RIATA-HGT: A Fast and Accurate Heuristic for Reconstructing Horizontal Gene Transfer

Luay Nakhleh[1], Derek Ruths[1], and Li-San Wang[2]

[1] Department of Computer Science, Rice University
Houston, TX 77005, USA
{nakhleh,druths}@cs.rice.edu
[2] Department of Biology, University of Pennsylvania
Philadelphia, PA 19104, USA
lswang@mail.med.upenn.edu

Abstract. Horizontal gene transfer (HGT) plays a major role in microbial genome diversification, and is claimed to be rampant among various groups of genes in bacteria. Further, HGT is a major confounding factor for any attempt to reconstruct bacterial phylogenies. As a result, detecting and reconstructing HGT events in groups of organisms has become a major endeavor in biology. The problem of detecting HGT events based on incongruence between a species tree and a gene tree is computationally very hard (NP-hard). Efficient algorithms exist for solving restricted cases of the problem.

We propose RIATA-HGT, the first polynomial-time heuristic to handle all HGT scenarios, without any restrictions. The method accurately infers HGT events based on analyzing incongruence among species and gene trees. Empirical performance of the method on synthetic and biological data is outstanding. Being a heuristic, RIATA-HGT may overestimate the optimal number of HGT events; empirical performance, however, shows that such overestimation is very mild.

We have implemented our method and run it on biological and synthetic data. The results we obtained demonstrate very high accuracy of the method. Current version of RIATA-HGT uses the PAUP tool, and we are in the process of implementing a stand-alone version, with a graphical user interface, which will be made public. The tool, in its current implementation, is available from the authors upon request.

1 Introduction

Horizontal (also known as lateral) gene transfer (HGT) plays a major role in microbial genome diversification [7, 20], and is claimed to be rampant among various groups of genes in bacteria [8]. M.-W. Ho also has written of the risks that HGT poses to humans, which include (1) antibiotic resistance genes spreading to pathogenic bacteria; (2) disease-associated genes spreading and recombining to create new viruses and bacteria that cause diseases; and (3) transgenic DNA inserting into human cell, triggering cancer [11]. Furthermore, the occurrence of HGT confounds or completely defeats any attempt to reconstruct evolution (especially for bacterial organisms) as has been famously summarized by Ford Doolittle [8]

L. Wang (Ed.): COCOON 2005, LNCS 3595, pp. 84–93, 2005.

> Molecular phylogeneticists will have failed to find the "true tree," not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot properly be represented as a tree.

In light of all this evidence supporting the significance of HGT as an evolutionary mechanism, its risks, and its confounding effects on evolution reconstruction, much research in biology today is dedicated to the problems of understanding the nature of HGT events and detecting and reconstructing them (their numbers as well as their donors and recipients).

In order to reconstruct genomic changes, what we usually seek is the species (organismal) phylogeny – the tree that traces the history of the replicating cell lineages that transmit genes and genomes to successive generations [15]. This species phylogeny provides the backdrop against which events such as HGT have occurred. In their study, Lerat *et al.* identified a set of genes resistant to HGT (those, according to the authors, are usually single-copy orthologous genes), combined them and built a species tree [15]. They tested whether a given gene had been horizontally transferred by comparing its tree (topology) against the species trees. Based on this study, the problem of detecting and reconstructing HGT is formulated as follows: given a species tree ST and a set G of gene trees, compute the minimum-cardinality set of HGT events whose occurrence on tree ST give rise to the gene trees in G (we give a mathematical definition of the problem in Section 2.1). The problem's formulation as an optimization problem, in which a minimum-cardinality set of HGT events is sought, is a reflection of Occam's razor: in the absence of any additional biological knowledge, HGT events should be used sparingly to explain data features otherwise explainable under a tree model. Further, the actual set of HGT events may not be computationally identifiable in certain cases since multiple (equally optimal) solutions may exist for the problem. The problem of finding a minimum-cardinality set of HGT events whose occurrence of species tree ST would give rise to the gene trees in set G is computationally NP-hard [4]. Efficient solutions for the problem exist, but for limited special cases [10, 19].

