DLIGHT – Lateral Gene Transfer Detection Using Pairwise Evolutionary Distances in a Statistical Framework

Christophe Dessimoz*, Daniel Margadant, and Gaston H. Gonnet

ETH Zurich, Institute of Computational Science,
CH-8092 Zurich and
Swiss Institute of Bioinformatics,
cdessimoz@inf.ethz.ch

Abstract. This paper presents an algorithm to detect lateral gene transfer (LGT) on the basis of pairwise evolutionary distances. The prediction is made from a likelihood ratio derived from hypotheses of LGT versus no LGT, using multivariate normal theory. In contrast to approaches based on explicit phylogenetic LGT detection, it avoids the high computational cost and pitfalls associated with gene tree inference, while maintaining the high level of characterization obtainable from such methods (species involved in LGT, direction, distance to the LGT event in the past). We validate the algorithm empirically using both simulation and real data, and compare its predictions with standard methods and other studies.

1 Introduction

Lateral gene transfer (LGT), or horizontal gene transfer (HGT), is widely recognized as a major force in prokaryotic genome evolution, but the study of its nature and extent is constrained by the limitations of current methods for LGT detection [1,2]. These methods can be divided in two broad categories: parametric methods and phylogenetic methods. In parametric methods, sequence properties such as nucleotide composition [3,4], dinucleotide frequencies [5], codon usage biases [6,7,8], or, more recently, nucleotide substitution matrices [9] are calculated for a specific gene and compared with the rest of the genome. A transferred gene has parameter values typical for its donor genome, which makes it distinguishable from the recipient genome. For this reason, the method can only detect LGT events taking place between organisms with significantly different patterns of evolution. Furthermore, parametric methods are limited to recent LGT transfers because the transfered sequences adapt to their new host relatively rapidly [3]. Lastly, some native genes may have atypical nucleotide composition for reasons other than LGT.

Phylogenetic methods identify LGT events by analyzing the discrepancy between the phylogeny of laterally transfered genes and their host genomes. Therefore, most phylogenetic methods consist of inference of gene and species trees,

[*] Corresponding author.

M. Vingron and L. Wong (Eds.): RECOMB 2008, LNBI 4955, pp. 315–330, 2008.

and their reconciliation [10,11]. Other methods, such as Lawrence's rank correlation test [12] or Clarke's phylogenetic discordance test [13] use unexpected sequence similarity scores to detect LGT, and do not require the inference of gene trees. To distinguish between the two types, we refer to the former by *explicit*, the latter by *implicit* phylogenetic methods. Explicit methods have the potential of describing in detail LGT events (involved species, direction of transfer, time of the transfer), but suffer from the difficulties associated with the inference of gene trees, a task both computationally expensive and error-prone. On the other hand, the two implicit phylogenetic methods mentioned here are fast and robust, though limited by their reliance on similarity scores, which do not always reflect phylogeny [14] in the first place, and by the relative coarseness of their underlying models, which limits their detection power.

In this manuscript, we introduce a new phylogenetic method for LGT detection, which we call DLIGHT (Distance Likelihood based Inference of Genes Horizontally Transfered). Based on evolutionary distances and applied in a probabilistic framework, it combines the speed, the lack of gene tree requirement, and the robustness of implicit methods with the high level of details obtained by explicit methods. The next section presents the algorithm, and is followed by validation using simulation and biological data.

2 Method

2.1 Preliminaries

Definition (family of orthologs). *A set of sequences (genes or proteins*[1]*) $f = \{x_1, x_2, ...\}$ is a family of orthologs if all pairs of sequences (x_i, x_j) in f are either orthologs or xenologs through orthologous replacement. We denote the set of all such families by F.*

DLIGHT's objective is to detect LGT in such families of orthologs. In the above definition, we require that the families have no paralogs (paralogy detection is beyond the scope of this article). This also ensures that there is at most one sequence per species in any family of orthologs. Thus, a sequence is also uniquely referenced by the pair (f, g), where f is a family that contains the sequence and g the species it belongs to (or the genome – the two terms are used here interchangeably). We denote by $G(f)$ the set of species of sequences of f. We denote the evolutionary distance between sequences of species i and j in family f by $d_f(i, j)$.

Assumption 1 (interspecies distance, family-specific rates). *We assume that, in the absence of LGT, all distances between orthologs of species i and j are proportional to an interspecies distance $d(i, j)$, with a family-specific proportionality constant τ_f. Formally, $d_f(i, j) = \tau_f \cdot d(i, j)$. Furthermore, we require that on average, the proportionality constant be one ($\frac{1}{|F|} \sum_{f \in F} \tau_f = 1$). This model is refered to as proportional branch lengths by [15].*

[1] In this work, we consider at most one protein sequence per gene.

Estimator $\hat{d}_f(i,j)$. The evolutionary distance $d_f(i,j)$ can be estimated from a pairwise alignment by maximum likelihood (ML) under a model of amino-acid substitution. We call this estimator $\hat{d}_f(i,j)$. The ML estimator is asymptotically unbiased and asymptotically normally distributed. The ML procedure also provides an estimate of its variance $\sigma^2(\hat{d}_f(i,j))$. Furthermore, covariance estimation is shown in [16].

