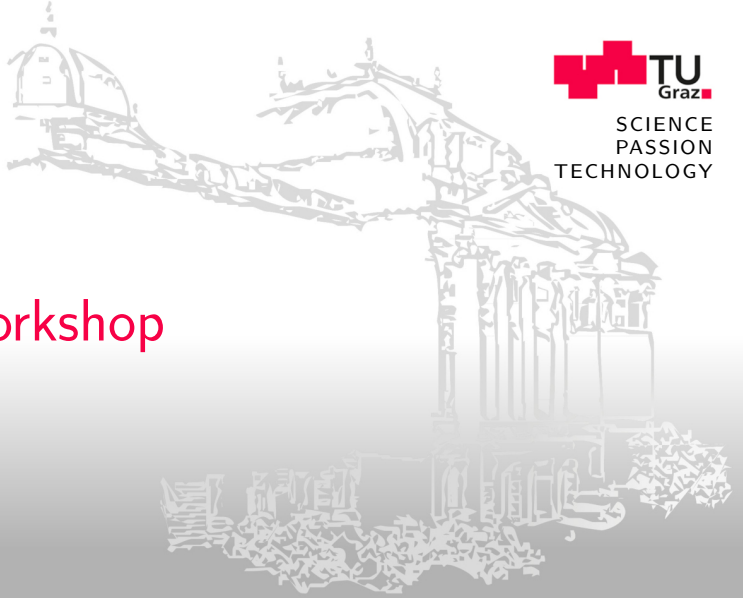


SCIENCE
PASSION
TECHNOLOGY

Bioinformatic workshop

Klaus Schliep, PhD

klaus.schliep@tugraz.at

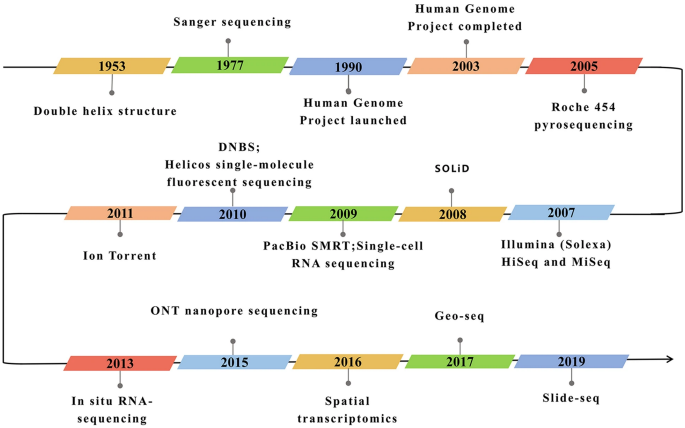
1. Sequencing technologies
2. Assembly
3. NCBI BLAST
4. Phylogenetics
   - Interpreting trees
   - Maximum Parsimony
5. Bootstrap
   - Maximum Likelihood

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Sequencing technologies

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

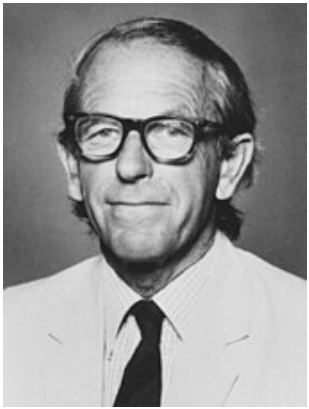
History



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

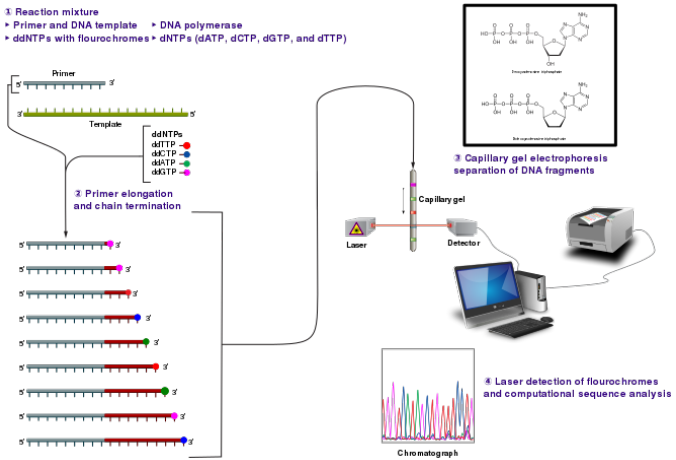
Frederick Sanger

- Two time Nobel laureate
- One in chemistry for his work on protein structure and Insulin
- One for his contributions concerning determination of base sequences in nucleic acids (together with Paul Berg and Walter Gilbert)



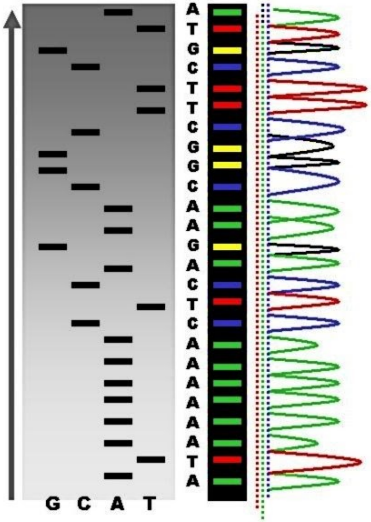
Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

How does Sanger sequencing work?



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Sequencing technologies



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Oxford nanopore sequencing

https://youtu.be/hs0FdiTHMbc, https://youtu.be/GUb1TZvMWsw

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

NGS platform trade-off

- Number of reads / Coverage
- Read lengths
- Cost
- Accuracy
- Application

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

NGS platform trade-off

This is just an indication and changing fast:

Platform	read lengths	error rate (not comparable)
Illumina	2 * 150bp	0.01%
PacBio	5-25kb	10-15%
Oxford Nanopore	5-20kb	5-15%

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

NGS reads

- Short subsequences of the genome
  - No idea on the original position in the genome
  - Orientation (strand) unknown
- Oversampled (high coverage)
  - Reads overlap: only clue for assembly
- Base errors
  - Indels, substitutions, characteristic biases
- Supposed to cover entire genome
  - Not always true
  - Coverage not uniform

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

What is a "read"? FASTQ-Format

- Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is optionally followed by the same sequence identifier.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

```
ACTTAAAAACTGGGTTCAGGCTACTCCCAAAACAGCATAATAAAAAATAATCTAAGAATGATCCCTTTTAGGACCAACC
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```


Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Phred Quality Scores

We can calculate the phred quality score from the probability of sequencing error (i.e. the base call is wrong) using:

$$Q = 10 \log_{10} p$$

Alternatively, we can rearrange to calculate the probability of error from the phred quality score using:

$$p = 10^{\frac{Q}{-10}}$$

Where Q is the phred quality score and p is the probability of error (i.e. the base call is wrong).

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Phred Quality Scores

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Assembly

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

We can only read a limited number of characters at a time from a random place:

```
good-natured, she thought: still
when it saw Alice. It looked
ought to be treated
good-natured, she thought, still
Cat only
a greet many
It looked good-
The Cat only grinned when it saw Alice.
be treated with respect.
sti1l it had very long claws
claws and a great many teeth, so she
so she felt that it ought
```


Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

We need an assembly algorithms to piece the story together.