In this paper, we propose a polynomial-time method, RIATA-HGT, for solving a relaxed version of the problem, where we drop the optimality criterion in the problem definition. Although the cardinality of the HGT set computed by our method is not guaranteed (theoretically) to be the minimum (among all such sets), experimental results of our method, on both biological and synthetic data, demonstrate that, in practice, the method almost always infers the correct set of HGT events. Whenever the method overestimates the minimum amount of HGT events, such an overestimation is very mild (often one or two additional HGT events). RIATA-HGT takes as input a species tree and a set of gene trees, and computes HGT events to explain all of those gene trees.

The rest of the paper is organized as follows. In Section 2 we briefly review the biology behind HGT, and give an overview of the techniques and tools for analyzing it. In Section 2.1 we mathematically define the problem of species-tree/gene-tree incongruence and HGT reconstruction. Section 3 describes RIATA-HGT and the underlying algorithms Further, we analyze the theoretical properties of the method. In Section 4 we show the performance of our method on biological as well as synthetic data. We conclude in Section 5 with final remarks and directions for future research.

2 Horizontal Gene Transfer and Its Detection: A Brief Overview

In horizontal gene transfer (HGT), genetic material is transferred from one lineage to another, such that certain sites (specified by a specific substring within the DNA sequence of the species into which the horizontally transferred DNA was inserted) are inherited through horizontal transfer from another species, while all others are inherited from the parent. Horizontal transfers are believed to be ubiquitous among bacteria and still quite common in other branches of the Tree of Life [7] – although this view has recently been challenged [9, 13, 23, 24]. The three major modes of HGT in the Archaea and Bacteria are *transformation* (uptake of naked DNA from the environment), *conjugation* (transfer of DNA by direct physical interaction between a donor and a recipient), and *transduction* (transfer of DNA by phage infection) [21].

The goal of much biological research has been to identify those genes that were acquired by the organism through horizontal transfers rather than inherited from its ancestors. In one of the first papers on the topic, Medigue [17] proposed the use of multivariate analysis of codon usage to identify such genes; since then various authors have proposed other intrinsic methods, such as using GC content, particularly in the third position of codons (e.g., [14]). An advantage of intrinsic approaches is their ability to identify and eliminate genes that do not obey a tree-like process of evolution and thus could prevent classical phylogenetic methods from reconstructing a good tree. With the advent of whole-genome sequencing, more powerful intrinsic methods become possible, such as the location of suspect genes with each genome: such locations tend to be preserved through lineages, but a transfer event can place the new gene in a more or less random location. Thus, biologists often consider the neighbors of a gene in the prokaryotic genome to identify horizontal transfers. However, differential selection pressure, uneven evolutionary rates, and biased gene sampling can all give rise to false identification of HGT [9].

Non-intrinsic approaches involve phylogenetic reconstruction. These methods use phylogenetic reconstructions to identify discrepancies that could tag transfer events. An old question in phylogenetic reconstruction has been "to combine or not to combine?" – that is, given DNA sequences for several genes, are we better off concatenating the sequences or analyzing each set separately (e.g., [6])?

The common sense conclusion that many genes inherited through lineal descent would override the confusing signal generated by a few genes acquired through horizontal transfer appears wrong [5, 27]. Of course, one must first resolve the old problem of gene trees vs. species trees: discrepancies between the trees derived from different genes do not necessarily indicate reticulate evolution, but may simply testify to the incongruent evolution of two or more genes, all within a valid, tree-shaped evolution of the species (e.g., [16, 22]). Distinguishing between the two is difficult in the absence of additional information.

With whole-genome sequencing, such information becomes available. Huynen [12] advocate two types of data: the fraction of shared orthologs and gene synteny. Synteny (the conservation of genes on the same chromosome) is not widely applicable with prokaryotes, but its logical extension, conservation of gene order, definitely is – and Huynen and Bork proposed to measure the fraction of conserved adjacencies.

