Applied Mycology and Biotechnology
An International Series
Volume 6. Bioinformatics

Phylogenetic Network Construction Approaches

Vladimir Makarenkov, Dmytro Kevorkov and Pierre Legendre
Département d'informatique, Université du Québec à Montréal, C.P. 8888, succ. Centre-Ville, Montréal (Québec) Canada H3C 3P8 (makarenkov.vladimir at uqam.ca); Département d'informatique, Université du Québec à Montréal, C.P. 8888, succ. Centre-Ville, Montréal (Québec) Canada H3C 3P8. (kevorkov at lacim.uqam.ca); Département de Sciences Biologiques, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal, (Québec) Canada H3C 3J7 (pierre.legendre at umontreal.ca).

This chapter presents a review of the mathematical techniques available to construct phylogenies and to represent reticulate evolution. Phylogenies can be estimated using distance-based, maximum parsimony, or maximum likelihood methods. Bayesian methods have recently become available to construct phylogenies. Reticulate evolution includes horizontal gene transfer between taxa, hybridization events, and homoplasy. Genetic recombination also creates reticulate evolution within lineages. Several methods are now available to construct reticulated networks of various kinds. Twelve such methods and the accompanying software are described in this review chapter.

1. INTRODUCTION

Evolution of species has long been assumed to be a branching process that could only be represented by a tree topology. In a tree topology, a species is linked to its closest ancestor; other interspecies relationships cannot be taken into account. Well-known evolutionary mechanisms such as hybridization or horizontal gene transfer can only be represented appropriately using a network model.

Patterns of reticulate evolution have been found in a variety of evolutionary contexts, giving rise to a number of recent studies. In bacterial evolution, lateral gene transfer (i.e. horizontal gene transfer) is the mechanism allowing bacteria to exchange genes across species (Sonea and Panisset 1976 1981; Doolittle 1999; Sonea and Mathieu 2000; Sneath

Corresponding author: Vladimir Makarenkov

2000). In plant evolution, allopolyploidy leads to the appearance of new species encompassing the chromosome complements of the two parent species. Reticulate patterns are also present in micro-evolution within species in sexually-reproducing eukaryotes (Smouse 2000). Examples of molecular data sets containing regions with reticulate histories can be found in Fitch et al. (1990). (multigene families), Robertson, Hahn and Sharp (1995). (virus strains), and Guttman and Dykhuizen (1994). (bacterial genes). For example, the phylogeny of 24 inbred strains of mice obtained by Atchley and Fitch (1991, 1993). included several strains with hybrid origins. Hatta et al. (1999). conducted a molecular phylogenetic analysis providing strong evidence that reef-building corals have evolved in repeated rounds of species separation and fusion, a process leading to a reticulate evolutionary history. Odorico and Miller (1997). discovered patterns of variation due to reticulate evolution in the ribosomal internal transcribed spacers and 5.8s rDNA among five species of *Acropora* corals. The reticulate origin of some root knot nematodes of the genus *Meloidogyne*, which are widespread agricultural pests, was discussed by Hugall, Stanton and Moritz (1999). Cheung et al. (1999). established clear evidence that the evolution of class-I alcohol dehydrogenase genes in catarrhine primates has been reticulate. Phylogenetic analyses of two archaeal genes in *Thermotoga maritima* revealed multiple transfers between archaea and bacteria (Nesbø et al. 2001). The latter analyses confirmed the hypothesis that lateral gene transfer (LGT) events have occurred between bacteria and archaea.

According to McDade (1995). analytical tools enabling one to generate reticulate topologies that accurately depict hybrid history represent a wide-open field for research. When traditional cladistic/phylogenetic methods are applied in such cases, they may produce confusing results since they are constrained to generate only tree-like patterns. Homoplasy is another source of confusion in the reconstruction of phylogenetic trees; it can be represented by supplementary branches added to phylogenetic trees (Makarenkov and Legendre 2000). In their review on reticulate evolution, Posada and Crandall (2001). considered several definitions of net-like evolution, accompanied by proposals of how the involved biological procedures should be represented mathematically. Nakhleh et al. (2003). reported a suite of useful techniques for studying the topological accuracy of methods for reconstructing phylogenetic networks. Linder et al. (2003, 2004). have recently provided an overview of the methods and software meant to depict reticulation events in different evolutionary contexts.

The present article is organized as follows: section 2 recalls the main approaches used to infer phylogenetic trees from sequence and distance data; section 3 describes different evolutionary contexts where patterns of reticulate evolution can occur; section 4 presents a number of algorithms and software for reconstructing evolutionary networks; we conclude with an extensive list of references.

2. PHYLOGENETIC TREE RECONSTRUCTION METHODS

A classical way to illustrate phylogenetic relationships among species is to model them using a phylogenetic tree (i.e. a phylogeny or an additive tree). In this section we

discuss the main approaches for inferring phylogenetic trees. For a comprehensive discussion of the methods for inferring phylogenies readers are referred to Swofford et al. (1996). Li (1997). and Felsenstein (2003).

There exist two main approaches for inferring phylogenies. The first one, called the *phenetic approach*, makes no reference to any historical relationship. It operates by measuring distances between species and reconstructs the tree using a hierarchical clustering procedure. The second one, called the *cladistic approach*, considers possible pathways of evolution, inferring the features of the ancestor at each node and choosing an optimal tree according to some model of evolutionary change. The phenetic approach is based on similarity whereas the cladistic approach is based on genealogy. Four basic types of methods for building phylogenies will be presented in detail in this section: distance-based methods (which belong to the phenetic approach), maximum parsimony, maximum likelihood, and Bayesian methods (which belong to the cladistic approach). The two most comprehensive software packages, widely used by the community of computational biologists, are PHYLIP (PHYLogeny Inference Package), a set of freeware programs developed by Felsenstein (2004). and PAUP (Phylogenetic Analysis Using Parsimony) developed by Swofford (1998). Both PAUP and PHYLIP contain the most popular distance-based, maximum likelihood and maximum parsimony methods. They also provide visualization tools as well as bootstrap and jackknife tree validation support. In addition, the user manuals available for both packages are recognized as essential guides, serving as a comprehensive introduction to phylogenetic analysis for beginners as well as important sources of references for experts in the field.

2.1. Distance-based Methods

Distance-based methods estimate pairwise distances prior to computing a branch-weighted phylogenetic tree. If the pairwise distances are sufficiently close to the number of evolutionary events between pairs of taxa, these methods reconstruct a correct tree (Kim and Warnow 1999). This assumption is true for many models of biomolecular sequence evolution, in which case distance-based methods give sufficiently accurate results (Li 1997). The main advantage of distance-based methods is their small time complexity that makes them applicable to the analysis of large data sets.

If the rate of evolution is constant over the entire tree and the "molecular clock" hypothesis holds, corrections to the pairwise distances required during inference of the phylogenetic tree may be small. However, the "molecular clock" assumption is usually inappropriate for distantly related sequences and the reconstruction of a correct phylogenetic tree becomes problematic under this hypothesis. If the molecular clock assumption does not hold, the observed differences among sequences do not accurately reflect the evolutionary distances. In that case, multiple substitutions at the same site obscure the true distances and make sequences seem artificially closer to each other then they really are. Correction of the pairwise distances that accounts for multiple substitutions at the same site should be used in such cases. There are many Markov

models for modeling sequence evolution; each of them implies a specific way to estimate and correct pairwise distances. Furthermore, these corrections have substantial variance when the distances are large. Among the most popular sequence-distance transformation models we have the Hamming, Jukes Cantor (Jukes and Cantor 1969). Kimura 2-parameter (Kimura 1981). and LogDet (Steel 1994). distances. When the goal is to infer relationships with high divergence between sequences, it can be difficult to obtain reliable values for the distance matrix; as consequence, the distance-based algorithms have little chance of succeeding. More detailed description of some distance-based methods is presented below:

UPGMA: The UPGMA [Unweighted Pair-Group Method using Arithmetic averages (Rohlf 1963).] method was originally proposed for taxonomic purposes. It could be used for phylogeny inferring as well, but one has to assume that the rate of nucleotide or amino acid substitution is the same for all evolutionary lineages. UPGMA always produces an ultrametric tree (i.e. a dendrogram). In practice, this method recovers the correct tree with reasonably high probability when the "molecular clock" hypothesis applies and the evolutionary distance is large for all pairs of sequences. This method can be useful to biologists interested in constructing species trees.

At present, however, many investigators use relatively short DNA sequences for which the "molecular clock" hypothesis is often not valid. Therefore, one should be cautious about UPGMA trees. This method produces a rooted tree because of the assumption of a constant rate of evolution, though it is possible to remove the root if necessary. We illustrate the application of the UPGMA procedure using a set of four species characterized by the sequences TAGG, TACG, AAGC, and AGCC. Using the number of differences as an estimate of the dissimilarity among species, we obtain the distance matrix shown in Table 1.

Table 1. Distance matrix for the four sequences TAGG, TACG, AAGC, and AGCC

	TAGG	TACG	AAGC	AGCC
TAGG	0	1	2	4
TACG		0	3	3
AAGC			0	2
AGCC				0

The smallest distance in Table 1 is 1 (between the sequences TAGG and TACG). Consequently, the first cluster to be formed is {TAGG, TACG} and the phylogeny will contain the tree fragment shown in Fig. 1.

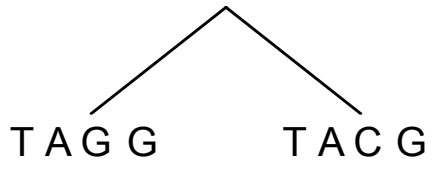

Fig. 1. The first cluster {TAGG, TACG} created by the UPGMA algorithm.

The combined node {TAGG, TACG}, formed by the nodes TAGG and TACG, replaces them in the initial distance matrix to obtain the reduced distance matrix shown in Table 2.

Table 2. Reduced distance matrix

	{TAGG, TACG}	AAGC	AGCC
{TAGG, TACG}	0	½ (2+3) = 2.5	½ (4+3) = 3.5
AAGC		0	2
AGCC			0

The next cluster with the closest nodes (distance = 2) is {AAGC, AGCC}. These two sequences have two differences in the homologous sites. The final cluster fusion links clusters {TAGG, TACG} and {AAGC, AGCC} (Fig. 2).

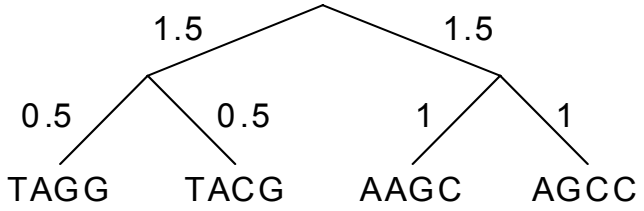

Fig. 2. Phylogenetic tree obtained by UPGMA for the set of sequences in Table 1.

Neighbor-joining (NJ): Neighbor-joining (Saitou and Nei 1987; Studier and Keppler 1988). is arguably the most popular among the distance-based methods. For some time, the success of NJ was inexplicable for computational biologists, due to the lack of approximation bounds. One of the first bounds was found by Atteson (1999). who showed that this method would be able to return the true phylogeny given that the observed distance is sufficiently close to the true evolutionary distance. Compared to UPGMA, NJ is designed to correct the unequal rates of evolution in different branches of the tree. NJ has a low $O(n^3)$ time complexity, where n is the number of species, and like other distance methods performs well when the divergence between sequences is low. In its first step, NJ considers a bush tree with n leaves and n branches. The tree is gradually transformed into a binary phylogenetic tree with the same n leaves and $2n$-3 branches by merging at each iteration a pair of branches corresponding to the shortest possible tree. Computationally, the tree generation by NJ is similar to UPGMA. When two nodes are linked, their common ancestral node is added to the reduced matrix and the terminal nodes with their respective branches are removed from it. Contrary to UPGMA, neighbor-joining does not produce a dendrogram (ultrametric distance) but an additive tree (additive distance).

Bio Neighbor-joining (BioNJ): The BioNJ (Gascuel 1997a). method is an improved version of the neighbor-joining method of Saitou and Nei (1987). The branch length estimation and distance matrix reduction formulae in NJ provide low variance estimators (Gascuel 1997a). In the paper describing BioNJ, Gascuel (1997a). showed how to improve the accuracy of NJ by incorporating minimum variance optimization in the

NJ reduction formula. BioNJ follows an agglomerative scheme similar to that of NJ. It works iteratively, picking a pair of taxa, creating a new node which represents the cluster of these taxa, and reducing the distance matrix by replacing the two taxa by this node. BioNJ uses a simple, first-order model of the variances and covariances of evolutionary distance estimates. This model is well adapted when the estimates are obtained from aligned sequences. At each step it permits the selection, from the class of admissible reductions, of the reduction that minimizes the variance of the new distance matrix. In this way, BioNJ obtains better estimates to choose the pair of taxa to be agglomerated during the next steps. Like NJ, the BioNJ method has a time complexity of $O(n^3)$ for n species. This makes it applicable to the analysis of large data sets. The performances of the two methods are similar when the substitution rates are low, or when they are the same in various lineages. When the substitution rates are high and varying among lineages, BioNJ outperforms NJ in terms of topological accuracy (Gascuel 1997a).

Among other popular distance-based methods, let us mention ADDTREE by Sattath and Tversky (1977). Unweighted Neighbor-Joining (UNJ) by Gascuel (1997b). the Method of Weighted least-squares (MW) by Makarenkov and Leclerc (1999). and FITCH by Felsenstein (1997).

Recommended software: PHYLIP (Felsenstein), PAUP (Swofford), MEGA (Kumar, Tamura and Nei), DAMBE (Xia), T-REX (Makarenkov), and BIONJ (Gascuel).

2.2. Maximum Parsimony

In contrast to the distance-based methods, parsimony infers phylogenetic trees by evaluating the possible mutations between sequences. In general terms, the aim of parsimony methods is to find the phylogenetic tree with minimum total length. That is the tree with the smallest number of evolutionary changes explaining the observed data. For instance, the phylogenetic tree with minimum total length for the sequences CAAG, CCAG, GCAT, and GCTT is presented in Fig. 3.

