Modelling phylogenetic relationships using reticulated networks

Vladimir Makarenkov, Pierre Legendre & Yves Desdevises

Accepted: 8 May 2003

Makarenkov, V., Legendre, P. & Desdevises, Y. (2004). Modelling phylogenetic relationships using reticulated networks. — *Zoologica Scripta*, *33*, 89–96.

Most traditional methods of phylogenetic analysis assume that species evolution can be represented by means of a bifurcating tree model. In many phylogenetic situations, however, some of the evolutionary links between species are due to reticulate evolution. For instance, reticulate models can adequately describe such complicated mechanisms as lateral gene transfer in bacteria or species hybridization. The theoretical concepts of reticulate evolution developed in the 1980s and 1990s need to be supported by appropriate analytical tools and software. In this paper, we present the main features of a new distance-based method for modelling phylogenetic relationships among species by means of reticulated networks (RNs). The method uses the least-squares model to build a RN by gradually improving upon the solution provided by a phylogenetic tree. A computer program facilitating the reconstruction and visualization of reticulate phylogenies is made available to researchers. In the application section, we illustrate the usefulness of the method by studying the evolution of honeybees (genus *Apis*). The method for reconstructing RNs has been included in the *T-Rex* (*Tree and Reticulogram Reconstruction*) package recently developed by the first-named author.

Vladimir Makarenkov. Département d'informatique, Université du Québec à Montréal, C.P. 8888, succursale Centre-Ville, Montréal (Québec), Canada, H3C 3P8. E-mail: makarenkov.vladimir@uqam.ca

Pierre Legendre, Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-Ville, Montréal (Québec), Canada, H3C 3J7. E-mail: pierre.legendre@umontreal.ca

Yves Desdevises, Laboratoire Arago, Université Pierre et Marie Curie, UMR CNRS 7628, BP 44, 66651 Banyuls-sur-Mer Cedex, France. E-mail: desdevises@obs-banyuls.fr

Introduction

Patterns of reticulate evolution have been found in a variety of evolutionary contexts, giving rise to a number of recent studies. For example, the phylogeny of 24 inbred strains of mice obtained by Atchley & Fitch (1991, 1993) includes several with hybrid origins. Examples of molecular data sets containing regions with reticulate histories can be found in Fitch *et al*. (1990) (multigene families), Robertson *et al*. (1995) (virus strains), and Guttman & Dykhuizen (1994) (bacterial genes). 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 & Miller (1997) discovered patterns of variation due to reticulate evolution in the ribosomal internal transcribed spacers and 5.8 s 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 *et al*. (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.

Following Sonea & Panisset (1976, 1981), who showed that LGT was a common evolutionary mechanism among bacteria, Doolittle (1999) emphasized the importance of LGT, which is a reticulate process, in the evolution of bacteria and higher groups of organisms. The fact that most archaeal and bacterial genomes contain genes from multiple sources is challenging for molecular biologists. According to Doolittle (1999), molecular phylogeneticists 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 properly be represented as a tree.

Another reticulate process, hybridization, prevails in plants. According to one estimate (Stace 1984), there are

about 70 000 naturally occurring interspecies plant hybrids in the world. 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 (Walter *et al*. 1999).

For centuries, traditional breeders have genetically modified plants and animals by crossing organisms with similar genetic makeup, transferring tens of thousands of genes at a time. Today, scientists can engineer transgenic crops and livestock by introducing one or more genes from a species that may not be closely related. It would be interesting to analyse the phylogenetic relationships of a group of genetically modified organisms; clearly, a reticulated network rather than a phylogenetic tree would be the appropriate model in this instance.

Reconstruction of reticulate evolution has long been difficult. Several methods have been proposed for uncovering it in nucleotide sequences. Existing studies have focused on 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 for recombination (Hein 1993). The popular method of split decomposition enables the representation of data in the form of a splits graph revealing conflicting signals contained in the data (Bandelt & Dress 1992a, b). In a splits graph, a pair of nodes may be linked by a set of parallel branches depicting alternative solutions.

In this paper, we describe an algorithm for modelling reticulate phylogenetic relationships among species by means of reticulated networks (RNs). Applications of this algorithm to problems of ecological biogeography (freshwater fishes), microgeographical morphological differentiation within a species (muskrats), and the study of plant hybrids (*Aphelandra*) have been published in Legendre & Makarenkov (2002). This method can be used to detect incompatibilities in phylogenetic trees; it may also indicate which species (or ancestors) have more similarities (e.g. genes in common) than might be depicted by the phylogenetic tree model. Finally, RNs can indicate the presence, or the absence, of possible reticulate events in the phylogenetic history of the group.

