Bioinformatik

Stefanie Hartmann

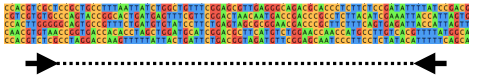
Wintersemester 2019 / 2020, Universität Potsdam

Phylogenetics
Nov 29, 2019

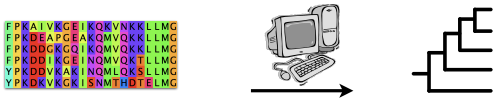
Now that you've got a MSA...

- characterization of conserved/functional domains

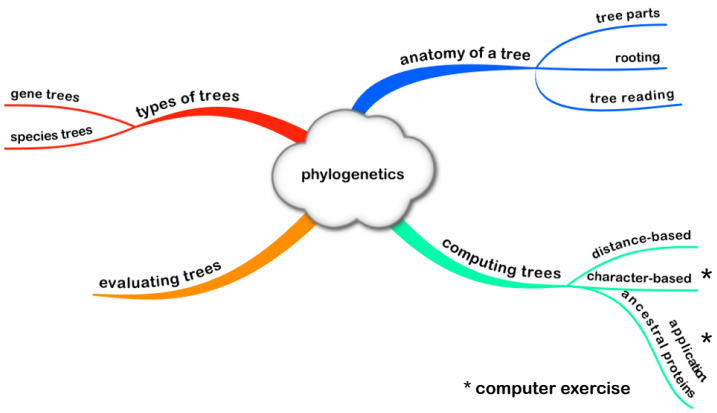
- design of PCR-primers



- evolutionary analysis of genes/organisms



Overview

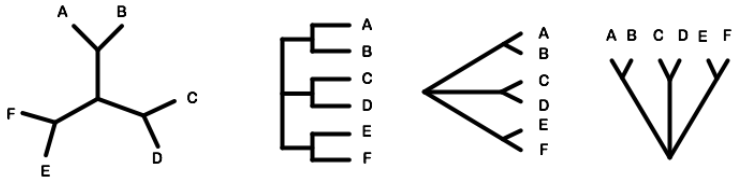


A phylogenetic tree (phylogeny)...

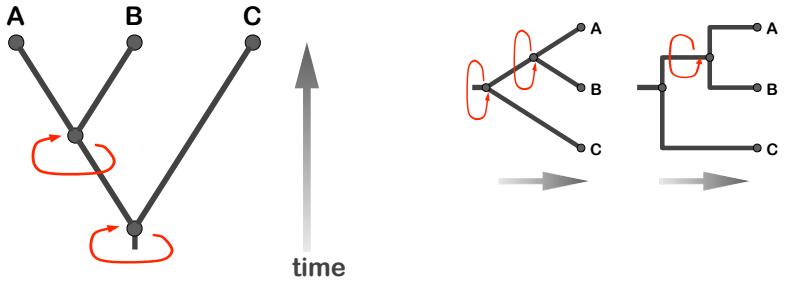
... is a hypothesis that depicts the historical relationships among entities in a branching diagram

DNA or protein sequences

organisms



Anatomy of a phylogeny

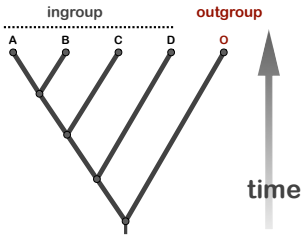


- leaves/tips (A, B, C): species, genomes, individuals, genes, ...
- internal nodes (D, E): hypothetical ancestors that split into two lineages
- branches: represent evolution of genes/taxa
- root: the branch leading to the common ancestor of all genes/taxa; the internal node that is the common ancestor of all genes/taxa

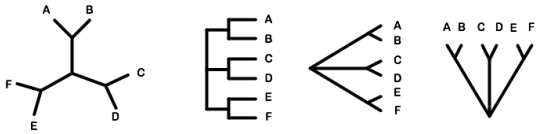
Rooted phylogenies

outgroup:

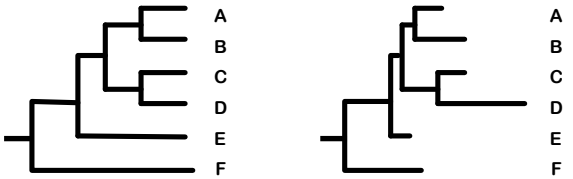
- a sequence that is more distantly related to each of the ingroup sequences than these are to each other, based on external information



unrooted:



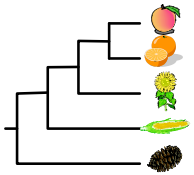
Cladogram vs. phylogram



Species tree vs. gene tree

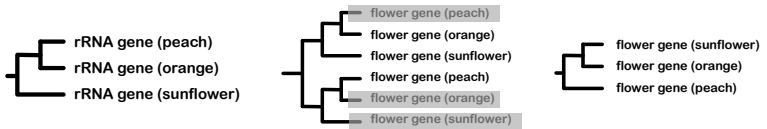
species tree:

- evolution of organisms, species
- derived from one or more data sources / data types

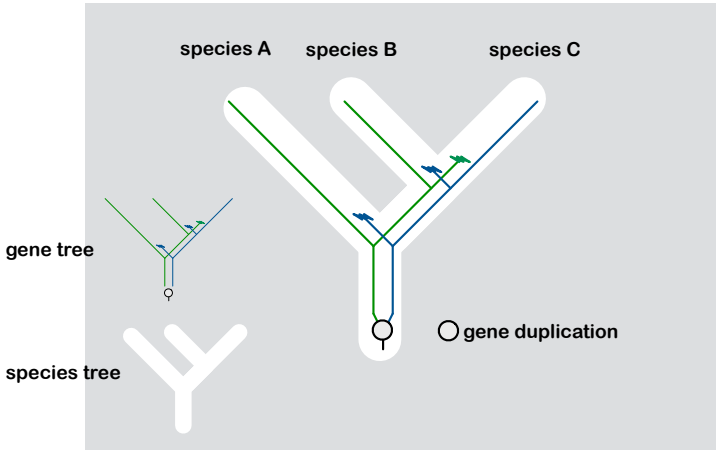


gene tree:

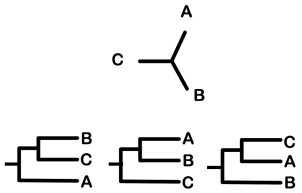
- evolution of gene sequences, gene function
- computed from gene sequences
- sometimes (!) corresponds to the species tree



Species tree vs. gene tree



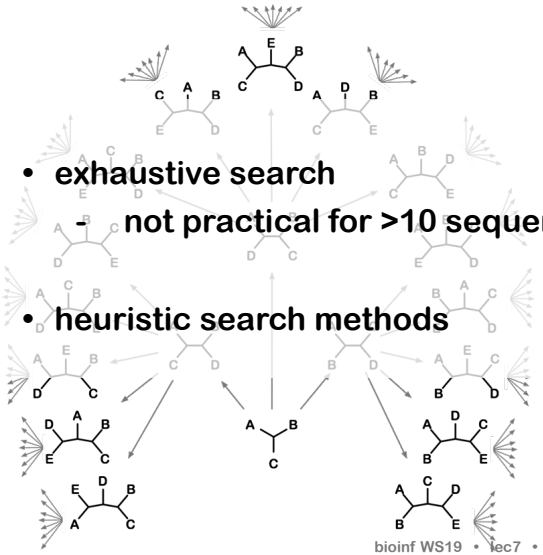
Number of possible trees



species / sequences	unrooted trees	rooted trees (bifurcated)
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425

$$\frac{(2n-5)!}{2^{n-3}(n-3)!} \qquad \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

