Bristol Bioinformatics Workshop 2019

Tom Williams, Gary Barker, Tom Batstone, Christopher Edsall

5th-6th November, 2019

Course aims

- Introduction to **a general approach** to doing bioinformatics
- **Core, transferable skills** for using computers to do science (UNIX, Python)
- **A taster** of some specific biological applications

Course structure

- **Day 1 (today)**
  Introduction to UNIX and using compute clusters
  Introduction to Python 3

- **Day 2:**
  Introduction to genome analysis
  Introduction to phylogenetics

Today's schedule

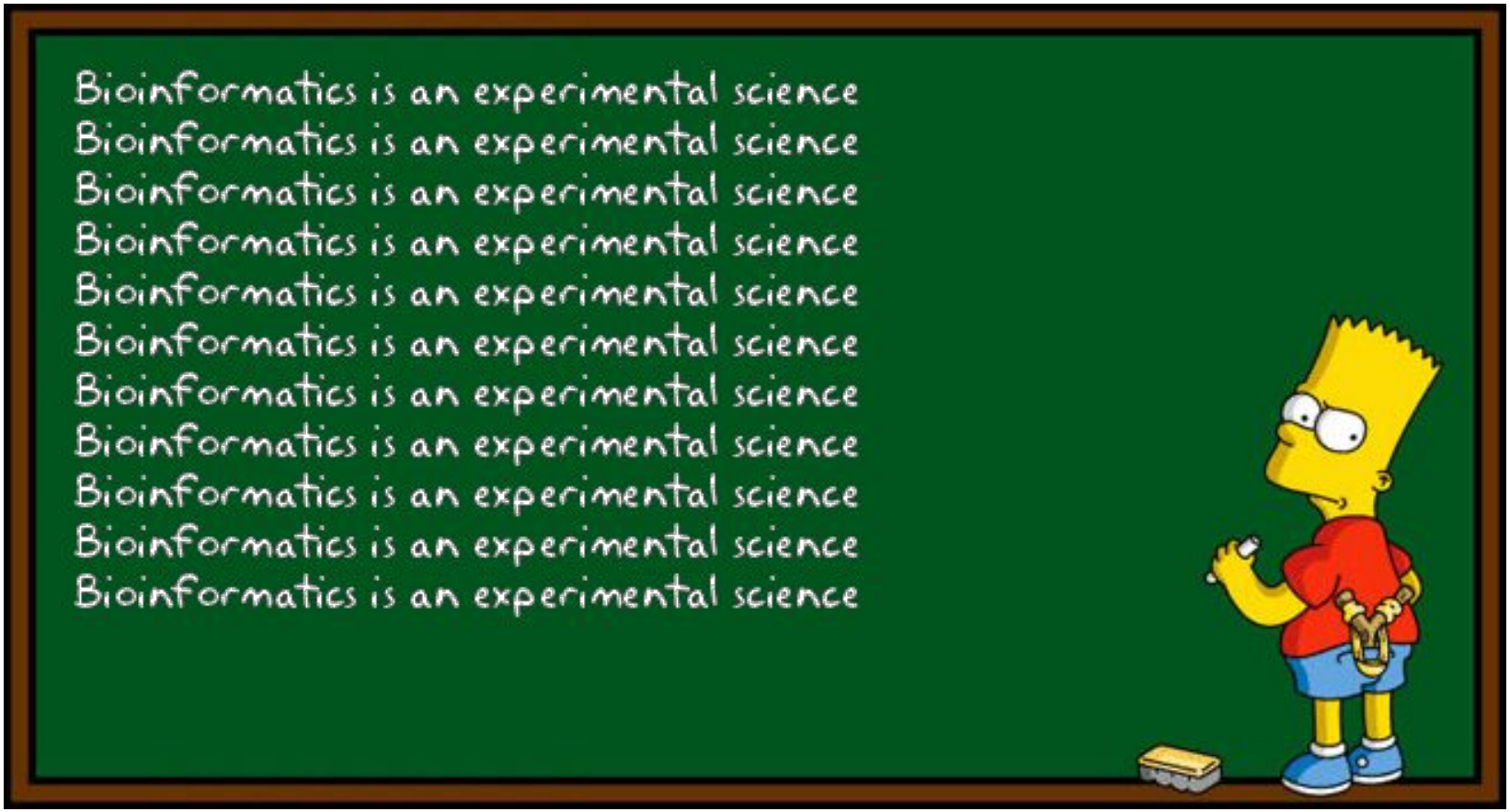
Time	Material	Lecturer
9.30-1	Unix and compute clusters	Tom Batstone
2-5	Introduction to Python	Christopher Edsall

Protip

Participation is key to getting the most out of this kind of course.

Please ask (and answer) questions!

The basic idea: doing bioinformatics



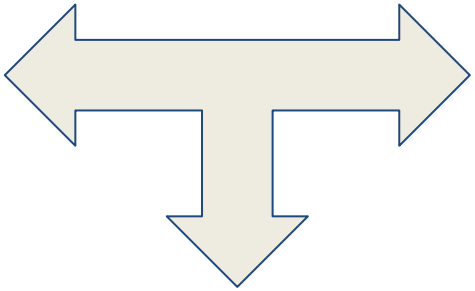
Doing science: a workflow

Collect, collect, collect, collect, collect, collect, collect, analyze, Nature

We use computers to explore data or test hypotheses

=

Doing science in an integrated way



“An **hour** of #bioinformatics **can save** a week in the **lab**. ALSO An **hour in the lab can save** a week of #bioinformatics” - Torsten Seeman

Two kinds of questions frequently encountered when doing bioinformatics

Scientific questions:

- Is my hypothesis testable, and with what data?
- What assumptions can I make when analyzing the data?
- Which of the available analysis methods is most appropriate?

Is what I'm doing sensible?

Technical questions:

- What software/methods are available for problem X?
- How do I install this software on my computer?
- How do I fix this error message?
- How do I make it give me the answer?

Why won't it $%$£"!# work?

Two kinds of questions frequently encountered when doing bioinformatics

Scientific questions:

- Is my hypothesis testable, and with what data?
- What assumptions can I make when analyzing the data?
- Which of the available analysis methods is most appropriate?

Is what I'm doing sensible?

*

Technical questions:

- What software/methods are available for problem X?
- How do I install this software on my computer?
- How do I fix this error message?
- How do I make it give me the answer?

Why won't it $%$£"!# work?

Bioinformatics: in sum

- Computers provide powerful tools for analyzing data
- They aren’t black boxes, into which you feed data and out of which the result comes
- They can speed up (not replace) our thinking, and make many types of analysis possible
- Be as critical as you would be in any other area of science

Introduction to molecular phylogenetics

(a) Core concepts

Tom Williams

Core concepts in molecular phylogenetics

- Some justification for doing phylogenetics
- Interpreting phylogenetic trees
- Inferring phylogenetic trees

A practical.

What is phylogenetics?

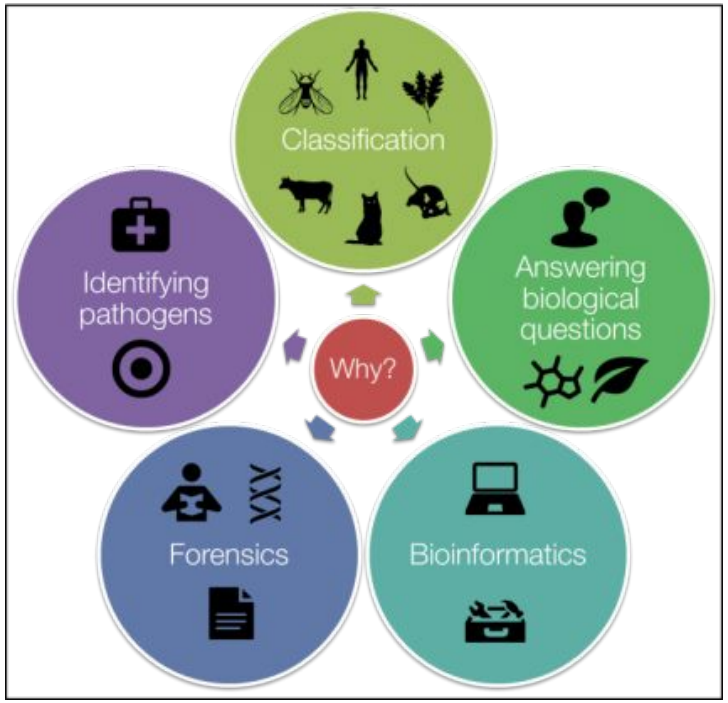
The study of the evolutionary relationships among evolving entities (species, genomes, genes, individuals).

"The time will come, though I shall not live to see it, when we shall have fairly true genealogical trees of each great kingdom of nature"

- Darwin (1857)

"One of the grand biological ideas is to be able to work out the complete detailed quantitative phylogenetic tree --- the history of the origins of all living species, back to the very beginning."

- Dayhoff and Eck (1972)



(Various practical applications)

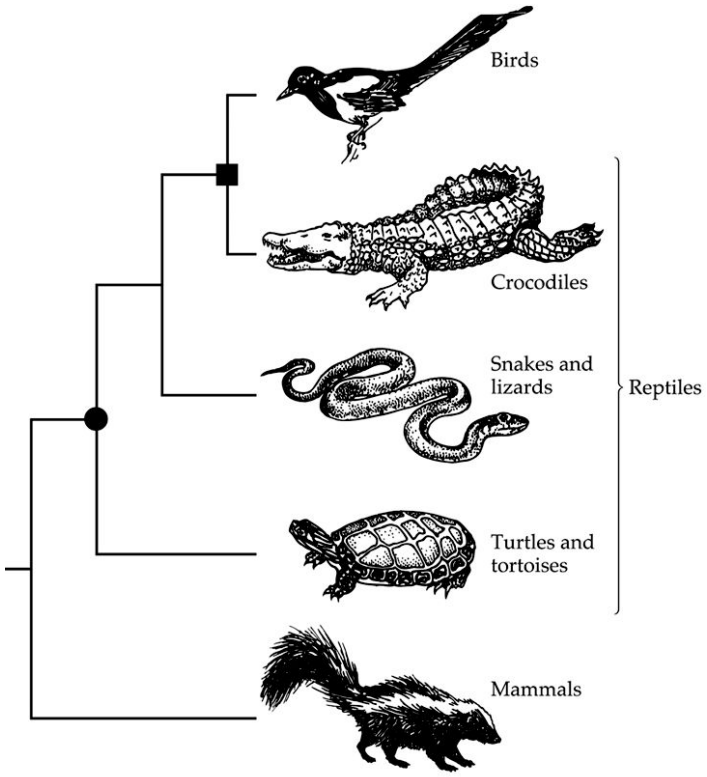
What is a phylogeny (evolutionary tree)?

A branching diagram representing the genealogical relationships among species

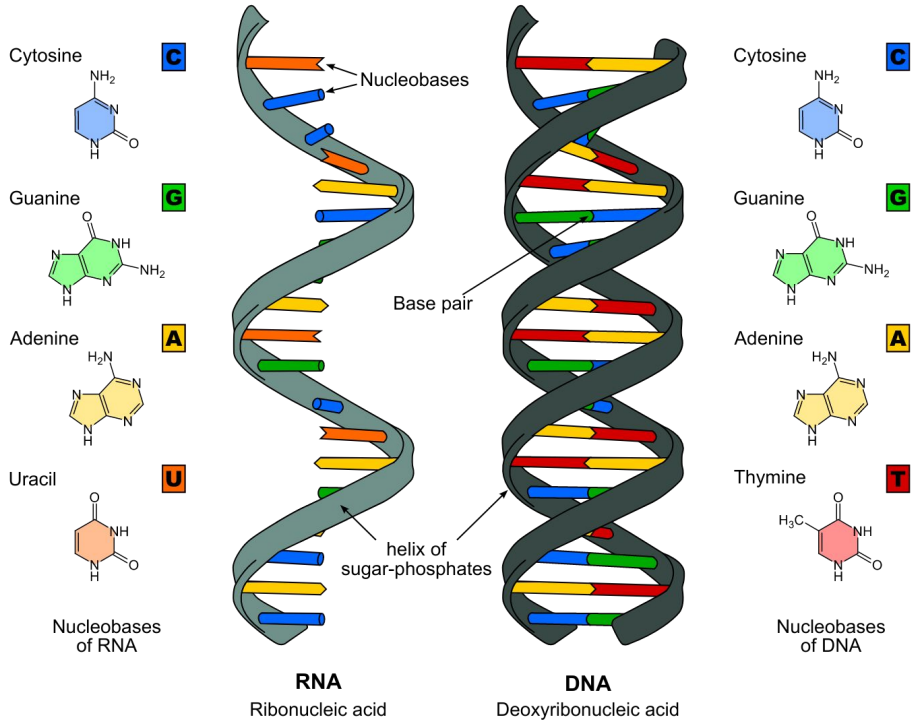
Can be inferred using:

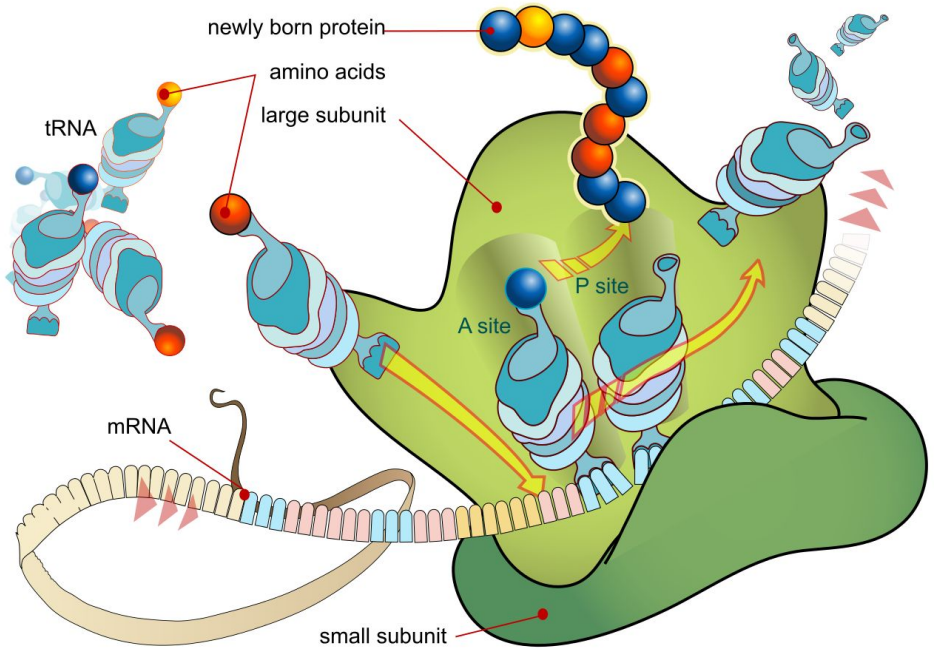
- Genetic (molecular) data
- Morphological data
- Any kind of character/trait data

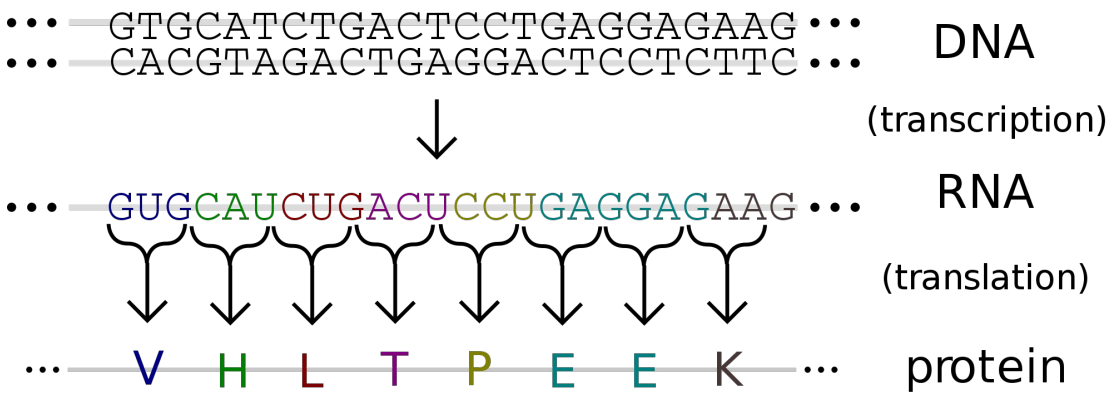
Normally, we want to infer a tree (or trees) that make sense of the character data



Molecular sequences provide digital information, many homologous characters





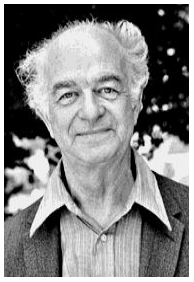


Wikimedia

DNA sequences can be used to make phylogenetic trees

Common ancestor

Gene A Sequence



GAACTCGAC

GATCTCGAC

GATCTGGGC

GCTCTGGGC

Common ancestor of vertebrates

GCTCTGCGTA

We use alignment to construct hypotheses of homology for sequence data

ATTGGCG

AGGAG

Sequences!

ATTGGCG

A--GGAG

Alignment!

AGGCG

+TT

C=>A

ATTGGCG

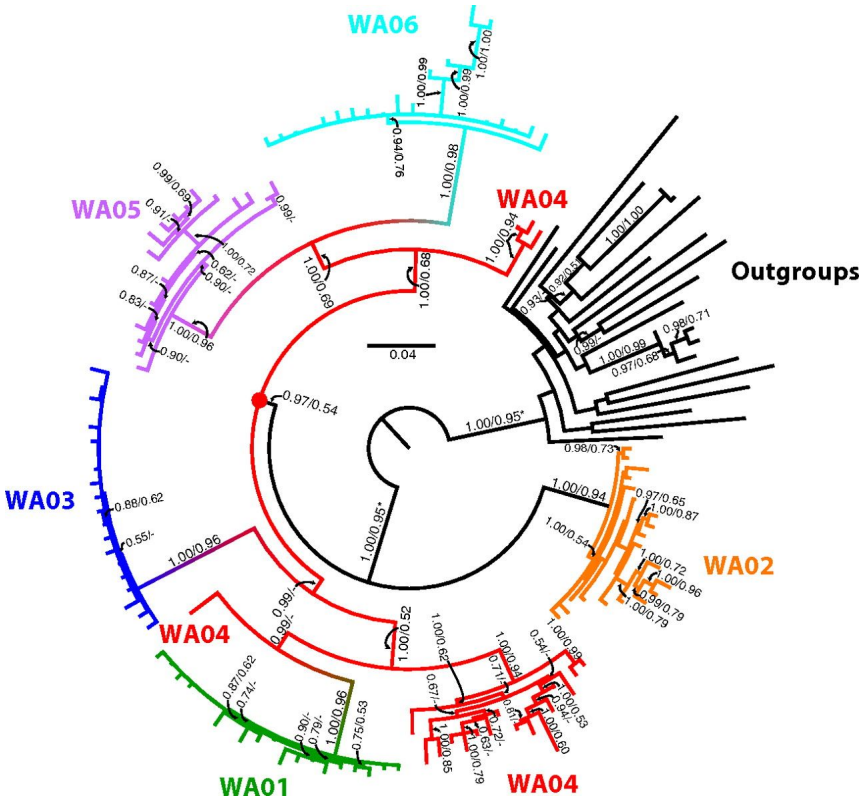
A--GGAG

Homology!

2. Interpreting phylogenetic trees

What kind of information can we learn from phylogenies?

- Information about relationships (who is mostly closely related to whom?)



