An Efficient Algorithm for the Detection and Classification of Horizontal Gene Transfer Events and Identification of Mosaic Genes

Alix Boc, Pierre Legendre, and Vladimir Makarenkov

Abstract In this article we present a new algorithm for detecting partial and complete horizontal gene transfer (HGT) events which may give rise to the formation of mosaic genes. The algorithm uses a sliding window procedure that analyses sequence fragments along a given multiple sequence alignment (MSA). The size of the sliding window changes during the scanning process to better identify the blocks of transferred sequences. A bootstrap validation procedure incorporated in the algorithm is used to assess the bootstrap support of each predicted partial or complete HGT. The proposed technique can be also used to refine the results obtained by any traditional algorithm for inferring complete HGTs, and thus to classify the detected gene transfers as partial or complete. The new algorithm will be applied to study the evolution of the gene *rpl12e* as well as the evolution of a complete set of 53 archaeal MSA (i.e., 53 different ribosomal proteins) originally considered in Matte-Tailliez et al. (Mol Biol Evol 19:631–639, 2002).

1 Introduction

Bacteria and viruses adapt to changing environmental conditions via horizontal gene transfer (HGT) and intragenic recombination leading to the formation of mosaic genes, which are composed of alternating sequence parts belonging either to the original host gene or stemming from the integrated donor sequence (Doolittle 1999; Zhaxybayeva et al. 2004). An accurate identification and classification of

A. Boc · P. Legendre
Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, QC H3C 3J7, Canada
e-mail: alix.boc@umontreal.ca; pierre.legendre@umontreal.ca

V. Makarenkov (✉)
Département d'Informatique, Université du Québec à Montréal, C.P.8888, succursale Centre Ville, Montreal, QC H3C 3P8, Canada
e-mail: makarenkov.vladimir@uqam.ca

B. Lausen et al. (eds.), *Algorithms from and for Nature and Life*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-319-00035-0_25,

mosaic genes as well as the detection of the related gene transfers are among the most important challenges posed by modern computational biology (Koonin 2003; Zheng et al. 2004). Partial HGT model assumes that any part of a gene can be transferred among the organisms under study, whereas traditional (complete) HGT model assumes that only an entire gene, or a group of complete genes, can be transferred (Makarenkov et al. 2006a,b). Mosaic genes can pose several risks to humans including cancer onset or formation of antibiotic-resistant genes spreading among pathogenic bacteria (Nakhleh et al. 2005). The term "mosaic" stems from the pattern of interspersed blocks of sequences having different evolutionary histories, but being combined in the resulting allele subsequent to recombination events. The recombined segments can derive from other strains in the same species or from other more distant bacterial or viral relatives (Gogarten et al. 2002; Hollingshead et al. 2000). Mosaic genes are constantly generated in populations of transformable organisms, and probably in all genes (Maiden 1998).

Many methods have been proposed to address the problem of the identification and validation of complete HGT events (e.g., Boc et al. 2010, Hallett and Lagergren 2001, and Nakhleh et al. 2005), but only a few methods treat the much more challenging problem of inferring partial HGTs and predicting the origins of mosaic genes (Denamur et al. 2000; Makarenkov et al. 2006b). We have recently proposed (Boc and Makarenkov 2011) a new method allowing for detection and statistical validation of partial HGT events using a sliding window approach.

In this article we describe an extension of the algorithm presented in Boc and Makarenkov (2011), considering sliding windows of variable size. We will show how the new algorithm can be used: (1) to estimate the robustness of the obtained HGT events; (2) to classify the obtained transfers as partial or complete; (3) to classify the species under study as potential donors or receivers of genetic material.

