Determining horizontal gene transfers in species classification: unique scenario

Vladimir Makarenkov[1,2], Alix Boc[1] and Abdoulaye Baniré Diallo[1]

[1] Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succ.
Centre-Ville, Montral (Québec), Canada, H3C 3P8
email: makarenkov.vladimir@uqam.ca, boc.alix@courrier.uqam.ca
and banire@math.uqam.ca

[2] Institute of Control Sciences, 65 Profsoyuznaya, Moscow 117806, Russia

Abtract. The problem of species classification taking into account the mechanisms of reticulate evolution such that horizontal gene transfer (HGT), species hybridization or gene duplication is very delicate. In this paper, we describe a new algorithm for determining a unique scenario of HGT events in a given additive tree (i.e. phylogenetic tree) representing evolution of a group of species. The algorithm first proceeds by establishing differences between topologies of species and gene additive trees. Then, it uses a least-squares optimization procedure to test the possibility of horizontal gene transfers between any pair of edges of the species tree, considering all previously added HGTs to determine the next one. In the application section, we show how the introduced algorithm can be used to represent possible ways of spread of the rubisco *rbcL* gene in a species classification including plastids, cyanobacteria, and proteobacteria.

1 Introduction

Species classification has been often modeled using an additive tree in which each species can only be linked to its closest ancestor and interspecies relationships are not allowed. However, such important evolutionary mechanisms as gene convergence, gene duplication, gene loss, and horizontal gene transfer (i.e. lateral gene transfer) can be appropriately represented only using a network model (see [OW97] or [Doo99]). This paper addresses the problem of detection of horizontal gene transfer events. Several attempts to use network-based evolutionary models to represent horizontal gene transfers can be found in the scientific literature (see for example [Pag94], [PC98], or [HL01]). Recently we proposed a new method [BM03] for detection of HGT events based on a mathematically sound model using a least-squares mapping of a gene tree into a species tree. The latter method proceeds by obtaining a classification of probable HGTs that may have occurred in course of the evolution; the biologists should then be able to choose appropriate HGTs from the classification established. In this article, we propose a new approach allowing one to establish a unique scenario of horizontal gene transfers. This approach exploits topological discrepancies existing in the species and gene additive trees. To reconcile topological differences between the species and gene trees, we recompute the edge lengths of the species tree with respect to the gene data. Then, each pair of edges of the species tree will be evaluated for the possibility of a horizontal gene transfer and the best one, according to the least-squares criterion, will be added to the species tree. All further HGT edges will be added to the species tree in the same way leading to the construction of an HGT network. A number of computational rules plausible from the biological point of view are incorporated in our model. In the application section, we discuss a unique scenario of the lateral gene transfers of the *rbcl* gene in a bacteria classification considered in [DP 96].

2 Description of the new method

The new algorithm allows one to determine the best possible least-squares scenario of horizontal gene transfer events for a group of considered organisms. It proceeds first by mapping gene data into a species tree followed by consecutive addition of new HGT edges with directions. Thus, the algorithms comprises the two main steps described below:

Step 1. Let T be an additive species tree whose leaves are labeled according to the set X of n taxa and T_1 a gene tree whose leaves are labeled according to the same set X of n taxa. T and T_1 can be inferred from sequence or distance data using an appropriate tree fitting algorithm. Without lost of generality we assume that T and T_1 are binaries trees, whose internal nodes are all of degree 3 and whose number of edges is $2n$-3. The species tree should be explicitly rooted; the position of the root is important in our model. If the topologies of T and T_1 are identical, we conclude that the evolution of the given gene followed that of the species, and no horizontal gene transfers between edges of the species tree should be indicated. However, if the two additive trees are topologically different it may be the result of horizontal gene transfers. In the latter case, the gene tree T_1 can be mapped into the species tree T by fitting by least-squares the edge lengths of T to the pairwise distances in T_1 (see [BG91] or [ML99] for more detail on this technique).