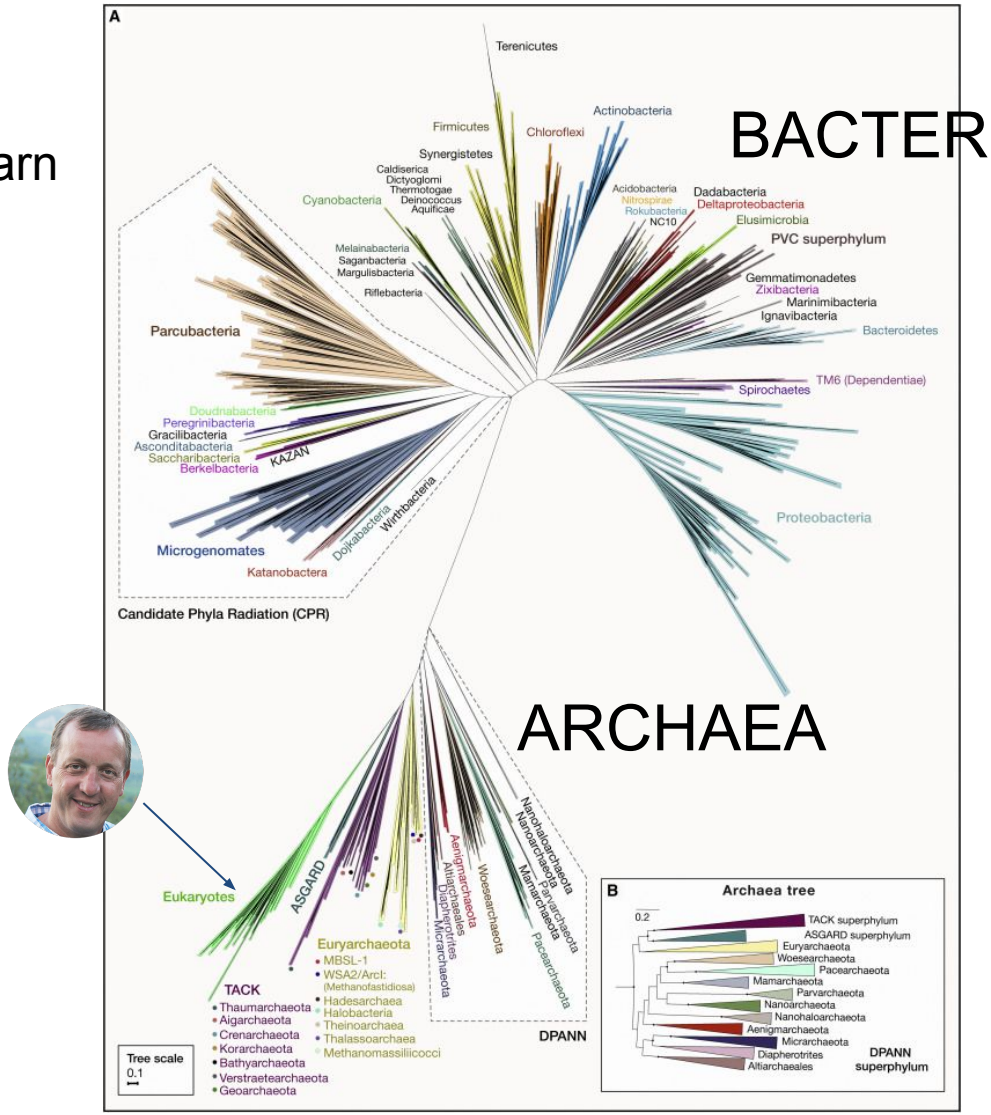
E.g. **epidemiology**
(HIV-1 env gene)

Scaduto et al. (2010)
PNAS

Interpreting phylogenetic trees

What kind of information can we learn from phylogenies?

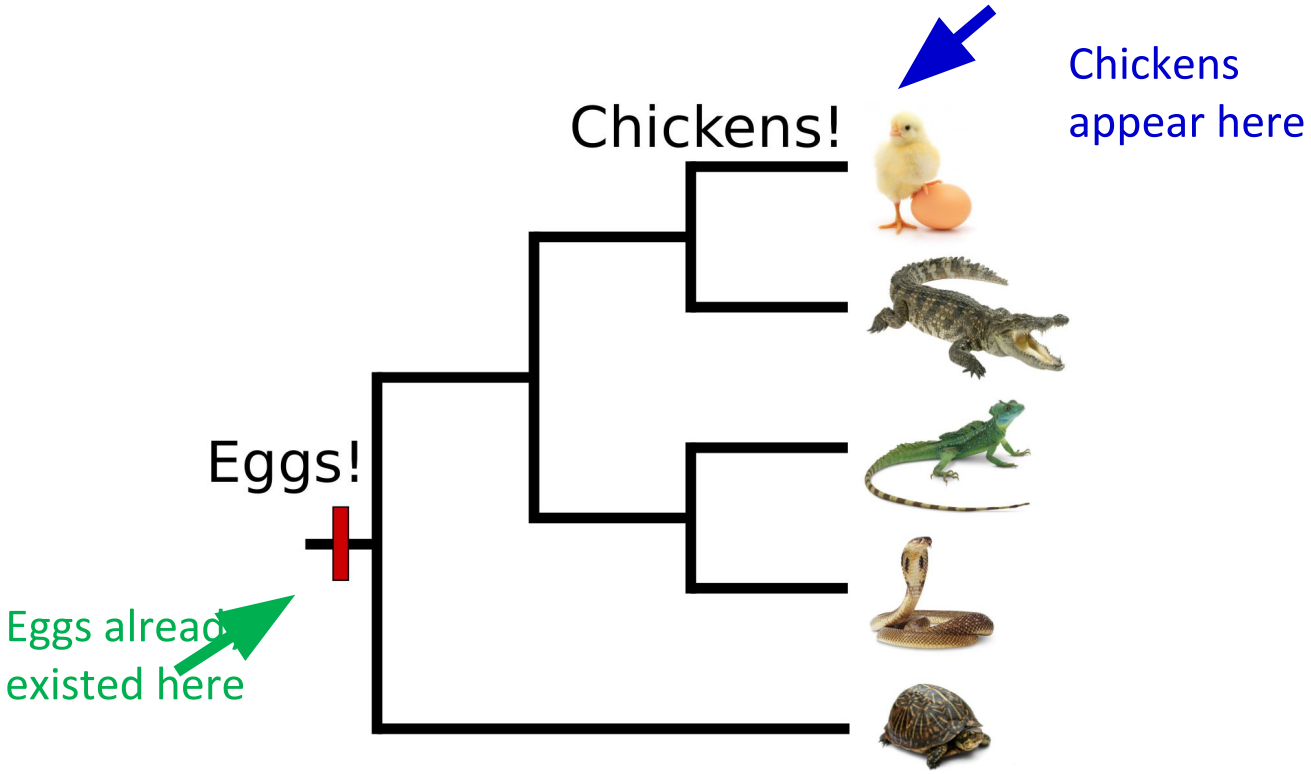
- Information about evolutionary relationships
- Information about genetic (or other) diversity



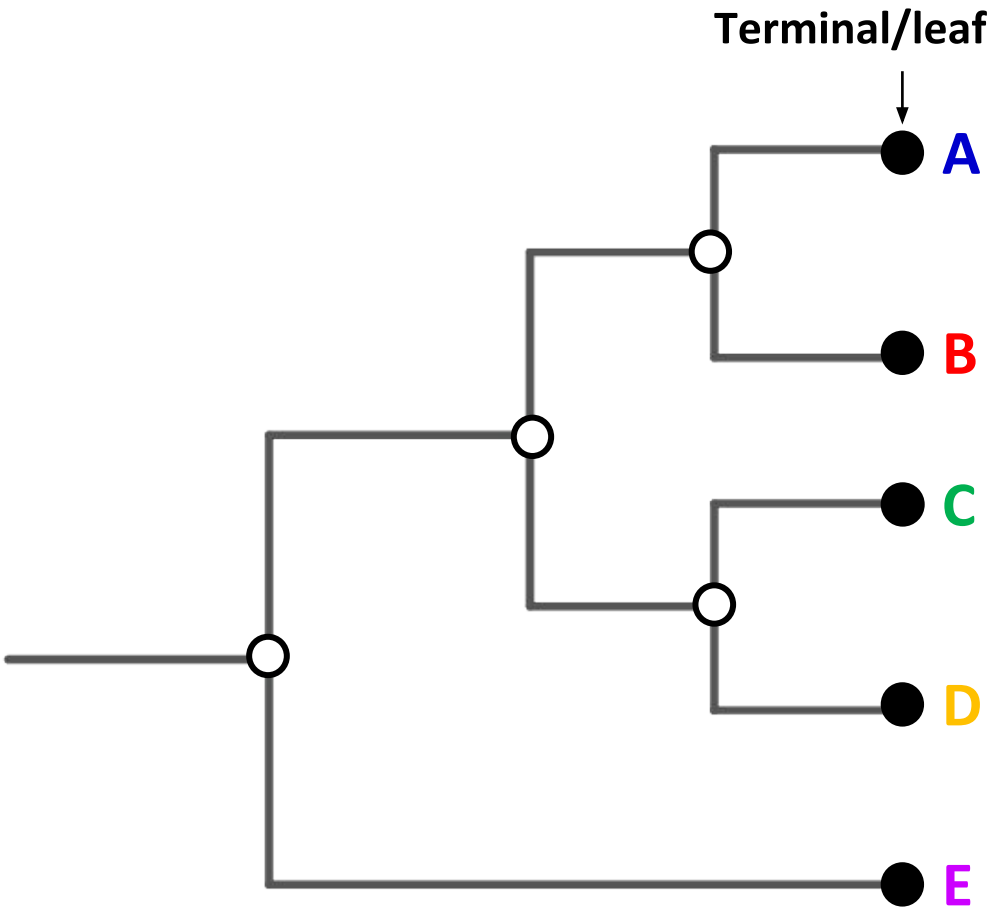
Interpreting phylogenetic trees

What kind of information can we learn from phylogenies?

- The history of trait evolution

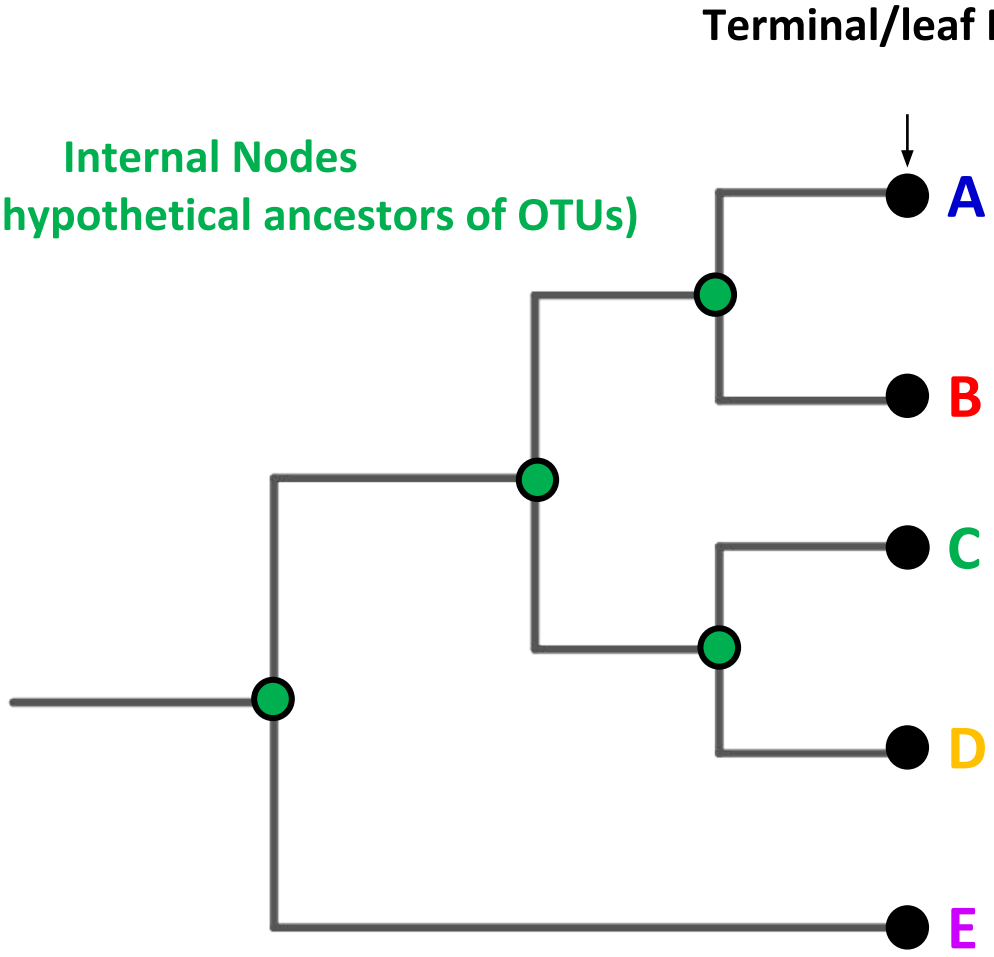


Anatomy of a phylogenetic tree



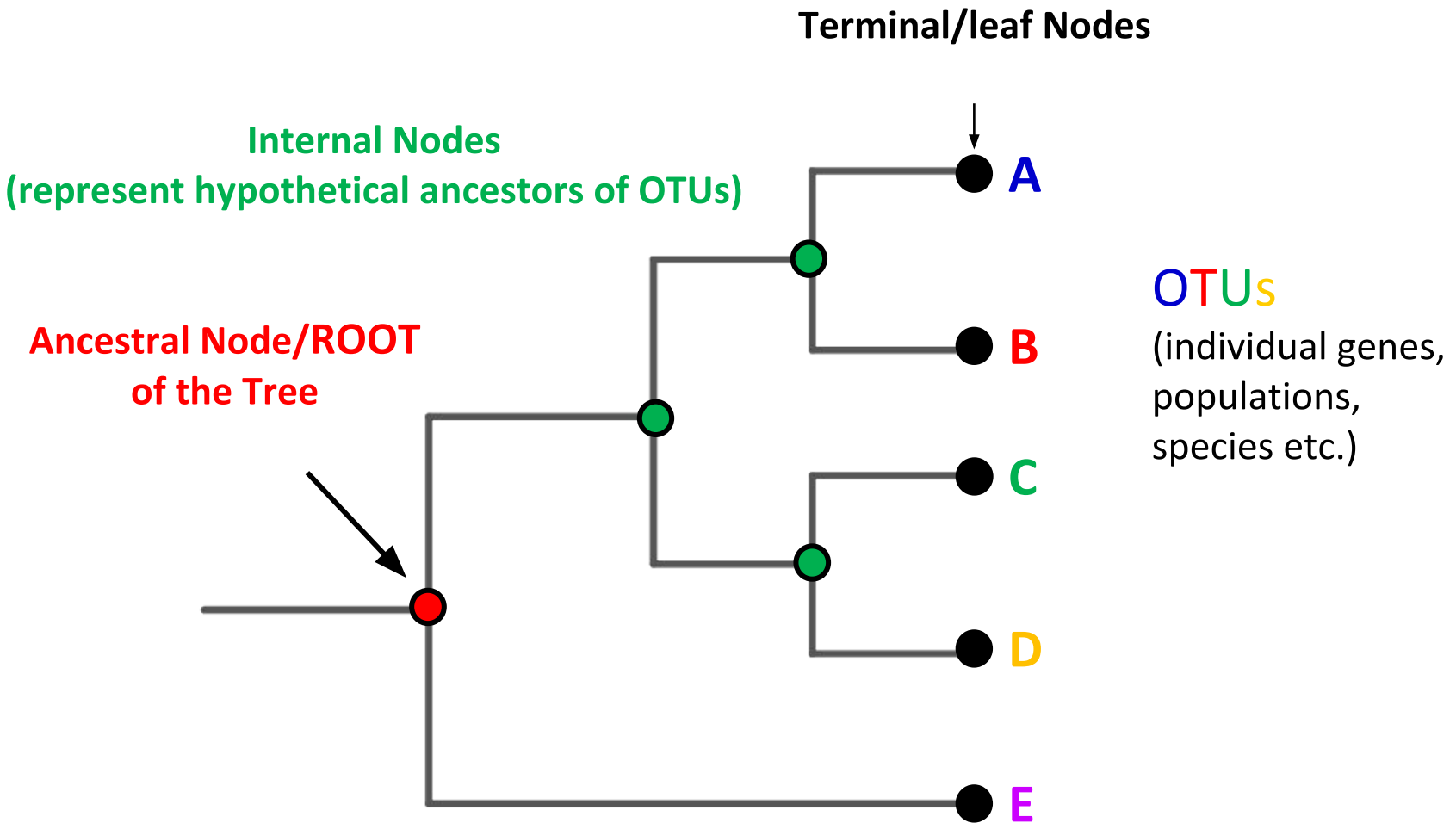
OTUs
(individual genes, populations, species etc.)

Anatomy of a phylogenetic tree

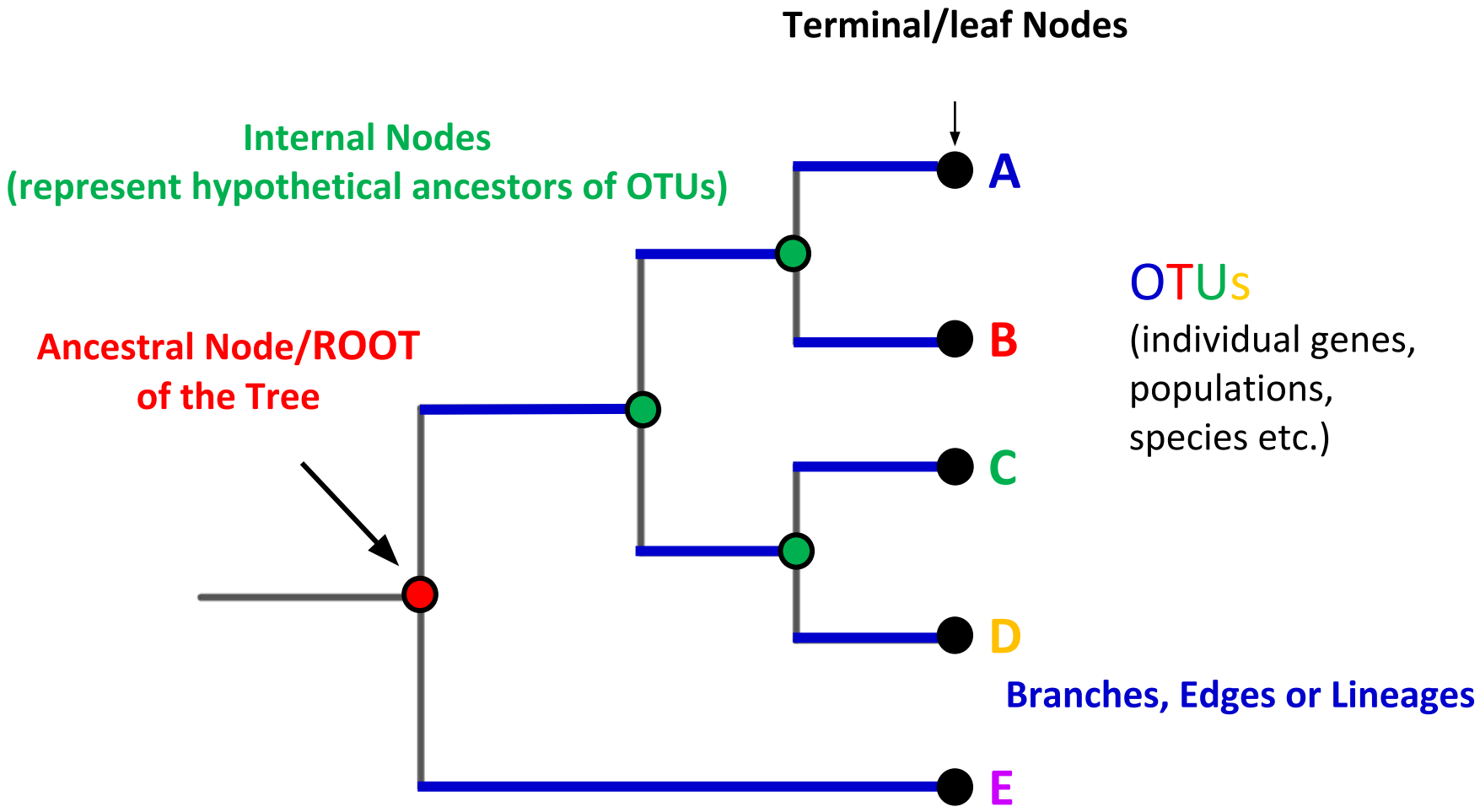


OTUs
(individual genes, populations, species etc.)

