New efficient algorithm for modeling partial and complete gene transfer scenarios

Vladimir Makarenkov[1], Alix Boc[1], Charles F. Delwiche[2], Alpha Boubacar Diallo[1], and Hervé Philippe[3]

[1] Département d'informatique, Université du Québec à Montréal,
C.P. 8888, Succ. Centre-Ville, Montréal (Québec), H3C 3P8, Canada,
[2] Cell Biology and Molecular Genetics, HJ Patterson Hall, Bldg. 073,
University of Maryland at College Park, MD 20742-5815, USA.
[3] Département de biochimie, Faculté de Médecine, Université de Montréal,
C.P. 6128, Succ. Centre-ville, Montréal, QC, H3C 3J7, Canada.

Abstract. In this article we describe a new method allowing one to predict and visualize possible horizontal gene transfer events. It relies either on a metric or topological optimization to estimate the probability of a horizontal gene transfer between any pair of edges in a species phylogeny. Species classification will be examined in the framework of the complete and partial gene transfer models.

1 Introduction

Species evolution has long been modeled using only phylogenetic trees, where each species has a unique most recent ancestor and other interspecies relationships, such as those caused by horizontal gene transfers (HGT) or hybridization, cannot be represented (Legendre and Makarenkov (2002)). HGT is a direct transfer of genetic material from one lineage to another. Bacteria and Archaea have sophisticated mechanisms for the acquisition of new genes through HGT, which may have been favored by natural selection as a more rapid mechanism of adaptation than the alteration of gene functions through numerous mutations (Doolittle (1999)). Several attempts to use network-based models to depict horizontal gene transfers can be found (see for example: Page (1994) or Charleston (1998)). Mirkin et al (1995) put forward a tree reconciliation method that combines different gene trees into a unique species phylogeny. Page and Charleston (1998) described a set of evolutionary rules that should be taken into account in HGT models. Tsirigos and Rigoutsos (2005) introduced a novel method for identifying horizontal transfers that relies on a gene's nucleotide composition and obviates the need for knowledge of codon boundaries. Lake and Rivera (2004) showed that the dynamic deletions and insertions of genes that occur during genome evolution, including those introduced by HGT, may be modeled using techniques similar to those used to model nucleotide substitutions (e.g. general Markov models). Moret et al (2004) presented an overview of the network modeling in phylogenetics. In this paper we continue the work started in Makarenkov et al (2004), where we described an HGT detection algorithm based on the least-squares optimization. To design a detection algorithm which is mathematically and biologically sound we will consider two possible approaches allowing for complete and partial gene transfer scenarios.

2 Two different ways of transferring genes

Two HGT models are considered in this study. The first model, assumes partial gene transfer. In such a model, the original species phylogeny is transformed into a connected and directed network where a pair of species can be linked by several paths (Figure 1a). The second model assumes complete transfer; the species phylogenetic tree is gradually transformed into the gene tree by adding to it an HGT in each step. During this transformation, only tree structures are considered and modified (Figure 1b).