Estimator $\hat{d}(i,j)$. We estimate the interspecies distance $d(i,j)$ using the unweighted sample average over all $|F|$ families of orthologs:

$$\hat{d}(i,j) = \frac{1}{|F|} \sum_{f \in F} \hat{d}_f(i,j)$$

The estimator is unbiased, because:

$$\mathbb{E}(\hat{d}(i,j)) = \frac{1}{|F|} \sum_{f \in F} \mathbb{E}(\hat{d}_f(i,j)) = \frac{1}{|F|} \sum_{f \in F} \tau_f \cdot d(i,j) = d(i,j) \underbrace{\frac{1}{|F|} \sum_{f \in F} \tau_f}_{=1} = d(i,j)$$

Assumption 2. *In the following, we will consider $\hat{d}(i,j)$ to be a point estimate, that is, we assume that $\sigma^2(\hat{d}(i,j)) = 0$.*

This assumption may appear to be quite strong, especially if the number of families under consideration is small. In most cases, however, the number of families is relatively large (larger than the size of a typical family), and the variances of interspecies distances are much smaller than those of the other estimators under consideration here. In terms of computation, the assumption considerably reduces the time complexity of our approach.

Estimator $\hat{\tau}_f$. We estimate the rate τ_f of family f using the following estimator:

$$\hat{\tau}_f = \frac{\frac{1}{n_f(n_f-1)} \sum_{i,j \in G(f), i \neq j} \hat{d}_f(i,j)}{\frac{1}{n_f(n_f-1)} \sum_{i,j \in G(f), i \neq j} \hat{d}(i,j)} = \frac{\sum_{i,j \in G(f), i \neq j} \hat{d}_f(i,j)}{\sum_{i,j \in G(f), i \neq j} \hat{d}(i,j)}$$

where $n_f = |G(f)|$. Due to assumption 2, the denominator is constant, and thus $\hat{\tau}_f$ follows a normal distribution with variance

$$\sigma^2(\hat{\tau}_f) = \frac{\sum_{i,j,k,l \in f, i \neq j, k \neq l} cov(\hat{d}_f(i,j), \hat{d}_f(k,l))}{(\sum_{i,j \in f, i \neq j} \hat{d}(i,j))^2}$$

Lateral Gene Transfer

Definition (lateral gene transfer). *In the present work, a lateral gene transfer (LGT) event is the transfer of a gene from a donor species d (or an ancestor thereof) to a recipient species r (or an ancestor thereof).*

Assumption 3. *Since the divergence of* d *and* r*, at most one LGT event per family of orthologs took place between the two lineages.*

Assumption 4. *The rate of evolution (the branch length on the phylogenetic tree) of a sequence after LGT is homogeneous among all donor and recipient lineages.*

Definition (δ). *Given a LGT event in family* f *between lineages of* d *and* r*, the evolutionary distance between the transfered sequence and the current sequences in* r *or* d *is expressed by* δ *(Fig. 1). The distance since LGT is the same for both species due to assumption 4.*

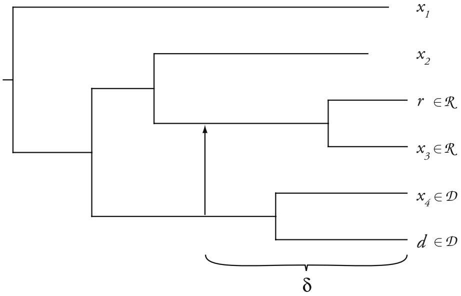

Fig. 1. Distance to LGT event as captured by the parameter δ. The LGT event is represented by the arrow.

Consequently, the expected distance between sequences in f of d and r is 2δ. For instance, if $\delta = 0$, the two proteins have not diverged since the LGT event, and thus the LGT is very recent.

2.2 Algorithm

DLIGHT identifies LGT events by considering, in all families of orthologs, all potential pairs of donor and recipient species. For each configuration, a likelihood ratio test is performed between the hypothesis of a LGT (alternative hypothesis) and the hypothesis of no LGT (null hypothesis). Formally, the set of significant LGT events is given by:

$$LGT = \left\{ (f,d,r) \,|\, f \in F;\; d,r \in G(f);\; \underset{\delta \geq 0 \in \mathbb{R}}{\operatorname{argmax}} \left(2\ln \frac{l(f,d,r,\delta)}{l(f,d,r,\delta=\infty)} \right) > \chi^2(\alpha,1) \right\}$$

where F is the set of all families of orthologs, d a potential donor species, r a potential recipient species and $l(f,d,r,\delta)$ is the likelihood of an LGT in f from lineages of d and r at distance δ in the past. $l(f,d,r,\delta=\infty)$ is the likelihood under the null hypothesis (in which δ is fixed to ∞, see below), and $\chi^2(\alpha,1)$ is the critical value of the chi-square distribution with significance level α and one degree of freedom. This test is known as the likelihood ratio test (see e.g. [17]). The ratio follows a chi-square distribution if the two models are nested, which is the case here, as we shall see below.

Below, we show how the likelihood of a LGT event $l(f,d,r,\delta)$ can be computed. The process can be split in three parts: first, given (f,d,r,δ), we infer which species of $G(f)$ belong to the set of donor species $\mathcal{D}$ and of recipient species $\mathcal{R}$. From these sets, we show how to compute the expected values of all $2|f|-3$ evolutionary distances of pairs in f that involve r and/or d, as well as their variances and covariances. Finally, we compute the likelihood of the event, which is based on the deviation of the observed distances from the expected distances.

