New Efficient Algorithm for Detection of Horizontal Gene Transfer Events

Alix Boc[1] and Vladimir Makarenkov[1,2]

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

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

Abstract. This article addresses the problem of detection of horizontal gene transfers (HGT) in evolutionary data. We describe a new method allowing to predict the ways of possible HGT events which may have occurred during the evolution of a group of considered organisms. The proposed method proceeds by establishing differences between topologies of species and gene phylogenetic trees. Then, it uses a least-squares optimization procedure to test the possibility of horizontal gene transfers between any couple of branches of the species tree. In the application section we show how the introduced method can be used to predict eventual transfers of the rubisco *rbcL* gene in the molecular phylogeny including plastids, cyanobacteria, and proteobacteria.

1 Introduction

Evolutionary relationships between species has long been assumed to be a tree-like process that has to be represented by means of a phylogenetic tree. In such a tree, each species can only be linked to its closest ancestor and interspecies relationships are not allowed. However, such important evolutionary mechanism as *horizontal gene transfer* (i.e. *lateral gene transfer*) can be represented appropriately only using a network model. Lateral gene transfer plays an key role in bacterial evolution allowing bacteria to exchange genes across species [6], [21]. Moreover, numerous bacterial sequencing projects reinforced the opinion that evolutionary relationships between species cannot be inferred from the information based on a single gene, i.e. single gene phylogeny, because of existence of such evolutionary events as gene convergence, gene duplication, gene loss, and horizontal gene transfer (see [10], [5], [11], [17]). This article is concerned with studying the possibility of horizontal gene transfers between branches of species trees inferred either from whole species genomes or based on genes that are not supposed to be duplicated, lost or laterally transferred. We will show how the discrepancy between a phylogenetic tree based on a particular gene family and a species tree can be exploited to depict possible scenarios of how this particular gene may have been laterally transferred in course of the evolution.

Several attempts to use network-based evolutionary models to represent lateral gene transfers can be found in the scientific literature (see for example [11],

[23]). Furthermore, a number of models based on the subtree transfer operations on leaf labeled trees have also been proposed (see [4], [12], [13]). Recently, a new lateral gene transfer model considering a mapping of a set of gene trees (not necessarily pairwise equal) into a species tree has been proposed [10]. The latter article completed an extensive study of the tree mapping problems considered amongst others in [3], [9], [10], [18], [19].

In this paper we define a mathematically sound model using a least-squares mapping of a gene tree into a species tree. The proposed model is based on the computation of differences between pairwise distances between species in both trees. First, a species phylogeny has to be inferred from available nucleotide or protein sequences using an appropriate tree inferring algorithm. Second, the matrix of evolutionary distances between species with respect to the gene data has to be computed using an appropriate sequence-distance transformation. Third, the length of the species tree should be readjusted with respect to the gene distance matrix (see [1], [15] for more detail). Then, each pair of branches of the species tree (with the original topology and branch lengths adjusted according to the gene distance matrix) has to be evaluated for the possibility of a horizontal gene transfer. A new model introduced in this paper takes into account all different situations sound from the biological point of view when the lateral gene transfer event can explain the discrepancy in positioning two taxa in the gene and species phylogenies.

The new method has been tested with both real and artificial data sets yielding very encouraging results for both types data. In the application section below we show how our method aids to detect possible horizontal gene transfers of the rubisco *rbcL* gene (ribulose-1.5-bisphosphate carboxylase/oxygenase) in the species phylogeny inferred from phylogenetic analysis of 16S rRNA and other evidence (see [5] for more detail on these data). For this data set, the new method provided us with the solution consisting of eight horizontal gene transfers accounting for the conflicts between the *rbcL* and species phylogenies. Among the eight transfers obtained one can find all eventual gene transfers indicated in Delwiche and Palmer (1996). In a latter study, four *rbcL* gene transfers between: cyanobacteria and γ-proteobacteria, α-proteobacteria and red and brown algae, γ-proteobacteria and α-proteobacteria, and γ-proteobacteria and β-proteobacteria, were suggested as the most probable cause (along with the hypothesis of gene duplication) of topological differences between the organismal and gene phylogenies.

