

Introduction to Phylogenetics

Martin Maiden & Keith Jolley,
Department of Biology



UNIVERSITY OF
OXFORD

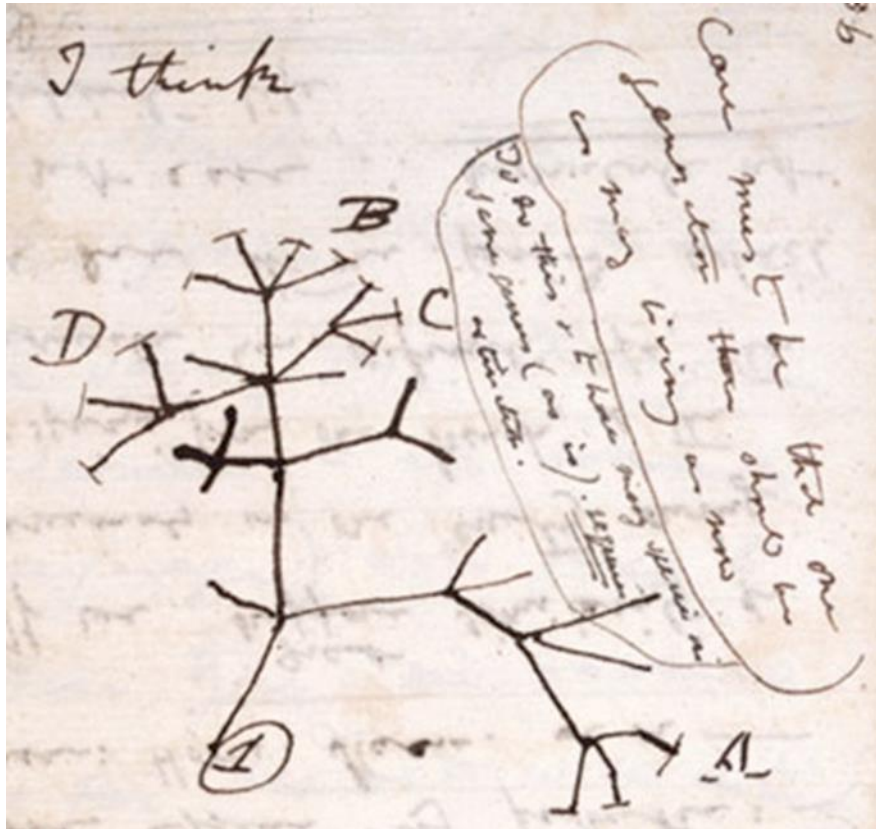
Learning outcomes

1. Define phylogenetics and describe how it relates to taxonomy and typing.
2. Introduce phylogenetic terminology and the concept of homology.
3. Describe molecular phylogenetics and mutational processes.
4. A practical exercise to illustrate the components of constructing a phylogeny: (i) alignment; (ii) calculating genetic distances; (iii) building a tree; (iv) assessing tree quality and phylogenetic uncertainty.

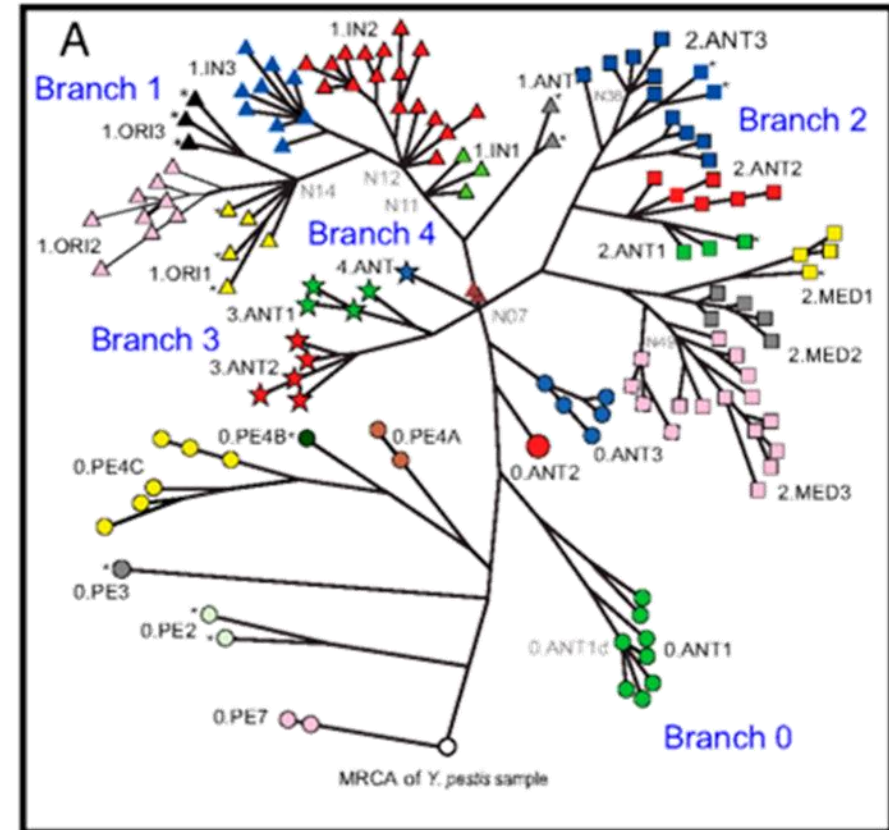
References

- **Yang, Z., and Rannala, B.** (2012) Molecular phylogenetics: principles and practice. *Nat Rev Genet* **13**, 303-314
- **Baldauf, S. L.** (2003). Phylogeny for the faint of heart: a tutorial. *Trends Genet.* **19**, 345-351.
- Online course:
 - <https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics/what-is-phylogenetics/>
- Free software – Windows, Mac OS, and LINUX (25th anniversary!):
 - <https://www.megasoftware.net/>

I think ... of trees ...

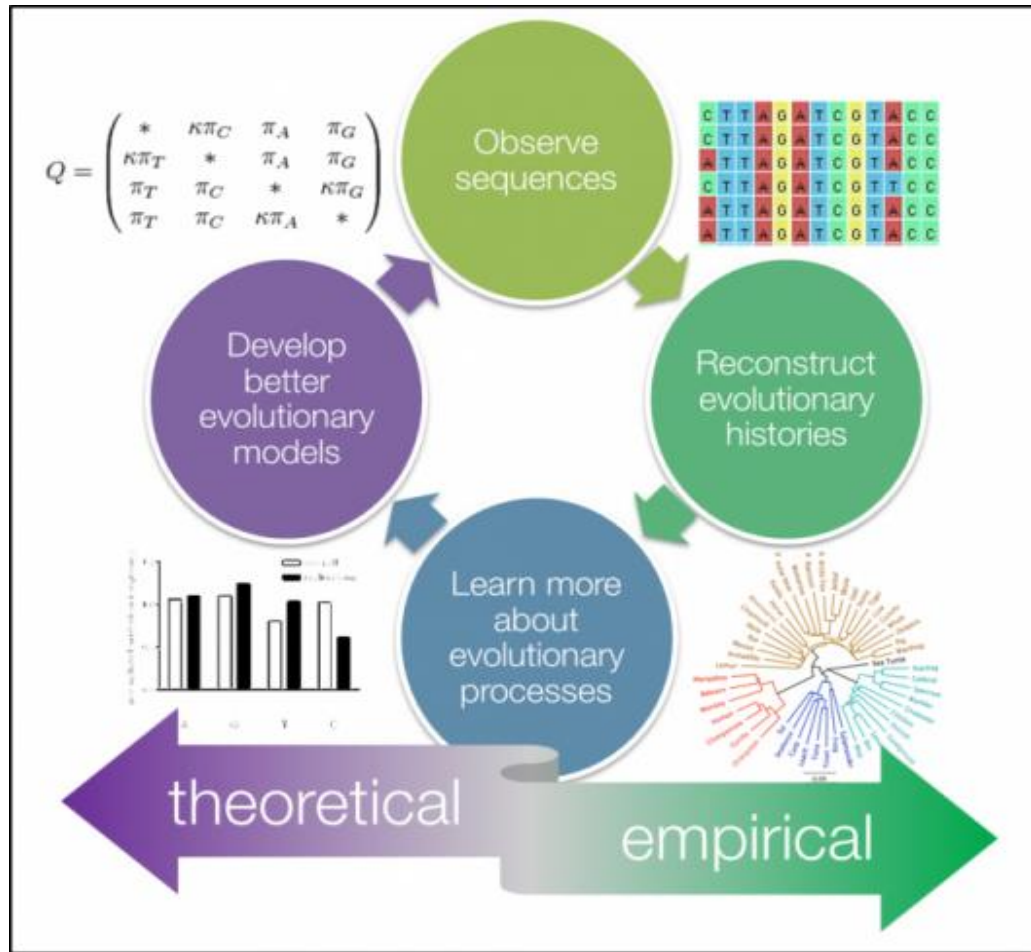


Darwin, C.R., Notebook B



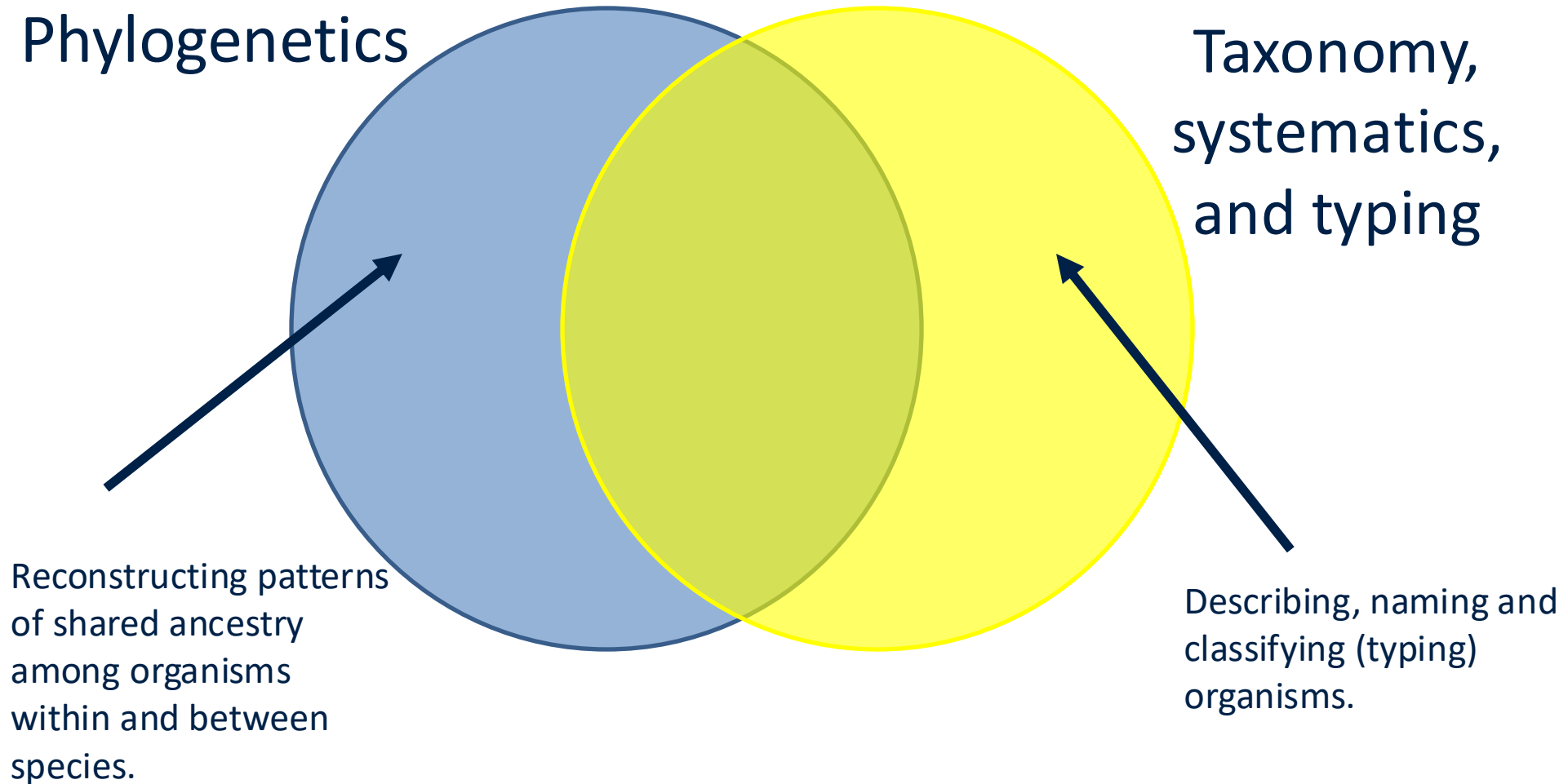
Cui, Y., *et al.* (2013). Historical variations in mutation rate in an epidemic pathogen, *Yersinia pestis*. *Proc. Natl. Acad Sci USA* **110**, 577-582.

