Centre de Recherches Mathématiques
CRM Proceedings and Lecture Notes
Volume **45**, 2008

Algorithms for Detecting Complete and Partial Horizontal Gene Transfers: Theory and Practice

Vladimir Makarenkov, Alix Boc, Alpha Boubacar Diallo,
and Abdoulaye Baniré Diallo

Abstract. We describe two methods for detecting horizontal gene transfers in the framework of the complete and partial gene transfer models. In case of a complete gene transfer model a new fast backward selection algorithm for predicting horizontal gene transfer events is presented. The latter algorithm can rely either on the metric or on the topological optimization to identify horizontal gene transfers between branches of a given species phylogeny. In case of the topological optimization, we use the well-known Robison and Foulds (RF) topological distance, whereas in case of the metric optimization, the least-squares (LS) criterion is considered. We also formulate and prove the NP-hardness of the partial gene transfer problem. Second, an efficient algorithm for predicting partial transfers, using the Gauss and Seidel optimization, is discussed. We also show how to assess the reliability of a specific gene transfer or a whole gene transfer scenario. In the application section, we apply the new algorithm to detect possible gene transfers occurred during the evolution of the gene `rpl12e`.

1. Introduction

Horizontal gene transfer (HGT) is a direct transfer of genetic material from one lineage to another. The understanding that horizontal gene transfer might have played a key role in biological evolution is one of the most fundamental changes in our perception of general aspects of molecular biology in recent years [6, 18, 19]. Bacteria and Archaea have sophisticated mechanisms for the acquisition of new genes through HGT which may have been favoured by natural selection as a more rapid mechanism of adaptation than the alteration of gene functions through numerous point mutations. If the donor DNA and the recipient chromosome display some homologous sequences, the donor sequences can be stably incorporated into the recipient chromosome by homologous recombination. The three main mechanisms of HGT are the following: transformation, consisting of uptake of naked DNA from the environment; conjugation, which is mediated by conjugal plasmids or conjugal transposons; and transduction, consisting of DNA transfer by phage.

2000 *Mathematics Subject Classification.* Primary: 90C35; Secondary: 90C27.
The authors are grateful to Dr. Hervé Philippe for his help in the analysis of the `rpl2e` data.
This is the final form of the paper.

©2008 American Mathematical Society

These transferring mechanisms can introduce sequences of DNA that display little similarity with the remaining DNA of the recipient cell [6].

There are a few ways to identify the genes that have been transferred horizontally. First, sequence analysis of the host genome may reveal areas with GC content or codon usage patterns atypical to it [17]. Second, if a sequence is found in only one organism and is absent from all other closely related organisms, it is more likely that it has been introduced horizontally into this organism rather than deleted from all the others. Third, the comparison of a morphology-based species tree or a molecular tree based on a molecule that is assumed to be refractory to horizontal gene transfer (e.g. 16S rRNA or 23S rRNA) against a phylogeny of an observed gene may reveal topological conflicts which can be explained by horizontal transfers.

Several attempts to use network-based models to depict horizontal gene transfers can be found (see for example: [35, 29, 4, 10], or [11]). A model of horizontal gene transfer that maps gene phylogenies into a species tree has been introduced by [10]. Mirkin *et al.* [25] and Hallett *et al.* [11] have developed algorithms allowing for simultaneous identification of gene duplications, gene losses, and horizontal gene transfers. The papers by Moret *et al.* [27, 28] give an overview of the network modeling in phylogenetics. In a recent paper published in the SFC2004 proceedings, [26] considered some approaches for biologically meaningful mapping of data of individual gene families into an evolutionary species tree. One approach first produces a gene tree, then maps it into the species tree, whereas the other approach first takes the gene phyletic profile, maps it into the species tree and then tunes it into a directed scenario based on the similarity data.

In this article we continue the work started in Ref. [3], where we described a HGT model based on least-squares, and in Ref. [22], where we showed the difference between complete and partial gene transfer models. First, we describe a polynomial-time HGT algorithm for the detection of complete transfers and test it with respect to the two optimization criteria: Least-squares (LS) and Robinson and Foulds (RF) topological distance. We also suggest how to assess the reliability of horizontal gene transfers identified by our algorithm. In the application section, we show how the new algorithm predicts transfers of the gene `rpl2e` for the group of 14 Archaea organisms which were originally examined in Ref. [24].

2. Algorithms for Predicting Horizontal Gene Transfers

2.1. Basic definitions. We start this section with some basic definitions about phylogenetic trees and tree metrics, generally following the terminology of Barthélemy and Guénoche [1, 2]. The distance $\delta(x, y)$ between two vertices x and y in a phylogenetic (i.e. additive) tree T is defined as the sum of the edge lengths in the unique path linking x and y in T. Such a path is denoted (x, y). A leaf is a vertex of degree one.

Definition 1. Let X be a finite set of n taxa. A *dissimilarity* d on X is a non-negative function on $(X \times X)$ such that for any x, y from X:

(1) $d(x, y) = d(y, x)$, and
(2) $d(x, y) = d(y, x) \geq d(x, x) = 0$.

	x2	x3	x4	x5
x1	6	6	4	2
x2		2	4	6
x3			4	6
x4				4

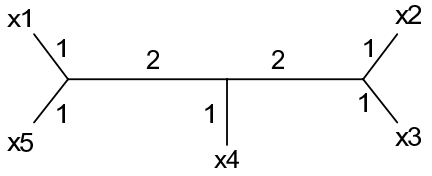

FIGURE 1. An example of a tree metric on the set X of 5 taxa.

Definition 2. A dissimilarity d on X satisfies the *four-point condition* if for any x, y, z, and w from X:

$$d(x,y) + d(z,w) \leq \text{Max}\{d(x,z) + d(y,w); d(x,w) + d(y,z)\}.$$

Definition 3. For a finite set X, a **phylogenetic tree** (i.e. an additive tree or a X-tree) is an ordered pair (T, ϕ) consisting of a tree T, with vertex set V, and a map ϕ :X$\rightarrow V$ with the property that, for all $x \in X$ with degree at most two, $x \in \phi(X)$. A phylogenetic tree is **binary** if ϕ is a bijection from X into the leaf set of T and every interior vertex has degree three.

The main theorem relating the four-point condition and dissimilarity representability by a phylogenetic tree (i.e., phylogeny) is as follows:

Theorem 2.1 (Zarestskii, Buneman, Patrinos & Hakimi, Dobson). *Any dissimilarity satisfying the four-point condition can be represented by a phylogenetic tree such that for any x, y from X, $d(x,y)$ is equal to the length of the path linking the leaves x and y in T. This dissimilarity is called a* tree metric. *Furthermore, this tree is unique.*

Figure 1 is an example of a tree metric on the set X of 5 taxa and the associated phylogenetic tree.

2.2. Optimization criteria. Here we present a fast greedy algorithm for predicting complete horizontal gene transfers. The algorithm for identifying HGTs proceeds by a progressive reconciliation of the given species and gene phylogenetic trees, denoted T and T' respectively. Usually, the species tree T is inferred from the genes that are refractory to horizontal gene transfer and genetic recombination (e.g., 16sRNA sequences). This tree represents the direct or tree-like evolution. The gene tree T' represents the evolution of a given gene which is supposed to undergo horizontal transfers.

At each step of the algorithm, all pairs of branches in T are tested against the hypothesis that a horizontal gene transfer has occurred between them. The considered HGT model assumes that the transferred gene supplants the entire homologous gene of the host or that the homologous gene is simply absent at the host genome. In such a model, the original species phylogenetic tree T is gradually transformed into the gene phylogenetic tree T' through a series of subtree moves (i.e., gene transfers or HGTs). The topology of the gene tree T' is kept fixed. The goal is to find the minimum possible sequence of trees $T, T_1, T_2, \ldots, T'$ that transforms T into T'. Obviously, a number of necessary biological rules should be taken into account. For instance, the transfers within the same lineage as well as some double-crossing transfers should be prohibited (for more detail, see [20, 30, 31, 10]).