2 Description of the new method

In this section, we describe a new method for detection of lateral gene transfer events obtained by mapping of a gene data into a species phylogeny. The new method allows to incorporate new branches with direction into the species phylogeny to represent gene transfers. Remember that any phylogenetic tree can be associated with a table of pairwise distances between its leaves which are labeled by the names of species; these distances are the minimum path-length distances

between the leaves of the tree. All other nodes of the tree are intermediates, they represent unknown ancestors. It has been shown that a distance matrix satisfying the four-point condition (1) defines a unique phylogenetic tree [2].

$$d(i,j) + d(k,l) \leq Max\{d(i,k) + d(j,l); d(i,l) + d(j,k)\}, \text{for any } i,j,k,l. \quad (1)$$

When the four-point condition is not satisfied, what is always the case for real data sets, a tree inferring method has to be applied. There exist a number of efficient methods for inferring phylogenies from distance data; see for example NJ of Saitou and Nei (1987), BioNJ of Gascuel (1997), FITCH of Felsenstein (1997), or MW of Makarenkov and Leclerc (1999).

The main objective of our method is to infer a species tree from sequence or distance data and then test all possible pairs of the tree branches against the hypothesis that a lateral gene transfer could take place between them. Our method consists of the three main steps described below:

Step 1. Let T be a species phylogeny whose leaves are labeled according to the set X of n taxa. T can be inferred from sequence or distance data using an appropriate tree fitting method. Without lost of generality we assume that T is a binary tree, whose internal nodes are all of degree 3 and whose number of branches is $2n$-3. This tree should be explicitly rooted because the position of the root is important in our model.

Step 2. Let T_1 be a gene tree whose leaves are labeled according to the same set X of n taxa used to label the species tree T. Similarly to the species tree, T_1 can be inferred from sequence or distance data characterizing this particular gene. If the topologies of T and T_1 are identical, no horizontal gene transfers between branches of the species tree should be indicated. However, if the two phylogenies are topologically different it may be the result of a horizontal gene transfer. In the latter 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 (for an overview of this fitting techniques, see Bryant and Wadell 1998 and Makarenkov and Leclerc 1999). These papers discuss two different ways of computing optimal branch lengths, according to the least-squares criterion, of a phylogenetic tree with fixed topology. After this operation the branch lengths of T will be modified, whereas its original topology will be kept unchanged.

Step 3. The goal of this step is to obtain an ordered list L of all possible HGT connections between pairs of branches in T. This list will comprise ($2n$-3)($2n$-4) entries, which is the possible number of different directed connections (i. e. number of possible HGTs) in a binary phylogenetic tree with n leaves. Each entry of L is associated with the value of the gain in fit obtained after addition of a new HGT branch linking a considered couple of branches. The first entries of L, those contributing the most to decrease the least-squares coefficient, will correspond to the most probable cases of the horizontal gene transfers.

Let us now show how to compute the value of the least-squares coefficient Q for an HGT branch (a,b) added to T to link the branches (x,y) and (z,w). In a phylogenetic tree there exists always a unique path linking any pair of the tree nodes, whereas addition of an HGT branch may create an extra path between

them. Fig. 1(a, b, and c) illustrate the three possible cases when the minimum path-length distance between taxa i and j are allowed to be changed after the addition of the new branch (a,b) directed from b to a. Fixing the position of i in the species tree T, these cases differ only by position of j. 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 and j if and only if either the node b is an ancestor of j (Fig. 1a) or the attachment point of j on the path (x,z) is located between the node z and Common Ancestor of x and z (Fig. 1b and 1c).