What is phylogenetics?

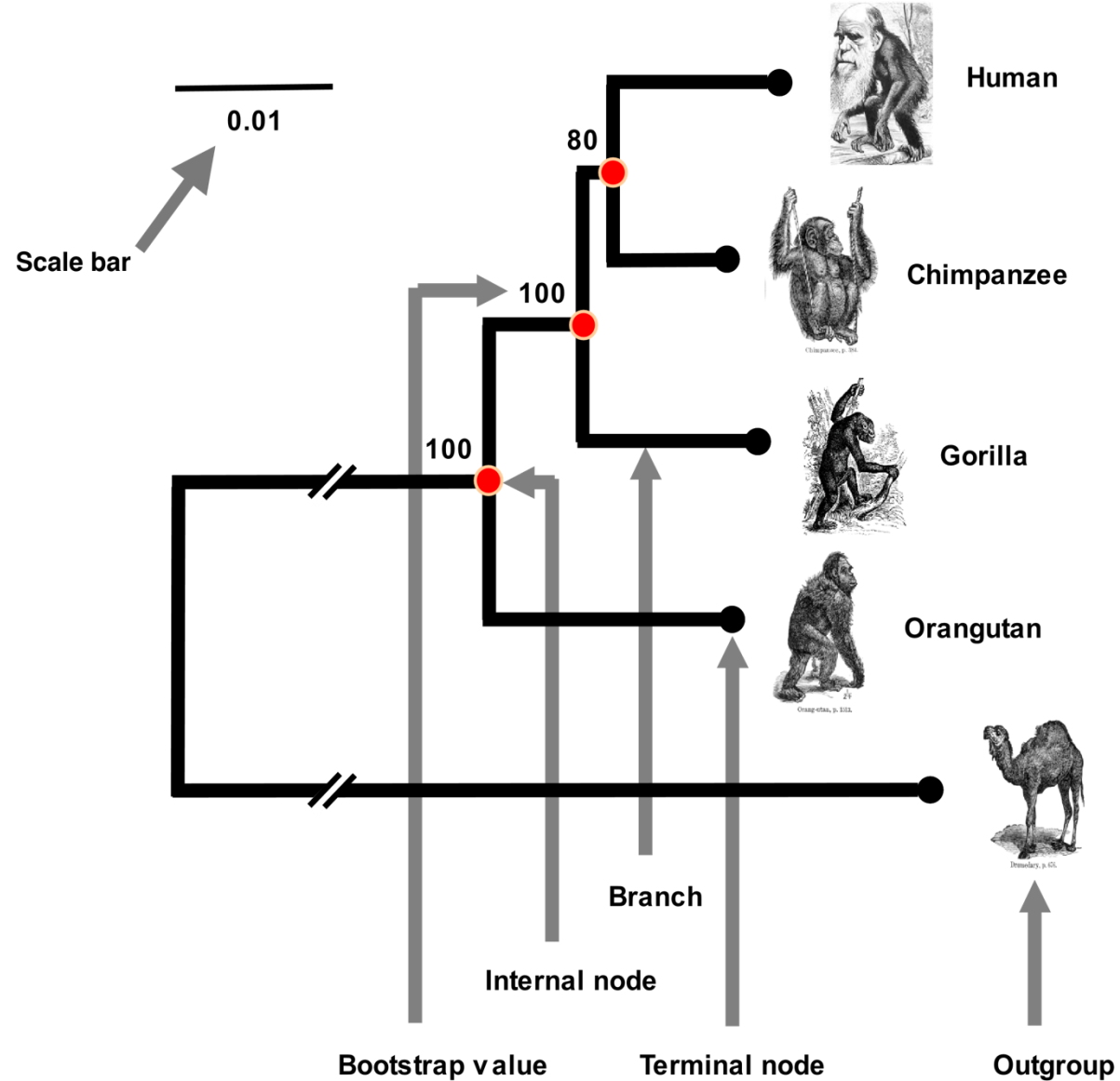


- “Phylogenetics is the study of evolutionary relationships among biological entities, often species, individuals or genes, which may be referred to as taxa.”
- In clinical microbiology phylogenetics (tree drawing) is used to infer relationships among pathogens, to:
 - investigate spread;
 - monitor the emergence of new phenotypes e.g. virulence, AMR.

Phylogenetics, Taxonomy, systematics, and typing

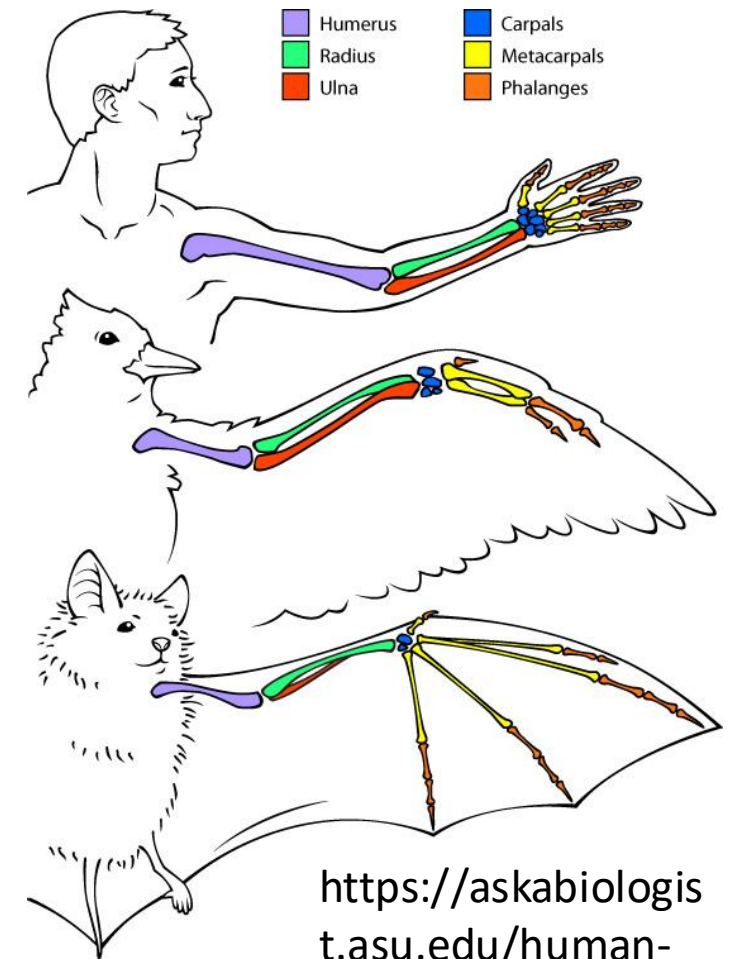


Phylogenetic terminology



Homology

- Phylogenetics is based on the principle of **homology**.
 - Characteristics of organisms are homologous if they are similar *and* have descended from a common ancestor.
- Characteristics are **analogous** if they are similar but have descended from different ancestors.
 - Bird and bat wings are homologous when considered as forelimbs, but analogous as wings.



<https://askabiologist.asu.edu/human-bird-and-bat-bone-comparison>

Molecular Phylogenetics

- Molecular sequences contain information about the evolutionary processes that produced them.
 - But this information is often scrambled, fragmentary, hidden or lost.
 - Molecular phylogenetics uses mathematical and statistical methods to recover and interpret this information.
- Different types of sequence comparisons reveal information about different evolutionary processes:
 - **Orthologous** sequences are from different species,
 - *speciation and extinction;*
 - **Homologous** sequences are from the same species,
 - *population genetic processes (selection, population dynamics, migration);*
 - **Paralogous** sequences are different genes in the same genome
 - *gene duplication and deletion.*

Molecules and morphology

- Molecular characters – protein and DNA sequences – arrived during the molecular biology revolution in the mid-20th century.
- Molecular characters have many advantages over morphological ones, they are:
 - very common;
 - objective, easy to quantify;
 - available when morphology is uninformative (micro-organisms);
 - can be generated quickly and inexpensively;
 - produced with generic, rather than specific, techniques.
- Molecular sequences have one significant disadvantage they are,
 - generally available for extinct species, but increasingly analysed from ‘ancient’ DNA.
- Phylogenetics progressed by investigating cases where molecular and morphological data initially disagreed.

There are only three types of mutation

1. Point mutation (small-scale change)

ATG GCT ATC GAC GAA AAC → ATG GC A ATC GAC A AA AAC

2. Insertion or deletion (small or large-scale)

ATG GCT ATC GAC GAA AAC ↔ ATG GCA ATC GAC AAA AAC

3. Rearrangement (small or large-scale)

ATG GCT ATC GAC GAA AAC ↔ ATG GAT TGC GAC AAA AAC

Mutations: Transitions and transversions

Transition mutation, purine-to-purine or pyrimidine-to-pyrimidine:



Transversion mutation, purine-to-pyrimidine:



Silent/synonymous mutation, encoded amino acid is unchanged.