Step 2. The goal of this step is to establish a classification of all possible HGT connections between pairs of edges in T. In our model, the HGT edges providing the greatest contribution to decreasing the least-squares coefficient will correspond to the most probable cases of the horizontal gene transfers. Thus, the first HGT in this list will be added to the species tree T transforming it into a phylogenetic network. Once the first HGT edge is added to T, all its edges including a new HGT edge, will be reestimated to fit the best the inter-leaves distances in the gene tree T_1. To add the first HGT edge to the network under construction, the algorithm considers (2n-3)(2n-4) possibilities, which is a maximum number of different directed inter-edge connections in a binary phylogenetic tree with n leaves. Then, the best second, third, and so forth, HGT edges are added to T in the same way. The addition of any new HGT edge starting from the second one is done taking into account all previously added HGTs. The al gorithm stops when a prefixed number of HGT edges are added to T. The obtained phylogenetic network will represent a possible scenario, which is the best one according to the least-squares criterion, of the horizontal transfers of the gene considered.

3 Computing the least-squares coefficient

The addition of a new edge may create an extra path between any pair of nodes in a phylogenetic network. Fig. 1 illustrates the only possible case when the minimum path-length distance between taxa (i.e. species) i and j is allowed to pass by the new HGT edge (a,b) directed from b to a. From the biological point of view it would be plausible to allow the horizontal gene transfer between b and a to affect the evolutionary distance between the pair of taxa i, whose position in T is fixed, and j if and only if j is located in the grayed area of Fig. 1. In all other cases, illustrated in Fig. 2 (a to d), the path between the taxa i and j should not be permitted to pass by the new edge (a,b) representing the gene transfer from b to a.

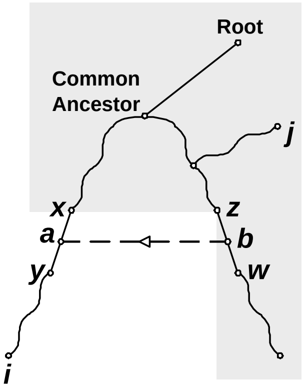

Fig. 1. The minimum path-length distance between the taxa i and j can be affected by addition of a new edge (a,b) representing the horizontal gene transfer between edges (z,w) and (x,y) in the species tree if the leaf j is located in the grayed area of the picture. The path between the taxa i and j can pass by the new edge (a,b).

To compute the value of the least-squares coefficient Q for a given HGT edge (a,b) the following strategy was adopted: first, we define the set of all pairs of taxa that can be allowed to pass by a new HGT edge (a,b); second, in this set we determine all pairs of taxa such that the minimum path-length distance between them may decrease after addition of (a,b); third, we look for an optimal value l of the length of (a,b), according to the least-squares criterion, while keeping fixed the lengths of all other tree edges; and finally, forth, all edge lengths are reassessed one at a time.

Let us define the set A(a,b) of all pairs of taxa ij such that the distances between them may change if an HGT edge(a,b) is added to the tree T. A(a,b) is the set of all pairs of taxa ij such that they are located in T as shown in Fig. 1 and:

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

where $d(i,j)$ is the minimum path-length distance between the nodes i and j in T; vertices a and b are located in the middle of the edges (x,y) and (z,w), respectively.
The following intermediate function can be defined:

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

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

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

$$+ \sum_{dist(i,j)\leq l} (d(i,j) - \delta(i,j))^2 \rightarrow min, \tag{3}$$

where $\delta(i,j)$ is the minimum path-length distance between the taxa i and j in the gene tree T_1. The function $Q(ab,l)$, measures the gain in fit when a new HGT edge (a,b) with length l is added to the species tree T.

When the optimal value of a new edge (a,b) is determined, this computation can be followed by an overall polishing procedure for all edge lengths in T. To reassess the length of any edge of T, one can use equations (1), (2), and (3) assuming that the lengths of all the other edges are fixed. These computations are repeated for all pairs of edges in the species tree T. When all pairs of edges in T are tested, only the HGT corresponding to the smallest value of Q is retained for addition in T. This algorithm requires $O(kn^4)$ operations to provide one with a unique HGT scenario including k transfer edges.