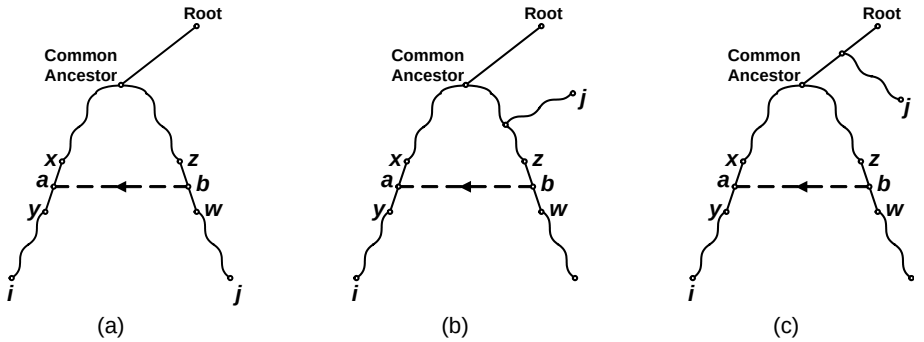

Fig. 1. Three situations when the minimum path-length distance between the taxa i and j can be affected by addition of a new branch (a,b) representing the horizontal gene transfer between branches (z,w) and (x,y) of the species tree. The path between the taxa i and j can now pass by the new branch (a,b).

In all other cases illustrated in Fig. 2 (a to f), the path between the taxa i and j cannot pass by the new branch (a,b) depicting the gene transfer from b to a. To compute the value of the least-squares coefficient Q for a given HGT branch (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 branch (a,b); second, in the latter 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 (a,b), according to the least-squares criterion, while keeping fixed the lengths of all other tree branches; and finally, forth, all branch 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 branch (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 (a, b, or c) and:

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

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

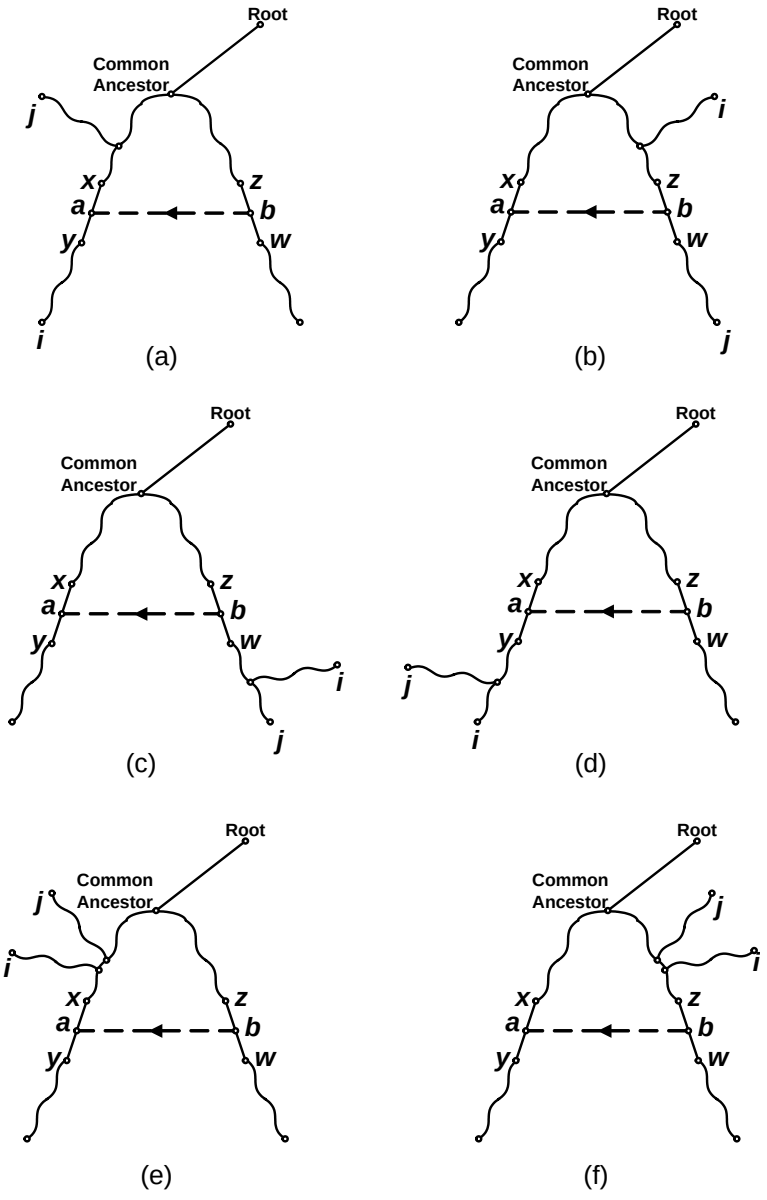

Fig. 2. Six situations when the minimum path-length distance between the taxa i and j is not affected by addition of a new branch (a,b) representing the horizontal gene transfer between branches (z,w) and (x,y) of the species tree. The path between the taxa i and j is not allowed to pass by the new branch (a,b).