```
The Cat only grinned when it saw Alice.
    Cat only
                     when it saw Alice. It looked
                                        It looked good-
                                                  good-natured, she thought: still
                                                  good-natured, she thought, still
sti1l it had very long claws
                       claws and a great many teeth, so she
                                 a greet many
                                                     so she felt that it ought
ought to be treated
         be treated with respect.
```


Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Assembly approaches

- Reference assembly
  - We have sequence of similar genome
  - Reads are aligned to the reference
  - Can guide, but can also mislead
  - Used a lot in human genomics
- *De novo* assembly
  - No prior information about the genome
  - Only supplied with read sequences
  - Necessary for novel genomes
  - Or where it differs from reference eg. Cancer

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Assembly Challanges

- Short reads
- Read errors
- Repeated sequences
- Ploidy
- Contamination

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

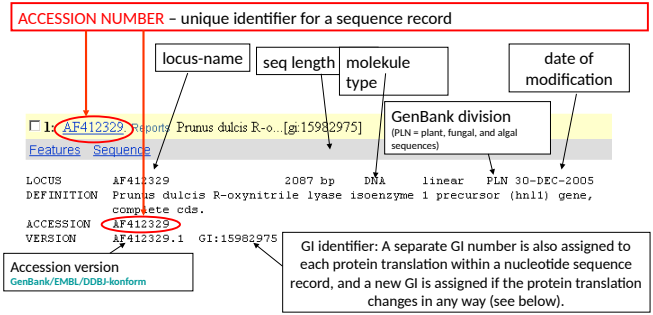
NCBI BLAST

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

NCBI record

Example „*Prunus dulcis* R-oxynitrile lyase isoenzyme 1"

Reference: http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

What is BLAST?

- Widely used sequence similarity search tool
- Finds high scoring alignments between two sequences (protein or DNA)
- Includes a model of score distributions for random local alignments
- Provides statistical significance for alignments

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

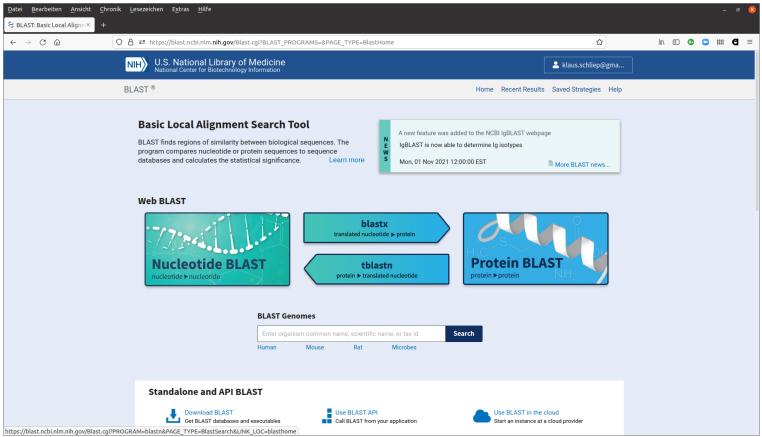
What BLAST tells you

We have a DNA or protein sequence

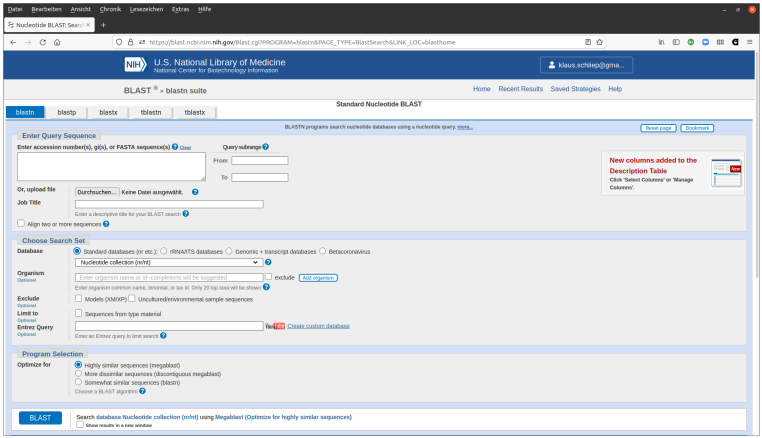
1. What is it related to? What does it do? (Homology)
2. Is it already in the database? (matching sequences , organism of origin)
3. Where is it located or how is it organized?

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

NCBI BLAST


Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
BLAST: Basic Local Align
https://blast.ncbi.nlm.nih.gov/Blast.cgi?BLAST_PROGRAMS=&PAGE_TYPE=BlastHome
U.S. National Library of Medicine
National Center for Biotechnology Information
klaus.schliep@gma...
BLAST ®
Home Recent Results Saved Strategies Help
Basic Local Alignment Search Tool
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.
Learn more
NEWS
A new feature was added to the NCBI IgBLAST webpage
IgBLAST is now able to determine Ig isotypes
Mon, 01 Nov 2021 12:00:00 EST
More BLAST news...
Web BLAST
Nucleotide BLAST
nucleotide ▶ nucleotide
blastx
translated nucleotide ▶ protein
tblastn
protein ▶ translated nucleotide
Protein BLAST
protein ▶ protein
BLAST Genomes
Enter organism common name, scientific name, or tax id
Search
Human Mouse Rat Microbes
Standalone and API BLAST
Download BLAST
Get BLAST databases and executables
Use BLAST API
Call BLAST from your application
Use BLAST in the cloud
Start an instance at a cloud provider
https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

NCBI BLAST



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Global vs. Local Alignment

- Global: Find the best overall alignment between sequences (end-to-end).
- Local: Find short regions of highly conserved sequence.

