Representation of a dissimilarity matrix using reticulograms

Pierre Legendre

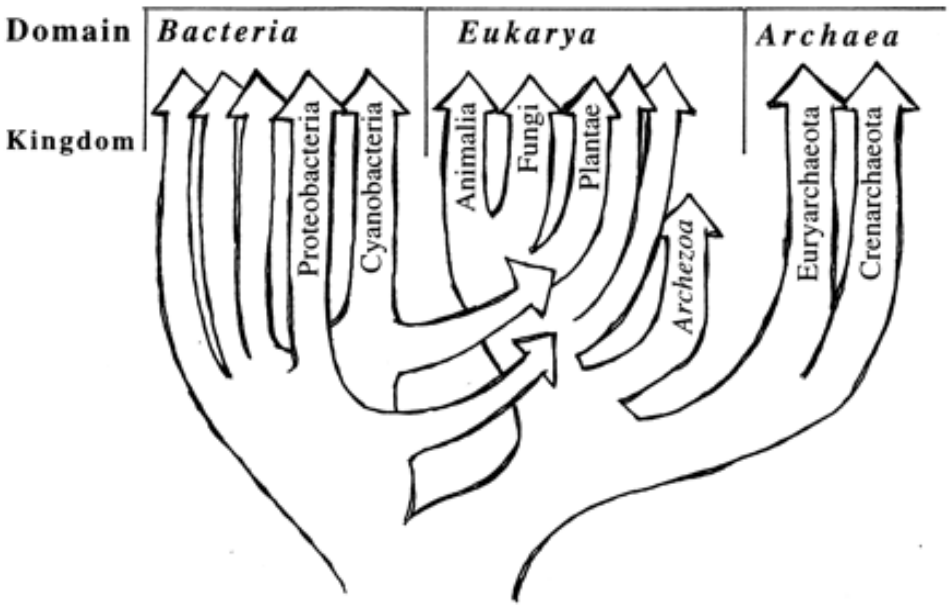
Université de Montréal

and

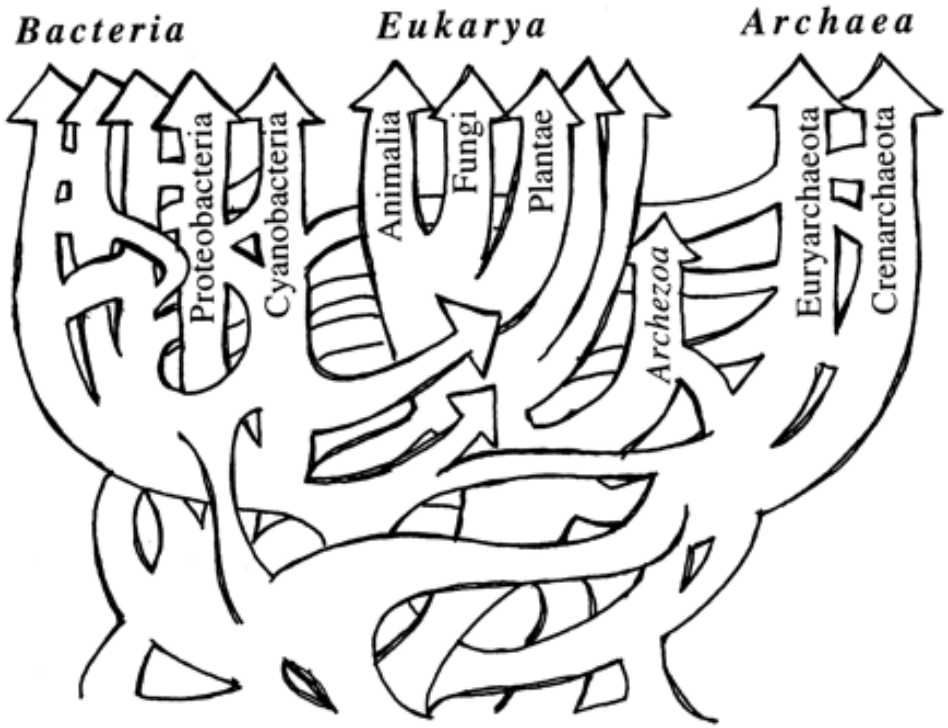
Vladimir Makarenkov

Université du Québec à Montréal

DIMACS Workshop on Reticulated Evolution, Rutgers University, September 20-21, 2004



The neo-Darwinian tree-like consensus about the evolution of life on Earth (Doolittle 1999, Fig. 2).



A reticulated tree which might more appropriately represent the evolution of life on Earth (Doolittle 1999, Fig. 3).

Reticulated patterns in nature

at different spatio-temporal scales

Evolution

1. Lateral gene transfer (LGT) in bacterial evolution.

2. Evolution through allopolyploidy in groups of plants.

3. Microevolution within species: gene exchange among populations.

4. Hybridization between related species.

5. Homoplasy, which produces non-phylogenetic similarity, may be represented by reticulations added to a phylogenetic tree.

Non-phylogenetic questions

6. Host-parasite relationships with host transfer.

7. Vicariance and dispersal biogeography.

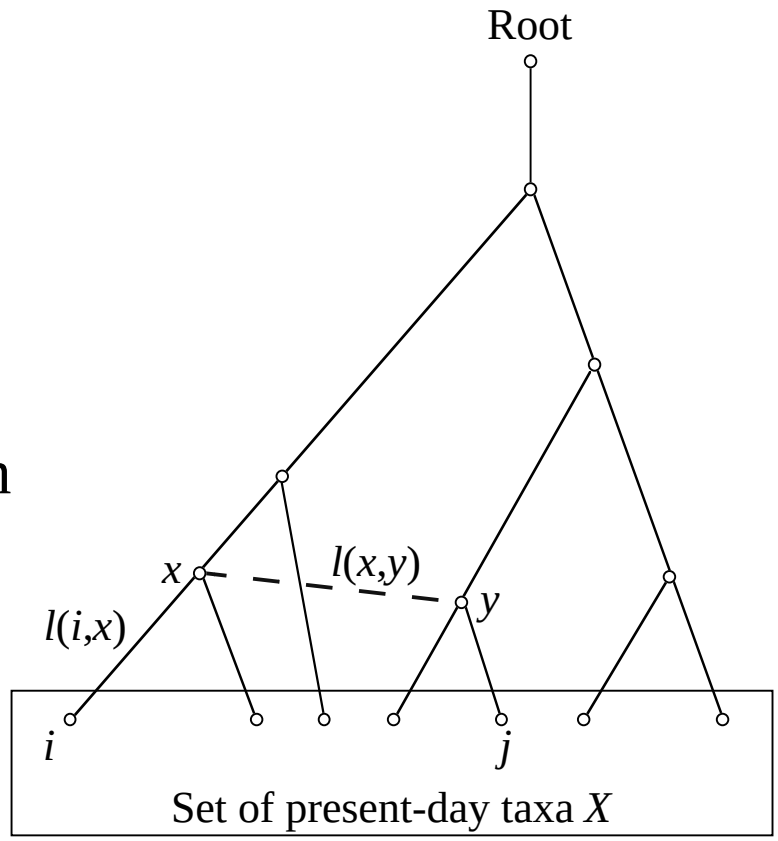
Reticulogram, or reticulated network

Diagram representing an evolutionary structure in which the species may be related in non-unique ways to a common ancestor.

A reticulogram R is a triplet (N, B, l) such that:

- N is a set of nodes (taxa, e.g. species);
- B is a set of branches;
- l is a function of branch lengths that assign real nonnegative numbers to the branches.

Each node is either a present-day taxon belonging to a set X or an intermediate node belonging to $N - X$.