Materials and Methods

Description of the algorithm

In this section we describe an algorithm for inferring a connected and undirected (when no directions are given to the branches) RN from a distance matrix; this algorithm gives the solution in polynomial time.

We used the following approach to build the network from a matrix of evolutionary distances among observed taxa: first, a phylogenetic tree is inferred from a distance matrix using one of the existing tree fitting algorithms; second, some extra branches, called *reticulation branches* (RBs), are added to the tree structure while optimizing a loss function which can be constructed with respect to either least-squares, or parsimony, or maximum likelihood criteria. In this study, we focus on network reconstruction with reference to the *least-squares criterion*. The addition of RBs stops when the minimum of a goodness-of-fit function (equation 4) is reached. This function takes into account the value of the least-squares criterion as well as the total number of branches of the reticulated network under construction. Because, in our study, the reconstruction technique is based on least squares, it is reasonable to consider as the starting solution a phylogenetic tree whose branch lengths have been also fitted to the given distances by least squares. For an overview of least-squares fitting techniques, see Barthélémy & Guénoche (1991), Bryant & Waddell (1998), or Makarenkov & Leclerc (1999).

A RN can be viewed as a weighted graph where some nodes are labelled by the names of the species (e.g. taxa); all other nodes of the network are intermediate: they represent unknown ancestors. The *minimum path-length distance* between pairs of nodes representing the observed species is called a *reticulation distance*. In a general weighted graph, several paths may exist linking a pair of nodes, whereas in a phylogenetic tree there exists a unique path linking any two nodes. Let d be a distance matrix on the set X of n taxa and *dist* an *additive distance* (i.e. a matrix of pairwise distances among taxa in a phylogenetic tree) inferred from d using an appropriate tree fitting algorithm. Note that any given phylogenetic tree can be transformed into a *binary tree*, whose internal nodes are all of degree 3, by adding branches of zero length. When this is done, a tree with n leaves has $n - 2$ internal nodes and $2n - 3$ branches. In this study, we consider binary phylogenetic trees as the foundation for the RN reconstruction algorithm. Thus, the RNs considered here always comprise $n - 2$ intermediate nodes in addition to the taxa in X; this makes the comparison of solutions provided by RNs and phylogenetic trees possible.

We now explore how to place the first RB into a tree. To add a new branch to a phylogenetic tree, we try out all possible pairs of nodes that are not already linked by a branch and select the one that most reduces the value of the least-squares function. Let us consider a binary phylogenetic tree T inferred from a distance matrix d and a pair of nodes x and y in T that are not linked by a branch (Fig. 1A). We look for an optimal value l of the least-squares loss function that will identify the new branch xy to be added to the tree T, while keeping fixed the lengths of all pre-existing branches (Fig. 1B).

We now describe in more detail how to determine the optimum value of the length of the first RB. First, we define the set $A(xy)$ of all pairs of taxa ij whose distances might change if a new RB connecting x and y was added to T. Specifically,

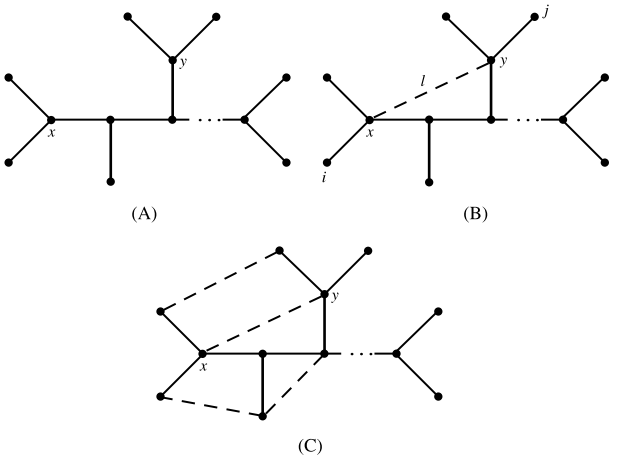

Fig. 1 Main steps of the algorithm for inferring reticulated networks: —A. Binary phylogenetic tree T is considered. —B. New branch of length l can be added to T to link nodes x and y. —C. Reticulate phylogeny inferred from T by addition of reticulation branches.