Step 1 – Assignment of Species to Sets of Donors ($\mathcal{D}$) and Recipients ($\mathcal{R}$). First, given a quartet (f,d,r,δ), we infer members of $G(f)$ belonging to the donor and recipient lineages, that is, the set of species that directly descend from the donor (set $\mathcal{D}$) and recipient species (set $\mathcal{R}$). These subsets of $G(f)$ can be defined as follows:

$$\mathcal{D} = \{j \in G(f) \mid \tau_f \cdot d(j,d) \leq 2\delta\}$$

$$\mathcal{R} = \{j \in G(f) \mid \tau_f \cdot d(j,r) \leq 2\delta\}$$

We shall now justify these definitions (illustrated in Fig. 1). First, note that as could be reasonably expected, $d \in \mathcal{D}, r \in \mathcal{R}$, because in both cases the distance to themselves is 0, and δ being a distance is non-negative. As for the other species of $G(f)$, the definitions use assumption 4 (we focus on the definition of $\mathcal{D}$; the rational for $\mathcal{R}$ is similar): if all sequences from the donor lineage in f evolve at the same rate, they will all be δ away from the LGT. Further, by definition, members of the donor lineage have speciated after the LGT event, and therefore, their sequences in f are separated by a distance of at most 2δ.

To build these sets, we must rely on the estimators $\hat{\tau}_f$ and $\hat{d}(j,d)$ (or $\hat{d}(j,r)$ in the case of $\mathcal{R}$). Since the interspecies distances are point estimates (assumption 2), we only need to consider the distribution of $\hat{\tau}_f$ (see Sect. 2.1): the sets of donors and recipients differ depending on the value of the estimator $\hat{\tau}_f$. Fig. 2 depicts the distribution with the critical values of $\hat{\tau}_f$ for the assignment of a species j to $\mathcal{D}$ and $\mathcal{R}$.

Thus, if we consider the two critical values for all species j in $G(f)$, the distribution of $\hat{\tau}_f$ will be partitioned into $2|f|+1$ ranges. Each of these ranges map to particular $\mathcal{D}_i$ and $\mathcal{R}_i$, whose probability is the area of the density function $\text{pdf}(\hat{\tau}_f)$ in that particular range. We refer to the probability of the ith range as p_i. We will compute for each of these sets of donors and recipients the corresponding likelihood, and then average them according to their probability. The next step is therefore repeated for all $2|f|+1$ possible assignments of $\mathcal{D}, \mathcal{R}$.

Step 2 – Pairwise Distance Statistics. Given a sextet $(f,d,r,\delta,\mathcal{D}_i,\mathcal{R}_i)$, the computation of the likelihood of a particular LGT event is based on the $2|f|-3$ pairwise distances in f that involve d or r. These distances are of interest because

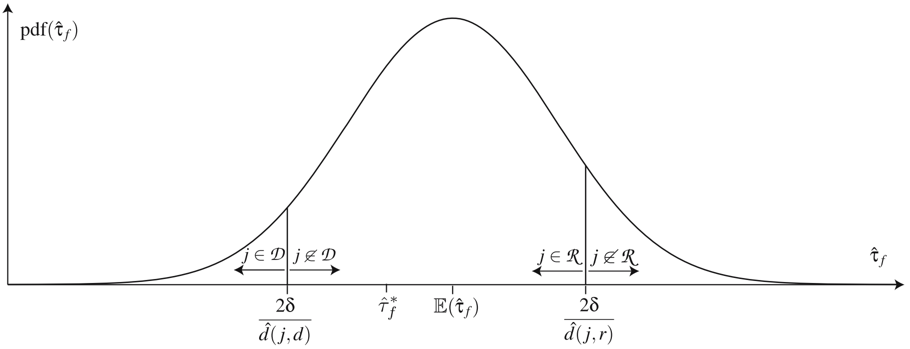

Fig. 2. The assignment of sequence j to sets $\mathcal{D}, \mathcal{R}$ depends on $\hat{\tau}_f$. For instance, at the point $\hat{\tau}_f^*$, j is in $\mathcal{R}$, but not in $\mathcal{D}$.

they are particularly altered by the LGT event, but the procedure could trivially be extended to all $\binom{|f|}{2}$ pairs in f.

The *observed* distances are simply the ML estimators for the relevant pairs of sequences of f. Estimators for the *expected* distances are provided in Table 1. Most distances involving the donor species d are unaffected by the LGT event, i.e. they are expected to follow the interspecies distances scaled by the family rate. Distances to the recipient species r however are mostly expected to follow the scaled interspecies distance to the donor d, because the sequence originated from the donor lineage, and after the LGT event, they evolved at the same rate as in the donor lineage (assumption 4). The special cases are: (i) distances between two recipients: they are unaffected by the LGT because the transfer happened before they speciated; (ii) distances between recipient and donor species: they are expected to be 2δ per definition; (iii) distances involving *inconsistent* species: the estimators and parameters can be such that a species is in both $\mathcal{D}_i$ and $\mathcal{R}_i$, for instance if δ is particularly large. In those cases, we treat the distance the same way as under the null hypothesis (no LGT transfer) and assign it an expected value that corresponds to the scaled interspecies distance. In terms of the model, this also has the advantage that the null hypothesis of no LGT is equivalent to the special case of a LGT with parameter $\delta = \infty$. This means that the models are nested, and therefore that the likelihood ratio follows a chi-square distribution with number of degree of freedom given by the difference in free parameters (one in our case).