Reticulogram distance matrix R = $\{r_{ij}\}$

The reticulogram distance r_{ij} is the minimum path-length distance between nodes i and j in the reticulogram:

$$r_{ij} = \min \{l_p(i,j) \mid p \text{ is a path from } i \text{ to } j \text{ in the reticulogram}\}$$

Problem

Construct a connected reticulated network, having a fixed number of branches, which *best* represents, according to *least squares* (LS), a dissimilarity matrix **D** among taxa. Minimize the LS function Q:

$$Q = \sum_{i \in X} \sum_{j \in X} (d_{ij} - r_{ij})^2 \quad \rightarrow \quad \min$$

with the following constraints:

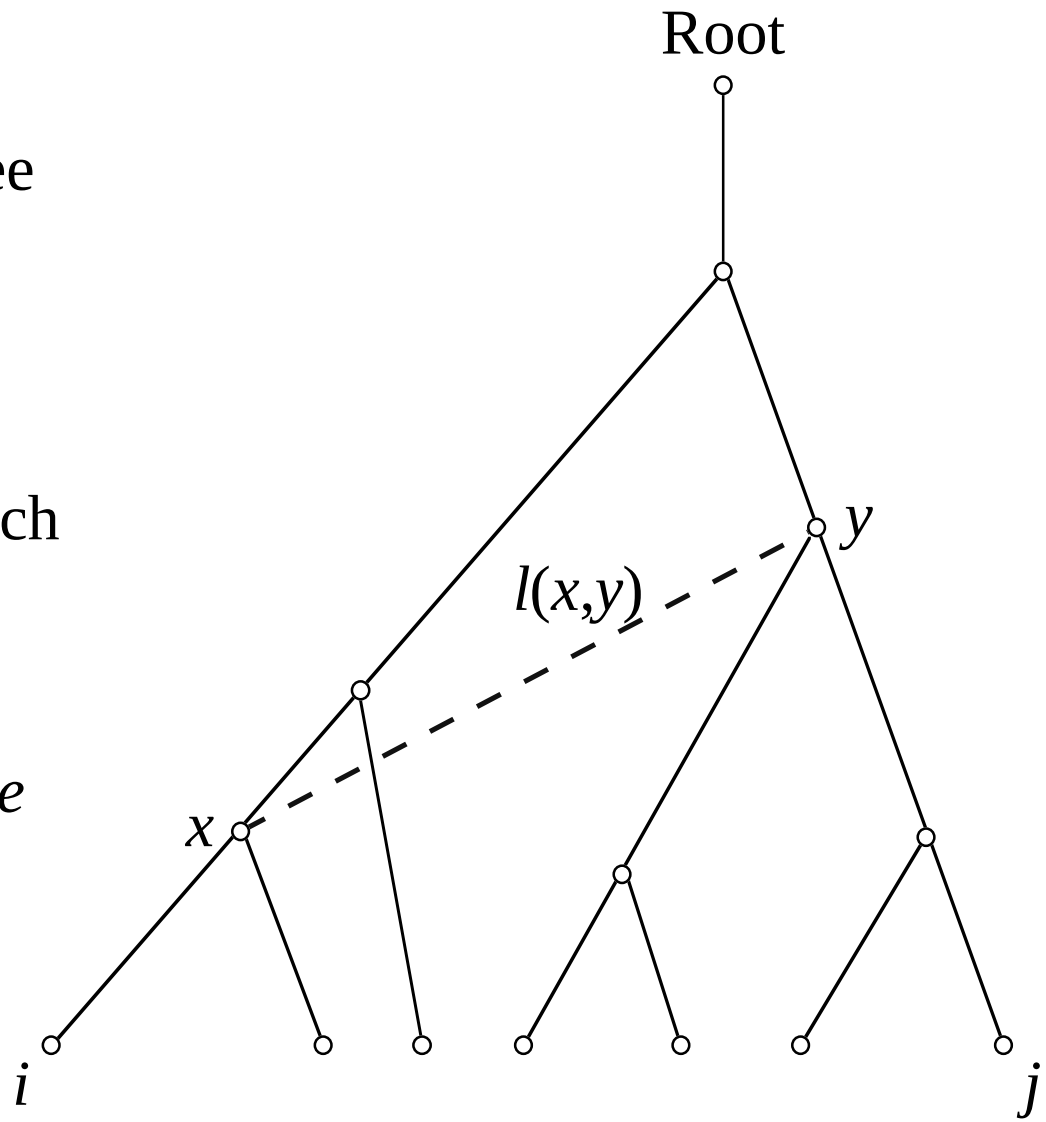
- $r_{ij} \geq 0$ for all pairs $i, j \in X$;
- **R** = $\{r_{ij}\}$ is associated with a reticulogram R having k branches.

Method

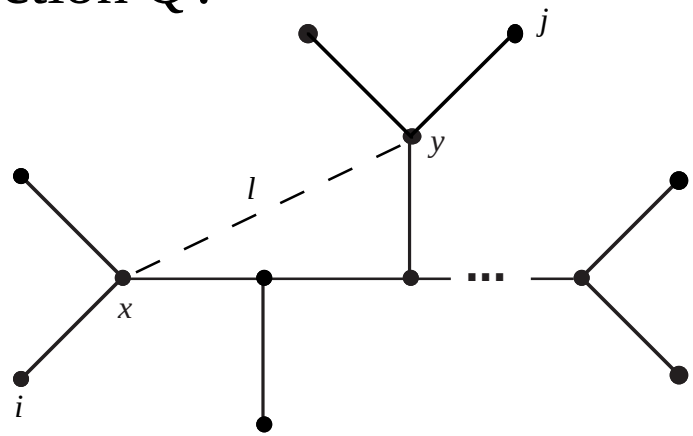
• Begin with a phylogenetic tree **T** inferred for the dissimilarity matrix **D** by some appropriate method.

• Add reticulation branches, such as the branch *xy*, to that tree.

Reticulation branches are annotations added onto the tree (B. Mirkin, 2004).



How to find a reticulated branch *xy* to add to **T**, such that its length *l* contributes the most to reducing the LS function *Q*?



Solution

1. Find a first branch *xy* to add to the tree

• Try all possible branches in turn:

Recompute distances among taxa $\in X$ in the presence of branch *xy*;

Compute $Q = \sum_{i \in X} \sum_{j \in X} (d_{ij} - r_{ij})^2$ incl. the candidate branch *xy*;

• Keep the new branch *xy*, of length *l*(*x*,*y*), for which *Q* is minimum.

2. Repeat for new branches.

STOP when the minimum of a stopping criterion is reached.

Reticulation branch lengths

The length of the reticulation branches is found by minimizing the quadratic sum of differences between the distance values (from matrix **D**) and the length of the reticulation branch estimates $l(x,y)$.

The solution to this problem is described in detail in Makarenkov and Legendre (2004: 199-200).

Stopping criteria

$$Q_1 = \frac{\sqrt{\sum_{i \in X} \sum_{j \in X} (d_{ij} - r_{ij})^2}}{\frac{n(n-1)}{2} - N} = \frac{\sqrt{Q}}{\frac{n(n-1)}{2} - N}$$

$$Q_2 = \frac{\sum_{i \in X} \sum_{j \in X} (d_{ij} - r_{ij})^2}{\frac{n(n-1)}{2} - N} = \frac{Q}{\frac{n(n-1)}{2} - N}$$

- $n(n-1)/2$ is the number of distances among n taxa
- N is the number of branches in the unrooted reticulogram

For initial unrooted binary tree: $N = 2n-3$

$$\mathrm{AIC} = \frac{\sum_{i \in X} \sum_{j \in X} (d_{ij} - r_{ij})^2}{\frac{(2n-2)(2n-3)}{2} - 2N} = \frac{Q}{\frac{(2n-2)(2n-3)}{2} - 2N}$$

$(2n-2)(2n-3)/2$ is the number of branches in a completely interconnected, unrooted graph containing n taxa and $(2n-2)$ nodes

$$\mathrm{MDL} = \frac{\sum_{i \in X} \sum_{j \in X} (d_{ij} - r_{ij})^2}{\frac{(2n-2)(2n-3)}{2} - N\log(N)} = \frac{Q}{\frac{(2n-2)(2n-3)}{2} - N\log(N)}$$

AIC: Akaike Information Criterion; MDL: Minimum Description Length.

Properties

1. The reticulation distance satisfies the triangular inequality, but not the four-point condition.

2. Our heuristic algorithm requires $O(kn^4)$ operations to add k reticulations to a classical phylogenetic tree with n leaves (taxa).