2 Algorithm

Here we present the new algorithm for inferring partial horizontal gene transfers using a sliding window of adjustable size. The idea of the method is to provide the most probable partial HGT scenario characterizing the evolution of the given gene. It takes as input a species phylogenetic tree representing the traditional evolution of the group of species under study and a multiple sequence alignment (MSA) representing the evolution of the gene of interest for the same group of species. A sliding window procedure, with a variable window size, is carried out to scan the fragments of the given MSA (see Fig. 1). In the algorithm Boc and Makarenkov (2011), the sliding window size was constant, thus preventing the method from detecting accurately the exact lengths of the transferred sequences (i.e. only an approximate length of the transferred sequence blocks was provided). In this study, the most appropriate size of the sliding window is selected with respect to the significance of the gene transfers inferred for different overlapping MSA intervals. The HGT significance is computed as the average HGT bootstrap support (Boc et al. 2010) obtained for the corresponding fixed MSA interval.

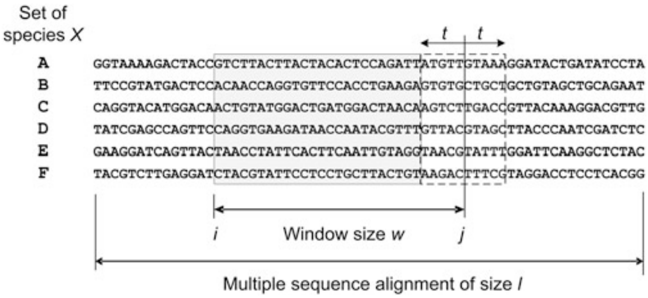
Set of
species X
t
t
A GGTAAAAGACTACCGTCTTACTTACTACACTCCAGATTATGTTGTAAAGGATACTGATATCCTA
B TTCCGTATGACTCCACAACCAGGTGTTCCACCTGAAGAGTGTGCTGCTGCTGTAGCTGCAGAAT
C CAGGTACATGGACAACTGTATGGACTGATGGACTAACAAGTCTTGACCGTTACAAAGGACGTTG
D TATCGAGCCAGTTCCAGGTGAAGATAACCAATACGTTTGTTACGTAGCTTACCCAATCGATCTC
E GAAGGATCAGTTACTAACCTATTCACTTCAATTGTAGGTAACGTATTTGGATTCAAGGCTCTAC
F TACGTCTTGAGGATCTACGTATTCCTCCTGCTTACTGTAAGACTTTCGTAGGACCTCCTCACGG
i
Window size w
j
Multiple sequence alignment of size l

Fig. 1 New algorithm uses a sliding window of variable size. If the transfers obtained for the original window position $[i; i + w - 1]$ are significant, we refine the obtained results by searching in all the intervals of types $[i; i + w - 1 - t + k]$ and $[i; i + w - 1 + t - k]$, where $k = 0, \dots, t$ and t is a fixed window contraction/extension parameter

The algorithm includes the three following main steps:

***Step* 1.** Let X be a set of species and l is the length of the given MSA. We first define the initial sliding window size w ($w = j - i + 1$, see Fig. 1) and the window progress step s. The species tree, denoted T, characterizing the evolution of the species in X can be either inferred from the available taxonomic or morphological data, or can be given. T must be rooted to take into account the evolutionary time-constraints that should be satisfied when inferring HGTs (Boc et al. 2010; Hallett and Lagergren 2001).

***Step* 2.** For i varying from 1 to $|l - w + 1|$, we first infer (e.g., using PhyML, Guindon and Gascuel (2003)) a partial gene tree T' from the subsequences located within the interval $[i; i+w-1]$ of the given MSA. If the average bootstrap support of the edges of T' constructed for this interval is significant (i.e. >60 % in this study), then we apply a standard HGT detection algorithm (e.g., Boc et al. 2010) using as input species phylogenetic tree T and partial gene tree T'. If the transfers obtained for this interval are significant, then we perform the algorithm for all the intervals of types $[i; i+w-1-t+k]$ and $[i; i + w - 1 + t - k]$, where $k = 0, \dots, t$ (see Fig. 1) and t is a fixed window contraction/extension parameter (in our study the value of t equal to $w/2$ was used). If for some of these intervals the average HGT significance is greater than or equal to the HGT significance of the original interval $[i; i+w-1]$, then we adjust the sliding window size w to the length of the interval providing the greatest significance, which may be typical for the dataset being analyzed. If the transfers corresponding to the latter interval have an average bootstrap score greater than a pre-defined threshold (e.g. when the average HGT bootstrap score of the interval is >50 %), we add them to the list of predicted partial HGT events and advance along the given MSA with the progress step s. The bootstrapping procedure for HGT is presented in Boc et al. (2010).