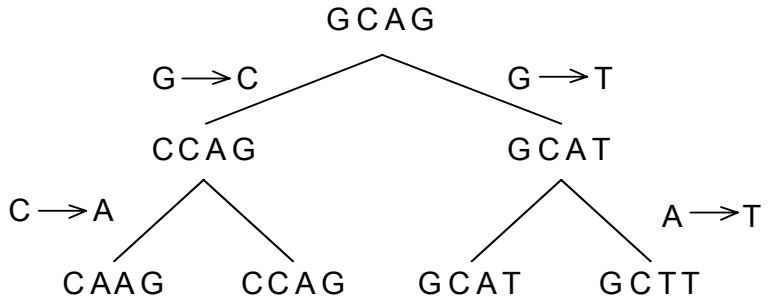

Fig. 3. The phylogenetic tree with minimum total length for the sequences CAAG, CCAG, GCAT, and GCTT.

There are several variations of parsimony. The two simplest and most widely used variations are the Fitch (Fitch 1971). and Wagner (Farris 1970). parsimonies. The Fitch parsimony uses no constraints at all, whereas the Wagner parsimony uses a minimum

of constraints on permissible character-state changes. The Wagner method assumes that characters are measured on an interval scale; thus, this method is appropriate for binary, ordered multistate and continuous characters. The Fitch method allows unordered multistate characters (e.g. in nucleotide or protein sequences). Wagner parsimony assumes that any transformation from one character state to another implies a transformation through any intervening states, as defined by the ordering relationship. The Fitch parsimony allows any state to transform directly into any other state. Both methods permit free reversibility. It means that the change of a character state in either direction is assumed to be equally probable, and character states may transform from one state to another and back again. A consequence of reversibility is that a tree may be rooted at any point with no change in tree length.

The Dollo (Farris 1977). and Camin-Sokal (Camin and Sokal 1965). parsimonies are less common. Dollo parsimony does not allow free reversibility. Each character state can appear only once in a tree. If the distribution of character states is not entirely accounted for by the tree, it must be explained by extra reversals (losses). This has been proposed as a way to analyze restriction site data, where the probability of a loss is much higher than that of a gain. Camin-Sokal was the first parsimony method described in the literature. In that method, the tree is rooted and the root contains all ancestral states. Evolution is assumed to be irreversible; only multiple gains are allowed.

Often, more than one tree with minimum total length may be found by maximum parsimony methods. In order to guarantee to find the best possible tree, an exhaustive evaluation of all possible tree topologies has to be carried out. Parsimony will correctly reconstruct a phylogenetic tree if the number of sequence changes per sequence position is small. In the case of a large number of changes, the proportion of homoplastic changes increases. This can cause errors during tree reconstruction, especially during the analysis of long unbranched lineages, or if the tree contains a mixture of short and long branches. Parsimony methods accurately reconstruct phylogenetic trees in which multiple changes at the same site rarely occur alongside a single branch (Hillis 1996; Kim 1996). Maximum parsimony methods are usually much slower than distance-based procedures.

Recommended software: PHYLIP (Felsenstein), PAUP (Swofford), MEGA (Kumar, Tamura and Nei), and NONA (Goloboff).

2.3. Maximum Likelihood

The maximum likelihood approach for inferring phylogenies from sequence data was introduced by Felsenstein (1981). The Felsenstein (1981). method does not impose any constraint on the constancy of evolutionary rate among lineages. It assigns quantitative probabilities to mutational events, rather than merely counting them. This method compares possible phylogenetic trees on the basis of their ability to predict the observed data. The tree that has the highest probability of producing the observed sequences is preferred. Similarly to maximum parsimony, maximum likelihood reconstructs ancestors at all nodes of each considered tree, but it also assigns branch

lengths based on the probabilities of mutations. For each possible tree topology, the assumed substitution rates are varied to find the parameters that give the highest likelihood of producing the observed sequences.

From many points of view, maximum likelihood seems to be an appealing way to estimate phylogenies (Whelan et al. 2001). All possible mutational pathways that are compatible with the data are considered. Likelihood functions are known to be a consistent and powerful basis for statistical inference (Edwards 1972). This method represents well the evolutionary relationships among sequences. It takes into account various parameters of the evolutionary process, such as the relative probabilities of transitions versus transversions, or the degree to which the rate of evolution differs across sites. The biologist does not need to know the correct values of these parameters; they are estimated in the tree evaluation process.

The main obstacle to the widespread use of maximum likelihood is computational time. Algorithms that find the maximum likelihood score must search through a multidimensional space of parameters. This makes the solution of large-scale problems (>100 sequences) extremely time consuming. Maximum likelihood estimation may be subject to systematic errors. This happens if the model of evolution used to evaluate the likelihood of given trees does not reflect the actual evolutionary processes.

Felsenstein has developed one of the first maximum likelihood programs, DNAML (DNA Maximum Likelihood program), which is included in the PHYLIP package. The program has been used extensively and has proved of great utility in phylogenetic analyses. Computer simulations have shown that the method is highly efficient in estimating true phylogenies under various situations involving violation of evolutionary rate constancy among lineages (see for instance, Hasegawa and Yano 1984; Hasegawa et al. 1991). An improved version of the DNAML program is based on the algorithm by Felsenstein and Churchill (1996). Several models of base substitution are available in DNAML; for example, a model allowing the expected frequencies of the four bases to be unequal and one allowing the expected frequencies of transitions and transversions to be different. DNAML has also several ways of allowing different rates of evolution to occur at different sites. Another program available in the PHYLIP package, DNAMLK (DNA Maximum Likelihood program with molecular clock), implements the maximum likelihood method for DNA sequences under the constraint that the derived phylogenies must be consistent with a molecular clock hypothesis.

Recommended software: PHYLIP (Felsenstein), PAUP (Swofford), MEGA (Kumar, Tamura and Nei), NONA (Goloboff), and PHYML (one of the fastest ML methods by Guindon and Gascuel).

2.4. Bayesian Phylogenetics

The Bayesian approach is relatively new in phylogenetics (Huelsenbeck and Ronquist 2001; Larget and Simon 1999; Li et al. 2000; Rannala and Yang 1996; Yang and Rannala 1997). This method is closely related to maximum likelihood. The optimal hypothesis is the one that maximizes the posterior probability. The posterior probability for a hypothesis is proportional to the likelihood multiplied by the prior probability of

that hypothesis. Prior probabilities of different hypotheses depend on the scientist's assumptions concerning the possible phylogenetic relationships in the data. In many cases, researchers have no information about prior probability distributions. One way of solving this is to specify a uniform prior, in which every possible value of a parameter is given the same probability *a priori*. Compared to maximum likelihood, the advantages of Bayesian methods are higher computational speed and a possibility to incorporate in them complex models of sequence evolution.

Complex parameter-rich models are a problem for maximum likelihood. When the ratio of data points to parameters is low, the estimation of parameters in maximum likelihood can be unreliable. In Bayesian analysis, the final result does not depend on one specific value, but considers all possible parameter values. Even if there are enough data to estimate many parameters, the hill-climbing algorithms that are used to find the maximum likelihood point can be slow or unreliable as the number of parameters increases (particularly if there are complex interactions among some of the parameters). This is not the case for Bayesian methods, because they rely on an algorithm that does not attempt to find the highest point in the space of all parameters.

The best-known Bayesian phylogenetic software programs are MRBAYES written by Huelsenbeck (Huelsenbeck and Ronquist 2001). and BAMBE written by Larget and Simon (1999). MRBAYES uses nucleic acid sequences, protein sequences, and morphological characters to derive phylogenies. It assumes a prior distribution of tree topologies and uses Markov Chain Monte Carlo (MCMC) methods to search the tree space and to infer the posterior distribution of topologies. The BAMBE package infers phylogenetic trees from DNA sequence data. The program uses a prior distribution of trees and implements an arrangement algorithm described in the paper by Mau et al. (1997). The resulting posterior distribution can be used to characterize the uncertainty about not only the tree, but the parameters of the substitution model as well.

Recommended software: MRBAYES (Huelsenbeck) and BAMBE (Larget and Simon).

3. EXISTING MECHANISMS OF RETICULATE EVOLUTION

Classically, the evolution of species has been depicted using phylogenetic trees. An example of such a tree, taken from a famous and controversial paper by Doolittle (1999). is shown in Fig. 4. This way of representing evolution has been questioned by recent developments in molecular phylogenetics. As pointed out by Doolittle (1999). molecular phylogeneticists will have failed to find the true tree of life, not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot be properly represented as a tree. Indeed, the mechanisms of horizontal gene transfer, hybridization, homoplasie, and homologous recombination necessitate the use of network models to illustrate them. Fig. 5 shows an example of a horizontal gene transfer network involving species from the kingdoms of *Bacteria*, *Eukarya*, and *Archaea.*

The fact that most archaeal and bacterial genomes contain genes from multiple sources is challenging for molecular biologists. Following Sonea and Panisset (1976, 1981, Sonea and Mathieu 2000). who showed that horizontal gene transfer (HGT) was a

common evolutionary mechanism among bacteria, Doolittle (1999). emphasized the importance of HGT in the evolution of bacteria and higher groups of organisms.

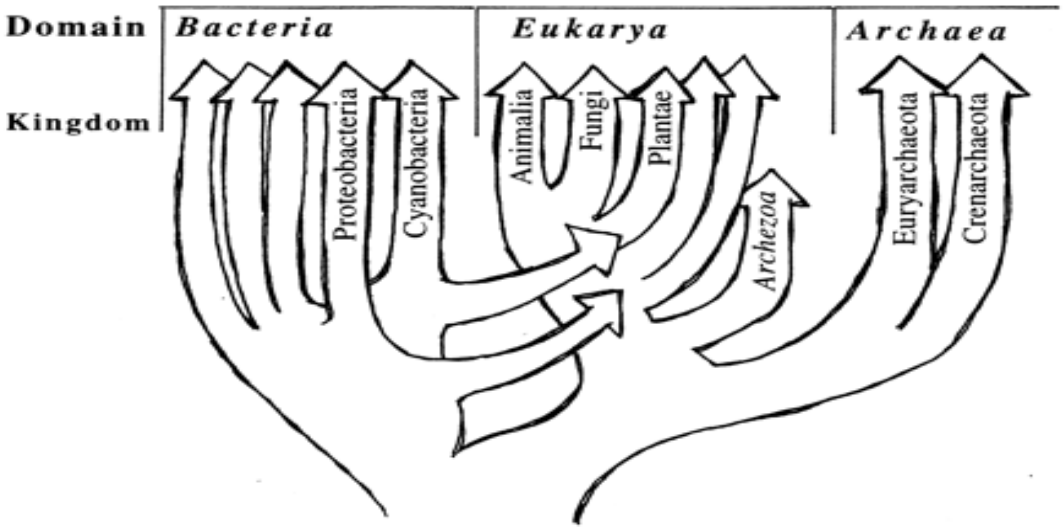

Fig. 4. An example of a phylogenetic tree with a strict hierarchical classification (from Doolittle[1] 1999).

Another reticulate process, hybridization, is prevailing in plants and some groups of animals. In plant evolution, hybridization is critically important as a source of novel gene combinations and as a mechanism of speciation. For instance, in plant breeding desirable traits can be moved from one cultivated or even wild species into another cultivated species (Walter *et al*. 1999). According to one estimate (Stace 1984). there are about 70 000 naturally occurring interspecies plant hybrids in the world.

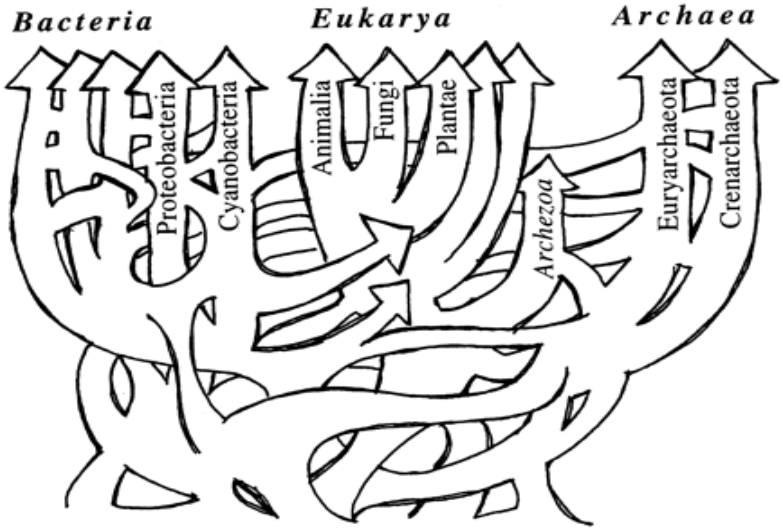

Fig. 5. A reticulated tree, or species network, which might more appropriately represent life's history (from Doolittle[1] 1999, Fig. 3).

[1] Reprinted with permission from Doolittle WF (1999). Phylogenetic classification and the universal tree. Science 284:2124-2128. Copyright 1999 AAAS.

Reticulate evolution shows the lack of independence between lineages. When a reticulation event occurs, two or more independent evolutionary lineages interact at some level of biological organization. In this section, we discuss the most important mechanisms of reticulate evolution which led to the development of the computational methods and software tools that will be described in the next section.

3.1. Horizontal Gene Transfer (HGT)

Horizontal gene transfer is a direct transfer of genetic material from one lineage to another. A HGT between the ancestors of Species 3 and 4 took place in the scenario shown in Fig. 6. Because only a few genes, and sometimes only a part of a gene, are transferred from one organism to another, two evolutionary scenarios (Fig. 7) can take place after a HGT event occurred. The first one, presented in Fig. 7a, is appropriate for the genes acquired through the horizontal transfer shown in Fig. 6, whereas the second one, shown in Fig. 7b, is plausible for all the other genes inherited from the direct species ancestors.