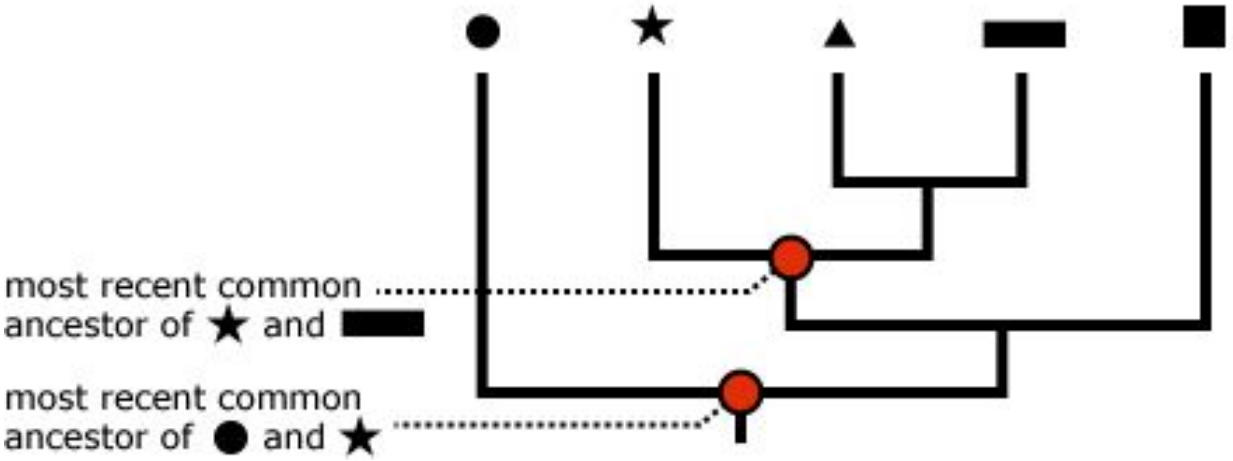
Anatomy of a phylogenetic tree



Anatomy of a phylogenetic tree

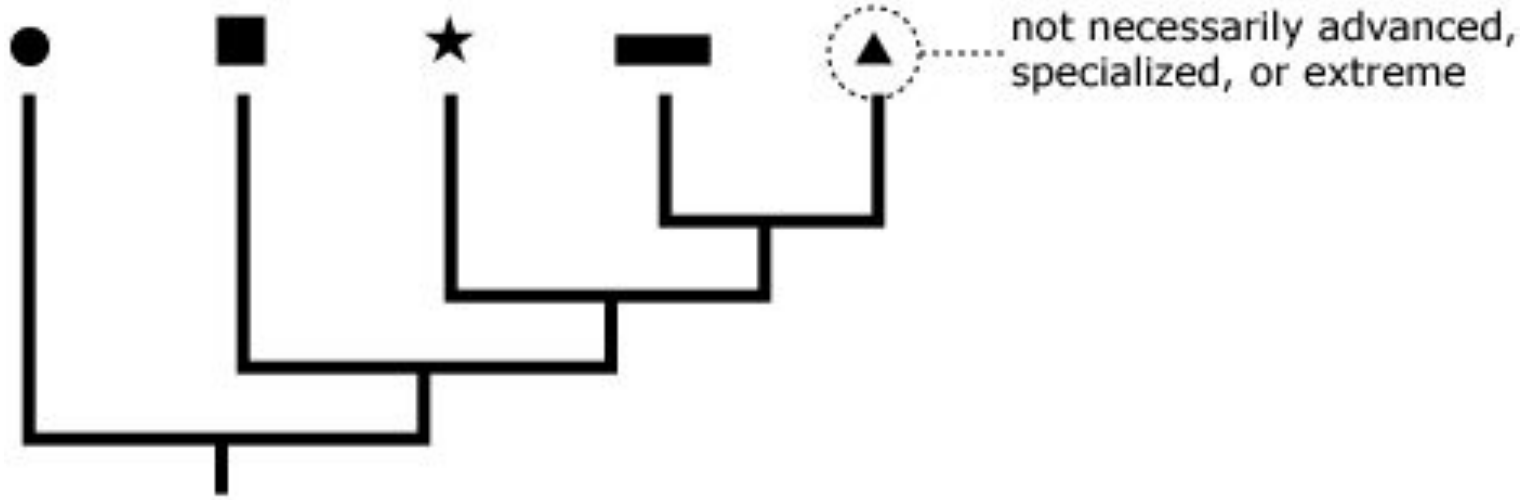


Reading trees: patterns of relatedness



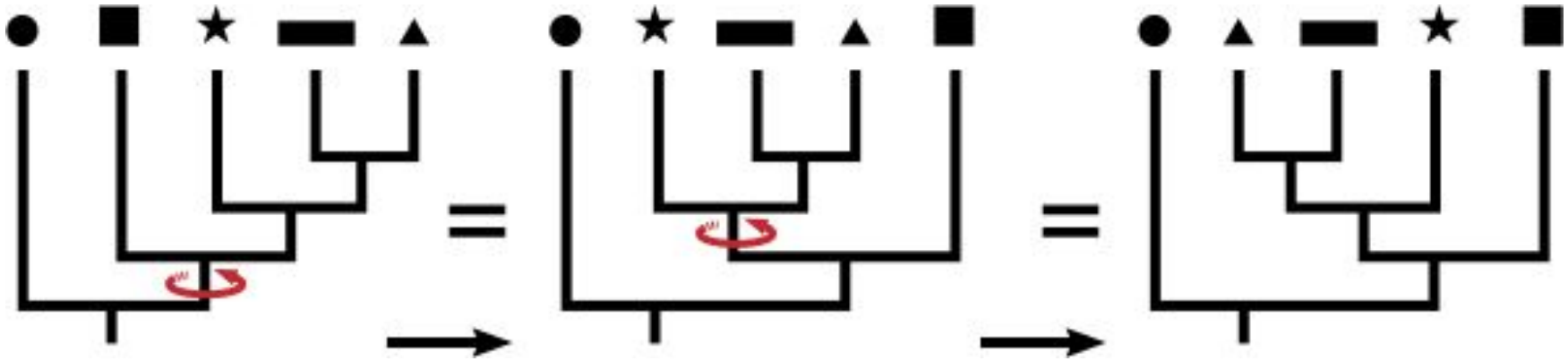
The more recently two species share a common ancestor, the more closely related they are.

Reading trees: patterns of relatedness



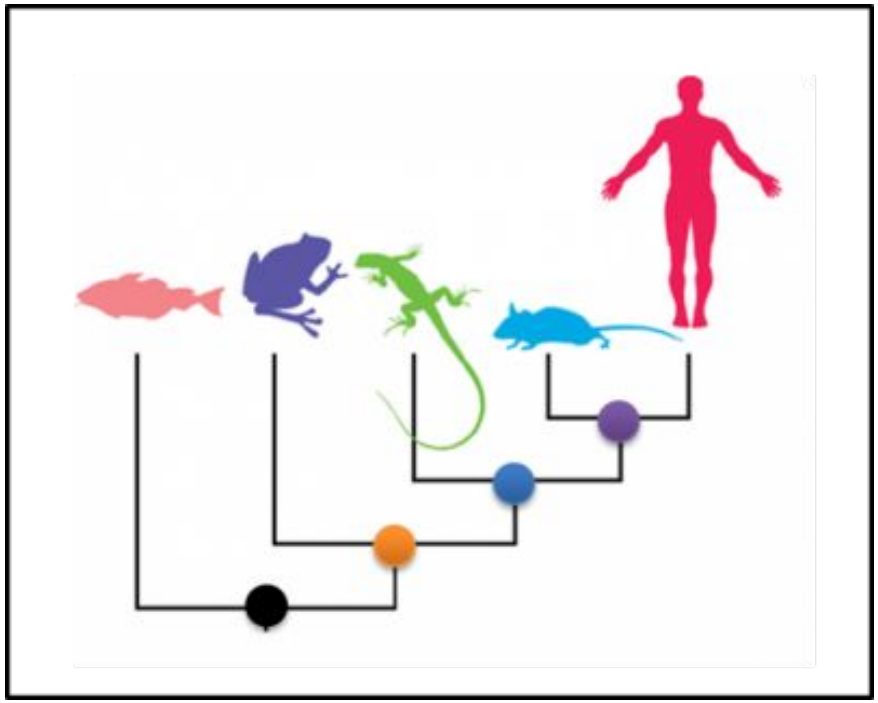
Trees depict evolutionary relationships, not evolutionary progress (!)

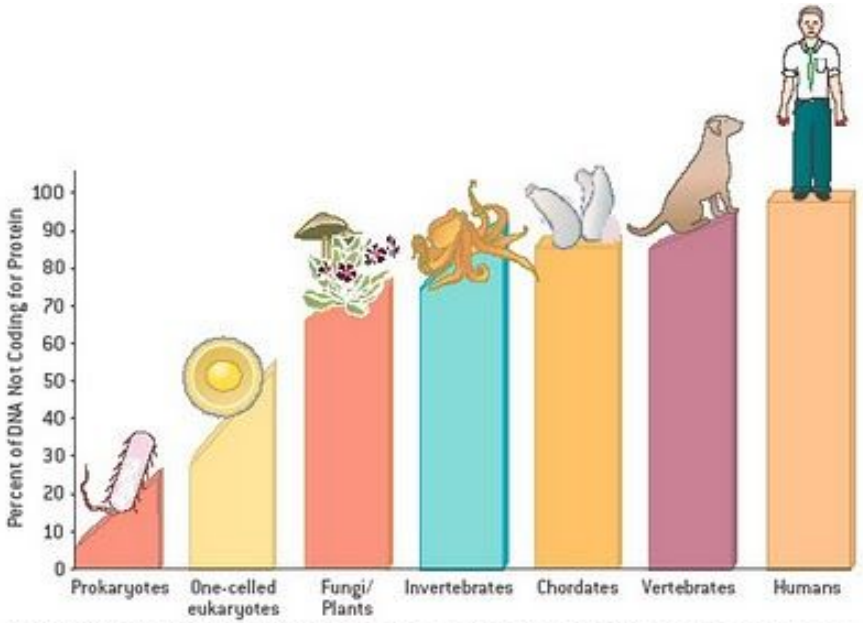
Reading trees: patterns of relatedness



Rotation around internal nodes produces equivalent trees.

Beware of “Dog’s Ass” plots

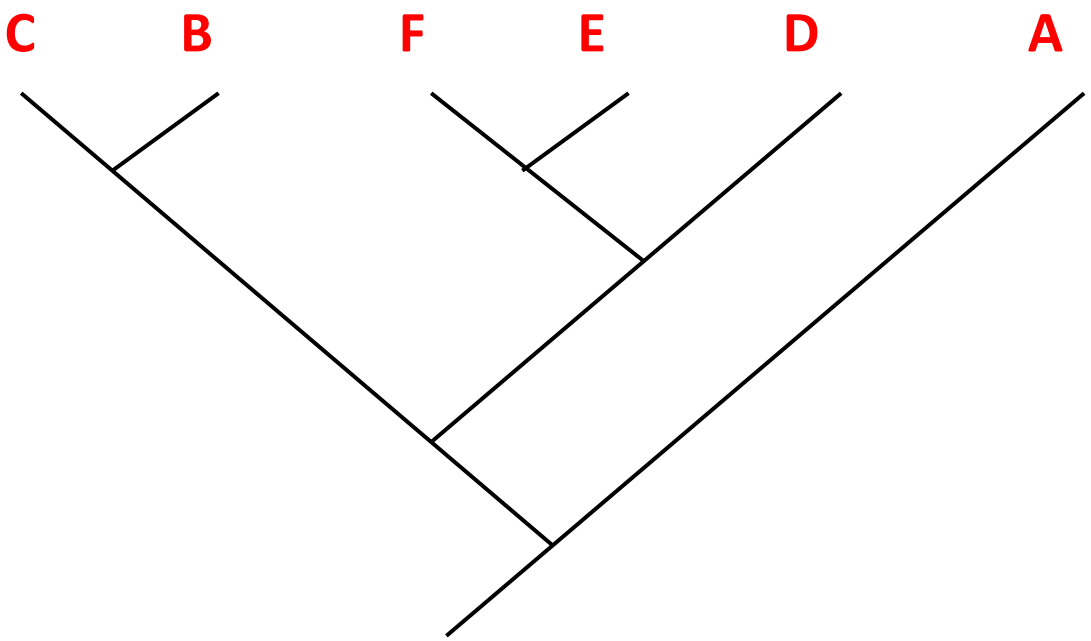




NONPROTEIN-CODING SEQUENCES make up only a small fraction of the DNA of prokaryotes. Among eukaryotes, as their complexity increases, generally so, too, does the proportion of their DNA that does not code for protein. The noncoding sequences have been considered junk, but perhaps it actually helps to explain organisms' complexity.

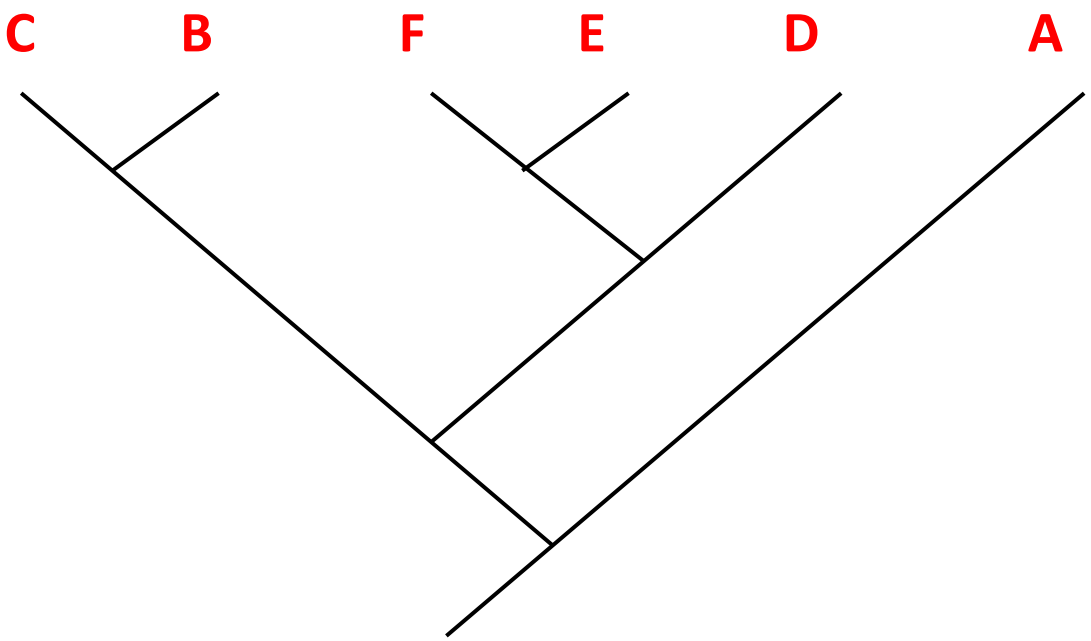
http://www.genomicron.evolverzone.com/2007/09/dogs-ass-plots-daps/

Quiz: reading trees



Is “E” more closely related to “D” or to “F”?
Is “E” more closely related to “B” or to “A”?
Is “E” more closely related to “B” or to “C”?

Quiz: reading trees



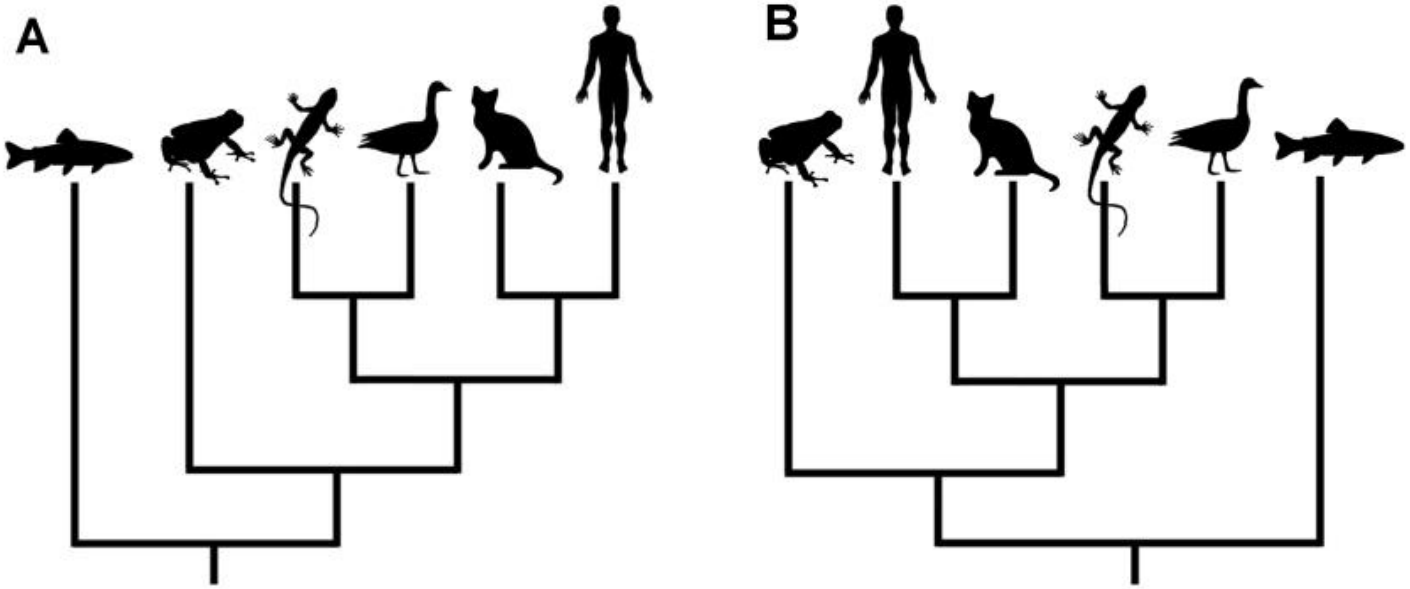
“E” more closely related to “F”

“E” more closely related to “B”

“E” more closely related to neither (equally to “B” & “C”)

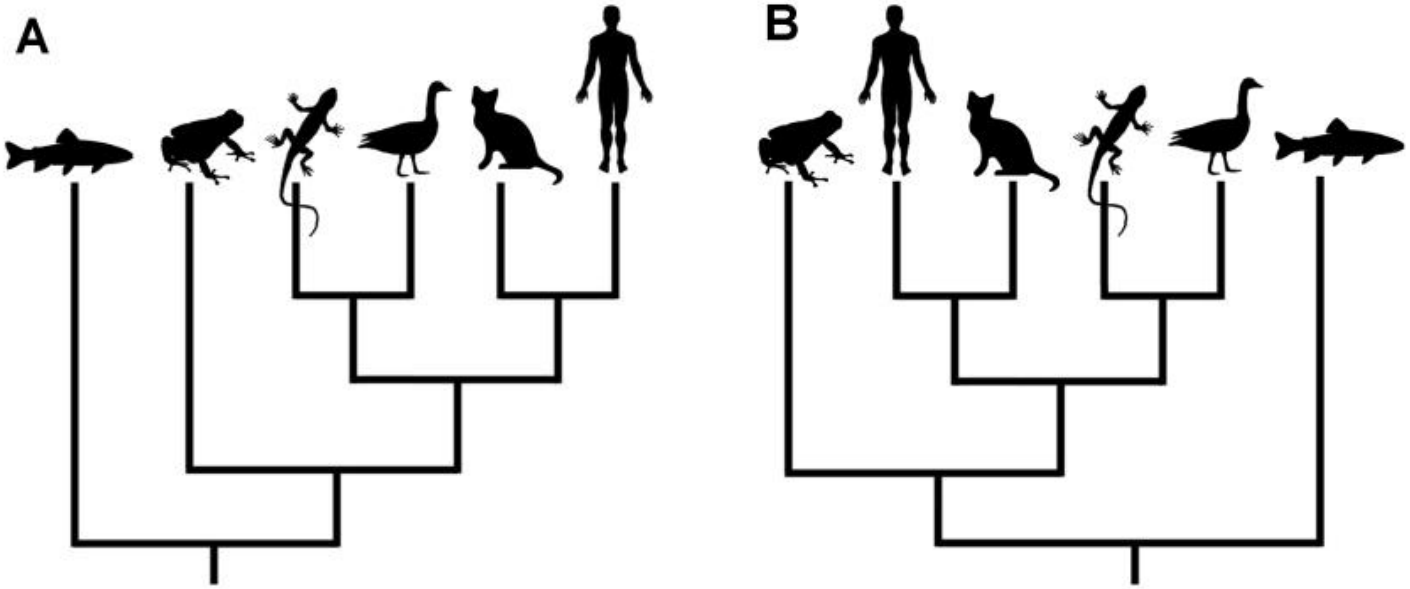
Quiz: reading trees

Find the differences between the two trees!