$A(xy)$ is the set of all taxon pairs ij such that:

$$Min\{dist(ix) + dist(jy);\ dist(jx) + dist(iy)\} < dist(ij) \quad (1)$$

where $dist(ij)$ is the minimum path distance between nodes i and j. Second, we define the following function

$$rho(ij) = dist(ij) - Min\{dist(ix) + dist(jy);\ dist(jx) + dist(iy)\} \quad (2)$$

The function to be minimized is the following:

$$Q(xy, l) = \sum_{rho(ij)>l} (Min\{dist(ix) + dist(jy);\ dist(jx) + dist(iy)\} + l - d(ij))^2 + \sum_{rho(ij)\le l} (dist(ij) - d(ij))^2 \rightarrow \min. \quad (3)$$

This function is $\sum_{i\in X}\sum_{j\in X}(dist\_new(ij, l) - d(ij))^2 \cdot dist\_new(ij, l)$ is the minimum path-length distance between the leaves i and j in T if the new branch xy with length l is added.

The function $Q(xy, l)$ defined by equation 3 is a quadratic polynomial spline. It has to be optimized on a number of intervals (l_k, l_{k+1}) defined by the distinct values of $rho(ij)$ for pairs ij in $A(xy)$. For a fixed interval (l_k, l_{k+1}) the function $Q(xy, l)$ is a quadratic polynomial; this makes its minimum value easy to find for each fixed pair of nodes xy. To obtain the optimum value of Q over the set of all possible new branches, these computations should be repeated for all pairs of tree nodes that are not linked by a branch. When all unlinked pairs of nodes are tested, only the best one, which is the one providing the global minimum of Q, will be linked by a new branch. When the first RB has been added to the network, the best second, the best third, and following RBs may be placed into it in the same way (Fig. 1C). This algorithm takes $O(kn^4)$ time for n taxa and k new RBs, since there are $O(n^2)$ taxon pairs ij for each pair of unlinked nodes xy and $O(n^2)$ node pairs xy.

Stopping rule for adding reticulation branches

A RN comprises more branches, and thus utilizes more estimated link-length 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 cost criterion should be introduced to estimate how many RBs have to be added to a network. We propose to use a goodness-of-fit criterion that takes into account the least-squares objective function as well as the number of degrees of freedom of the RN. When the exact number of RBs is unknown in advance, as it is often the case in evolutionary problems, one can stop the addition of new branches when the minimum of the criterion is reached.

The total number of nodes in a binary phylogenetic tree with n leaves is $2n - 2$. Therefore, the maximum number of branches one might place into a RN, constructed by adding RBs to a phylogenetic tree with n leaves, is $(2n - 2)(2n - 3)/2$. However, any metric distance can be represented by a complete graph with $n(n - 1)/2$ branches. Therefore, the latter limit can be considered as the maximum possible number of branches in a RN. Thus, the number of degrees of freedom of a RN with N branches is $n(n - 1)/2 - N$. It is reasonable to consider a penalty function opposing the loss in degrees of freedom to the gain in fit. The numerator of this function is the sum of quadratic differences between the values of the distance d and the corresponding reticulation estimates $dist$:

$$Q_2 = \frac{\sum_{i\in X}\sum_{j\in X}(dist(ij) - d(ij))^2}{n(n-1)/2 - N} = \frac{Q}{n(n-1)/2 - N} \quad (4)$$

Interestingly, as confirmed by a simulation study reported in Legendre & Makarenkov (2002), the function Q_2 usually has only one minimum over the interval $[2n - 3,\ n(n - 1)/2]$ of values of N. This minimum can be used as a stopping rule for addition of new branches to the reticulate phylogeny.

Results

Study of honeybee evolution using RNs

Honeybees (subfamily Apinae) belong to the family of social bees. The subfamily includes a single genus, *Apis*, which is characterized by the building of vertical combs of hexagonal cells constructed bilaterally from a midrib, using only wax secreted by the worker bees (see Milner 1996; Baudry *et al.* 1998). *Apis* has been able to colonize a wide variety of environments ranging from tropical to cool temperate. While the

species of most genera were indigenous to all continents, bees belonging to *Apis* were originally found only in Asia, Africa, and Europe, suggesting that the genus appeared much later. It includes four main species: *A. florea*, *A. dorsata*, *A. cerana* and *A. mellifera* (little, giant, eastern and western honeybee, respectively). The lifestyle of *A. cerana* is similar to that of *A. mellifera* and both are used in apiculture with modern moveable comb hives (Milner 1996). However, the numerical strength of *A. cerana* colonies is usually much less, and honey yields are smaller, contributing to their rapid replacement by imported *A. mellifera* races.

