Research Article

A Likelihood Framework to Measure Horizontal Gene Transfer

Simone Linz*, Achim Radtke*, and Arndt von Haeseler†‡$¶

Institute of Research: Department of Bioinformatics, Heinrich-Heine-University, Universitätsstr. 1, 40225 Düsseldorf, Germany

*Department of Bioinformatics, Heinrich-Heine-University, Universitätsstr. 1, 40225 Düsseldorf, Germany; †Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, Dr. Bohr-Gasse 9, 1030 Vienna, Austria; ‡University of Vienna; $Medical University of Vienna; and ¶University of Veterinary Medicine Vienna

Corresponding author:
Simone Linz, Department of Bioinformatics, Heinrich-Heine-University, Universitätsstr. 1, 40225 Düsseldorf, Germany, Tel.: +49 (0)211 8112163, Fax: +49 (0)211 8115767, Email: linz@cs.uni-duesseldorf.de

Key words: horizontal gene transfer, gene tree, species tree, non-tree-like evolution

Running head: A rate of horizontal gene transfer

Abbreviations: HGT, horizontal gene transfer; COG, Cluster of Orthologous Groups of Proteins

Abstract

We suggest a likelihood based approach to estimate an overall rate of horizontal gene transfer (HGT) in a simplified setting. To this end, we assume that the number of occurring HGT events within a given time interval follows a Poisson process. To obtain estimates for the rate of HGT, we simulate the distribution of tree topologies for different numbers of HGT events on a clocklike species tree. Using these simulated distributions, we estimate an HGT rate for a collection of gene trees representing a set of taxa. As an illustrative example, we use the "Clusters of Orthologous Groups of proteins (COGs)". We also perform a correction of the estimated rate taking into account the inaccuracies due to gene tree reconstructions. The results suggest a corrected HGT rate of about 0.36 per gene and unit time. In other words, eleven HGT events have occurred on average among the 44 taxa of the COG species tree. A software package to estimate an HGT rate is available online (http://www.cibiv.at/software/hgt/).

Introduction

It is well known that gene trees reconstructed for different genetic loci for the same set of taxa do not necessarily agree. Their branching pattern may be different from each other and different from the species tree (Pamilo and Nei, 1988). These discrepancies are not always due to the uncertainty of the phylogenetic inference method, but rather due to biological processes like hybridization, gene duplication and deletion, or horizontal gene transfer (HGT) (Syvanen, 1994). In the following, we will focus on the latter of these processes.

The effect of one HGT event is visualized in figure 1 which shows a species tree (fig. 1A) of the five taxa A, B, C, D and E. This tree indicates a close relation between A and the cluster of B and C. In many cases, the species tree also explains the phylogeny of single genes, but sometimes a gene has a different evolutionary history than the species tree (Pamilo and Nei, 1988). For such a gene the gene tree is displayed in figure 1B. One possible explanation for this kind of difference is HGT, in this case from A to D. During such a process, a piece of DNA (e.g. a gene) is transferred from one organism to another which is not its offspring. The genetic material is stably incorporated in the acceptor genome, in contrast to the vertical inheritance of genes by descent from one's parents (Bushman, 2002). In the depicted case, the arrow shows gene transfer from species A to species D. As a consequence, the gene tree for this gene shows a close relationship between A (donor) and D (acceptor).

HGT is known as an important mechanism to shape the genomes of bacteria (Ochman *et al.*, 2000; Boucher *et al.*, 2003), but recently there is also an accumulation of data indicating that this process occurs in the evolution of eukaryotes (de la Cruz and Davies, 2000; Andersson, 2005) and archaea (Nelson *et al.*, 1999; Diruggiero *et al.*, 2000) too.

Several approaches have been published that discover single HGT events (Lerat *et al.*, 2003, 2005), whereas another kind of approach estimates the amount of genes that are acquired through HGT for a given genome. The latter type of analysis is reviewed in Ochman *et al.* (2000) for 19 completely sequenced genomes. In these species, the amount of adopted genes varies between virtually none in organisms with small genome size, for example

Rickettsia prowazekii, *Borrelia burgdorferi*, and *Mycoplasma genitalium*, to nearly 17 % in *Synechocystis* PCC6803 (Ochman *et al.*, 2000). Another way of detecting horizontally transferred genes uses bacterial genome sequences to examine the nucleotide composition (GC content) and usage of different codons (Lawrence and Ochman, 1997).

In contrast to these approaches, we estimated an overall rate of HGT for a given set of species based on simulating a likelihood curve for the reconstructed species tree. We constructed a clocklike species tree reflecting the actual evolutionary pathways of the investigated organisms (Pamilo and Nei, 1988) and simulated different numbers of HGT events, implemented as series of subtree prune and regraft processes on the species tree (Hillis *et al.*, 1996). Simulations with different numbers of HGT lead to a distribution of tree topologies that are comparable with the gene tree distribution to estimate an HGT rate supported by a likelihood framework.