We consider two optimization criteria which can be used at each algorithmic step to select the best HGT. The first optimization criterion that we consider is the *least-squares* (LS) *function* Q. It is computed as follows:

$$Q = \sum_i \sum_j \big(d(i,j) - \delta(i,j)\big)^2, \tag{2.1}$$

where $d(i,j)$ is the pairwise distance between the leaves i and j in the species tree T (or in the tree T_1 obtained from T after the first subtree move) and $\delta(i,j)$ the pairwise distance between i and j in the gene tree T'. The second criterion that can be useful for assessing discrepancy between the species and gene phylogenies is the Robinson and Foulds (RF) topological distance. The RF metric (Robinson and Foulds 1981) is an important and frequently used tool to compare the topologies of phylogenetic trees. This distance is equal to the minimum number of elementary operations, consisting of merging and splitting nodes, necessary to transform one tree into the other. This distance is also the number of bipartitions or Buneman's splits belonging to exactly one of the two trees. When the RF distance is considered, we can use it as an optimization criterion as follows: all possible transformations of the species tree, consisting of transferring one of its subtrees from one branch to another, are evaluated in a way that the RF distance between the transformed species tree T_1 and the gene tree T' is computed. The subtree transfer providing the minimum of the RF distance between T_1 and T' is retained. Note that the problem asking to find the minimum number of subtree transfer operations necessary to transform one tree into another (i.e. also known as *Subtree Transfer Problem*) *has been shown to be* NP*-hard* [12].

2.3. Greedy backward algorithm for predicting complete horizontal gene transfers. In this section we discuss the main features of our algorithm based on the backward selection of horizontal gene transfers. Consider a gene transfer in the species tree T going from b to a and transforming it into the tree T_1 (Fig. 2). The following timing constraint is considered (see also Ref. [22]): to allow the transfer between the branches (z,w) and (x,y) of the species tree T, the cluster combining the subtrees rooted by the vertices y and w must be present in the gene tree T'. Such a constraint enables us, first, to arrange the topological conflicts between T

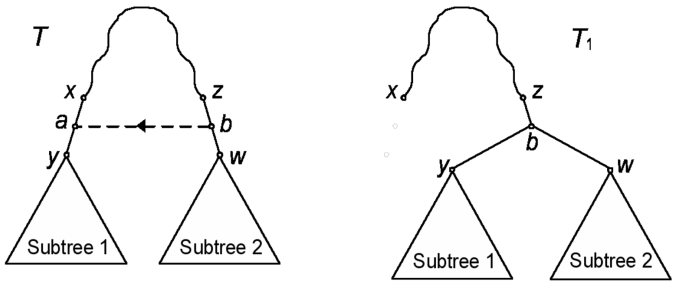

FIGURE 2. Subtree constraint: the transfer between the branches (z,w) and (x,y) of the species tree T can be allowed if and only if the cluster regrouping both affected subtrees is present in the gene tree; here, a single branch is depicted by a plane line and a path is depicted by a wavy line.

and T' that are due to the transfers between single species or their close ancestors and, second, to identify the transfers that have occurred deeper in the phylogeny (i.e., closer to the tree root). The usage of this constraint allows the method to follow the order that is opposite to the order of evolution and infer first the most recent HGTs which are easier to detect.

Proposition 2.1. *If all bipartitions corresponding to the branches of the path (x, z) in the transformed species tree T_1 (Fig. 2) can be found in the bipartition table of the gene tree T', then the transfer from b to a, transforming the species tree T into T_1, is a part of a minimum cost* HGT *scenario transforming T into T'.*

This Proposition can be easily proved by induction on the number of branches of the path (x, z).

The main steps of the HGT detection algorithm are the following:

Preliminary step. Infer species and gene phylogenies, denoted respectively T and T', whose leaves are labeled by the same set of n taxa. Both species and gene trees must be rooted. If there exist identical subtrees with two or more leaves belonging to both T and T', reduce the size of the problem by replacing these subtrees with the same auxiliary taxa in both T and T'.

Step 1 ($\ldots k$). Test all possible HGTs between pairs of branches in T_{k-1} ($T_{k-1} = T$ at Step 1) except the transfers between adjacent branches and those violating the evolutionary and subtree constraints. If no such a transfer exists, relax the subtree constraint. In our simulations described in the section Simulation study, this relaxation was necessary on average in 1.2% of cases. Search for the transfers satisfying the conditions of Proposition 2.1. If no such transfers exist, choose the best HGT with respect to the selected optimization criterion that can be in our case: the least-squares (LS) or the Robinson and Foulds (RF) metric. Reduce the size of the problem by contracting the newly-formed subtree in the transformed species tree T_k and the gene tree T'. In the list of the obtained HGTs, search for and eliminate the idle transfers using a backward procedure. An idle transfer is the transfer whose removal does not change the topology of the tree T_k.

Stopping condition and time complexity. The procedure stops when the LS or RF coefficient equals zero. Such a computation requires $O(kn^4)$ time to generate k transfers in a phylogenetic tree with n leaves. However, because of the progressive size reduction of the species and gene trees, the practical time complexity of this algorithm is rather $O(kn^3)$.

Proposition 2.2. *If the subtree constraint is not relaxed, the* HGT *detection algorithm requires at most $n - 3$ steps to transform a binary species tree with n leaves into a binary gene tree with the same set of n leaves.*

The proof of this proposition is based on the fact that the maximum value of the RF distance between two binary trees with n leaves is $2n - 6$ and that each subtree transfer satisfying the subtree constraint decreases the value of the RF distance by at least 2.

2.4. Partial gene transfer model. The partial gene transfer model is more general, but also more complex and challenging. It presumes that only a part of the transferred gene has been acquired by the host species through the process

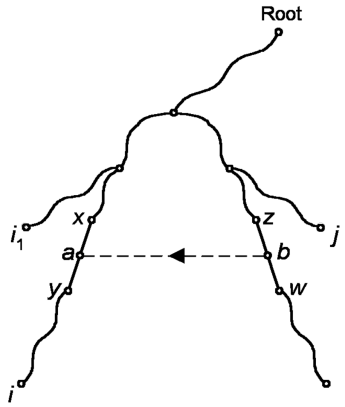

FIGURE 3. Evolutionary distance between the taxa i and j can be allowed to change after the addition of the branch (b, a) representing a partial HGT between the branches (z, w) and (x, y). Evolutionary distance between the taxa i_1 and j must not be affected by the addition of (b, a).

of homologous recombination [22]. This means that the traditional species phylogenetic tree is transformed into a directed phylogenetic network (i.e. a directed connected graph). For example, Denamur *et al.* [5] proposed a method to identify gene segments being transferred horizontally. This method was applied to detect partial HGTs of the mut$U$ and mut$S$ genes within E. coli evolutionary trees. Because many analyzes are now directed at understanding the evolution of complete genomes, the partial gene transfer model could be also useful if one wanted to model the transfer of a portion of a genome.

In a phylogenetic tree, there is always a unique path connecting a pair of nodes. Adding to it a HGT branch creates an extra path between certain nodes. Figure 3 illustrates the case where the evolutionary distance between the taxa i and j can be affected by the addition of the HGT branch (b, a) representing partial gene transfer from b to a. It is relevant to assume that the HGT from b to a can affect the evolutionary distance between the taxa i and j if and only if the destination point a is located on the path between i and the root of the tree; the position of j is fixed. Thus, in the reticulate phylogeny T in Fig. 3 the evolutionary distance $d_1(i, j)$ between the taxa i and j can be computed as follows:

$$d_1(i,j) = (1-\alpha)d(i,j) + \alpha\big(d(i,a) + d(j,b)\big), \tag{2.2}$$

where α indicates the fraction, unknown in advance, of the transferred gene and d is the internode distance in the species tree before the addition of the HGT branch (b, a).

On the contrary, the distance between the taxa i_1 and j (Fig. 3) must not be affected by the addition of (b, a). Figure 4 illustrates the other cases where the addition of a HGT branch must not affect the length of the evolutionary path between i and j.

The least-squares loss function Q to be minimized with the *unknown vector of edge lengths* ℓ in T and the *unknown fraction of the transferred gene* α is as follows:

$$
\begin{aligned}
(2.3)\quad & Q(L,\alpha) \\
&= \sum_{ij\in S}\left((1-\dot{a})\sum_{k\in \mathrm{path}(ij)}\ell_{ij}^{k}+\alpha\left(\sum_{k\in \mathrm{path}(ia)}\ell_{ia}^{k}+\sum_{k\in \mathrm{path}(jb)}\ell_{jb}^{k}\right)-\delta(i,j)\right)^{2} \\
&\qquad +\sum_{ij\notin S}\left(\sum_{k\in \mathrm{path}(ij)}\ell_{ij}^{k}-\delta(i,j)\right)^{2}\longrightarrow \min,
\end{aligned}
$$

where $\delta(i,j)$ is the given gene dissimilarity between i and j; ℓ_{ij}^{k} is the length of the branch k of the path (ij) in T; α is the fraction of the transferred gene $(0\leq\alpha\leq 1)$; and S is the set of pairs of taxa $\{ij\}$ such that the transfer (ba) can affect the evolutionary distance between them.