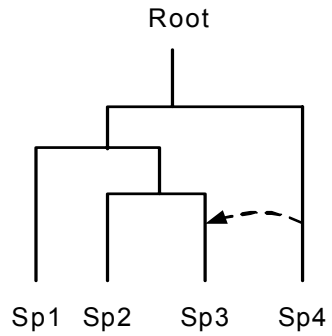

Fig. 6. Horizontal gene transfer.

Horizontal gene transfer is common among bacteria. *Bacteria* and *Archaea* developed the ability to adapt to new environments using the acquisition of new genes through horizontal transfer rather than by the alteration of gene functions through numerous point mutations. Because they are unable to reproduce sexually, bacterial organisms have adopted several mechanisms to exchange genetic materials. The major mechanisms of HGT are the following:

- *Transformation* — This process is most common in bacteria that are naturally transformable. Bacteria take up naked DNA fragments from the environment. This is a common mode of horizontal gene transfer; it can mediate the exchange of any part of a chromosome. Typically, only short DNA fragments are exchanged in this way.
- *Conjugation* — This type of DNA transfer is mediated by conjugal plasmids or conjugal transposons. Even though conjugation requires cell-to-cell contact, it can occur between distantly related bacteria or even between bacteria and eukaryotes. Long fragments of DNA can be transferred by conjugation.

- *Transduction* — This is the transfer of DNA by phage. It requires that the donor and recipient share cell surface receptors for phage binding. It is typically limited to closely related bacteria. The length of DNA transferred by transduction is limited by the size of the phage head.

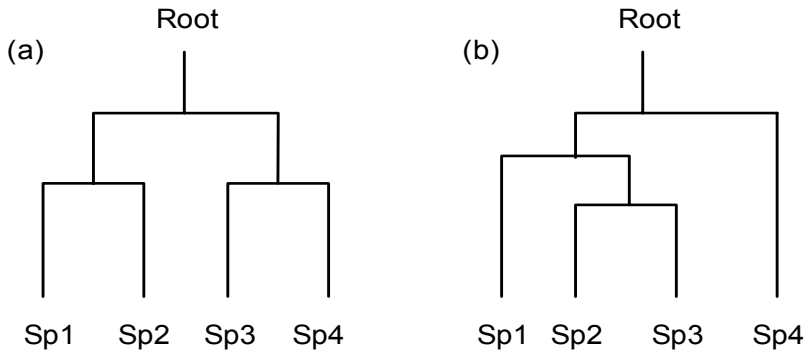

Fig. 7. Horizontal gene transfer: the two possible gene trees.

These mechanisms of horizontal gene transfer can introduce sequences of DNA that have little homology with the remaining DNA of the recipient cell. If the donor DNA and the recipient chromosome share some homologous sequences, the donor sequences can be stably incorporated into the recipient chromosome by homologous recombination. If the homologous sequences are located near sequences that are absent in the recipient, the recipient may acquire an insertion from another strain of unrelated bacteria; such insertions can be of any size.

3.2. Hybridization

Hybridization is another example of reticulate evolution. In Fig. 8, two lineages (Root-Species 2 and Root-Species 3) recombine to create a new species (Species 4). If the new species have the same number of chromosomes as the parent species, the process is called *diploid hybridization*. When it has the sum of the number of its parents' chromosomes, it is called *polyploid hybridization*. The three main mechanisms of hybridization are the following:

- *Autopolyploidization* is a speciation event involving the doubling of the chromosomes within a single species. It produces a bifurcating speciation event in a phylogenetic tree.
- *Allopolyploidization* is a type of hybridization between two species, when an offspring acquires the complete diploid chromosome complements of the two parents. In this case the parents do not need to have the same number of chromosomes. Allopolyploidization results in instantaneous speciation because any backcrossing to the diploid parents is likely to produce a sterile triploid offspring.
- *Diploid hybrid speciation* is a normal sexual event taking place between parents from different but related species. In nearly all cases, the two parents need to have the same number of chromosomes. In this case, successful backcrossing to the parents is possible, so the hybrids have to be isolated from the parents to become new species.

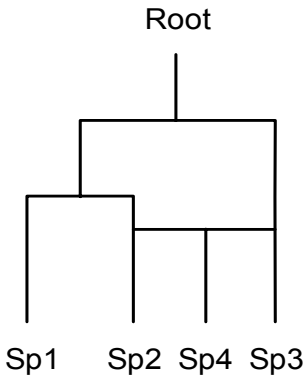

Fig. 8. Hybridization.

In sexually reproducing organisms, hybridization may lead to an entirely female hybrid population. It can sometimes reproduce either by parthenogenesis, or by gynogenesis, forming a new species consisting only of females. Gynogenesis, found among fish, amphibians and reptiles, is a mode of reproduction that allows a unisexual female hybrids population to reproduce, using the sperm from a related bisexual ancestor species to stimulate the development of the eggs (Dawley 1989).

Consider the problem of modeling reticulate evolution after diploid hybrid speciation. In normal diploid organisms, each chromosome consists of a pair of homologs. In the process of diploid hybridization, the hybrid inherits one of the two

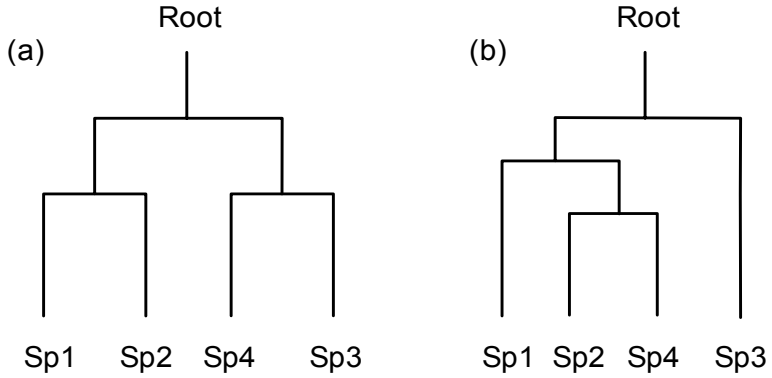

Fig. 9. Hybridization: two possible gene trees for the hybridization event shown in Fig. 8.

homologs for each chromosome from each of its two parents. Since the genes from both parents are contributed to the hybrid, the evolution of genes inherited from each parent can be represented on separate trees inside a network model. Classical phylogenetic analysis of the four species involved in a hybrid speciation event (Fig. 8) will produce either the tree in Fig. 9a or the one in Fig. 9b.

Hybridization is very common in plants, fish, amphibians and reptiles, and is virtually absent in other groups, particularly in birds, mammals, and most arthropods.

The latter groups are only occasionally affected by hybrid speciation. They usually produce triploids which can only reproduce by asexual modes.

3.3. Homoplasy

Homoplasy is the development of organs or other bodily structures within different species, which resemble each other and have the same functions, but did not have a common ancestral origin. These organs arise via convergent evolution and are thus analogous, not homologous to each other. For example, the wings of insects, birds and bats, which are all used for flying, are homoplastic (meaning: similar in form and structure, but not in origin). As shown in Fig. 10, the wings of birds and bats are structurally different: the bird wing (a) is supported by digit number 2, the bat wing (b) by digits 2-5.

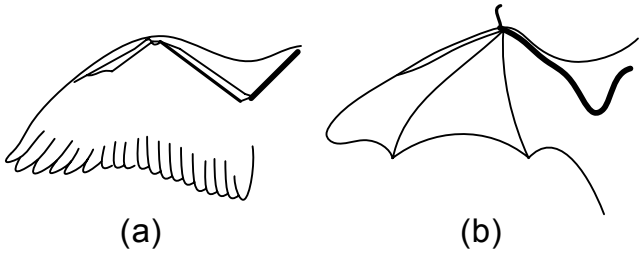

Fig. 10. The wings of birds and bats.

Since homoplasy is a feature shared by a set of species but not present in their common ancestor, it can cause problems during phylogenetic reconstruction (Smouse 2000). As pointed out by Legendre (2000b). in the case of homoplasie, the objective of reticulation analysis is not to model actual reticulation events, but to produce a diagram containing reticulation branches to describe more accurately the common patterns found in the data. If distant species seem to be artificially close to one another, the addition of reticulation branches to a tree produces a reticulogram (i.e. reticulated cladogram) which describes the data better than a tree would do.

Fig. 11, from Makarenkov and Legendre (2000). is an example of a reticulogram built for the primates data originally considered by Hayasaka et al. (1998). First, a distance matrix over 12 species of primates was computed on the base of protein-coding mRNA (898 bases). The phylogenetic tree was constructed from the distance matrix using the neighbor-joining method (Saitou and Nei 1987). The NJ tree is represented by solid lines in Fig. 11. Four groups of primates were found in the phylogeny. The reticulogram building algorithm (Makarenkov and Legendre 2000). added 5 reticulation branches (dashed lines) to the primate phylogeny. From the mathematical point of view, each reticulation branch improved the least-squares fit of the distance matrix, compared to the classical phylogenetic tree. From the biological point of view, the reticulation branches are long and they are formed between distant groups, so, they most likely represent homoplasy. For example, consider *Tarsius*: its position in the phylogeny of primates is uncertain (E. Douzery, personal communication). *Tarsius* is clustered with *Lemur catta* in the NJ phylogenetic tree (solid lines), but it is also close to **Hominoidea**

(reticulation branch between *Tarsius* and *Pongo*) and **Cercopithecoidea** (reticulation branch between *Tarsius* and *Macaca fascicularis*). Thus, modeling phylogenetic relationships among primates with reticulograms allowed the authors to depict alternative evolutionary features, homoplasy in this case, which cannot be represented by means of a classical tree model.

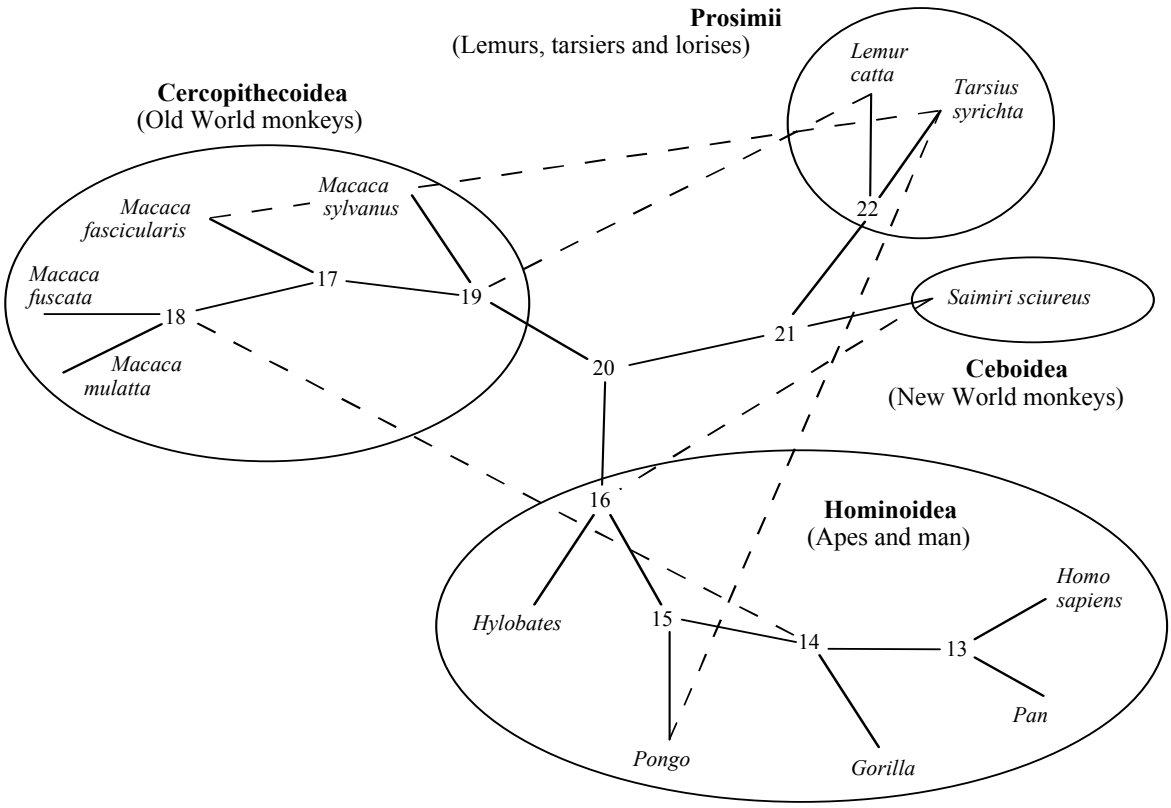

Fig. 11. Reticulogram representing homoplasy among primates (Makarenkov and Legendre[2] 2000, Fig. 2).

3.4. Genetic Recombination

Recombination refers to any process that gives rise to new combinations of genetic material, such as the reassortment of parental genes through crossing over during meiosis, which leads to the formation of gametes. Recombination creates reticulate evolution within lineages. Homologous chromosomes become paired during the prophase of meiosis, as shown diagrammatically in Fig. 12a. In crossing over, two homologous chromosomes swap a portion of their genetic material (Fig. 12b). After separation, each member of a pair of homologues contains parts of its partner's genetic material (Fig. 12c).

[2] Reprinted with permission from Makarenkov V and Legendre P. (2000). Improving the additive tree representation of a dissimilarity matrix using reticulations. In: HAL Kiers, J-P Rasson, PJF Groenen and M Schader, eds. Data Analysis Classification and Related Methods. Berlin: Springer, pp 35–40. Copyright 2000 Springer Verlag.

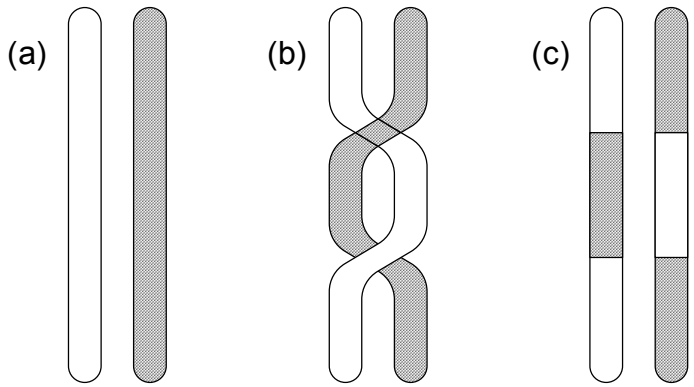

Fig. 12. Homologous chromosomes exchanging genetic material (their central portions) by crossing over.