Define the following function:

$$dist(i,j) = d(i,j) \; - \; Min\{d(i,a) + d(j,b); d(j,a) + d(i,b)\}, \tag{3}$$

so that $A(a,b)$ is a set of all leaf pairs ij with $dist(i,j) > 0$.

The least-squares loss function to be minimized, with l used as an unknown variable, is formulated as follows:

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

$$+ \sum_{dist(i,j)\leq l} (d(i,j) - \delta(i,j))^2 \rightarrow min, \tag{4}$$

where $d(i,j)$ is the minimum path-length distance between the taxa i and j in the gene tree T_1. The function $Q(ab,l)$, which is a quadratic polynomial spline, measures the gain in fit when a new HGT branch (a,b) with length l is added to the species tree T. The lower is the value of $Q(ab,l)$, the more likely that a horizontal gene transfer event has occurred between the branches (x,y) and (z,w).

Once the optimal value of a new branch (a,b) is computed, this computation can be followed by an overall polishing procedure for branch lengths reevaluation. In fact, the same calculations can be used to reevaluate the lengths of all other branch in T. To reassess the length of any branch in T, one can use equations (2), (3), and (4) assuming that the lengths of all the other branches are fixed. Thus, the polishing procedure can be carried out for branch number one, then branch number two, and so on, until all branch lengths are optimally reassessed. Then, one can return to the added HGT branch to reassess its length for the second time, and so forth.

These computations are repeated for all pairs of branches in the species tree T. When all pairs of branches in T are tested, an ordered list L providing their classification with respect to the possibility of a horizontal gene transfer can be established. This algorithm takes $O(n^4)$ to compute the optimal value of Q for all possible pairs of branches in T, the algorithmic complexity increases up to $O(n^5)$ when the branch lengths polishing procedure is also carried out.

3 Lateral gene transfers of the *rbcL* gene

The method introduced in the previous section was applied to analyze the plastids, cyanobacteria, and proteobacteria data considered in Delwiche and Palmer (1996). The latter paper discusses the hypotheses of lateral gene transfer events of the rubisco genes between the three above-mentioned groups of organisms. Delwiche and Palmer inferred a maximum parsimony phylogeny of the *rbcL* (large subunit of rubisco) gene for 48 species. This phylogeny is shown in Fig. 3. The latter authors found that the gene classification based on the *rbcL* gene contains a number of conflicts compared to the species classification based on the 16S

ribosomal RNA and other evidence. The aligned *rbcL* protein sequences for these species are available at www.life.umd.edu/labs/delwiche/publications.html.