It is believed that bees originally evolved from hunting wasps that acquired a taste for nectar and became vegetarians. Fossil evidence is sparse but bees probably appeared at about the same time as the flowering plants, during the Cretaceous period, 146–74 Mya. Fossils of the true *Apis* type were first discovered from the Lower Miocene (22–25 Mya) of western Germany. A bee resembling *A. dorsata*, but much smaller, is thought to be present in the Upper Miocene (ca.12 Mya). *A. florea* and *A. dorsata* may have existed as separate species as early as the Oligocene. It remains to be discovered when bees of the *A. mellifera/A. cerana* type first appeared. It is thought that they must have acquired separate identities during the latter part of the Tertiary. The two species were probably physically separated at the time of the Pleistocene glaciations (1 million to about 10 000 years ago); there was no subsequent contact between them until that imposed by human intervention in recent times. In the postglacial period, *A. mellifera* and *A. cerana* (and to a less extent *A. dorsata* and *A. florea*) have shown similar evolution into geographical subspecies, or races (Koeniger *et al.* 1993; Milner 1996).

The ultimate western boundary of the *A. cerana* territory was in Afghanistan, some 600 km to the east of the nearest *A. mellifera* colonies in Iran. It is not possible to cross *A. cerana* with *A. mellifera* even using instrumental insemination, because the two species are now genetically incompatible, and viable eggs do not result from cross-fertilization (Milner 1996). Other differences include their reactions to diseases, infestations, and predators.

The most urgent problem in apiculture throughout the world is that of protecting *A. mellifera* against the varroa mite which threatens to exterminate it (Milner 1996). The ultimate hope is that varroa-resistant strains of bees may evolve. There is a danger that the development of resistance among apiary stocks might be concealed by the normal antivarroa treatments and that a resistant strain might be lost through the death of the queens. If, as has been hypothesized, the separation of the *A. cerana* and *A. mellifera* species occurred in relatively recent times, the gene which enabled *A. cerana* to develop a defence against varroa may still be among the genes of the *A. mellifera* races. This is why a comprehensive study of honeybee evolution is a matter of great importance.

Table 1 Original distance matrix between six species of honeybees (genus *Apis*). The pairwise distances among species were obtained by means of the Hamming distance computed over DNA sequences (677 bases).

A. andreniformis	0					
A. mellifera	0.090	0				
A. dorsata	0.103	0.093	0			
A. cerana	0.096	0.090	0.117	0		
A. florea	0.004	0.093	0.106	0.099	0	
A. koschevnikovi	0.075	0.100	0.103	0.099	0.078	0

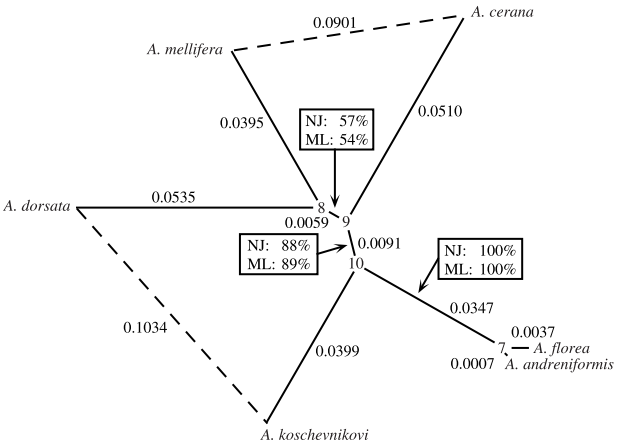

Fig. 2 Reticulate phylogeny representing the evolution of honeybees (genus *Apis*). It was constructed by adding two reticulation branches (dashed lines) to a phylogenetic tree (solid lines) inferred from distance data in Table 1 using the neighbour-joining fitting algorithm. Boxes: bootstrap support values for the clades (NJ: neighbour-joining; ML: maximum likelihood). Decimal numbers: branch lengths.

We applied the new method for detection of reticulate evolution to the DNA data of six species of honeybees. (We have not reported here the DNA sequences that we used, each of which comprised 677 characters; readers are referred to the popular SplitsTree package by Huson (1998) which includes an example with the complete set of bee DNA sequences.) A distance matrix (Table 1) for *A. andreniformis*, *A. mellifera*, *A. dorsata*, *A. cerana*, *A. florea* and *A. koschevnikovi* was obtained by computing Hamming distances (proportion of mismatches, called 'uncorrected-p' in PAUP) among the sequences. Since this was the distance matrix provided by Huson (1998) for the bee data, it is the one that we used.