```
Sequence 1: KLAUS
Sequence 2: NIKOLAUS

Sequence 1: --K-LAUS
Sequence 2: NIKOLAUS

Sequence 1: LAUS
Sequence 2: LAUS
```


Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

NCBI – BLAST Search: S- and E-value

Calculating alignment scores

- The raw score S for an alignment is calculated by summing the scores for each aligned position and the scores for gaps.
- In amino acid alignments, the score for an identity or a substitution is given by the specified substitution matrix (e.g. BLOSUM62).
- the higher the alignment score the more similar are sequences.

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

NCBI – BLAST Search: S- and E-value

E value

- The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size.
- It decreases exponentially as the Score (S) of the match increases.
- E value describes the random background noise.
- The lower the E-value, or the closer it is to zero, the more "significant" the match is. However, keep in mind that virtually identical short alignments have relatively high E values. This is because the calculation of the E value takes into account the length of the query sequence. These high E values make sense because shorter sequences have a higher probability of occurring in the database purely by chance.

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

16S rRNAs as markers

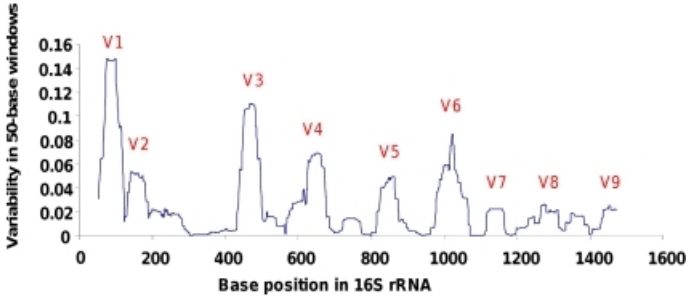
16S rRNA most commonly used Conserved, but sufficiently different across organisms

- Exist in all living organisms
- Play important roles in protein translation
- Can be used as molecular markers
  - For phylogenetic analysis
  - Used to build tree-of-life
- 16S rRNA most commonly used (prokaryotes)
  - Conserved, but sufficiently different across organisms
  - 18s and ITS are common marker for fungi

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

16s rRNAs as markers

- Highly conserved, but different across organisms
- Conserved regions (for primer construction) alternate with variable regions



Bodilis et al. (2012) 16S variable regions in Pseudomonas PLoS ONE

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Identification of uncultured bacteria using 16S

We first search for these nucleotide sequences in NCBI https://www.ncbi.nlm.nih.gov/:
KU524801.1 AB759680.1 GQ158974.1 DQ904997.1 EU488411.1 KX431275.1 EU556993.1 FM873915.1 FM874039.1 HM124388.1 FJ624883.1 FJ625334.1 EU236261.1 EF508875.1 DQ814438.1 HM779760.1 HM780090.1 EF604165.1 EF604435.1 EF604230.1

- Press Run BLAST
- Database: rRNA/ITS databases
- Press BLAST

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Phylogenetics

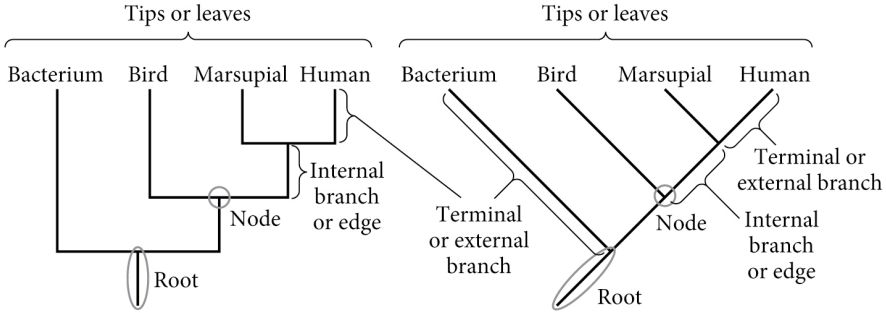
Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

What is a Phylogenetic Tree

A phylogenetic tree shows relationship among evolutionary related objects. *Objects* can be biological sequences, species, languages, manuscripts, ...
Mathematical:
A tree is a connected graph without cycles. Nodes of degree one are called tips, external nodes or leaves. Nodes o degree higher than one are called internal nodes. A *phylogenetic tree* or phylogeny is a tip-labelled tree with label set X of species.

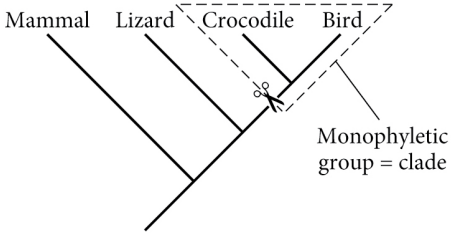
Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

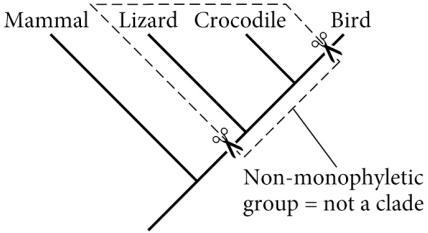
Terminology



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

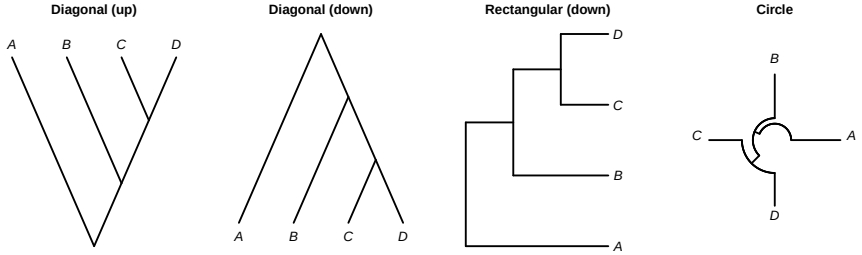
Terminology





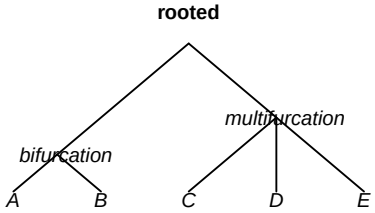
Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

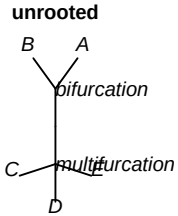
Styles



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Birfucating trees, Polytomies

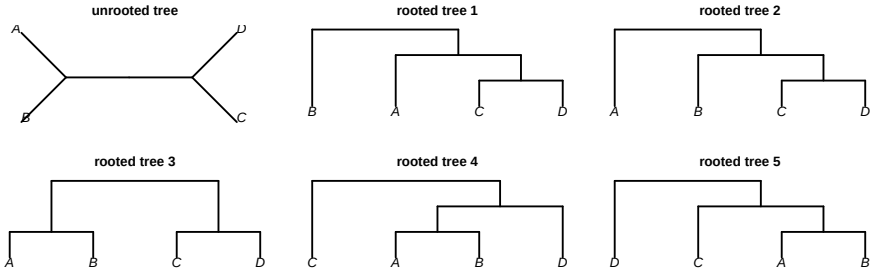




Some software only accepts bifurcating trees, i.e. trees without multifurcations / polytomies.

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

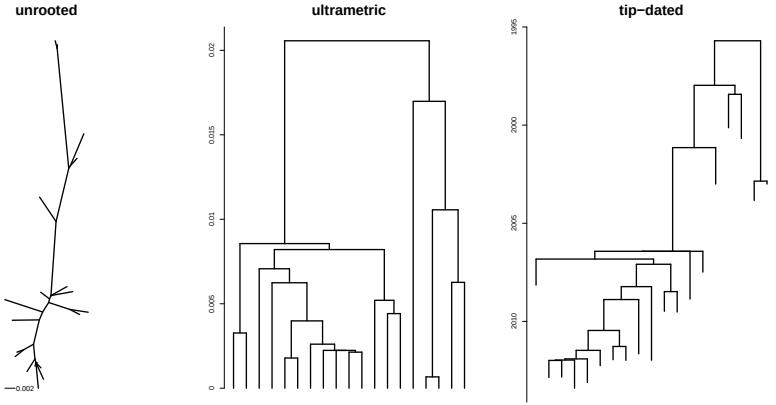
Unrooted trees



Many inference programs return unrooted trees. There are many ways to root a tree (on every internal edge).

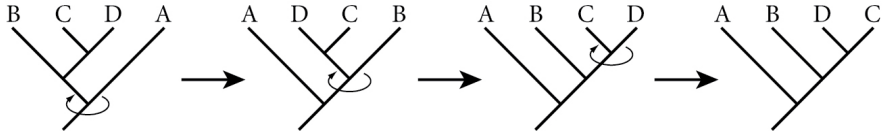
Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Trees & Time, Molecular clock



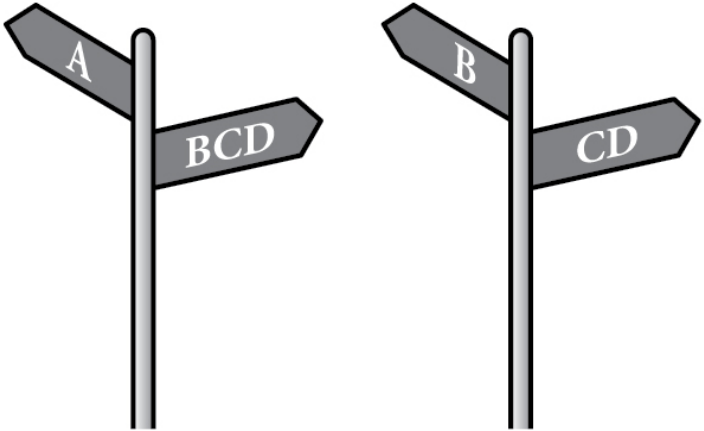
Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Topology



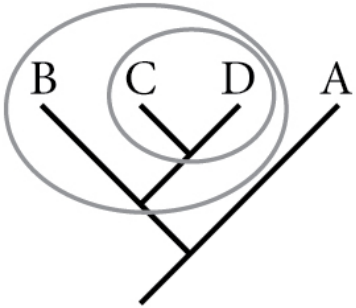
Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

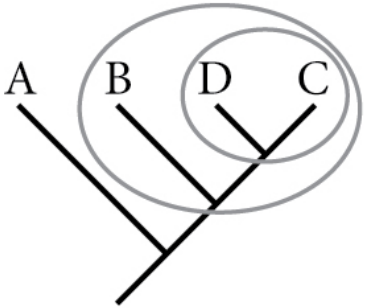
Topology



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

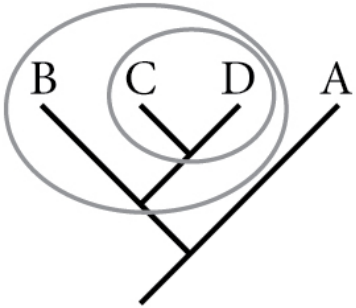
Topology

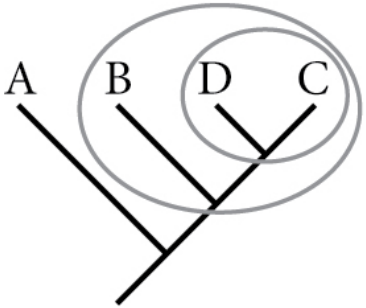




Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Newick