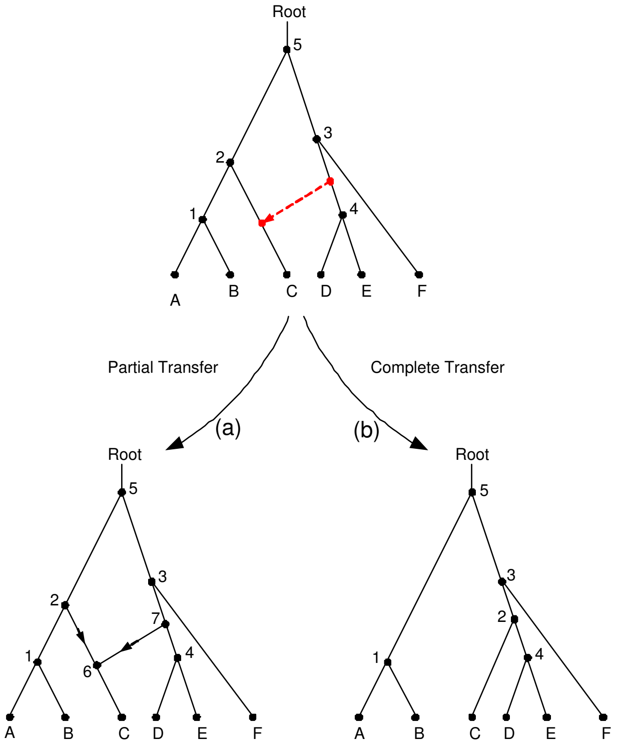

Fig. 1. Two evolutionary models, assuming that either a partial (a) or complete (b) HGT has taken place. In the first case, only a part of the gene is incorporated into the recipient genome and the tree is transformed into a directed network, whereas in the second, the entire donor gene is acquired by the host genome and the species tree is transformed into a different tree.

3 Complete gene transfer model

In this section we discuss the main features of the HGT detection algorithm in the framework of the complete gene transfer model. This model assumes that the entire transferred gene is acquired by the host (Figures 1b). If the homologous gene was present in the host genome, the transferred gene can supplant it. Two optimization criteria will be considered. The first of them is the least-squares (LS) function Q:

$$Q = \sum_i \sum_j (d(i,j) - \delta(i,j))^2, \tag{1}$$

where $d(i,j)$ is the pairwise distance between the leaves i and j in the species phylogenetic tree T and $\delta(i,j)$ the pairwise distance between i and j in the gene tree T_1. The second criterion that can be useful to assess the incongruence between the species and gene phylogenies is the Robinson and Foulds (RF) topological distance (1981). When the RF distance is considered, we can use it as an optimization criterion as follows: All possible transformations (Figure 1b) of the species tree, consisting of transferring one of its subtrees from one edge to another, are evaluated in a way that the RF distance between the transformed species tree T' and the gene tree T_1 is computed. The subtree transfer providing the minimum of the RF distance between T' and T_1 is retained as a solution. Note that the problem asking to find the minimum number of subtree transfer operations necessary to transform one tree into another has been shown to be NP-hard but approximable to within a factor of 3 (Hein et al (1996)).

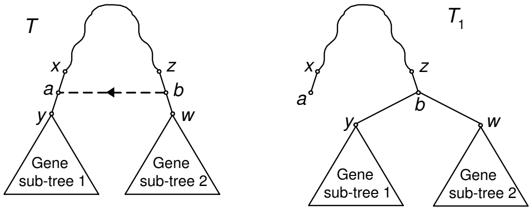

Fig. 2. Timing constraint: the transfer between the edges (z,w) and (x,y) of the species tree T can be allowed if and only if the cluster regrouping both affected subtrees is present in the gene tree T_1.

Several biological rules have to be considered in order to synchronize the way of evolution within a species phylogeny (Page and Charleston (1998)). For instance, transfers between the species of the same lineage must be prohibited. In addition, our algorithm relies on the following timing constraint: The cluster combining the subtrees rooted by the vertices y and w must be present in the gene tree T_1 in order to allow an HGT between the edges (z,w) and (x,y) of the species tree T (Figure 2). Such a constraint enables us, first, to arrange the topological conflicts between T and T_1 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. The main steps of the HGT detection algorithm are the following:

Step 0. This step consists of inferring the species and gene phylogenies denoted respectively T and T_1 and labeled according to the same set X of n taxa (e.g. species). Both species and gene trees should be explicitly rooted. If the topologies of T and T_1 are identical, we conclude that HGTs are not required to explain the data. If not, either the RF difference between them can be used as a phylogeny transformation index, or the gene tree T_1 can be mapped into the species tree T fitting by least-squares the edge lengths of T to the pairwise distances in T_1 (see Makarenkov and Leclerc (1999)).

Step 1. The goal of this step is to obtain an ordered list L of all possible gene transfer connections between pairs of edges in T. This list will comprise all different directed connections (i.e. HGTs) between pairs of edges in T except the connections between adjacent edges and those violating the evolutionary constraints. Each entry of L is associated with the value of the gain in fit, computed using either LS function or RF distance, found after the addition of the corresponding HGT connection. The computation of the ordered list L requires $O(n^4)$ operations for a phylogenetic tree with n leaves. The first entry of L is then added to the species tree T.