To show the NP-hardness of the least-squares optimization in the context of the partial gene transfer the following problem can be stated:

Given: Species phylogenetic tree T (with the associated tree metric d on the set of taxa X), gene dissimilarity $\boldsymbol{\delta}$ on X, and a fixed non-negative value ε.

Problem. Find the minimum number of partial gene transfers k such that:

$$
(2.4)\qquad Q=\sum_{i}\sum_{j}\big(d_k(i,j)-\delta(i,j)\big)^{2}\leq\varepsilon,
$$

where $d_k(i,j)$ is the network distance between i and j, computed using Formulae 2.2 and 2.3, in the phylogenetic network T_k obtained from T after the addition of k partial gene transfers.

Theorem 2.2. *The minimum number of partial gene transfer problem* (*MNPGT problem*) *is* NP*-hard.*

The proof of this theorem is based on a polynomial-time reduction from the *Subtree Transfer Problem* (STR problem) that consists of finding the minimum number of complete gene transfers to transform a given species tree T into a given gene tree T'. The STR problem is identical to the problem of adding to T the minimum number of complete gene transfers such that $Q=\sum_i\sum_j\big(d_k(i,j)-\delta(i,j)\big)^2\leq 0$ (i.e., the case of $\varepsilon=0$ is considered), where $d_k(i,j)$ is the pairwise distance between i and j in the phylogenetic tree (i.e., a particular case of a phylogenetic network). Here, the tree T_k is obtained from T after the addition of k complete gene transfers (i.e., a particular case of a partial transfer) and $\delta(i,j)$ is the given tree metric associated with T'.

Several important timing constraints have to be incorporated into this model, in addition to those taken into account in the complete HTS model, to identify the interactions between HGTs that are not intelligible from an evolutionary point of view. Some of these constraints, but not all of them, were initially pointed out by Page and Charleston [30, 31]. For instance, double-crossing transfers between two lineages (Figs. 5a and b) must be forbidden. In this case, the HGT events affect the ancestor of the species from the previous transfer. Making the source and destination lineages contemporaneous for one HGT makes the other transfer impossible (Fig. 5).

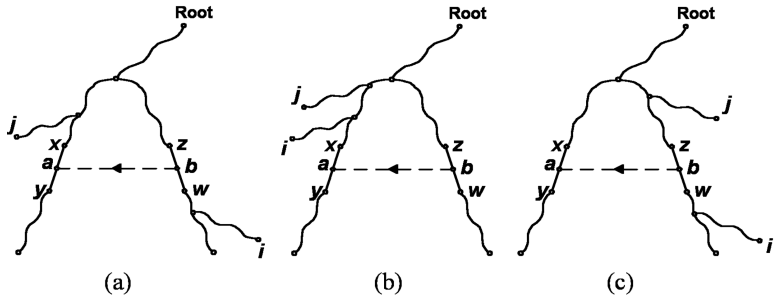

FIGURE 4. Three situations when the evolutionary distance between the taxa i and j must not be affected by the addition of the new branch (b, a) representing a partial HGT between the branches (z, w) and (x, y). Path between the taxa i and j cannot to go through the branch (b, a).

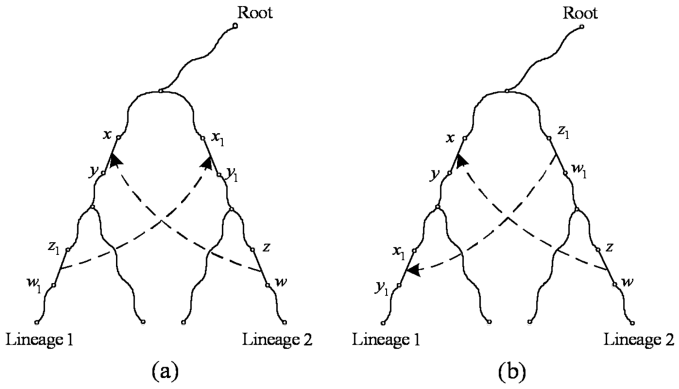

FIGURE 5. Transfers between two lineages crossing in such ways must be prohibited.

Note that the rule illustrated in Figure 5a is automatically taken into account in the complete gene transfer model, where its violation would be equivalent to the violation of the same lineage constraint (see Page and Charleston [30, 31]). For instance (Figure 5a), the HGT from (z, w) to (x, y) cannot be followed by the transfer from (z_1, w_1) to (x_1, y_1) because after the first HGT the branches (z_1, w_1) and (x_1, y_1) will be located on the same lineage (Lineage 2). We also identify two cases, where the evolutionary distance between the taxa i and j can be affected by multiple transfers (Figures 6a and b); and, two cases, where this distance must not be affected by them (Figures 6c and d). Failure to take these constraints into account can result in postulating transfers that are mutually incompatible.

Assume that a partial gene transfer between the branches (z, w) and (x, y) (i.e., from b to a in Fig. 3) of the species tree T has taken place. The lengths of all branches in T are reassessed in the least-squares sense after the addition of (b, a), whereas the length of (b, a) is assumed to be 0. To reassess the branch lengths of T, we have first to make an assumption about the value of the parameter α (eq. 2.2),

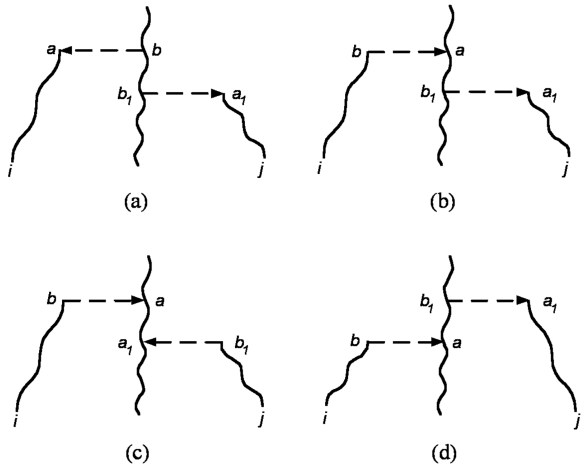

FIGURE 6. Cases (a) and (b): evolutionary path between the taxa i and j can go through both HGT branches (b, a) and (b_1, a_1). Cases (c) and (d): evolutionary path between the taxa i and j cannot go through both HGT branches (b,a) and (b_1, a_1).

indicating the gene fraction being transferred. This parameter can be estimated either by comparing sequence data corresponding to the subtrees rooted by the vertices y and w, or different values of α can be tested in the optimization problem.

Fixing the parameter α, we reduce to a linear system the system of equations establishing the correspondence between the experimental gene distances and the path-length distances in the HGT network. This system having generally more variables (i.e. branch lengths of T) than equations (i.e. pairwise distances in T; the number of equations is always $n(n-1)/2$ for n taxa) can be solved by approximation in the least-squares sense. Let us now show how the approximation problem can be stated and efficiently solved.

Let $\mathbf{A}_\alpha$ be the matrix of dimension $n(n-1)/2 \times m$, each row of which is associated with one pair of taxa of X, where n is the number of taxa and m is a number of edges in T. The value $a_{ij,e}$ of this matrix corresponding to the pair of taxa ij and the edge e is equal either to 1, or to α, or to $1-\alpha$ if the edge e is in the path (ij) in T, and is equal to 0 if not. Let ℓ be the vector of edge lengths of m elements and $\mathbf{d}$ be given vector of gene distances of $n(n-1)/2$ elements.

Fixing the value of α (e.g., values 0, 0.1, 0.2, ..., and 1.0 can be tested in turn), we obtain a linear system of $n(n-1)/2$ equations with m unknowns: $\mathbf{A}_\alpha \times \ell = \mathbf{d}$.