2.1 The Graph-Theoretic Approach

From a graph-theoretic point of view, the problem can be formulated as pure phylogenetic network reconstruction [18, 19]. In the case of HGT, a phylogenetic network is a pair (T, Ξ), where T is the species (organismal) tree, and Ξ is a set of HGT edges whose addition to T creates a directed acyclic graph (DAG) N, referred to as a phylogenetic network. We say that a network N (interchangeably, pair (T, Ξ)) induces a tree T' if T' can be obtained from N by the following two steps: (1) for each node in N that has two edges coming into it, remove one of the two edges, and (2) suppress each node that has one incoming edge and one outgoing edge.

The problem of resolving incongruence between a species tree ST and a set G of gene trees is then defined as follows.

Problem 1. (The HGT Reconstruction Problem)

> **Input:** A species tree ST and a set G of gene trees.
> **Output:** Set Ξ of minimum cardinality such that the pair (ST, Ξ) induces each of the gene trees in G.

As mentioned before, the minimization criterion reflects the fact that the simplest solution is sought; in this case, the simplest solution is one with the minimum number of HGT events. Hallett and Lagergren [10] gave an efficient algorithm for solving the HGT Reconstruction Problem; however, their algorithm handles limited special cases of the problem in which the number of HGT events is very small, and the number of times a gene is transferred is very low (also, their tool handles only binary trees; [2]). Boc and Makarenkov [3] solve the problem by reconstructing the species tree from a sequence alignment, and then add edges to the tree to minimize a distance-based optimization criterion. Since the original alignment contains HGT events, the starting tree may be inaccurate, which results in the addition of arbitrary HGT events. Further, distance-based approaches suffer from a lack of accurate techniques for estimating branch lengths. Nakhleh *et al.* [19] gave efficient algorithms for solving the problem, but for limited special cases referred to as gt-networks; further, they handled only binary trees. In the next section, we describe our method, RIATA-HGT, which is a heuristic for solving the HGT Reconstruction Problem, and demonstrate its empirical performance in Section 4. RIATA-HGT is the first method for solving the general case of the HGT Reconstruction Problem.

3 RIATA-HGT

3.1 Terminology and Notation

A rooted phylogenetic tree T over set S of taxa is a rooted tree with $|S|$ leaves, each labeled by a unique element of S. We denote by $r(T)$ the root of T and by $L(T)$ the leaf set of T. Let T be a rooted phylogenetic tree over set S of taxa, and let $S' \subseteq S$. We denote by $T(S')$ the minimal rooted subtree of T that connects all the element of S'. Furthermore, the restriction of T to S', denote $T|S'$ is the rooted subtree that is obtained from $T(S')$ by suppressing all vertices (except for the root) whose number of incident

edges is 2. Let S' be a maximum-cardinality set of leaves such that $T_1|S' = T_2|S'$, for two trees T_1 and T_2; we call $T_1|S'$ (equivalently, $T_2|S'$) the maximum agreement subtree of the two trees, denoted $MAST(T_1, T_2)$. A *clade* of a tree T is a complete subtree of T. Let $T' = MAST(T_1, T_2)$; then, $T_1 - T'$ is the set of all maximal clades whose pruning from T_1 yields T' (we define $T_2 - T'$ similarly). In other words, there do not exist two clades u and u' in $T_1 - T'$ such that either u is a clade in u', or u' is a clade in u.

We say that node x reaches node y in tree T if there is a directed path from x to y in T. We denote the root of a clade t by $r(t)$. We say that clade t_1 reaches clade t_2 (both in tree T) if $r(t_1)$ reaches $r(t_2)$. The sibling of node x in tree T is node y, denoted $sibling_T(x) = y$ whenever x and y are children of the same node in T. We denote by T_x the clade rooted at node x in T. The least common ancestor of a set X of taxa in tree T, denoted $lca_T(X)$ is the root of the minimal subtree of T that contains the leaves of X. The edge incoming into node x in tree T is denoted by $inedge_T(x)$.