Note that in our model, both observed and expected pairwise distances are normally distributed random variables, which can be expressed using two $2|f|-3$ dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. In both case, we have estimators for their variance-covariance matrices $\Sigma_{\boldsymbol{x}}$ and $\Sigma_{\boldsymbol{y}}$: for observed distances, the diagonal entries can be obtained by ML theory, and the covariances can be computed as described in [16]. As for the expected distances, the variance is either that of $\hat{\tau}_f$ scaled appropriately, or else null when $\hat{\tau}_f$ does not appear in the expression. The expected distances do not covary, and thus all off-diagonal entries are null.

Table 1. LGT event: expected distances to r and d in f. Note that the last row (*inconsistent*) can occur in our model if δ is large; the adverse impact of such inherently inconsistent case is limited by using the same expected distances as under the null hypothesis (no LGT event).

label	$\in \mathcal{D}_i$	$\in \mathcal{R}_i$	$\hat{\mathbb{E}}(\hat{d}_f(j,d))$	$\hat{\mathbb{E}}(\hat{d}_f(j,r))$
outgroup	no	no	$\hat{\tau}_f \cdot \hat{d}(j,d)$	$\hat{\tau}_f \cdot \hat{d}(j,d)$
donor	yes	no	$\hat{\tau}_f \cdot \hat{d}(j,d)$	2δ
recipient	no	yes	2δ	$\hat{\tau}_f \cdot \hat{d}(j,r)$
inconsistent	yes	yes	$\hat{\tau}_f \cdot \hat{d}(j,d)$	$\hat{\tau}_f \cdot \hat{d}(j,r)$

Let $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{y}$. The vector $\boldsymbol{z}$ is normally distributed, with expected value $\mathbb{E}(\boldsymbol{z}) = \mathbf{0}$. If we now assume that $\boldsymbol{x}, \boldsymbol{y}$ are independent, $\Sigma_{\boldsymbol{z}} = \Sigma_{\boldsymbol{x}} + \Sigma_{\boldsymbol{y}}$. In reality, they are not strictly independent, because $\boldsymbol{x}$ is a component (albeit a minor one) of $\hat{\tau}_f$, which itself is used in the computation of $\boldsymbol{y}$.

Step 3 – Computation of the Likelihoods, and Estimation of $\boldsymbol{\delta}$. The likelihood of the LGT event $(f, d, r, \delta, \mathcal{D}_i, \mathcal{R}_i)$ can be computed from the multivariate normal probability density function of the vector $\boldsymbol{z}$ and covariance matrix $\Sigma_{\boldsymbol{z}}$ from the previous section:

$$l(f, d, r, \delta, \mathcal{D}_i, \mathcal{R}_i) = \frac{exp(-\frac{1}{2}\boldsymbol{z}^T \Sigma_{\boldsymbol{z}}^{-1} \boldsymbol{z})}{\sqrt{(2\pi)^{2|f|-3} |\Sigma_{\boldsymbol{z}}|}}$$

We can now marginalize over the $2|f|+1$ different sets of donors and recipients (see step 1) to compute the likelihood of the LGT event (f, d, r, δ):

$$l(f, d, r, \delta) = \sum_i p_i \cdot l(f, d, r, \delta, \mathcal{D}_i, \mathcal{R}_i)$$

Furthermore, the parameter δ can be estimated by maximizing the likelihood. As mentioned above, the likelihood for the null hypothesis of no LGT event is obtained by the special case with parameter $\delta = \infty$.

2.3 Model Violations and Test of Multivariate Normality

DLIGHT is based on assumptions that do not always hold, in particular when dealing with biological sequences whose evolution strongly deviates from the markovian model. To limit the adverse effect of such model violations, we test the multivariate normality of the data by computing a p-value based on the squared Mahalanobis distance $\boldsymbol{z}^T \Sigma_{\boldsymbol{z}}^{-1} \boldsymbol{z}$, which is known to be chi-square distributed if $\boldsymbol{z}$ is multivariate normal. Data falling in extreme quantiles are considered dubious. In experiments reported here, predictions with data falling in the $(1 - 10^{-10})$ quantile were considered artifacts due to model violation, and were disregarded.

Furthermore, in case of poorly estimated variances or covariances, the matrix $\Sigma_{\boldsymbol{z}}$ may not be positive definite, or it may be singular if the sequences of two species are identical. In our implementation, we still try to identify LGT events by working with a subset of the family that constitute a well-posed problem (the problematic sequences are excluded on the basis of a simple greedy approach).

2.4 Combination of Results and Correction for Multiple Testing

As we presented above, DLIGHT computes a likelihood ratio test in all families of orthologs, for all different possible pairs of potential donor and recipient species. This raises the issues of combining results and correcting for multiple testing. Currently, we take the conservative approach of combining results that are consistent, for instance when a LGT event happened before speciation of the recipient species into two species g_1 and g_2: the algorithm may detect a transfer when run with both species as recipient, but if in both cases the estimated δ suggests a transfer prior to their speciation, the prediction is consistent and can be combined. Another common case for combination are pairs of results that report LGT between consistent sets of donor and recipient genomes, but with reverse direction. The direction of some LGT events, such as transfers between close species, is inherently difficult to assess. Nevertheless, if one direction has a significantly higher probability, and provided that the estimated parameter δ is consistent, the direction of the LGT can be infered.