When $n \geq 4$, this system has more equations than unknowns. It can be solved by approximation in the least-squares sense:

$$(\mathbf{A} \times \ell - \mathbf{d})^2 \to \min\,. \tag{2.5}$$

After taking the gradient we have:

$$\mathbf{A}_\alpha^t \times (\mathbf{A}_\alpha \times l - \mathbf{d}) = 0. \tag{2.6}$$

Following algebraic manipulations, we obtain:

$$\mathbf{A}_\alpha^t \times \mathbf{A}_\alpha \times l = \mathbf{A}_\alpha^t \times \mathbf{d}. \tag{2.7}$$

Thus, we have: $\mathbf{B} \times \ell = \mathbf{c}$, where $\mathbf{B}$ is a $(m \times m)$ matrix, and $\mathbf{c}$ is a vector with m components.

Following Barthélemy and Guénoche [1] and Makarenkov and Leclerc [23], we apply a slightly modified Gauss-Seidel method to solve the above system. The method consists of decomposing $\mathbf{B}$ into its diagonal ($\mathbf{\Delta}$), its strictly upper triangular component ($-\mathbf{F}$), and its strictly lower triangular component ($-\mathbf{E}$):

$$\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \dots & \dots & \dots & \dots \\ b_{m1} & b_{m2} & \dots & b_{mm} \end{pmatrix} = \begin{pmatrix} & \mathbf{\Delta} & -\mathbf{F} \\ -\mathbf{E} & & \end{pmatrix} = \mathbf{\Delta} - \mathbf{E} - \mathbf{F}. \tag{2.8}$$

Then, we apply the iterative procedure:

$$\mathbf{\Delta} \times \ell^{k+1} = \mathbf{E} \times \ell^{k+1} + \mathbf{F} \times \ell^{(k)} + \mathbf{c}, \tag{2.9}$$

which allows us to compute gradually the components of the vector $\ell(j)^{(k+1)}$, corresponding to the edge lengths at the $k+1$th iteration, from those of $\ell(j)^k$. If the computed value of $\ell(j)^{(k+1)}$ is negative, it is replaced with the value 0. This operation is equivalent to the projection on the cone $\mathbf{L} \geq 0$, which ensures an appropriate solution.

The exact equation used in this method is the following for all $j = 1, 2, \dots, m$:

$$\ell(j)^{(k+1)} = \left(- \left(\sum_{j+1 \leq i \leq m} b_{ij} \ell(j)^{(k)} \right) - \left(\sum_{1 \leq i \leq j-1} b_{ij} \ell(j)^{(k+1)} \right) + c_j \right) \Big/ b_{jj}. \tag{2.10}$$

Thus, the main steps of the partial gene transfer algorithm can be stated as follows:

Preliminary step. This step corresponds to the preliminary step discussed in the context of the complete gene transfer model. It consists of inferring the species and gene phylogenies denoted respectively T and T' whose leaves are labeled by the same set X of n taxa. Because the classical Robinson and Foulds distance is defined only for tree topologies, we use the least-squares as a unique optimization criterion when modeling partial HGTs.

Step 2. Test all connections between pairs of branches in the species tree $\mathbf{T}$. For each HGT connexion satisfying evolutionary constraints, carry out the following optimization:

(a) Fix the value of the fraction of the gene being transferred α (e.g., one can try in turn the values of 0, 0.1, 0.2, ..., and 1.0). Compute using the Gauss–Seidel method the optimal lengths $\mathbf{l}$ of the edges in the species tree (or network, starting from Step 2) $\mathbf{T}$.

(b) Go back to the original equation system: $\mathbf{A}_\alpha \times \mathbf{l} = \mathbf{d}$. Fix the values of the vector $\mathbf{l}$ found using the Gauss-Seidel method and solve this problem by least-squares considering as unknown the parameter α.

(c) Then, fix the optimal value of α found and repeat the computation until both unknown parameters l and α converge to a certain solution.

All eligible pairs of branches in T can be processed in this way. The HGT connection providing the smallest value of the LS coefficient Q and satisfying the defined evolutionary constraints should be selected for the addition to the species tree T, transforming it into a phylogenetic network.

Step 3 $(2, \ldots, k)$**.** Run the algorithm until a fixed number k of partial gene transfers is found and added to T or the value of the LS criterion Q is lower than a pre-established threshold ε.

Time complexity of this algorithm is $O(kn^5)$ to add k partial horizontal gene transfers to the species tree with n leaves.

2.5. Bootstrap validation of horizontal gene transfers. Bootstrap analysis can be used to place confidence intervals on internal branches of evolutionary trees [7]. We designed a bootstrap validation procedure for computing the bootstrap scores either for a specific gene transfer or a whole gene transfer scenario. The following strategy was adopted to assess the reliability of obtained HGTs. Because we are mostly interested in the evolution of a given gene or a group of genes, the sequences used to build the species tree are not resampled. The species tree is taken as an *a priori* assumption of the method and held constant. The sequence data used to build the gene tree are drawn with replacement in order to create a series of pseudo-replicates. The HGT detection algorithm is then carried out on the bootstrapped pseudo-replicates. Thus, for all HGT branches appearing in the original scenario, we verify if they appear in the obtained transfer scenarios, using as input the original species tree and the gene tree inferred from the sets of pseudo-replicates. It is worth noting that among resampled datasets only those that give rise to a gene phylogenetic tree such that it contains the root branch separating this tree into exactly the same bipartition sets as the root branch of the original gene tree does, are eligible for the HGT bootstrap analysis.

Simulation study. A Monte Carlo study was conducted to test the ability of the new method to recover correct gene transfers. In the framework of *the complete* HGT *model only* we examined how the detection procedure performed depending on the model of sequence evolution, number of observed species, and sequence length. The results illustrated in Figs. 7 and 8, and reported in Tables 1 and 2 (see Appendix) were obtained from simulations carried out with random binary phylogenetic trees with 8, 16, 24, 32, 48, and 64 leaves, whereas the sequence length varied from 125 to 1000 sites. The simulation procedure consisted of the five basic steps described below:

1. A true tree topology, denoted T, was obtained using the random tree generation procedure proposed by Kuhner and Felsenstein [16]. The branch lengths of T were computed using an exponential distribution. Following the approach of Guindon and Gascuel [9], we added some noise to the branches of the true phylogenies to create a deviation from the molecular clock hypothesis. All the branch lengths of T were multiplied by $1 + \alpha x$, where the variable x was obtained from a standard exponential distribution $\big(P(x > k) = \exp(-k)\big)$, where the constant a was a tuning factor for the deviation intensity. Following Guindon and Gascuel [9], a was fixed to 0.8. The random trees generated by this procedure are chosen to have the depth of $O\big(\log(n)\big)$, where n is the number of species (i.e. number of leaves in a binary phylogenetic tree).

2. Each random phylogeny was then submitted to the SeqGen program [32] to simulate sequence evolution along its branches according to the Jukes and Cantor [14], Kimura 2-parameter [15], and Jin–Nei Gamma[13] models.

3. To assess the quality of HGT detection by the new method, we developed a simulation program using the results of SeqGen. For each considered rooted tree,

viewed as an organismal phylogeny, our program created one random horizontal gene transfer that respected the evolutionary constraints discussed in the algorithmic section. During this operation, the program regenerated the DNA sequences for each tree node located in the subtree affected by the HGT. As the simulations were carried out for the complete gene transfer model, the HGT destination sequence was set identical to the source sequence and the new sequences were regenerated from it according to the selected evolutionary model.

4. The sequence to distance transformation corresponding to the considered model of evolution was then applied to the DNA sequences associated with the leaves of the phylogeny affected by the gene transfer. The NJ method [34] was used to infer the gene trees from the obtained distance matrix. The topology of the organismal phylogeny (i.e. true tree T) was supposed to be known.

5. The HGT detection method was then carried out to infer the transfer. The experiments were conducted using the procedures based on the RF and LS optimization. The simulations were carried out for 500 random rooted phylogenies with 8 and 16 leaves and 100 random rooted phylogenies with 24 to 64 leaves.

Figures 7 and 8 present the average simulation results obtained for random phylogenies with 8 to 64 leaves, using as optimization criteria the RF topological distance and LS function, respectively. These figures illustrate how the detection rate changes as the number of sites varies from 125 to 1000. As expected, the detection rate grows as the number of sites increases and the number of species decreases. Note that for the phylogenies with 8 to 32 leaves the best results were obtained under the Kumura and Jukes–Cantor models. For the phylogenies with 48 to 64 species the best performances were regularly obtained under the Kimura model, whereas the results found under the Jukes–Cantor model were the worst of the three evolutionary models.