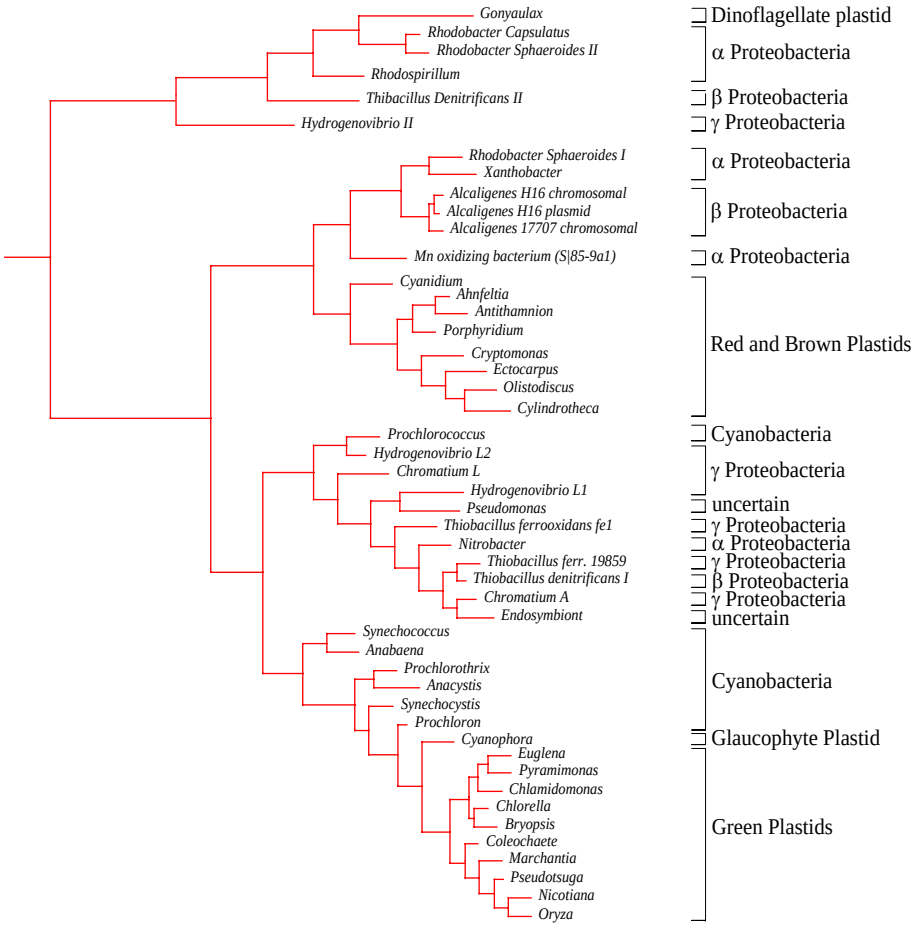

Fig. 3. Maximum parsimony tree of *rbcL* amino acids for 48 bacteria and plastids (from Fig. 2 of Delwiche and Palmer 1996). The tree shown is one of 24 shortest trees of length 2.422, selected arbitrarily. This phylogeny is inferred from 48 *rbcL* amino acid sequences with 497 bases. Classification of taxa based on 16S rRNA and other evidence is indicated to the right.

To apply our method we first attempted to construct the species tree of these 48 organisms based on the sequences data from NCBI [22], Ribosomal Database Project [14] and other bioinformatics databases. However, the 16S rRNA data were available only for 28 of 48 species considered.