Simulations

to test the capacity of our algorithm to correctly detect reticulation events when present in the data.

Generation of distance matrix

Method inspired from the approach used by Pruzansky, Tversky and Carroll (1982) to compare additive (or phylogenetic) tree reconstruction methods.

- Generate additive tree with random topology and random branch lengths.

- Add a random number of reticulation branches, each one of randomly chosen length, and located at random positions in the tree.

- In some simulations, add random errors to the reticulated distances, to obtain matrix **D**.

Tree reconstruction algorithms to estimate the additive tree

1. ADDTREE by Sattath and Tversky (1977).

2. Neighbor joining (NJ) by Saitou and Nei (1987).

3. Weighted least-squares (MW) by Makarenkov and Leclerc (1999).

Criteria for estimating goodness-of-fit

1. Proportion of variance of **D** accounted for by **R**:

$$Var\% = 100 \times \left(1 - \frac{\sum_{i \in X} \sum_{j \in X} (d_{ij} - r_{ij})^2}{\sum_{i \in X} \sum_{j \in X} (d_{ij} - \bar{d})^2} \right)$$

2. Goodness of fit Q_1, which takes into account the least-squares loss (numerator) and the number of degrees of freedom (denominator):

$$Q_1 = \frac{\sqrt{\sum_{i \in X} \sum_{j \in X} (d_{ij} - r_{ij})^2}}{\frac{n(n-1)}{2} - N}$$

Simulation results (1)

1. Type 1 error

• Random trees without reticulation events and without random error: no reticulation branches were added to the trees.

• Random trees without reticulation events but with random error: the algorithm sometimes added reticulation branches to the trees. Their number increased with increasing *n* and with the amount of noise $\sigma^2 = \{0.1, 0.25, 0.5\}$. Reticulation branches represent incompatibilities due to the noise.

2. Reticulated distance **R**

The reticulogram always represented the variance of **D** better than the non-reticulated additive tree, and offered a better adjustment (criterion Q_1) for all tree reconstruction methods (ADDTREE, NJ, MW), matrix sizes (*n*), and amounts of noise $\sigma^2 = \{0.0, 0.1, 0.25, 0.5\}$.

Simulation results (2)

3. Tree reconstruction methods and reticulogram

The closer the additive tree was to **D**, the closer was also the reticulogram (criterion Q_1). It is important to use a good tree reconstruction method before adding reticulation branches to the additive tree.

4. Tree reconstruction methods

MW (*Method of Weights*, Makarenkov and Leclerc 1999) generally produced trees closer to **D** than the other two methods (criterion Q_1).

Application 1: Homoplasy in phylogenetic tree of primates[1]

Data: A portion of the protein-coding mitochondrial DNA (898 bases) of 12 primate species, from Hayasaka et al. (1988).

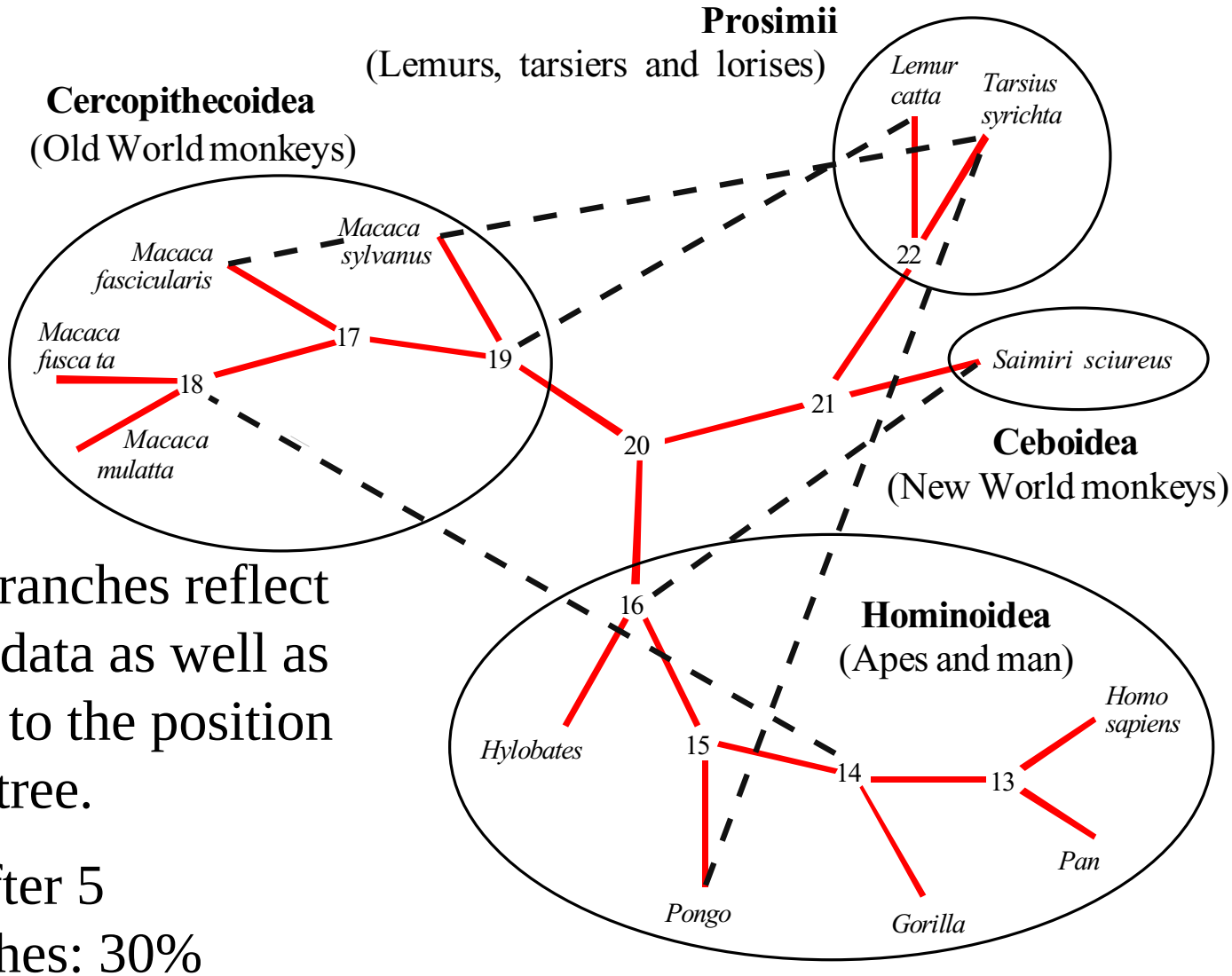
Distance matrix

	1	2	3	4	5	6	7	8	9	10	11
1. *Homo sapiens*	0.000										
2. *Pan*	0.089	0.000									
3. *Gorilla*	0.104	0.106	0.000								
4. *Pongo*	0.161	0.171	0.166	0.000							
5. *Hylobates*	0.182	0.189	0.189	0.188	0.000						
6. *Macaca fuscata*	0.232	0.243	0.237	0.244	0.247	0.000					
7. *Macaca mulatta*	0.233	0.251	0.235	0.247	0.239	0.036	0.000				
8. *Macaca fascicularis*	0.249	0.268	0.262	0.262	0.257	0.084	0.093	0.000			
9. *Macaca sylvanus*	0.256	0.249	0.244	0.241	0.242	0.124	0.120	0.123	0.000		
10. *Saimiri sciureus*	0.273	0.284	0.271	0.284	0.269	0.289	0.293	0.287	0.287	0.000	
11. *Tarsius syrichta*	0.322	0.321	0.314	0.303	0.309	0.314	0.316	0.311	0.319	0.320	0.000
12. *Lemur catta*	0.308	0.309	0.293	0.293	0.296	0.282	0.289	0.298	0.287	0.285	0.252

[1] Example developed in Makarenkov and Legendre (2000).

1. A phylogenetic tree was constructed from **D** using the neighbor-joining method (NJ). It separated the primates into 4 groups.