This trend can be observed in the case of both optimization criteria. Obviously, with the short sequences we have a bigger phylogenetic error that can either appear like a HGT, when it does not occur, or disguise a real HGT. Tables 1 and 2 (see Appendix) report the false positive and false negative (indicated in parentheses) detection rates obtained using as optimization criteria the RF distance and LS function, respectively. A false positive HGT is an incorrect transfer found by the algorithm and a false negative HGT is the right transfer that has not been detected. A false positive HGT will always occur if the gene tree inferred by NJ (see Step 4 above) is different from the true gene tree (see Step 3 above), but it can also take place when both trees are identical but a transfer going to the direction opposite to the correct HGT disguises it, leading to the same gene tree (see [20]).

False negative HGTs are mostly due to the error of inferring the gene tree, but can also happen when a transfer going to the opposite direction disguises the correct HGT. As defined, the false positive detection rate is always bigger or equal to the negative one. The analysis of Tables 1 and 2 shows that the false negative rate is almost as big as the false positive rate when the tests were conducted with large phylogenies (48 and 64 species) and short sequences (125 and 250 sites). The false negative rate was noticeably lower than the false positive one in the case of the large phylogenies and long sequences. Furthermore, we have measured the recovery rates for the HGT source, destination, and source and destination combined (i.e. the latter parameter corresponds to the detection rate depicted in Figs. 7 and 8). These tests were carried out under the Jukes and Cantor model of sequence evolution and

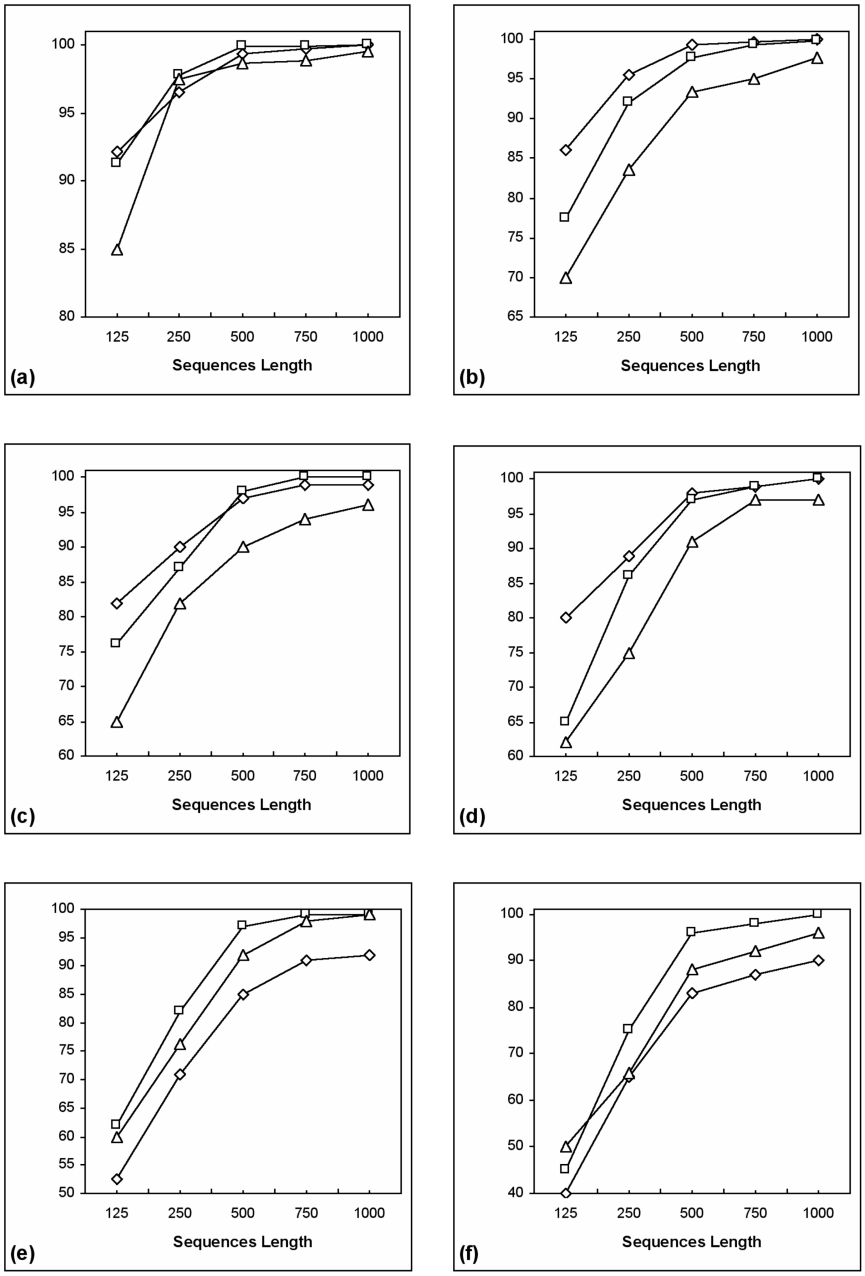

FIGURE 7. HGT detection rates obtained for random phylogenies with 8 to 64 leaves ($8-a$, $16-b$, $24-b$, $32-d$, $48-e$, $64-f$) using the RF topological distance for optimization. Jukes and Cantor ($\Diamond$), Kimura 2-parameter ($\Box$), and Jin–Nei Gamma (Δ) models were used for the tree generation.

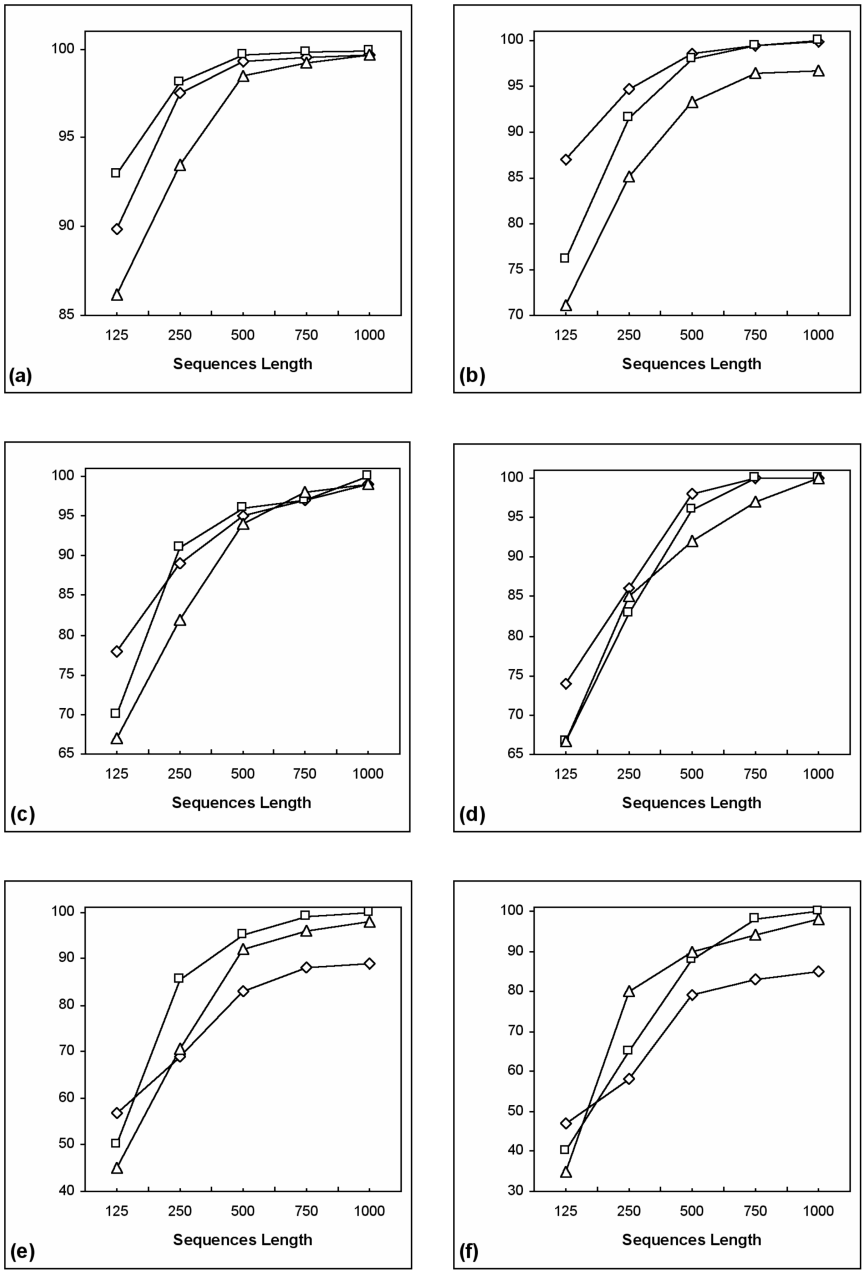

FIGURE 8. HGT detection rates obtained for random phylogenies with 8 to 64 leaves ($8-a$, $16-b$, $24-b$, $32-d$, $48-e$, $64-f$) using the LS function for optimization. Jukes and Cantor ($\Diamond$), Kimura 2-parameter ($\Box$), and Jin–Nei Gamma (Δ) models were used for the tree generation.