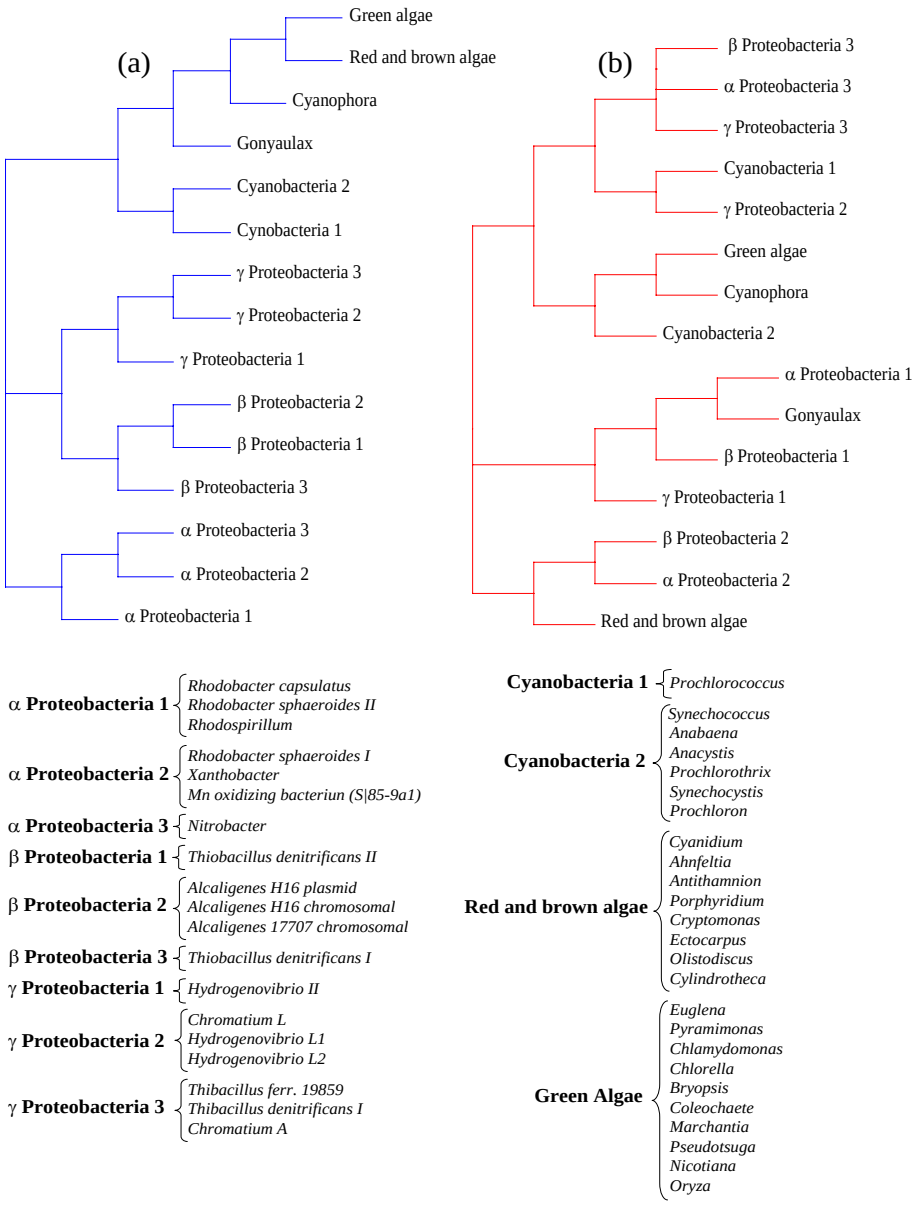

Fig. 4. (a) Species tree for 15 taxa representing different groups of bacteria and plastids from Fig. 3. Each taxon represents a group of organisms reported in the bottom. Species tree is built on the base of 16S rRNA sequences and other evidence. (b) *rbcL* gene tree for 15 taxa representing different groups of bacteria and plastids from Fig. 3. Gene tree is constructed by contracting nodes of the 48 taxa phylogeny in Fig. 3.

Thus, to carry out the analysis we reduced the number of species to 15 (see trees in Fig. 4a and b). Each species shown in Fig. 4 represents a group of bacteria or plastids from Fig. 3. We decided to conduct our study with three α-proteobacteria, three β-proteobacteria, three γ-proteobacteria, two cyonobacteria, one green plastids, one red and brown plastids, and two single species Gonyaulax and Cyanophora. The species tree in Fig. 4a was built using 16S rRNA sequences and other evidence existing in the scientific literature. The gene tree (Fig. 4b) was derived by contracting nodes of the 48 taxa phylogeny in Fig. 3. While observing the topologies of the species and gene trees one can note an important discrepancy between them.

