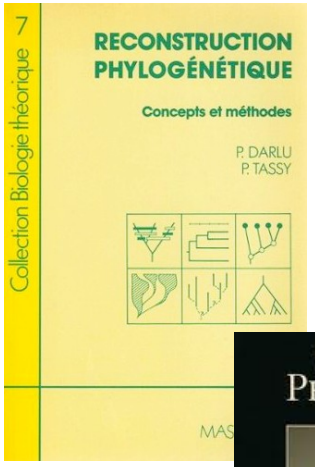
Master d'informatique, option ABC – Institut Gaspard Monge

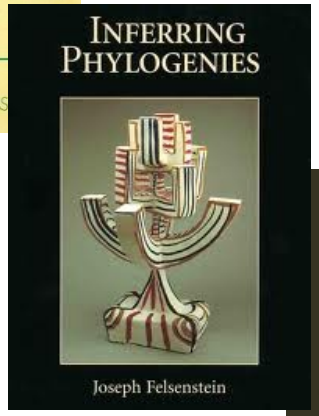
Algorithmique pour la bioinformatique

Algorithmics & Phylogenetics

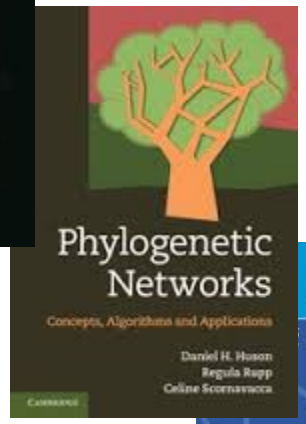


Darlu & Tassy, 1993

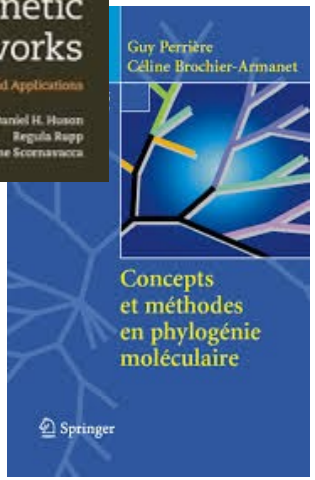
http://sfs.snv.jussieu.fr/?q=node/11



Felsenstein, 2002

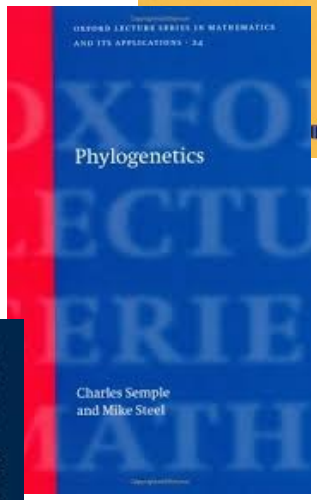


Huson, Rupp & Scornavacca, 2010

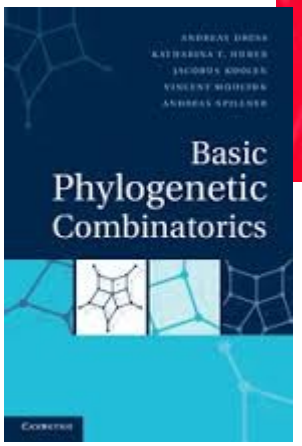


Perrière & Brochier-Armanet, 2010

Barthelémy & Guénoche, 1988



Semple & Steel, 2003



Dress, Huber, Koolen, Moulton & Spillner, 2011

UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

GASPARD MONGE INSTITUT

Philippe Gambette

References

Finding **bibliographic references**:

- Google Scholar: large coverage, including noise (preprints, fake papers...)
- Bibliographic resources with university access:

http://www.u-pem.fr/bibliotheque/consulter-les-ressources-en-ligne/ressources-en-ligne-de-a-a-z/

→ Science Direct for Elsevier journals, Springer, JSTOR, etc.