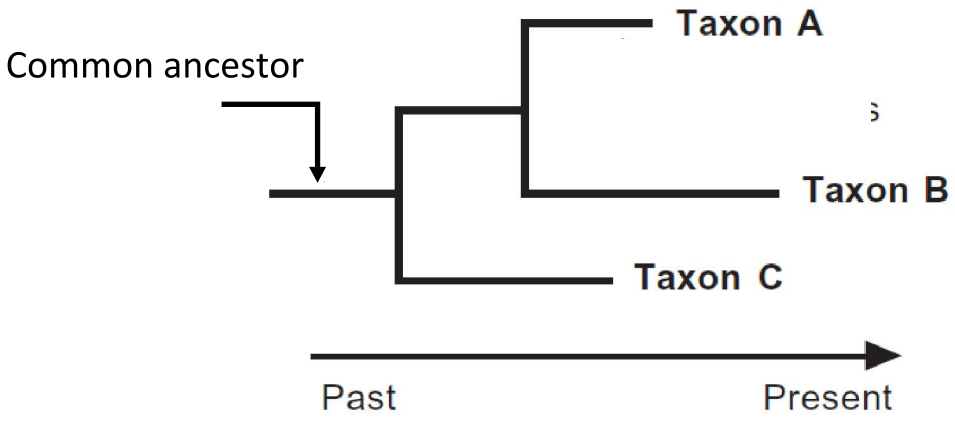
Quiz: reading trees

Find the differences between the two trees!



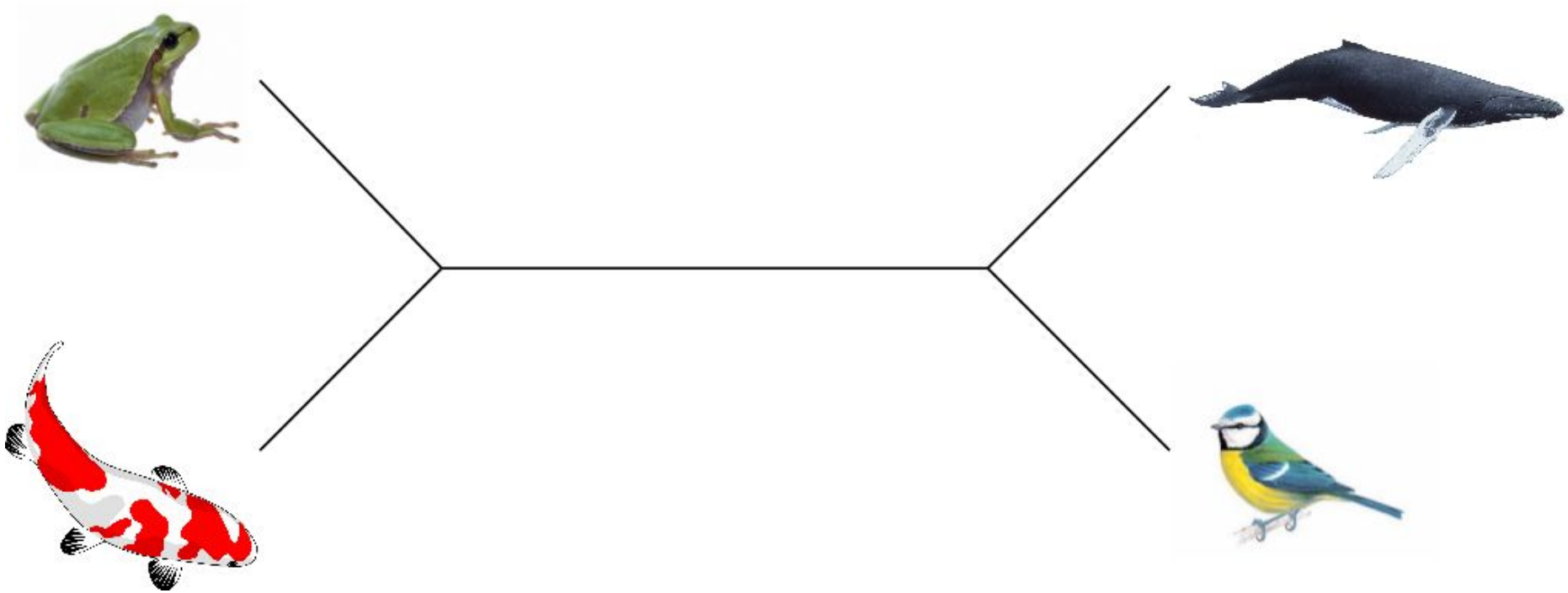
They are the same.

4. The root: a very special node on the tree

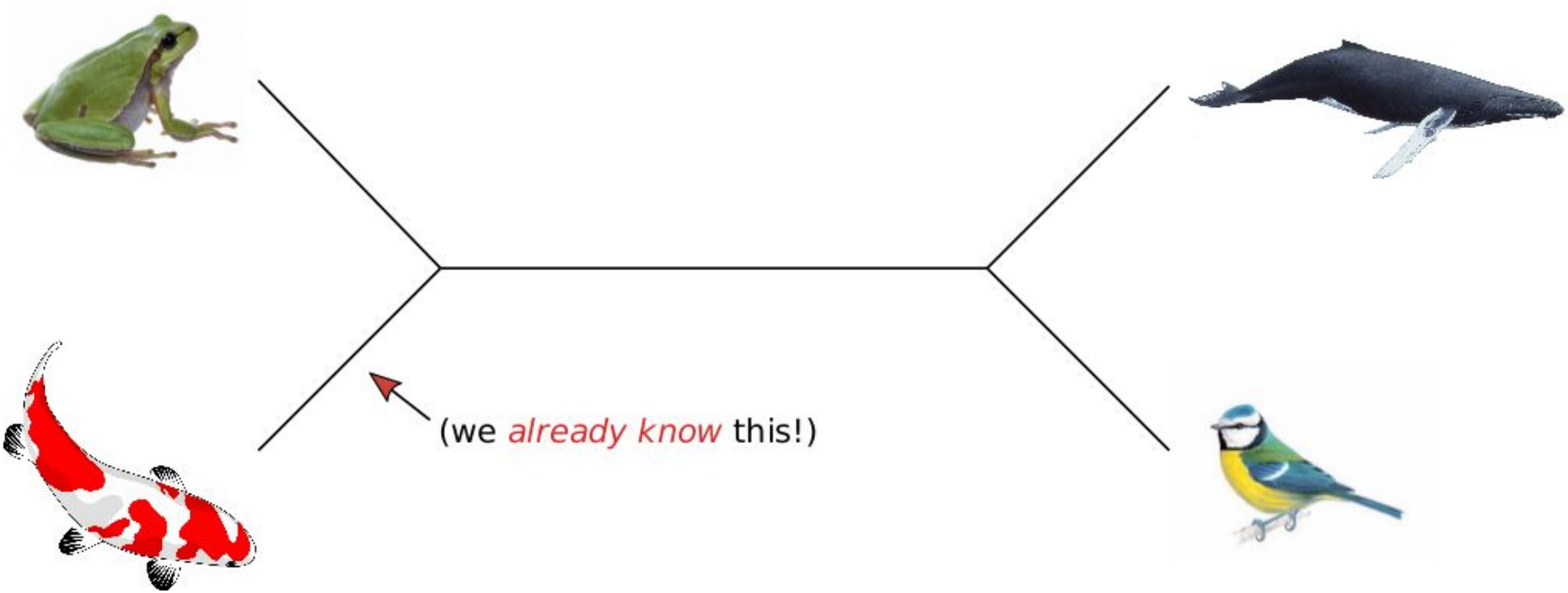


- The root is the oldest point on the tree (most recent common ancestor of all taxa). It orients the tree in time.
- All inferences about ancestor-descendant relationships **and relatedness** depend on the root position.
- Most phylogenetic methods do not estimate the root!

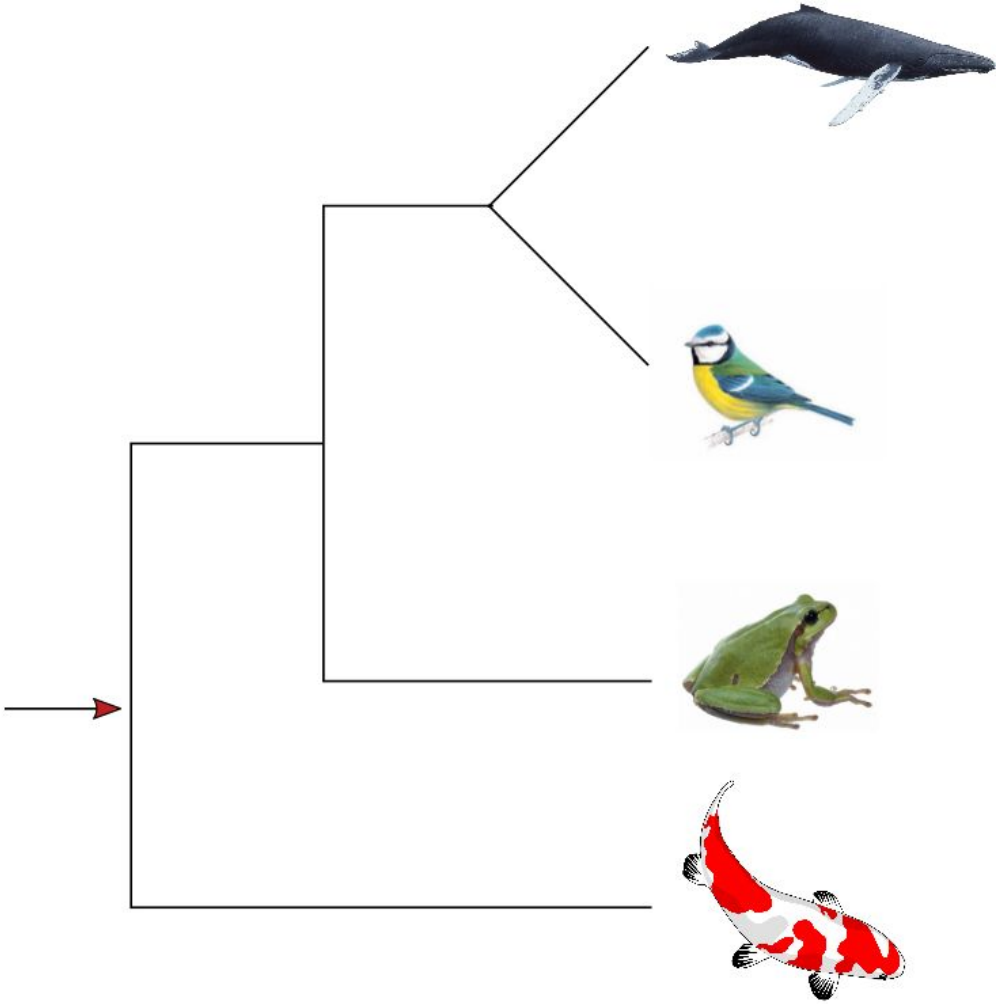
Rooting an unrooted tree using *a priori* knowledge (outgroup)



Rooting an unrooted tree using *a priori* knowledge (outgroup)



Rooting an unrooted tree using *a priori* knowledge (outgroup)



- Most packages will arbitrarily root the tree on some branch, for no particular reason
- You need to re-root on the outgroup branch

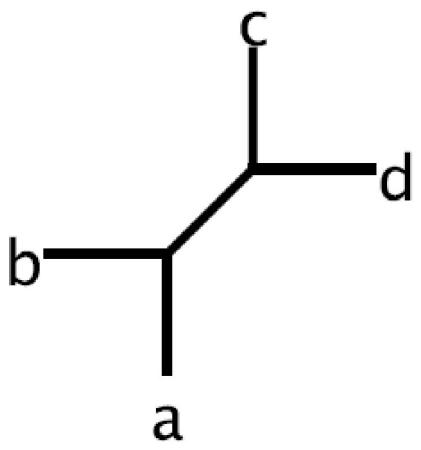
This doesn't apply if you use a method that can infer the root:

- Molecular clock
- Gene tree-species tree reconciliation
- Non-reversible/ non-stationary phylogenetic model
- Some pop. gen. settings (if you know the ancestral/derived allele)

Quiz: interpreting roots

d is more closely related to:

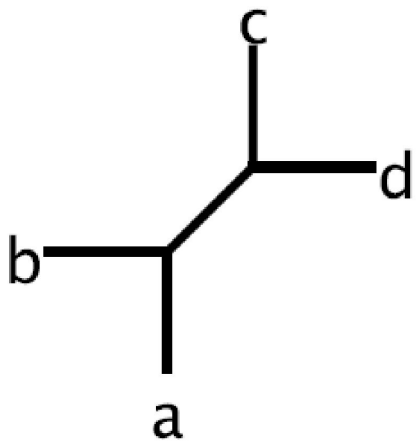
1. a than it is to b or c
2. b than it is to a or c
3. c than it is to a or b



Quiz: interpreting roots

d is more closely related to:

1. a than it is to b or c
2. b than it is to a or c
3. c than it is to a or b



It depends on where the root is!

Now: draw a set of rooted trees where (3) is True, and a set where (3) is False.

Introduction to molecular phylogenetics

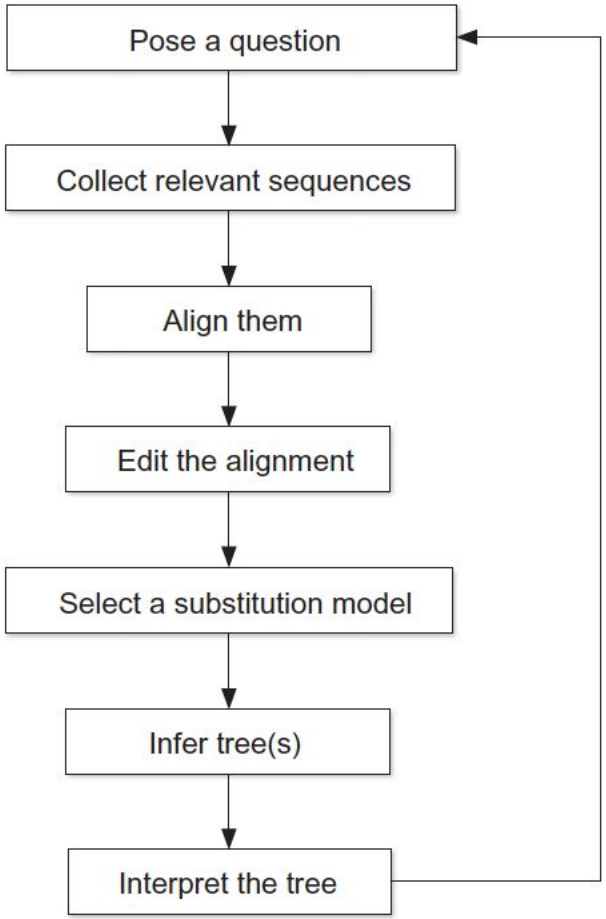
(a) A basic workflow

Tom Williams

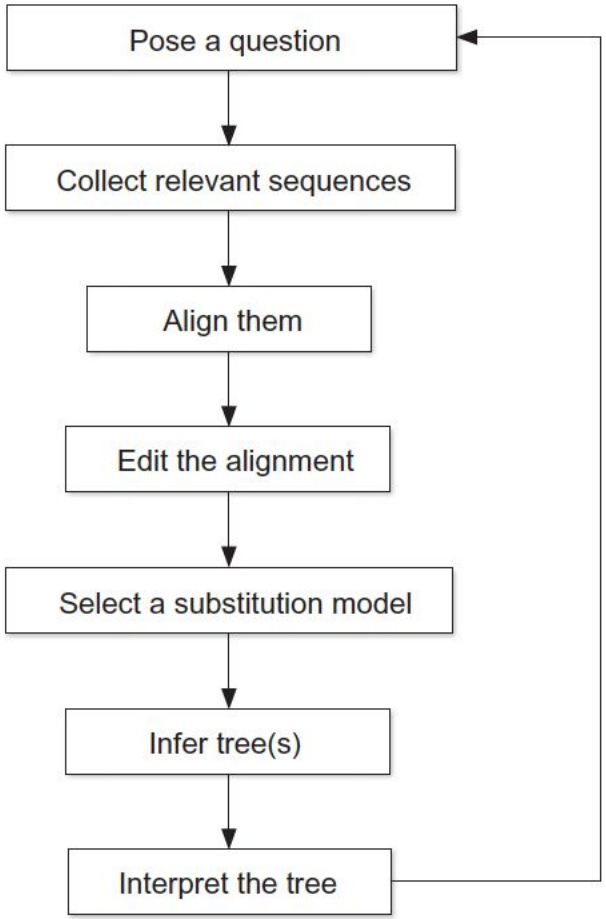
A basic workflow for molecular phylogenetics

- Typical analysis steps
- Some discussion of the methodology at each step

Molecular phylogenetics: a possible flowchart

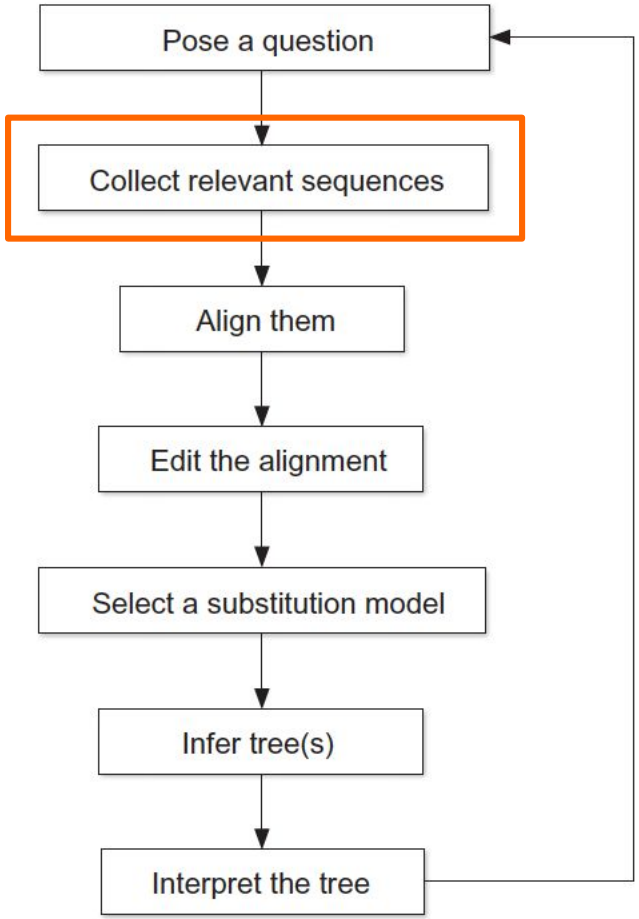


Molecular phylogenetics: a possible flowchart



A critical approach to analysis is key: don't fall into the black box!