We address the issue of multiple testing by using the Bonferroni adjustment, a common approach that discounts the significance by a factor corresponding to the total number of tests. If the tests are not independent from each other, which is the case here, the correction is excessive and some sensitivity is wasted.

3 Validation and Results

DLIGHT was tested using four different approaches: simulation, artifical LGT events, real biological data and comparison with previous results from the literature. The results of simulation are also reported for three simple LGT detection methods that serve as benchmark: methods based on GC-content, best-hits, and perturbed-distances. They are described in the *Appendix.*

3.1 *In Silico* Evolution Scenarios

Although a simulation will never fully capture the complexity and diversity of natural evolutionary processes, it allows the evaluation of algorithms with knowledge of the history of events, and therefore constitutes a tracable baseline. Synthetic genomes were generated using the software *EVA* (manuscript in preparation). EVA starts from a single organism and simulates the following evolutionary mechanisms: codon mutation based on empirical substitution probabilities [18], with biased genome-specific GC contents and gene-specific mutation rates, codon insertion and deletion, gene duplication, gene loss, LGT (both orthologous replacement and novel gene acquisition), and speciation. The probabilities of LGTs, gene duplications and gene loss were set to a proportion of 1:2:3, thereby keeping the expected number of genes constant (as suggested in [19]). The two types of LGT events, novel gene acquisition and orthologous replacement, were set to have an equal probability of occurence. Table 2 details the remaining parameters of the two different evolutionary scenarios investigated here. Genes from the resulting genomes were grouped in orthologous familes using the OMA algorithm [20].

Table 2. Overview of the simulation parameters. In *simulation 1* closely related organisms are used while in *simulation 2* more distantly related organisms are analysed.

Name	# of species	Avg. # of genes	Avg. genome distance (expect. identity)	# LGT	# of families
simulation 1	9	197	16 PAM (85.4%)	50	241
simulation 2	9	202	74 PAM (50.7%)	42	295

The different algorithms were run on the two datasets and the performances were analysed in terms of both sensitivity and specifity, at three levels of precision: first, the ability to report families of orthologs that contain at least one laterally transfered gene; second, the ability to identify the protein involved in a LGT event, that is, either report a *donor* or a *recipient* species; and third, the ability to correctly identify the direction of the LGT, in addition to the species involved. The six resulting ROC curves are presented in Fig. 3. Overall, DLIGHT showed significantly higher sensitivity and specificity than the other methods. It also performed more consistently than the other methods, with curves of similar shape across all experiments. The significance threshold is rather conservative (a consequence of the stringent Bonferoni correction) and led to 100% specifity in most cases. In the case in which the direction of LGTs was required, in distantly related species, the GC content and the perturbed distance approach outperformed DLIGHT. This may be due to the difficulties in estimating distances and variances when organisms are so far apart. In those cases, simpler methods may prove to be more robust.

3.2 Artificial LGT Events in Real Data

LGT events between real biological genomes can be simulated by introducing a gene from one species into another, either as substitute for its ortholog ("orthologous replacement") or as additional sequence. Such *artificially introduced* LGT event allows the testing of the algorithm on real biological data while having a positive control. However, only the specific case of very recently introduced genes can be simulated. Furthermore, real occurences of LGTs may already be present in the dataset and their signals may conflict with the artificially introduced ones.

The biological data consisted of 15 archaea with 2273 gene families, of which 727 families had at least 6 genes. 200 cases of LGT events from random donors to random recipients were introduced, as orthologous replacement, in families with at least 6 genes. Fig. 4 presents the results of the tests. The 200 top scoring

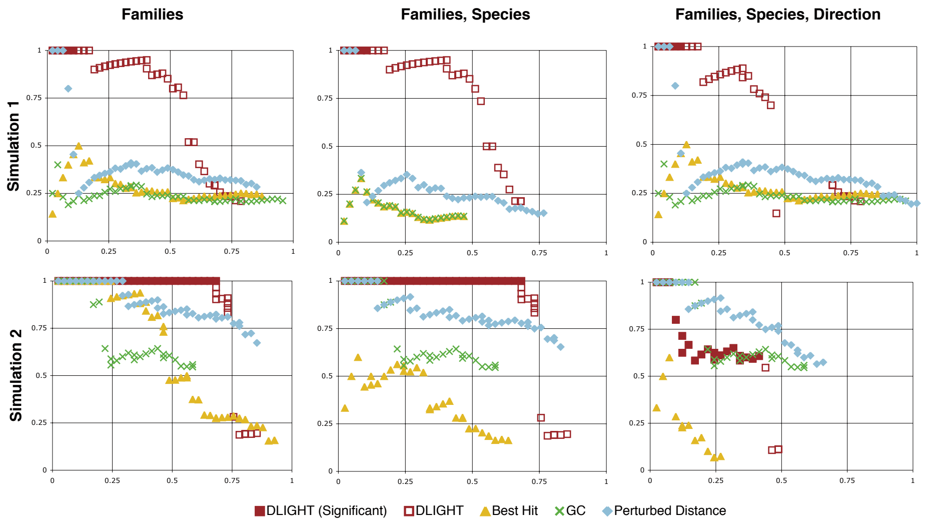

Fig. 3. ROC analyses. Sensitivity is ploted along the X axis, specifity along the Y axis. Plots on the first line were obtained from a simulation with closer species, plots on the second line from more distantly related ones. The left column shows results of identifying families with LGT events. The middle column shows results of identifying families with LGT events and the involved species. The right column shows results of identifying families with LGT events, the involved species and the direction of the transfers.

predictions were compared to the set of artificially introduced LGTs. Of all four methods, our performed best. Given the relatively good results obtained with the perturbed-distance approach in the previous test, its performance here is surprisingly poor, with only 7 artificial LGTs recovered. Note also that being recent, transfers introduced here constitue ideal conditions for both the GC method (the composition has not had time to adapt to the new host) and the best-hit approach (transfer after all speciations).