2. Five reticulation branches were added to the tree (stopping criterion Q_1).



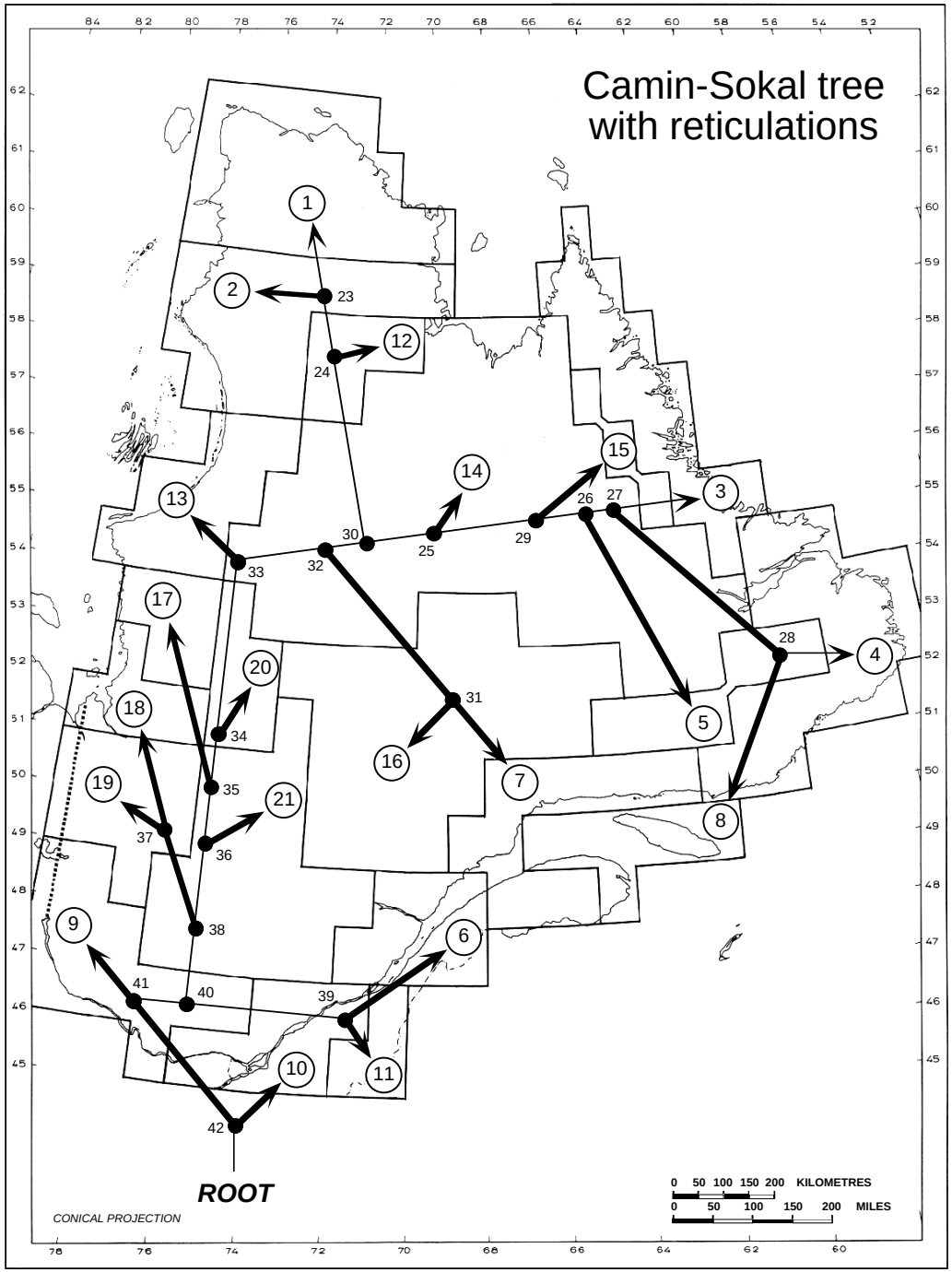
The reticulation branches reflect homoplasy in the data as well as the uncertainty as to the position of Tarsiers in the tree.

Reduction of Q after 5 reticulation branches: 30%

Application 2: Postglacial dispersal of freshwater fishes[1]

Question: Can we reconstruct the routes taken by freshwater fishes to reinvade the Québec peninsula after the last glaciation?

The Laurentian glacier melted away between –14000 and –5000 years.



[1] Example developed in Legendre and Makarenkov (2002).

Step 1

Presence-absence of 109 freshwater fish species in 289 geographic units (1 degree x 1 degree). A Sørensen similarity matrix was computed among units, based on fish presence-absence data. The 289 units were grouped into 21 regions by clustering under constraint of spatial contiguity (Legendre and Legendre 1984)[1].

Step 2

Using only the 85 species restricted to freshwater (stenohaline species), a phylogenetic tree was computed (Camin-Sokal parsimony), depicting the loss of species from the glacial refugia on their way to the 21 regions (Legendre 1986)[2].

[1] Legendre, P. and V. Legendre. 1984. Postglacial dispersal of freshwater fishes in the Québec peninsula. *Canadian Journal of Fisheries and Aquatic Sciences* 41: 1781-1802.

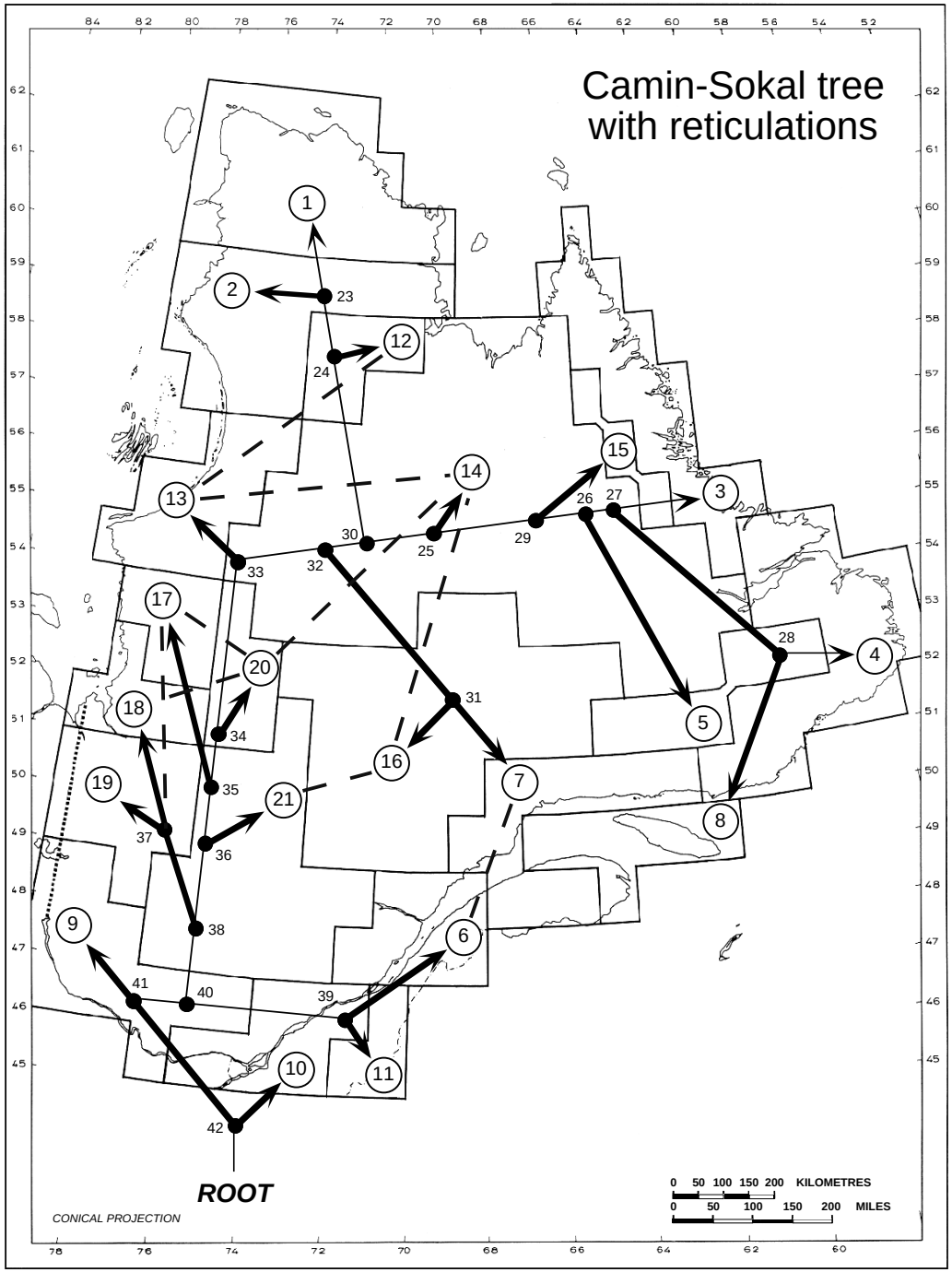
[2] Legendre, P. 1986. Reconstructing biogeographic history using phylogenetic-tree analysis of community structure. *Systematic Zoology* 35: 68-80.