- University library

Conferences with computer science papers applied to phylogenetics:

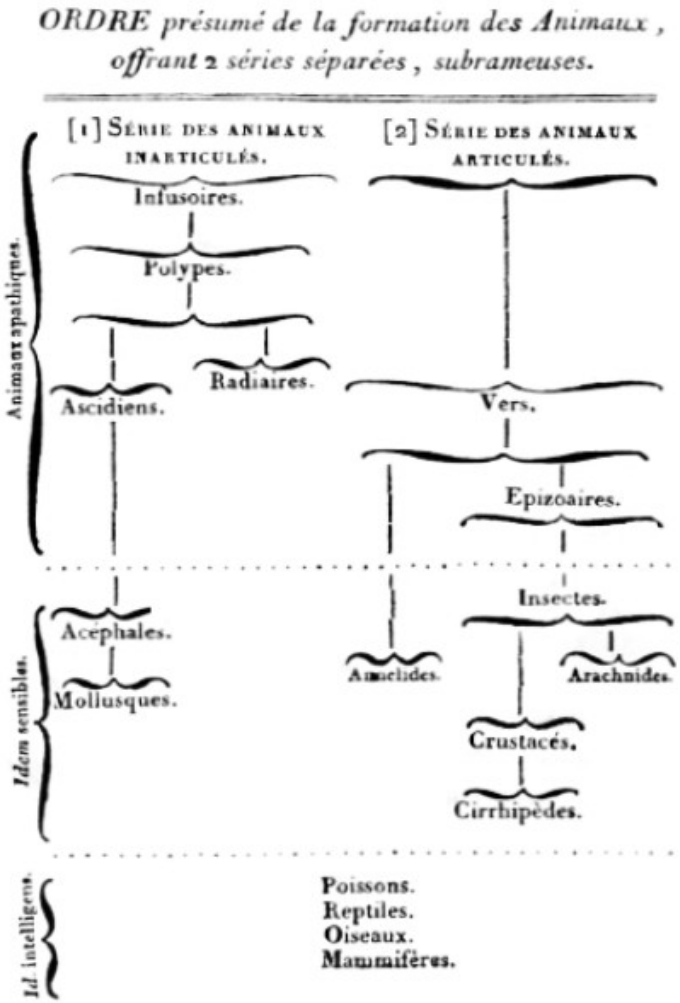
- International: ISMB, RECOMB, WABI, ECCB, ISBRA,
+ algorithmics conferences: SODA, CPM, ISAAC, COCOON, etc.
- In France: JOBIM, Alphy

Scholarly organizations:

Phylogenetic trees

Phylogenetic tree of a set of species:

- **organizing** them according to common characters **classification**
- describing their evolution



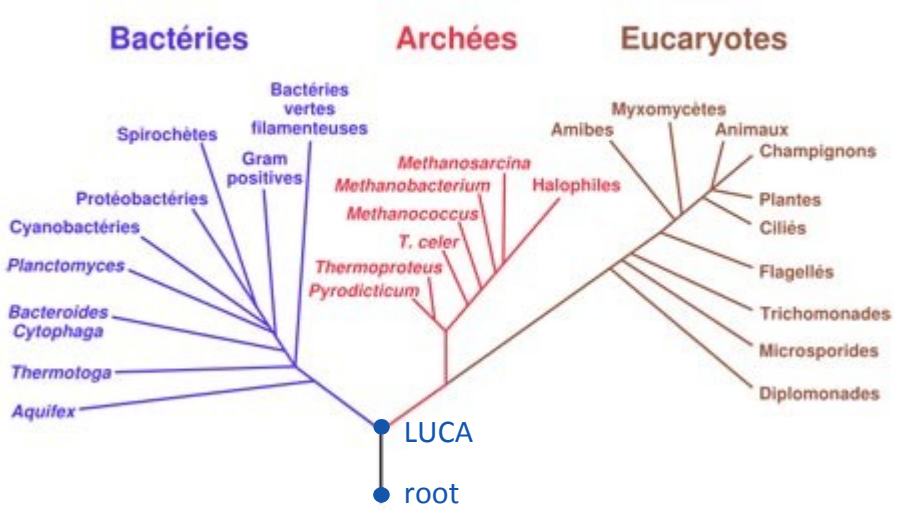
Lamarck : Histoire naturelle des animaux sans vertèbres (1815)

Phylogenetic trees

Phylogenetic tree of a set of species:

- organizing them according to common characters
- **describing** their evolution

modelization



Woese, Kandler, Wheelis : Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, Proceedings of the National Academy of Sciences, 87(12), 4576–4579 (1990)

Encoding a tree

"Newick" format:

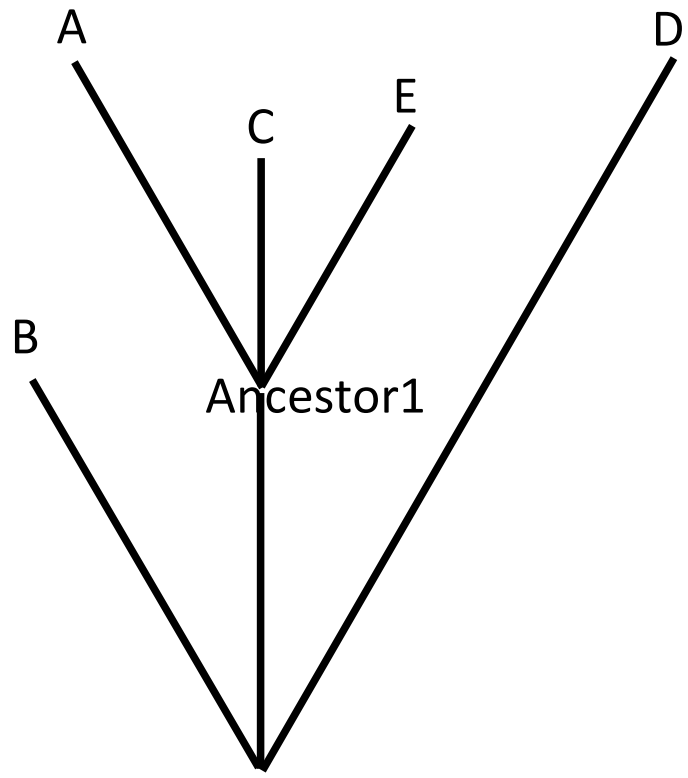
(B:6.0,(A:5.0,C:3.0,E:4.0)Ancestor1:5.0,D:11.0);

Not unique:

((A:5.0,C:3.0,E:4.0)Ancestor1:5.0,B:6.0,D:11.0);

Possible to have nodes which are not binary:

- uncertainty
- or speciation known to be at the same time



http://evolution.genetics.washington.edu/phylip/newicktree.html

Properties of phylogenetic trees

Characterizing a tree with:

- its “clusters”: one cluster of T = the set of leaves below one vertex of T
- its “triplets”: one triplet of T = a tree on 3 leaves contained in T

Properties of rooted and unrooted trees

Clusters: “laminar family”, i.e. it contains no overlapping sets

→ reconstruction from clusters: Hasse Diagram of the cluster inclusion graph

Triplets: do not contain {ab|c, b|cd, a|bd} or {ab|c, b|cd, ad|b}

Guillemot & Mnich, Kernel and fast algorithm for Dense Triplet Inconsistency, 2013

Splits: “compatible split system”, i.e. for any pair of splits A1|B1, A2|B2, at least one of the sets $A1 \cap A2$, $A1 \cap B2$, $B1 \cap A2$, $B1 \cap B2$ is empty

Quartets: for any leaf *e*, $ab|cd \in Q \Rightarrow ab|ce \in Q$ or $ae|cd \in Q$

Bandelt & Dress, Reconstructing the shape of a tree from observed dissimilarity data, 1986

Properties of rooted and unrooted trees

Tree distances:

Characterized by the **four-point condition**:

For all a, b, c, d, $d(a,b)+d(c,d) \leq \max\{d(a,c)+d(b,d), d(a,d)+d(b,c)\}$.

Given a tree distance, only one possible tree

Tree distances when the tree contains a center at equal distance from all leaves:

Characterized by the **ultrametric inequality**:

For all a, b, c, $d(a,b) \leq \max\{d(a,c), d(b,c)\}$

molecular clock hypothesis!

Reconstructing a tree from an ultrametric

UPGMA algorithm (Unweighted Pair Group Method with Arithmetic Mean):

- Initialize all clusters with leaf singletons
- While there are more than 2 clusters:

    - pick the nearest two clusters

    - combine them and update the distance matrix with average values (average weighted by the size of each of the two clusters)

Sokal & Michener, A statistical method for evaluating systematic relationships, 1958

→ Correctly reconstructs ultametric distances, but not all tree distances

→ Neighbor-Joining...

Reconstructing a tree from its triplets

BUILD algorithm:

- Build the following graph: leaves as vertices; for each triplet a|bc, add edge bc.
- While there is more than one connected component:
  - each connected component corresponds to one subtree
  - recursively apply the algorithm on the leaf set of each connected component

Aho, Sagiv, Szymanski & Ullman, Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions, 1981.