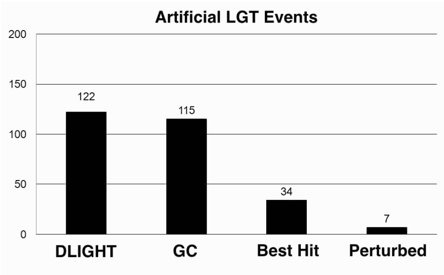

Fig. 4. Artificially introduced LGT. The number of such LGTs among the top 200 predictions is given.

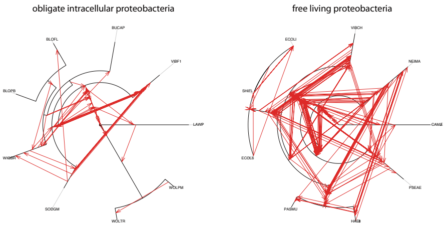

Fig. 5. LGT flow among proteobacteria. LGTs are drawn with arrows indicating the direction of the transfer. DLIGHT was run with the same parameters on both datasets individually.

3.3 Real Biological Data

LGT events are believed to happen throughout the prokaryotes, but not uniformly so. Some organisms are considered to be little affected by LGT while others are thought to have acquired many genes from distant species. Endosymbionts and endoparasites are micro-organisms that spend most of their life inside a host cell. As a consequence, for an LGT event to happen, foreign DNA would need to cross the membrane and defensive system of both the organism and its host. Therefore, such organisms are expected to have very few genes aquired through LGT compared to free living micro-organisms [21].

Our algorithm was verified against these observations by comparing predictions of two different datasets. We inferred LGTs for 9 endosymbionts[2] and for 9 free living pathogenic proteobacteria[3]. The organisms were classified according to HAMAP [22].

DLIGHT detected between 1 and 22 foreign genes (6.3 in average) in endosymbionts, and between 2 and 70 genes (40.7 in average) in free living bacteria. Normalized with the genome sizes, this gives between 0.15% and 0.89% percent of foreign genes in endosymbionts, versus 0.12% to 2.43% in free living. Thus, endosymbionts appear indeed to have lower LGT rates than their free living counterparts. In figure 5 the LGT events are indicated in both trees as thin lines and there too, the difference in LGT occurences is clearly visible.

The detected percentages of foreign genes is much lower than the values of 2% to 60% found in previous reports [23,24,25]. However, these higher numbers represent all genes received by any organism outside the vertical genealogy, while our data reflect only gene transfer among 9 bacteria.

A larger set with 15 archaea[4] consisting of 2273 orthologous families was analysed in a similar way. The average LGTs per gene was at 1.07%, with 292 detected LGT events in all 15 archaea. The number of acquired genes varies from 1 for *Nanoarchaeum equitans* to 37 for *Methanosarcina mazei* . Looking at the relative gene uptake with regard to the genome size, *Nanoarchaeum equitans* still recieved the fewest genes with 0.19%. *Thermoplasma volcanium* received the most genes with 2.4%. It has been proposed previously that LGT is common between Thermoplasmatales and Sulfolobales [1]. In our dataset, *Thermoplasma*

[2] Candidatus Blochmannia floridanus, Blochmannia pennsylvanicus (strain BPEN), Buchnera aphidicola (subsp. Schizaphis graminum), Lawsonia intracellularis (strain PHE/MN1-00), Sodalis glossinidius (strain morsitans), Vibrio fischeri (strain ATCC 700601 / ES114), Wigglesworthia glossinidia brevipalpis, Wolbachia pipientis wMel, Wolbachia sp. (subsp. Brugia malayi) (strain TRS)

[3] Campylobacter jejuni, Escherichia coli O6, Escherichia coli, Haemophilus influenzae (strain 86-028NP), Neisseria meningitidis serogroup A, Pasteurella multocida, Pseudomonas aeruginosa, Shigella flexneri, Vibrio cholerae

[4] Methanocaldococcus jannaschii, Methanosarcina mazei, Pyrobaculum aerophilum, Sulfolobus solfataricus, Methanosarcina acetivorans, Aeropyrum pernix, Archaeoglobus fulgidus, Halobacterium salinarium, Methanobacterium thermoautotrophicum, Methanopyrus kandleri, Pyrococcus horikoshii, Thermoplasma volcanium, Nanoarchaeum equitans, Thermoplasma acidophilum, Methanococcus maripaludis

volcanium exchanged 14 genes with *Sulfolobus solfataricus* and *Thermoplasma acidophilum* also 14 genes with *Sulfolobus solfataricus*. This is significantly more than the 3.6 average LGTs between archaea.

