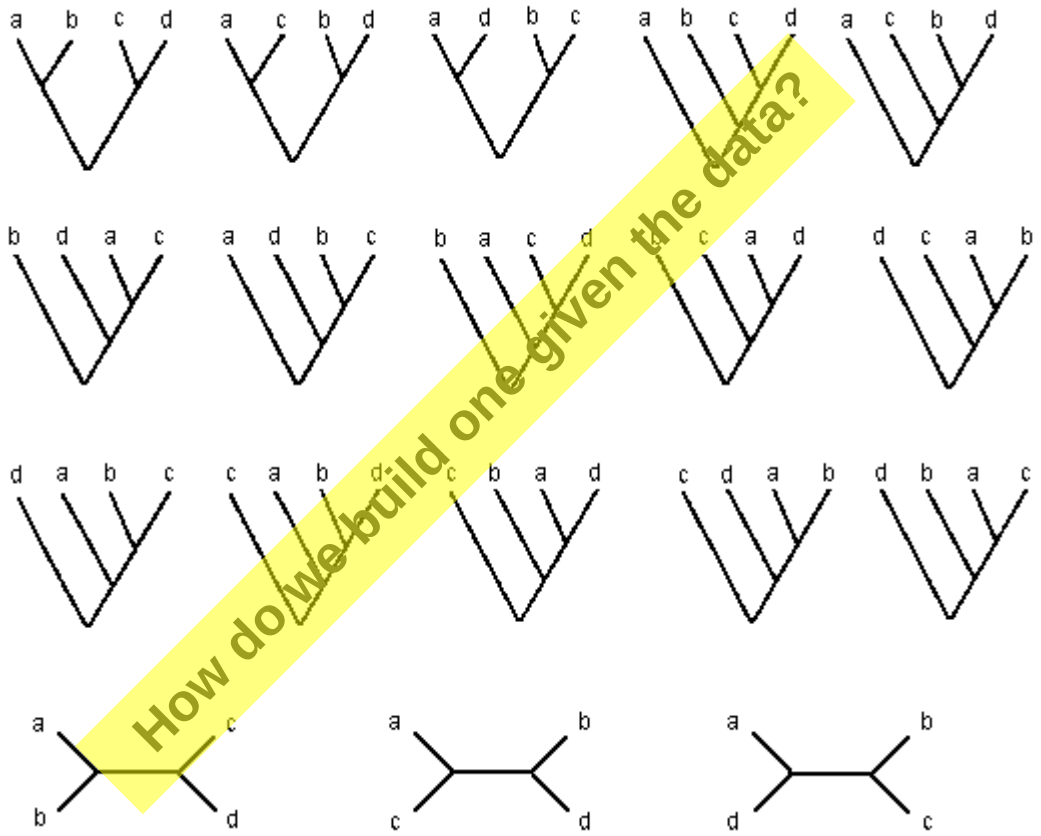
It depends on the number (m) of taxa

ROOTED TREES	UNROOTED TREES
$N(m) = [(2m-3)!] / [2^{m-2}(m-2)!]$	$N(m) = [(2m-5)!] / [2^{m-3}(m-3)!]$
N(4)= 15 N(10) = 34,459, 425	N(4)= 3 N(10) = 2,027, 025

and the number of possible trees grows very quickly with the number of taxa!