***Step* 3**. Using the established list of all predicted significant partial HGT events, we identify all overlapping intervals giving rise to the identical partial transfers (i.e., the same donor and recipient and the same direction) and re-execute the algorithm separately for all overlapping intervals (considering their total length in each case). If the same partial significant transfers are found again when concatenating these overlapped intervals, we assess their bootstrap support and, depending on the obtained support, include them in the final solution or discard them. If some significant transfers are found for the intervals whose length is greater than 90 % the total MSA length, those transfers are declared complete.

The time complexity of the described algorithm is as follows:

$$O(r \times (\frac{t \times (l - w)}{s} \times (C(Phylo\_Inf) + C(HGT\_Inf)))), \qquad (1)$$

where $C(Phylo\_Inf)$ is the time complexity of the tree inferring method used to infer partial gene phylogenetic trees, $C(HGT\_Inf)$ is the time complexity of the HGT detection method used to infer complete transfers and r is the number of replicates in the HGT bootstrapping. The simulations carried out with the new algorithm (due the lack of space, the simulation results are not presented here) showed that it outperformed the algorithm described in Boc and Makarenkov (2011) in terms of HGT prediction accuracy, but was slower than the latter algorithm, especially in the situations when large values of the t parameter were considered.

3 Application Example

We first applied the new algorithm to analyze the evolution of the gene *rpl12e* for the group of 14 organisms of Archaea originally considered in Matte-Tailliez et al. (2002). The latter authors discussed the problems encountered when reconstructing some parts of the species phylogeny for these organisms and indicated the evidence of HGT events influencing the evolution of the gene *rpl12e* (MSA size for this gene was 89 sites). In Boc et al. (2010), we examined this dataset using an algorithm for predicting complete HGTs and found five complete transfers that were necessary to reconcile the reconstructed species and gene *rpl12e* phylogenetic trees (see Fig. 2a). These results confirm the hypothesis formulated in Matte-Tailliez et al. (2002). For instance, HGT 1 between the cluster (*Halobacterium sp.*, *Haloarcula mar.*) and *Methanobacterium therm.* as well as HGTs 4 and 5 between the clade of *Crenarchaeota* and the organisms *Thermoplasma ac.* and *Ferroplasma ac.* have been characterized in Matte-Tailliez et al. (2002) as the most likely HGT events occurred during the evolution of this group of species. In this study, we first applied the new algorithm allowing for prediction of partial and complete HGT event to confirm or discard complete horizontal gene transfers presented in Fig. 2a,