In addition to these tests, DLIGHT was applied to a dataset of 10 mammals[5]. Although LGT between higher eukaryotes and bacteria are found by some authors, we are not aware of any case of LGT between two mammals. Mammals serve therefore as negative control for our LGT detection method. Indeed, DLIGHT did not detect any LGT among the 10 mammals.

3.4 Comparision with Previous Results

Results from different LGT inference approaches can be very inconsistent, with overlaps at times smaller than expected by random [26]. This is particularily true when comparing the results of parametric and phylogenetic methods. Thus, the results of DLIGHT were compared with two studies based on phylogenetic approaches.

Comparison with Zhaxybayeva *et al.* (2006). In [27], the authors used an embedded quartet decomposition analysis to search events of LGT in 11 completey sequenced cyanobacteria. Orthologs were grouped via reciprocal top-scoring blast hits, resulting in families with few paralogs. A set of 1128 ortholgous genes was found to be present in at least nine of the 11 cyanobacterial genomes and taken as input for the LGT search. Within the group of cyanobacteria, 135 LGTs were detected, mostly between *Gloeobacter violaceus* and *Synechococcus elongatus* (45) and *Prochlorococcus marinus* SS120 and *Prochlorococcus marinus* (strain MIT 9313) (28).

We tried to confirm the predictions of LGT in these 135 families using DLIGHT. In 54 families (40%), significant LGTs were reported. In 32 of them, the species predicted to be involved were either the same, or in agreement with the trees constructed by [27]. The 22 other predictions were conflicting with their trees. Additionally, it should be noted that the interspecies distances estimated by DLIGHT were computed on the basis of these 135 families, none of which is congruent to the species tree according to [27]; this suggests that DLIGHT is relatively robust with respect to perturbations in the data.

Comparison with Beiko *et al.* (2005). DLIGHT was compared with results from [10], a large scale LGT inference study using an explicit phylogenetic method. For 22,437 families of proteins in 144 genomes, they constructed gene trees and compared in each tree all bifurcations to a reference species tree. They reported bifurcations with significant posterior probability (PP), classified in either consistent or conflicting with the species tree.

A subset of their 8,315 protein families of size up to 15 sequences was randomly selected. Based on their bifurcation analysis, these familes were partitioned in

[5] Homo sapiens, Mus musculus, Canis familiaris, Rattus norvegicus, Bos taurus, Pan troglodytes, Monodelphis domestica, Macaca mulatta, Loxodonta africana, Oryctolagus cuniculus

four categories: *i.* 28.5% families with strong support of no LGT (all bifurcations consistent with species tree with $PP \geq 0.95$), *ii.* 38.4% families with mild support of no LGT (no conflicting bifurcation with $PP \geq 0.5$), *iii.* 15.2% families with mild support of LGT (at least one conflicting bifurcation with $PP \geq 0.5$, none with $PP \geq 0.95$), and *iv.* 17.8% families with strong support of LGT ($PP \geq 0.95$).

DLIGHT was run on this dataset, with, as sole input, the protein sequences labeled with family and species identifiers. The computation of all pairwise evolutionary distances within families required about 2 days on a single AMD Opteron 1.8 GHz. DLIGHT used another day to predict significant LGT events, which were found in 634 families. The distribution of inferred LGT events among the four categories defined from their predictions was as follows: *i.* 7.1%, *ii.* 13.1%, *iii.* 19.2%, and *iv.* 60.6%. As almost 80% of the predictions are the same, the level of agreement between the two methods is quite high, especially considering the large differences in methodologies.

4 Conclusion

In this article, we introduce a new implicit phylogenetic method for LGT detection, based on pairwise evolutionary distances in a probabilistic framework. Validation shows that it compares favorably with existing parametric and implicit phylogenetic methods. Furthermore, its advantages over explicit phylogenetic methods include speed and lack of reliance on multiple sequence alignments and gene tree inference.

There are, though, a number of aspects that could be the object of further improvement: the sensitivity could be increased by the computation of the likelihoods using all pairwise distances within gene families, and not only the distances to the transfered genes; confidence intervals in the estimation of the interspecies distances. Instead of the approximation of multivariate normality, and at expense of increased time complexity, the distribution of the distances could possibly be estimated in an MCMC framework.

Acknowledgements

The authors thank Manual Gil, Adrian Schneider, Sukman Dessimoz, Gina Cannarozzi, Maria Anisimova, Susan Holmes, Dan Graur, and three anonymous reviewers for helpful comments and suggestions.

References

1. Philippe, H., Douady, C.J.: Horizontal gene transfer and phylogenetics. Curr. Opin. Microbiol. 6, 498–505 (2003)
2. Lawrence, J.G., Ochman, H.: Reconciling the many faces of lateral gene transfer. Trends Microbiol. 10, 1–4 (2002)