FINDING THE TRUE TREE IS A DIFFICULT PROBLEM ALREADY FOR FEW SPECIES

Tree topologies

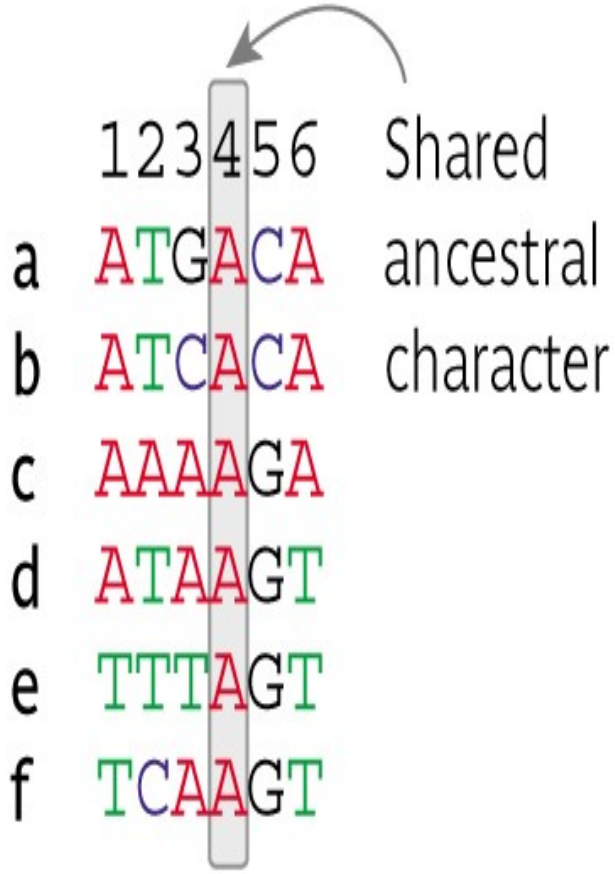


How do we build one given the data?

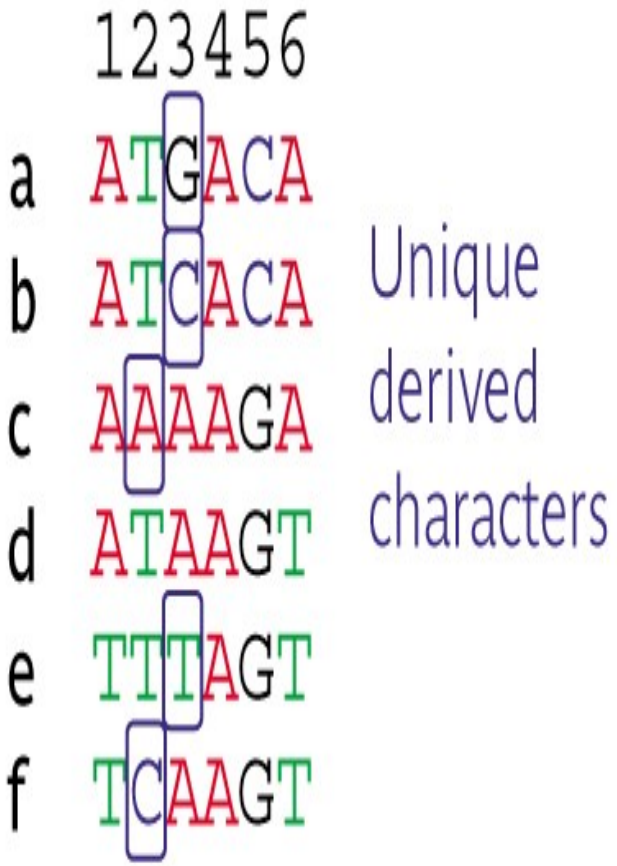
Alignment: a key step to recover evolutionary history from DNA sequences

```
  123456
a ATGACA
b ATCACA
c AAAAGA
d ATAAGT
e TTTAGT
f TCAAGT
```

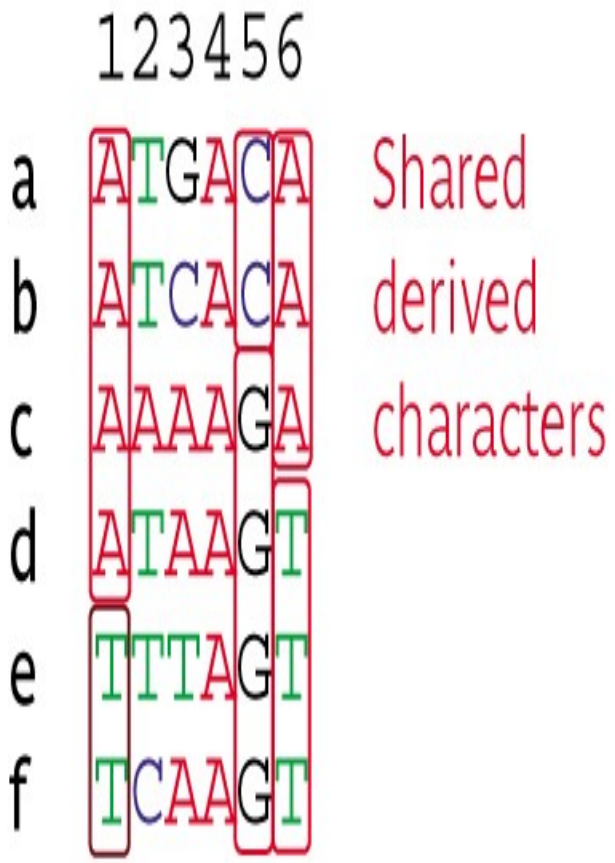
Alignment: a key step to recover evolutionary history from DNA sequences