The bee phylogenetic tree was reconstructed using a distance method, neighbour-joining (NJ; Fig. 2, full lines), and by maximum likelihood (ML, which produced the same tree topology as NJ). The obtained trees were validated by bootstrapping (Felsenstein 1985) using 100 replicates for ML, and 1000 replicates for NJ. All computations were performed with PAUP*

4.0d8 (Swofford 2001). For the ML analysis, we used the evolutionary model selected by Modeltest (Posada & Crandall 1998) via a hierarchical likelihood ratio test. The selected model was Kimura 81 with unequal base frequencies (Kimura 1981), taking into account substitution rate heterogeneity using a γ distribution with the a parameter equal to 0.16. The bootstrap support values for the clades are shown in Fig. 2. Using the same model to correct the distances in the NJ analysis as in ML gave approximately the same bootstrap support values.

The phylogeny clearly separated two groups of bees, with *A. mellifera*, *A. dorsata*, and *A. cerana* forming the first group and *A. andreniformis*, *A. florea* and *A. koschevnikovi* the second; the bootstrap support for the separation branch is 88% for NJ and 89% for ML. The branch lengths of the phylogenetic tree were then optimally adjusted to the distances using least-squares (see Bryant & Wadell 1998 or Makarenkov & Leclerc 1999). The values of the least-squares criterion Q and the goodness-of-fit criterion Q_2 obtained for the phylogenetic tree inferred by NJ were 0.000143 and 0.000024, respectively. The new method for detecting reticulate evolution was then used with the phylogenetic tree provided by NJ. Q_2 was chosen as the stopping rule for addition of new branches. Two new RBs (dashed lines in Fig. 2) were added to the phylogenetic tree by our algorithm. The minimum of Q_2 was reached at the second step of the algorithm, decreasing its value to 0.000020, whereas the value of Q dropped to 0.000078. The decrease of Q after addition of only two RBs was dramatic for these data. The gain in fit was 27.3% (Q = 0.000104) after addition of the first branch, linking *A. mellifera* and *A. cerana*, and the total gain was 45.5% (Q = 0.000078) after addition of the second, linking *A. dorsata* and A. *koschevnikovi*. These results demonstrate the relevance of the reticulation model to the data, where RBs bring to light conflicting features that are embedded in the phylogenetic tree. The poor bootstrap support (57% or 54%) obtained for the branch linking nodes 8 and 9 of the tree, before the reticulations were added, is consistent with a close relationship between *A. mellifera* and *A. cerana*.

When interpreting the RBs, variation in length is of great importance. If a RB is short with respect to the tree branches, this may be interpreted as a possible hybridization event occurring late during evolution. If very long, it may represent homoplasy (information representing convergent evolution, i.e. parallel evolution and evolutionary reversal). Obviously, one cannot illustrate these phenomena using a classical phylogenetic tree topology. For instance, the first RB linking species *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. It shows that the two species are genetically closer to each other than suggested by the phylogenetic tree. Fig. 3 depicts

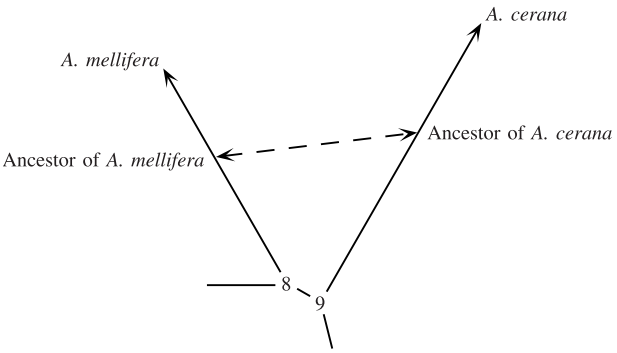

Fig. 3 Upper part of the reticulate phylogeny from Fig. 2 focusing on the evolution of *A. mellifera* and *A. cerana*. The branch depicted by a dashed line represents a possible hybridization event involving ancestors of the two species.

what may have happened: a recent ancestor of *A. cerana* may have hybridized with one of the recent ancestors of *A. mellifera* to produce the new *A. mellifera* bee; conversely, a recent ancestor of *A. mellifera* may have hybridized with one of the recent ancestors of *A. cerana* to produce the new *A. cerana* bee. This hypothesis agrees with the statement that *A. mellifera* and *A. cerana* must have shared a close common ancestor in relatively recent times. The other RB linking *A. dorsata* and *A. koschevnikovi* also reveals that the relationship between these two species is closer than that depicted by the phylogenetic tree.