```
( ( B, ( C,   D )), A);
```

```
( A, ( B, ( C, D ) ) );
```

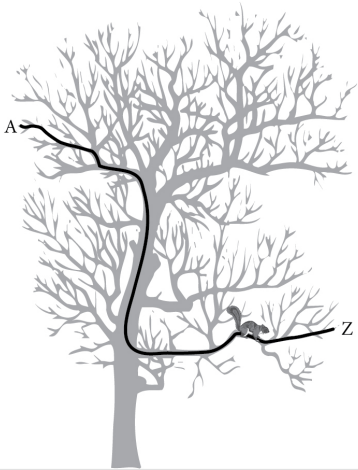
Parentheses enclose elements (terminal or internal branches), separated by commas that emerge from the same internal node

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Newick's lobster house

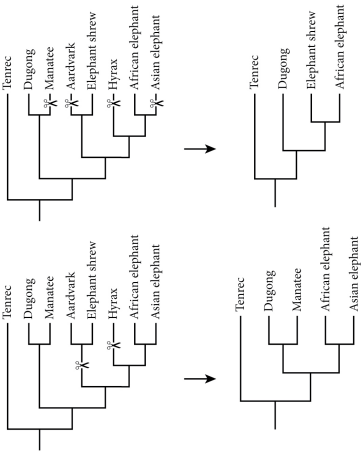
Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Merging and Pruning



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Merging and Pruning



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Now let's have a quiz

https://klash.shinyapps.io/tree_thinking/

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Software for tree visualisation

- Dendroscope https://www.wsi.uni-tuebingen.de/lehrstuehle/algorithms-in-bioinformatics/software/dendroscope/
- TreeView https://code.google.com/archive/p/treeviewx/
- R packages: ape https://CRAN.R-project.org/package=ape, phytools https://CRAN.R-project.org/package=phytools & ggtree https://bioconductor.org/packages/ggtree/
- Python: ETE http://etetoolkit.org/

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Maximum Parsimony

- The basic idea behind maximum parsimony phylogeny inference is that the best phylogenetic tree is the one that explains the character data with the smallest number of changes.
- The principle underlying this idea is a pervasive one in science (and, in fact, in our day to day lives, even if we usually don't realize it).
- This principle is called the principle of parsimony or Occam's razor.
- The principle states that given two competing hypotheses, the one with the smaller number of assumptions (i.e., the simpler hypothesis) is more likely to be the correct one.

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Maximum Parsimony

There are two problems:

- How do we figure out how many evolutionary changes are implied by a particular data pattern on the tree?
- How do we find the tree with the smallest number of evolutionary changes?

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Inferring Phylogenies: Parsimony

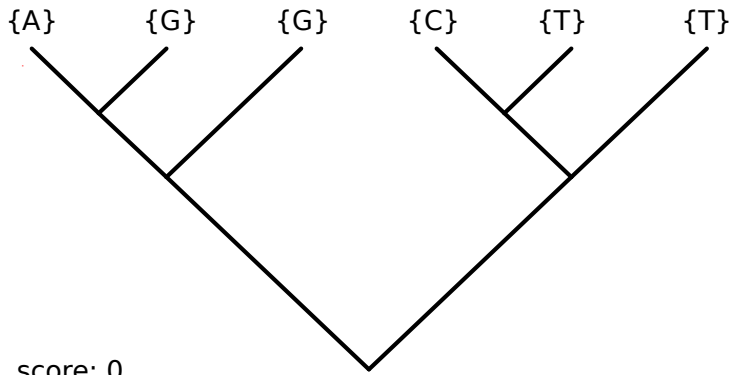
The first algorithm to compute the minimum number of changes implied by a data pattern on a tree is called the Fitch algorithm because it was identified by Walter Fitch (1929-2011).

1. To apply the Fitch algorithm we start at the tips of the tree, and then we descend root-ward (post-order traversal) through internal nodes.
2. For each node, we compute the intersection between the sets for the nodes above.
3. If the set is empty, we compute the union and $+1$ to the parsimony score.

We have to do these steps for each column in an alignment and sum the scores up.

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

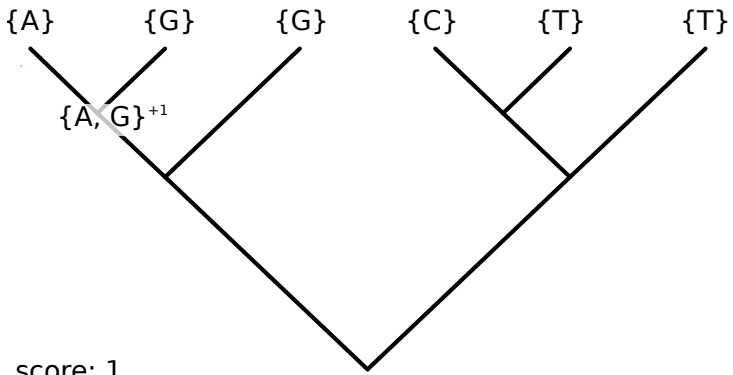
Fitch algorithm



score: 0

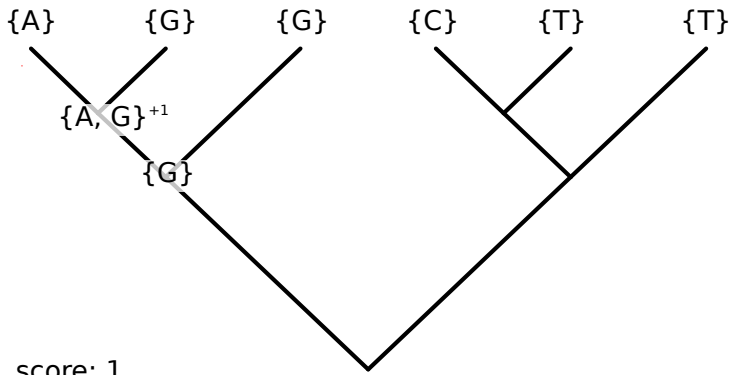
Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Fitch algorithm



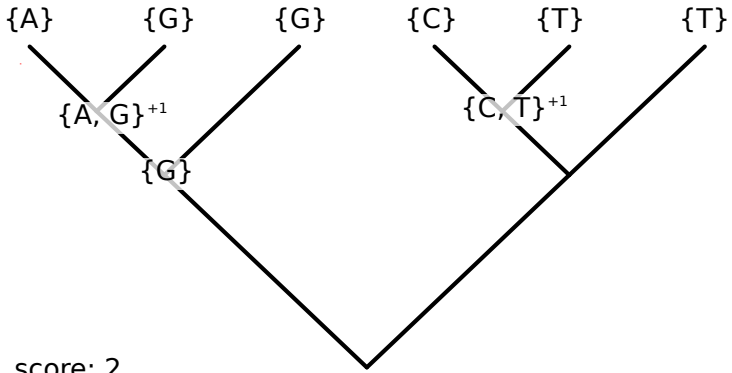
Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Fitch algorithm



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

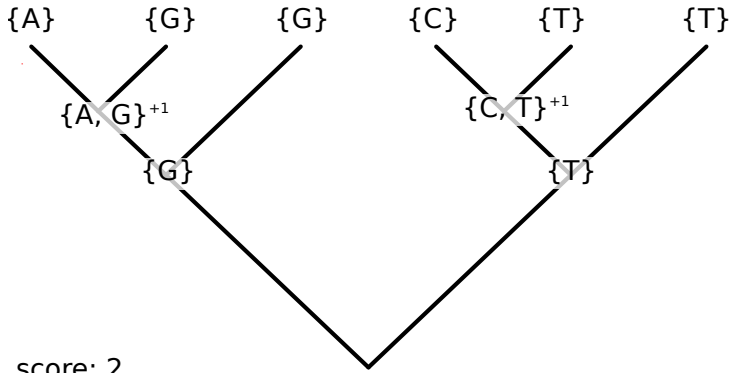
Fitch algorithm



score: 2

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

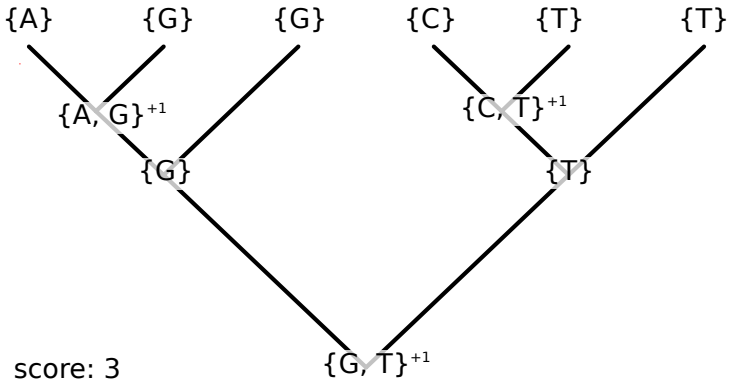
Fitch algorithm



score: 2

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Fitch algorithm



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Maximum Parsimony

There are two problems:

- How do we figure out how many evolutionary changes are implied by a particular data pattern on the tree?
- How do we find the tree with the smallest number of evolutionary changes?

First problem solved! So let's find the best tree!

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

A tree for the number of trees

# species unrooted trees	# of trees	# species rooted trees
3	1	2
4	3	3
5	15	4
6	105	5
7	945	6
8	10,395	7
9	135,135	8
10	2,027,025	9
11	34,459,425	10
12	654,729,075	11
13	13,749,310,575	12
14	316,234,143,225	13
15	7,905,853,580,625	14
50	$2.753 \cdot 10^{76}$	49

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Tree rearrangements

- The strategy of evaluating the maximum parsimony criterion for all trees (brute force) in order to find the best tree topology is in most cases highly impracticable.
- Instead, (local) tree rearrangements are used to search locally within the tree space. The idea behind such a heuristic is to use a starting tree and search locally for improved scores (parsimony, maximum likelihood, Least-Squares), until no further rearrangements can lead to a tree with a better score.

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

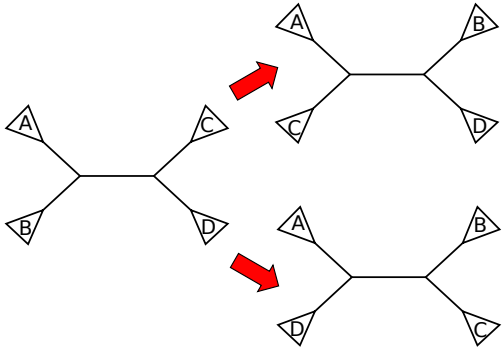
Maximum parsimony with phangorn