Alignment: a key step to recover evolutionary history from DNA sequences

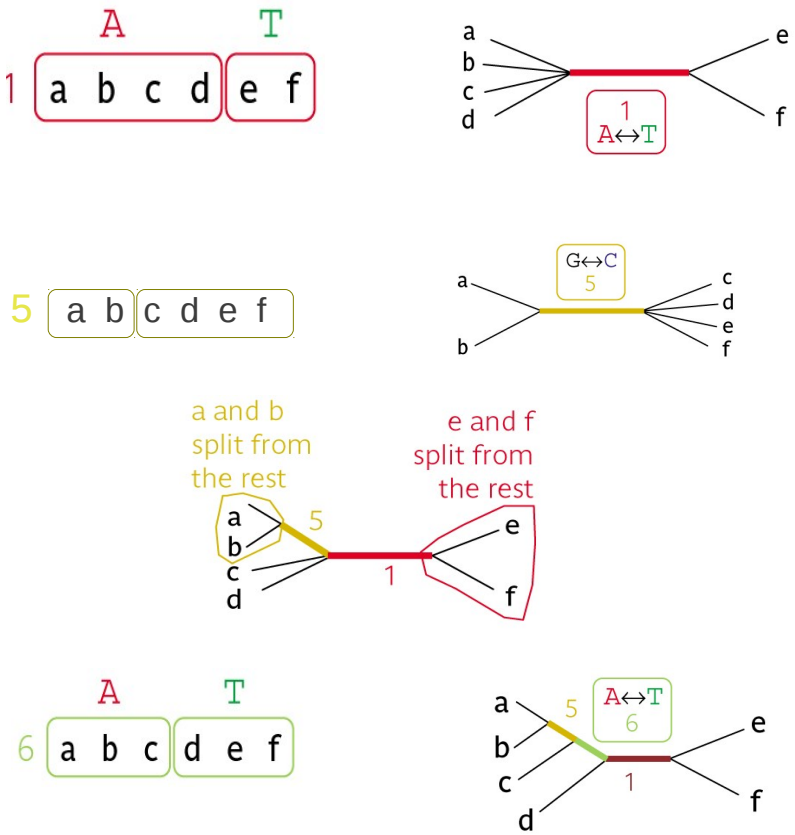


Alignment: a key step to recover evolutionary history from DNA sequences

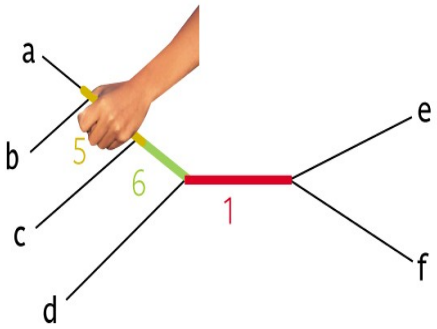


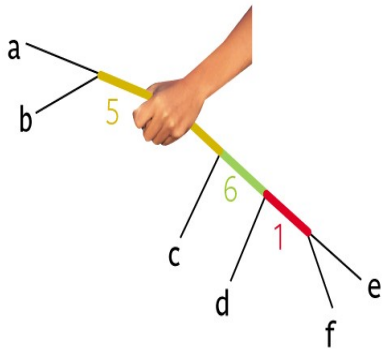
Alignment: a key step to recover evolutionary history from DNA sequences

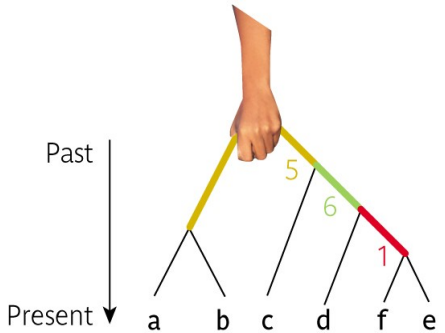
```
  123456
a ATGACA
b ATCACA
c AAAAGA
d ATAAGT
e TTTAGT
f TCAAGT
```