Replacement/non-synonymous mutation, encoded amino acid is changed

Genes and alleles

ATG **ATT** **GAC** GAC GAA AAC

M **I** **D** D E N



ATG GCT ATC GAC GAA AAC

M A I D E N

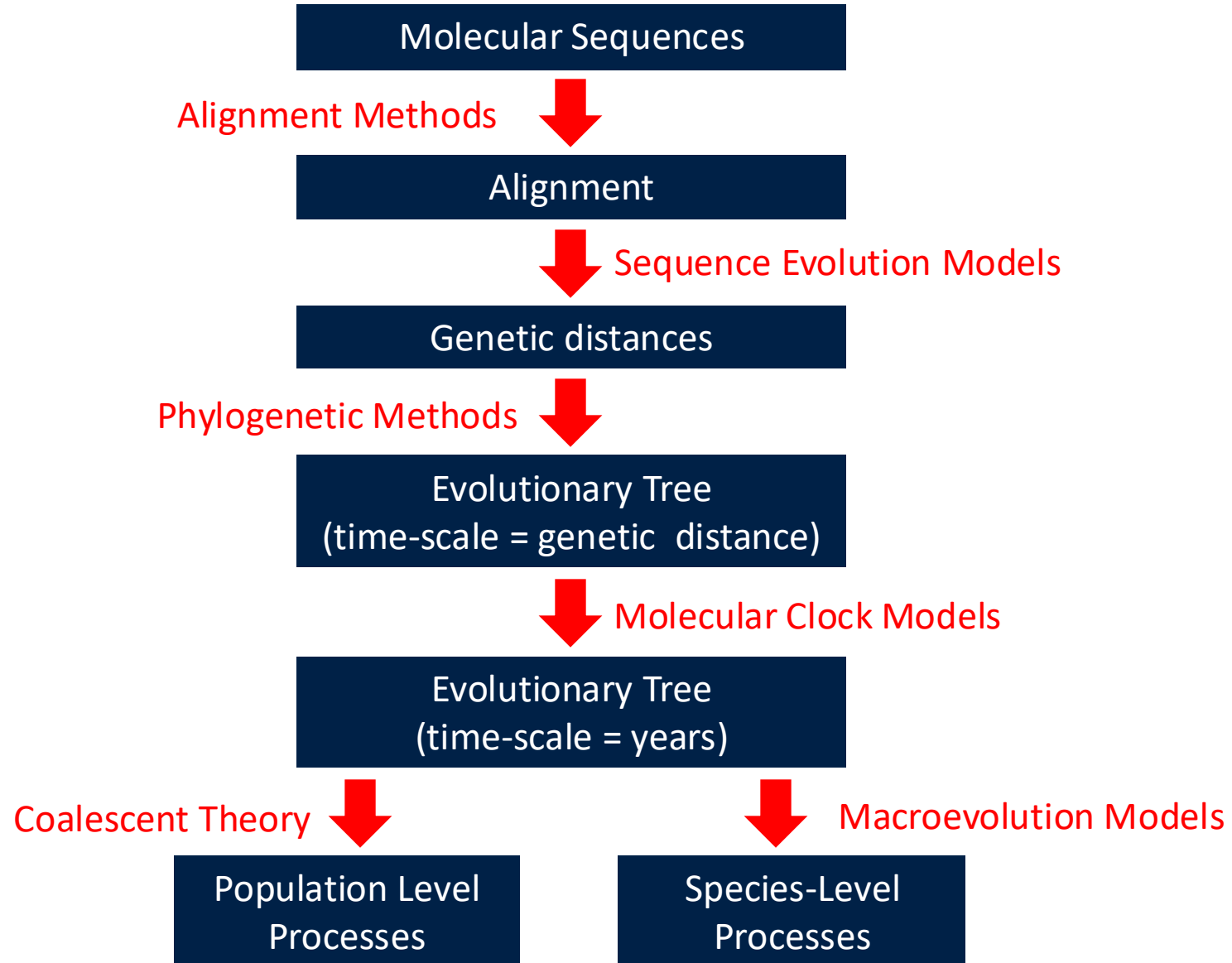


ATG GCT **CGC** **ACC** **ATA** AAC

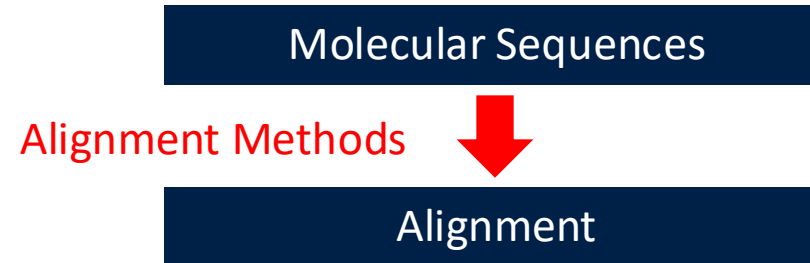
M A **R** **T** **I** N



How to build a tree



How to build a tree



Alignment

- Molecular sequence alignment is based on the concept of **positional homology**.
- Nucleotides (or amino acids) exhibit positional homology if they exist at equivalent positions in their respective sequences.
- A set of nucleotide or amino acid sequences is converted into an **alignment** by proposing positional homologies for each site.
- How should these two sequences be aligned?

Seq1 : ATGCGTCGTT

Seq2 : ATCCGCGTC

Two possible alignments



Two mismatches one indel

Sequence 1: ATGCGTCGTT

|| • || ||| •

Sequence 2: ATCCG-CGTC



No mismatches five indels

Sequence 1: AT--GCGTCGTT

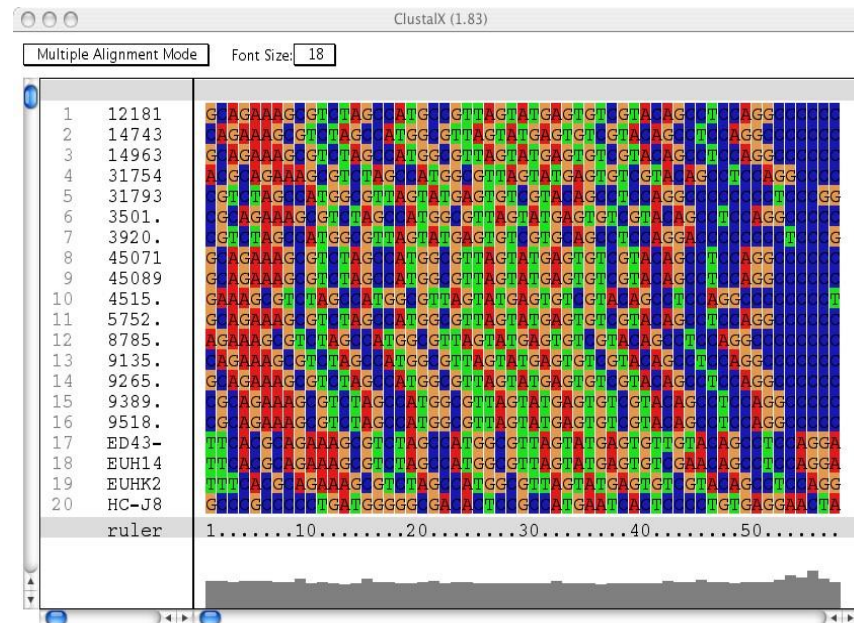
|| |||||

Sequence 2: ATCCGCGTC---

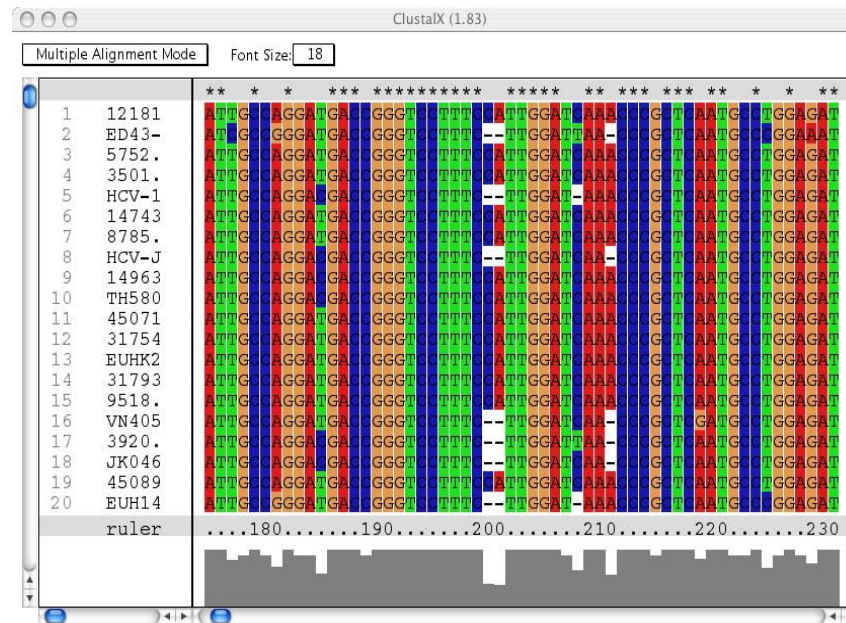
The 'costs' of alignment

- Most alignment methods start by assigning a different “cost” to each type of sequence difference,
 - transitions, transversions, insertions, deletions, etc.
- Each possible alignment therefore has a total cost.
- Algorithms then identify the alignment with the lowest cost.
 - This, of course is ‘the clever bit’; however, it is mostly computer science, not biology.
- CLUSTAL and MUSCLE are popular alignment applications.
- A good alignment is *essential* for good phylogenetics.

Alignments rapidly get complicated...



File /Users/Stephane/New_life/HK_HCV_paper/HCV6a_phylo tree.fas loaded.

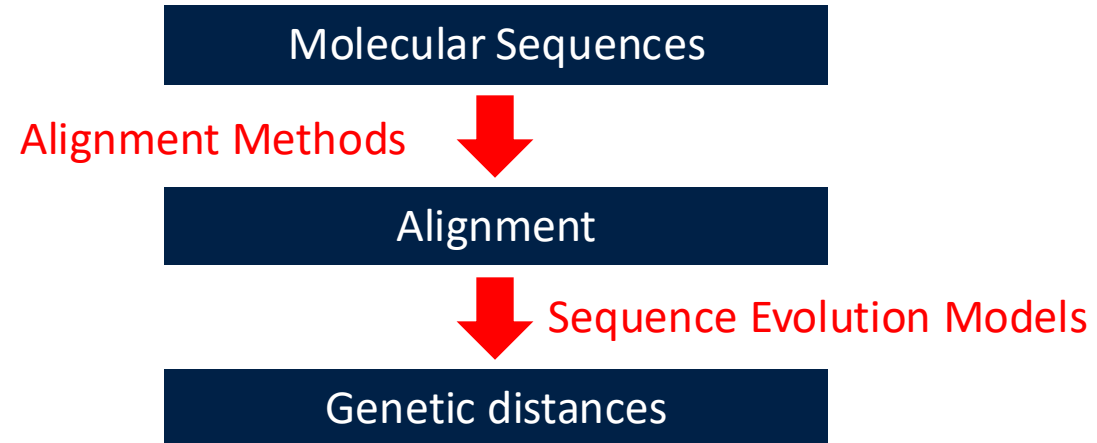


Elapsed time : 831.80 Secs

Alignment programs often make mistakes, particularly when the sequences are diverse or contain long insertions/deletions.