Number of possible trees



- exhaustive search
  - not practical for >10 sequences
- heuristic search methods

How to infer a phylogeny?

start with an optimal multiple sequence alignment

- distance methods
  - Neighbor Joining

- character-based methods
  - parsimony
  - statistical methods
    - maximum likelihood
    - Bayesian inference

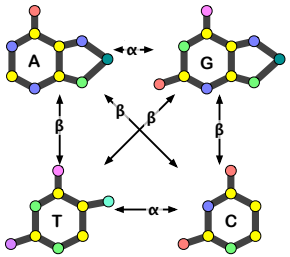
How to obtain a distance matrix

models of amino acid replacement

- PAM, BLOSUM, etc

models of DNA replacement

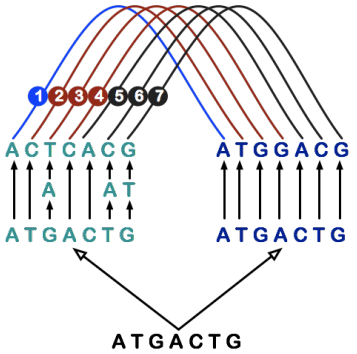
- use a model that will correct for multiple substitutions



	A	C	G	T
A				
C				
G				
T				

	A	C	G	T
A				
C				
G				
T				

Substitution patterns



ACTCACG
ATGGACG

<u>one apparent change:</u>

2. single substitution
3. multiple substitution
4. coincidental substitution

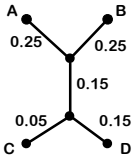
<u>no apparent change:</u>

1. no change
5. parallel substitution
6. convergent substitution
7. back substitution

From a distance matrix to a tree

topology & branch lengths of the computed tree should reflect values in distance matrix

	A	B	C	D
A	-	0.5	0.4	0.6
B		-	0.5	0.5
C			-	0.2
D				-



Neighbor Joining

- a very fast heuristic algorithm
- finds the tree that best fits a distance matrix
- outputs a single unrooted tree

Character-based methods

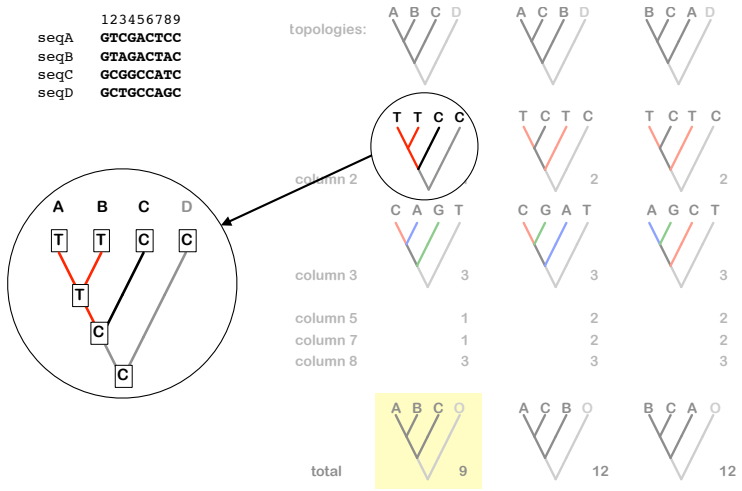
maximum parsimony

- the best tree is the one that requires the fewest mutations in the sequence data

maximum likelihood

- statistical approach
- probability of the sequence data, given a tree topology and a substitution model
- the best tree is the one with the highest probability, <u>under a given substitution model</u>

Maximum parsimony



Maximum parsimony

disadvantages

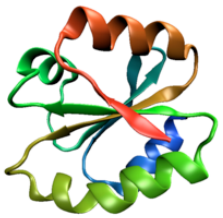
- non-probabilistic: it is difficult to evaluate results in a statistical framework
- does not correct for multiple substitutions

advantages

- can be used for non-molecular characters (morphology)
- provides exact mapping of characters along branches