The exchange of genetic material between homologous chromosomes, called *homologous genetic recombination* (also known as *general recombination* or *general homologous recombination*), may occur at any part of a chromosome. This event can take place in bacteriophage recombination, in recombination following bacterial conjugation, and during the formation of plasmid multimers. Site-specific recombination involves the exchange of genetic material at very specific sites only. Examples include the integration of a bacteriophage lambda into a host chromosome to form a prophage and the rearrangement of chromosomal DNA prior to expressing antibody genes.

Recombination has an important influence on genomes and on the genetic structure of populations. It affects biological evolution at many different levels and explains a considerable amount of genetic diversity in natural populations of sexually-reproducing species. In general, genes located in regions of the genome with low levels of recombination have low levels of polymorphism. Recombination reshuffles the existing variation and even creates new gene variants at the amino acid level. It shapes the genetic structure of natural populations (Anderson and Kohn 1998; Feil et al. 2001). and the action of natural selection (Marais et al. 2001).

Many applications in biology today are based on the estimation of phylogenetic trees. Since recombination leads to mosaic genes, where different regions may have different phylogenetic histories, it is important to take this process into account during the tree reconstruction. A number of statistical methods for the detection of recombination in DNA sequences are available. Their detailed description can be found in Posada and Crandall (2001a). who estimated the performance of 14 different algorithms dealing with recombination.

4. ALGORITHMS AND SOFTWARE FOR DETECTING RETICULATE EVOLUTION

In this section we discuss the algorithms and related software that have been created for the detection and visualization of patterns of reticulate evolution. The web page (http://evolution.genetics.washington.edu/phylip/software.html) supported by J. Felsenstein contains a comprehensive list of phylogeny reconstruction tools, which includes 251 software packages and 29 servers (available on January 12, 2006). In this paper we focus on the software that include algorithms for building and visualizing

reticulate phylogenies. For a review of network-like structures used to detect reticulate evolution, readers can also consult the papers by Posada and Crandall (2001b). and Linder et al. (2003 and 2004). A special section dedicated to reticulate evolution and related problems has been published by the Journal of Classification (Legendre 2000a). with contributions from Sneath, Smouse, Lapointe, Rohlf, and Legendre.

Reticulate evolution has long been neglected in phylogenetic analyses. The first methods for studying the mechanisms of reticulate evolution started to appear in the mid-1970s (Sneath et al. 1975; Sonea and Panisset 1976). Several tentative methods have been proposed for the identification of reticulate evolution in nucleotide sequences. They include displays of compatibility (Sneath et al. 1975). tests for clustering (Stephens 1985). a randomization approach (Sawyer 1989). and an extension of the parsimony method of phylogenetic reconstruction that allows recombination (Hein 1993). Rieseberg and Morefield (1995). developed a computer program, RETICLAD, allowing one to identify hybrids based on the expectation that they would combine the characters of their parents. However, this program can only find reticulation events between terminal branches of a tree. Rieseberg and Ellstrand (1993). showed examples where the program appears to work well. The popular method of split decomposition enables the representation of data in the form of a splitsgraph revealing the conflicting signals contained in the data (Bandelt and Dress 1992a, 1992b). In a splitsgraph, a pair of nodes may be linked by a set of parallel edges depicting alternative evolutionary hypotheses. Hallet and Lagergren (2001). showed how lateral gene transfer events can be detected by evaluating topological differences between species and gene trees. Bryant and Moulton (2002, 2004). introduced a network-inferring method, NeighborNet, allowing the reconstruction of planar phylogenetic networks. Each of these methods has features that make them useful for the analysis of particular types of data, and they all have a role to play in detecting and describing reticulate evolution. Legendre and Makarenkov (2002). and Makarenkov and Legendre (2004). proposed to use reticulograms for detecting reticulation events in evolutionary data. They developed a distance-based method to infer reticulate phylogenies. That method uses the topology of a phylogenetic tree as a supporting structure for building a reticulogram. The other network-inferring techniques considered in the present paper are the following: HGT detection of Boc and Makarenkov (2003). and Makarenkov et al. (2004, 2006). Statistical parsimony (Templeton et al. 1992). Netting (Fitch 1997). Median networks (Bandelt et al. 1995 and 2000). Median-joining networks (Foulds et al. 1979; Bandelt et al. 1999). Molecular-variance parsimony (Excoffier and Smouse 1994). Pyramids (Diday and Bertrand 1986). and Weak hierarchies (Bandelt and Dress 1989).

4.1. Horizontal Gene Transfer Detection (Hallet and Lagergren)

Hallet and Lagergren (2001). and Addario-Berry et al. (2003). developed a model of horizontal gene transfer which compares the evolution of a set of gene trees to a species

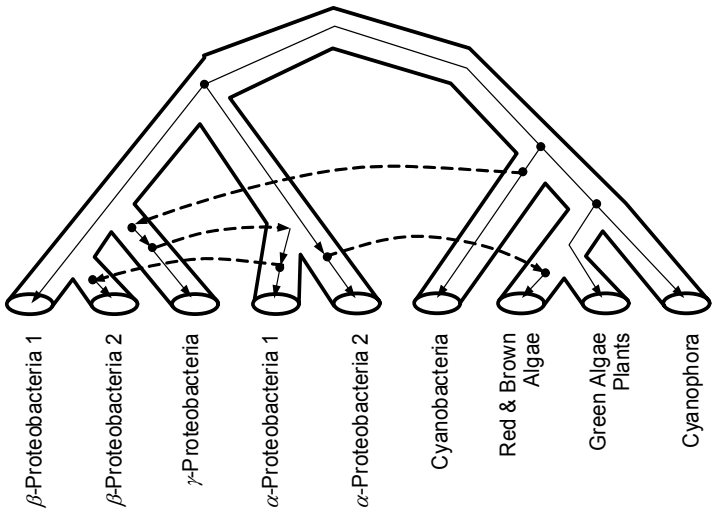

Fig. 13. Horizontal gene transfer scenario of the *rbcL* gene identified by Hallet and Lagergren (2001).

tree. The algorithm proceeds by mapping given gene trees into the species tree. A number of constraints are introduced in the model to make this mapping biologically meaningful. If a multiple copy of a gene appears in the species tree, the algorithm recognizes it as a possible lateral gene transfer. A scenario of lateral transfer of the *rbcL* gene is presented in Fig. 13 (example taken from Hallet and Lagergren 2001). This model also includes an activity parameter α that defines the number of genes allowed to be simultaneously active.

The algorithm is implemented in the Lateral Transfer software available at: http://cgm.cs.mcgill.ca/~laddar/lattrans/. This program also includes an option allowing one to seek scenarios under a combined lateral transfer/gene duplication model.

4.2. Horizontal Gene Transfer Detection (Boc and Makarenkov)

Two models for detection of horizontal gene transfer have been considered by Boc and Makarenkov (2003). Makarenkov et al. (2004, 2006). Both models use a distance approach and are based on the reconciliation of the topologies of the gene and species phylogenetic trees built for the same set of species.

The first model (Boc and Makarenkov 2003; Makarenkov et al. 2004). assumes partial gene transfer; it is based on the computation and optimization of the minimum path-length distances in a directed network (Fig. 14a). In this model, the phylogenetic tree is transformed into a connected and directed graph in which a pair of species can be linked by several paths. The second model (Makarenkov et al. 2006). assumes complete transfer: the species phylogenetic tree is gradually transformed into the gene phylogenetic tree by adding to it a horizontal gene transfer in each step. During this transformation, only a tree topology is taken into account and modified (Fig. 14b). Though the second model is less general, a fast and effective algorithm has been described to solve the problem. Moreover, two criteria, one metric and the other topological, can be combined in the optimization procedure (Makarenkov et al. 2006).

Both models produce scenarios of horizontal transfers of the given gene. According to Makarenkov et al. (2006). the use of the topological criterion, which is the Robinson and Foulds (1981). topological distance, enables a better detection of gene transfers compared to the metric criterion (least-squares function); one of the considered examples concerned the well-known *rbcL* dataset from Delwiche and Palmer (1996). Among the recent developments in the field of HGT detection techniques, a validation procedure (bootstrapping) for gene transfer have been designed to measure the reliability of an individual transfer as well as that of a whole gene transfer scenario; see Makarenkov et al. (2006) for more detail. These methods were included in the T-REX package (Makarenkov 2001). which provides users with a friendly visualization support. T-REX is available at the following URL: http://www.trex.uqam.ca.

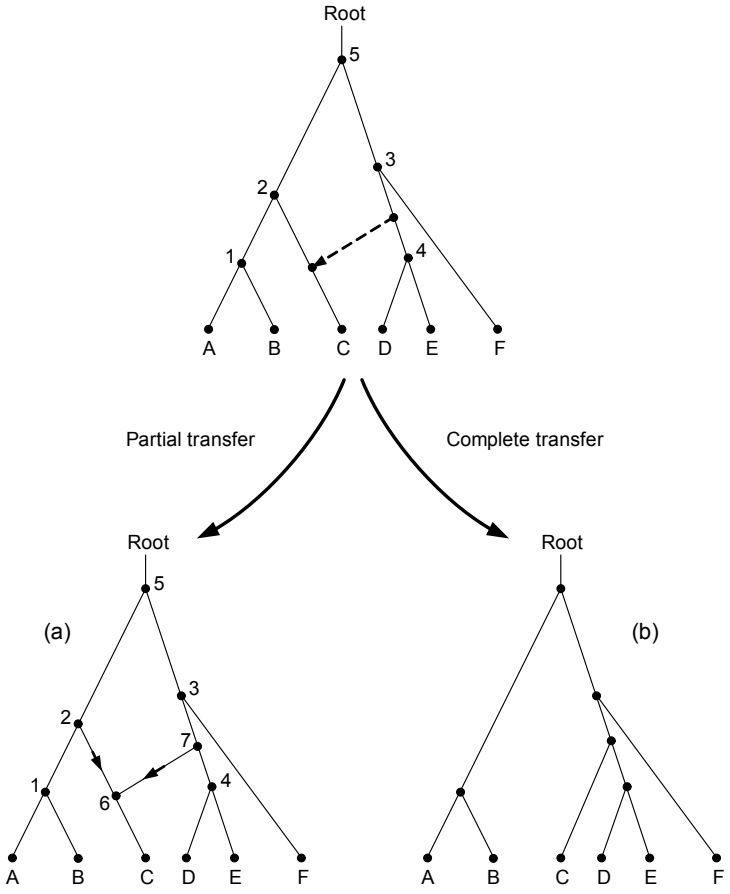

Fig. 14. Two evolutionary models, assuming that either a partial (a, model 1) or a complete (b, model 2) horizontal gene transfer has taken place. In the first case, only a part of the gene is transferred and the tree is transformed into a directed network, whereas in the second, the donor gene replaces the homologous gene of the host and the initial tree is transformed into a different phylogenetic tree.

The main steps of the HGT detection algorithm (model 1) described in Boc and Makarenkov (2003). and Makarenkov et al. (2004). are the following. The algorithm first identifies the topological differences between the species and gene phylogenies. Then, it uses a least-squares optimization procedure to find where horizontal gene transfers

between branches of the species tree may have taken place. A species phylogenetic tree T whose leaves are labeled according to the set of n taxa must have been constructed before starting the HGT detection algorithm. Tree T can be inferred from sequence or distance data using an appropriate tree fitting method. The tree should be explicitly rooted; the position of the root is important in this model. Likewise, a gene tree T_1 must have been inferred using a similar procedure; the leaves of T_1 are labeled according to the same set of n taxa labels as in the species tree T. Without loss of generality, the method assumes that T and T_1 are binary trees whose internal nodes are all of degree 3 and whose number of branches is $2n$–3.

If the topologies of T and T_1 are identical, the algorithm concludes that the evolution of the gene followed that of the species, and no horizontal gene transfers between branches of the species tree have taken place. However, if the two phylogenies are topologically different, it may be due to horizontal gene transfers. In this case, the gene tree T_1 can be mapped into the species tree T by fitting, by least squares, the branch lengths of T to the pairwise distances in T_1 [details on this least-squares fitting technique are available in Barthélemy and Guénoche (1991). and Makarenkov and Leclerc (1999)].

The goal of the next step is to determine the order of addition of HGT branches to the tree, considering all possible HGT connections between pairs of branches in T. There are $(2n–3)(2n–4)$ possibilities for the addition of the first HGT branch. This is the maximum number of different directed inter-branch connections in a binary phylogenetic tree with n leaves. The HGT connection providing the largest contribution to the decrease of the least-squares coefficient Q is the most probable case, in the least-squares sense, of horizontal gene transfer. That connection is added to the tree, transforming T into a network. After the first HGT branch has been added to T, all its branches, including the new HGT branch, are reassessed to fit optimally the inter-leaf distances from the gene tree T_1. Then, the best second, third, and so forth, HGT branches are added to T in the same way. Starting from the second HGT branch, addition of any new HGT connection takes into account all previously added HGTs. The algorithm stops when a predetermined number of HGT branches have been added to T. The phylogenetic network obtained in this way represents the best possible scenario, according to least squares, of horizontal transfer of the gene under study.

The following strategy was adopted to estimate the value of the least-squares coefficient Q for a given HGT branch (a,b). First, the algorithm lists all pairs of taxa such that the path between them can include the new HGT branch (a,b); this is controlled by a number of biological rules incorporated into the model. Second, the algorithm lists the pairs of taxa for which the minimum path-length distance may decrease after the addition of the branch (a,b). Third, the algorithm looks for the optimal value l of the length of branch (a,b), keeping fixed the lengths of all the other tree branches; see below. Fourth, all tree branch lengths are reassessed one at a time to improve the fit.