As the number of tree topologies increases exponentially with the number of leaves, the probability to get a specific topology is very low. To overcome this, we worked with quartet subtrees instead of the complete gene tree topologies (see the Materials and Methods section).

To apply the method to real data, we used the "Cluster of Orthologous Groups of proteins (COGs)" (Tatusov *et al.*, 2001).

Materials and Methods

Notation

In the following, we introduce some mathematical notation needed for the analysis. For more detailed definitions, we refer the reader to Semple and Steel (2003).

A phylogenetic tree $T = (\mathcal{V}, \mathcal{E})$ is described by a set of vertices $\mathcal{V}$ (also called nodes) and a set of edges $\mathcal{E} \subset \{(x, y) : x \neq y, x, y \in \mathcal{V}\}$. The degree of a vertex v, denoted by $deg(v)$, is the number of edges incident with v. Any node with $deg(v) = 1$ is called a leaf and all other vertices are called internal. With $\mathcal{L}(T)$ we denote the set of all leaves. An unrooted phylogenetic tree T is called binary if each node $v \in \mathcal{V}$ is either a leaf or has

$deg(v) = 3$. If the binary phylogenetic tree is rooted, this definition must be extended to the most recent common ancestor of all leaves, which is called the root r. This vertex r is the only one in a rooted binary phylogenetic tree with $deg(r) = 2$.

Let $S = \{s_1, \ldots, s_n\}$ be a set of n taxa and let X be an arbitrary subset of S. To describe all binary tree topologies with leaves which are labeled with taxa of X, we use $\mathcal{T}(X)$ whereas $\tau(X)$ denotes an element of $\mathcal{T}(X)$. The tree $T(S) \in \mathcal{T}(S)$ is the species tree of $s_1, \ldots, s_n$ with a length function l, which defines the branch length of each edge $e \in \mathcal{E}_{T(S)}$. The size $L(T(S))$ of a tree is the sum over all branch lengths. We assume a species tree $T(S)$ to be binary, rooted, leaf-labeled, and clocklike (each species (leaf of the tree) has the same distance to the root), and we interpret the distance between any node and the root as time which has passed since the first split at the root. A gene tree is a tree topology of a leaf-labeled tree which evolves within a species tree and comprises at most all taxa of S.

The restriction of S to X, denoted by $\tau(S|X)$, is a tree topology derived from $T(S)$ where all leaves in $S \setminus X$ are ignored and all vertices with $deg(v) = 2$ are suppressed, except the root (fig. 1C).

Modeling Horizontal Gene Transfer

To model the process of HGT, we have to make some pivotal assumptions: 1. A binary, leaf-labeled, rooted, and clocklike species tree $T(S)$ is known, as well as all splitting times along this tree. 2. Differences between a gene tree and $T(S)$ are only caused by HGT events. 3. The transfer rate λ is homogeneous per gene and unit time. 4. Genes are transferred independently. 5. One copy of the transferred gene still remains in the donor genome. 6. The transferred gene replaces any existing ortholog counterpart in the acceptor genome.

As described before, the effect of HGT can result in a branching pattern of a gene tree which differs from the species tree (fig. 1). From a computational point of view, we model each HGT event as a subtree prune and regraft process (Hillis *et al.*, 1996). This means an HGT event is modeled in the following way: As we assume a homogeneous

HGT rate λ, the transfer events are uniformly distributed along all branches of the tree. For each HGT event, we randomly choose a starting point in the clocklike species tree, determine the corresponding time in this tree, search for all branches that exist at that point of time, and randomly select one as acceptor branch. Consequently, single transfer events between species are only possible if they coexist in time in order to prevent gene transfer from present-day species to fossils. This biologically motivated restriction is not considered in most current research on HGT models (e.g. Suchard, 2005). Furthermore, it is easily seen that not every single HGT event changes the branching pattern of the species tree, for example, if the process takes place between branches that share the most recent common ancestor.