→ When missing triplets, efficient implementation in $O(|T|+n^2 \log n)$

Henzinger, King & Warnow, Constructing a tree from homeomorphic subtrees, with applications to computational evolutionary biology, 1999.

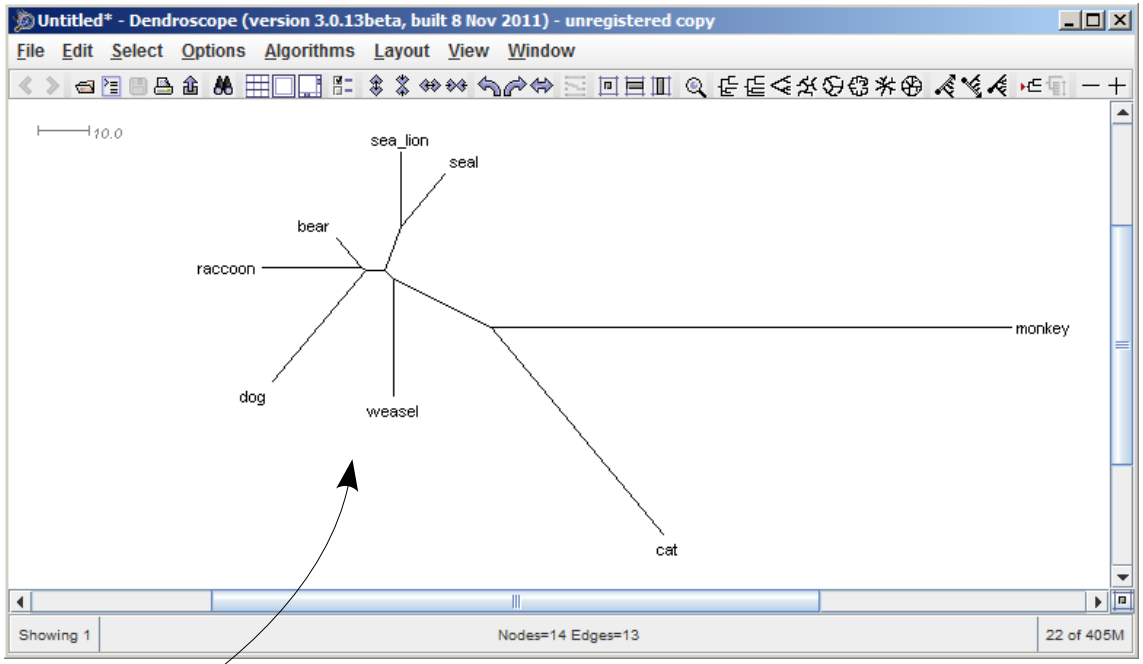
Visualizing phylogenetic trees

Visualize branch lengths or not

Several kinds of visualizations:

- rectangular phylogram
- rectangular cladogram
- slanted cladogram
- circular phyogram
- circular cladogram
- inner circular cladogram
- radial phylogram
- radial cladogram

((raccoon:19.19959,bear:6.80041):0.84600,((sea_lion:11.99700, seal:12.00300):7.52973,((monkey:100.85930,cat:47.14069):20.59201, weasel:18.87953):2.09460):3.87382,dog:25.46154);

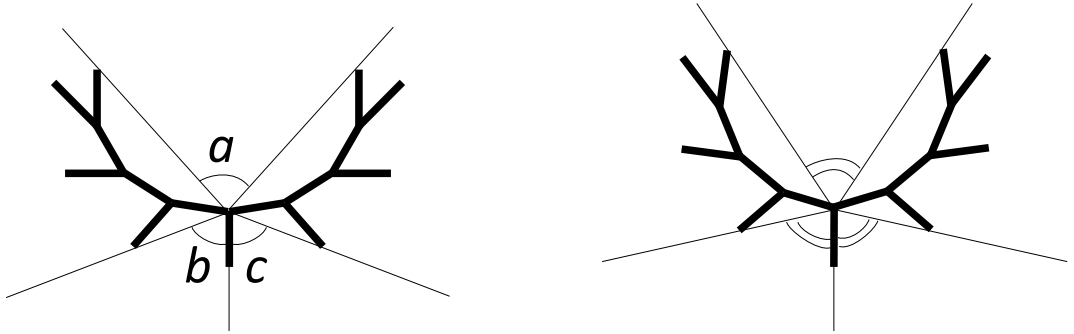


Visualizing phylogenetic trees: radial phylogram

"Equal angle" algorithm to draw a radial phylogram on *n* leaves:

- Compute the angles "bottom-up" starting with angle $2i\pi/n$ for leaf *I*
- Locate the nodes "top-down" using:
  - the angles
  - the edge lengths
- Add the labels (avoiding overlap)

"Equal daylight" algorithm to optimize used space:



Felsenstein, Inferring phylogenies, 2002, 579-583

Comparing trees

- **Maximum Agreement Subtree** (MAST):

  - Given T1 and T2 on the set X of leaves, an *agreement subtree* T of T1 and T2, on a subset X' of leaves, is such that T1 and T2 restricted to X' are equal to T.

  - a *maximum agreement subtree* is an agreement subtree of maximum size

- **Maximum Compatible Tree** (MCT):

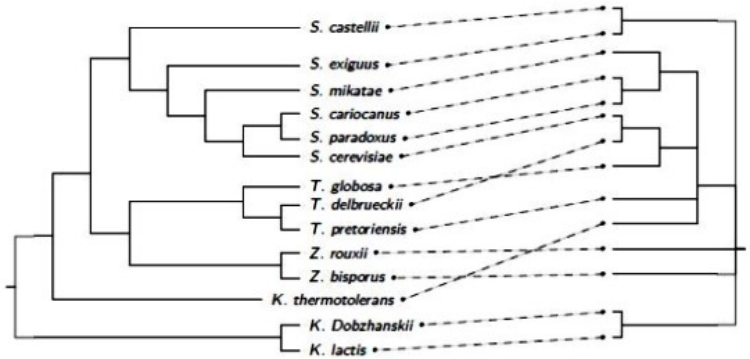
  - if a tree is not binary, several binary trees *refine* it

  - a compatible tree on a subset X' of leaves is a binary tree which refines the trees T1 and T2 restricted to the leaves of X'.

Berry & Nicolas, Maximum Agreement and Compatible Supertrees, 2004

- **Tanglegrams**:

  - display both trees for visual comparison, linking their leaves with edges, minimizing edge crossings.
  → general problem NP-complete
  → planar embedding in linear time
  → if one tree is fixed, O(n log n)



Venkatachalam, Apple, St John, Gusfield, Untangling tanglegrams: comparing trees by their drawings, 2010

Comparing trees

Distances between trees:

- **Robinson Foulds distance** between T1 and T2:

  - Number of different splits ("symmetric difference metric")

  Robinson and Foulds, Comparison of phylogenetic trees, 1981

  - Minimum number of edge contractions/decontractions to go from T1 to T2

- **quartet distance** between T1 and T2:

  - Number of different quartets

- **SPR distance** between T1 and T2:

  - Minimum number of SPR moves to go from T1 to T2

Exploring the tree space

NNI: nearest neighbor interchange

Consider an edge *e* and exchange the adequate subtrees connected to *e.*

SPR: subtree pruning and regrafting

Disconnect a subtree and reattach it somewhere else.

TBR: tree bisection and reconnection

Delete an edge in the tree, reconnect the two parts with a new edge anywhere.

An NNI is a special kind of SPR, which is a special kind of TBR.

NNIs allow to explore the whole tree space. *Proof: induction on…*

Felsenstein, Inferring phylogenies, 2002, 39-44

Exploring the tree space... to find the optimal tree

Exploring the tree space is useful to find the optimal topology for:

Felsenstein, Inferring phylogenies, 2002, 39-44
http://evolution.genetics.washington.edu/phylip/software.html

- **Parsimony**

Given the tree topology, find the scenario which explains current genetic sequences with the minimum number of operations along the tree edges

- **Likelihood**

Given the tree topology and a statistical model of evolution, find the scenario which produces current genetic sequences with the highest probability

Models of evolution: Jukes Cantor'69, Kimura'80, Felsenstein'81

http://en.wikipedia.org/wiki/Models_of_DNA_evolution

- **Distance optimization**

Given the tree topology, find edge lengths which best explain distance data between current genetic sequences

Tree quality: Is the obtained tree "robust"?