Step 3

• A new **D** matrix (1 – Jaccard similarity coefficient) was computed for the 85 stenohaline species.

• Reticulation edges were added to the Camin-Sokal tree using a weighted least-squares version of the algorithm. Weights were 1 for adjacent, or 0 for non-adjacent regions.

• Stopping criterion Q_1: 9 reticulation branches were added to the Camin-Sokal tree.



Biogeographic interpretation of the reticulations

The reticulation branches added to the tree represent faunal exchanges by fish migration between geographically adjacent regions using interconnexions of the river network, in addition to the main exchanges described by the additive tree.

Application 3: Evolution of photosynthetic organisms[1]

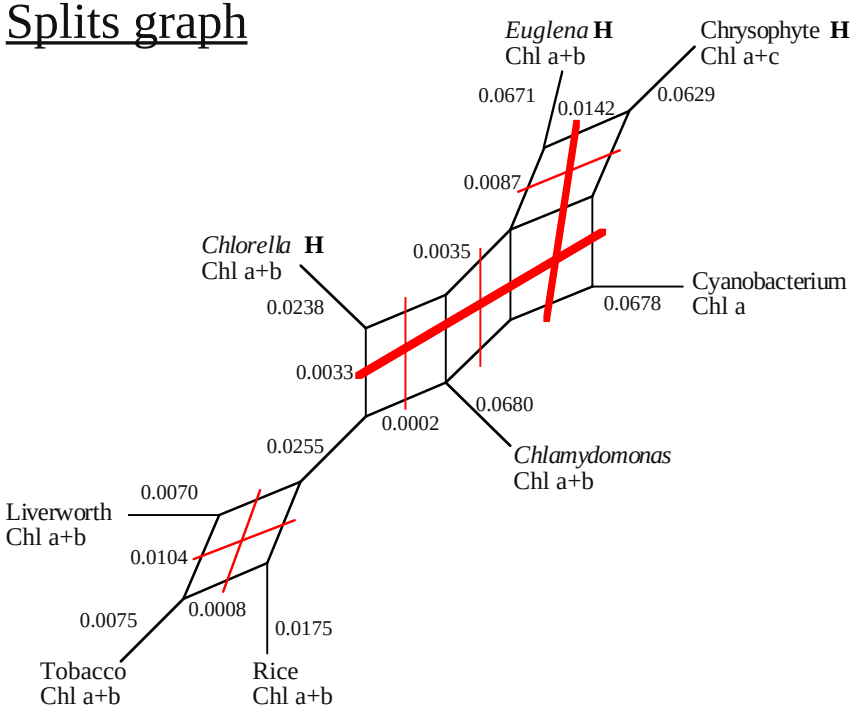
Compare reticulogram to splits graph.

Data: LogDet distances among 8 species of photosynthetic organisms, computed from 920 bases from the 16S rRNA of the chloroplasts (sequence data from Lockhart et al. 1993).

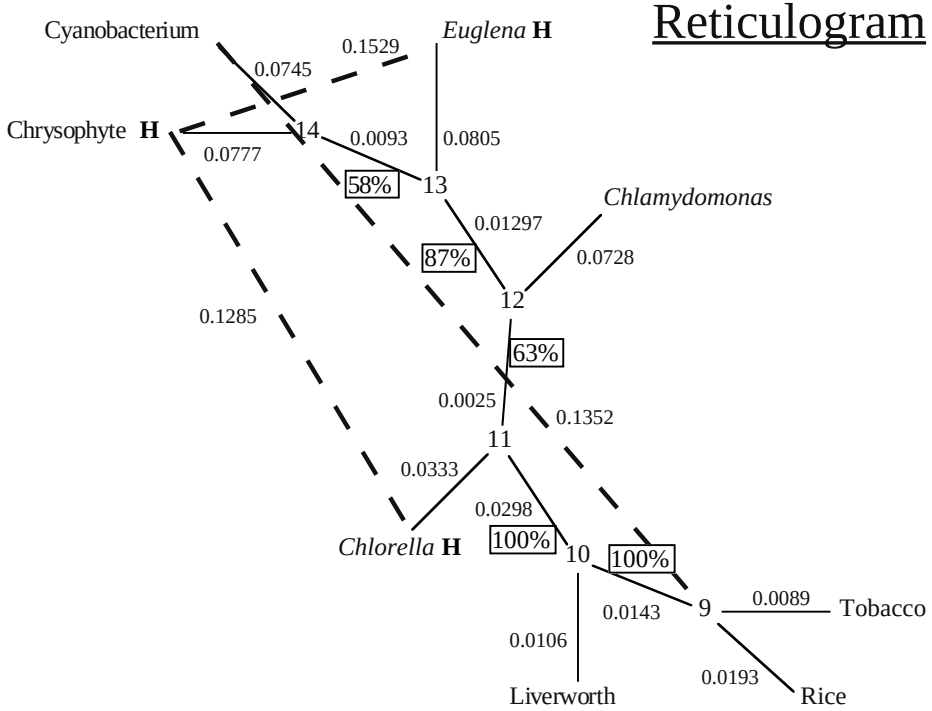
	1	2	3	4	5	6	7	8
1. Tobacco	0.0000							
2. Rice	0.0258	0.0000						
3. Liverworth	0.0248	0.0357	0.0000					
4. *Chlamydomonas*	0.1124	0.1215	0.1014	0.0000				
5. *Chlorella*	0.0713	0.0804	0.0604	0.0920	0.0000			
6. *Euglena*	0.1270	0.1361	0.1161	0.1506	0.1033	0.0000		
7. Cyanobacterium	0.1299	0.1390	0.1190	0.1535	0.1128	0.1611	0.0000	
8. Chrysophyte	0.1370	0.1461	0.1261	0.1606	0.1133	0.1442	0.1427	0.0000

[1] Example developed in Makarenkov and Legendre (2004).

Splits graph



Reticulogram



Interpretation of the splits

- Separation of organisms with or without chlorophyll *b*.
- Separation of facultative heterotrophs (**H**) from the other organisms.

Interpretation of the reticulation branches

- Group of facultative heterotrophs.
- Endosymbiosis hypothesis: chloroplasts could be derived from primitive cyanobacteria living as symbionts in eukaryotic cells.