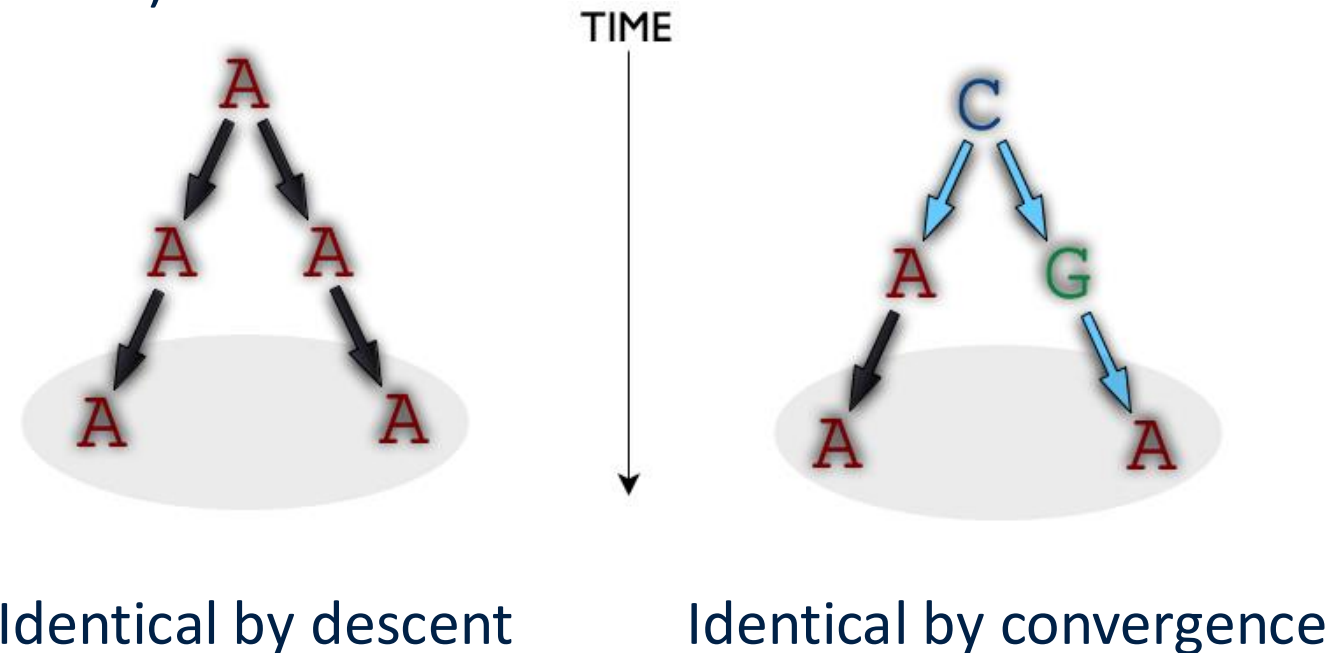
Exercise 1: Sequence alignment

How to build a tree



Genetic Distances

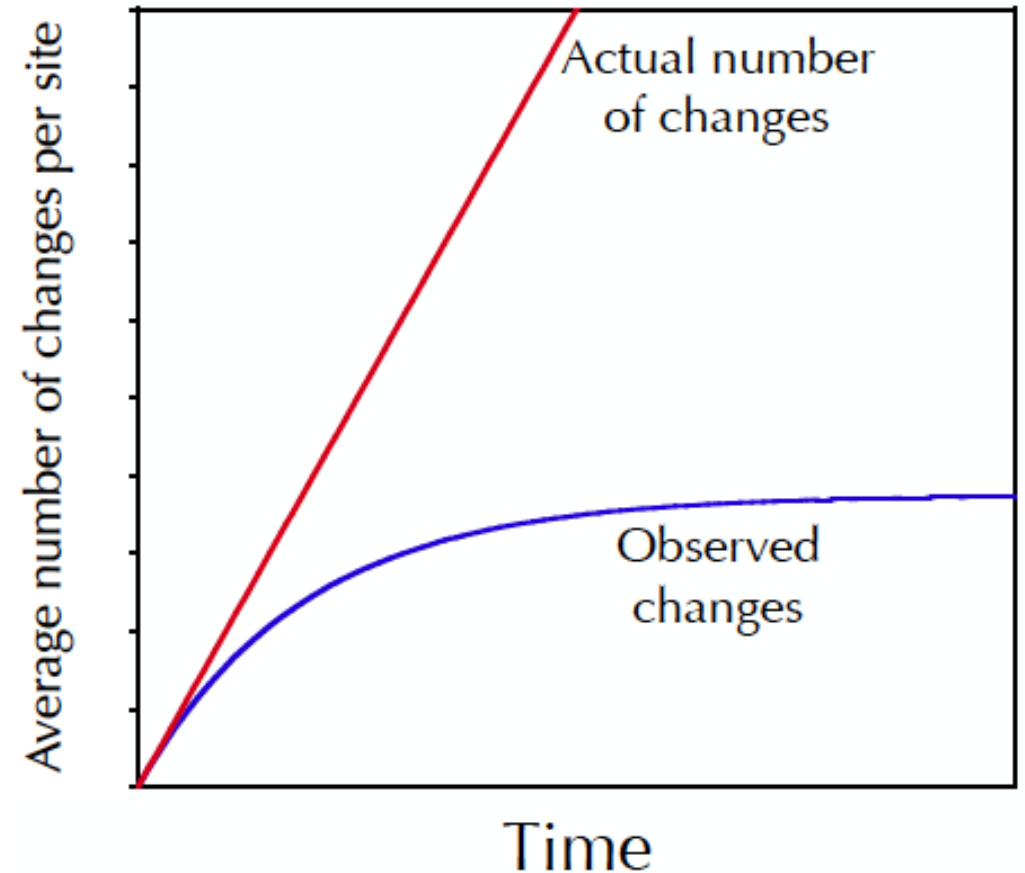
- Given two sequences, we would like a measure of how different they are.
- Why not simply count the proportion of sites that are mismatched (sometimes called the p-distance)?



How can we tell how many mutations actually occurred?

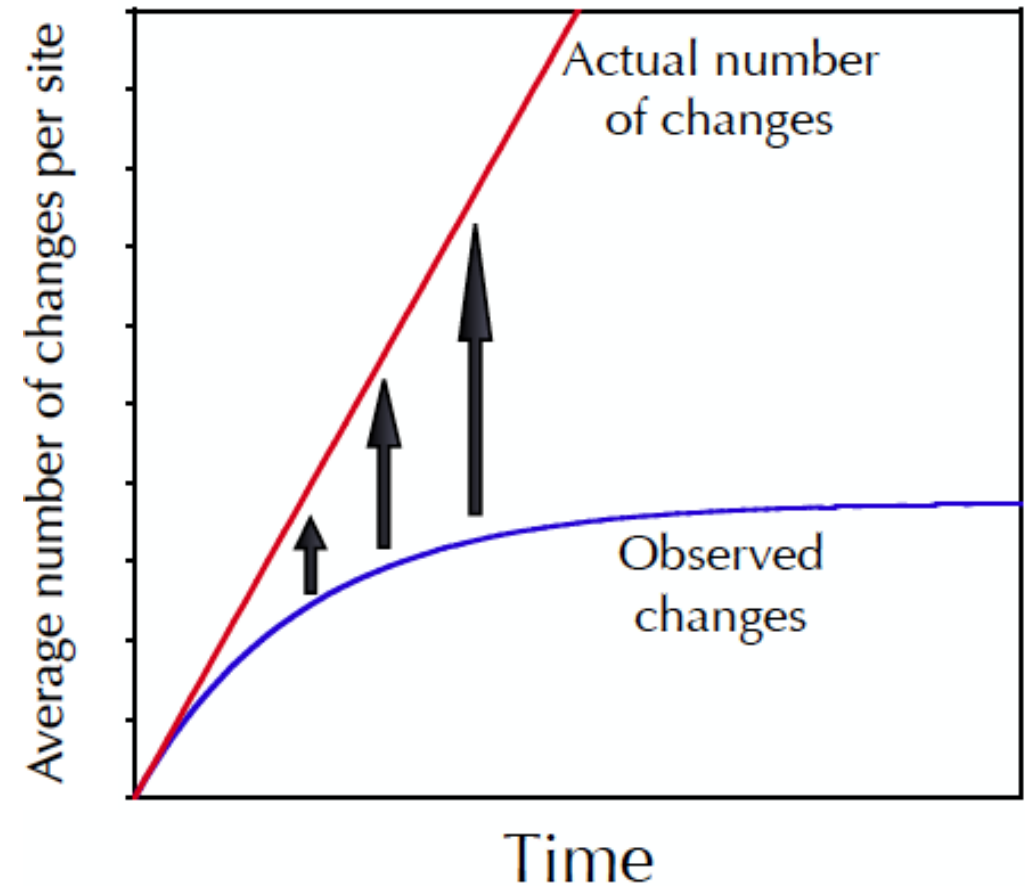
Genetic distances

- When divergence is low, the observed number of changes is similar to the true number.
- When divergence is high, the observed number underestimates the true genetic distance.
- This is the **multiple hits** problem.



Genetic distances

- **Nucleotide substitution models** are used to estimate the true genetic distance from the observed changes.
- They mathematically represent the stochastic process of sequence evolution through time.

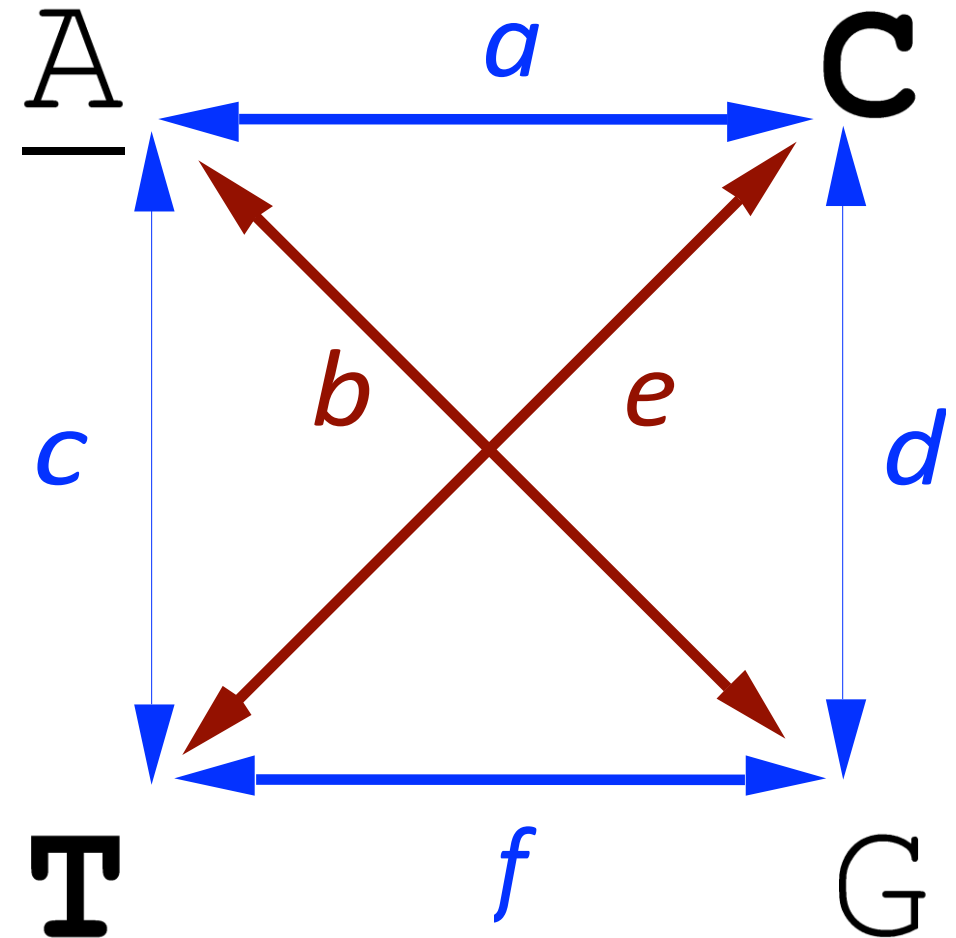


Nucleotide substitution models

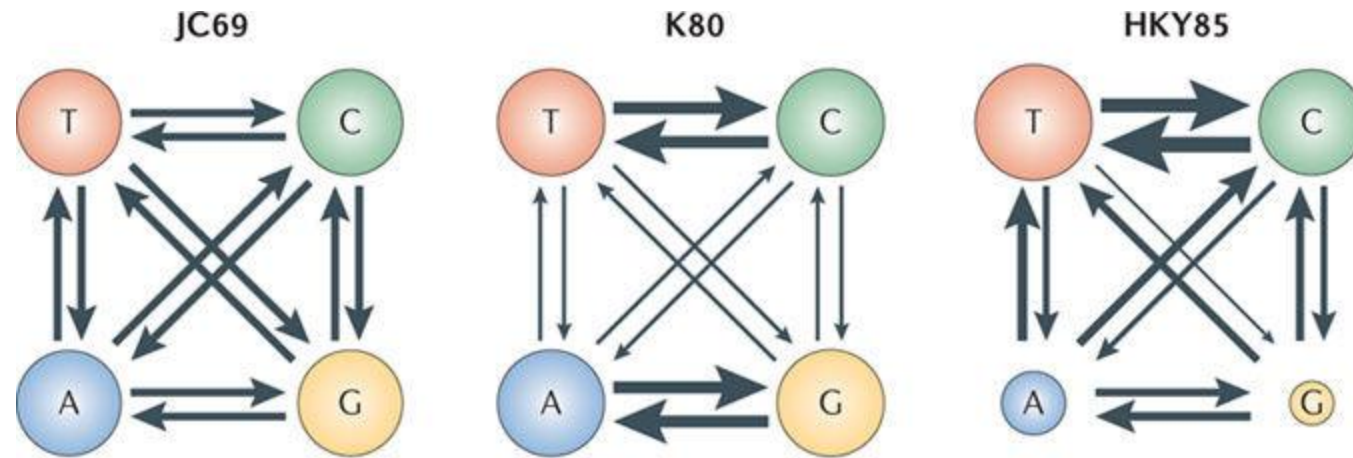
Letters a - f represent the relative rates of different types of mutation.

- The simplest **Jukes-Cantor (JC)** model assumes $a=b=c=d=e=f$.
- The **HKY** model assumes *transitions* occur at one rate and *transversions* at another (i.e. $a=c=d=f$ and $b=e$).
- The **General Time-Reversible (GTR)** model assumes each type of mutation occurs at a different rate.

Realistic models also include the relative frequency of each nucleotide.