3. Lawrence, J.G., Ochman, H.: Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. 44, 383–397 (1997)
4. Lawrence, J.G., Ochman, H.: Molecular archaeology of the Escherichia coli genome. Proc. Natl. Acad. Sci. U S A 95, 9413–9417 (1998)
5. Karlin, S.: Global dinucleotide signatures and analysis of genomic heterogeneity. Curr. Opin. Microbiol. 1, 598–610 (1998)
6. Moszer, I., Rocha, E.P., Danchin, A.: Codon usage and lateral gene transfer in Bacillus subtilis. Curr. Opin. Microbiol. 2, 524–528 (1999)
7. Mrazek, J., Karlin, S.: Detecting alien genes in bacterial genomes. Ann. N. Y. Acad. Sci. 870, 314–329 (1999)
8. Medigue, C., Rouxel, T., Vigier, P., Henaut, A., Danchin, A.: Evidence for horizontal gene transfer in Escherichia coli speciation. J. Mol. Biol. 222, 851–856 (1991) (Comparative Study)
9. Hamady, M., Betterton, M.D., Knight, R.: Using the nucleotide substitution rate matrix to detect horizontal gene transfer. BMC Bioinformatics 7, 476 (2006)
10. Beiko, R.G., Harlow, T.J., Ragan, M.A.: Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. U S A 102, 14332–14337 (2005) (Comparative Study)
11. Gophna, U., Ron, E.Z., Graur, D.: Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events. Gene 312, 151–163 (2003)
12. Lawrence, J.G., Hartl, D.L.: Inference of horizontal genetic transfer from molecular data: an approach using the bootstrap. Genetics 131, 753–760 (1992)
13. Clarke, G.D.P., Beiko, R.G., Ragan, M.A., Charlebois, R.L.: Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. J. Bacteriol. 184, 2072–2080 (2002)
14. Koski, L.B., Golding, G.B.: The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. 52, 540–542 (2001)
15. Pupko, T., Huchon, D., Cao, Y., Okada, N., Hasegawa, M.: Combining multiple data sets in a likelihood analysis: which models are the best?. Mol. Biol. Evol. 19, 2294–2307 (2002)
16. Susko, E.: Confidence regions and hypothesis tests for topologies using generalized least squares. Mol. Biol. Evol. 20, 862–868 (2003)
17. Felsenstein, J.: Inferring Phylogenies. Sinauer Associates Inc., Sunderland (2004)
18. Schneider, A., Cannarozzi, G.M., Gonnet, G.H.: Empirical codon substitution matrix. BMC Bioinformatics 6 (2005)
19. Kunin, V., Ouzounis, C.A.: The balance of driving forces during genome evolution in prokaryotes. Genome Res. 13, 1589–1594 (2003)
20. Dessimoz, C., Cannarozzi, G., Gil, M., Margadant, D., Roth, A., Schneider, A., Gonnet, G.: OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: Introduction and first achievements. In: McLysaght, A., Huson, D.H. (eds.) RECOMB 2005. LNCS (LNBI), vol. 3678, pp. 61–72. Springer, Heidelberg (2005)
21. Lawrence, J.G., Hendrickson, H.: Lateral gene transfer: when will adolescence end? Mol. Microbiol. 50, 739–749 (2003)
22. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M.: The swiss-prot protein knowledgebase and its supplement trembl in 2003. Nucleic Acids Res. 31, 365–370 (2003)

23. Lerat, E., Daubin, V., Ochman, H., Moran, N.A.: Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. 3, e130 (2005)
24. Ge, F., Wang, L.S., Kim, J.: The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. 3, e316 (2005)
25. Dagan, T., Martin, W.: Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. Proc. Natl. Acad. Sci. USA 104, 870–875 (2007)
26. Ragan, M.A.: On surrogate methods for detecting lateral gene transfer. FEMS Microbiol. Lett. 201, 187–191 (2001)
27. Zhaxybayeva, O., Gogarten, J.P., Charlebois, R.L., Doolittle, W.F., Papke, R.T.: Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. 16, 1099–1108 (2006)

Appendix

4.1 Benchmark Methods

The three benchmark methods used in the validation section are described here. All three consist of a scoring function which is used to rank all genes as potentially laterally transfered candidates.

GC Content. The GC method used in this paper is a basic implementation of this common parametric approach. A more advanced implementation can be found in [3]. The version used here considers the GC content on the first and third codon position, without performing a codon usage analysis. The score for a gene x in a species X is computed as follows:

$$S_{GC}(x) = \frac{(GC(x,1) - \mu_{GC}(X,1))^2}{\sigma^2_{GC}(X,1)} + \frac{(GC(x,3) - \mu_{GC}(X,3))^2}{\sigma^2_{GC}(X,3)}$$

where $GC(x,i)$ is the average GC content of the gene x at its ith codon position, and $\mu_{GC}(X,i), \sigma^2_{GC}(X,i)$ the average and variance of GC content among all ith codon position of genes in species X.

Best Hit Approach. The best hit method infers LGT when the highest scoring hit of a particular sequence is in a distant species [13]. Our implementation improves this idea by considering the shortest evolutionary distance rather than the top similarity score. More precisely, the score of a gene x from a species X and family of orthologs f is computed as follows:

$$S_{BH}(x) = \frac{Rank_f(T)}{|f|}$$

where T is the organism in which x has its closest homolog, $Rank_f(T)$ the rank of T among the species represented in f ordered by increasing average interspecies distance to X.

Perturbed-Distances Approach. The third method detects LGT using the same underlying idea as DLIGHT – the discrepancy between gene and interspecies pairwise distances that results from an LGT event – but in a much cruder way: the score of a gene x from an species X, in family f is

$$S_{PD}(x) = \frac{1}{|f| - 1} \sum_{y \in f, y \neq x} (d(x,y) - d(X,Y))$$

where $d(x,y)$ denotes the evolutionary distance between genes x and y, $d(X,Y)$ the interspecies distance between X and Y.