Steps 2 ... k. In the step k, a new tree topology is examined to determine the next transfer by computing the ordered list L of all possible HGTs. The procedure stops when the RF distance equals 0 or the LS coefficient stops decreasing (ideally dropping to 0). Such a procedure requires $O(kn^4)$ operations to add k HGT edges to a phylogenetic tree with n leaves.

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 genome through the process of homologous recombination. Mathematically, this means that the traditional species phylogenetic tree is transformed into a directed evolutionary network (Figure 1a). Figure 3 illustrates the case where the evolutionary distance between the taxa i and j may change after the addition of the edge (b,a) representing a partial gene transfer from b to a.

From a biological point of view, it is relevant to consider that the HGT from b to a can affect the distance between the taxa i and j if and only if a is located on the path between i and the root of the tree; the position of j is assumed to be fixed. Thus, in the network T (Figure 3) the evolutionary distance $dist(i,j)$ between the taxa i and j can be computed as follows:

$$dist(i,j) = (1-\mu)d(i,j) + \mu(d(i,a) + d(j,b)), \tag{2}$$

where μ indicates the fraction (unknown in advance) of the gene being transferred and d is the distance between the vertices in T before the addition of the HGT edge (b,a). A number of biological rules, not discussed here due to the space limitation, have to be incorporated into this model (see Makarenkov et al (2004) for more details). Here we describe the main features of the network-building algorithm:

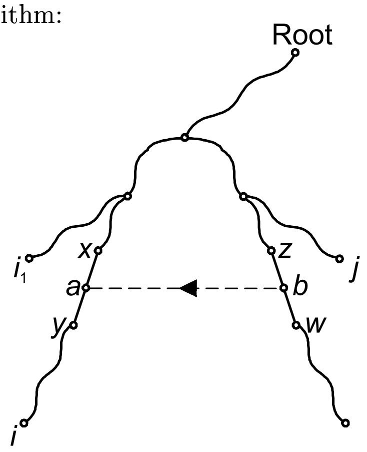

Fig. 3. Evolutionary distance between the taxa i and j can be affected by the addition of the edge (b,a) representing a partial HGT between the edges (z,w) and (x,y). Evolutionary distance between the taxa i_1 and j cannot be affected by the addition of (b,a).

Step 0. This step corresponds to Step 0 defined for the complete gene transfer model. It consists of inferring the species and gene phylogenies denoted respectively T and T_1. Because the classical RF distance is defined only for tree topologies, we use the LS optimization when modeling partial HGT.

Step 1. Assume that a partial HGT between the edges (z,w) and (x,y) (Figure 3) of the species tree T has taken place. The lengths of all edges in T should be reassessed after the addition of (b,a), whereas the length of (b,a) is assumed to be 0. To reassess the edge lengths of T, we have first to make an assumption about the value of the parameter μ (Equation 2) 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 by testing different values of μ in the optimization problem. Fixing this parameter, 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. edge lengths of T) than equations (i.e. pairwise distances in T; number of equations is always $n(n\text{-}1)/2$ for n taxa) can be solved by approximation in the least-squares sense. All pairs of edges in T can be processed in this way. The HGT connection providing the smallest value of the LS coefficient and satisfying the evolutionary constraints will be selected for the addition to the tree T transforming it into a phylogenetic network.

Steps 2 ... k. In the same way, the best second, third and other HGT edges can be added to T, improving in each step the LS fit of the gene distance. The whole procedure requires $O(kn^5)$ operations to build a reticulated network with k HGT edges starting from a species phylogenetic tree with n leaves.

5 Detecting horizontal transfers of PheRS synthetase

In this section, we examine the evolution of the PheRS protein sequences for 32 species including 24 Bacteria, 6 Archaea, and 2 Eukarya (see Woese et al (2000)). The PheRS phylogenetic tree inferred with PHYML (Guindon and Gascuel (2003)) using G-law correction is shown in Figure 4.