The set $A(a,b)$ of all pairs of taxa, such that the minimum path-length distances between them may change if the HGT branch (a,b) is added to the tree T (Fig. 15), is found as follows: $A(a,b)$ is the set of all pairs of taxa ij such that:

$$Min\{d(i,a) + d(j,b);\ d(j,a) + d(i,b)\} < d(i,j), \quad (1)$$

where $d(i,j)$ is the minimum path-length distance between taxa i and j in T; vertices a and b are located in the center of branches (x,y) and (z,w), respectively.

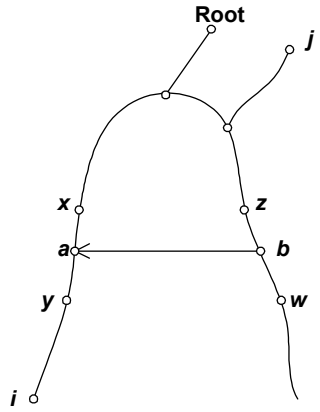

Fig. 15. The minimum path-length distance between taxa i and j can be affected by the addition of a new branch (a,b) representing the horizontal gene transfer between branches (z,w) and (x,y) in the species tree.

The following function is used:

$$dist(i,j) = d(i,j) - Min\{d(i,a) + d(j,b);\ d(j,a) + d(i,b)\}. \quad (2)$$

Thus, $A(a,b)$ is the set of all leaf pairs ij such that $dist(i,j) > 0$. The least-squares objective function to be minimized, with l used as an unknown variable, is formulated as follows:

$$Q(ab,l) = \sum_{dist(i,j)>l} (Min\{d(i,a) + d(j,b);\ d(j,a) + d(i,b)\} + l - \delta(i,j))^2 + \sum_{dist(i,j)\le l} (d(i,j) - \delta(i,j))^2, \quad (3)$$

where $\delta(i,j)$ is the minimum path-length distance between taxa i and j in the gene tree T_1. The function $Q(ab,l)$, measures the gain in fit when a new HGT branch (a,b) of length l is added to the species tree T. When the optimal value (i.e. the one that minimizes the function Q) of a new branch (a,b) is found, this computation is followed by an overall polishing procedure for all branch lengths in T. To reassess the length of any branch of T, one can use Equations 1, 2, and 3, assuming that the lengths of all the other branches are fixed. These computations are repeated for all pairs of branches in the species tree T. After all pairs of branches in T have been reassessed, only the HGT corresponding to the smallest value of Q is retained for addition to T. This algorithm requires $O(kn^4)$ operations to produce a HGT scenario with k HGT branches.

4.3. Reticulogram Reconstruction and the T-REX Package

In this subsection, we discuss the method for inferring connected and undirected reticulated networks (also called *reticulograms* or *reticulated networks*) from matrices of

evolutionary distances between species. This method was used in several biological problems and turned up to be relevant for detecting hybrids, homoplasy and HGT, as well as biogeographic networks; see the papers by Makarenkov and Legendre (2000 and 2004). Legendre and Makarenkov (2002). and Makarenkov et al. (2004). The method is distance-based and works according to the following scheme: first, it infers a phylogenetic tree from a distance matrix using one of the existing tree fitting algorithms. Supplementary branches, called *reticulation branches*, are then added to the tree structure, one at a time, each one minimizing a least-squares or weighted least-squares loss function. The addition of reticulation branches stops when the minimum of a special goodness-of-fit function is reached. Four such functions have been proposed; each one takes into account the value of the least-squares criterion as well as the total number of branches of the reticulated network under construction. This algorithm requires $O(kn^4)$ time to add k reticulation branches to a phylogenetic tree with n leaves.

We will now describe the main features of this technique and show how it can be applied to study the evolution of a group of honeybees of the genus *Apis*. Let δ be a distance function used to estimate phylogenetic distances between the elements of the set X containing n taxa, and T a phylogenetic tree inferred from δ by means of an appropriate tree reconstruction method. Let d be an expression of the distances in T between the taxa of X (i.e. pairwise distances between the leaves of T). A reticulated network comprises more branches and thus uses more parameters than a phylogenetic tree. As in all statistical models, more parameters mean better fit, but fewer degrees of freedom and a loss of simplicity. A special cost criterion should be used to estimate how many reticulation branches have to be added to a network. The authors proposed four goodness-of-fit criteria to determine when to stop adding branches to a reticulogram (Makarenkov and Legendre 2004). When the exact number of reticulation branches is unknown, as it is often the case in evolutionary problems, one can stop the addition of new branches when the minimum of the selected criterion is reached.

The total number of nodes in a binary unrooted phylogenetic tree with n leaves is $2n–2$; this includes $n–2$ intermediate nodes and n terminal nodes (leaves, taxa). The maximum number of undirected branches one might place in a reticulated network inferred from a binary phylogenetic tree with n leaves is $(2n–2)(2n–3)/2$. Here we counted all possible connections between leaves, between nodes, and between leaves and nodes. However, any metric distance can be represented by a complete graph with $n(n–1)/2$ branches between the leaves. Thus, any of these two limits, $(2n–2)(2n–3)/2$ or $n(n–1)/2$, can be considered as the maximum possible number of branches in a reticulated network. If the latter limit is considered, the number of degrees of freedom of a reticulated network with N branches is $n(n–1)/2 – N$.

It would be reasonable to consider a penalty function opposing the loss in degrees of freedom to the gain in fit. The first proposed goodness-of-fit function is called Q_1:

$$Q_1 = \frac{\sqrt{\sum_{i \in X} \sum_{j \in X} (d(i,j) - \delta(i,j))^2}}{n(n-1)/2 - N} = \frac{\sqrt{Q}}{n(n-1)/2 - N}. \quad (4)$$

The numerator of this function is the square root of Q, which is the sum of squared differences between the values of the given distance δ and the corresponding reticulation estimates d. Interestingly, as was confirmed by a simulation study carried out by Legendre and Makarenkov (2002), function Q_1 usually has only one minimum over the interval $[2n–3, n(n–1)/2]$ of possible values of N. This minimum defines a stopping rule for addition of new branches to the reticulate phylogeny.

The least-squares function itself may be used as the numerator for a goodness-of-fit measure. Thus, one can consider a slightly different criterion, called Q_2, which usually adds more reticulation branches to the network than Q_1:

$$Q_2 = \frac{\sum_{i \in X} \sum_{j \in X} (d(i,j) - \delta(i,j))^2}{n(n-1)/2 - N} = \frac{Q}{n(n-1)/2 - N}. \quad (5)$$

One can also consider the Akaike Information Criterion (AIC) which is a useful and well-known statistic (Akaike 1987). A model with a minimum value of AIC may be chosen to be the best-fitting solution among several competing models. In our algorithm, the Akaike rule would select the model that minimizes the following quantity:

$$\mathrm{AIC} = \frac{Q}{(2n-2)(2n-3)/2 - 2N}. \quad (6)$$

Another popular statistical estimator, the Minimum Description Length (MDL) criterion introduced by Rissanen (1978), can be also used as stopping rule for the reticulogram construction algorithm. The MDL criterion, which is closely related to the AIC statistics, is computed as follows:

$$\mathrm{MDL} = \frac{Q}{(2n-2)(2n-3)/2 - N\log(N)}. \quad (7)$$

The reticulogram in Fig. 16 represents the evolutionary relationships within a group of honeybees. Makarenkov et al. (2004). applied the method for detection of reticulate evolution to the DNA sequence data of six species of honeybees (genus *Apis*). The DNA sequences (677 bases) considered in this work were taken from the SPLITSTREE package (Huson 1998). The bee phylogenetic tree was reconstructed by neighbor-joining (NJ; Fig. 16, full lines), and by maximum likelihood (ML which produced the same tree topology as NJ). The tree was validated by bootstrapping (Felsenstein, 1985) using 100 replicates for ML, and 1000 replicates for NJ. The phylogeny clearly separated two groups of bees,

with the species *A. mellifera*, *A. dorsata*, and *A. cerana* forming the first group and species *A. andreniformis*, *A. florae,* and *A. koschevnikovi* the second group. The bootstrap support for the group separation branch was 88% for NJ and 89% for ML.

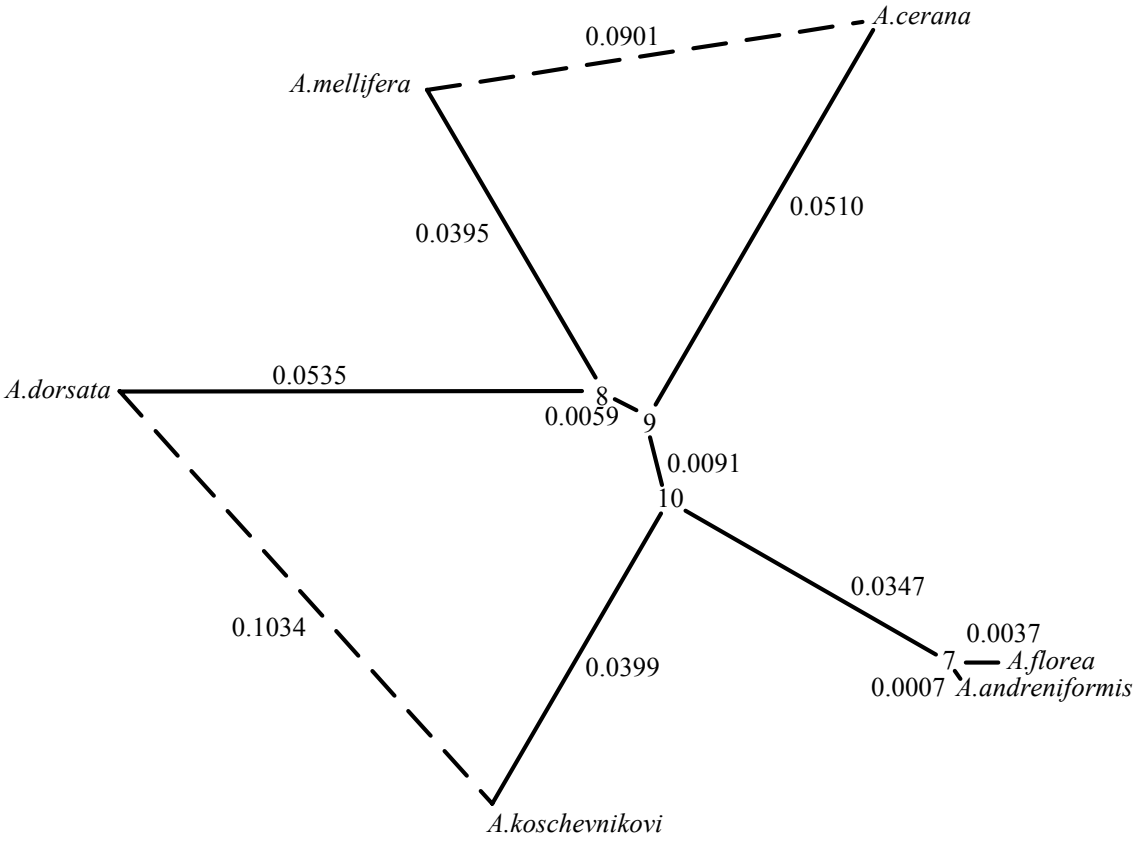

Fig. 16. Reticulogram representing the possible evolution of *Apis* honeybees.

The reticulogram construction algorithm was then applied to the phylogenetic tree provided by NJ. The goodness-of-fit function Q_2 was chosen as the stopping rule for addition of new branches. Two reticulation branches (dashed lines in Fig. 16) were added to the phylogenetic tree by the algorithm. The minimum of the goodness-of-fit function Q_2 was reached at the second step of the algorithm, decreasing the value of Q_2 from 0.000024 to 0.000020, whereas the value of the least-squares loss function Q dropped from 0.000143 to 0.000078. The decrease of Q after addition of only two reticulation branches was dramatic for these data. The gain in fit was 27.3% (Q = 0.000104) after the addition of the first branch, linking bees *A. mellifera* and *A. cerana*, and the total gain was 45.5% (Q = 0.000078) after the addition of the second branch, linking species *A. dorsata* and A. *koschevnikovi*. These results indicate the relevance of the reticulogram model for the honeybee data, where reticulation branches bring to light conflicting features that are embedded in the phylogenetic tree. The poor bootstrap support (57% and 54% for NJ and ML, respectively) obtained for the branch linking

nodes 8 and 9 of the tree is an indication of a close relationship between *A. mellifera* and *A. cerana*.

How should the reticulation branches be interpreted? The first reticulation branch linking *A. mellifera* and *A. cerana* is only about twice the length of the branches of the tree. It may be interpreted as a possible hybridization event involving the ancestors of the two species which occurred during the evolutionary process. This reticulation branch shows that the two species are genetically closer to each other than it is represented by the phylogenetic tree. Fig. 16 depicts what may have happened during evolution: a recent ancestor of *A. cerana* may have hybridized with one of the recent ancestors of *A. mellifera* to produce the modern *A. mellifera* bee. Or, conversely, a recent ancestor of *A. mellifera* may have hybridized with one of the recent ancestors of *A. cerana* to produce the modern *A. cerana* species. This hypothesis is in agreement with the belief, based on biological and behavioral data, that *A. mellifera* and *A. cerana* have shared a close common ancestor in relatively recent times (Milner 1996). The other reticulation branch, linking the species *A. dorsata* and *A. koschevnikovi*, also reveals that the relationship between these two species is closer than depicted by the phylogenetic tree.