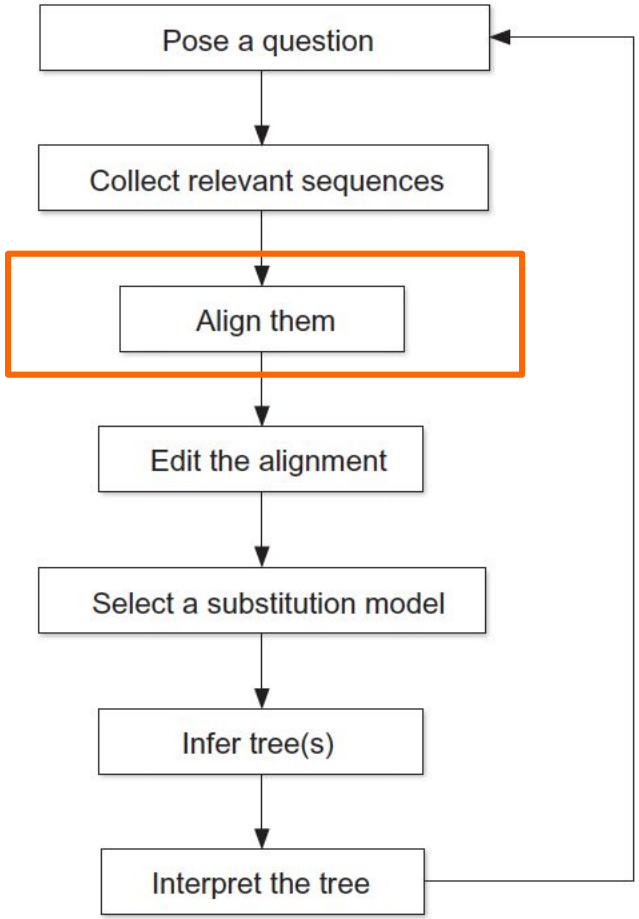
Molecular phylogenetics: a possible flowchart



Did I miss any important ones?

A critical approach to analysis is key: don't fall into the black box!

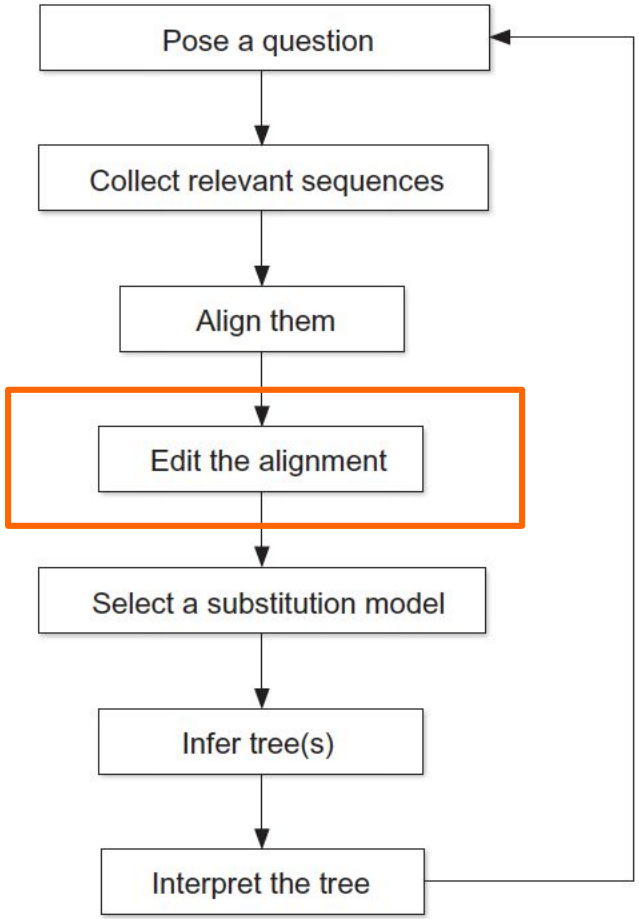
Molecular phylogenetics: a possible flowchart



How certain is my alignment?
Would other alignments give a different tree?

A critical approach to analysis is key: don't fall into the black box!

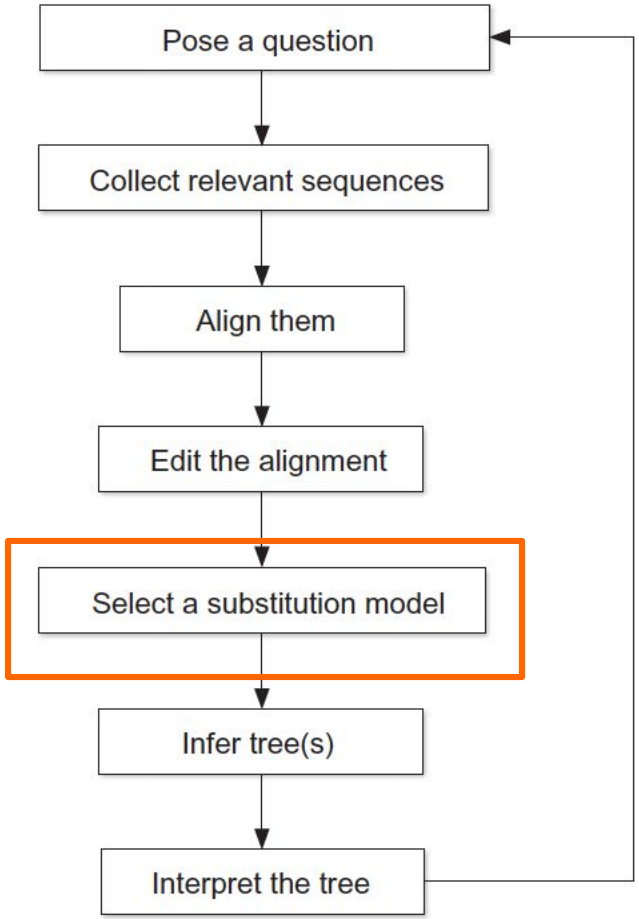
Molecular phylogenetics: a possible flowchart



Which bits, if any, should I remove?

A critical approach to analysis is key: don't fall into the black box!

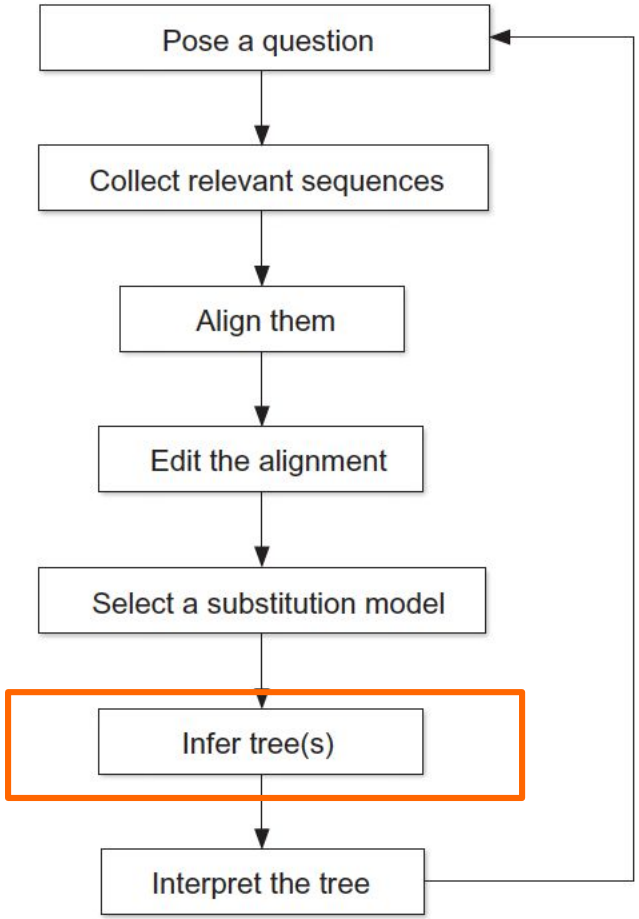
Molecular phylogenetics: a possible flowchart



Is this the best available model?
Is it good enough?

A critical approach to analysis is key: don't fall into the black box!

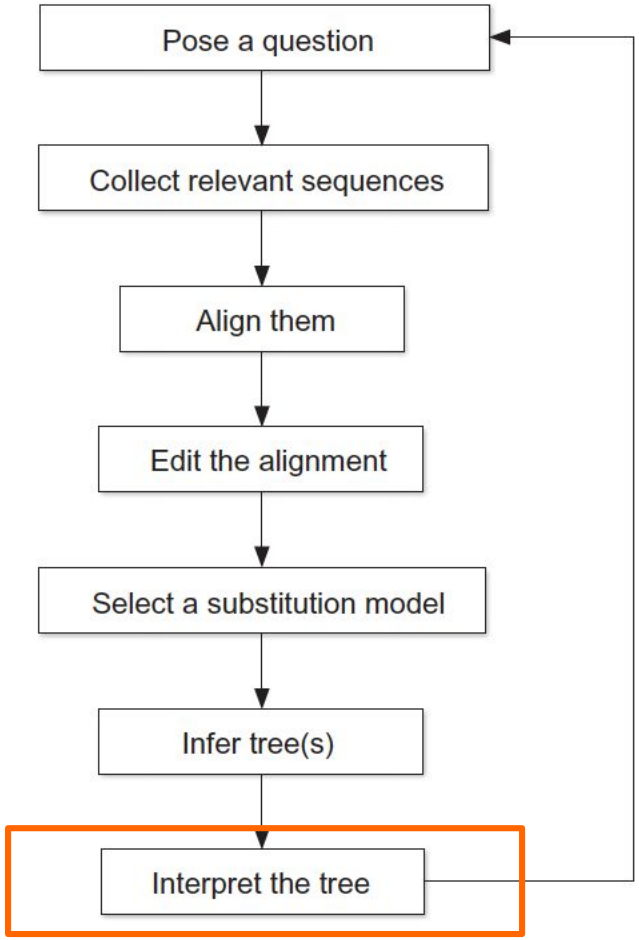
Molecular phylogenetics: a possible flowchart



Are my ML or Bayesian analyses working properly? (Local maxima, convergence)

A critical approach to analysis is key: don't fall into the black box!

Molecular phylogenetics: a possible flowchart



How well-supported are the key branches?

What does the tree mean for my hypothesis?

Can alternative hypotheses be rejected?

A critical approach to analysis is key: don't fall into the black box!

1. Pose a question

- Systematics, classification, character evolution in a clade of interest
- Use phylogenies as part of modelling a broader process

2. Assembling a dataset

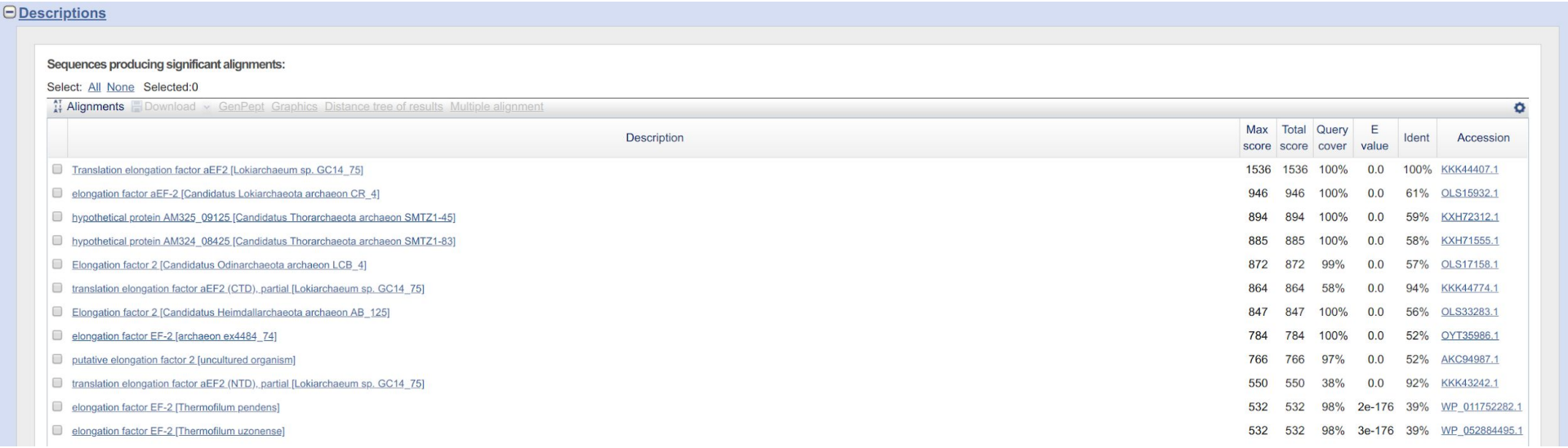
- Where do I get the data?
- Use an existing dataset
- Use your own data
- Use published data to test a new hypothesis (GenBank, Ensembl, PDB, Dryad, FigShare...)

2. Assembling a dataset

- Starting with a query, search sequence databases.
- An iterative process: start broad and refine using initial trees
- Should I include it or not? Do the experiment.

2. Assembling a dataset

- Starting with a query, search sequence databases (BLAST, PSI-BLAST, HMMER).



Descriptions

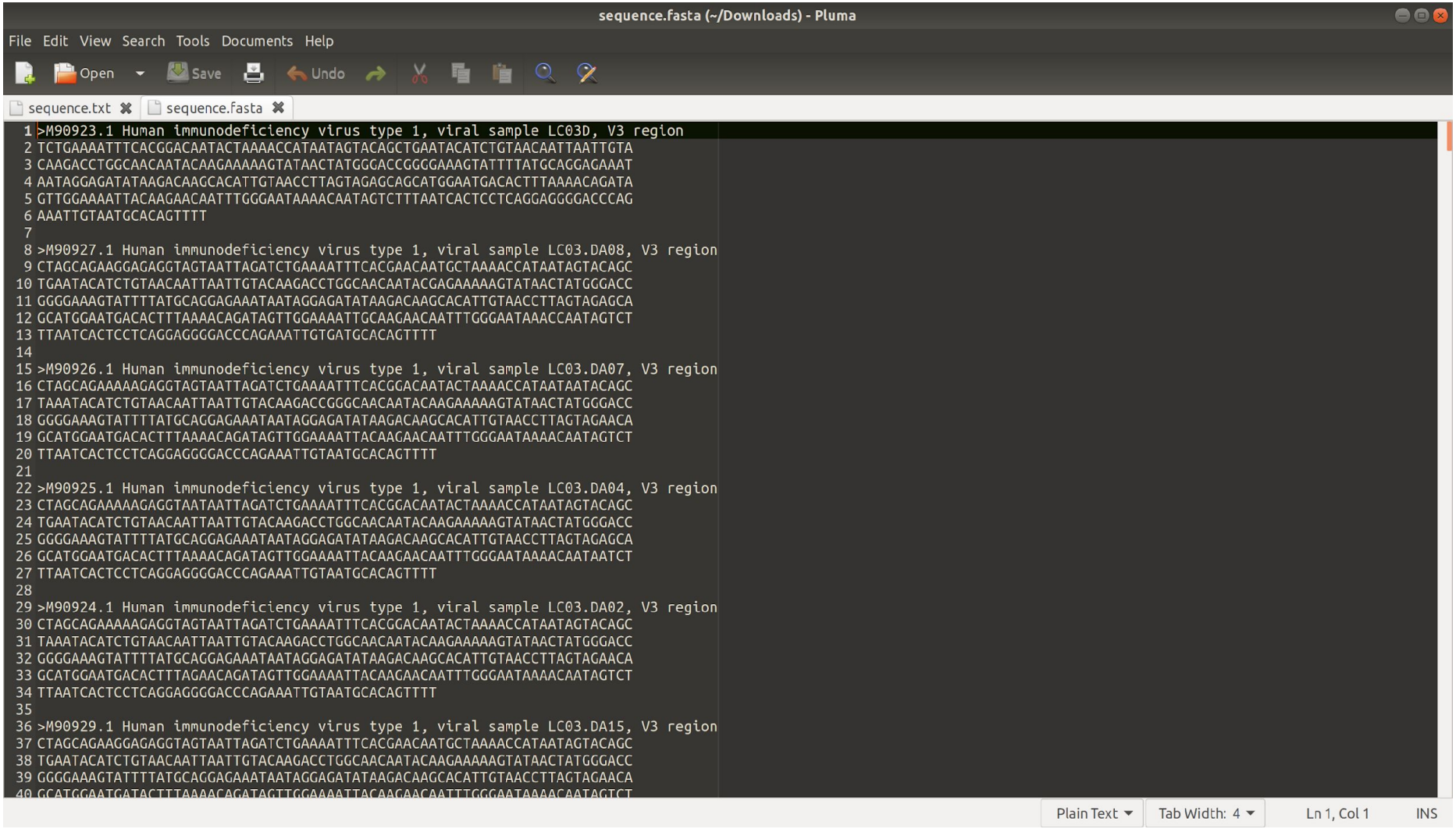
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
Translation elongation factor aEF2 [Lokiarchaeum sp. GC14_75]	1536	1536	100%	0.0	100%	KKK44407.1
elongation factor aEF-2 [Candidatus Lokiarchaeota archaeon CR_4]	946	946	100%	0.0	61%	OLS15932.1
hypothetical protein AM325_09125 [Candidatus Thorarchaeota archaeon SMTZ1-45]	894	894	100%	0.0	59%	KXH72312.1
hypothetical protein AM324_08425 [Candidatus Thorarchaeota archaeon SMTZ1-83]	885	885	100%	0.0	58%	KXH71555.1
Elongation factor 2 [Candidatus Odinarchaeota archaeon LCB_4]	872	872	99%	0.0	57%	OLS17158.1
translation elongation factor aEF2 (CTD), partial [Lokiarchaeum sp. GC14_75]	864	864	58%	0.0	94%	KKK44774.1
Elongation factor 2 [Candidatus Heimdallarchaeota archaeon AB_125]	847	847	100%	0.0	56%	OLS33283.1
elongation factor EF-2 [archaeon ex4484_74]	784	784	100%	0.0	52%	QYT35986.1
putative elongation factor 2 [uncultured organism]	766	766	97%	0.0	52%	AKC94987.1
translation elongation factor aEF2 (NTD), partial [Lokiarchaeum sp. GC14_75]	550	550	38%	0.0	92%	KKK43242.1
elongation factor EF-2 [Thermofilum pendens]	532	532	98%	2e-176	39%	WP_011752282.1
elongation factor EF-2 [Thermofilum uzonense]	532	532	98%	3e-176	39%	WP_052884495.1

2. Assembling a dataset



3. Alignments: hypotheses of homology

ATTGGCG

AGGAG

Sequences!