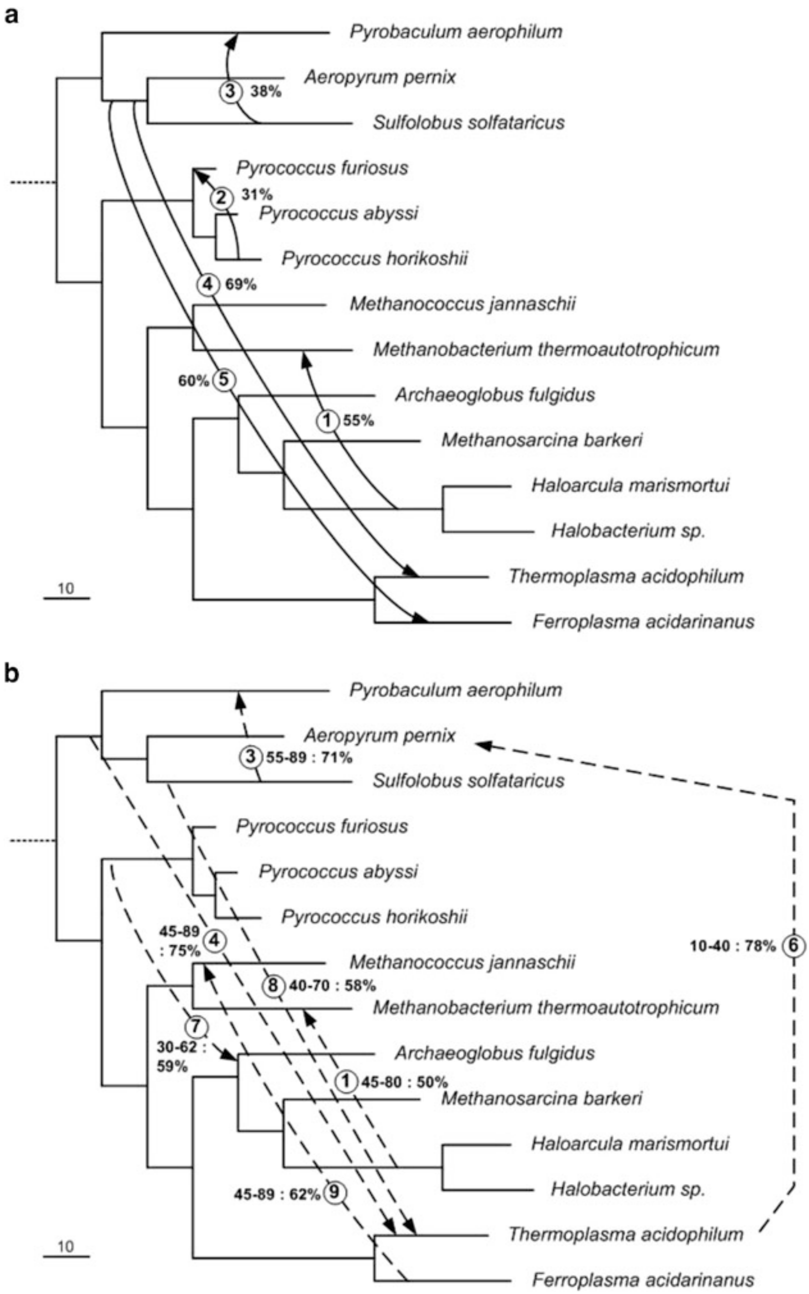

Fig. 2 Species tree (Matte-Tailliez et al. 2002, Fig. 1a) encompassing: (**a**) five complete horizontal gene transfers, found by the algorithm described in Boc et al. (2010), indicated by *arrows*; numbers on HGTs indicate their order of inference; HGT bootstrap scores are indicated near each HGT *arrow*; and (**b**) seven partial HGTs detected by the new algorithm; the identical transfers have the same numbers in the positions A and B of the figure; the interval for which the transfer was detected and the corresponding bootstrap score are indicated near each HGT *arrow*

and thus to classify the detected HGT as partial or complete. We used an original window size w of 30 sites (i.e. 30 amino acids), a step size s of 5 sites, the value of $t = w/2$ and a minimum acceptable HGT bootstrap value of 50 %; 100 replicates were used in the HGT bootstrapping. The new algorithm found seven partial HGTs represented in Fig. 2b. The identical transfers in Fig. 2a, b have the same numbers.

The original lengths of the transfers 1, 3 and 7 (see Fig. 2b) have been adjusted (with the values +4, +5 and +2 sites, respectively) to find the interval length providing the best average significance rate, the transfers 4 and 9 have been detected on two overlapping original intervals, and the transfers 6 and 8 have been detected using the initial window size. In this study, we applied the partial HGT detection algorithm with the new dynamic windows size feature to bring to light the possibility of creation of mosaic gene during the HGT events described above. The proposed technique for inferring partial HGTs allowed us to refine the results of the algorithm predicting complete transfers (Boc et al. 2010). Thus, the transfers found by both algorithms (i.e. HGTs 1, 3 and 4) can be reclassified as partial. They are located approximately on the same interval of the original MSA. The complete HGTs 2 and 5 (Fig. 2a) were discarded by the new algorithm. In addition, four new partial transfers were found (i.e. HGTs 6, 7, 8 and 9). Thus, we can conclude that no complete HGT events affected the evolution of the gene *rpl12e* for the considered group of 14 species, and that the genes of 6 of them (i.e. *Pyrobaculum aer.*, *Aeropyrum pern.*, *Methanococcus jan.*, *Methanobacterium therm.*, *Archaeoglobus fulg.* and *Thermoplasma acid.*) are mosaic.