The reticulogram reconstruction algorithm has been implemented in the T-REX (*tree and reticulogram reconstruction*) package (Makarenkov 2001) available for the Windows and Macintosh platforms and as a free web server. The program includes a number of popular algorithms for the reconstruction of phylogenetic trees and reticulograms from a distance matrix. Phylogenetic trees can also be inferred from data matrices containing missing values. T-REX provides a window with the tree or reticulogram fitting statistics and a window with the tree or reticulogram drawing. For tree reconstruction, the program includes six methods for fitting a tree metric (distance representable by a tree with non-negative branch lengths) to a distance matrix: the ADDTREE method of Sattath and Tversky (1977). the Neighbor-Joining (NJ) method of Saitou and Nei (1987). the BioNeighbor-Joining (BioNJ). method of Gascuel (1997a). the Unweighted Neighbor-Joining (UNJ) method of Gascuel (1997b). the Circular order reconstruction method of Makarenkov and Leclerc (1997). and Yushmanov (1984). and the Weighted least-squares method (MW) of Makarenkov and Leclerc (1999). Four fitting methods are offered for reconstruction of phylogenies from partial distance matrices (i.e. matrices containing missing values): the Triangle method of Guénoche and Leclerc (2001). the Ultrametric procedure for missing values estimation of De Soete (1984). and Landry and Lapointe (1997). the Additive procedure for missing values estimation of Landry and Lapointe (1997). and the Modified weighted least-squares method MW* of Makarenkov and Lapointe (2004). With the reticulogram inferring option, the program first computes a phylogenetic tree using one of the six available tree-fitting algorithms. Then, at each step of the reticulogram building procedure, a reticulation branch minimizing the least-squares or weighted least-squares loss function is added to the network. When the horizontal gene transfer option is selected, the program maps the gene tree into the species tree following the procedures by Boc and Makarenkov (2003). and Makarenkov et al. (2006).

4.4. Statistical Parsimony

The statistical parsimony method was developed by Templeton et al. (1992). It estimates the maximum number of differences among haplotypes which are caused by single substitution events. This estimation is complemented with a 95% statistical confidence. Multiple substitutions at a single site are neglected. The maximum number of differences is called the parsimony limit. The algorithm initially connects haplotypes differing by one change, then those differing by two, by three, and so on. The algorithm stops when either all the haplotypes are connected in a network or the parsimony connection limit is reached. Since the statistical parsimony method connects haplotypes with small differences, it shows the similarities rather than the dissimilarities between the haplotypes and provides an empirical assessment of deviations from parsimony. This method enables the identification of putative recombinants by looking at the spatial distribution, in the sequence, of the homoplasies defined by the network.

This method is implemented in the TCS Java computer program which estimates gene genealogies including multifurcations and/or reticulations. The corresponding software is described in the paper by Clement et al. (2000). It is available at the following web site: http://inbio.byu.edu/Faculty/kac/crandall_lab/tcs.htm. An example of the network generated by statistical parsimony for the *Apis* honeybees of Fig. 16 is shown in Fig. 17.

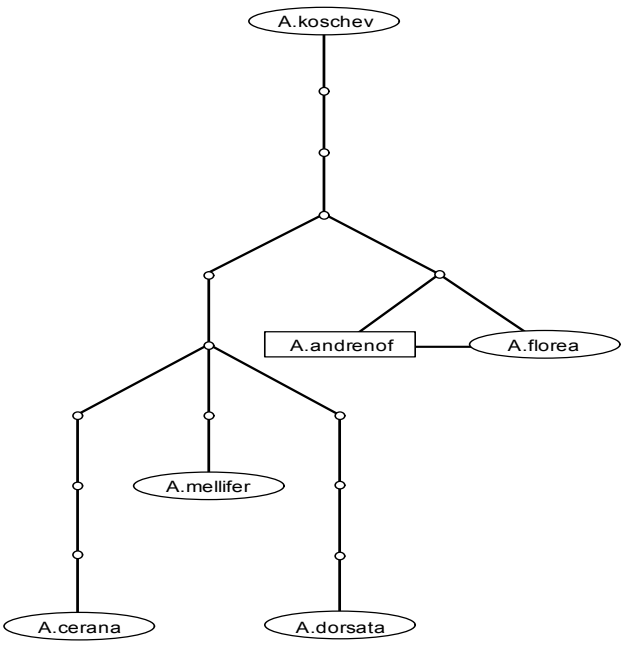

Fig. 17. Phylogenetic network for the *Apis* honeybees, generated by the TCS program.

4.5. Netting

This distance-based method (Fitch 1997). generates all the equally most parsimonious trees for a given data set and connects the leaves (sequences) into a single network. First, the algorithm connects the pair of sequences having the largest similarity. Then, it

connects the joined sequences with the sequence having the largest similarity to them. This connection is made in such a way that the three pairwise differences are satisfied. Thus, the patristic distance between two sequences is necessarily equal to the number of differences. A new connection is added to the network if homoplasy is encountered. Gaps and invariant positions are not considered in the analysis. Since the method tends to satisfy all distances among haplotypes, the number of dimensions may be high and the representation of the network may become difficult.

4.6. Median Network

In the median-network method (Bandelt et al. 1995; Bandelt et al. 2000). sequences are first transformed into binary data, whereas constant sites are excluded from the analysis. Each split is encoded as a binary character taking values 0 and 1. Sites supporting the same split are clustered into one site which is then weighted by the number of clustered sites. Thus, this method represents haplotypes as binary vectors. Consensus or median vectors are estimated for each triplet of vectors until the median network is derived. With more than 30 haplotypes, the resulting median networks are very difficult to display due to the presence of high-dimensional hypercubes. Luckily, the size of a median network can be reduced using predictions from coalescence theory. All the most parsimonious trees are represented in a median network. Initially designed for the analysis of mtDNA data, median networks can be built for other kinds of data, as long as the data are binary or can be reduced to that form.

4.7. Molecular-variance Parsimony

The molecular-variance parsimony method developed by Excoffier and Smouse (1994). uses population statistics to select an optimal network. The algorithm generates a number of minimum-spanning trees which are translated into matrices of patristic distances among haplotypes. These matrices are used to compute some of relevant population statistics such as: squared patristic distances among haplotypes, geographic partitioning of populations, and functions of haplotype frequencies. The algorithm chooses the optimal trees by minimizing the molecular variance. This method makes explicit use of the sample haplotype frequencies and geographic subdivisions, and presents the solution in the form of a set of optimal networks.

Excoffier, Schneider, and Roessli have released the ARLEQUIN package, the program for carrying out the population genetics analysis. ARLEQUIN contains a number of useful methods including estimation of gene frequencies, testing of linkage disequilibrium, and analysis of diversity between populations. Another relevant feature of this program consists in its ability to compute a variety of evolutionary measures including the Jukes and Cantor (1969). Kimura 2-parameter (1981), and Tamura and Nei (1984). distances with or without correction for gamma-distributed rates of evolution. ARLEQUIN also computes minimum spanning tree networks. The executable for Windows, MacOS and Linux, Java source code, and a comprehensive documentation for this software are available at the following web site: http://acasun1.unige.ch/arlequin.

4.8. Median-joining Network

The median-joining network algorithm (Bandelt et al. 1999; Foulds et al. 1979). starts by combining the minimum-spanning trees within a single network. Using a parsimony criterion, the procedure adds to the network median vectors representing missing intermediates. Median-joining networks can be used to analyze large datasets and multistate characters. This technique is extremely fast and is able to process thousands of haplotypes in reasonable time. It can also be applied to amino acid sequences. However, the method cannot cope with recombinations, which restricts its application to the population level.

Röhl, Forster and Bandelt have written the NETWORK 4.1 program, the software for inferring median-joining networks from non-recombining DNA, STR, amino acid, and RFLP data. The networks can be constructed using either the reduced median network or the median-joining network method. Windows and DOS executables of the program are freely available at: http://www.fluxus-engineering.com/sharenet.htm. An example of the reduced median-joining network presented in Fig. 18 was calculated using NETWORK 4.1. This network was inferred for the dataset of *Apis* honeybees from Fig. 16.

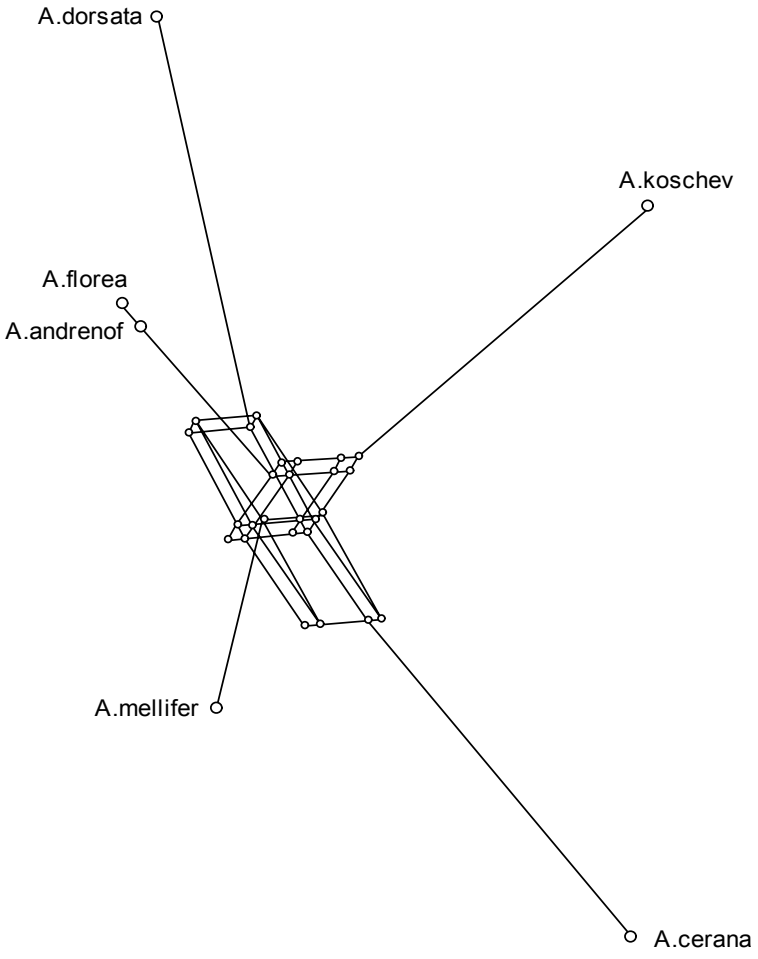

Fig. 18. Median-joining network for the *Apis* honeybees, generated by NETWORK 4.1.

4.9. Split Decomposition

Bandelt and Dress (1992a). designed the technique of split decomposition which transforms evolutionary distances into a sum of weakly compatible splits. There exist a number of algorithms for carrying out the split decomposition. The most popular is implemented in the SPLITSTREE program by Huson (1998).

We recall some basic definitions related to the split decomposition and splitsgraphs. Let X be a set of taxa. A split $S = \{B, B'\}$ is defined as a partition of X into two nonempty sets B and B' such that $B \cup B' = X$. For instance, any branch in a phylogenetic tree introduces a split consisting of all the taxa found on one side (set B) and on the other (set B') of this branch. A set S of splits is called *weakly compatible* if, for any three splits S_1, S_2, and S_3 from S and all $B_i \in S_i$ (i = 1, 2 and 3), at least one of the four intersections:

$$B_1 \cap B_2 \cap B_3, B_1 \cap B'_2 \cap B'_3, B'_1 \cap B_2 \cap B'_3, \text{or } B'_1 \cap B'_2 \cap B_3$$

is empty (see Bandelt and Dress 1992a, b). A *splitsgraph* representing a weakly compatible split system S is a graph $G(S) = (V, E)$ whose vertices $v \in V$ are labeled by the set of taxa in X and whose edges (i.e. branches) $e \in E$ are straight line segments representing the splits in S (Fig. 19). In such a graph, each split $\{B, B'\}$ in S is depicted by a group of parallel branches of equal lengths, so that deleting all branches in such a group splits the graph into exactly two parts, one containing all vertices labeled by the taxa in B and the other containing all vertices labeled by the taxa in B'. This method requires an accurate estimation of pairwise distances. Any deviation from the optimal conditions leads to too many splits returned by the method.

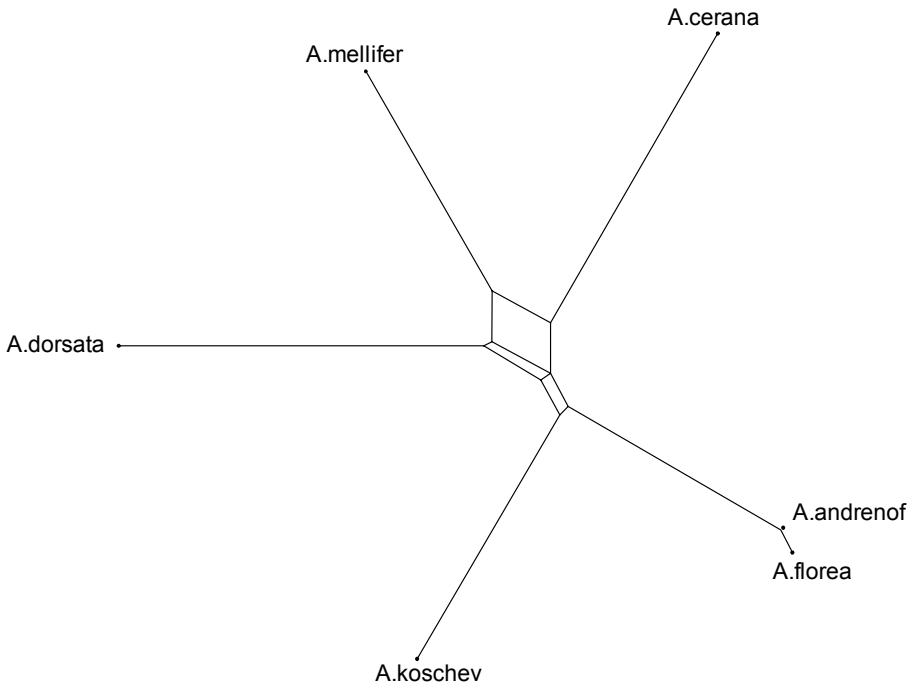

Fig. 19. SPLITSTREE network for the *Apis* honeybees.

The split decomposition method is fast, which means that a reasonable number of haplotypes can be analyzed. It can be applied to nucleotide or protein data. The program suuports the inclusion of models of the nucleotide substitution or amino acid replacement. The method is also suitable for bootstrap evaluation. Fig. 19 represents a splitsgraph built for the dataset of *Apis* honeybees using the LogDet (Steel 1994). evolutionary model selected to compute distances between species.