3.2 The Algorithm

We describe the algorithm underlying RIATA-HGT in terms of a species tree and a gene tree. Our implementation of RIATA-HGT allows the user to specify a set of gene trees, and it iterates over each pair of the species tree and a gene tree, and summarizes the results for all trees. The core of RIATA-HGT is the divide-and-conquer algorithm ComputeHGT algorithm (outlined in Fig. 1). The algorithm starts by computing the $MAST$, T', of the species tree ST and gene tree GT; tree T' forms the basis for detecting and reconstructing the HGT events (computing T' is done in Step 1 in Fig. 1). The algorithm then decomposes the clade sets U_1 and U_2 (whose removal from ST and GT, respectively, yields T') into maximal clades such that each maximal clade in one of the two sets is "matched" by a maximal clade on the same leaf set in the second set. The algorithm for this decomposition is outlined in Fig. 2. The algorithm then recurses on each maximal clade and its matching maximal clade (Steps 5.c.(1) and 5.d.(5).(1) in Fig. 1) to compute the HGT events whose recipients form sub-clades of those maximal clades. Finally, we add a single HGT event per each maximal clade to connect it to its "donor" in the ST; this is achieved through the calls to AddSingleHGT (Fig. 3) in Steps 5.c.(2) and 5.d.(5).(3) in Fig. 1. We have the following properties of the method, which we present without proofs due to space constraints: (1) ComputeHGT always terminates, (2) The pair (ST, Ξ) (the output of ComputeHGT) induces the tree GT, and (3) $ComputeHGT$ takes $O((h^2 + \log n)n^2)$ time.

4 Experimental Results and Discussion

We have implemented RIATA-HGT using the Python language [1], and used the PAUP* tool [26] for computing the maximum agreement subtree of two trees. We studied the empirical performance of our method on synthetic as well as biological datasets. We tried to run the tool of [10], but, unfortunately, the program crashed on almost all datasets.

To obtain synthetic datasets, we have written a simulator that takes a (species) tree ST as input, and adds a randomly generated set Ξ of HGT events to T, where the

PROCEDURE COMPUTEHGT(ST,GT)
Input: Species tree ST, and gene tree GT, both on the same set S of taxa.
Output: Computes the set Ξ of HGT events such that the pair (ST, Ξ) induces GT.

1. $T' = MAST(ST, GT)$;
2. **If** $T' = ST$ **then**
   (a) **Return**;
3. $U_1 = ST - T'$; $U_2 = GT - T'$;
4. $V = \emptyset$;
5. **Foreach** $u_2 \in U_2$
   (a) $Decompose(U_1, u_2, T', V)$;
6. $U_2 = V$;
7. **While** $V \neq \emptyset$
   (a) Let u_2 be an element of V;
   (b) Let $u_1 \in U_1$ be such that $L(u_2) \subseteq L(u_1)$;
   (c) $Y = \{y \in U_2 : L(y) \cap L(u_1) \neq \emptyset\}$;
   (d) $Z = \{y | (L(y) - L(u_1)) : \ y \in Y\}$;
   (e) $V = V - Y$; $V = V \cup Z$;
   (f) $X = \{u_1 | L(y) : y \in Y\}$;
   (g) **Foreach** $y \in Y$
      i. Let $x \in X$ be such that $L(x) \cap L(y) \neq \emptyset$;
      ii. $ComputeHGT(x, y)$;
      iii. $AddSingleHGT(ST, GT, y, U_2, T')$;

Fig. 1. The main algorithm for detecting and reconstructing HGT events based on a pair of species tree and gene tree.

number of events in Ξ is specified by the user. The simulator also implements certain constraints so that the network N resulting from adding Ξ to ST is a directed acyclic graph. Once the pair (ST, Ξ) is generated, the simulator outputs a (gene) tree GT that uses all the edges in Ξ. In other words, GT results from N by

1. for each node v in N such that there are two edges incoming into v, remove the edge that is not in Ξ; and
2. suppress all nodes that has only one incoming edge and one outgoing edge.

To generate a species tree, ST, we used the r8s tool [25], which generates random birth-death trees, with a number of leaves specified by the user.

We studied the performance of RIATA-HGT on ten different sizes k of Ξ, where k ranges from 1 to 10. For each k, we generated 30 triplets (ST, Ξ, GT), with $|\Xi| = k$, as described above, and ran RIATA-HGT on (ST, GT). We studied the performance of RIATA-HGT in terms of the number of HGT events it predicts compared to the actual number of HGT events. We looked at the predictions in each of the runs (Fig. 4(a)), and at the averages (Fig. 4(b)).