Rates of substitution



Nature Reviews | **Genetics**

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

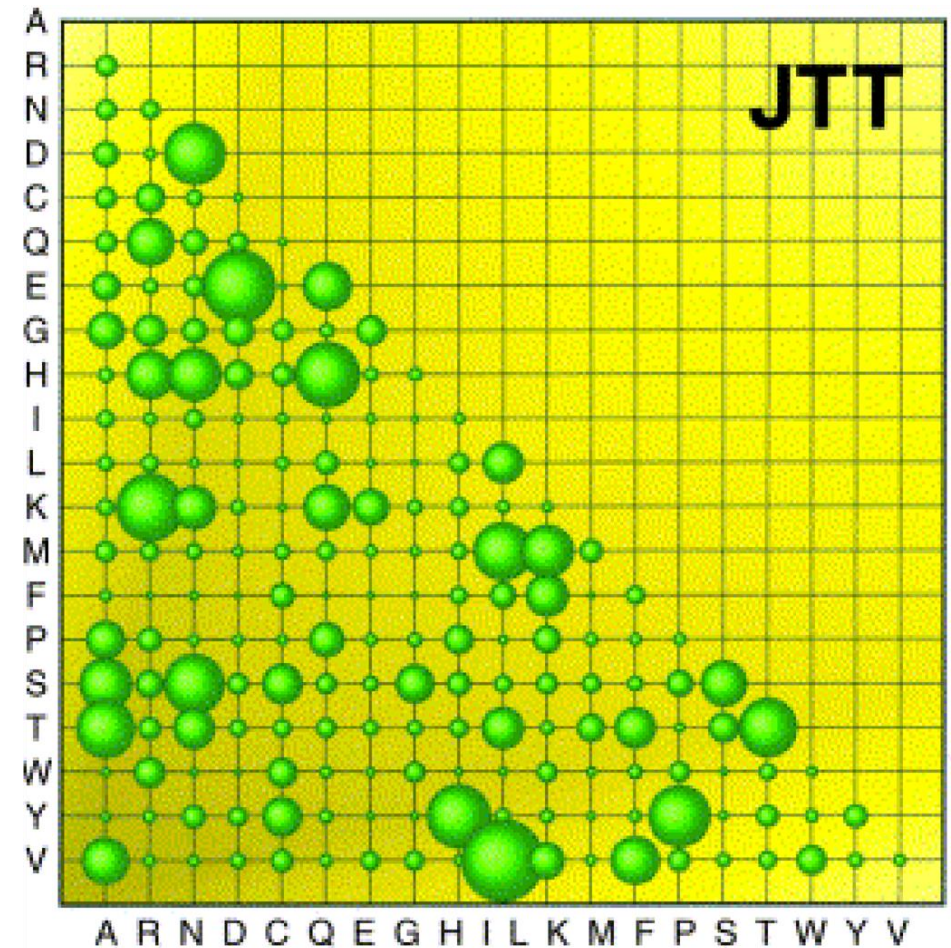
Yang, Z. & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nat Rev Genet* **13**, 303-314.

Nucleotide substitution models

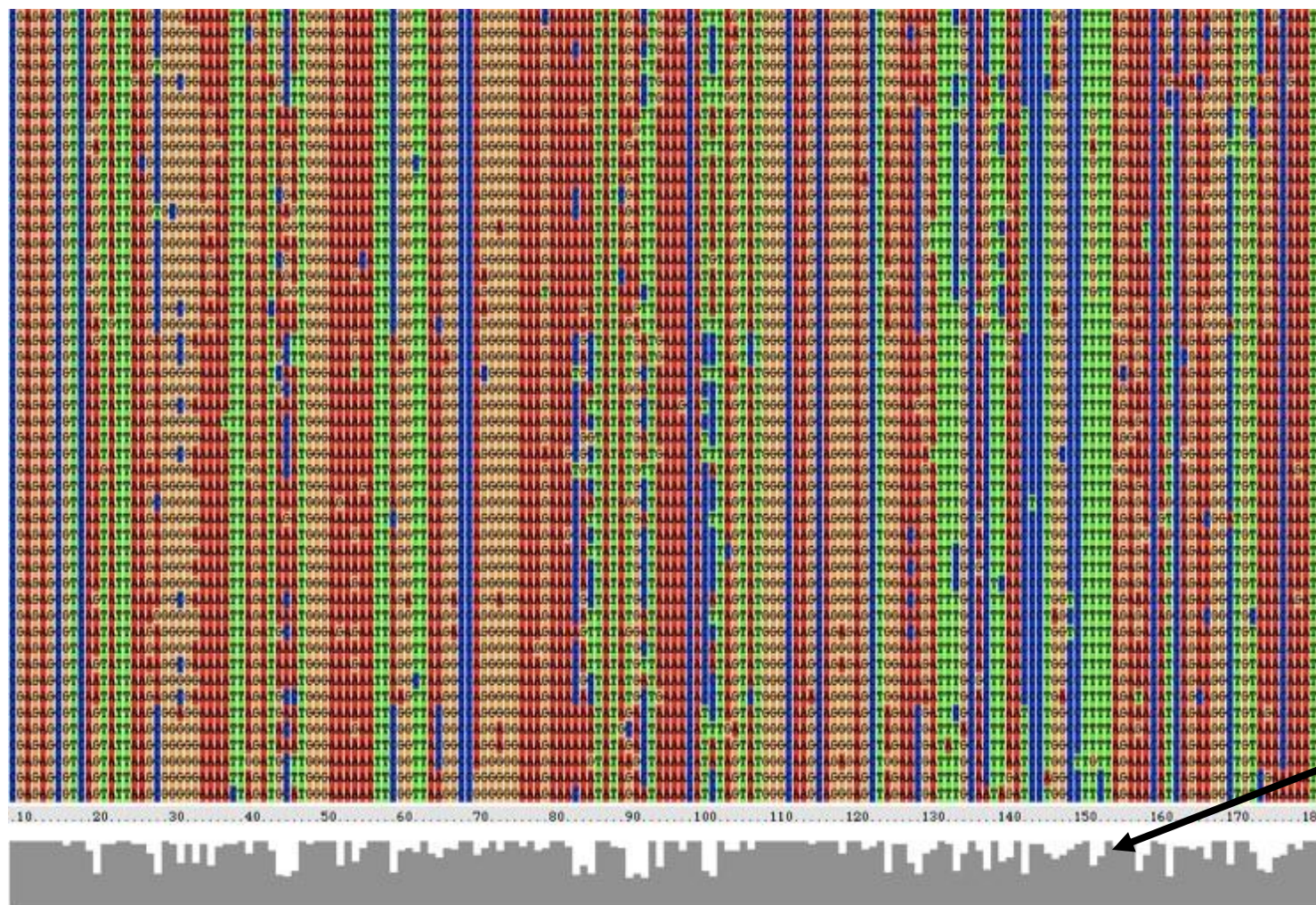
- The models make important biological assumptions:
 - evolution at each site occurs at the same rate;
 - in nature, sites change at very different rates due to functional constraint. Easily fixed, using models of among-site rate heterogeneity (usually the gamma model).
- Nucleotide base frequencies are the same for all sequences,
 - this can lead to problems if organisms with very different base compositions are compared.
- Evolution at each site is independent
 - Mathematically, this assumption is very hard to avoid. In nature, changes at sites under selection are often correlated (particularly for RNA genes with secondary RNA structure).

Amino acid substitution models

- Protein sequences have 20 states:
A, C, D, E, F, G, H, I, K, L, M, N, P, Q,
R, S, T, V, W, Y.
- Protein models therefore require a huge 20x20 matrix.
- These rates are obtained from large surveys of protein variation (not from your data set).
- The popular JTT matrix is estimated from a range of globular proteins.



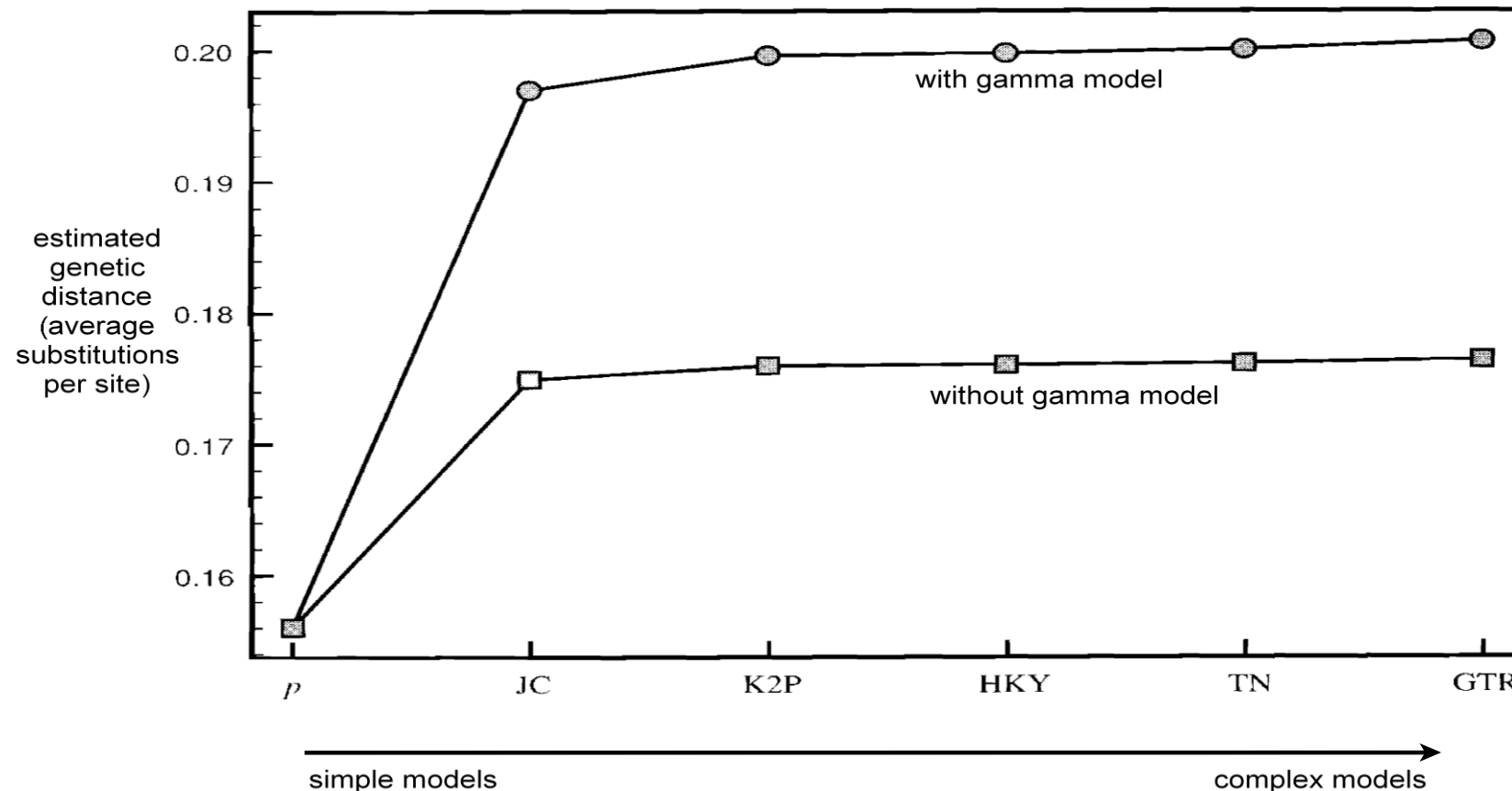
Among-site variation



nucleotide
similarity
varies
among sites

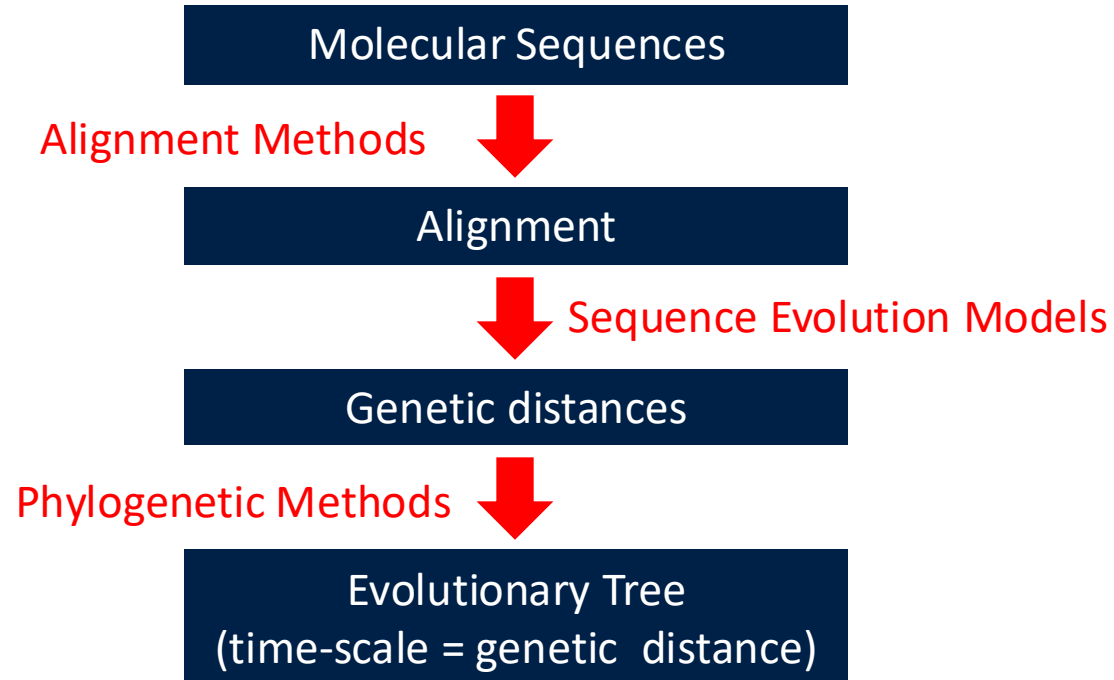
Genetic distance between human and squirrel TRIM5 α genes

Statistical models are used to capture the variation in evolutionary rate among sites.
The gamma-distribution model is most commonly used.