Discussion

We have developed a new algorithm to infer reticulate phylogenies from evolutionary distances among observed species. It reconstructs a reticulated network (RN) by adding supplementary branches to a phylogenetic tree. Any new branch added to a phylogenetic tree represents *unresolved conflicting information* contained in it. Two species or clusters that are linked by a reticulated branch (RB) are more closely related to one another than is shown by the phylogenetic tree model. The main challenge consists in giving plausible explanations for each of the extra relations represented by RBs. These new branches should be interpreted differently under different evolutionary circumstances. First, we suggest that long RBs linking nodes located far away from one another in the phylogenetic tree reveal incompatibilities of a tree structure with respect to the observed evolutionary distances. Two explanations are possible: first, the phylogenetic tree does not provide a good representation of the evolutionary distances; second, long RBs may represent homoplasy among the observed species. On the other hand, short RBs may reflect either hybridization events that occurred between related species or their ancestors, or allopolyploidy if plant genetic distances are considered. The case of lateral gene transfer

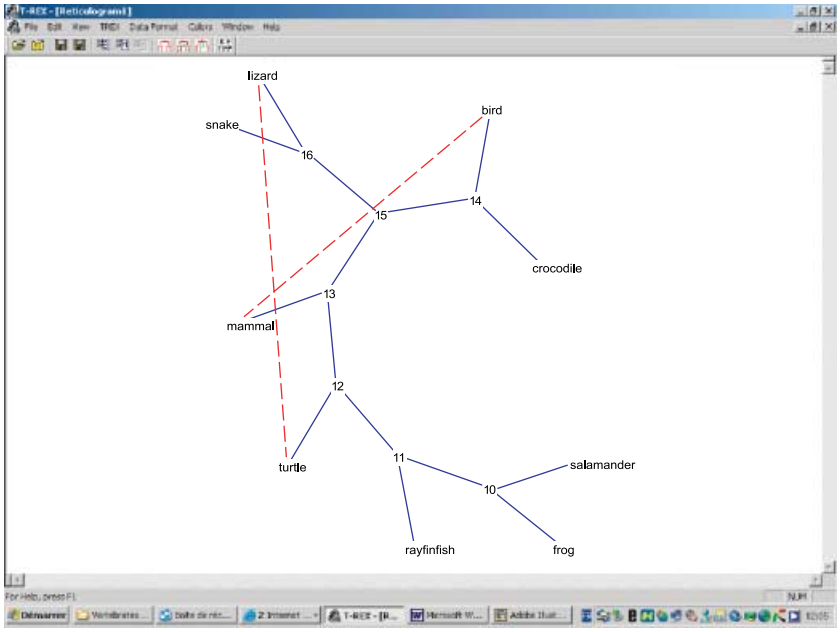

Fig. 4 *T-Rex* screenshot showing a reticulated network representing phylogenetic relationships among vertebrate organisms (for details on the vertebrate morphological dataset, see Maddison & Maddison 2000). First, a phylogenetic tree (solid lines) were inferred from a distance matrix among vertebrates using NJ (Saitou & Nei 1987); then, two reticulation branches (dashed lines) were added to the tree using the method discussed in this paper. The reticulation branches linking 'lizard' and 'turtle', and 'mammal' and 'bird' show that they are more closely related than is shown in the phylogenetic tree.

(LGT) seems to be the most complicated because RBs depicting gene exchange may be of any length. In this situation, investigation of the characters causing a reticulation might assist in the interpretation: if the characters responsible are contiguous in the nucleic acid sequence, LGT might be indicated.

Recommendations to researchers who have access to sets of molecular sequences for a clade of species and want to test the data for the presence of reticulate evolution are as follows: (1) compute a matrix of evolutionary distances among the species using an appropriate sequence-distance transformation, e.g. Hamming, Kimura 3ST (Kimura 1981), Jukes Cantor (Jukes & Cantor 1969), or LogDet (Steel 1994); (2) infer a phylogenetic tree from the matrix using a tree-fitting algorithm; (3) launch the new algorithm for reconstruction of reticulate phylogenies using goodness-of-fit criteria as a stopping rule for addition of RBs. Interpretation of the latter should be based on available biological or evolutionary knowledge.