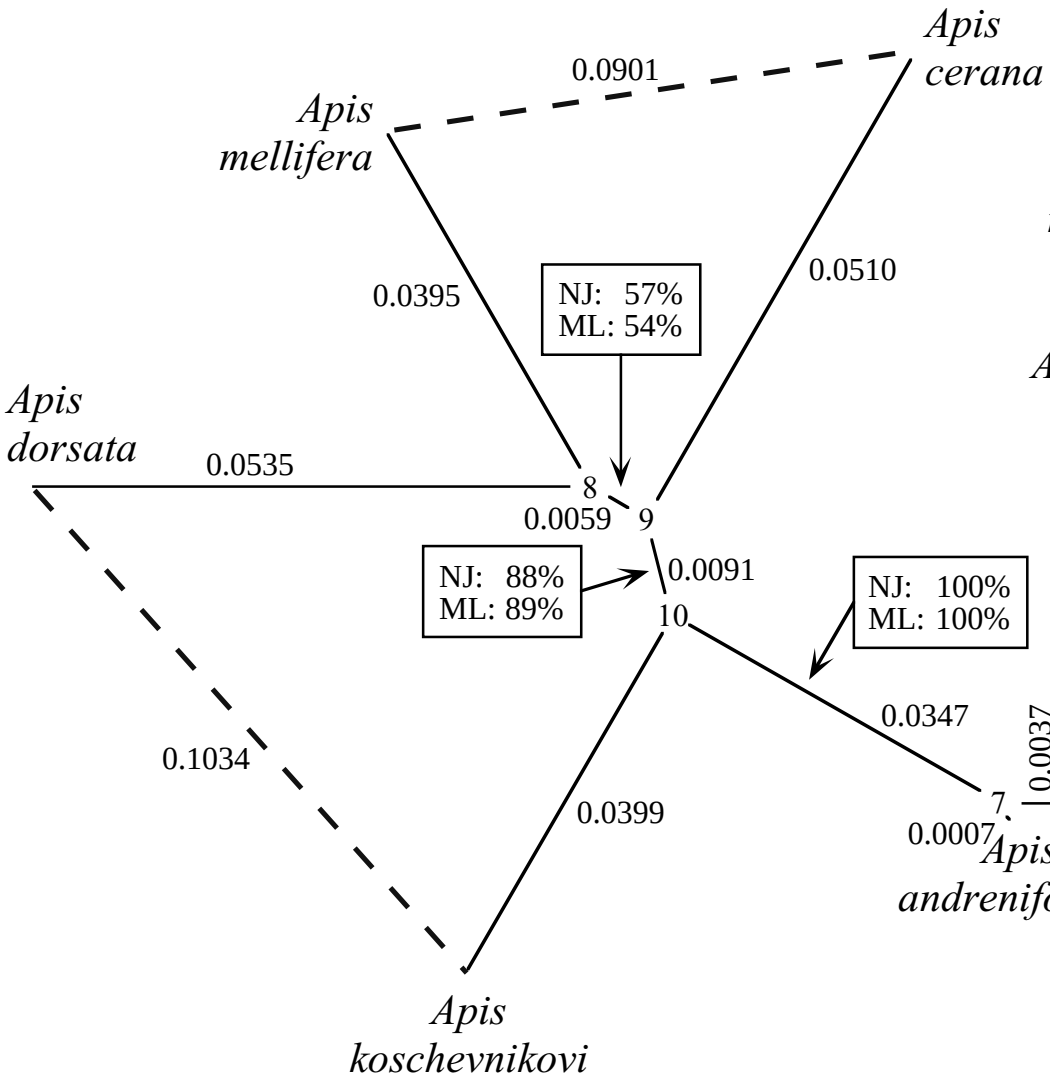
Application 4: Phylogeny of honeybees[1]

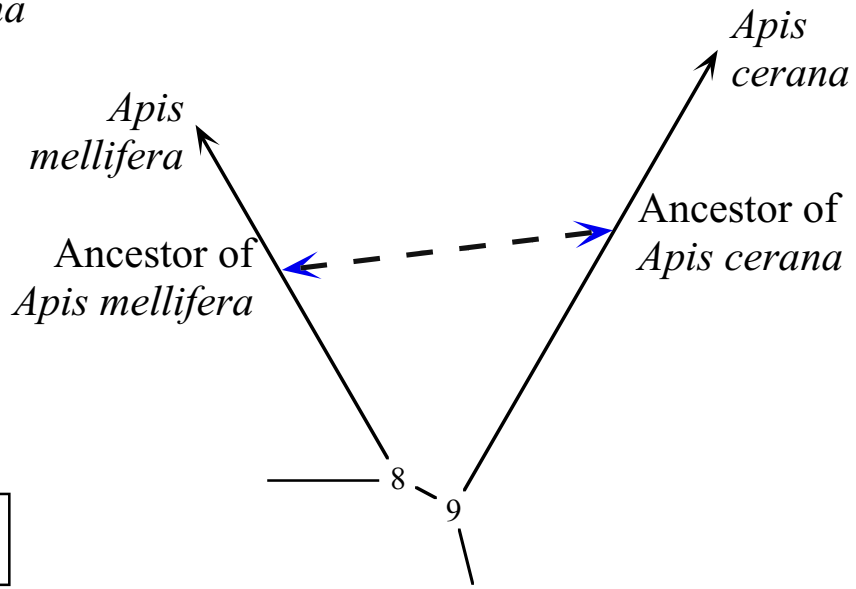
Data: Hamming distances among 6 species of honeybees, computed from DNA sequences (677 bases) data. **D** from Huson (1998).

	1	2	3	4	5	6
1. *Apis andreniformis*	0.000					
2. *Apis mellifera*	0.090	0.000				
3. *Apis dorsata*	0.103	0.093	0.000			
4. *Apis cerana*	0.096	0.090	0.117	0.000		
5. *Apis florea*	0.004	0.093	0.106	0.099	0.000	
6. *Apis koschevnikovi*	0.075	0.100	0.103	0.099	0.078	0.000

Phylogenetic tree reconstruction method: Neighbor joining (NJ).

[1] Example developed in Makarenkov, Legendre and Desdevises (2004).





	Least-squares loss Q	Criterion Q_2
Phylogenetic tree	0.000143	0.000024
+ 1 reticulation	0.000104	0.000021
+ 2 reticulations	0.000078	0.000020 (min)

Application 5: Microgeographic differentiation in muskrats[1]

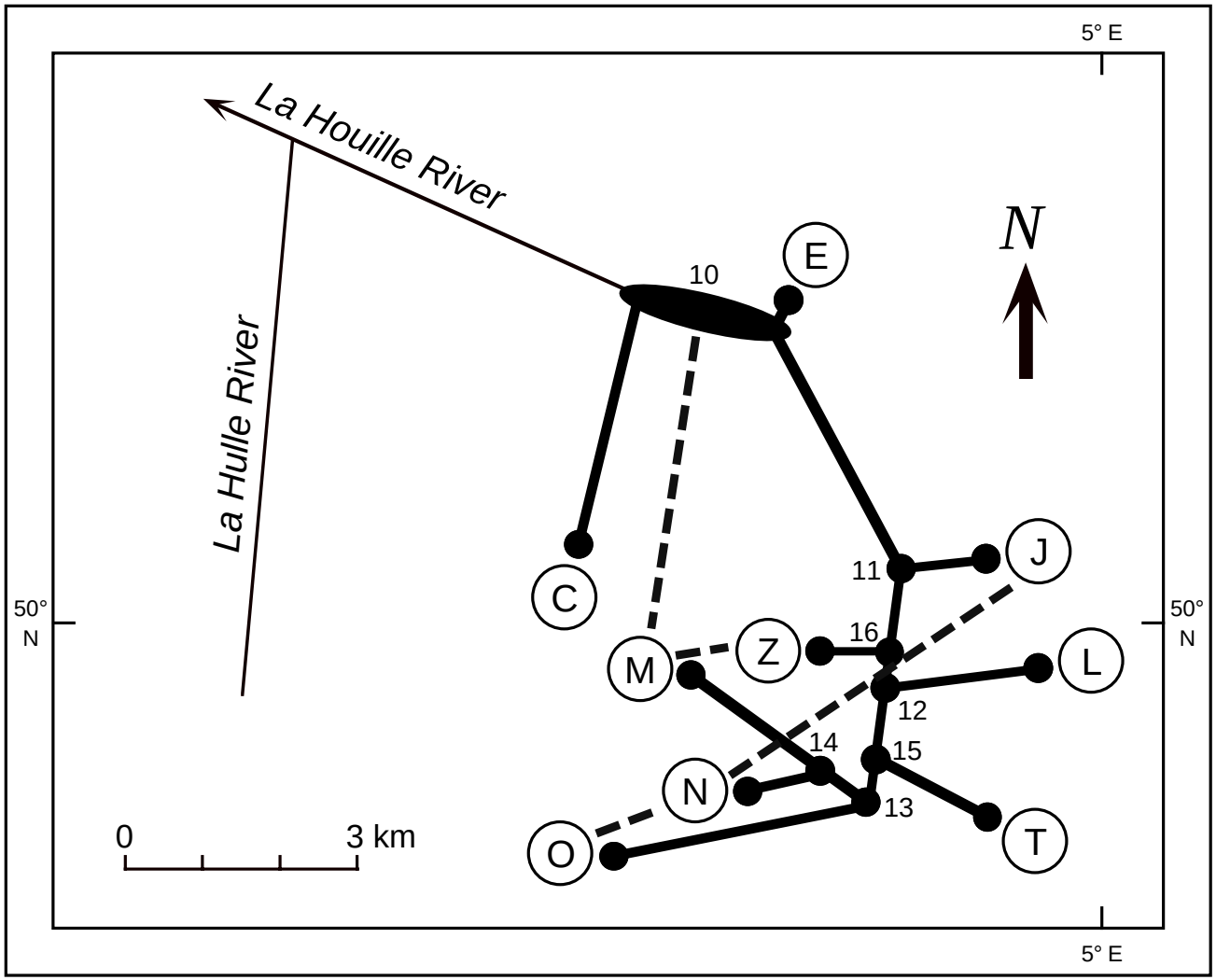
The morphological differentiation among local populations of muskrats in La Houille River (Belgium) was explained by "isolation by distance along corridors" (Le Boulengé, Legendre et al. 1996).