ATTGGCG

A--GGAG

Alignment!

AGGCG

+TT

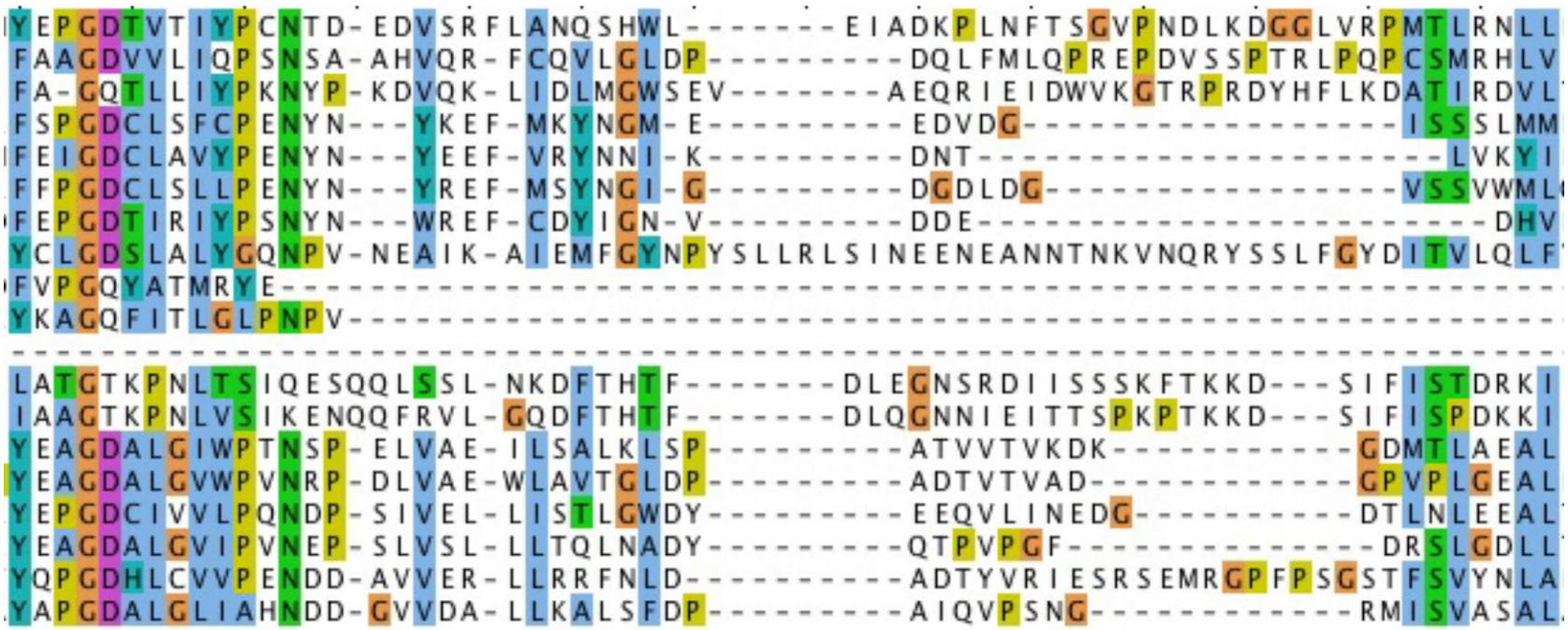
C=>A

ATTGGCG

A--GGAG

Homology!

Some bits are easier to align than others



Homologous!

???

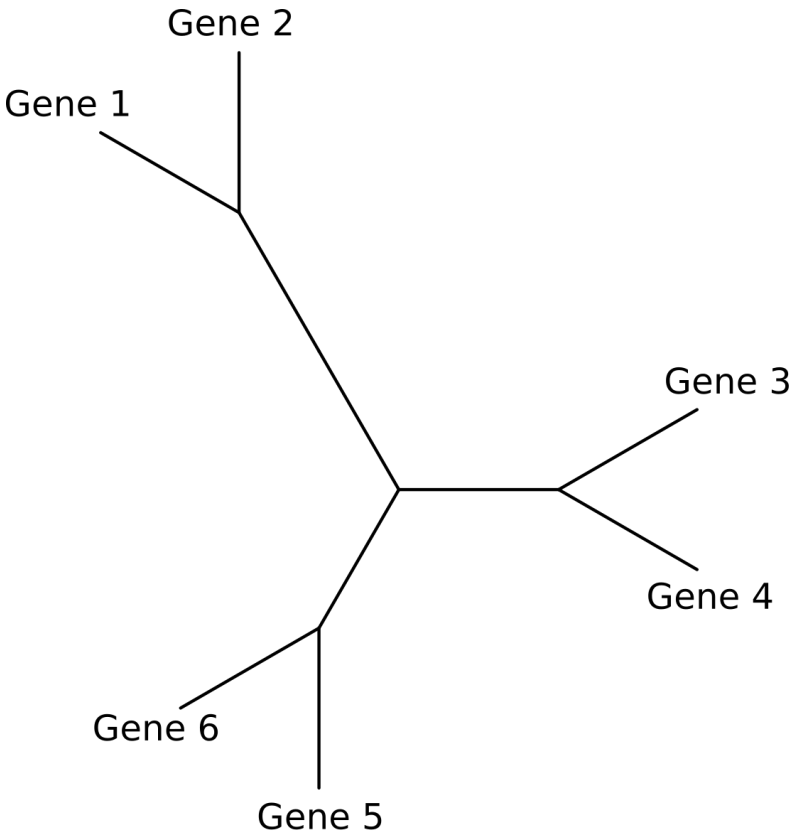
3. Alignments: some considerations

- Tree inference methods assume characters in each alignment column are homologous
- Automated alignment methods often disagree
- The alignment is uncertain, but often fixed in downstream analyses. **It matters!**

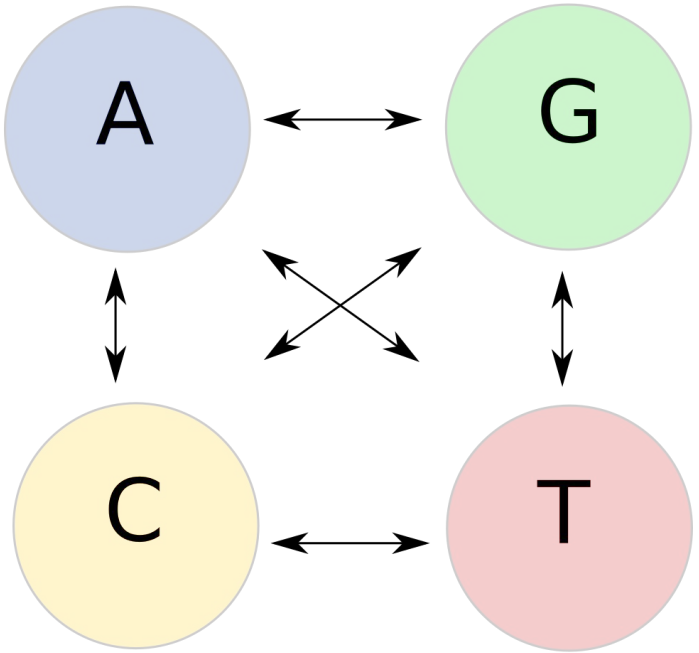
4. Modelling sequence evolution and inferring trees

- Modern phylogenetic analysis is model-based
- Fitting this model to the sequence alignment helps us learn about the parameters of the model (including the tree topology)
- Inference depends on **likelihood**: the probability of the data given the model.

A basic model



(a) Phylogenetic tree
(with branch lengths)

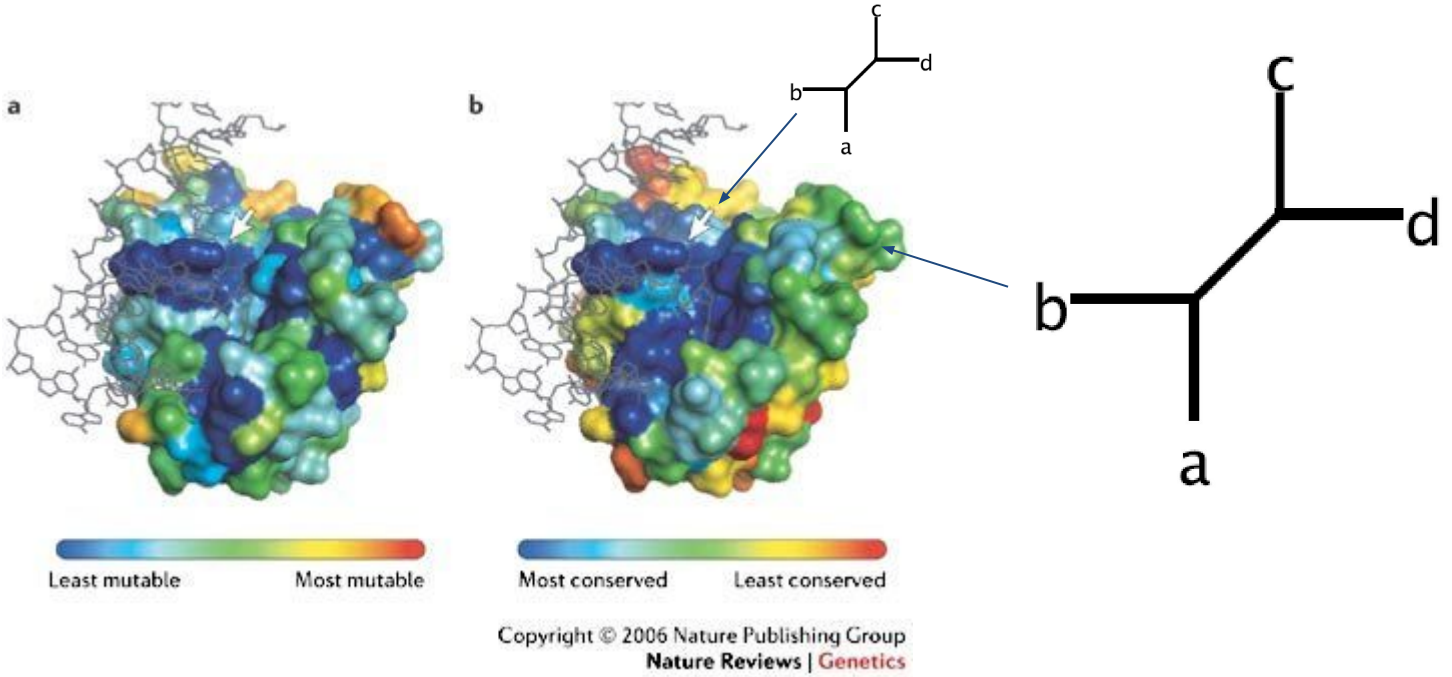


(b) Substitution process
Rates, compositions

Adding complexity (and parameters) to the basic model

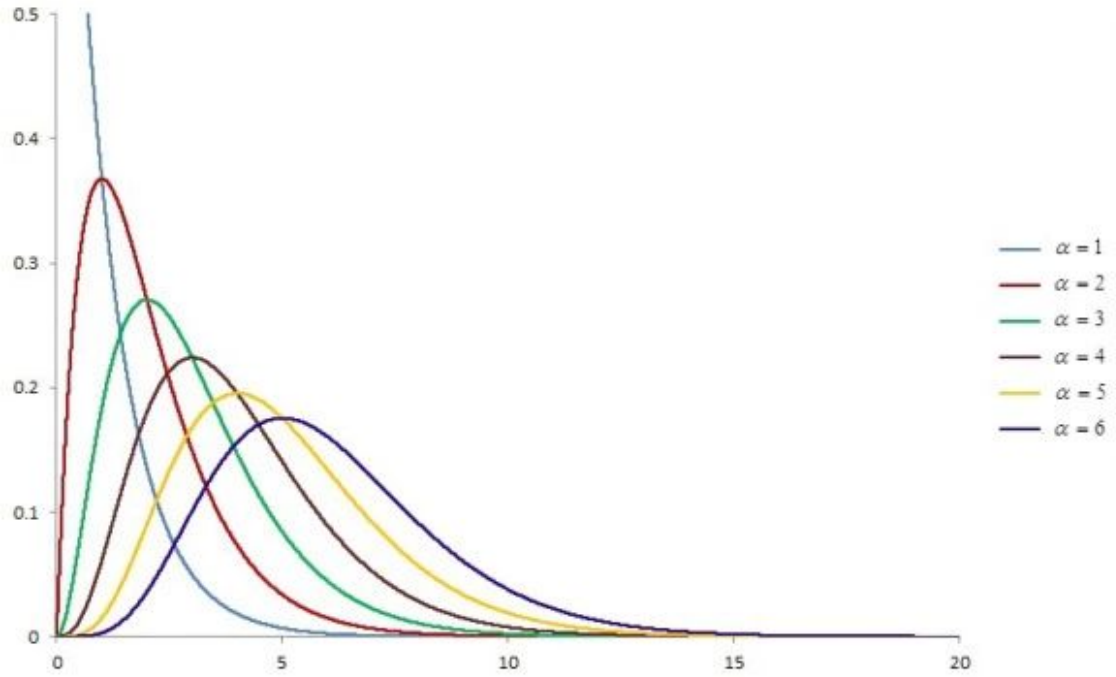
- Different sites evolve at different rates
- Different genes evolve at different rates
- The substitution process might vary across the tree (descent with modification)

Adding complexity: site rates



- Some sites are more important for function than others; they tend to evolve more slowly.
- Model this by having several different rates (tree lengths)

Adding complexity: site rates



- Gamma distribution can take many shapes, depending on a single parameter (alpha). So can model the distribution of rates-across-sites with just one more parameter!
- A ****mixture model****: average the likelihood per site over the possible rates!

Choosing an appropriate model

Likelihood = *P*(alignment | model)

Tree
Branch lengths
Evolutionary rates
Exchangeabilities
etc.

- The evidence we extract from the data totally depends on the model.
- **All models are wrong!**
  ….but some are useful (George Box).

How can we pick a useful model for our data?

Choosing an appropriate model

Likelihood = P(alignment | model)

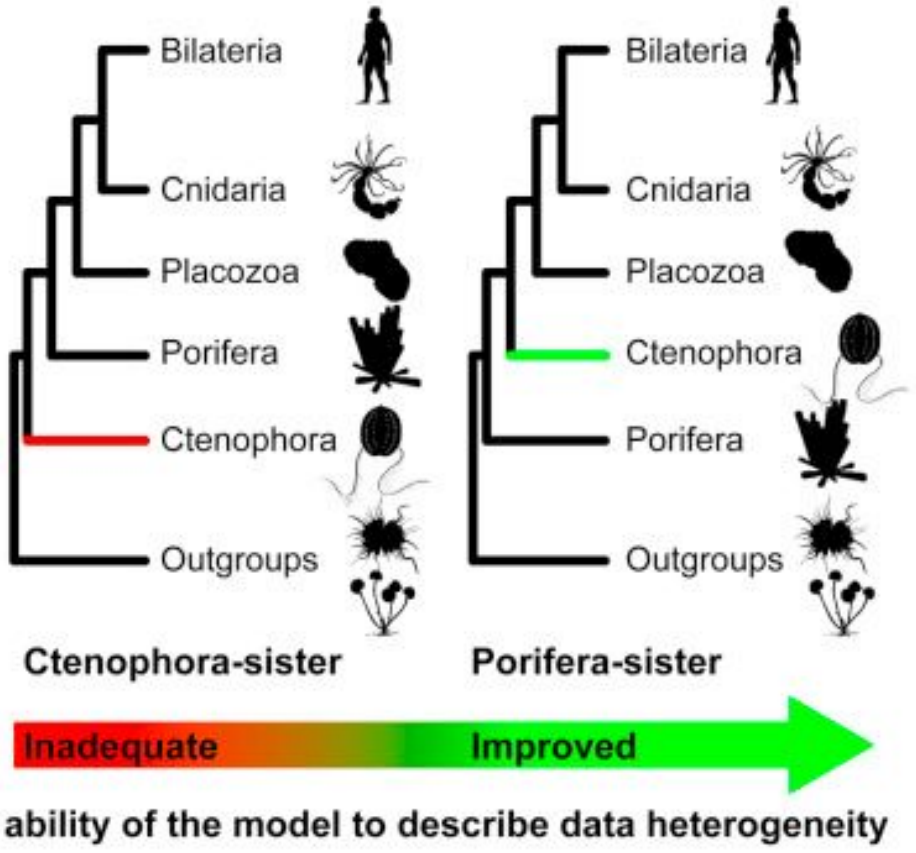
Tree
Branch lengths
Evolutionary rates
Exchangeabilities
etc.

- The likelihood provides a natural way of comparing models.
- Which model makes the data most likely?
- Subtle difficulties of maximum likelihood: model fit is not quite so easy. Use scores that are modifications of the maximum likelihood.

Model selection

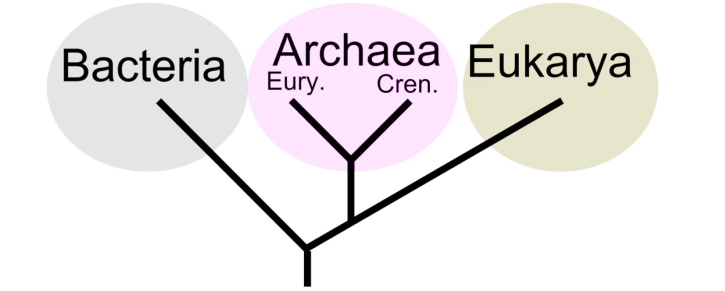
- Phylogeny packages implement a set of candidate models.
- Use statistical tests (AIC, BIC - based on the maximum likelihood, penalised for the number of parameters) to pick **the best model from that set** for your data.
- It's possible that none of the implemented models is adequate for your data (use simulations?).
- ***Many published analyses are of very poor quality: **be critical and don't fall into this trap.**

Model selection matters: nervous system origins

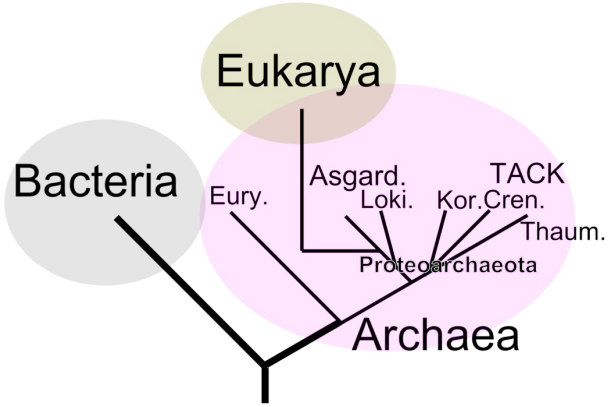


Feuda et al. (2017) *Curr Biol;* Whelan et al. (2017) *Nat Ecol Evol*

Model selection matters: The tree of life



Carl Woese's three domain tree



two domain tree

Cox et al. (2008) PNAS, Williams et al. (2012) PRSB

Fitting models: maximum likelihood vs. Bayesian inference

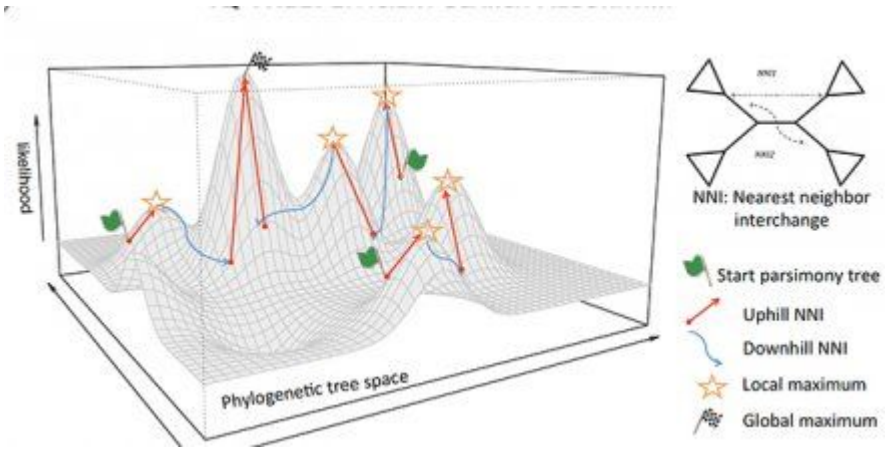
- **To a first approximation, the fit of the model is more important!**

Maximum likelihood

Task: find values for all of the parameters of the model that make the probability of the alignment (likelihood) as high as possible.