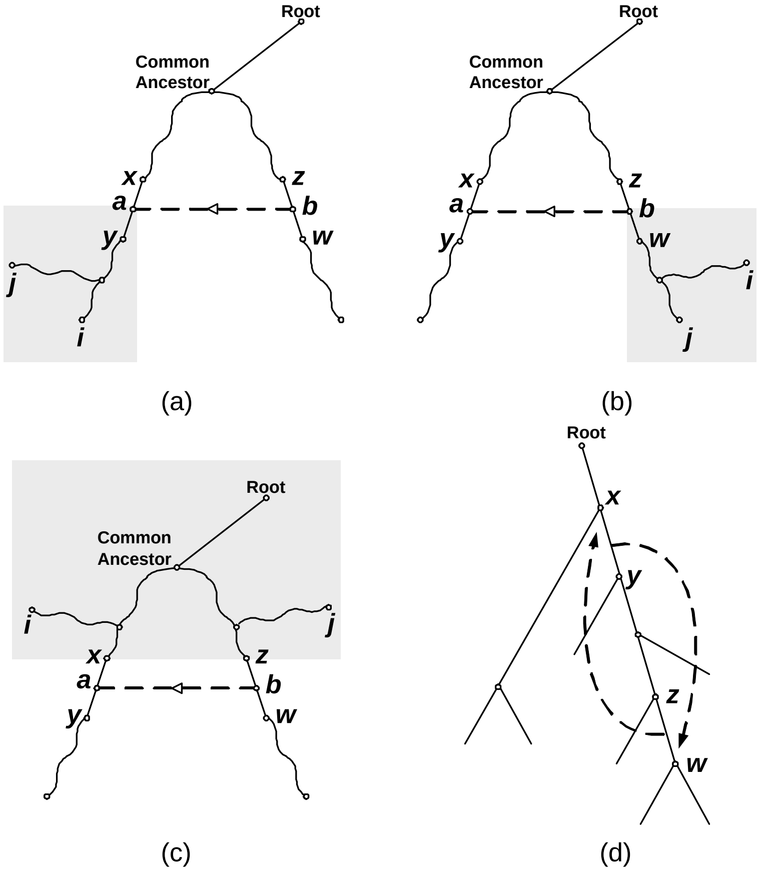

Fig. 2. (a-c) Three situations when the minimum path-length distance between the taxa i and j is not affected by addition of a new edge (a,b) representing the horizontal gene transfer between edges (z,w) and (x,y) of the species tree. The path between the taxa i and j is not allowed to pass by the new edge (a,b). In figures (a), (b), and (c) both leaves i and j must be located in the grayed area. Figure (d) shows that no HGT can be considered when edges (x,y) and (z,w) are located on the same lineage (i.e. on the same path coming from the root).

4 Horizontal gene transfers of the *rbcL* gene: unique scenario

The new algorithm discussed in previous sections was applied to analyze the plastids, cyanobacteria, and proteobacteria data considered in [DP96]. The latter authors found that the gene classification based on the rbcL gene contains a number of conflicts compared to the species classification (for 48 species) based on the 16S ribosomal RNA and other evidence. To carry out the analysis we reduced the number of species to 15 (see trees in Fig. 3a and b). Each species shown in Fig. 3 represents a group of bacteria or plastids from the original phylogeny provided by Delwiche and Palmer (1996, Fig.2). We decided to conduct our study with three α-proteobacteria, three β-proteobacteria, three γ-proteobacteria, two cyonobacteria, one green plastid, one red and brown plastid and two single species *Gonyaulax* and *Cyanophora*. The new algorithm used as input the species and gene additive trees in Fig. 3 (a and b) and provided us with a unique scenario of horizontal gene transfers of the *rbcL* gene. The topological conflicts between the trees in Fig. 3 can be explained either by lateral gene transfers that may have taken place between the species indicated or by ancient gene duplication followed by gene loss; two hypotheses which are not mutually exclusive (see [DP96] for more detail). In this paper, the lateral gene transfer hypothesis was examined to explain the conflicts between the species and gene classifications.