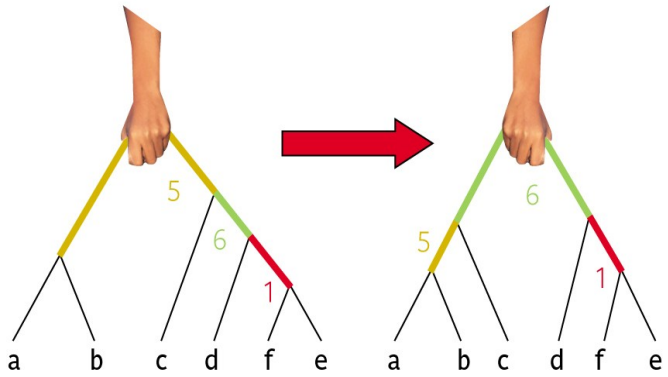


Reading trees: rooting a tree

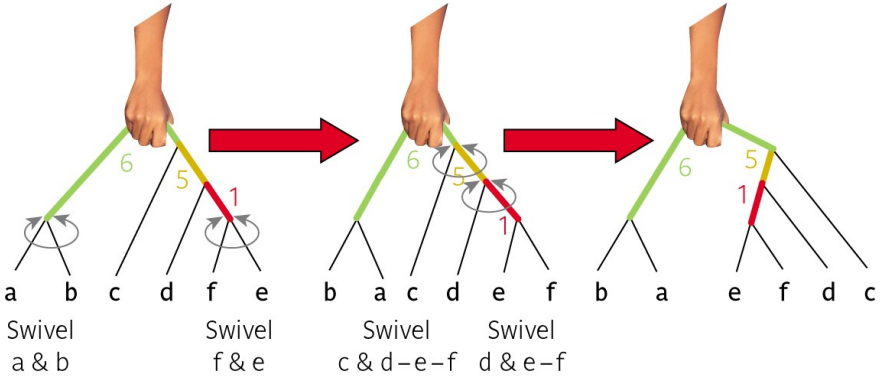


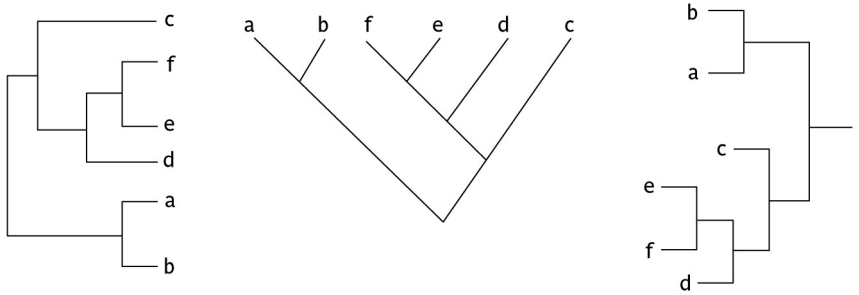






Reading trees: changes that do not make a difference





Statistical methods for phylogenetic inference

1. Distance based methods

2. Parsimony methods

3. Maximum likelihood methods

Distance-based methods for phylogenetic inference

1. Compute evolutionary distances for all pairs of taxa

	Human	Horse	Cow	Kangaroo	Newt	Carp
Human		0.129	0.129	0.205	0.572	0.665
Horse	0.134		0.129	0.232	0.638	0.651
Cow	0.134	0.134		0.197	0.598	0.624
Kangaroo	0.216	0.246	0.207		0.638	0.708
Newt	0.662	0.751	0.697	0.751		0.752
Carp	0.789	0.770	0.733	0.849	0.913	

PC correction, and gamma distance (alpha=2)

Distance-based methods for phylogenetic inference

1. Compute evolutionary distances for all pairs of taxa

2. Construct trees from distance data
   Examine different possible topologies and chose the best as the “true “ topology

Distance-based methods for phylogenetic inference

1. Compute evolutionary distances for all pairs of taxa

2. Construct trees from distance data

UPGMA: Unweighted pair-group method (using arithmetic averages)

Hp) Constant rate of evolution between lineages
Ts) Wrong topology when Hp does not hold

Distance-based methods for phylogenetic inference

1. Compute evolutionary distances for all pairs of taxa

2. Construct trees from distance data

UPGMA: Unweighted pair-group method (using arithmetic averages)

ME: Minimum evolution method

Idea: Minimization of the sum of all branch length estimates

Very intense and slow computations!