ML programs implement (hill climbing) algorithms for joint estimation of these values.

The ML tree is the best point estimate of the true tree, if the model is correct.



Maximum likelihood: assessing support

- ML tree is best estimate of the true tree. But is it much better than any other estimate? (Hypothesis testing, e.g. **AU-test**)
- How certain are we of the relationships on the tree? Are some better supported than others? (**Bootstrapping**)

Maximum likelihood: assessing support with the bootstrap

Bootstrapping is a technique for estimating the confidence interval around a parameter by **resampling and reanalysis** of the original data (alignment).

1. Sample, with replacement, the same number of characters as the original dataset.
2. Repeat tree inference on each of many bootstrap replicates.
3. Assess confidence as the proportion of samples in which each split on the tree appears.

Taxon	1	2	3	4
A	V	W	R	A
B	V	W	R	L
C	I	Y	K	M
D	M	F	K	A

Original dataset (4 characters)

Taxon	1	1	3	2
A	V	V	R	W
B	V	V	R	W
C	I	i	K	Y
D	M	M	K	F

One bootstrap dataset

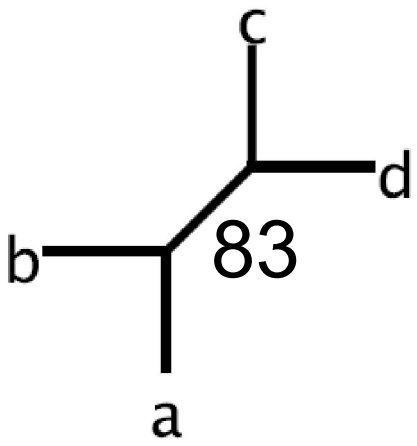
Maximum likelihood: assessing support with the bootstrap

Bootstrapping is a technique for estimating the confidence interval around a parameter by **resampling and reanalysis** of the original data (alignment).

1. Sample, with replacement, the same number of characters as the original dataset.
2. Repeat tree inference on each of many bootstrap replicates.
3. Assess confidence as the proportion of samples in which each **split** on the tree appears.

Taxon	1	2	3	4
A	V	W	R	A
B	V	W	R	L
C	I	Y	K	M
D	M	F	K	A

Original dataset (4 characters)



Maximum likelihood: assessing support with the bootstrap

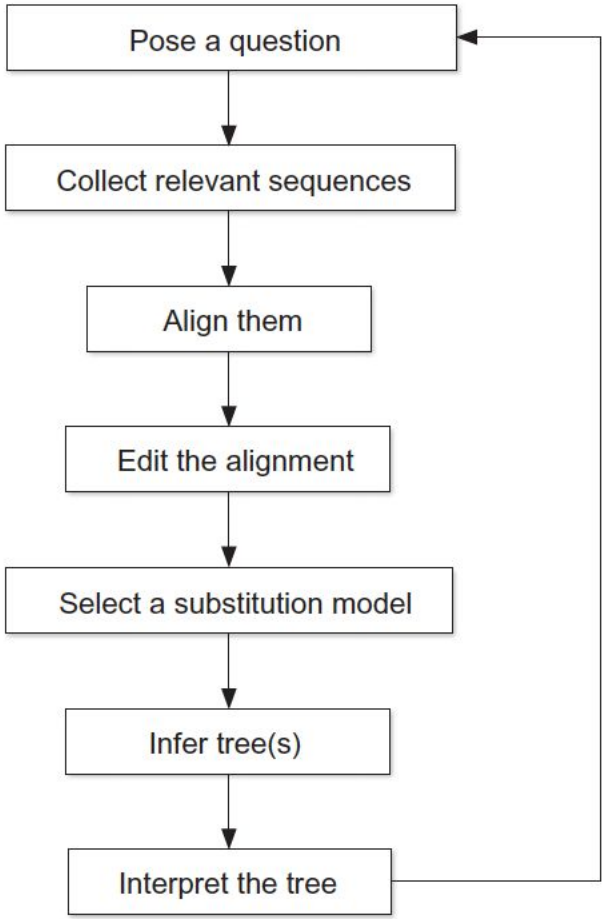
Interpreting bootstrap values:

- ***The bootstrap is not the probability of a split being correct!**
- Early on, some justifications by comparison to bootstrapping confidence intervals in other settings (v**ariance of bootstrap trees around ML tree ~ variance of ML trees around true tree**)
- These days, mostly thought of as a way of getting at the robustness of the signal for relationships in the alignment.
- Values of >70 usually taken as being somewhat reliable.

Fitting phylogenetic models: summary

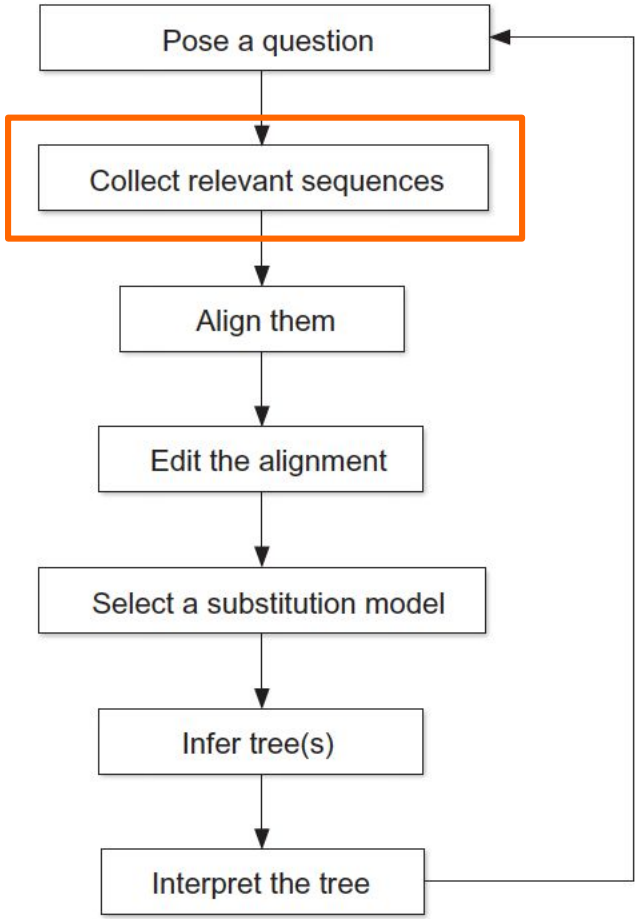
- Phylogenetic trees are statistical inferences
- The inferences are based on a model: it's important, so do experiments!
- Quantifying the uncertainty in our inferences is critically important for interpretation
- ML & Bayesian methods both provide a framework for assessing support

Molecular phylogenetics: a possible flowchart



A critical approach to analysis is key: don't fall into the black box!

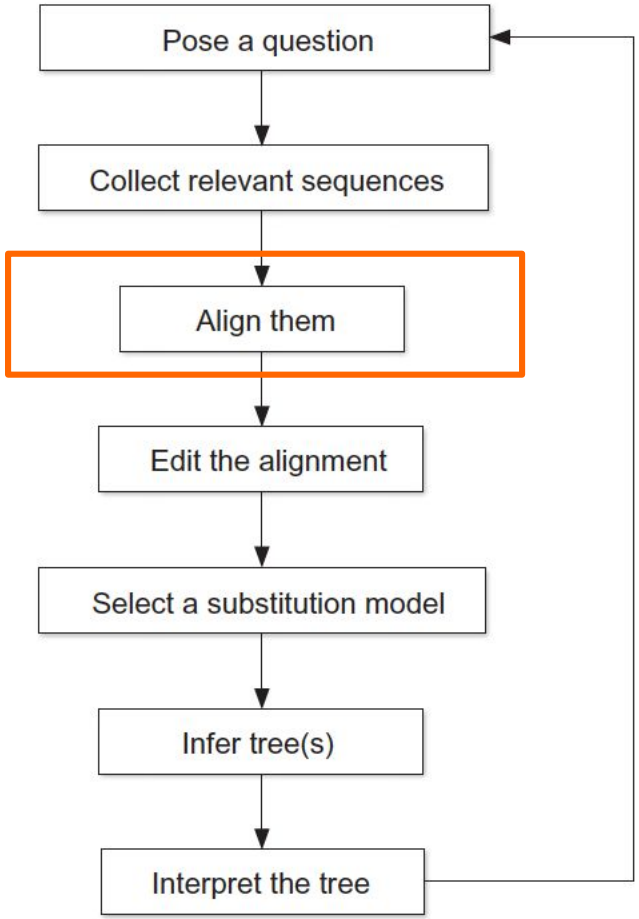
Molecular phylogenetics: a possible flowchart



Did I miss any important ones?

A critical approach to analysis is key: don't fall into the black box!

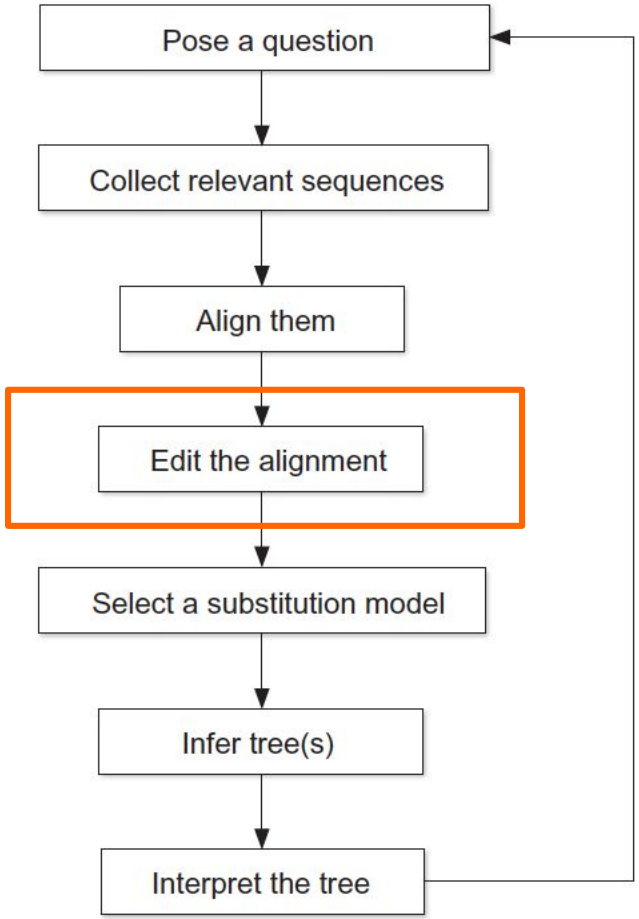
Molecular phylogenetics: a possible flowchart



How certain is my alignment?
Would other alignments give a different tree?

A critical approach to analysis is key: don't fall into the black box!

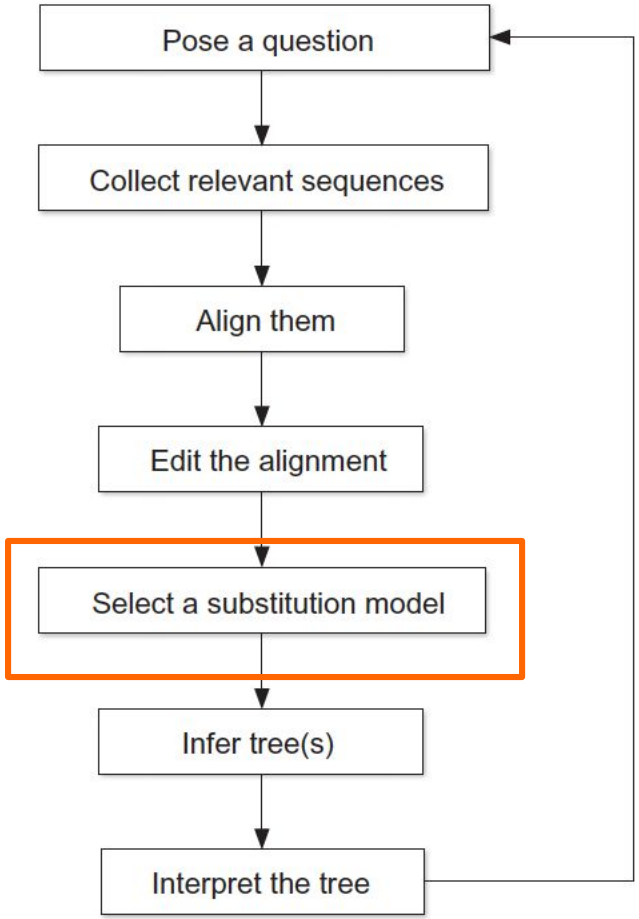
Molecular phylogenetics: a possible flowchart



Which bits, if any, should I remove?

A critical approach to analysis is key: don't fall into the black box!

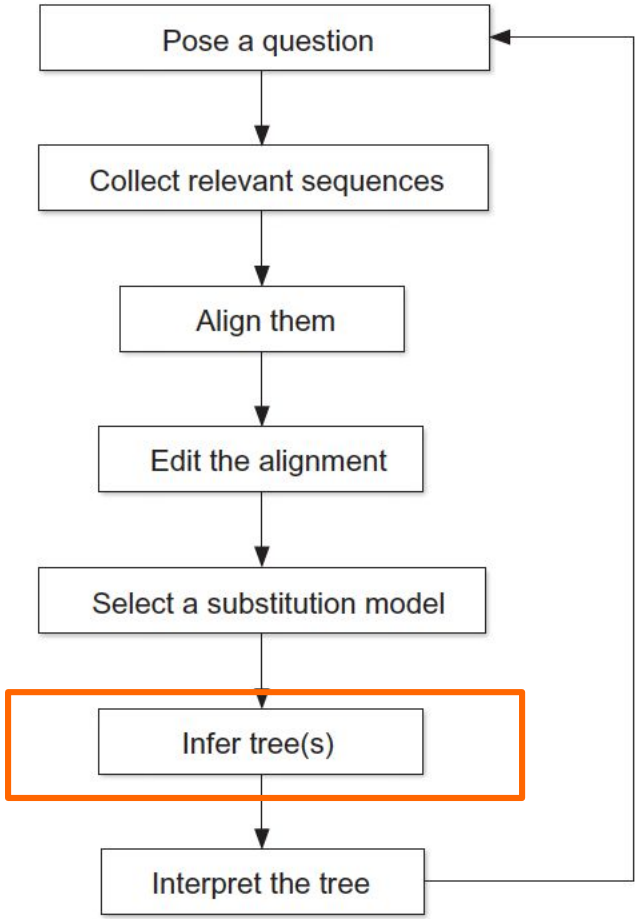
Molecular phylogenetics: a possible flowchart



Is this the best available model?
Is it good enough?

A critical approach to analysis is key: don't fall into the black box!

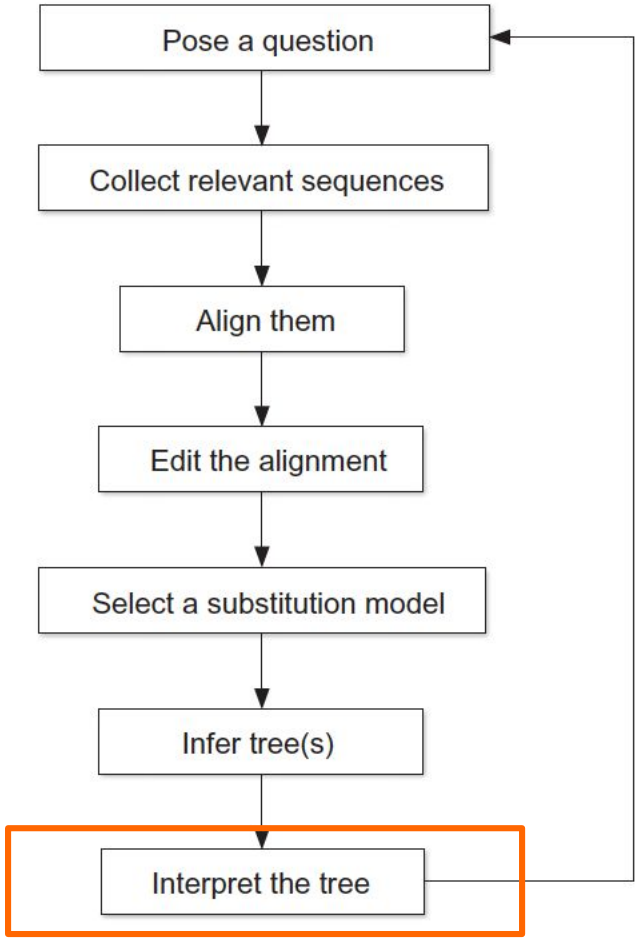
Molecular phylogenetics: a possible flowchart



Are my ML or Bayesian analyses working properly? (Local maxima, convergence)

A critical approach to analysis is key: don't fall into the black box!

Molecular phylogenetics: a possible flowchart



How well-supported are the key branches?

What does the tree mean for my hypothesis?

Can alternative hypotheses be rejected?

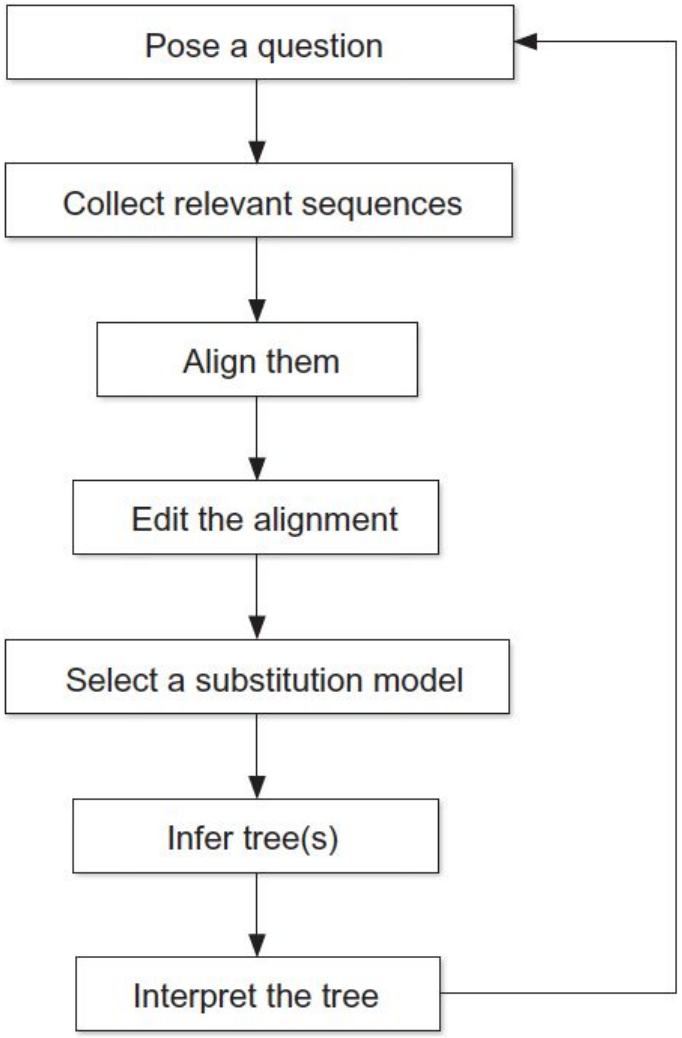
A critical approach to analysis is key: don't fall into the black box!