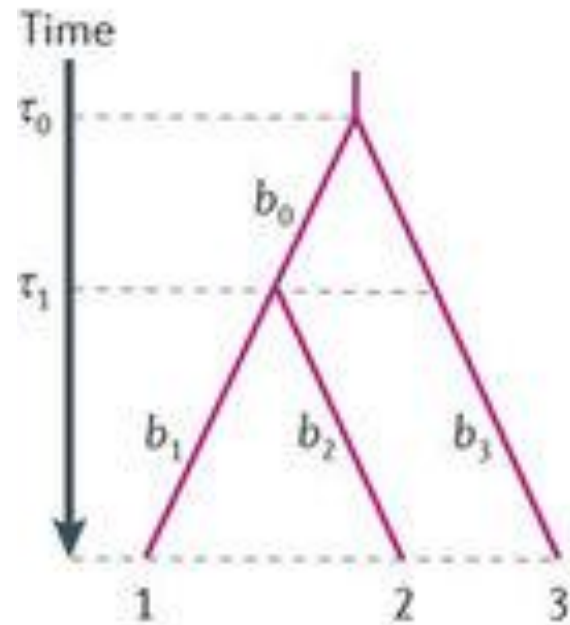
Exercise 2: Calculating genetic distances

How to build a tree

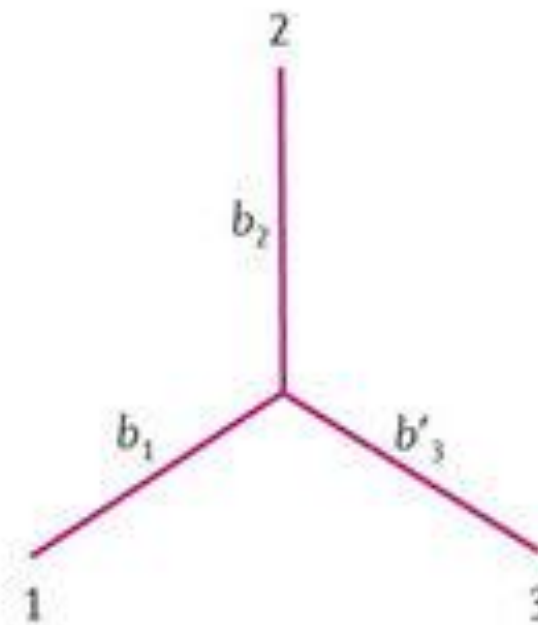


Tree concepts

a Rooted tree




b Unrooted tree



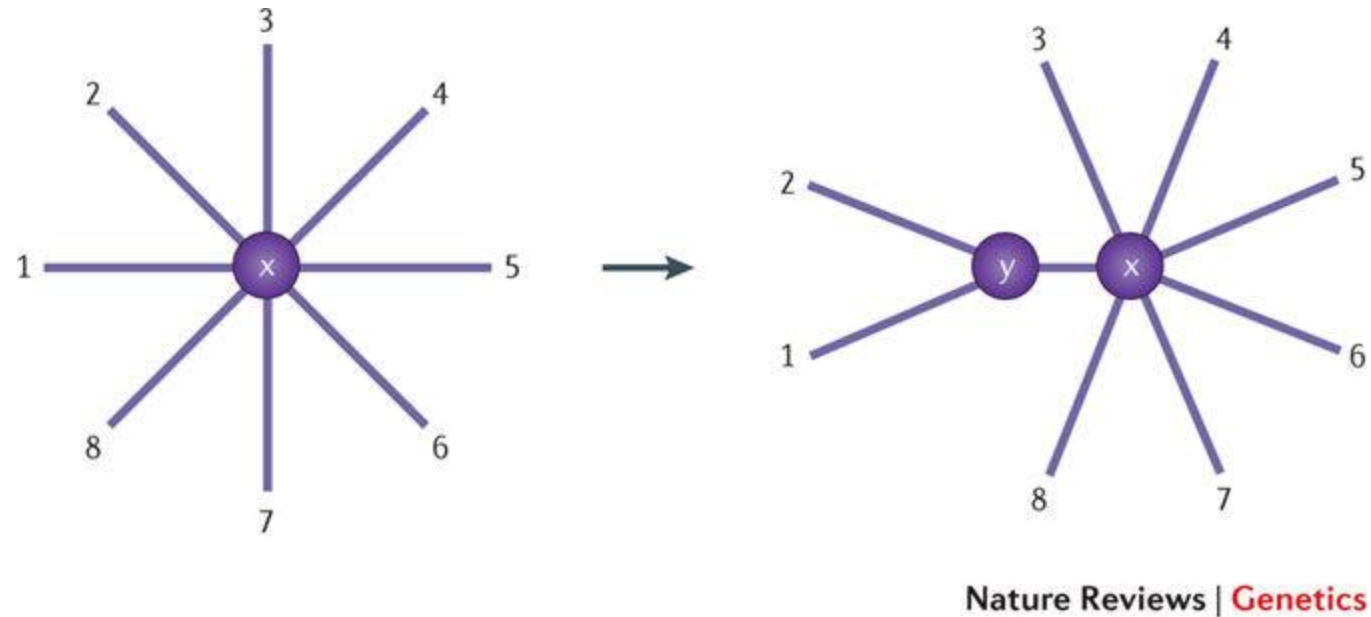
Nature Reviews | **Genetics**

Yang, Z. & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nat Rev Genet* **13**, 303-314.

Common phylogenetic methods

| | | |
|--|--|------------------------|
|  Fast | UPGMA Neighbour-joining | Algorithmic methods |
| | Maximum parsimony | Optimality methods |
| | Maximum likelihood Bayesian inference | Statistical methods |
| Slow | | |

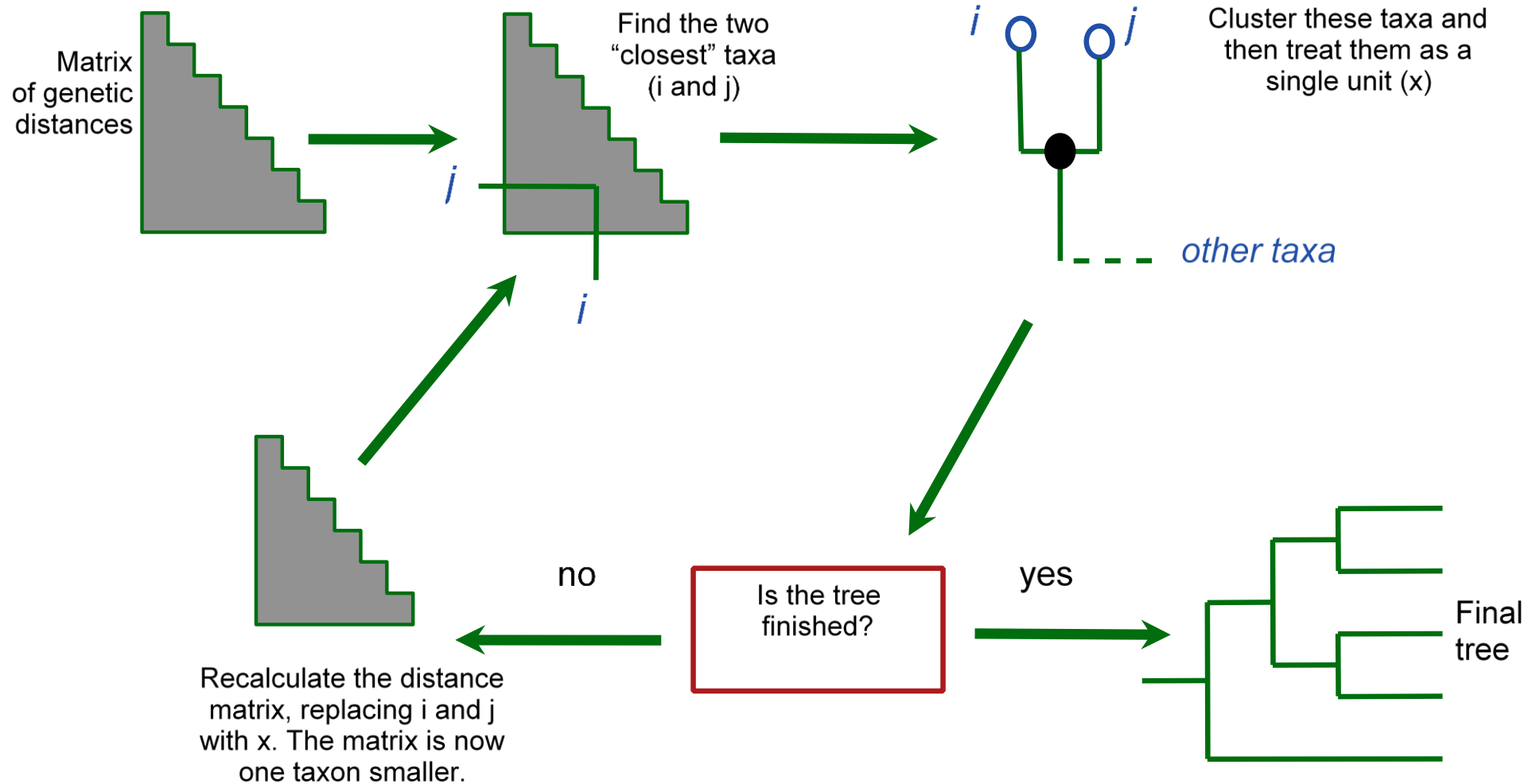
Algorithmic methods: Neighbour joining



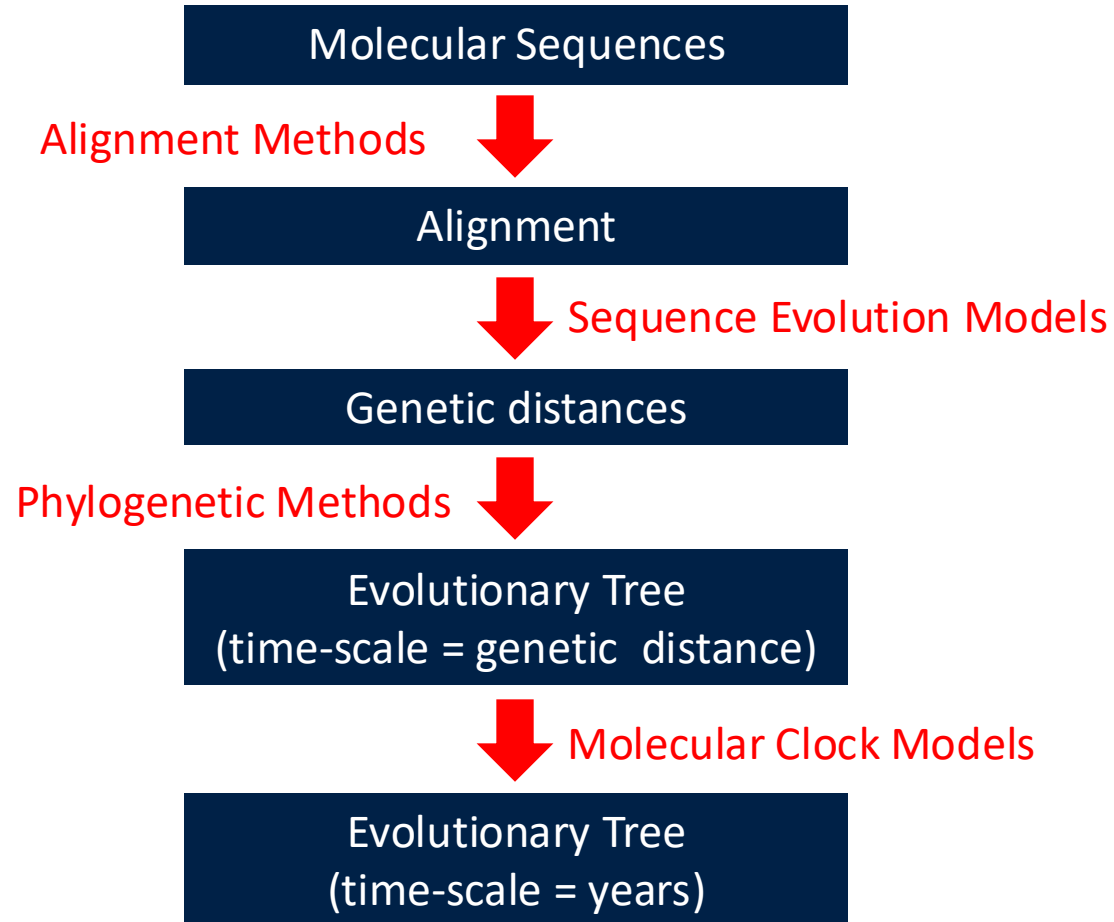
The neighbour joining algorithm is a divisive cluster algorithm. It starts from a star tree: two nodes are then joined together on this tree (in this example, nodes 1 and 2), reducing the number of nodes at the root (node x) by one. The process is repeated until a fully resolved tree is generated.