The box-and-whisker plot in Fig. 4(a) show that RIATA-HGT computed the correct set of HGT events in most cases. In particular, RIATA-HGT obtained the exact number of HGT events (with the exception of very few outliers) in the cases when

PROCEDURE DECOMPOSE(U_1,u_2, T, U')
Input: Set U_1 of clades from ST, clade u_2 from GT, the backbone clade u_2, and U' which will contain the "refined" clades of u_2.
Output: Decompose u_2 so that no clade in U' has a leaf set that is the union of leaf sets of more than one clade in U_1.

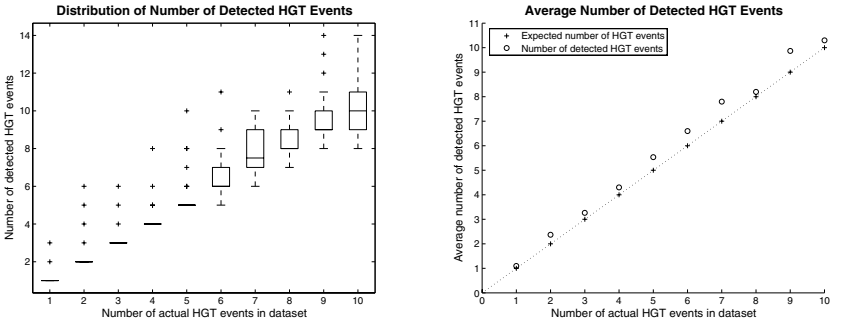
1. **If** $\exists u_1 \in U_1$ such that $L(u_2) \subseteq L(u_1)$ **then**
   (a) $U' = U' \cup \{u_2\}$;
   (b) $B(u_2) = T$;
   (c) **Return** u_2;
2. **Else**
   (a) **If** $\exists u_1 \in U_1$ such that $r(u_2) = r(u_2|L(u_1))$
      i. $t = u_2|L(u_1)$;
      ii. $U' = U' \cup \{t\}$;
      iii. $B(t) = T$;
      iv. Let $X = u_2 - t$;
      v. **Foreach** $x \in X$
         A. $Decompose(U_1, x, t, U')$;
      vi. **Return** t;
   (b) **Else**
      i. Let $c_1, \ldots, c_k$ be the children of $r(u_2)$;
      ii. $x = Decompose(U_1, T_{c_1}, T, U')$;
      iii. **For** $i = 2$ to k
         A. $Decompose(U_1, T_{c_i}, x, U')$;
      iv. **Return** x;

Fig. 2. The algorithm for decomposing the clades in U_1 and U_2 such that for all $u_1 \in U_1$ and $u_2 \in U_2$ we have $L(u_1) \not\subset L(u_2)$.

PROCEDURE ADDSINGLEHGT(ST, GT, u_2, U_2, T')
Input: Species tree ST, gene tree GT, clade u_2 of GT, set U_2 of clades of GT, and $MAST$ T' of ST and GT.
Output: Add to Ξ a single HGT event whose donor is determined in this procedure and whose recipient is clade u_2.

1. $Q = L(u_2) \cup L(B(u_2))$;
2. $T'' = GT|Q$; $\quad p = lca_{T''}(L(u_2))$;
3. $tq = lca_{ST}(L(u_2))$; $\quad te = inedge_{ST}(tq)$;
4. **If** p is a child of $r(T'')$ and $|L(B(u_2))| > 1$ **then**
   (a) $sq = lca_{ST}(L(B(u_2)))$;
   (b) $\Xi = \Xi \cup (sq \rightarrow te)$;
5. **Else**
   (a) $O = \bigcup_{\{p':p'=sibling_{T''}(p)\}} L(T_{p'})$;
   (b) $sq = lca_{ST}(O)$; $\quad se = inedge_{ST}(sq)$;
   (c) $\Xi = \Xi \cup (se \rightarrow te)$;

Fig. 3. The algorithm for detecting and reconstructing the single HGT event in which clade u_2 is the recipient.