Second, we applied the presented HGT detection algorithm to examine a complete dataset of 53 ribosomal archaeal proteins (i.e. 53 different MSAs for the same group of species were considered; see Matte-Tailliez et al. (2002) for more details). Our main objective here was to compute complete and partial HGT statistics and to classify the observed organisms as potential donors or receivers of genetic material. The same parameter settings as in the previous example were used. Figure 3 illustrates the 10 most frequent partial (and complete) transfer directions found for the 53 considered MSAs. The numbers near the HGT arrows indicate the rate of the most frequent partial HGTs, which is followed by the rate of complete HGTs. Matte-Taillez and colleagues (Matte-Tailliez et al. 2002) pointed out that only about 15 % (8 out of 53 genes; the gene *rpl12e* was a part of these 8 genes) of the ribosomal genes under study have undergone HGT events during the evolution of archaeal organisms. The latter authors also suggested that the HGT events were rather rare for these eight proteins. Our results (see Fig. 3) shows, however, that about 36 % of the genes analyzed in this study can be considered as mosaic genes. Also, we found that about 7 % of genes were affected by complete gene transfers. The most frequent partial HGTs were found within the groups of *Pyrococcus* (HGTs 1, 2 and 5) and *Crenarchaeota* (HGTs 3 and 4). We can also conclude that partial gene transfers were about five times more frequent than partial HGT events.

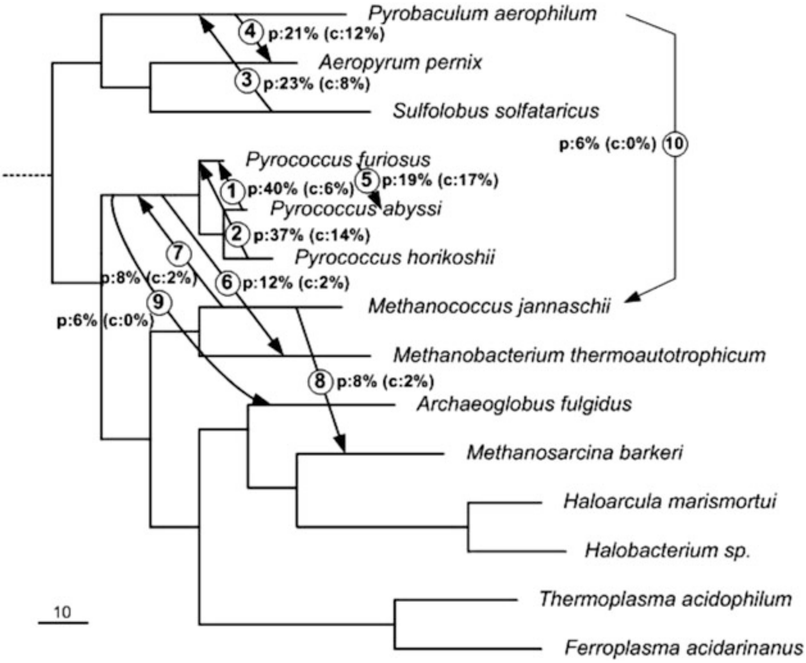

Fig. 3 Species tree (Matte-Tailliez et al. 2002, Fig. 1a) with the 10 most frequent HGT events obtained by the new algorithm when analyzing separately the MSAs of 53 archaeal proteins. The first value near each HGT arrow indicates the rate of partial HGT detection (p) and the second value indicates the rate of complete HGT detection (c)