The algorithm for reconstructing reticulate phylogenies described in this paper has been included in the *T-Rex* (tree and reticulogram reconstruction) package (Makarenkov 2001). Developed for both Macintosh and Windows platforms (Fig. 4 shows a screenshot of the Windows version), it is freely available for researchers at the following URL: http://www.fas.umontreal.ca/biol/casgrain/en/labo/t-rex. The package also includes some popular phylogenetic tree-fitting algorithms: ADDTREE by Sattath & Tversky (1977); neighbour-joining (NJ) by Saitou & Nei (1987); unweighted neighbour-joining (UNJ) by Gascuel *et al.* (1997); circular order reconstruction by Makarenkov *et al.* (1997) and Makarenkov & Leclerc (2000); the method of weighted least-squares (MW) by Makarenkov & Leclerc (1999), and others. *T-Rex* allows users to infer and visualize reticulate phylogenies by adding extra branches to phylogenetic trees obtained by the above-mentioned tree-fitting algorithms. The algorithm for reticulogram reconstruction implemented in the program analyses data sets in polynomial time, like most other methods of phylogenetic reconstruction. The honeybee data set used as example in this paper is very small (comprising only six species); the data sets analysed by Legendre & Makarenkov (2002) were also relatively small (biogeographical example: 21 regions; muskrats: nine population zones; *Aphelandra*: 12 species plus hybrids). The program can, in fact, analyse much larger data sets sufficiently quickly to provide a reconstruction in reasonable time.

In the present study, minimum path-length distances among nodes in reticulate phylogenies were used to approximate empirical evolutionary distances among the species. It is important to note that the minimum path-length distance is simply another expression of the principle of parsimony which is widely applied to phylogenetic reconstruction problems. The principle of parsimony, also called 'Ockham's razor', was formulated by the English logician and philosopher William Ockham (1290–1349). It states: '*Pluralites non est ponenda sine necessitate* [Multiplicity ought not to be posited without necessity]'. In other words, unnecessary assumptions should be avoided when formulating hypotheses.

Following this principle, parameters should be used with parsimony in modelling, and any parameter or assumption that is not necessary should be eliminated. Other models could

also be used to calculate path lengths. For instance, one might consider a model that allows splitting at each node with a certain probability to the next descendants; the distance between two taxa through different paths can be weighted by the probabilities of these paths. Such a probabilistic approach should constitute an interesting and relevant subject for further development of phylogenetic reticulation analysis.

Acknowledgements

The authors are grateful to Philippe Casgrain for his contribution to programming *T-Rex*, which makes the new method of reticulation analysis available to the scientific community.

References

Atchley, W. R. & Fitch, W. M. (1991). Gene trees and the origins of inbred strains of mice. *Science*, *254*, 554–558.

Atchley, W. R. & Fitch, W. M. (1993). Genetic affinities of inbred mouse strains of uncertain origin. *Molecular Biology and Evolution*, *10*, 1150–1169.

Bandelt, H.-J. & Dress, A. W. M. (1992a). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution*, *1*, 242–252.

Bandelt, H.-J. & Dress, A. W. M. (1992b). A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, *92*, 47–65.

Barthélémy, J. P. & Guénoche, A. (1991). *Trees and Proximity Representations*. New York: Wiley.

Baudry, E., Solignac, M., Garnery, L., Gries, M., Cornuet, J. M. & Koeniger, N. (1998). Relatedness among honeybees *Apis mellifera* of a drone congregation. *Proceedings of the Royal Society of London B*, *265*, 2009–2014.

Bryant, D. & Waddell, P. (1998). Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. *Molecular Biology and Evolution*, *15*, 1346–1359.

Cheung, B., Holmes, R. S., Easteal, S. & Beacham, I. R. (1999). Evolution of class I alcohol dehydrogenase genes in catarrhine primates: Gene conversion, substitution rates, and gene regulation. *Molecular Biology and Evolution*, *16*, 23–36.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science*, *284*, 2124–2128.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, *39*, 783–791.

Fitch, D. H. A., Mainone, C., Goodman, M. & Sligh-Tom, J. L. (1990). Molecular history of gene conversions in the primate fetal y-globin genes. *Journal of Biological Chemistry*, *265*, 781–793.

Gascuel, O. (1997). Concerning the NJ algorithm and its unweighted version, UNJ. In B. Mirkin, F. R. McMorris, F. Roberts & A. Rzhetsky (Eds) *Mathematical Hierarchies and Biology* (pp. 149–170). DIMACS Series in Discrete Mathematics and Theoretical Computer Science. Providence, RI: American Mathematical Society.

Guttman, D. S. & Dykhuizen, D. E. (1994). Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science*, *266*, 1380–1383.