(a) (b)

Fig. 4. (a) A box-and-whisker plot for the predictions of HGT event numbers made by RIATA-HGT. (b) The averages of HGT event numbers estimated by RIATA-HGT vs. the actual number of HGT events.

the actual number of HGT events was between 1 and 5. For datasets with 6 to 10 HGT events, RIATA-HGT predicted the correct number in a large number of the cases. When RIATA-HGT underestimated the actual number, the method found the minimum number of HGT events, as opposed to the actual number (for reasons outlined in Section 2.1). In the absence of any additional biological knowledge, computing the minimum number of such events amounts to finding the simplest solution. The cases where RIATA-HGT overestimates the number of HGT events are a reflection of the heuristic nature of the method. Nonetheless, the overestimation is very mild, as Fig. 4(a) shows. In Fig. 4(b), we plot the average predictions of RIATA-HGT versus the expected number of HGT events in the input. The figure shows a very mild deviation of the average predicted numbers of HGT events from the expected numbers.

For the biological dataset, we considered the species tree and two gene trees of the γ-Proteobacteria group, as reported in [15]. The species tree (shown in Figs. 5(a) and 5(b)) was reconstructed by the authors using a phylogenetic analysis on a sequence dataset obtained by concatenating 203 orthologous gene datasets, all of whose gene trees were concordant. Fig. 5(c) shows the gene tree of the biotin synthase enzyme (BioB), and Fig. 5(d) shows the gene tree of the virulence factor MviN. Both gene trees are incongruent with the species tree of Fig. 5(a).

RIATA-HGT computed the set Ξ of HGT events shown in Fig. 5(a) when invoked on the species tree and the biotin gene tree, and the set Ξ of HGT events shown in Fig. 5(b) when invoked on the species tree and the virulence gene tree. Those two sets of HGT events were hypothesized in [15], and RIATA-HGT computed them. In summary, RIATA-HGT performed very well on the synthetic datasets we generated, as well as on the biological dataset we used.

5 Conclusions and Future Work

We proposed a new method, RIATA-HGT, for detecting and reconstructing horizontal gene transfer events. Our method is a polynomial-time heuristic that, given a species

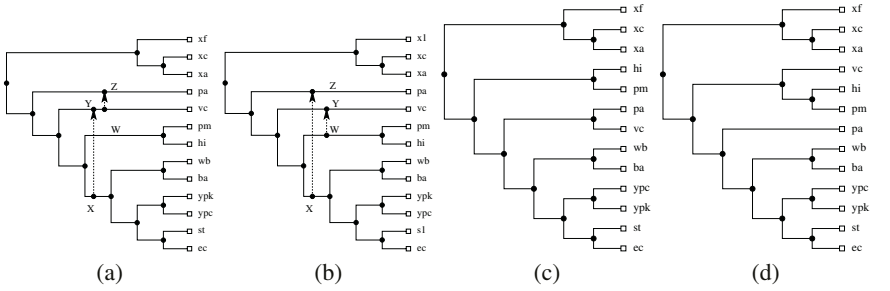


Fig. 5. The species tree of the γ-Proteobacteria group, as reported in [15], is shown by the solid lines in (a) and (b). The two HGT events were computed by RIATA-HGT using the gene tree in (c) are shown by the dotted arrows in (a), and the two HGT events were computed by RIATA-HGT using the gene tree in (d) are shown by the dotted arrows in (b).

tree and a set of gene trees as input, attempts to compute the set of HGT events that explain all the gene trees. Despite the lack of theoretical bounds on the performance of our method, we demonstrated, using synthetic and biological datasets, that RIATA-HGT has excellent performance in practice. RIATA-HGT is the first fast heuristic, with proven empirical performance, that handles the general case of the HGT Reconstruction Problem.

We plan to provide a standalone version of RIATA-HGT (along with a graphical user interface) to the research community. Future work includes testing the method on more biological datasets, more efficient handling of multiple gene trees (instead of the iterative process currently implemented), and extending the method to handle cases in which not all homologs of genes are found across all species.