Data: Mahalanobis distances among 9 local populations, based on 10 age-adjusted linear measurements of the skulls. Total: 144 individuals.

Populations	C	E	J	L	M	N	O	T	Z
C	0.0000								
E	2.1380	0.0000							
J	2.2713	2.9579	0.0000						
L	1.7135	2.3927	1.7772	0.0000					
M	1.5460	1.9818	2.4575	1.0125	0.0000				
N	2.6979	3.3566	1.9900	1.8520	2.6954	0.0000			
O	2.9985	3.6848	3.4484	2.4272	2.6816	2.3108	0.0000		
T	2.3859	2.3169	2.4666	1.4545	1.7581	2.2105	2.5041	0.0000	
Z	2.3107	2.3648	1.8086	1.6609	2.0516	2.2954	3.4301	2.0413	0.0000

[1] Example developed in Legendre and Makarenkov (2002).

Tree: The river network of La Houille.



4 reticulation branches were added to the tree (minimum of Q_2).
Interpretation of O-N, M-Z, M-10: migrations across wetlands.
N-J = type I error (false positive)?

Application 6: Detection of *Aphelandra* hybrids[1]

L. A. McDade (1992)[2] artificially created hybrids between species of Central American *Aphelandra* (*Acanthus* family).

Data: 50 morphological characters, coded in 2-6 states, measured over 12 species as well as 17 hybrids of known parental origins.

Distance matrix: $D_{ij} = (1 - S_{ij})^{0.5}$ where S_{ij} is the simple matching similarity coefficient between species *i* and *j*.

[1] Example developed in Legendre and Makarenkov (2002).

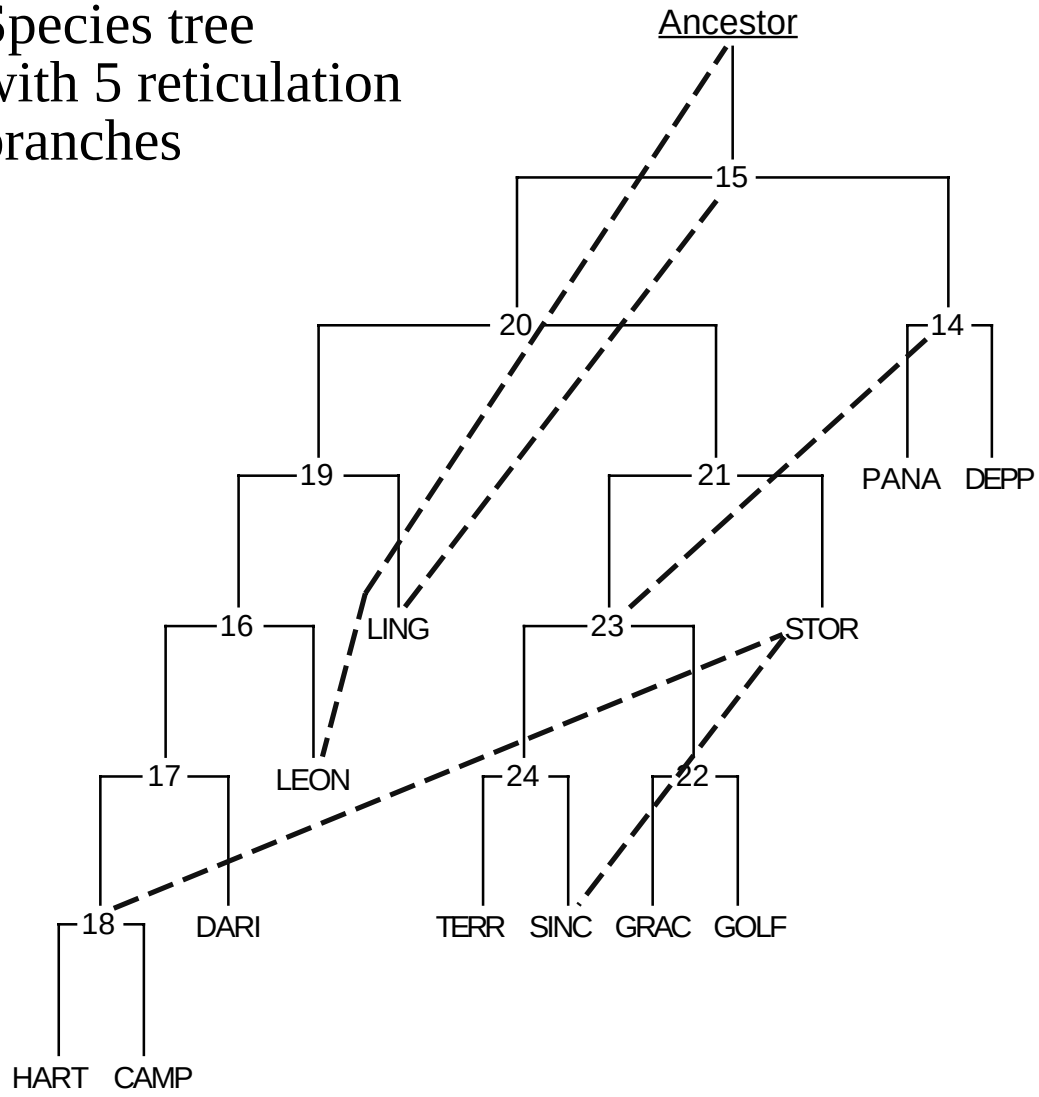
[2] McDade, L. A. 1992. Hybrids and phylogenetic systematics II. The impact of hybrids on cladistic analysis. *Evolution* 46: 1329-1346.

Step 1

Calculation of a neighbor-joining phylogenetic tree and a reticulogram among the 12 *Aphelandra* species.

The minimum of Q_1 was reached after addition of 5 reticulated branches.

Species tree with 5 reticulation branches



Step 2: Addition of one of McDade's hybrids to the distance matrix and recalculation of the reticulated tree.