The SPLITSTREE package, which includes the split decomposition method, is available at: http://www-ab.informatik.uni-tuebingen.de/software/splits/welcome_en.html. The more recent SPLITSTREE 4.0 version includes also the NeighborNet method (Bryant and Moulton 2002, 2004). discussed in the next paragraph.

4.10. NeighborNet

NeighborNet (Bryant and Moulton 2002 and 2004). is a network construction and data representation method that combines the principles of the neighbor-joining and split decomposition techniques. Similarly to neighbor-joining, NeighborNet uses data agglomeration: taxa are combined into progressively larger and larger overlapping clusters.

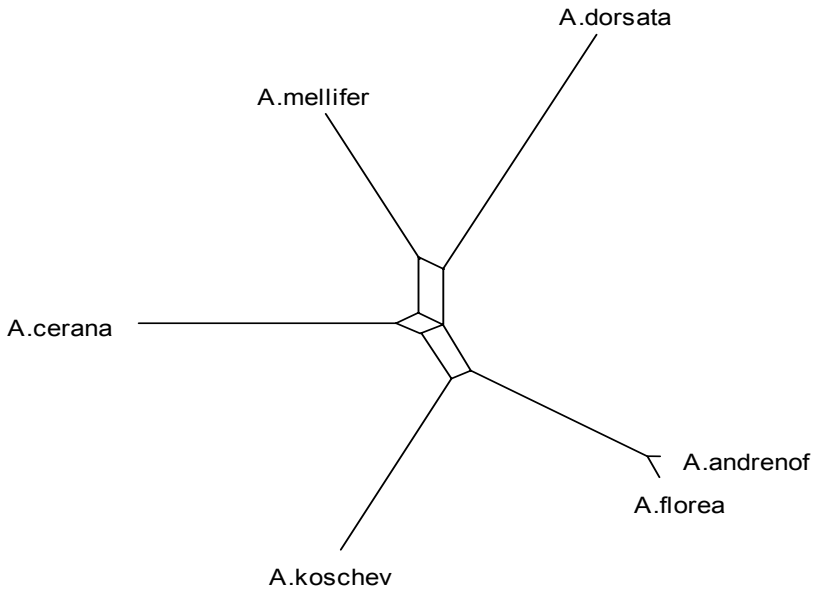

Fig. 20. NeighborNet network for the *Apis* honeybees, generated by SPLITSTREE 4.0.

This strategy has paid dividends in the tree building business with algorithms such as NJ (Saitou and Nei 1987). and BioNJ (Gascuel 1997a). The NeighborNet method can be used to represent multiple phylogenetic hypotheses simultaneously, or to detect complex evolutionary processes like recombination, lateral gene transfer or hybridization. NeighborNet tends to produce networks that are generally more resolved than those built by split decomposition. More precisely, NeighborNet generates a weighted circular split system rather than a hierarchy or a tree, which can

subsequently be represented by a planar splitsgraph; for more detail see Bryant and Moulton (2002, 2004). In such graphs, bipartitions or splits of the taxa are represented by classes of parallel lines; conflicting signals or incompatibilities appear as boxes. The method runs in $O(n^3)$ time, for n species, and is well suited for the preliminary analysis of large phylogenetic data sets and for carrying out intensive validation techniques such as bootstrapping. A NeighborNet network for the *Apis* honeybee data is shown in Fig. 20. The LogDet (Steel 1994). evolutionary model was selected to compute distances between species. The NEIGHBORNET package, created by D. Bryant, implementing the method for the Linux and MacOS X platforms is available at the following website: http://www.mcb.mcgill.ca/~bryant/NeighborNet/. As mentioned in the previous paragraph, this method is also available in the SPLITSTREE 4.0 package.

4.11. Pyramids

The Pyramids method was introduced by Diday and Bertrand (1986). Its theoretical description can also be found in Diday (1984 and 1986). The pyramidal clustering model generalizes hierarchies by allowing non-disjoint classes at a given level instead of partitions. The classical hierarchical methods reconstruct a set of the non-overlapping, nested clusters. In contrast to them, pyramids represent a set of clusters that may overlap, with no need for them to be nested. Pyramids can be useful for depicting reticulation events among species. The method infers a pyramid by an agglomerative bottom-up algorithm. It is based on the computation of a Robinsonian dissimilarity matrix between species under study (set X). This means that X admits an ordering such that for any triplet (i, j, k) the dissimilarity value d_{ik} must be larger than or equal to the maximum of d_{ij} and d_{jk}.

The software, running on the Sun, Linux and Unix platforms, carrying out the Pyramids method, is available at the following website: http://195.221.65.10:1234/Pyramids. Fig. 21 shows a pyramid constructed for the *Apis* honeybee data. It was generated using the on-line software available at: http://bioweb.pasteur.fr/seqanal/interfaces/pyramids.html.

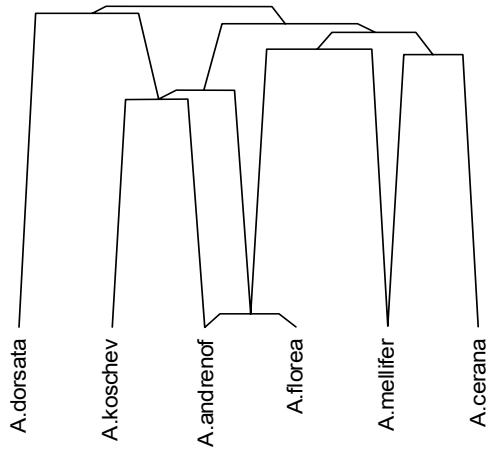

Fig. 21. Pyramid topology representing evolution of the *Apis* honeybees.

4.12. Weak Hierarchies

The method of Weak Hierarchies was introduced by Bandelt and Dress (1989). The method first uses the similarity matrix to infer a dendrogram (strong clusters), and then adds to it weak clusters representing supplementary inter-species relationships. Consequently, a weak hierarchy is an extension of dendrograms that includes both the weak and strong clusters. A subset C of the set X is regarded as a weak cluster if any two objects a, b in C are more similar to each other than any other object x from X-C is similar to either a or b.

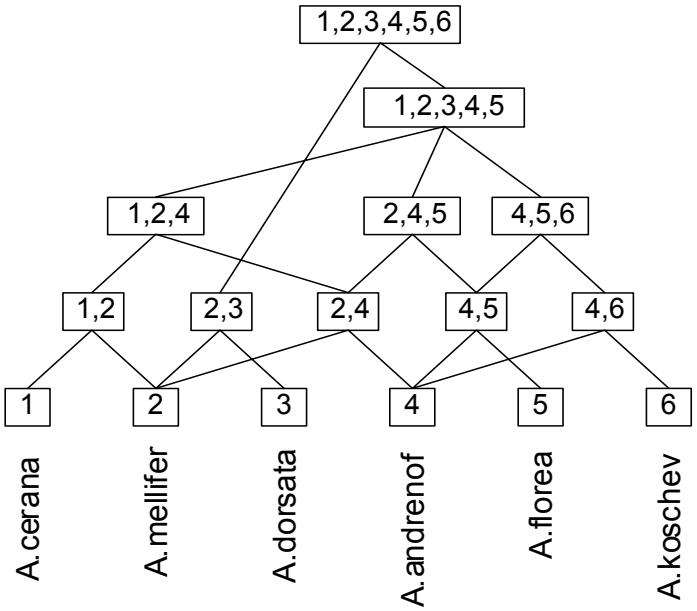

Fig. 22. Weak hierarchy representing the relationships among the *Apis* honeybees.

The mathematical definitions presented by Bandelt and Dress (1989). are as follows. Let S be a similarity function on a set X of objects. This function perfectly corresponds to a dendrogram if and only if it satisfies the ultrametric inequality (8):

$$S(a,b) \geq Min\{S(a,x), S(b,x)\}, \text{ for all } a, b, x \in X. \quad (8)$$

However, the ultrametric inequality is rarely satisfied for similarity measures encountered in reality. For an arbitrary similarity measure S, a subset C of the set X is called a strong cluster if it satisfies the inequality (9):

$$S(a,b) > Max\{S(a,x), S(b,x)\}, \text{ for all } a, b \in C \text{ and } x \in X\text{-}C. \quad (9)$$

If all objects in a subset C satisfy inequality (10), C is called a weak cluster:

$$S(a,b) > Min\{S(a,x), S(b,x)\}, \text{ for all } a, b \in C \text{ and } x \in X\text{-}C. \quad (10)$$

As pointed out by Bandelt and Dress (1989). potential applications of this method include fitting of dendrograms with few additional non-nested clusters and simultaneous representation of families of multiple dendrograms. Figure 21 shows a weak hierarchy for the *Apis* honeybee data also considered in the previous sections. Programs for computing weak hierarchies are available from either H-J. Bandelt (upon request) or V. Makarenkov (the C source code of the program is available at: http://www.info2.uqam.ca/~makarenv/software/Weak_Hierarchies.cpp).

5. CONCLUSION

Phylogenies can be estimated using distance-based, maximum parsimony, maximum likelihood, and Bayesian approaches. Methods and software for phylogenetic tree inferring have been developed since the seminal paper by Cavalli-Sforza and Edwards (1964). who described a tree reconstruction method for continuous characters. A standard format for representing phylogenies in computer-readable form, called the *Newick Standard*, was adopted by an informal committee convened during the Society for the Study of Evolution conference in Durham, New Hampshire, on June 26, 1986; see http://evolution.genetics.washington.edu/phylip/newicktree.html for more details. This format has enhanced the portability of results among computer packages and greatly facilitated the life and work of evolutionary biologists.

Patterns of reticulate evolution have been found in a variety of evolutionary contexts: lateral gene transfer, allopolyploidy, hybridization, as well as mechanisms operating at the micro-evolutionary level. These patterns can be modelled and analysed using methods of reticulate network reconstruction. Homoplasy can also be modelled using reticulate networks. Contrary to the tree inferring, the network building methods are still in their infancy. More refined methods need to be developed to address a variety of situations and research issues. Some of these issues have to be translated into mathematical and statistical form, requiring the help of mathematicians and statisticians. Development of new methods will involve collaboration between evolutionary biologists and computer scientists, as it has been the case for some of the presently available algorithms and models. The new and existing methods will have to be tested against carefully annotated benchmark data, representing different types of reticulate patterns, which should be made available to researchers in a remotely accessible repository. These methods should also be statistically validated and tested against simulated evolutionary data. The development of adequate simulation benchmarks should be discussed at length among evolutionary biologists. Software developers should also get together and develop a common format for the representation of reticulated networks, inspired by the *Newick* format mentioned in the previous paragraph. For the time being, many biologists conducting phylogenetic analysis still interpret their results in a conservative way, while the emerging field of reticulate evolution is trying to gain some level of confidence in the new methods.

REFERENCES

Addario-Berry L, Hallett M and Lagergren J (2003). Towards identifying lateral gene transfer events. Pac Symp Biocomput 8:279–290.

Anderson JB and Kohn LM (1998). Genotyping, gene genealogies and genomics bring fungal population genetics above ground. Trends Ecol Evol 13:444–449.

Atchley WR and Fitch WM (1991). Gene trees and the origins of inbred strains of mice. Science 254:554–558.

Atchley WR and Fitch WM (1993). Genetic affinities of inbred mouse strains of uncertain origin. Mol Biol Evol 10:1150–1169.

Atesson K (1999). The performance of Neighbor-Joining methods of phylogenetic reconstruction. Algorithmica 25: 251-278.

Aude JC, Diaz-Lazcoz Y, Codani JJ and Risler JL (1999). Application of the pyramidal clustering method to biological objects. Comput. Chem. 23:303-315.

Bandelt H-J and Dress AWM (1989). Weak hierarchies associated with similarity measures – an additive clustering technique. Bull Math Biol 51(1):133-166.

Bandelt H-J and Dress AWM (1992a). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol Phylogenet Evol 1:242–252.

Bandelt H-J and Dress AWM (1992b). A canonical decomposition theory for metrics on a finite set. Adv Math 92:47-65.

Bandelt H-J, Forster P, Sykes BC and Richards MB (1995). Mitochondrial portraits of human populations using median networks. Genetics 141:743–753.

Bandelt H-J, Forster P and Rohl A (1999). Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37-48.

Bandelt H-J, Macaulay V and Richards M (2000). Median networks:speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. Mol Phylogenet Evol 16:8–28.

Barthélemy J-P and Guénoche A (1991). Trees and Proximity Representations. Wiley, New York.

Baudry E, Solignac M, Garnery L, Gries M, Cornuet JM and Koeniger N (1998). Relatedness among honeybees *Apis mellifera* of a drone congregation. Proc R Soc Lond B 265:2009-2014.

Boc A and Makarenkov V (2003). New efficient algorithm for detection of horizontal gene transfer events. In: Algorithms in Bioinformatics, Springer, WABI 2003, pp 190-201.

Bryant D and Moulton V (2002). NeighborNet: an agglomerative method for the construction of planar phylogenetic networks. Algorithms in Bioinformatics: Second International Workshop, WABI 2002, Rome, Italy, September 17-21, pp 375 – 391.

Bryant D and Moulton V (2004). Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21:255-265.

Camin JH and Sokal RR (1965). A method for deducing branching sequences in phylogeny. Evolution 19: 311-326.

Cavalli-Sforza LL and Edwards AWF (1964). Analysis of human evolution. In: Genetics Today: Proc XI Int Congr Genet, pp 923-933.

Cheung B, Holmes RS, Easteal S and Beacham IR (1999). Evolution of class I alcohol dehydrogenase genes in catarrhine primates: gene conversion, substitution rates, and gene regulation. Mol Biol Evol 16:23-36.

Clement M, Posada D and Crandall KA (2000). TCS: a computer program to estimate gene genealogies. Mol Ecol 9:1657-1660.

Crandall KA (1995). Intraspecific phylogenetics: Support for dental transmission of human immunodeficiency virus. J Virol 69:2351-2356.

Dawley RM (1989). An introduction to unisexual vertebrates. In: RM Dawley and JP Bogart, eds. Evolution and Ecology of Unisexual Vertebrates. Albany, New York: New York State Museum, Bulletin 466, pp 1-18.