Distance-based methods for phylogenetic inference

1. Compute evolutionary distances for all pairs of taxa

2. Construct trees from distance data

UPGMA: Unweighted pair-group method (using arithmetic averages)

ME: Minimum evolution method

NJ: Neighbor joining method

Hp) ME method + Applied neighbor joining strategy

Statistical methods for phylogenetic inference

1. Distance based methods

2. Parsimony methods

3. Maximum likelihood methods

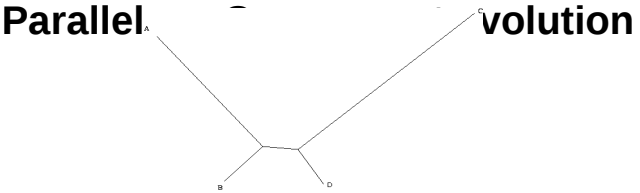
Maximum Parsimony methods for phylogenetic inference

1. Alignment of the sequences

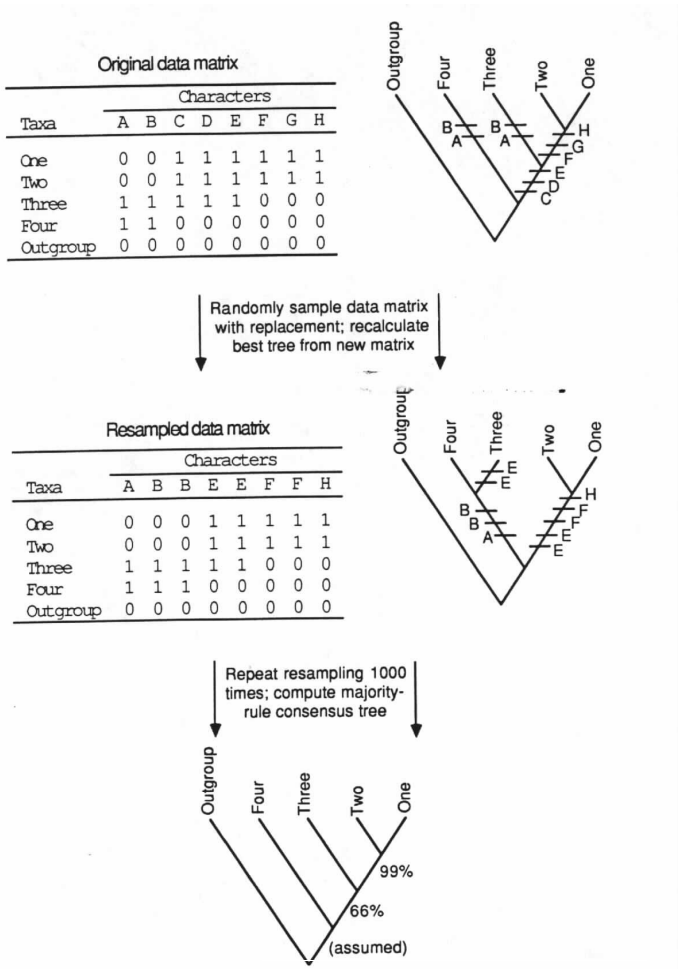
2. Select the tree that requires the smallest number of substitutions to explain the entire evolutionary process

Complex and slow computations!
Specific methods have been designed to reduce the search space