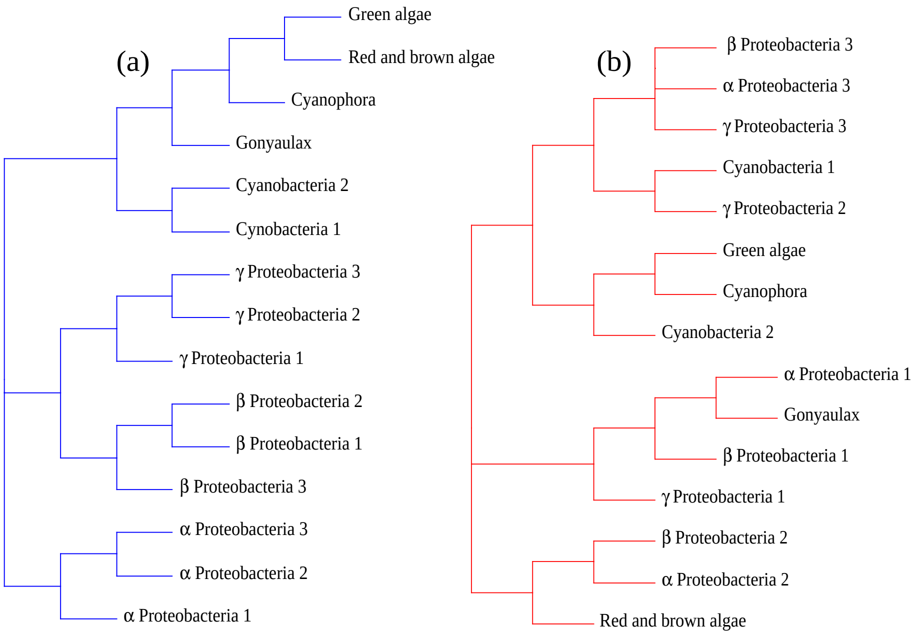

Fig. 3. (a) Species tree for 15 taxa representing different groups of bacteria and plastids from [DP96, Fig. 2]. Each taxon represents a group of organisms reported in [BM03]. Species tree is built on the base of 16S rRNA sequences and other evidence. (b) *rbcL* gene tree for 15 taxa representing different groups of bacteria and plastids constructed by contracting nodes of the 48 taxa phylogeny from [DP96, Fig. 2].

The solution network depicting the species tree with eight horizontal gene transfers is shown in Fig. 4. The numbers at the HGT edges correspond to their position in the scenario of transfers. Thus, the transfer between α-proteobacteria1 and *Gonyaulax* was found in the first iteration of the algorithm, then, the transfer between α-proteobacteria1 and β-proteobacteria1, followed by that from γ-proteobacteria2 to cyanobacteria1, and so forth. Delwiche and Palmer (1996, Fig. 4) indicated as the most probable four HGT events of the rubisco genes including those between cyanobacteria and γ-proteobacteria, γ-proteobacteria and α-proteobacteria, γ-proteobacteria and β-proteobacteria, and finally, between α-proteobacteria and plastids. The three last transfers can be found in our unique scenario in Fig 4, whereas the first one, between cyanobacteria and γ-proteobacteria, goes in the opposite direction. It is worth noting that the obtained solution is also different from that found in [BM03].