4 Conclusion

In this article we described a new algorithm for inferring partial and complete horizontal gene transfer events using a sliding window approach in which the size of the sliding window is adjusted dynamically to fit the nature of the sequences under study. Such an approach aids to identify and validate mosaic genes with a better precision. The main advantage of the presented algorithm over the methods used to detect recombination in sequence data is that it allows one to determine the source (i.e. putative donor) of the subsequence being incorporated in the host gene. The discussed algorithm was applied to study the evolution of the gene *rpl12e* and that of a group of 53 ribosomal proteins in order to estimate the pro-portion of mosaic genes as well as the rates of partial and complete gene transfers characterizing the considered group of 14 archaeal species. In the future, this algorithm could be adapted to compute several relevant statistics regarding the functionality of genetic fragments affected by horizontal gene transfer as well as to estimate the rates of intraspecies (i.e. transfers between strains of the same species) and interspecies (i.e. transfers between distinct species) HGT.

References

Boc, A., Philippe, H., & Makarenkov, V. (2010). Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Systematic Biology, 59*, 195–211.

Boc, A., & Makarenkov, V. (2011). Towards an accurate identification of mosaic genes and partial horizontal gene transfers. *Nucleic Acids Research*. doi:10.1093/nar/gkr735.

Denamur, E., Lecointre, G., & Darlu, P., et al. (12 co-authors) (2000). Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell, 103*, 711–721.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science, 284*, 2124–2129.

Gogarten, J. P., Doolittle, W. F., & Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution, 19*, 2226–2238.

Guindon, S., & Gascuel, O. (2003). A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology, 52*, 696–704.

Hallett, M., & Lagergren, J. (2001). Efficient algorithms for lateral gene transfer problems. In N. El-Mabrouk, T. Lengauer, & Sankoff, D. (Eds.), *Proceedings of the fifth annual international conference on research in computational biology*, Montréal (pp. 149–156). New-York: ACM.

Hollingshead, S. K., Becker, R., & Briles, D. E. (2000). Diversity of PspA: mosaic genes and evidence for past recombination in Streptococcus pneumoniae. *Infection and Immunity, 68*, 5889–5900.

Koonin, E. V. (2003). Horizontal gene transfer: the path to maturity. *Molecular Microbiology, 50*, 725–727.

Maiden, M. (1998). Horizontal genetic exchange, evolution, and spread of antibiotic resistance in bacteria. *Clinical Infectious Diseases, 27*, 12–20.

Makarenkov, V., Kevorkov, D., & Legendre, P. (2006a). Phylogenetic network reconstruction approaches. In *Applied mycology and biotechnology* (International Elsevier series: Bioinformatics, Vol. 6, pp. 61–97). Amsterdam: Elsevier.

Makarenkov, V., Boc, A., Delwiche, C. F., Diallo, A. B., & Philippe, H. (2006b). New efficient algorithm for modeling partial and complete gene transfer scenarios. In V. Batagelj, H. H. Bock, A. Ferligoj, A. Ziberna, (Eds.), *Data science and classification* (pp. 341–349). Berlin: Springer.

Matte-Tailliez, O., Brochier, C., Forterre, P., & Philippe, H. (2002). Archaeal phylogeny based on ribosomal proteins. *Molecular Biology and Evolution, 19*, 631–639.

Nakhleh, L., Ruths, D., & Wang, L. S. (2005). RIATA-HGT: A fast and accurate heuristic for reconstructing horizontal gene transfer. In L. Wang, (Ed.), *Lecture notes in computer science* (pp. 84–93). Kunming: Springer.

Zhaxybayeva, O., Lapierre, P., & Gogarten, J. P. (2004). Genome mosaicism and organismal lineages. *Trends in Genetics, 20*, 254–260.

Zheng, Y., Roberts, R. J., & Kasif, S. (2004). Segmentally variable genes: a new perspective on adaptation. *PLoS Biology, 2*, 452–464.