Bootstrap: apply the same algorithm on "resampled" data

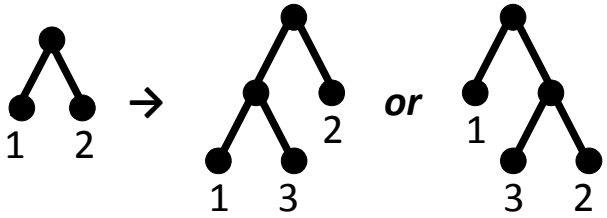
Exploring the tree space... by randomly generating trees

Which model do you choose to randomly generate (rooted binary) trees?

→ *Random tree generation also used to simulate data to test algorithms!*

- Labeled tree **equiprobability**

- **Yule-Harding model** :



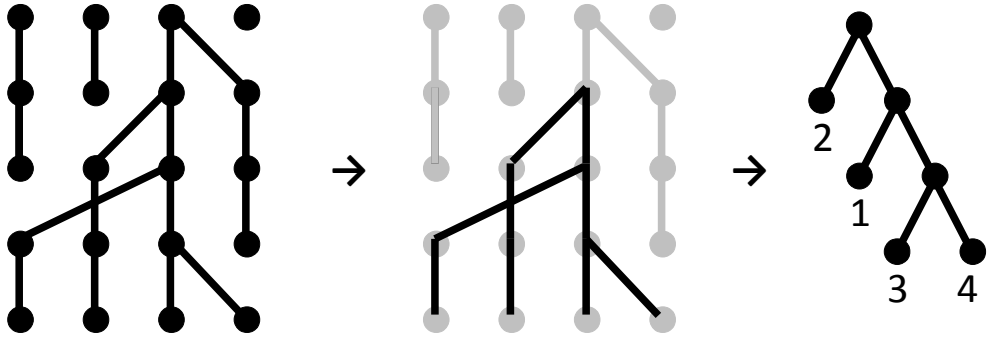
  - start from a root with two labeled children
  - choose one leaf at random to split it, creating a new labeled leaf

Felsenstein, Inferring phylogenies, 2002, 559-562

- **Kingman's coalescent model** (population genetics) :

  - start from a population of *n* leaves (each leaf representing a gene copy)
  - the probability that two gene copies come from the same copy in the previous generation is 1/2*n*

equivalent to repeating, for *k* gene copies:
- go back ≈4*n*/(*k*(*k*-1)) generations in time
- combine 2 random lineages
- decrease *k* by 1



Felsenstein, Inferring phylogenies, 2002, 450-460

Exercises

Exercise 1 – Tree shapes and random generation models

Q1. Evaluate the probability of each rooted binary tree shape on 4 leaves, for each of the three random generation models

Q2. Evaluate the probability of the rooted binary caterpillar tree (i.e. a tree where no node has two children having two children) on n leaves with the Yule-Harding model as well as with Kingman's coalescent model

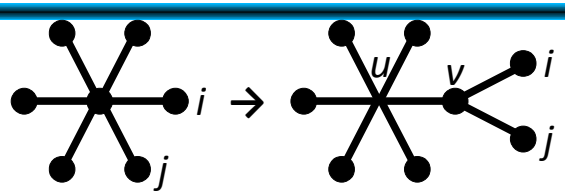
Exercise 2 – Characterization of laminar families

Given the clusters C_i of a rooted tree (not necessarily binary):

Q1. Give a characterization of the boolean matrix M_{ij} such that each line represents a cluster C_i, each column represents a leaf x_j, and $M_{ij} = 1$ if $x_j \in C_i$

Q2. Give a characterization of the intersection graph of $\{C_i\}$, i.e. the graph whose vertices represent the clusters, whose edges link all pairs of intersecting clusters

Exercises



Exercise 3 – Neighbor-Joining algorithm

The Neighbor-Noining algorithm (NJ) is a tree reconstruction algorithm which identifies at each step the **two neighbor leaves** which **minimizes the total expected length of the tree**, and **replaces them by their parent**.

Q1. Given a tree T_{ij} made of a central vertex u with n-2 leaf neighbors, as well as a neighbor v of u having two leaf neighbors i and j, and an additive metric d corresponding to T_{ij}, evaluate the total length L_{ij} of T depending on: the distance d(i,j) between i and j, the sum of all distances between leaf neighbors of u, and the sum of all distances between leaf neighbors of u on the one side and i and j on the other side.

Q2. Rewrite L_{ij} to express this total length depending on the sum of distances between all pairs of leaves of T_{ij}, as well as r_i and r_j, where r_x is the sum between leaf x and all other leaves of T_{ij}.