Yang, Z. & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nat Rev Genet* **13**, 303-314.

Algorithmic methods: UPGMA



How to build a tree

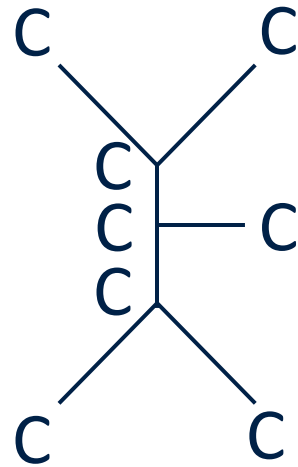
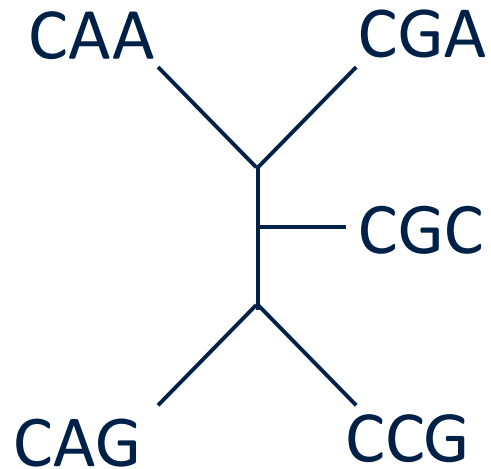


Optimality methods

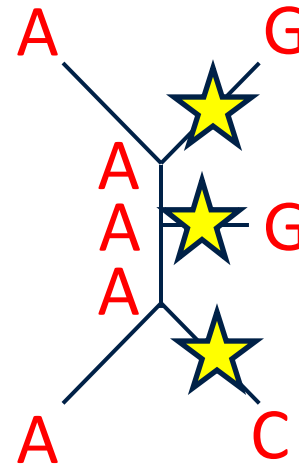
- **Maximum Parsimony**
 - The tree which requires the fewest evolutionary changes to explain the observed sequences is the best tree.
 - Fast, but inapplicable to fast-evolving or highly-divergent sequences. Most useful when applied to morphological character data.
- **Maximum Likelihood**
 - The tree which is probabilistically most likely to have given rise to the observed sequences is the best tree.
 - Slower. The probabilities are given by a nucleotide substitution model. Most common approach for sequence data.
- **Bayesian Inference**
 - Each tree has a probability given the data. We should consider the whole probability distribution, not just focus on the single most probable tree.
 - Slowest. Closely related to Maximum Likelihood. Most useful for testing evolutionary hypotheses. Becoming increasingly popular

Parsimony

- For any given tree and set of characters, the **parsimony score** is the minimum number of evolutionary changes (denoted —) required to explain the observed characters.
- The score is calculated separately for each character, then summed across characters.
- Suppose we have 5 sequences, which clockwise from top-left are CAA, CGA, CGC, CCG, and CAG:



Character
(base) 1
score=0



Character
(base) 2
score=3

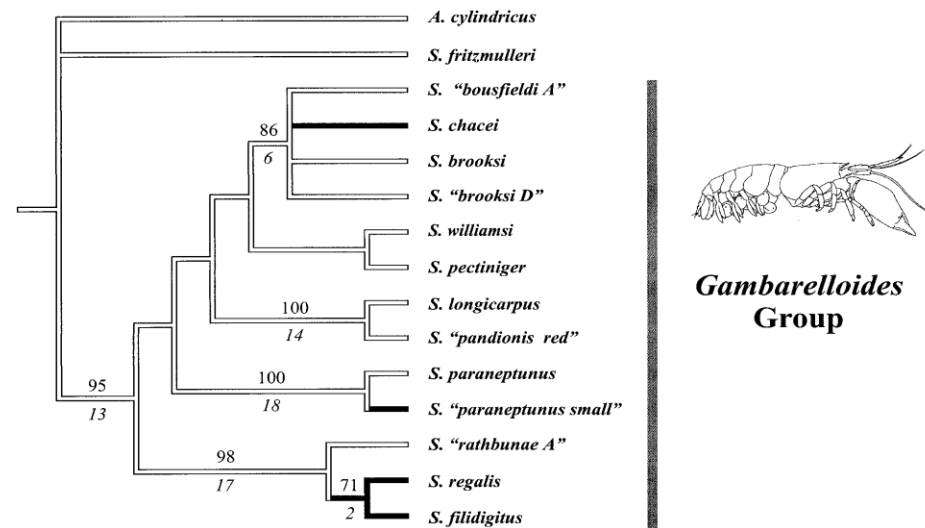


Character
(base) 3
score=2

TOTAL PARSIMONY SCORE FOR THIS TREE = 0+3+2 = 5

Parsimony

- The **most parsimonious tree** is that with the lowest parsimony score. However, there may be very many trees that share this property.
- Historically, parsimony was used to estimate phylogenies from morphological data.
- Today, parsimony is commonly used in **phylogenetic character mapping**.



Evolution of eusociality in coral-reef shrimp (black=eusocial, white=not eusocial).

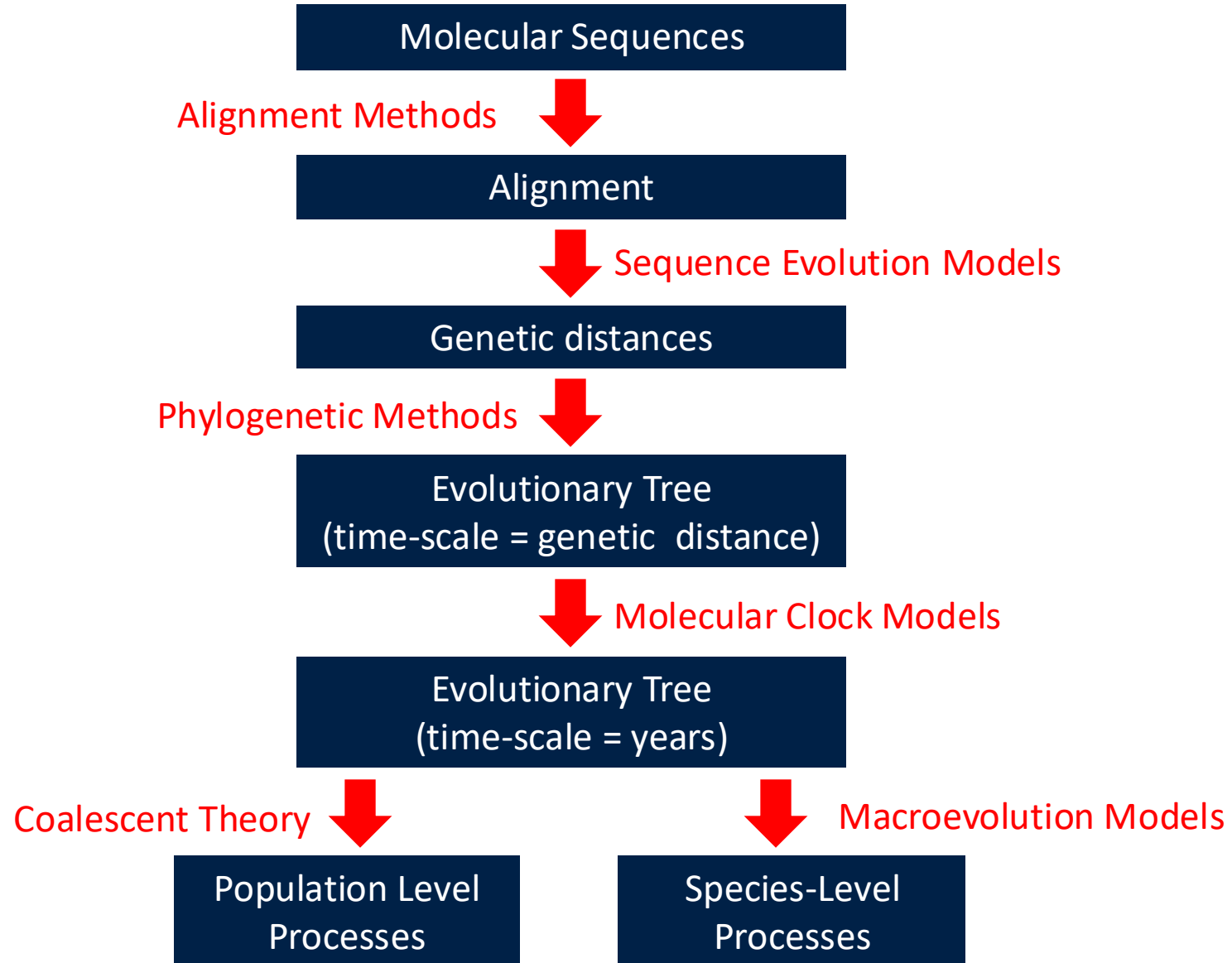
Likelihood

- Nucleotide (or amino acid) substitution models enable us to calculate $P(seqs/T, B, Q)$, that is, the probability of the observed sequences given:
 - a tree topology (T);
 - a set of branch lengths (B), each of which represents a genetic distance;
 - rate parameters of the substitution model (Q).
- The tree likelihood is proportional to this probability*. Calculating this requires some fairly heavy-duty maths.
- A **tree search** is used to find the topology T with the highest likelihood.

* You can think of likelihood and probability as the same thing; if you want to understand the distinction then read A.W.F Edwards (1972) *Likelihood*. Cambridge University Press.