Delwiche CF and Palmer JD (1996). Rampant horizontal transfer and duplication of rubisco genes in Eubacteria and plastids. Mol Biol Evol 13:873-882.

De Soete G (1984). Additive-tree representations of incomplete dissimilarity data. Qual Quant 18:387-393.
Diday E (1984). Une representation des classes empiétantes : les pyramides. Research report INRIA 291.
Diday E (1986). Orders and overlapping clusters by pyramids. In: J.De Leeuw *et al*., ed. Multidimensional Data Analysis Proc., DSWO Press, Leiden.
Diday E and Bertrand P (1984). An extension of hierarchical clustering: the pyramidal representation. In: ES Gelsema and LN Kanal eds., Pattern Recognition in Practice, Amsterdam, North-Holland, pp 411-424.
Doolittle WF (1999). Phylogenetic classification and the universal tree. Science 284:2124-2128.
Edwards AWF (1972). Likelihood. Oxford Univ. Press, Oxford, UK, pp 252.
Excoffier L and Smouse PE (1994). Using allele frequencies and geographic subdivision to reconstruct gene trees within a species:molecular variance parsimony. Genetics 136:343–359.
Farris JS (1970). Methods for computing Wagner trees. Syst Zool 19:83-92.
Farris JS (1977) Phylogenetic analysis under Dollo's Law. Syst Zool 26:77-88.
Feil EJ, Holmes EC, Bessen DE, Chan M-S, Day NPJ, Enright MC, Goldstein R, Hood DW, Kalia A, Moore CE, Zhou J and Spratt BG (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci USA 98:182-187.
Felsenstein J (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368-376
Felsenstein J (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791.
Felsenstein J (1997). An alternating least-squares approach to inferring phylogenies from pairwise distances. Syst Zool 46:101-111.
Felsenstein J (2003). Inferring Phylogenies. Sinauer Assoc pp 664.
Felsenstein J. (2004). PHYLIP (http://evolution.genetics.washington.edu/phylip.html - software download page and software manual) - PHYLogeny Inference Package.
Fitch WM (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. Syst Zool 20:406-416.
Fitch WM (1997). Networks and viral evolution. J Mol Evol 44:65–75.
Fitch DHA, Mainone C, Goodman M and Sligh-Tom JL (1990). Molecular history of gene conversions in the primate fetal y-globin genes. J Biol Chem 265:781-793.
Foulds LR, Hendy MD and Penny D (1979). A graph theoretic approach to the development of minimal phylogenetic trees. J Mol Evol 13:127–149.
Gascuel O (1997a). BIONJ:an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14:685-695.
Gascuel O (1997b). Concerning the NJ algorithm and its unweighted version, UNJ. In: B Mirkin, F R McMorris, F Roberts and A Rzhetsky, eds. Mathematical hierarchies and Biology. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. Providence, RI: American Mathematical Society, pp 149–170.
Guénoche A and Leclerc B (2001). The triangles method to build X-trees from incomplete distance matrices. RAIRO Oper Res 35:283--300.
Guindon S and Gascuel O (2003). A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. Syst Biol 52:696-704.
Guttman DS and Dykhuizen DE (1994). Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science 266:1380-1383.
Hallet MT and Lagergren J (2001). Efficient algorithms for lateral gene transfer problems. In: Proceedings of the 5th Ann Int Conf Compt Mol Biol (RECOMB 01), New York, ASM Press. pp 149-156.
Hatta M, Fukami H, Wang W, Omori M, Shimoike K, Hayashibara T, Ina Y and Sugiyama T (1999). Reproductive and genetic evidence for a reticulate evolutionary history of mass-spawning corals. Mol Biol Evol 16:1607-1613.
Hayasaka K, Gojobori T and Horai S (1998). Molecular phylogeny and evolution of primate mitochondrial DNA. Mol Biol Evol 5:626-644.

Hein J (1993). A heuristic method to reconstruct the history of sequences subject to recombination. J Mol Evol 36:396-405.
Hillis DM (1996). Inferring complex phylogenies. Nature 383:130–131.
Huelsenbeck JP, Ronquist F, Nielsen R and Bollback JP (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.
Hugall A, Stanton J and Moritz C (1999). Reticulate evolution and the origins of ribosomal internal transcribed spacer diversity in apomictic meloidogyne. Mol Biol Evol 16:157-164.
Huelsenbeck JP and Ronquist FR (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinf 17:754–755.
Huson DH (1998). SplitsTree: a program for analyzing and visualizing evolutionary data. Bioinf 141:68-73.
Jukes TH and Cantor CR (1969). Evolution of protein molecules. In: H. N. Munro, eds. Mammalian Protein Metabolism, Academic Press, New York, pp 21-132.
Kim JH (1996). General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Syst Biol 45:363–374.
Kim J and Warnow T (1999). Tutorial on phylogenetic tree estimation. In: Proc. 7th Int'l Conf. on Intelligent Systems for Molecular Biology (ISMB99).
Kimura M (1981). Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454-458.
Koeniger G, Koeniger N, Mardan M and Wongsiri S (1993). Variance in weight of sexuals and workers within and between 4 Apis species (*A. florea, Apis dorsata, Apis cerana* and *Apis mellifera*). Asian Apicult 1:106-111.
Landry PA and Lapointe FJ (1997). Estimation of missing distances in path-length matrices: problems and solutions. In: B Mirkin, FR McMorris, F Roberts, A Rzhetsky eds., Mathematical hierarchies and Biology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Amer Math Soc, Providence, RI, pp 209-224.
Lapointe F-J (2000). How to account for reticulation events in phylogenetic analysis: a comparison of distance-based methods. J Classif 17:175-184.
Larget B and Simon DL (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol 16:750–759.
Legendre P (Guest Editor) (2000a). Special section on reticulate evolution. J Classif 17:153–195.
Legendre P (2000b). Biological applications of reticulation analysis. J Classif 17:191–195.
Legendre P and Makarenkov V (2002). Reconstruction of biogeographic and evolutionary networks using reticulograms. Syst Biol 51:199–216.
Li W-H (1997). Molecular Evolution. Sunderland, Massachusetts: Sinauer Assoc, pp 487.
Li S, Pearl DK and Doss H (2000). Phylogenetic tree construction using Markov chain Monte Carlo. J Am Stat Assoc 95:493–508.
Linder CR, Moret BME, Nakhleh L and Warnow T (2003). Network (reticulate) evolution: biology, models, and algorithms. A tutorial presented at the Ninth Pacific Symposium on Biocomputing (PSB 2004).
Linder CR, Moret BME, Nakhleh L and Warnow T (2004). Reconstructing networks part II: computational aspects. A tutorial presented at the Ninth Pacific Symposium on Biocomputing (PSB 2004).
Makarenkov V and Leclerc B (1997). Tree metrics and their circular orders:some uses for the reconstruction and fitting of phylogenetic trees. In: B Mirkin, F R McMorris, F Roberts and A Rzhetsky, eds. Mathematical hierarchies and Biology. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. Providence, RI: American Mathematical Society, pp 183–208.
Makarenkov V and Leclerc B (1999). An algorithm for the fitting of a tree metric according to a weighted least-squares criterion. J Classif 16:3–26.
Makarenkov V and Leclerc B (2000). Comparison of additive trees using circular orders. J Comput Biol 7:731–744.

Makarenkov V and Legendre P (2000). Improving the additive tree representation of a dissimilarity matrix using reticulations. In: HAL Kiers, J-P Rasson, PJF Groenen and M Schader, eds. Data Analysis Classification and Related Methods. Berlin: Springer, pp 35–40.
Makarenkov V (2001). T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks, Bioinf 17:664–668.
Makarenkov V and Legendre P (2004). From a phylogenetic tree to a reticulated network. J Comput Biol 11:195–212.
Makarenkov V, Legendre P and Desdevises Y (2004). Modeling phylogenetic relationships using reticulated networks. Zool Scrip 33:89–96.
Makarenkov V, Boc A and Diallo AB (2004). Representing lateral gene transfer in species classification. Unique scenario. In: Classification, Clustering, and Data Mining Applications, IFCS 2004, Chicago: Springer, pp 439-446.
Makarenkov V and Lapointe F-J (2004). A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. Bioinformatics 20:2113-2121.
Makarenkov V, Boc A, Delwiche CF and Philippe H (2006). A new efficient method for detecting horizontal gene transfers: Modeling partial and complete gene transfer scenarios, submitted.
Marais G, Mouchiroud D and Duret L (2001). Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci USA 98:5688–5692.
Mau B, Newton MA and Larget B (1997). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Mol Biol Evol 14:717-724.
McDade L (1995) Hybridization and phylogenetics. In PC Hoch and AG Stephenson, eds., Experimental and Molecular Approaches to Plant Biosystematics, Monographs in Systematic Botany from the Missouri Botanical Garden. pp 305–331.
Milner A (1996). An introduction to understanding honeybees, their origins, evolution and diversity. Available via Bibba electronic journal, URL:<http://www.bibba.com>.
Nei M and Kumar S (2000). Molecular Evolution and Phylogenetics. Oxford Univ. Press, New York, pp 333.
Nesbø CL, L'Haridon S, Stetter KO and Doolittle WF (2001). Phylogenetic analyses of two "archaeal" genes in *Thermotoga maritima* reveal multiple transfers between archaea and bacteria. Mol Biol Evol 18:362-375.
Odorico DM and Miller DJ (1997). Variation in the ribosomal internal transcribed spacers and 5.8s rDNA among five species of *Acropora* (*cnidaria*; *scleractinia*): Patterns of variation consistent with reticulate evolution. Mol Biol Evol, 14:465-473.
Posada D and Crandall KA (1998). Modeltest: testing the model of DNA substitution. Bioinf 14, 817-818.
Posada D and Crandall KA (2001a). Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc Natl Acad Sci USA 98(24):13757-13762.
Posada D and Crandall KA (2001b). Intraspecific gene genealogies: trees grafting into networks. Trends Ecol Evol 16 (1):37-45.
Rannala B and Yang Z (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol 43:304–311.
Rieseberg LH and Ellstrand NC (1993). What can morphological and molecular markers tell us about plant hybridization? Crit Rev Plant Sci 12:213-241.
Rieseberg LH and Morefield JD (1995). Character expression, phylogenetic reconstruction, and the detection of reticulate evolution. In: PC Hoch and AG Stephenson, eds., Experimental and Molecular Approaches to Plant Biosystematics. Monographs in Systematic Botany from the Missouri Botanical Garden 53, pp 333-354.
Robertson DL, Hanh BH and Sharp PM (1995). Recombination in AIDS viruses. J Mol Evol 40:249-259.
Robinson DR and Foulds LR (1981). Comparison of phylogenetic trees. Math Biosci 53:131-147.
Rohlf FJ (1963). Classification of *Aedes* by numerical taxonomic methods (Diptera: Culicidae). Ann Entomol Soc Am 56:798–804.
Rohlf FJ (2000). Phylogenetic models and reticulations. J Classif 17(2):185-189.

Saitou N and Nei M (1987). The neighbor-joining method:a new method for reconstructing phylogenetic trees. Mol Biol Evol 4, 406–425.
Sattath S and Tversky A (1977). Phylogenetic similarity trees. Psychometrika 42:319-345.
Sawyer S (1989). Statistical tests for detecting gene conversion. Mol Biol Evol 6:526-536.
Schmidt HA and von Haeseler A (2003). Maximum-Likelihood Analysis Using TREE-PUZZLE. In A.D. Baxevanis, D.B. Davison, R.D.M. Page, G. Stormo, and L. Stein (eds.) Current Protocols in Bioinformatics, Unit 6.6, Wiley and Sons, New York.
Smouse PE (2000). Reticulation inside the species boundary. J Classif 17:165-173.
Sneath PHA, Sackin MJ and Ambler RP (1975). Detecting evolutionary incompatibilities from protein sequences. Syst Zool 24:311-332.
Sneath PHA (2000). Reticulate evolution in bacteria and other organisms: how can we study it? J Classif 17:159-163.
Sonea S and Mathieu LG (2000). Prokaryotology – A coherent view. Presses de l'Université de Montréal, Montréal.
Sonea S and Panisset M (1976). Pour une nouvelle bactériologie. Rev Can Biol 35:103-167.
Sonea S and Panisset M (1981). Introduction à la nouvelle bactériologie. Presses de l'Université de Montréal, Montréal and Masson, Paris, pp 127.
Stace CA (1984). Plant taxonomy and biosystematics. Edward Arnold, London, pp 272.
Steel MA (1994). Recovering a tree from the leaf colorations it generates under a Markov model. Appl Math Lett 72:19-24.
Stephens JC (1985). Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol Biol Evol 2:539-556.
Studier JA and Keppler KJ (1988). A note on the neighbor-joining algorithm of Saitou and Nei. Mol Biol Evol 5:729–731.
Swofford DL, Olsen GL, Waddell PJ and Hillis MD (1996). Phylogenetic Inference. In: D. M. Hill ed. Molecular Systematics. Sinauer, pp 407-514.
Swofford DL (2001). PAUP: Phylogenetic analysis using parsimony and other methods. Version 4.0d8. Champaign, Illinois: Illinois Natural History Survey.
Tajima F and Nei M (1984). Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 1:269.
Templeton AR, Crandall KA and Sing CF (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics 132:619-633.
Walter SJ, Campbell CS, Kellogg EA and Stevens PF (1999). Plant systematics. A phylogenetic approach. Sinauer Associates. Inc. Sunderland, Massachusetts, USA, pp 576.
Whelan S Lio P and Goldman N (2001). Molecular phylogenetics:state-of-the-art methods for looking into the past. Trends Genet 17:262–272.
Xia X and Xie Z (2001). DAMBE: Data analysis in molecular biology and evolution. Journal of Heredity 92:371-373.
Yang ZH and Rannala B (1997). Bayesian phylogenetic inference using DNA sequences:a Markov chain Monte Carlo method. Mol Biol Evol 14:717–724.
Yushmanov SV (1984). Construction of a tree with p leaves from 2p-3 elements of its distance matrix (in Russian). Matematicheskie Zametki 35:877-887.