```
> library(phangorn)
> txt <- read.csv("cats_dogs.csv")
> head(txt)
> align <- read.phyDat("cats_dogs_mafft.fasta", format="fasta"
> names(align) <- txt$Latin_name
> image(align)
> tree_mp <- pratchet(align)
> tree_mp <- acctran(tree)
> plotBS(tree_mp, type = "phylogram")
> add.scale.bar()
```


Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

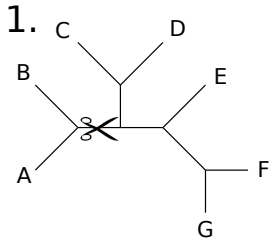
Nearest neighbor interchange

For any internal edge of a binary tree there exist three different ways to connect its four subtrees, one of which is the current tree.



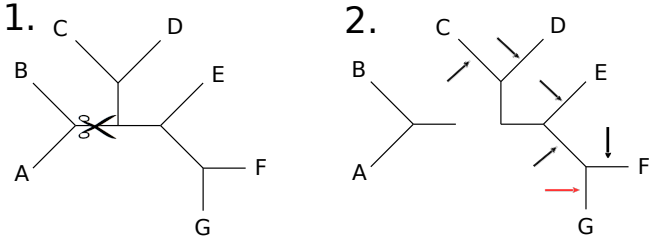
Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Subtree pruning and regrafting



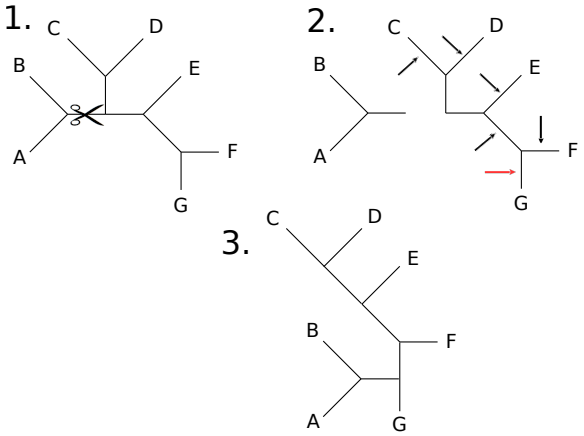
Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Subtree pruning and regrafting



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Subtree pruning and regrafting



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

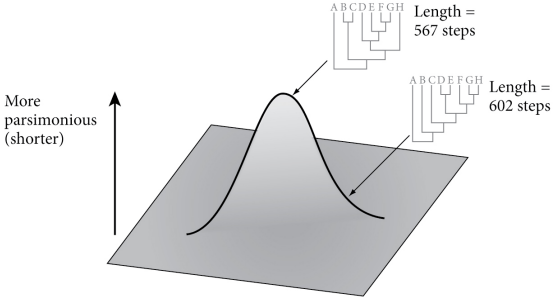
Tree space

- Tree rearrangements only guarantee to find a local optimum!
- To better explore the tree space often the search with the rearrangements is started from different starting trees. These can be random trees, random addition trees or for example the current best trees which was itself pertubated with several tree rearrangements.

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Tree space

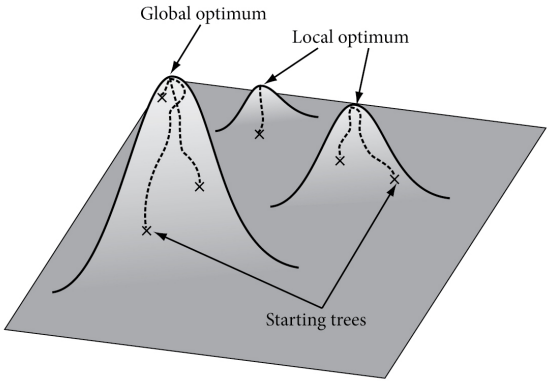
Visualization of tree spacewith the best (shortest) trees sit at the peak



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Tree space

Searching the tree space with multiple optima.



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Bootstrap

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

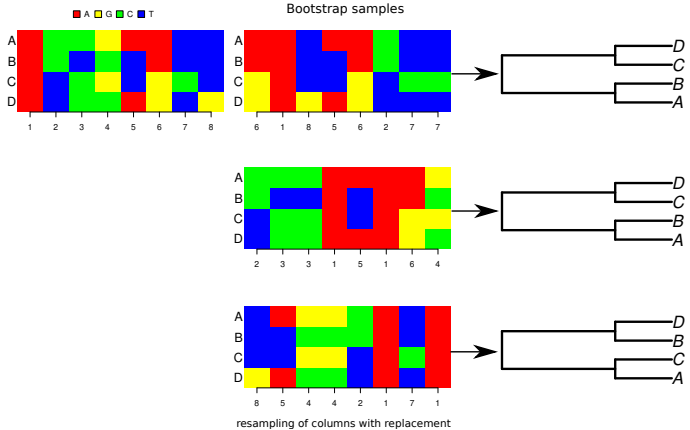
Bootstrap

General principle of the bootstrap:

1. Estimate a parameter from the data: $\hat{\theta}$
2. Resample with replacement the data.
3. Estimate the parameter $\hat{\theta}^*$ for this “bootstrap” sample.
4. Repeat steps 2 and 3 many times.
5. Assess the distribution of the $\hat{\theta}^*$'s: this gives an estimate of the error of $\hat{\theta}$ This procedure "mimicks" the process of sampling repeatedly a distribution, and makes no assumption on the shape of this distribution.

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Bootstrap



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Bipartitions / Splits

Each edge of a tree defines a bipartition (or split).
This tree has $n = 4$ tips (or leaves) and thus defines 6 splits:

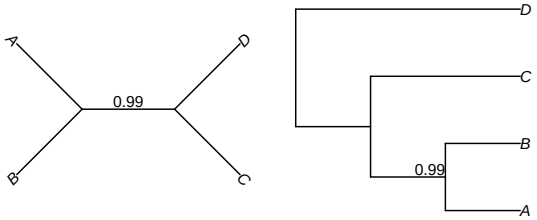
- The internal branch (edge) defines one non-trivial split: AB|CD
- The terminal branches define four trivial splits: A|BCD, B|ACD, C|ABD, D|ABC
- One additional trivial split is defined by the empty set: $\varnothing$|ABCD

An unrooted tree with n tips has $n3$ internal branches, and thus defines $2n - 2$ splits: $n - 3$ non-trivial and $n + 1$ trivial.
The number of possible splits grows exponentially with n: 2^{n-1}

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Bootstrap

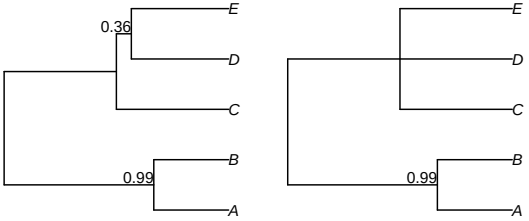
The bootstrap statistic in phylogenetics considers the $n - 3$ non-trivial splits and counts how many times they appear in the bootstrap trees. These are the bootstrap proportions (BP) and should be interpreted as measures of confidence in the estimated tree (not as probabilities).
The result can be represented graphically with the BP on the internal branch of the estimated tree – usually close to the node defining an MRCA after rooting the tree.



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Bootstrap