Exercise 3: Building a tree

How to build a tree



Tree searching

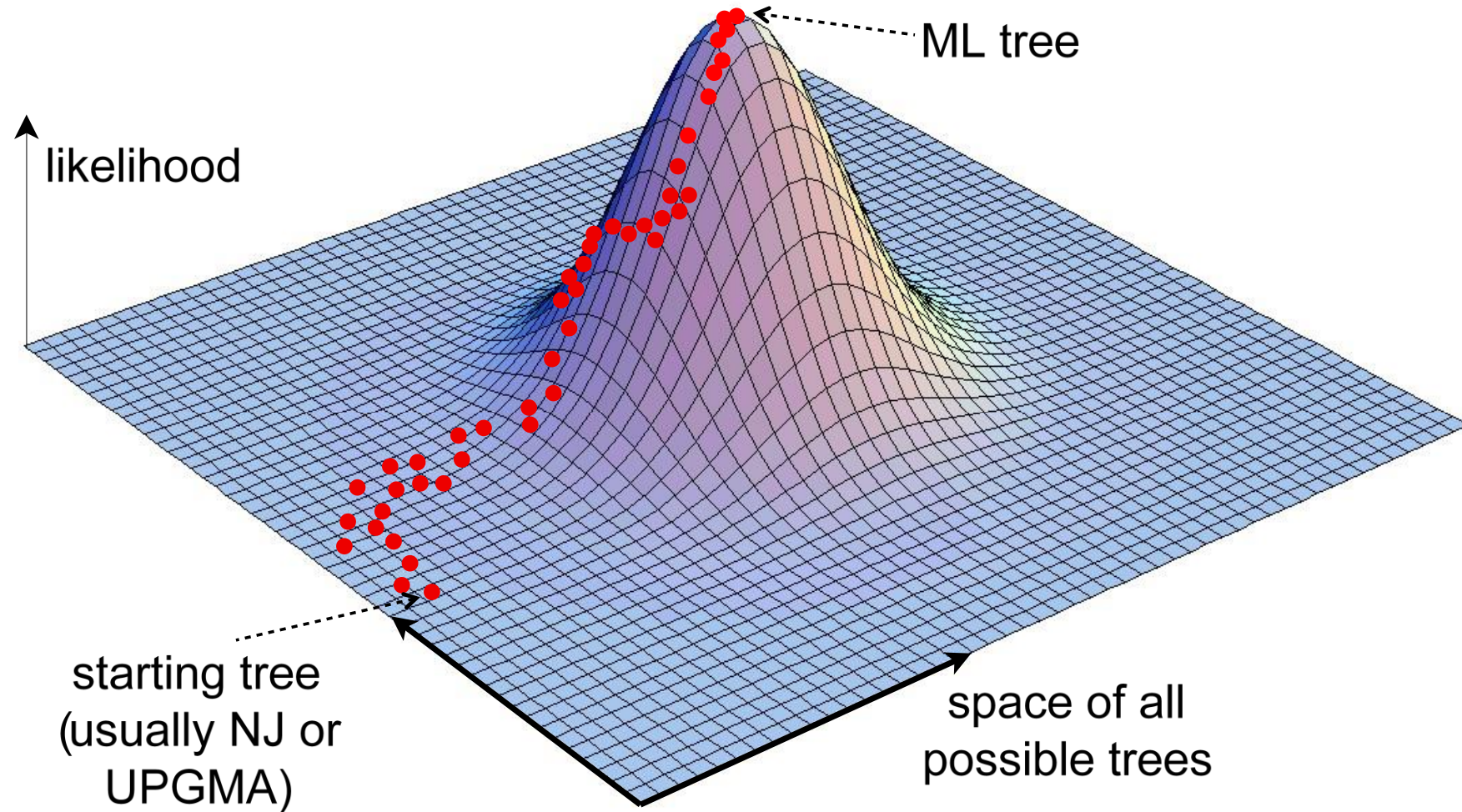
For n taxa there are $(2n-3)!$ rooted tree topologies:

| n | Number of trees | Comment |
|-----|------------------------|--|
| 4 | 15 | Enumerable by hand |
| 5 | 105 | Rainy day! |
| 6 | 945 | ... computer needed ... |
| 7 | 10395 | ... more so... |
| 8 | 135135 | > Hairs on your head |
| 9 | 2027025 | Population of a small country |
| 10 | 34459425 | Upper limit for exhaustive searching |
| 20 | 8.20×10^{21} | Upper limit for branch-and bound searching |
| 48 | 3.21×10^{70} | ~ particles in the Universe... |
| 136 | 2.11×10^{267} | A typical data set (!) |

How do we search through the set of all possible trees to find the best tree?

- **Exhaustive Search:** Tries every possible tree.
 - This is the only feasible with very small number of taxa.
- **Hill Climbing:** Searches through trees in a semi-random way. Doesn't check all possible trees and isn't guaranteed to find the optimal one.
 - But this is the only option for large data sets.

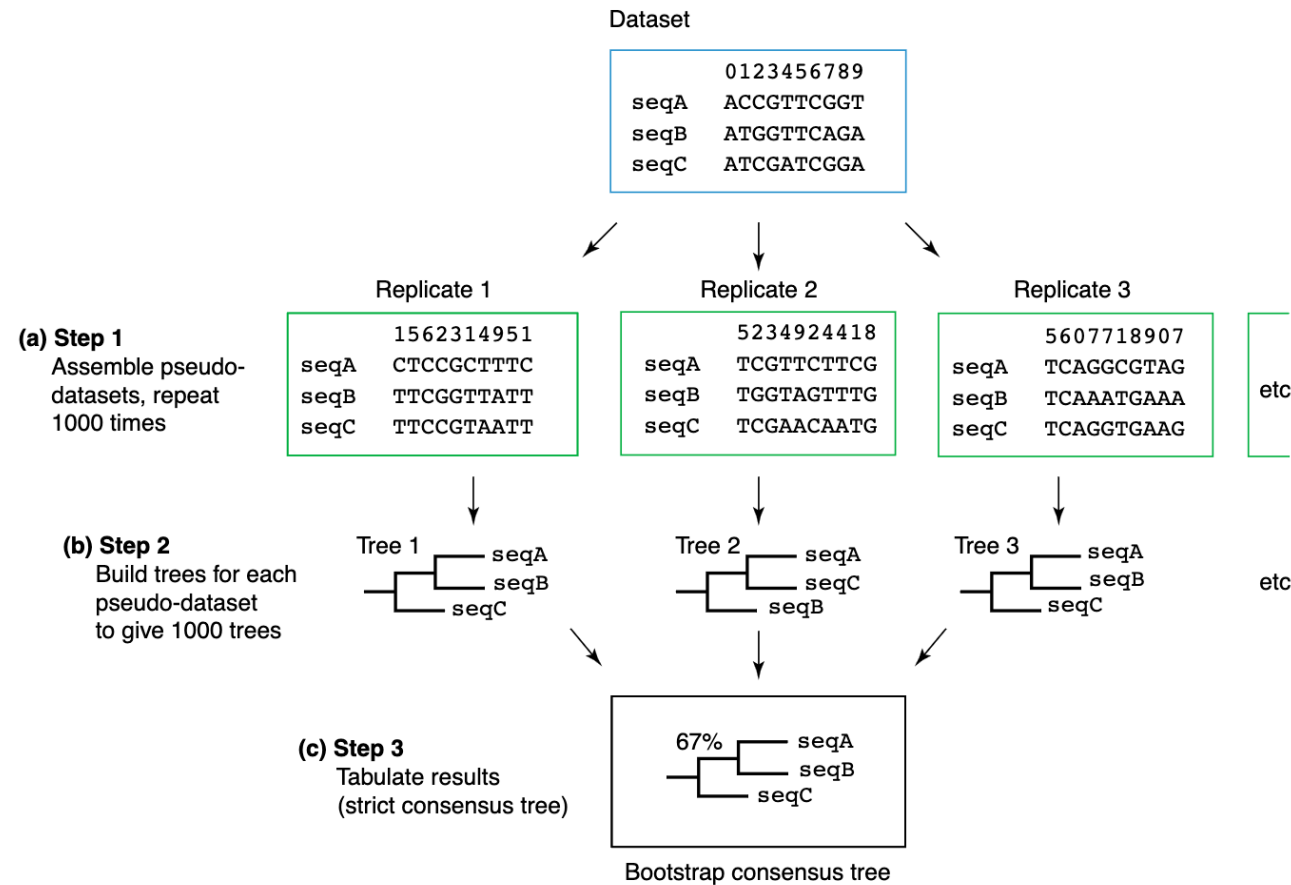
Hill Climbing



Phylogenetic uncertainty

- Most phylogenetic methods (UPGMA, NJ, ML) provide a single estimate of the ‘true’ tree.
- How do we measure the uncertainty of this estimate?
- Different parts of a tree (clusters) can be assessed individually for their reliability.
- “**Bootstrapping**” is the most common technique. It involves permutation of the original data to create a large number of *pseudoreplicate* data sets.

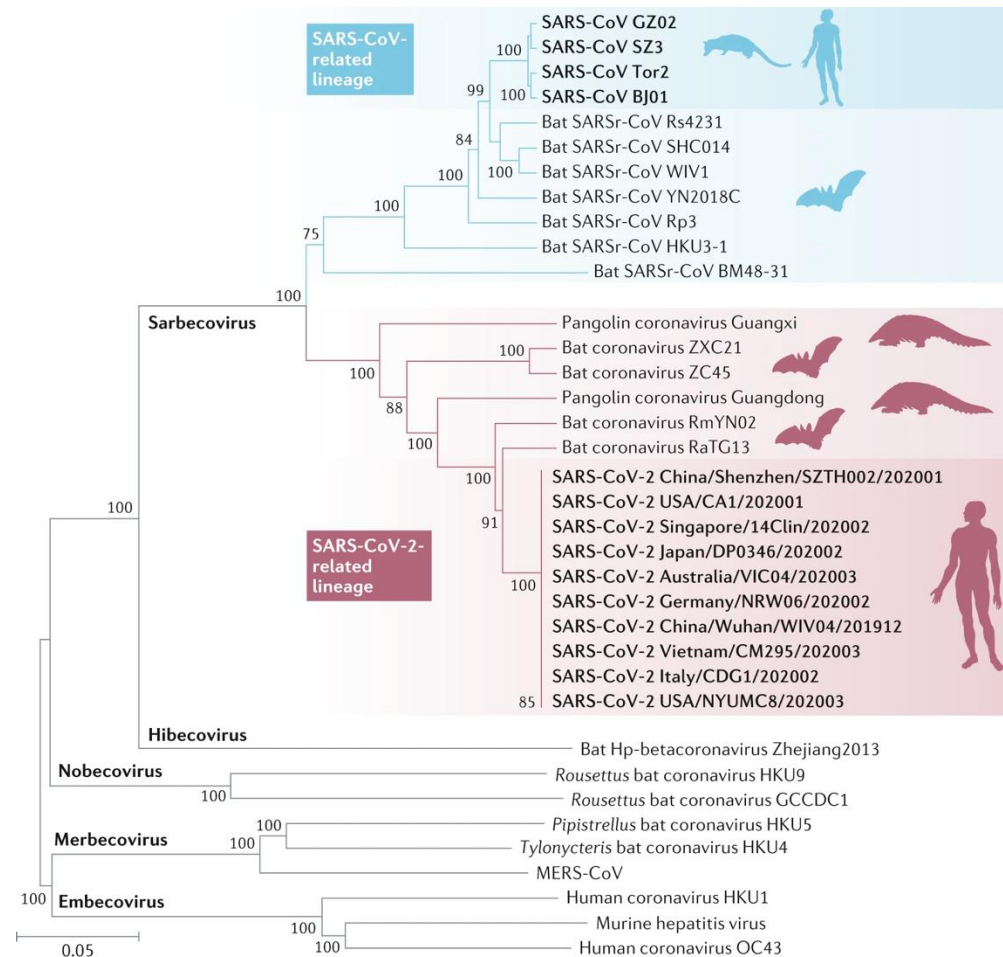
Bootstrapping



TRENDS in Genetics

Exercise 4: Calculating bootstraps

Example: SARS-CoV2 phylogeny



Hu, B., Guo, H., Zhou, P. & Shi, Z. L. (2021). Characteristics of SARS-CoV-2 and COVID-19. *Nat Rev Microbiol.* **19**, 141-154.

Tree construction was performed by the neighbour joining method with use of the program MEGA6 with bootstrap values being calculated from 1,000 trees. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) clusters with closely related viruses in bats and pangolins and together with SARS-CoV and bat SARS-related coronaviruses (SARSr-CoVs) forms the sarbecoviruses. The sequences were downloaded from the GISAID database and GenBank. MERS-CoV, Middle East respiratory syndrome coronavirus.

Take Home messages

- Once the general principles are understood, constructing phylogenies is straightforward,
 - MEGA software can perform most applications.
- The question must be defined, and the data appropriate,
 - a phylogeny is a model based on assumptions, not 'the truth'.
- For many applications, the precise models or tree building method will make little difference,
 - multiple phylogenies can be constructed and compared.
- Conventional phylogenetics does not account for horizontal gene transfer (HGT),
 - which is widespread in bacterial pathogens (to be continued).