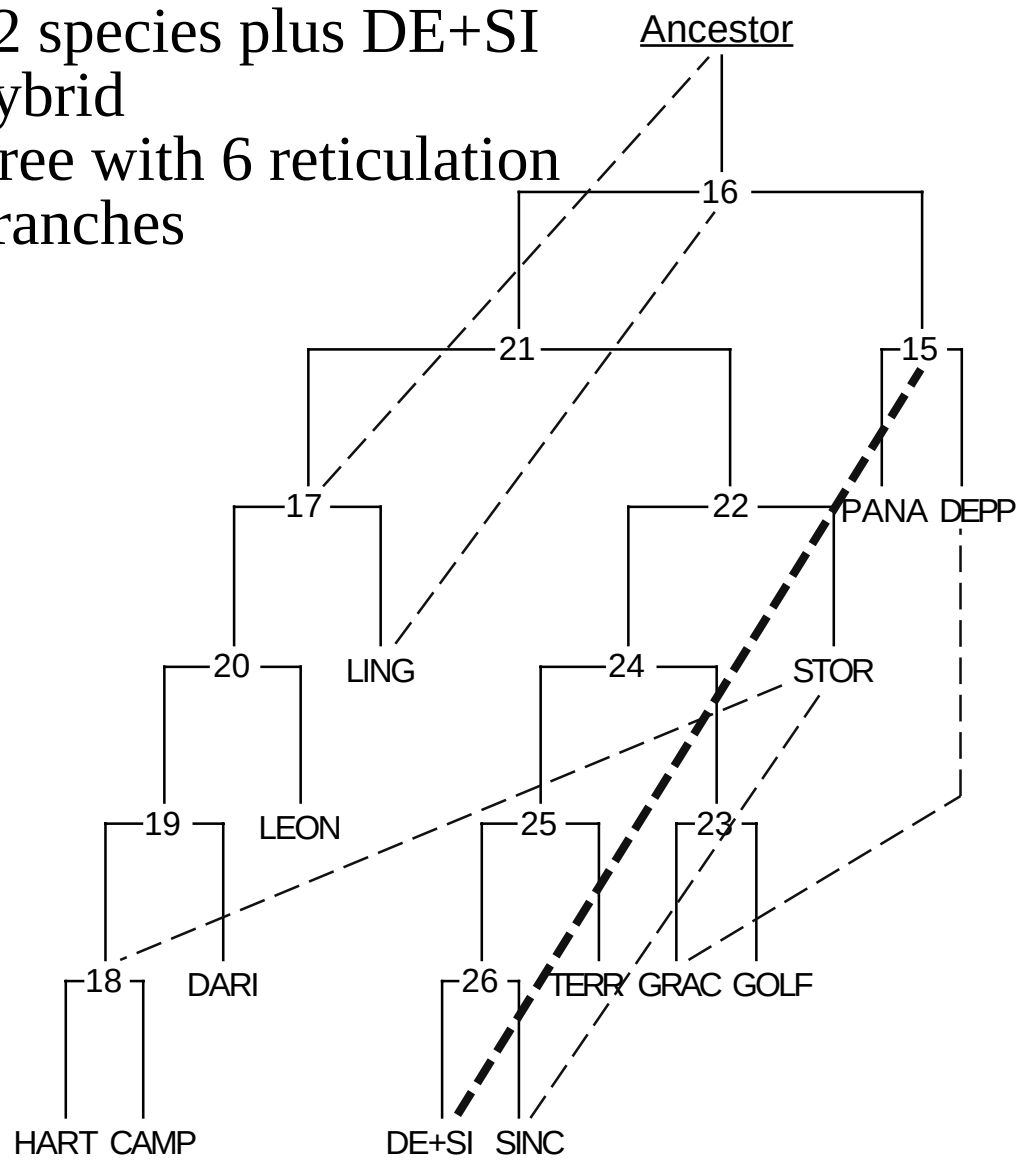
Hybrid: DExSI
Ovulate parent: DEPP
Staminate parent: SINC

6 reticulation branches were added to the tree.

• DExSI is the sister taxon of SINC in the tree.

• DExSI is connected by a new edge (bold) to node 15, the ancestor of DEPP.

12 species plus DE+SI hybrid
Tree with 6 reticulation branches



References

Available in PDF at http://www.fas.umontreal.ca/biol/legendre/reprints/
and http://www.info.uqam.ca/~makarenv/trex.html

Legendre, P. (Guest Editor) 2000. Special section on reticulate evolution. *Journal of Classification* 17: 153-195.

Legendre, P. and V. Makarenkov. 2002. Reconstruction of biogeographic and evolutionary networks using reticulograms. *Systematic Biology* 51: 199-216.

Makarenkov, V. and P. Legendre. 2000. Improving the additive tree representation of a dissimilarity matrix using reticulations. In: *Data Analysis, Classification, and Related Methods.* Proceedings of the IFCS-2000 Conference, Namur, Belgium, 11-14 July 2000.

Makarenkov, V. and P. Legendre. 2004. From a phylogenetic tree to a reticulated network. *Journal of Computational Biology* 11: 195-212.

Makarenkov, V., P. Legendre and Y. Desdevises. 2004. Modelling phylogenetic relationships using reticulated networks. *Zoologica Scripta* 33: 89-96.

T-Rex — Tree and Reticulogram Reconstruction[1]

Downloadable from http://www.info.uqam.ca/~makarenv/trex.html

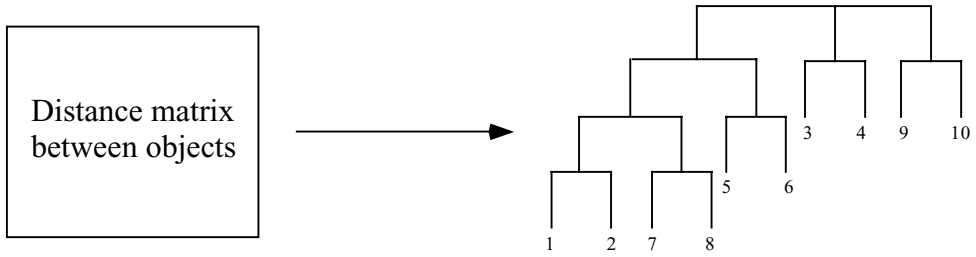
Authors: Vladimir Makarenkov

Versions: Windows 9x/NT/2000/XP and Macintosh

With contributions from A. Boc, P. Casgrain, A. B. Diallo, O. Gascuel, A. Guénoche, P.-A. Landry, F.-J. Lapointe, B. Leclerc, and P. Legendre.

Methods implemented

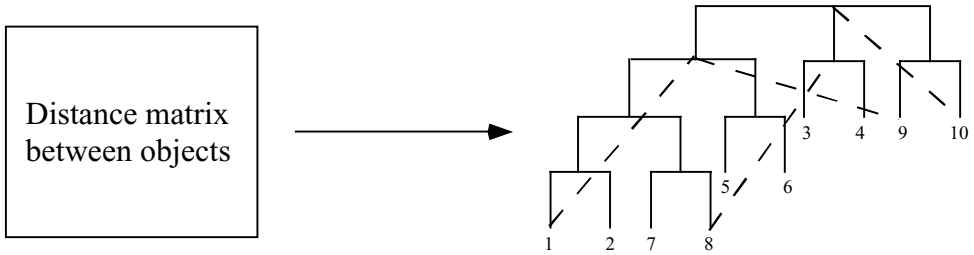
- 6 fast distance-based methods for additive tree reconstruction.



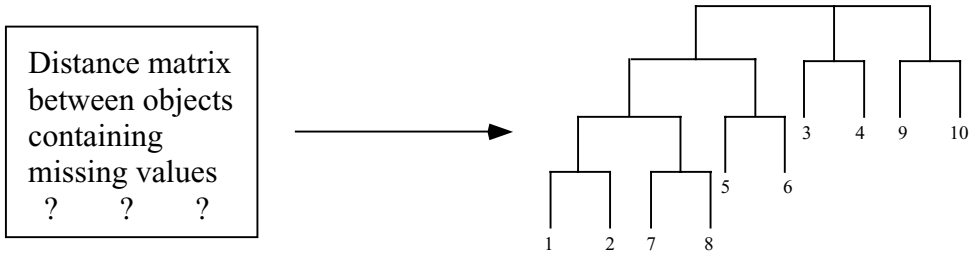
---

[1] Makarenkov, V. 2001. T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics* 17: 664-668.

• Reticulogram construction, weighted or not.



• 4 methods of tree reconstruction for incomplete data.



• Reticulogram with detection of reticulate evolution processes, hybridization, or recombination events.

• Reticulogram with detection of horizontal gene transfer among species.

• Graphical representations: hierarchical, axial, or radial. Interactive manipulation of trees and reticulograms.