A typical issue: IGNORING MULTIPLE HITS!
Long branch attraction



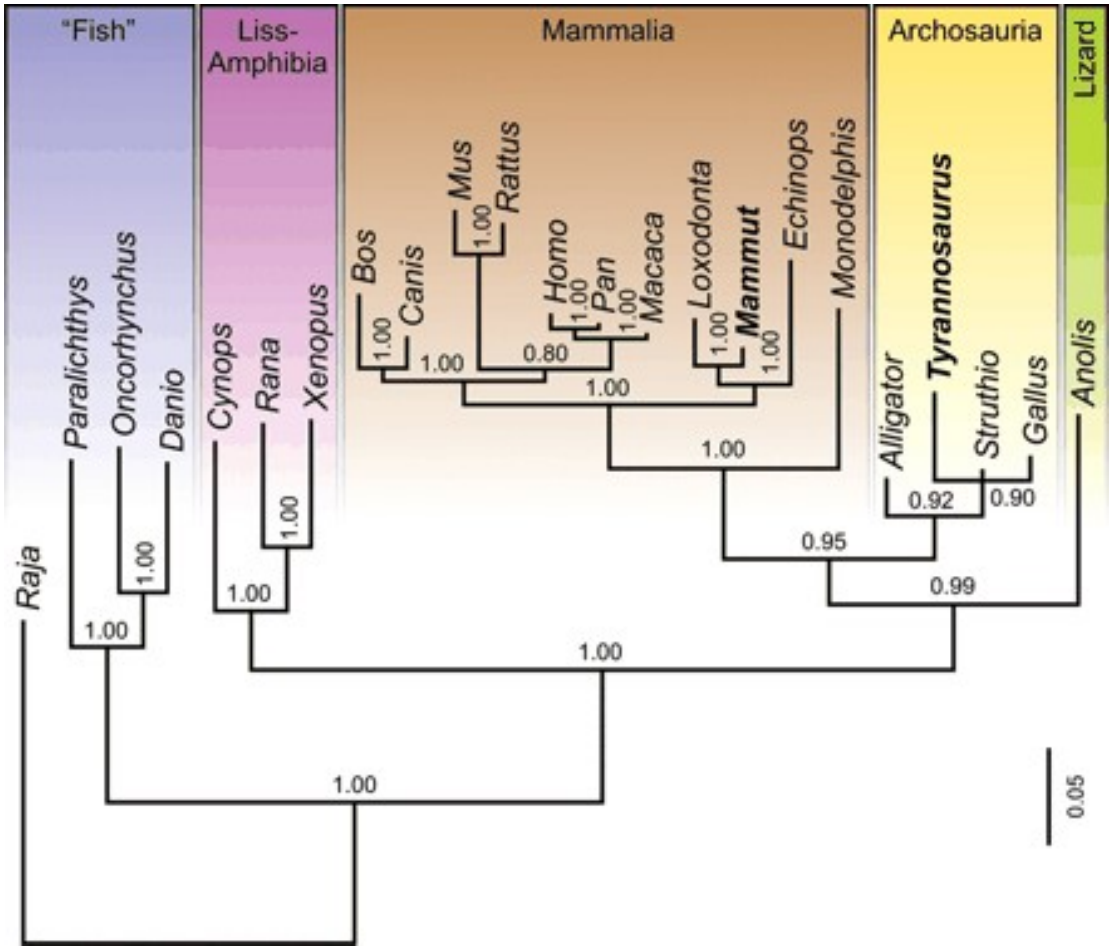
I have a tree now.
How do I know that I can trust it?
BOOTSTRAP
The percentage of bootstrap trees that support a given node

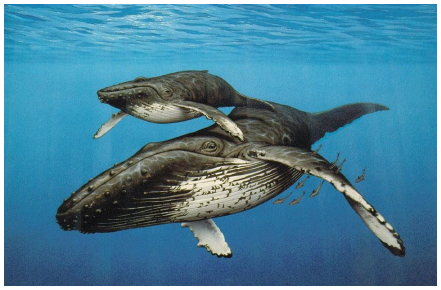


http://akira.ruc.dk/~olesk/sekvens/

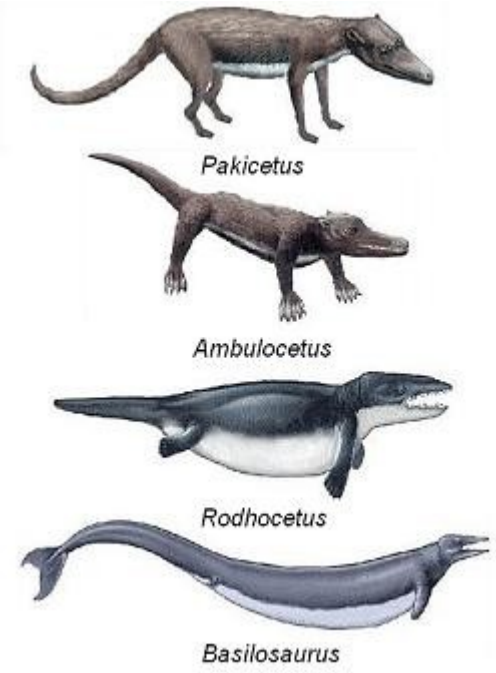
Revisiting the evolutionary history of dinosaurs with a molecular approach

Molecular Phylogenetics of Mastodon and Tyrannosaurus rex





YOUR PRACTICAL:
Using molecular phylogenetics to reveal a “back to the sea tale”



Ambulocetus