using the RF distance for the algorithmic optimization. Note that the transfer destinations were generally better detectable than their sources. The difference in the source-destination detection was more important for the short sequence. For example, for the sequences with 125 sites it varied, on average, from 6% (for 8 species) to 1% (for 64 species). However, for the longer sequences the source and destination rates were very similar.

Generally, the procedure based on the RF distance provided better results than that based on the LS function. Nevertheless, some noticeable exceptions (e.g. under the Kimura model for the phylogenies with 8 leaves or under the Jin–Nei model in the case of the short sequences) can be pointed out. The simulation study suggested that the accuracy of the transfer detection is highly dependable on the model of sequence evolution, number of considered species, and length of observed sequences.

Results and discussion. [Detecting horizontal transfers of the gene `rpl2e`] We first tested our algorithm on the phylogeny of 14 species of Archaea originally considered by Matte-Tailliez *et al* [24]. The latter authors discuss problems encountered when reconstructing some parts of the archaeal phylogeny, pointing out the evidence of HGT events perturbing the evolution of a number of considered genes. Matte-Tailliez *et al.* inferred the maximum likelihood tree (Figure 10, undirected lines) based on the concatenated 53 ribosomal proteins (7,175 positions) and compared it to the maximum likelihood phylogeny of the gene `rpl2e` (Figure 9) built for the same 14 organisms. The calculations of the best ML tree and its branch lengths for the 53 concatenated proteins were conducted using the PUZZLE program with Γ-law correction.

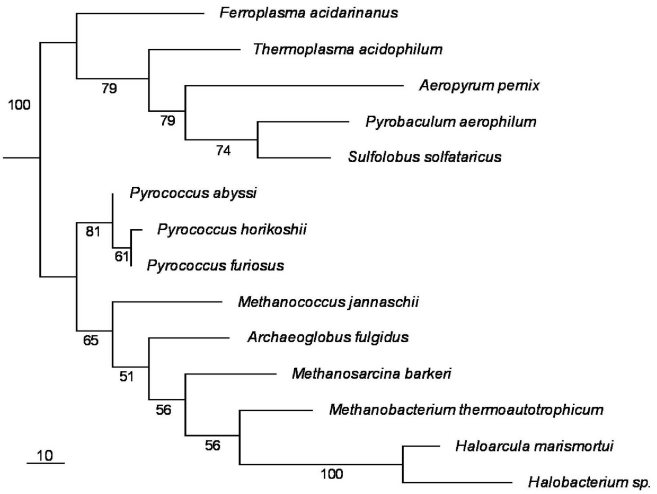

FIGURE 9. Maximum likelihood phylogenetic tree for the protein `rpl2e` (89 positions). Numbers close to branches are ML bootstrap scores obtained from the sampled protein sequences using the SeqBoot and Proml (JTT model) programs from the PHYLIP package (Felsenstein, 1989). Its topology is identical to the tree found by Matte-Taillez *et al* [24, Figure 3].

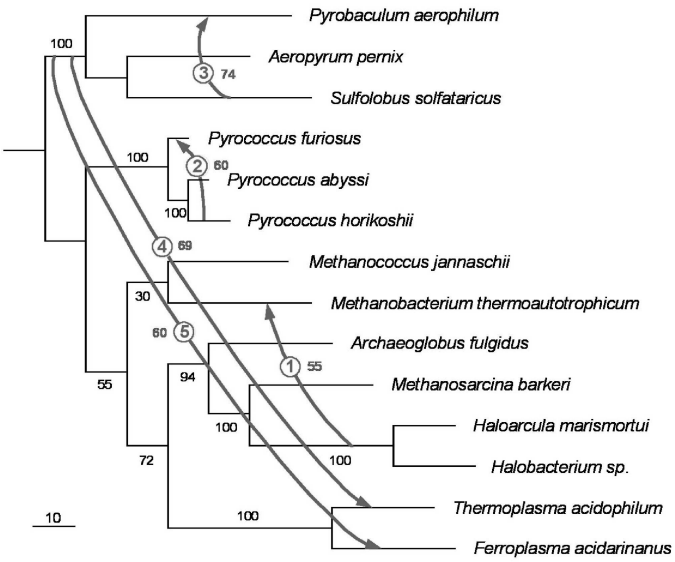

FIGURE 10. Species tree (Matte-Taillez *et al.* [24, Fig. 1a], with five reconciliation branches (denoted by arrows). Numbers close to branches are ML bootstrap scores computed by the RELL method upon 2,000 top-ranking trees using the *MOLPHY* program without correction for among-site variation. Numbers on HGT arrows indicate their order of appearance in the unique gene transfer scenario found by the HGT detection method. Bootstrap scores for transfers are indicated by numbers close to arrow circles. Arrows 4 and 5 depict the HGTs between the clades of *Thermoplasmatales* and *Crenarchaeota* also predicted by Matte-Taillez *et al* [24].

Given the topological incongruence of the obtained phylogenies, the authors hypothesized a few cases of lateral transfers of the gene `rpl2e`. More precisely, the case of the transfer between the clades of *Thermoplasmatales* (*Ferroplasma acidarmanus* and *Thermoplasma acidophilum*) and *Crenarchaeota* (*Aeropyrum pernix*, *Pyrobaculum aerophilum* and *Sulfolobus solfataricus*) was indicated as the most evident one.

In order to apply our method, we first reconstructed from the original sequences the topologies of the gene (Figure 9) and species trees (Figure 10, undirected lines). The computations were conducted in the framework of the complete gene transfer model, using the RF optimization and subtree constraint options (Figure 2). Five directed branches needed to reconcile the species and gene topologies have been found (Figure 10). The connection representing the transfer between the cluster of *Halobacterium sp.* and *Haloarcula marismortui* and the species *Methanobacterium thermoautotrophicum* was found in the first iteration. This transfer provided the biggest drop of the RF distance between the species and gene phylogenies; its bootstrap score is 55%.

In the second and third iterations, we found the reconciliation branches between the species *Pyrococcus horikoshii* and *Pyrococcus furiosus* and between *Sulfolobus*

solfataricus and *Pyrobaculum aerophilum*. Both of these reconciliation branches link closely related species. Such kind of connections may be due to HGT as well as to local topological rearrangements necessary because of the tree reconstruction artifacts (e.g. attraction of long branches, unequal evolutionary rates, etc). The transfer branches 4 and 5 linking the cluster of *Crenarchaeota* to the species *Thermoplasma acidophilum* and *Ferroplasma acidarmanus* can be interpreted as HGT events that might have taken place between *Thermoplasmatales* and *Crenarchaeota*.

In the second and third iterations, we found the reconciliation branches between the species *Pyrococcus horikoshii* and *Pyrococcus furiosus* and between *Sulfolobus solfataricus* and *Pyrobaculum aerophilum*. Both of these reconciliation branches link closely related species. Such kind of connections may be due to HGT as well as to local topological rearrangements necessary because of the tree reconstruction artifacts (e.g. attraction of long branches, unequal evolutionary rates, etc). The transfer branches 4 and 5 linking the cluster of *Crenarchaeota* to the species *Thermoplasma acidophilum* and *Ferroplasma acidarmanus* can be interpreted as HGT events that might have taken place between *Thermoplasmatales* and *Crenarchaeota*.