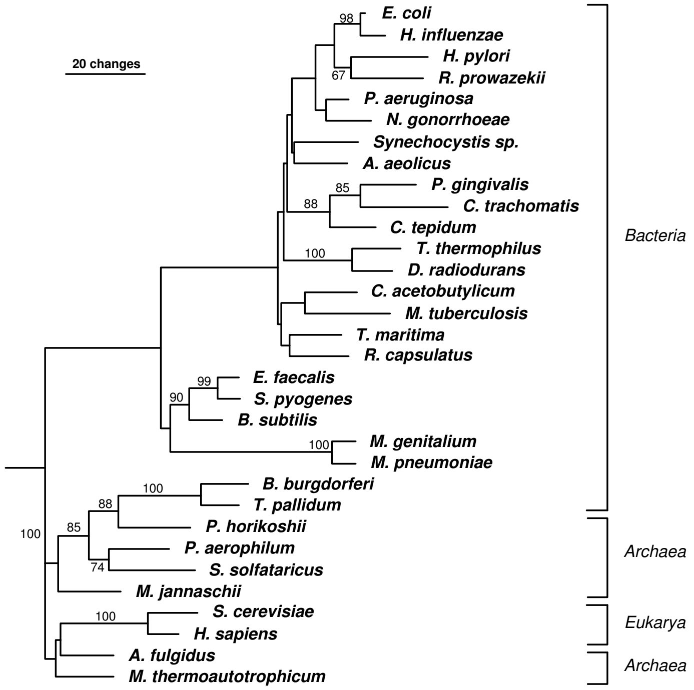

Fig. 4. Phylogenetic tree of PheRS sequences (i.e. gene tree). Protein sequences with 171 bases were considered. Bootstrap scores above 60% are indicated.

This tree is slightly different from the phylogeny obtained by Woese et al (2000, Fig. 2); the biggest difference involves the presence of a new cluster formed by two Eukarya (*H. sapiens* and *S. cerevisiae*) and two Archaea (*A. fulgidus* and *M. thermoautotrophicum*). This 4-species cluster with a low bootstrap support is probably due to the reconstruction artifacts. Otherwise, this tree shows the canonical pattern, the only exception being the spirochete PheRSs (i.e. *B. bugdorferi* and *T. pallidum*). They are of the archaeal, not the bacterial genre, but seem to be specifically related to *P. horokoshii* within that grouping (Figure 4). The species tree corresponding to the NCBI taxonomic classification was also inferred (Figure 5, undirected lines). The computation of HGTs was done in the framework of the complete gene transfer model. The five transfers with the biggest bootstrap scores are represented.

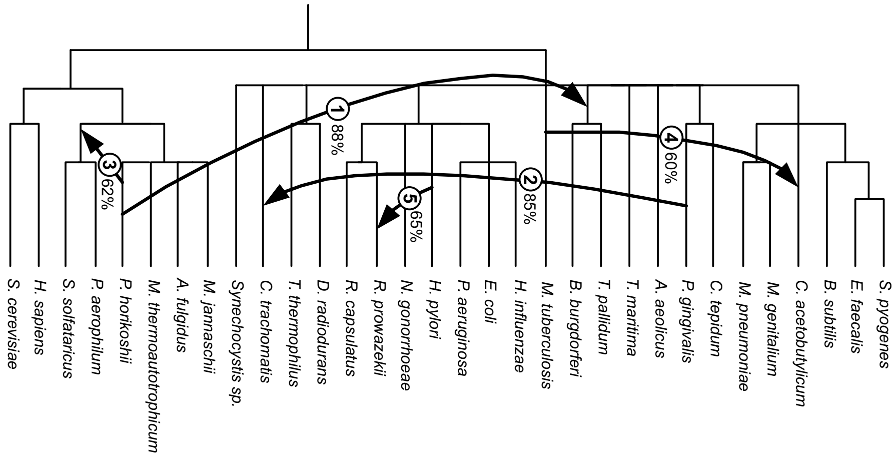

Fig. 5. Species phylogeny corresponding to the NCBI taxonomy for the 32 species in Figure 4. HGTs with bootstrap scores above 60% are depicted by arrows. Numbers on the HGT edges indicate their order of appearance in the transfer scenario.