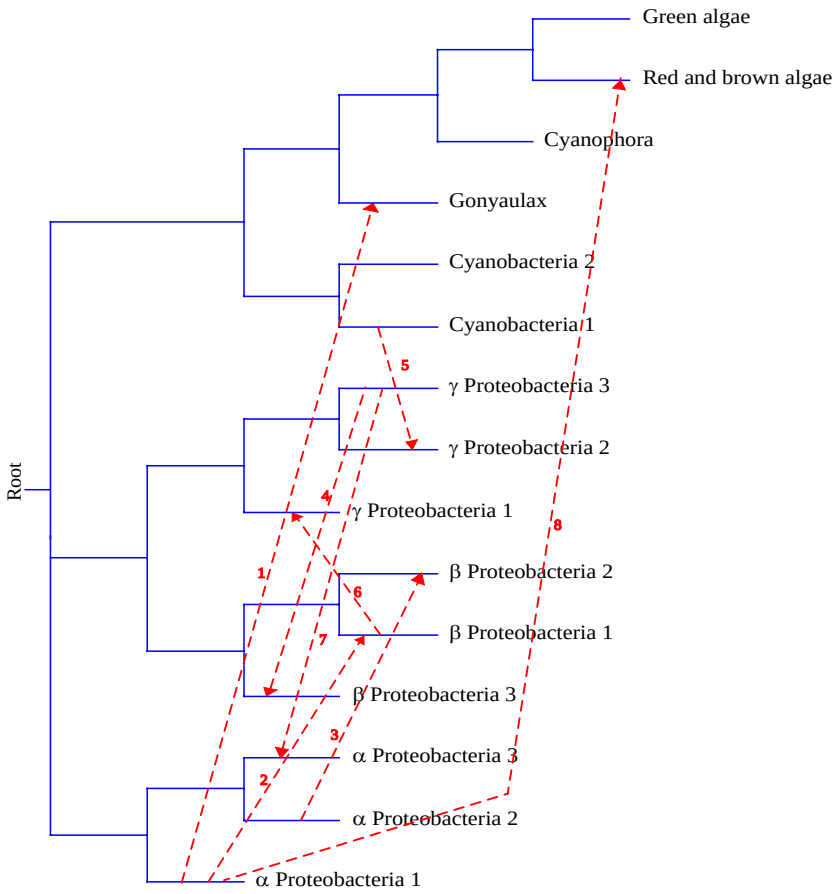

Fig. 5. Species tree from Fig. 4a with 8 dashed, arrow-headed lines representing possible horizontal gene transfers of the *rbcL* gene found by the new algorithm. Numbers on the arrow-headed lines indicate their order of appearance (i.e. order of importance) in the list of all possible HGT transfers found.

For instance, the gene tree comprises a cluster regrouping α-proteobacteria3, β-proteobacteria3, and γ-proteobacteria3, as well as cynanobacteria1 is clustering with γ-proteobacteria2, and so on.

These contradictions can be explained either by lateral gene transfers that may have taken place between the species indicated or by ancient gene duplication; two hypotheses which are not mutually exclusive (see [5] for more detail). In this paper, the lateral gene transfer hypothesis is examined to explain the conflicts between the species and gene phylogenies.

The method introduced in this article was applied to the two phylogenies in Fig. 4(a and b) and provided us with a list of lateral gene transfers, ordered according their likelihood, between branches of the species tree. To reduce the algorithmic time complexity the links between adjacent branches were not considered. The solution network depicting the gene tree with eight horizontal gene transfers is shown in Fig. 5. The numbers at the arrows representing HGT events correspond to their position in the ordered list of transfers. Thus, the transfer between α-proteobacteria1 and Gonyaulax was considered as the most significant, then, the transfer between α-proteobacteria1 and β-proteobacteria1, followed by that from α-proteobacteria2 to β-proteobacteria2, and so forth. Delwiche and Palmer (1996, fig. 4) indicated four HGT events of the rubisco genes between cyanobacteria and γ-proteobacteria, γ-proteobacteria and α-proteobacteria, γ-proteobacteria and β-proteobacteria, and finally, between α-proteobacteria and plastids. All these transfers can be found in our model in Fig. 5.

4 Conclusion

We have developed a method for detection of horizontal gene transfers events in evolutionary data. The new method exploits the discrepancy between the species and gene trees built for the same set of observed species to map the gene tree into the species tree and then estimate the possibility of a horizontal gene transfer for each pair of branches of the species tree. As result, the new method gives an ordered list of the horizontal gene transfers between branches of the species tree. Entries of this list should be carefully analyzed using all available information about data in hand to select the gene transfers to be represented as final solution. Any gene transfer branch added to the species phylogeny aids to resolve a discrepancy between it and the gene tree. 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 lateral gene transfers in real data sets. In this paper a model based on the least-squares was considered. It would be interesting to extend and test this procedure in the framework of the maximum likelihood and maximum parsimony models. Future developments of this method allowing for different scenarios for addition of several horizontal gene transfers at the same time will be also necessary. The method for detection of horizontal gene transfers events described in this paper will be included (for the May 2003 release) in the *T-Rex* (tree and reticulogram reconstruction) package (see [16] for more detail). The *T-Rex* program, which is implemented for Windows and

Macintosh plat-forms, is freely available for researchers at the following URL: <http://www.fas.umontreal.ca/biol/casgrain/en/labo/t-rex>.