Note, that HGT between these two groups was also predicted by Matte-Taillez *et al* [24]. In fact, the transfers 4 and 5 could consist of a unique transfer between the clades of *Thermoplasmatales* and *Crenarchaeota* that was separated into two transfers by our method due to the application of the subtree constraint (Figure 2) and the presence of the tree reconstruction artifacts. Figure 11 illustrates the evolution of the newly formed *Thermoplasmatales-Crenarchaeota* clade involving

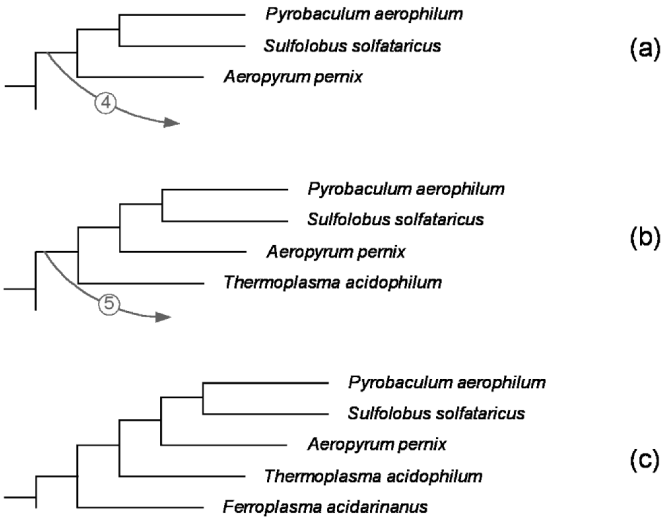

FIGURE 11. Changes in the *Crenarchaeota-Thermoplasmatales* cluster occurring after the addition of HGT branches 4 and 5. (a) This cluster after the transfer 3; the species *Thermoplasma acidophilum* joins the *Crenarchaeota* cluster. (b) This cluster after the transfer 4; the species *Ferroplasma acidarmanus* is added to the clade comprising three *Crenarchaeota* and *Thermoplasma acidophilum*. (c) This cluster after the transfer 5.

the HGTs 4 and 5. The usage of the LS criterion instead of RF leads to the solution consisting of 6 HGTs including all transfers from Figure 10 except the HGT number 2 that goes in the opposite direction. Note that a new reconciliation branch found with LS brings the species *Methanococcus jannaschii* to the cluster of 4 species including *Archaeoglobus fulgidus*. This reconciliation branch turns out to be useless and have a low bootstrap score of 14%.

3. Conlusion

We presented two polynomial-time algorithms for detecting horizontal gene transfer events. We considered the complete and partial gene transfer models, implying at each step, either the transformation of a species phylogeny into another tree or its transformation into a network structure. The algorithm for inferring complete gene transfers exploits the discrepancies between the species and gene phylogenies either to map the gene tree into the species tree by least-squares or to compute a topological distance between them and then estimate the possibility of a HGT event between each pair of branches of the species phylogeny. The models based on the optimization of the least-squares function and the Robinson and Foulds topological distance were introduced.

Inferred HGTs should be carefully analyzed using all available information about the data in hand in order to select the transfers that will be represented as a final solution. Each gene transfer branch added to the species phylogeny aids to resolve a conflict between it and the gene tree (i.e. helps to reconcile the species and gene phylogenies). A bootstrap validation procedure allowing one to assess the reliability of a specific gene transfer or whole gene transfer scenario was proposed. A comprehensive Monte Carlo study was carried out to test the ability of the new method to recover correct HGTs. It provided very encouraging results especially when the Robinson and Foulds distance was used as an optimization criterion. The example of the evolution of the gene `rpl2e` was considered in the application section. More simulation work is required to investigate the properties of the algorithm intended to infer partial gene transfers.

As any method of phylogenetic inferring, the new HGT detection method is subject to a number of artifacts which generally affect phylogenetic analysis; the main of them being: attraction of long branches, unequal evolutionary rates, and situations when the occurrence of some HGT events almost coincides with speciation events located closely to the recipient species. It is important to investigate in greater details the impact of these artifacts on the HGT detection technique introduced in this article. It would be also interesting to extend the presented model to the case, where the gene and species trees have different numbers of taxa; this situation can take place when some species have more than one copy of the gene under consideration.

The software implementing the new algorithms for detecting complete and partial horizontal gene transfers is freely available at the following URL address: `http://www.info2.uqam.ca/~boca05/software/` (this is a consol version running on the Unix and Windows platforms; it is distributed along with its `C++` source code). A graphical version of this program has been also implemented and included in the *T*-Rex web server [21] at the following URL: `http://www.trex.uqam.ca`.

References

1. J.-P. Barthélemy and A. Guénoche, *Les arbres et les représentations des proximités*, Paris, Masson, 1988.
2. ———, *Trees and proximity representations*, New York, Wiley, 1991.
3. A. Boc and V. Makarenkov, *New efficient algorithm for detection of horizontal gene transfer events*, 3rd Workshop on Algorithms in Bioinformatics (Budapest, 2003) (G. Benson and R. Page, eds.) WABI, Lecture Notes in Comput. Sci., Springer Verlag, 2003, pp. 190–201.
4. M. A. Charleston, *Jungle: a new solution to the host/parasite phylogeny reconciliation problem*, Math. Biosci. **149** (1998), 191–223.
5. E. Denamur, G. Lecointre, and P. Darlu *et al*, *Evolutionary implications of the frequent horizontal transfer of mismatch repair genes*, Cell. **103** (2000), 711–721.
6. W. F. Doolittle, *Phylogenetic classification and the universal tree*, Science **284** (1999), 2124–2129.
7. J. Felsenstein, *Confidence limits on phylogenies: an approach using the bootstrap*, Evolution **39** (1985), 738–791.
8. ———, PHYLIP—*Phylogeny inference package*, v. 3.2, Cladistics **5** (1989), 164–166.
9. S. Guindon and O. Gascuel, *Efficient biased estimation of evolutionary distances when substitution rates vary across sites*, Mol. Biol. Evol. **19** (2002), 534–543.
10. M. Hallett and J. Lagergren, *Efficient algorithms for lateral gene transfer problems*, RECOMB (N. El-Mabrouk, T. Lengauer, and D. Sankoff, eds.), ACM, New York, 2001, pp. 149–156.
11. M. Hallett, J. Lagergren, and A. Tofigh, *Simultaneous identification of duplications and lateral transfers*, RECOMB (P. E.Bourne and D. Gusfield, eds.), ACM, San Diego, 2004, pp. 347–356.
12. J. Hein, T. Jiang, L. Wang, and K. Zhang, *On the complexity of comparing evolutionary trees*, Discrete Appl. Math. **71** (1996), 153–169.
13. L. Jin and M. Nei, *Limitations of the evolutionary parsimony method of phylogenetic analysis*, Mol. Biol. Evol. **7** (1990), 82–102.
14. T. H. Jukes and C. Cantor, *Mammalian protein metabolism*, Evolution of Protein Molecules (H. N. Munro, ed.), New York Academic Press, 1969, pp. 21–132.
15. M. A Kimura, *Simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences*, J. Mol. Evol. **16** (1980), 111–120.
16. M. Kuhner and J. Felsenstein, *A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates*, Mol. Biol. Evol. **11** (1994), 459–468.
17. J. G. Lawrence and H. Ochman, *Amelioration of bacterial genomes: rates of change and exchange*, J. Mol. Evol. **44** (1997), 383–397.
18. P. Legendre (guest ed.), *Special section on reticulate evolution*, J. Classification **17** (2000), 153–195.
19. P. Legendre and V. Makarenkov, *Reconstruction of biogeographic and evolutionary networks using reticulograms*, Syst. Biol. **51** (2002), 199–216.
20. W. P. Maddison, *Gene trees in species trees*, Syst. Biol. **46** (1997), 523–536.
21. V. Makarenkov, T-Rex: *reconstructing and visualizing phylogenetic trees and reticulation networks*, Bioinformatics **17** (2001), 664–668.
22. V. Makarenkov, A. Boc, C. F. Delwiche, A. B. Diallo, and H. Philippe, *New efficient algorithm for modeling partial and complete gene transfer scenarios*, Data Science and Classification (V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Ziberna, eds.), IFCS 2006, Studies in Classification, Data Analysis, and Knowledge Organization, part VIII, Springer Verlag, 2006, pp. 341–349.
23. V. Makarenkov and B. Leclerc, *An algorithm for the fitting of a phylogenetic tree according to a weighted least-squares criterion*, J. Classification **16** (1999), 3–26.
24. O. Matte-Tailliez, C. Brochier, P. Forterre, and H. Philippe, *Archaeal phylogeny based on ribosomal proteins*, Mol. Biol. Evol. **19** (2002), 631–639.
25. B. Mirkin, T. I. Fenner, M. Galperin, and E. Koonin, *Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes*, BMC Evol. Biol. **3** (2003); 2.
26. B. Mirkin, *Mapping gene family data onto evolutionary trees*, Comptes rendus des 11es Rencontres de la Sociéte Francophone de Classification, (M. Chavent, O. Dordan, C. Lacomblez, M. Langlais, and B. Patouille, eds.), University of Bordeaux, pp. 61-68.