The bootstrap scores for HGT edges were found fixing the topology of the species tree and resampling the PheRS sequences used to obtain the gene tree. The transfer number 1, having the biggest bootstrap support, 88%, links *P. horokoshii* to the clade of spirochetes. This bootstrap score is the biggest one that could be obtained for this HGT, taking into account the identical 88% score of the corresponding 3-species cluster in the PheRS phylogeny (Figure 4). In total, 14 HGTs, including 5 trivial connections, were found; trivial transfers occur between the adjacent edges. Trivial HGTs are necessary to transform a non-binary tree into a binary one. The non-trivial HGTs with low bootstrap score are most probably due to the tree reconstruction artifacts. For instance, two HGT connections (not shown in Figure 5) linking the cluster of Eukarya to the Archaea (*A. fulgidus* and *M. thermoautotrophicum*) have a low bootstrap support (16% and 32%, respectively). In this example, the solution found with the RF distance was represented. The usage of the LS function leads to the identical scenario differing from that shown in Figure 5 only by the bootstrap scores found for the HGT edges 3 to 5.

6 Conclusion

We described a new distance-based algorithm for the detection and visualization of HGT events. It 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 probability of HGT between each pair of edges of the species phylogeny. In this study 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 examples of the evolution of the PheRS synthetase considered in the application section showed that the new algorithm can be useful for predicting HGT in real data. In the future, it would be interesting to extend and test this procedure in the framework of the maximum likelihood and maximum parsimony models. The program implementing the new algorithm was included to the T-Rex package (Makarenkov (2001), http://www.trex.uqam.ca).

References

CHARLESTON, M. A. (1998): Jungle: a new solution to the host/parasite phylogeny reconciliation problem. *Math. Bioscience, 149, 191-223.*

DOOLITTLE, W. F. (1999): Phylogenetic classification and the universal tree. *Science, 284, 2124-2129.*

GUINDON, S. and GASCUEL, O. (2003): A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol., 52, 696-704.*

LAKE, J. A. and RIVERA, M. C. (2004): Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Mol. Biol. Evol., 21, 681-690.*

LEGENDRE, P. and V. MAKARENKOV. (2002): Reconstruction of biogeographic and evolutionary networks using reticulograms. *Syst. Biol., 51, 199-216.*

MAKARENKOV, V. and LECLERC, B. (1999): An algorithm for the fitting of a tree metric according to a weighted LS criterion. *J. of Classif., 16, 3-26.*

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

MAKARENKOV, V., BOC, A. and DIALLO, A. B. (2004): Representing lateral gene transfer in species classification. Unique scenario. In: D. Banks, L. House, F. R. McMorris, P. Arabie, and W. Gaul (eds.): *Classification, Clustering and Data Mining Applications.* Springer Verlag, proc. IFCS 2004, Chicago 439-446

MIRKIN, B. G., MUCHNIK, I. and SMITH, T.F. (1995): A Biologically Consistent Model for Comparing Molecular Phylogenies. *J. of Comp. Biol., 2, 493-507.*

MORET, B., NAKHLEH, L., WARNOW, T., LINDER, C., THOLSE, A., PADOLINA, A., SUN, J. and TIMME, R. (2004): Phylogenetic Networks: Modeling, Reconstructibility, Accuracy. *Trans. Comp. Biol. Bioinf., 1, 13-23.*

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

PAGE, R. D. M. and CHARLESTON, M. A. (1998): Trees within trees: phylogeny and historical associations. *Trends Ecol. Evol., 13, 356-359.*

ROBINSON, D. R. and FOULDS, L. R. (1981): Comparison of phylogenetic trees. *Math. Biosciences, 53, 131-147.*

TSIRIGOS, A. and RIGOUTSOS, I. (2005): A Sensitive, Support-Vector-Machine Method for the Detection of Horizontal Gene Transfers in Viral, Archaeal and Bacterial Genomes. *Nucl. Acids Res., 33, 3699-3707.*

WOESE, C., OLSEN, G., IBBA, M. and SÖLL, D. (2000): Aminoacyl-tRNA synthetases, genetic code, evolut. process. *Micr. Mol. Biol. Rev., 64, 202-236.*