References

1. Python software foundation, 2005. www.python.org.
2. L. Addario-Berry, M.T. Hallett, and J. Lagergren. Towards identifying lateral gene transfer events. In *Proc. 8th Pacific Symp. on Biocomputing (PSB03)*, pages 279–290, 2003.
3. A. Boc and V. Makarenkov. New efficient algorithm for detection of horizontal gene transfer events. In *Proc. 3rd Int'l Workshop Algorithms in Bioinformatics (WABI03)*, volume 2812, pages 190–201. Springer-Verlag, 2003.
4. M. Bordewich and C. Semple. On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics*, pages 1–15, 2005. In press.
5. J.R. Brown, C.J. Douady, M.J. Italia, W.E Marshall, and M.J. Stanhope. Universal trees based on large combined protein sequence data sets. *Nat. Genet.*, 28:281–285, 2001.
6. J.J. Bull, J.P. Huelsenbeck, C.W. Cunningham, D. Swofford, and P. Waddell. Partitioning and combining data in phylogenetic analysis. *Syst. Biol.*, 42(3):384–397, 1993.
7. F. de la Cruz and J. Davies. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.*, 8:128–133, 2000.
8. W.F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284:2124–2129, 1999.
9. J.A. Eisen. Horizontal gene transfer among microbial genomes: New insights from complete genome analysis. *Curr Opin Genet Dev.*, 10(6):606–611, 2000.

10. M.T. Hallett and J. Lagergren. Efficient algorithms for lateral gene transfer problems. In *Proc. 5th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB01)*, pages 149–156, New York, 2001. ACM Press.
11. M.-W. Ho. Recent evidence confirms risks of horizontal gene transfer, 2002. http://www.i-sis.org.uk/FSAopenmeeting.php.
12. M.A. Huynen and P. Bork. Measuring genome evolution. *Proc. Nat'l Acad. Sci., USA*, 95:5849–5856, 1998.
13. C.G. Kurland, B. Canback, and O.G. Berg. Horizontal gene transfer: A critical view. *Proc. Nat'l Acad. Sci., USA*, 100(17):9658–9662, 2003.
14. J.G. Lawrence and H. Ochman. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.*, 44:383–397, 1997.
15. E. Lerat, V. Daubin, and N.A. Moran. From gene trees to organismal phylogeny in prokaryotes: The case of the γ-proteobacteria. *PLoS Biology*, 1(1):1–9, 2003.
16. W. Maddison. Gene trees in species trees. *Syst. Biol.*, 46(3):523–536, 1997.
17. C. Medigue, T. Rouxel, P. Vigier, A. Henaut, and A. Danchin. Evidence for horizontal gene transfer in E. coli speciation. *J. Mol. Biol.*, 222:851–856, 1991.
18. 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 Transactions on Computational Biology and Bioinformatics*, 1(1):13–23, 2004.
19. L. Nakhleh, T. Warnow, and C.R. Linder. Reconstructing reticulate evolution in species–theory and practice. In *Proc. 8th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB04)*, pages 337–346, 2004.
20. H. Ochman, J.G. Lawrence, and E.A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 2000.
21. P.J. Planet. Reexamining microbial evolution through the lens of horizontal transfer. In R. DeSalle, G. Giribet, and W. Wheeler, editors, *Molecular Systematics and Evolution: Theory and Practice*, pages 247–270. Birkhauser Verlag, 2002.
22. A. Rokas, B.L. Williams, N. King, and S.B. Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425:798–804, 2003.
23. S.L. Salzberg and J.A. Eisen. Lateral gene transfer or viral colonization? *Science*, 293:1048, 2001.
24. S.L. Salzberg, O. White, J. Peterson, and J.A. Eisen. Microbial genes in the human genome – lateral transfer or gene loss? *Science*, 292(5523):1903–1906, 2001.
25. M. Sanderson. `r8s` software package. Available from http://loco.ucdavis.edu/r8s/r8s.html.
26. D.L. Swofford. PAUP*: Phylogenetic analysis using parsimony (and other methods), 1996. Sinauer Associates, Underland, Massachusetts, Version 4.0.
27. S.A. Teichmann and G. Mitchison. Is there a phylogenetic signal in prokaryote proteins? *J. Mol. Evol.*, 49:98–107, 1999.