27. B. M. E. Moret, L. Nakhleh, T. Warnow, C. R. Linder, A. Tholse, A. Padolina, J. Sun, and R. Timme, *Phylogenetic networks*: *modeling, reconstructibility, and accuracy*, IEEE/ACM Trans. on Comput. Biol. and Bioinf. **1** (2004), 13–23.
28. L. Nakhleh, D. Ruths, and H. Innan,*Gene trees, species trees, and species networks*, Meta-analysis and Combining Information in Genetics (R. Guerra and D. Allison, eds.), 2005, pp. 1–27.
29. R. D. M. Page, *Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas*, Systematic Biol. **43** (1994), 58–77.
30. R. D. M. Page and M. A. Charleston, *From gene to organismal phylogeny*: *reconciled trees*, Bioinformatics **14** (1998), 819–820.
31. _____, *Trees within trees*: *phylogeny and historical associations*, Trends Ecol. Evol. **13** (1998b), 356–359.
32. A. Rambaut and N. C. Grassly, *Seq-Gen*: *an application for the Monte Carlo simulation of* DNA *sequence evolution along phylogenetic trees*, Comput. Appl. Biosci. **13** (1996), 235–238.
33. D. R. Robinson and L. R. Foulds, *Comparison of phylogenetic trees*, Math. Biosciences **53** (1981), 131–147.
34. N. Saitou and M. Nei, *The neighbor-joining method*: *a new method for reconstructing phylogenetic trees*, Mol. Biol. Evol. **4** (1987), 406–425.
35. A. von Haeseler and G. A. Churchill, *Network models for sequence evolution*, J. Mol. Evol. **37** (1993), 77–85

Appendix A

This Appendix includes the results of the tests described in the section Simulation Study. The results reported in Tables 1 and 2 correspond to the graphics represented in Figures 7 (optimization using the RF distance) and 8 (optimization using the LS function). They were obtained from simulations carried out for random binary phylogenies with 8, 16, 24, 32, 48, and 64 leaves, whereas the sequence length varied from 125 to 1000 sites. Note that the sum of the HGT detection rate shown in Figures 7 and 8 and of the false negative detection rate reported in Tables 1 and 2 is always 100%.

Département d'Informatique, Université du Québec à Montréal, C.P. 8888, Succ. Centre-ville, Montréal (Québec), H3C 3P8, Canada.
E-mail address, V. Makarenkov: makarenkov.vladimir@uqam.ca
E-mail address, A. Boc: boc.alix@courrier.uqam.ca
E-mail address, A. B. Diallo: diallo.alpha_boubacar@courrier.uqam.ca

McGill Centre for Bioinformatics and School of Computer Science, McGill University, 3775 University Street, Montréal (Québec), H3A 2B4, Canada.
E-mail address: banire@mcb.mcgill.ca

TABLE 1. False positive and false negative (in parentheses) detection rates obtained for random phylogenies with 8 to 64 leaves using the RF distance as an optimization criterion. A false positive HGT is an incorrect transfer found by the algorithm and a false negative HGT is the right transfer that has not been found. For each sequence length, the simulations were carried out for 500 random phylogenies with 8 and 16 leaves and 100 random phylogenies with 24 to 64 leaves.

RF rates (in %)			Sequence length				
			125	250	500	750	1000
Species number	8	Jukes-Cantor	14.9(7.8)	5.9(3.5)	1.1(0.7)	0.3(0.3)	0.0(0.0)
		Kimura	12.9(8.7)	3.3(2.2)	0.2(0.1)	0.1(0.1)	0.0(0.0)
		Jin-Nei	20.1(15.0)	3.9(2.5)	1.6(1.3)	1.1(1.1)	0.5(0.5)
	16	Jukes-Cantor	25.7(14.0)	7.1(4.5)	1.2(0.7)	0.4(0.3)	0.0(0.0)
		Kimura	35.1(22.5)	11.9(7.9)	3.2(2.3)	0.6(0.6)	0.1(0.0)
		Jin-Nei	43.0(30.0)	22.5(16.5)	7.6(6.6)	5.3(4.9)	2.3(2.3)
	24	Jukes-Cantor	36(18)	15(10)	4(3)	1(1)	1(1)
		Kimura	43(24)	24(13)	4(2)	2(0)	0(0)
		Jin-Nei	55(35)	33(18)	19(10)	9(6)	5(4)
	32	Jukes-Cantor	37(20)	29(11)	4(2)	1(1)	1(0)
		Kimura	60(35)	31(14)	8(3)	3(1)	2(0)
		Jin-Nei	70(38)	47(25)	16(9)	8(3)	8(3)
	48	Jukes-Cantor	65(48)	49(29)	28(15)	1(1)	1(0)
		Kimura	55(38)	46(18)	9(3)	3(1)	2(0)
		Jin-Nei	70(40)	58(24)	19(8)	8(3)	8(3)
	64	Jukes-Cantor	70(60)	45(35)	27(17)	23(13)	20(10)
		Kimura	65(55)	35(25)	14(4)	12(2)	10(0)
		Jin-Nei	60(50)	44(34)	22(12)	18(8)	14(4)

TABLE 2. False positive and false negative (in parentheses) detection rates obtained for random phylogenies with 8 to 64 leaves using the LS function as an optimization criterion. A false positive HGT is an incorrect transfer found by the algorithm and a false negative HGT is the right transfer that has not been found. For each sequence length, the simulations were carried out for 500 random phylogenies with 8 and 16 leaves and 100 random phylogenies with 24 to 64 leaves.

RF rates (in %)			Sequence length				
			125	250	500	750	1000
Species number	8	Jukes-Cantor	17.2(10.1)	5.0(2.5)	0.8(0.7)	0.8(0.5)	0.3(0.3)
		Kimura	10.8(7.0)	2.8(1.9)	0.3(0.3)	0.2(0.2)	0.1(0.1)
		Jin-Nei	18.6(13.8)	7.8(6.5)	1.7(1.5)	0.9(0.8)	0.5(0.3)
	16	Jukes-Cantor	25.5(13.0)	7.6(5.3)	2.2(1.4)	0.8(0.5)	0.1(0.1)
		Kimura	37.6(23.8)	11.9(8.4)	2.3(2.0)	0.6(0.6)	0.0(0.0)
		Jin-Nei	40.9(28.8)	20.9(14.8)	8.1(6.7)	3.8(3.6)	3.3(3.3)
	24	Jukes-Cantor	43(22)	13(11)	5(5)	3(3)	1(1)
		Kimura	59(30)	26(9)	7(4)	4(3)	1(0)
		Jin-Nei	67(33)	26(18)	12(6)	6(2)	3(1)
	32	Jukes-Cantor	47(26)	21(14)	5(2)	0(0)	0(0)
		Kimura	56(33)	31(17)	9(4)	0(0)	0(0)
		Jin-Nei	50(33)	31(15)	12(8)	11(3)	4(0)
	48	Jukes-Cantor	53(43)	38(31)	33(7)	22(12)	19(11)
		Kimura	60(50)	34(14)	16(5)	5(1)	2(0)
		Jin-Nei	65(55)	50(29)	25(8)	12(4)	10(3)
	64	Jukes-Cantor	63(53)	52(42)	41(21)	27(17)	25(15)
		Kimura	70(60)	45(35)	22(12)	15(2)	10(0)
		Jin-Nei	75(65)	40(20)	20(10)	16(6)	12(2)