Ancestral sequence reconstruction

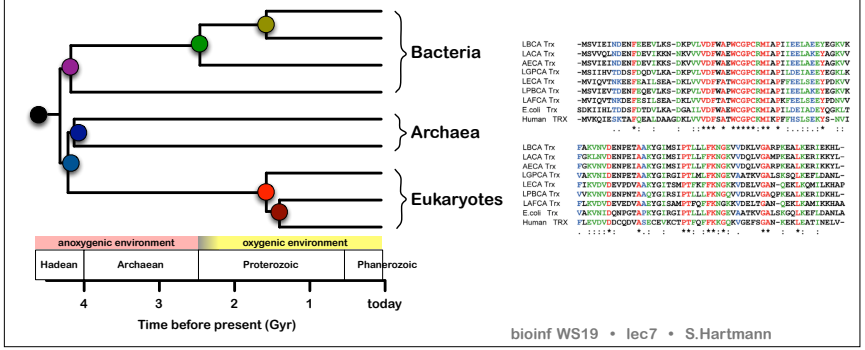
- reconstruction of enzymes from extinct species
- study properties of these enzymes
- examine ancient environmental conditions
- understand how life has evolved



- example: thioredoxin (Trx)
  - oxidoreductase activity, found in all domains of life
  - conserved active site and fold
  - probably present in primitive forms of life
- R Perez-Jimenez et al., Nat. Struct. Mol. Biol. 2011

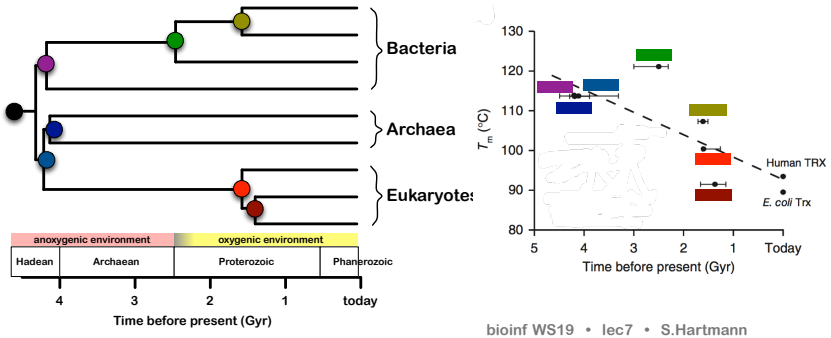
Ancestral sequence reconstruction

- retrieval of 203 thioredoxins in GenBank
- multiple sequence alignment
- phylogenetic analysis
- inference of ancestral amino acid sequences
  (not parsimony but maximum likelihood)



Resurrection of ancestral proteins

- analysis of enzymes
  - thermal stability (higher temps!)
  - activity under various conditions (more acidic!)
  - chemical mechanism of reduction (unchanged!)



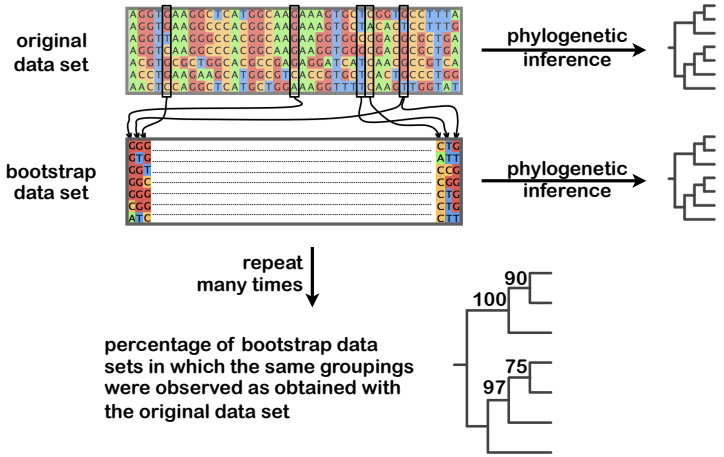
Assessing confidence

- phylogenetic methods will always compute at least one tree
- even random data will have some 'best' tree

approaches:

- test for phylogenetic signal in the data
- evaluate reliability of reconstructed branches/clades
  - bootstrapping

Bootstrapping



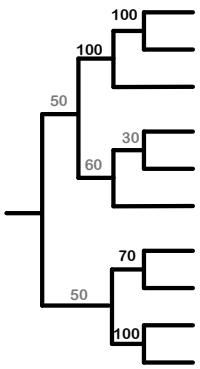
percentage of bootstrap data sets in which the same groupings were observed as obtained with the original data set

Evaluating confidence in clades: bootstrapping

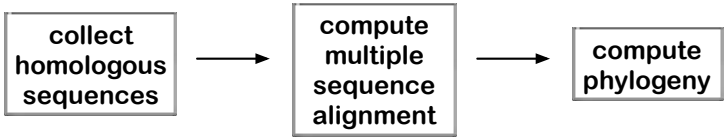
- resample the alignment (columns) with replacement
  - the original data are the columns of the multiple alignment
  - sample X number of new alignments, of the same length as the original alignment
- compute a phylogeny for each alignment
  - count how many times a branch appears that also exist in the original tree
- label branches from the original (best) tree with bootstrap proportions/percentage

Evaluating confidence in clades: bootstrapping

- bootstrapping measures how consistently the data support certain clades
- high bootstrap values (close to 100%) mean uniform support
- low bootstrap values (below 50%) are meaningless
- bootstrap values do not indicate whether or not the tree is correct



Many decisions!

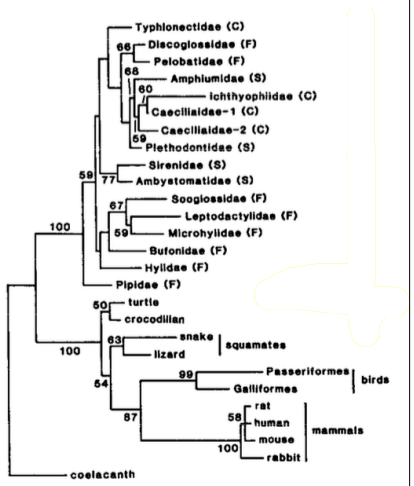


- several steps
- for each step: several general approaches / programs
- for each approach / program: several parameters

Do it right!

1990 study of 18S rRNA

- PCR-amplify
- isolate & sequence
- align
- phylogenetic analysis
  - MP, NJ
  - bootstraps
- birds cluster with mammals!



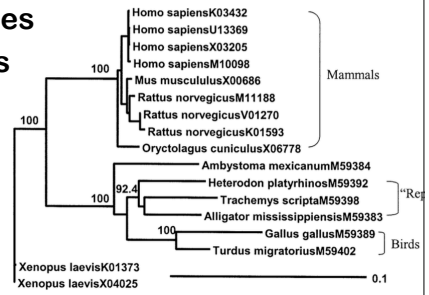
Do it right!

1990 study

- contradicts other studies

re-analysis by other authors

- correct MSA
- correct substitution parameters
- higher-quality sequences
- birds cluster with reptilians



2000). This study highlights the problem of applying a battery of computer programs to the data without first checking the quality of the data and emphasizes the importance of becoming intimately familiar with the data.

Today's exercise

the data

- 5 globin homologs

the analysis

- compute a phylogeny using maximum parsimony
- view and interpret the phylogeny
- reconstruct ancestral amino acids

Key terms and concepts

- anatomy of a phylogeny
- rooted vs. unrooted trees
- gene tree vs. species tree
- main differences between distance-based and character-based methods
- why are multiple substitutions a problem for phylogenetic inference?
- principle of Neighbor Joining
- principle of Maximum Parsimony
- bootstrap support: goal, procedure, interpretation