Acknowledgements

The authors are grateful to Dr. Charles F. Delwiche for help in gathering and analyzing species and gene data considered in this study. This research was supported by the Natural Sciences and Engineering Research Council of Canada research grants to Vladimir Makarenkov, OGP 249644.

References

1. Bryant, D., and P. Waddell.: Rapid evaluation of least-squares and minimum evolution criteria on phylogenetic trees. Mol. Biol. Evol. **15** (1998) 1346-1359
2. Buneman, P.: A note on metric properties of trees. Jl Comb. Theory B. **17** (1974) 48-50
3. Charleston, M. A.: Jungle: a new solution to the host/parasite phylogeny reconciliation problem. Math. Biosci. **149** (1998) 191-223
4. DasGupta, B., He, X., Jiang, T., Li, M., Tromp, J. and Zhang, L.: On distances between phylogenetic trees. Proc. 8th Annual ACM-SIAM Symposium on discrete Algorithms, SODA'97 (1997) 427-436
5. Delwiche, C.F. and Palmer, J.D.: Rampant horizontal transfer and duplication of Rubisco genes in Eubacteria and Plastids. Mol. Biol. Evol. **13(6)** (1996) 873-882
6. Doolittle, W. F.: Phylogenetic classification and the universal tree. Science **284** (1999) 2124-2128
7. Felsenstein, J.: An alternating least-squares approach to inferring phylogenies from pairwise distances. Syst. Biol. **46** (1997) 101-111
8. Gascuel O.: BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data, Mol. Biol. Evol. **14(7)** (1997) 685-695
9. Guig, R. I. Muchnik, and T.F. Smith.: Reconstruction of ancient molecular phylogenies. Mol. Phyl. Evol. **6,2** (1996) 189-213
10. Hallet, M., and Lagergreen, J.: Efficient algorithms for lateral gene transfer problems. RECOMB (2001) 149-156
11. Hein, J.: A heuristic method to reconstructing the evolution of sequences subject to recombination using parsimony. Math. Biosci. (1990) 185-200
12. Hein. J., Jiang, T., Wang, L. and Zhang, K.: On the complexity of comparing evolutionnary treees. Combinatorial Pattern Matching (CPM)95, LLNCS **937** (1995) 177-190
13. Hein, J., Jiang, T., Wang, L. and Zhang, K.: On the complexity of comparing evolutionnary trees. Discr. Appl. Math. **71** (1996) 153-169
14. Maidak, B.L., Cole, J.R., Lilburn, T.G., Parker, C.T., Saxman, P.R.,Farris, R.J., Garrity, G.M., Olsen, G.J., Schmidt, T.M., and Tiedje, J.M.: The RDP-II (ribosomal database project). Nucleic Acids Research **29** (2001) 173-174
15. Makarenkov, V. and Leclerc, B.: An algorithm for the fitting of a phylogenetic tree according to a weighted least-squares criterion, J. of Classif. **16,1** (1999) 3-26
16. Makarenkov,V.: T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks. Bioinformatics, **17** (2001) 664-668

17. Olsen, G. J. and Woese, C. R.: Archael genomics an overview. Cell **89** (1997) 991-994
18. Page, R. D. M.: Maps between trees and cladistic analysis of historical associations among genes, organism and areas. Syst. Biol. **43** (1994) 58-77
19. Page, R. D. M. and Charleston, M. A.: From gene to organismal phylogeny: Reconciled trees. Bioinformatics **14** (1998) 819-820
20. Saitou, N. and Nei, M.: The neighbour-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4** (1987) 406-425
21. Sneath, P. H.: Reticulate evolution in bacteria and other organisms: How can we study it? J. Classif. **17** (2000) 159-163
22. The NCBI handbook [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2002 Oct. Chapter 17, The Reference Sequence (RefSeq) Project
23. von Haseler, A. and Churchill, G. A.: Network models for sequence evolution. J. Mol. Evol. **37** (1993) 77-85