Q3. The NJ algorithm consists in repeating, starting from a star tree: choose two vertices i and j which minimize L_{ij} and replace them by node v in the distance matrix corresponding to d. Give an appropriate formula to compute $d(v,k)$ for each leaf k of T_{ij} (including i and j: for them, use $d(i,j)$, r_i and r_j).

Dealing with real data to build the tree of life

The model of evolution seen so far is **too simple**, not only mutations but also:

- deletions
- insertions
- duplications (paralogs), tandem duplications
- inversions
- translocations
- gene transfer across species / hybridization

Differences (number of leaves, tree topology, etc.):

- between gene trees
- between gene tree and species tree

→ "Tree of 1 percent" (31-protein tree of life) Dagan & Martin, The tree of one percent, 2006
→ Consensus tree (same leaf set) / supertree (partial leaf sets)
→ Reconciliation between trees
→ Duplication/Loss/Transfer models

Maximizing triplet consistency

We have seen BUILD algorithm to reconstruct a tree from its triplets.

Can we reconstruct the tree if there are errors in the triplets?

Triplet edition problem:
Input: set X of leaves, set R of triplets, positive integer $k \leq n$.
Output: yes if there exists a tree containing k triplets of R, no otherwise.

NP-complete: reduction from Cyclic ordering

Cyclic ordering problem:
Input: set A of elements, set C of ordered triples (a,b,c) of distinct elements of A
Output: yes if there exists a bijection $f: A \rightarrow [1..|A|]$ such that for each (a,b,c) in C, either $f(a)<f(b)<f(c)$ or $f(b)<f(c)<f(a)$ or $f(c)<f(a)<f(b)$

Jansson, On the Complexity of inferring rooted evolutionary trees, 2001

Maximizing triplet consistency

Triplet edition problem:
Input: set X of leaves, set R of triplets, positive integer $k \leq n$.
Output: yes if there exists a tree containing k triplets of R, no otherwise.

Cyclic ordering problem:
Input: set A of elements, set C of ordered triples (a,b,c) of distinct elements of A
Output: yes if there exists a bijection $f: A \rightarrow [1..|A|]$ such that for each (a,b,c) in C, either $f(a)<f(b)<f(c)$ or $f(b)<f(c)<f(a)$ or $f(c)<f(a)<f(b)$

Check that Triplet edition is in NP (solution checked in polynomial time):
→ BUILD algorithm!

Reduction:
Given an instance of the Cyclic ordering problem, build an instance of the Triplet edition problem:
- $X=A \cup \{x_0,x_1,x_2,...,x_{|C|}\}$, $k=|A|(|A|-1)/2+2|C|$;
- for all $a \neq b$ in X, add $b|ax_0$ and $a|bx_0$ to R;
- for each i in $[1..|C|]$, add $b|ax_i$, $c|bx_i$ and $a|cx_i$ to R.

Jansson, On the Complexity of inferring rooted evolutionary trees, 2001

Maximizing triplet consistency

Removing *k* triplets to obtain a triplet set consistent with a tree?

NP-complete

But fixed-parameter tractable using the “obstructions” (or “conflicts”):
“do not contain {ab|c, b|cd, a|bd} or {ab|c, b|cd, ad|b}”

“Bounded search tree” algorithm:
while there is a conflict, solve it in all possible ways

$\Rightarrow O(3^k)$ time algorithm

Further work for the second part of the semester...

Programming project

Software for dynamic representations of phylogenetic trees:
Morphing algorithm to transform one unrooted tree into another

In-depth study of research articles

Jesper Nielsen, Anders K Kristensen, Thomas Mailund & Christian NS Pedersen (2011) A sub-cubic time algorithm for computing the quartet distance between two general trees, *Algorithms for Molecular Biology* 6:15 doi:10.1186/1748-7188-6-15

Jakub Truszkowski, Yanqi Hao & Daniel G Brown (2012) Towards a practical O(*n* log *n*) phylogeny algorithm, Algorithms for Molecular Biology, 7:32 doi:10.1186/1748-7188-7-32

MR Henzinger, V King & T Warnow (1999) Constructing a tree from homeomorphic subtrees, with applications to computational evolutionary biology. *Algorithmica* 24:1-13, 1999. doi: 10.1007/PL00009268.