Hatta, M., Fukami, H., Wang, W., Omori, M., Shimoike, K., Hayashibara, T., Ina, Y. & Sugiyama, T. (1999). Reproductive and genetic evidence for a reticulate evolutionary history of mass-spawning corals. *Molecular Biology and Evolution*, *16*, 1607–1613.

Hein, J. (1993). A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, *36*, 396–405.

Hugall, A., Stanton, J. & Moritz, C. (1999). Reticulate evolution and the origins of ribosomal internal transcribed spacer diversity in apomictic *Meloidogyne*. *Molecular Biology and Evolution*, *16*, 157–164.

Huson, D. H. (1998). SplitsTree: a program for analyzing and visualizing evolutionary data. *Bioinformatics*, *141*, 68–73.

Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. In H. N. Munro (Ed.) *Mammalian Protein Metabolism* (pp. 21–132). New York: Academic Press.

Kimura, M. (1981). Estimation of evolutionary distances between *homologous* nucleotide sequences. *Proceedings of the National Academy of Sciences of the USA*, *78*, 454–458.

Koeniger, G., Koeniger, N., Mardan, M. & Wongsiri, S. (1993). Variance in weight of sexuals and workers within and between 4 *Apis* species (*Apis florea*, *Apis dorsata*, *Apis cerana* and *Apis mellifera*). *Asian Apiculture*, *1*, 106–111.

Legendre, P. & Makarenkov, V. (2002). Reconstruction of biogeographic and evolutionary networks using reticulograms. *Systematic Biology*, *51*, 199–216.

Maddison, D. R. & Maddison, W. P. (2000). *Macclade 4: Analysis of Phylogeny and Character Evolution*. Sunderland, Massachusetts: Sinauer Associates.

Makarenkov, V. (2001). T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, *17*, 664–668.

Makarenkov, V. & 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 & A. Rzhetsky (Eds) *Mathematical Hierarchies and Biology* (pp. 183–208). DIMACS Series in Discrete Mathematics and Theoretical Computer Science. Providence, RI: American Mathematical Society.

Makarenkov, V. & Leclerc, B. (1999). An algorithm for the fitting of an additive distance according to a weighted least-squares criterion. *Journal of Classification*, *16*, 3–27.

Makarenkov, V. & Leclerc, B. (2000). Comparison of additive trees using circular orders. *Journal of Computational Biology*, *7*, 731–744.

Milner, A. (1996). An introduction to understanding honeybees, their origins, evolution and diversity. *Available via Bibba Electronic Journal*. URL: http://www.bibba.com.

Nesbø, C. L., L'Haridon, S., Stetter, K. O. & Doolittle, W. F. (2001). Phylogenetic analyses of two 'archaeal' genes in *Thermotoga maritima* reveal multiple transfers between archaea and bacteria. *Molecular Biology and Evolution*, *18*, 362–375.

Odorico, D. M. & Miller, D. J. (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. *Molecular Biology and Evolution*, *14*, 465–473.

Posada, D. & Crandall, K. A. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics*, *14*, 817–818.

Robertson, D. L., Hahn, B. H. & Sharp, P. M. (1995). Recombination in AIDS viruses. *Journal of Molecular Evolution*, *40*, 249–259.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*, 406–425.

Sattath, S. & Tversky, A. (1977). Phylogenetic similarity trees. *Psychometrika*, *42*, 319–345.

Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Molecular Biology and Evolution*, *6*, 526–536.

Sneath, P. H. A., Sackin, M. J. & Ambler, R. P. (1975). Detecting evolutionary incompatibilities from protein sequences. *Systematic Zoology*, *24*, 311–332.

Sonea, S. & Panisset, M. (1976). Pour une nouvelle bactériologie. *Revue Canadienne de Biologie*, *35*, 103–167.

Sonea, S. & Panisset, M. (1981). *Introduction à la Nouvelle Bactériologie*. Montréal: Presse de l'Université de Montréal.

Stace, C. A. (1984). *Plant Taxonomy and Biosystematics*. London: Edward Arnold.

Steel, M. A. (1994). Recovering a tree from the leaf colorations it generates under a Markov model. *Applied Math Letters*, *72*, 19–24.

Stephens, J. C. (1985). Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Molecular Biology and Evolution*, *2*, 539–556.

Swofford, D. L. (2001). *PAUP. Phylogenetic Analysis Using Parsimony and Other Methods*, Version 4.0d8, Champaign, Illinois: Illinois Natural History Survey.

Walter, S. J., Campbell, C. S., Kellogg, E. A. & Stevens, P. F. (1999). *Plant Systematics A Phylogenetic Approach*. Sunderland, Massachusetts: Sinauer Associates.