Phylogenetics: some key questions

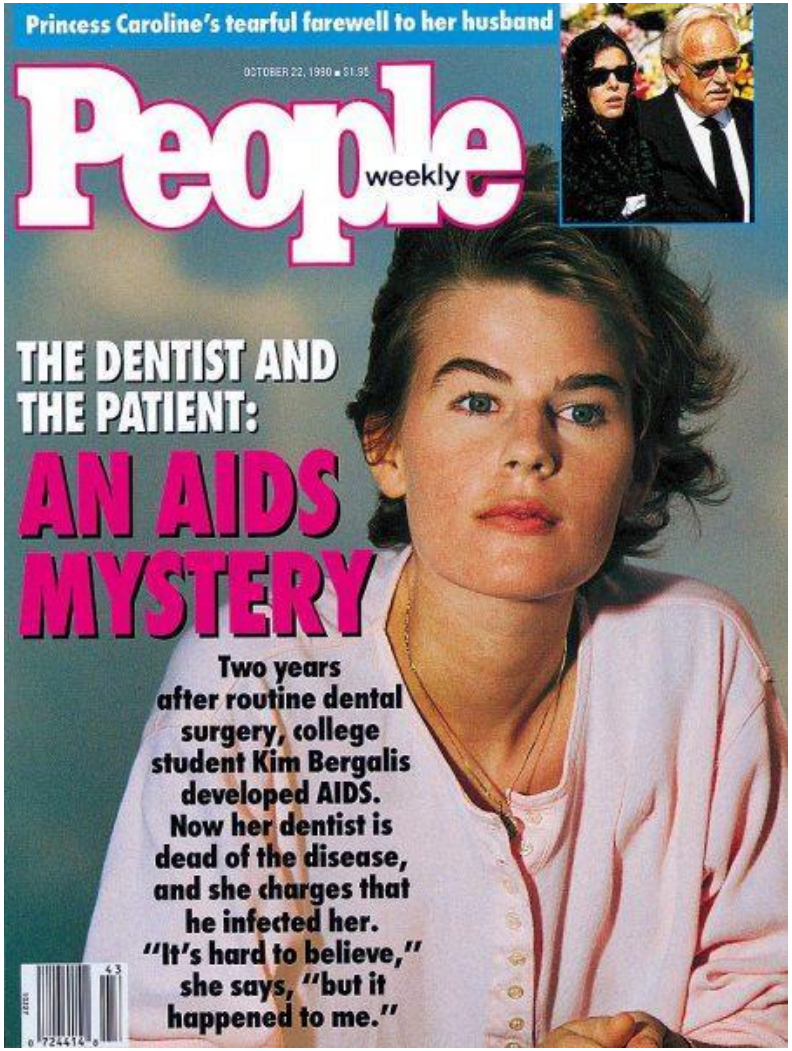
- Did I analyse all the relevant data?
- Does the model fit the data?
- How well-resolved is the tree? Can alternative hypotheses be rejected?
- What do these analyses mean for my question?

Molecular phylogenetics practical

Practical: A first phylogenetic analysis



Dentist-patient transmission of HIV?



In the early 90s, a dentist was accused of infecting several of his patients with HIV during surgical procedures.

After a "low-risk" patient was diagnosed with HIV, other patients were screened. 10 had HIV.

Did the dentist infect them? We will do a phylogenetic analysis to evaluate these claims.

The data

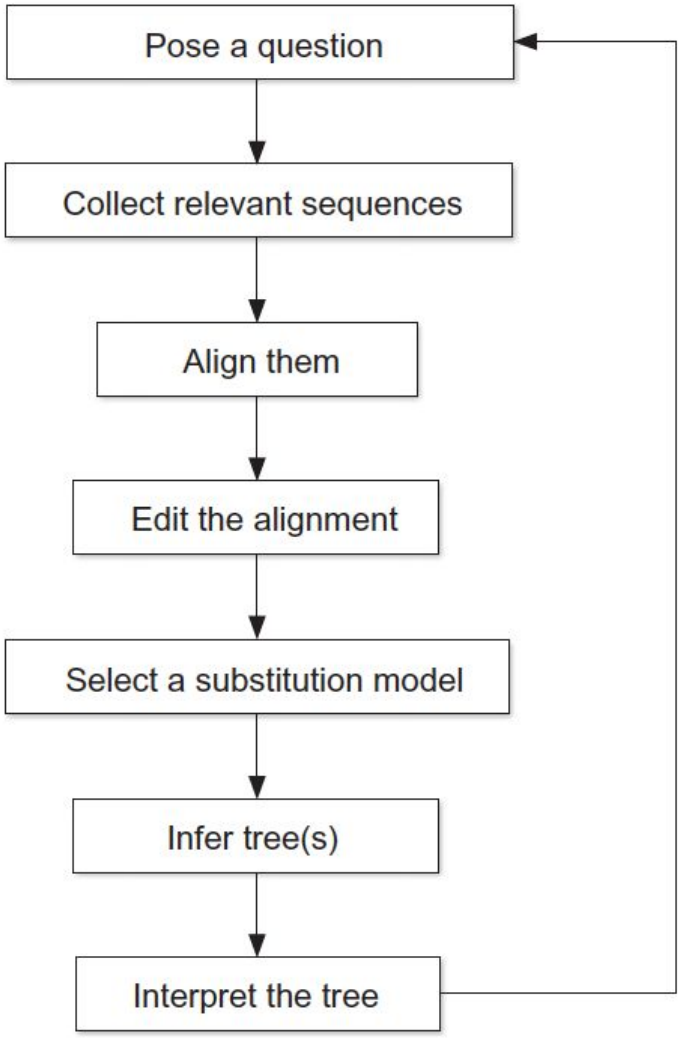
Molecular Epidemiology of HIV Transmission in a Dental Practice

Chin-Yih Ou, Carol A. Ciesielski, Gerald Myers, Claudiu I. Bandea, Chi-Cheng Luo, Bette T. M. Korber, James I. Mullins, Gerald Schochetman, Ruth L. Berkelman, A. Nikki Economou, John J. Witte, Lawrence J. Furman, Glen A. Satten, Kersti A. MacInnes, James W. Curran, Harold W. Jaffe, Laboratory Investigation Group,* Epidemiologic Investigation Group†

HIV *env* sequences from:

- The dentist
- His patients
- Local controls

The workflow



Q: Are patient sequences *descended* from the dentist's sequences?

Let's use NCBI…

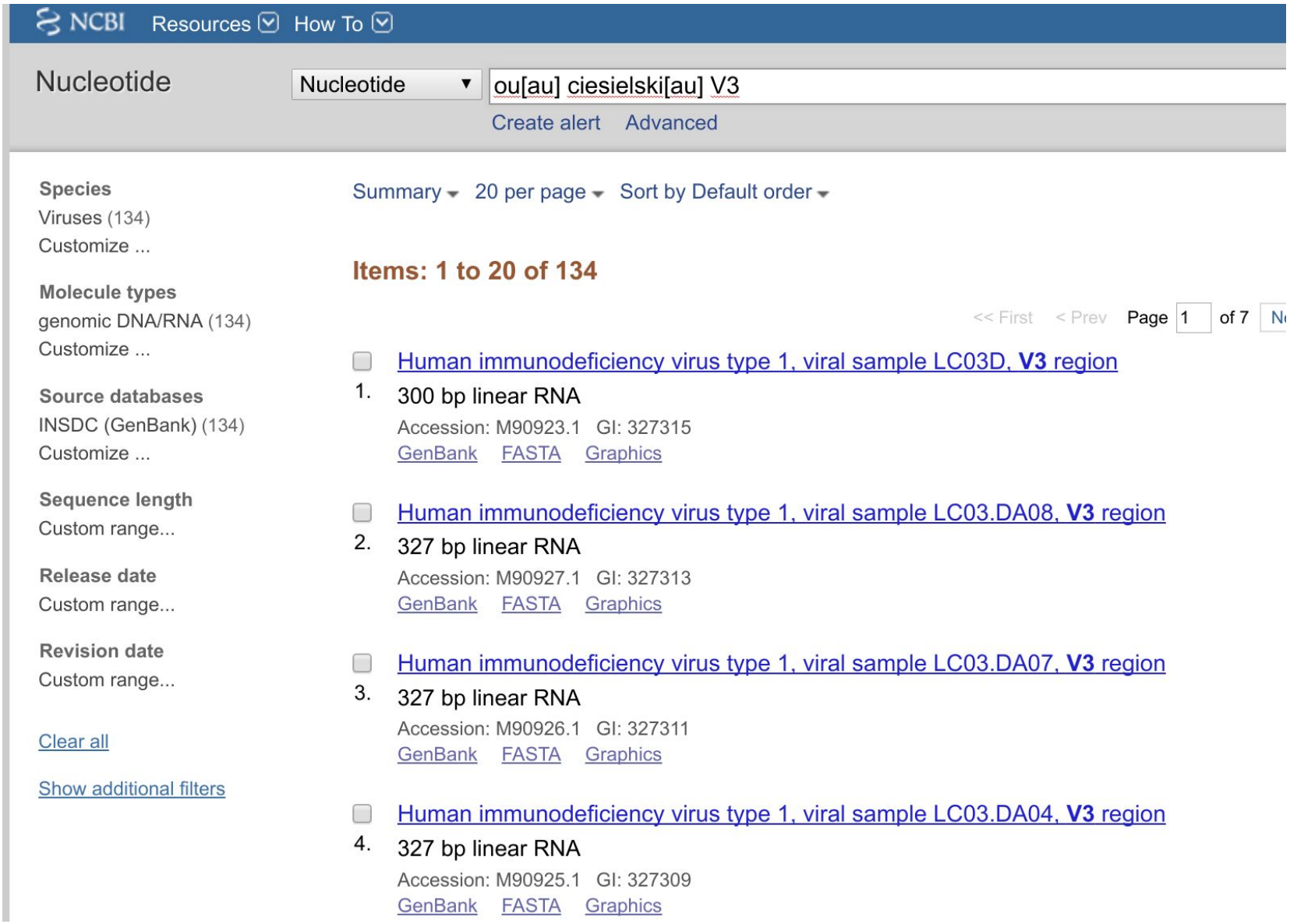
We'll use mafft.

***

We'll use the best-fitting model in IQ-Tree...

Does the phylogeny support the claims?

How to retrieve the data?

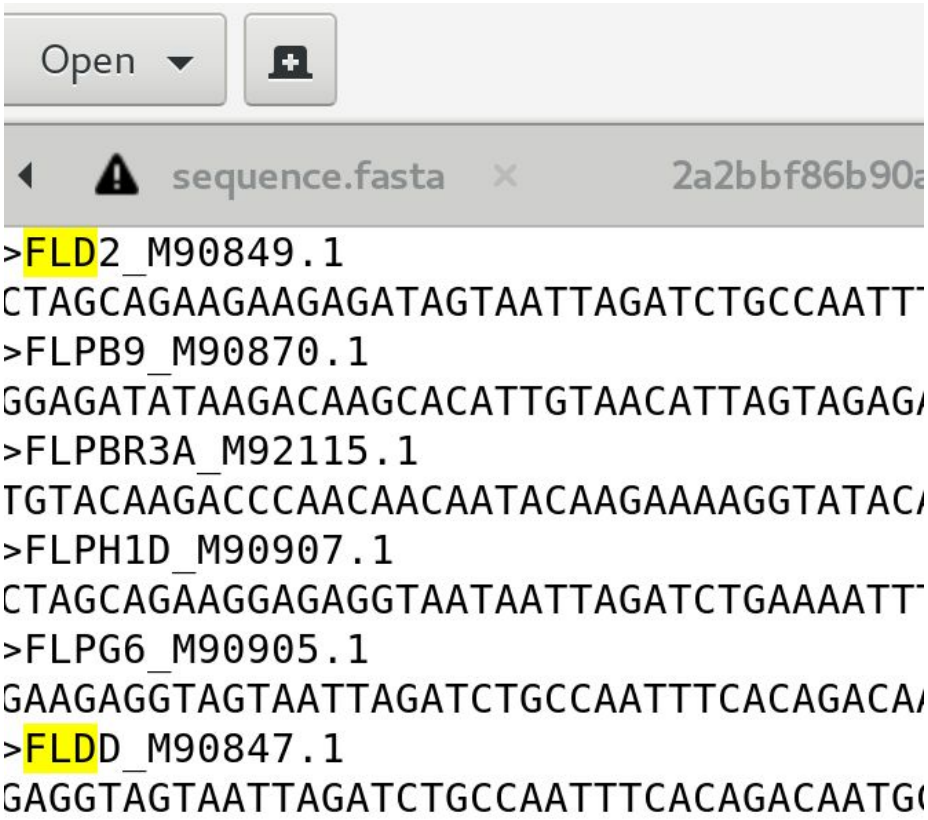


How to retrieve the data?

- **Dentist sequences: FLD1,2,4,5,7,8**
- **Patient sequences: FLP[A-H]: take 3 isolates/patient**
- **Local controls (LCXX, FLQ): take at least 10.**

The workflow

- **Download the sequences in FASTA format, save in a text file.**
- **Rename with**
- **sensible tags**
- **FASTA format:**



Some useful commands

Alignment:

```
mafft --auto mySequences > myAlignment
```

View the alignment in `belvu/Jalview`.

Tree inference (black box):

```
iqtree -s myAlignment -m MFP -bb 1000
```

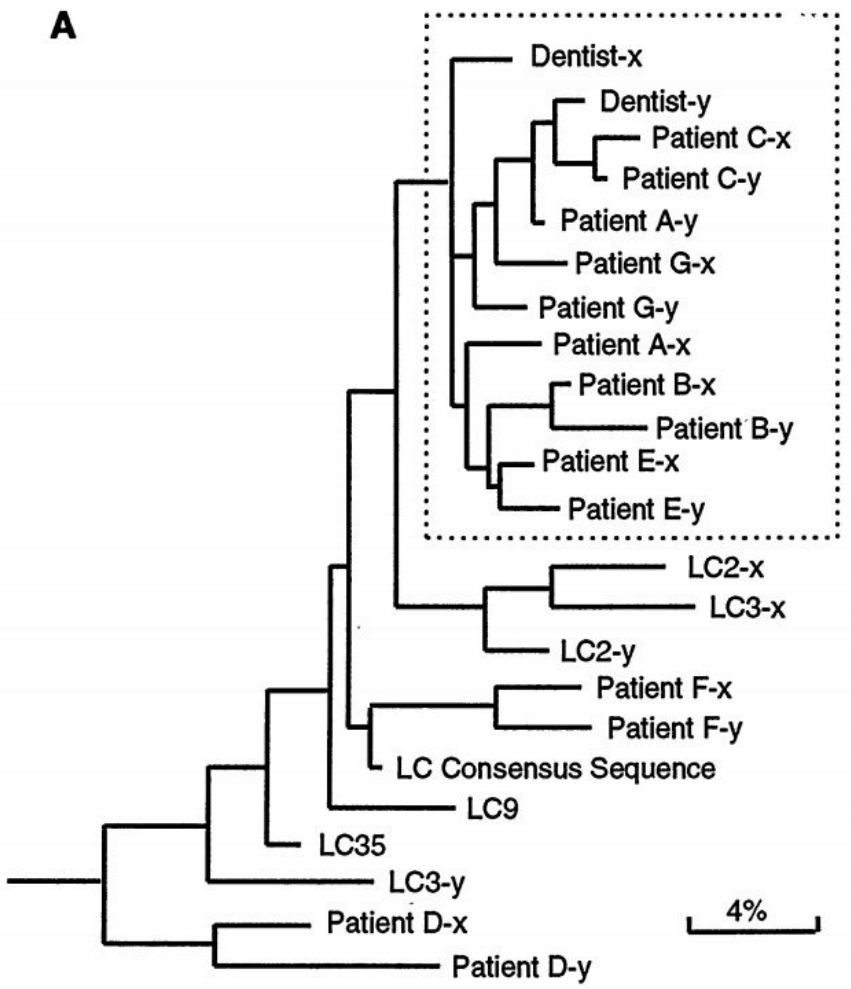
Questions

1. Did the patients get their viruses from the dentist?
2. How much would you bet on it?
3. Why include local controls in the analysis?

For a bonus practical, see:

`https://github.com/Tancata/bioinf_workshop`

Ou et al. (1992)


A
Dentist-x
Dentist-y
Patient C-x
Patient C-y
Patient A-y
Patient G-x
Patient G-y
Patient A-x
Patient B-x
Patient B-y
Patient E-x
Patient E-y
LC2-x
LC3-x
LC2-y
Patient F-x
Patient F-y
LC Consensus Sequence
LC9
LC35
LC3-y
Patient D-x
Patient D-y
4%

Some properties of the patients

Table 1. Dental cohort clinical information and HIV nucleotide variation in the C2-V3 domain of the envelope gene.

Person	Sex	Known risk factor	Clinical status*	M13 clones (no.)	Intraperson variation† (%)	Interperson variation† (%)	
						To dentist	To 30 LCs‡
Dentist	M	Yes	AIDS	6	3.3 (0.8–5.4)		11.0 (5.8–16.0)
Patient A	F	No	AIDS	6	2.0 (0.0–4.5)	3.4 (0.8–6.2)	10.9 (5.4–14.8)
Patient B	F	No	Asymptomatic (CD4 = 222/μl)	12	1.9 (0.4–3.7)	4.4 (2.1–7.0)	11.2 (6.2–16.5)
Patient C	M	No§	Asymptomatic (CD4 = <50/μl)	5	1.2 (0.4–1.6)	3.4 (2.1–4.9)	11.1 (7.0–15.6)
Patient E	F	No	Asymptomatic (CD4 = 567/μl)	6	2.1 (0.4–3.7)	3.4 (1.2–6.6)	10.8 (5.8–14.8)
Patient G	M	No	Asymptomatic (CD4 = 400/μl)	5	2.8 (1.6–3.7)	4.9 (2.9–7.0)	11.8 (6.2–16.9)
Patient D	M	Yes	AIDS	5	7.5 (0.0–9.9)	13.6 (11.5–15.6)	13.1 (7.8–17.3)
Patient F	M	Yes	Asymptomatic (CD4 = 253/μl)	6	3.0 (0.8–5.8)	10.7 (8.2–13.6)	11.9 (7.0–17.3)

Some further reading

Papers:

- Yang and Rannala (2008) “Molecular phylogenetics: principles and practice”. *Nat Rev Genet*
- Holder and Lewis (2003) “Phylogeny estimation: traditional and Bayesian approaches.” *Nat Rev Genet*

Books:

- Bromham L. “An introduction to molecular evolution and phylogenetics.”
- Yang Z. “Computational molecular evolution: a statistical approach.”