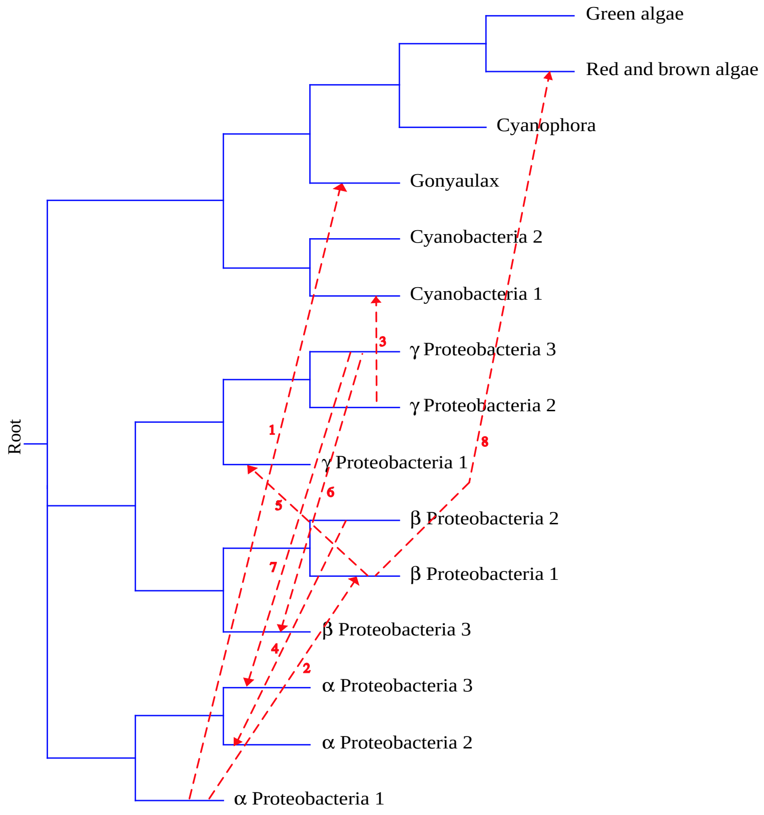

Fig. 4. Species tree from Fig. 3a with 8 dashed, arrow-headed edges representing horizontal gene transfers of the *rbcL* gene. Numbers on the HGT edges indicate their order of appearance.

5 Conclusion

The algorithm described in this article allows one to determine a unique scenario of horizontal gene transfer events that may have occurred in course of the evolution. The algorithm exploits the discrepancy between the species and gene additive trees built for the same set of observed species by mapping the gene data into the species tree and then by estimating the possibility of a horizontal gene transfer between each pair of edges in the species tree. A number of plausible biological rules not considered in [BM03] are incorporated in the reconstruction process. The best HGT, according to the least-squares model, found in the first iteration is used to calculate the next HGT in the second iteration, and so one. The example of evolution of the *rbcL* gene considered in the previous section clearly shows that the new method can be useful for prediction of horizontal gene transfers in real data sets. In this paper, a model based on the least-squares criterion was considered. Future developments including extending and testing this procedure in the framework of the maximum likelihood and maximum parsimony models are necessary. The algorithm for detection of horizontal gene transfers described in this article was included in the T-Rex package (Tree and Reticulogram reconstruction, see [Mak01] for more detail). This software is freely available for researchers at the following URL: <http://www.info.uqam.ca/ makarenv/trex.html>.

References

[BG91] Barthélemy, J.-P., Guénoche A.: Trees and Proximity Relations. Wiley, New York (1991)

[BM03] Boc, A. and Makarenkov, V.: New Efficient Algorithm for Detection of Horizontal Gene Transfer Events. In: G. Benson and R. Page (Eds) Lecture Notes in BioInformatic, Springer, Berlin Heidelberg New York (2003)

[Doo99] Doolittle, W. F.: Phylogenetic classification and the universal tree. Science, **284**, 2124–2128 (1999)

[DP96] Delwiche, C.F., Palmer, J.D.: Rampant horizontal transfer and duplication of Rubisco genes in Eubacteria and Plastids Mol. Biol. Evol. **13(6)** 873–882 (1996)

[HL01] Hallet, M., Lagergreen, J.: Efficient algorithms for lateral gene transfer problems. In: Emery, M. (ed) Proceedings of the 5th Conference on Computational Molecular Biology. ACM-Press, New York (2001)

[Mak01] Makarenkov,V.: T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks. Bioinformatics, **17**, 664–668 (2001)

[ML99] Makarenkov, V., Leclerc, B.: An algorithm for the fitting of a phylogenetic tree according to weighted least-squares. J. of Classif., **16**, 3–26 (1999)

[OW98] Olsen, G., Woese, C.: Archael genomics overview. Cell, **89**, 991–994 (1998)

[Pag94] Page, R. D. M.: Maps between trees and cladistic analysis of historical associations among genes, organism and areas. Syst. Biol., **43**, 58–77 (1994)

[PC98] Page, R. D. M., Charleston, M. A.: From gene to organismal phylogeny: Reconciled trees. Bioinformatics, **14**, 819–820 (1998)