Often bootstrap proportions are only presented when these are high (e.g. >80%) and sometimes edge with low support are deleted (consensus tree).



Bootstrapping works for ML, MP and distance.

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Parsimony software

- PHYLIP https://evolution.genetics.washington.edu/phylip.html by Joseph Felsenstein
- Paup* https://paup.phylosolutions.com/ by David Swofford
- TNT http://www.lillo.org.ar/phylogeny/tnt/ by Pablo Goloboff
- R package phangorn https://CRAN.R-project.org/package=phangorn

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

What is likelihood?

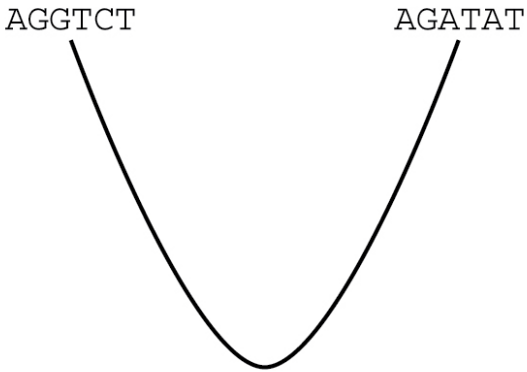
Imagine we have a bag of coins and we know they are of two types, one type is fair or unbiased. The other coins are fairly biased, you through with them head 75% of the time.

Imagine you throw a coin ten times and get the following result:

Toss	1	2	3	4	5	6	7	8	9	10	Likelihood
Result											
Prob. if fair	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.001
Prob. if biased	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.056

$L(data|hypothesis)$, here data are the results from the throws and hypothesis is the type of coin.

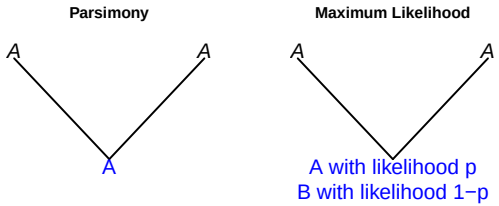
What's distance between two sequences



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

What's distance between two sequences?

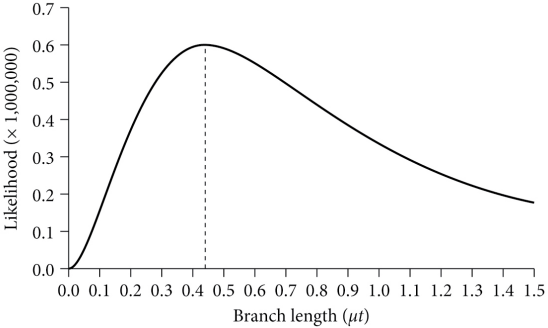
Let's take a discrete character with two states: A and B. Suppose we observe two species in state A: what was the state of their ancestor?



p will depend on the model of evolution of the character and the branch lengths.

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

What's distance between two sequences



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Bootstrap

Sp. A AGGTCT
Sp. B AGATCT

The "pattern" GA is because of:
G $\longleftrightarrow$ A
or: G $\longleftrightarrow$ C $\longleftrightarrow$ A
or: else

The "pattern" AA is because of:
No mutation
or: A $\longleftrightarrow$ C $\longleftrightarrow$ A
or: else ?
Observations in molecular biology suggest that mutations are random.

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

The substitutions of nucleotides can be modelled with a Markov chain: $A \rightleftarrows B$. One or two parameters control this model: the rates r_1 and r_2. The rate matrix is (Q):

$$\begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{pmatrix} \end{array}$$

r_1 and r_2 measure the "quantity of change" during a very short time (so short that multiple changes are impossible). Time does not appear in this formulation.

For a given time interval t the *transition* $A \longrightarrow B$ can be due to only one change or several ($A \longrightarrow B \longrightarrow A \longrightarrow B, \ldots$) Similarly if no change is observed during t ($A \longrightarrow B \longrightarrow A, \ldots$).

The *probability matrix* is computed with the matrix exponential $P(t) = e^{tQ}$ and takes these multiple changes into account.

The method is generalised to more than 2 states, e.g. { A,G,C,T }.

$$\begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \end{array}$$

This is the Jukes–Cantor (1969) model. This model is simple enough, so P can be calculated directly: its off-diagonal elements are $(1e^{4\alpha t})/4$. The expected number of substitutions along t is also simple to calculate and gives the Jukes–Cantor distance (x: proportion of sites with an observed change, or "raw" distance):

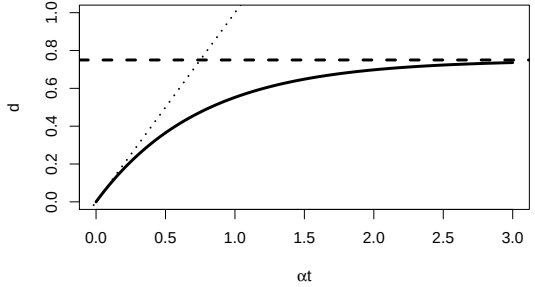
$$d = -0.75 \ln(1 - \frac{4}{3}x)$$

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Example: 20% of different sites between two sequences, so $x = 0.2$:

```
> -0.75 * log(1 - 0.2 * 4/3)
[1] 0.2326162
```

Interpretation: each site has experienced, on average, 0.23 change since both sequences separated. So, it is incorrect to write "23% divergence".



Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Other substitution models

The K80, K2P, or Kimura two parameter model.
This model has a different rate for transitions (C$\longleftrightarrow$T and A$\longleftrightarrow$G) and transversions (all other substitutions).

$$\begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{pmatrix} \end{array}$$

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Other substitution models

Model:	Summary
JC69 (Jukes & Cantor 1969)	Equal rates; equal nucleotide frequencies.
K80 (Kimura 1980)	Different rates for transitions and transversions.
F81 (Felsenstein 1981)	Equal rates; different nucleotide frequencies.
HKY85 (Hasegawa et al. 1985)	Different rate for transitions and transversions; unequal base frequencies.
F84 (Felsenstein 1984)	Similar to HKY85 but with two additional rates (for pyrimidines and purines).
TN93 (Tamura & Nei 1993)	Two types of transversions; one type of transition; unequal base frequencies.
GTR (several refs.)	All rates different; all base frequencies different.

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Model Selection

Use programs like ModelTest, ProtTest to choose the best fitting substitution model!

	Model	df	logLik	AIC	AICw	AICc	AICcw	BIC
1	JC	91.00	-54303.67	108789.35	0.00	108794.77	0.00	109341.20
2	JC+I	92.00	-50672.85	101529.71	0.00	101535.25	0.00	102087.63
3	JC+G	92.00	-48684.10	97552.19	0.00	97557.74	0.00	98110.11
4	JC+G+I	93.00	-48588.86	97363.73	0.00	97369.39	0.00	97927.71
5	F81	94.00	-54212.64	108613.27	0.00	108619.06	0.00	109183.32
6	F81+I	95.00	-50548.97	101287.94	0.00	101293.86	0.00	101864.05
7	F81+G	95.00	-48500.49	97190.99	0.00	97196.90	0.00	97767.10
8	F81+G+I	96.00	-48401.46	96994.92	0.00	97000.96	0.00	97577.10
9	HKY	95.00	-51275.86	102741.72	0.00	102747.64	0.00	103317.83
10	HKY+I	96.00	-47450.80	95093.59	0.00	95099.64	0.00	95675.77
11	HKY+G	96.00	-44893.04	89978.08	0.00	89984.13	0.00	90560.26
12	HKY+G+I	97.00	-44762.63	89719.27	0.00	89725.44	0.00	90307.51
13	GTR	99.00	-50758.41	101714.83	0.00	101721.26	0.00	102315.20
14	GTR+I	100.00	-47079.80	94359.60	0.00	94366.16	0.00	94966.03
15	GTR+G	100.00	-44746.72	89693.44	0.00	89700.00	0.00	90299.87

Klaus Schliep, PhD, IMBT
klaus.schliep@tugraz.at

Software

General software for estimating distance an ML trees

- PHYLIP https://evolution.genetics.washington.edu/phylip.html
- Paup* https://paup.phylosolutions.com/
- R package phangorn https://CRAN.R-project.org/package=phangorn

Fast and more specialized Likelihood tools:

- RAxML https://cme.h-its.org/exelixis/web/software/raxml/
- iqtree http://www.iqtree.org/

Bayesian phylogentic software:

- RevBayes https://revbayes.github.io/
- BEAST2 https://paup.phylosolutions.com/