For a given species tree with total length $L(T(S))$ and fixed λ, the tree topology $\tau(S)$ occurs with a certain probability $P(\tau(S) \mid T(S), \lambda, L(T(S)))$. In the following, we write $P(\tau(S) \mid \lambda)$ instead, because λ is the parameter of interest. As stated in the introductory part, the number of HGT events is Poisson distributed with parameter $\Lambda = \lambda \cdot L(T(S))$ for a fixed species tree. Thus, the probability for $\tau(S)$ given λ is

$$P(\tau(S)|\lambda) = \sum_{h=0}^{\infty} \Big(\frac{e^{-\Lambda} \cdot \Lambda^h}{h!} \cdot P(\tau(S) \mid \mathrm{HGT} = h) \Big). \tag{1}$$

The Poisson distribution describes the probability for h HGT events to happen on the species tree $T(S)$ with $L(T(S))$ and λ, whereas the second factor is the probability to observe $\tau(S)$ as tree topology after h HGT events. Although the Poisson distribution is easy to calculate, the probability distribution of the gene trees for a fixed number of HGT events appears hard to calculate, except for trivial cases like $h \in \{0, 1\}$. Moreover, for a fixed arbitrary subset $X \subset S$, we can compute the probability for each subtree $\tau(X)$ as follows:

$$P(\tau(X)|\lambda) = \sum_{\tau(S) \in \mathcal{T}(S)} \Big(\delta_{(\tau(X), \tau(S|X))} \cdot P(\tau(S)|\lambda) \Big). \tag{2}$$

The Kronecker delta $\delta_{(\tau(X),\tau(S|X))}$ is 1 if the topology of the induced subtree $\tau(S|X)$ with respect to $X \subset S$ is identical to $\tau(X)$, otherwise it is 0.

Equations 1 and 2 allow the estimation of λ in a likelihood framework. Therefore, we assume that λ acts on each gene independently. If m gene trees $\tau_1(S), \ldots, \tau_m(S)$ are

reconstructed, the likelihood of λ is

$$\text{lik}(\lambda|\tau_1(S),\ldots,\tau_m(S)) = \prod_{i=1}^{m} P(\tau_i(S)|\lambda). \tag{3}$$

We maximize equation 3 with respect to λ, which is interpreted as the most likely transfer rate.

This approach turns out to be computationally infeasible because a reliable estimation of $P(\tau(S)|\lambda)$ is only possible for a small number of taxa. Hence, we resort to an approximation of the likelihood. We consider a collection of subsets $X_1,\ldots,X_m \subseteq S$ together with the probability distribution induced by equation 2 and the simplified situation that the occurrencies of gene trees $\tau(X_1),\ldots,\tau(X_m)$ are mutually independent for different randomly chosen subsets $X_1,\ldots,X_m$. In this case, the joint probability of $\tau(X_1),\ldots,\tau(X_m)$ is given by

$$P(\tau(X_1),\ldots,\tau(X_m)) \approx \prod_{i=1}^{m} P(\tau(X_i)|\lambda). \tag{4}$$

Although equation 4 is an approximation to equation 3, the simulations show that it is good enough for the practical application and we can also apply the described estimation scheme to estimate $\hat{\lambda}$ and $\hat{\Lambda}$, respectively.

Estimating the Probability Distribution of Gene Trees

From the previous paragraph, it is obvious that it would be very difficult to find an analytical expression for any of the equations. However, equation 1 suggests an efficient simulation. For any fixed number h of HGT events, we can approximate the distribution $P(\tau(S)|\text{ HGT} = h)$ reasonably well. Therefore, we simulate $N = 100,000$ times h HGT events on the species tree with $0 \leq h \leq 60$ and calculate how often each gene tree occurs in the simulated trees. We end up with a probability distribution $P^*(\cdot)$ in which each column represents one gene tree and each row a fixed number of HGT events. The final likelihood estimation is based on $P^*(\cdot)$.

Although $P(\tau(S)|\lambda)$ can be estimated for small taxa sets, it gets intractable for biologically interesting numbers because too many tree topologies exist and it is almost impossible to simulate enough trees for a reliable estimation within a reasonable time

span. In such situations, the probability for different subsets $X \subseteq S$ proves more successful. Thus, we reduce the calculated probability distribution $P^*(\cdot)$ to a subset of randomly chosen quartet topologies of the given set of gene trees.

The COG Data

The whole data set, which is available via the NCBI website (http://www.ncbi.nlm.nih.gov/), comprises 3,167 protein families of 44 species (2 eukaryotes, 9 archaea, and 33 bacteria). As we concentrated on single-copy genes up to now, we only extracted those families which fulfill this criterion. To obtain enough phylogenetic information (Nei, 1996) to reconstruct the gene trees, we only used COG families with a minimum alignment length of 100 amino acids for each of the corresponding proteins. We also required at least four species per COG family. After applying these three criteria, 780 protein families still remained (see Supplementary Material S1 and S2). For each of these families a gene tree was reconstructed with TREE-PUZZLE (Schmidt *et al.*, 2002), using the Dayhoff substitution model (Dayhoff *et al.*, 1978).

Species Tree Reconstruction

To construct the species tree of the mentioned 780 protein families, we built all three binary trees for all possible quartets (A, B, C, D) and computed the corresponding log-likelihood values ℓ as sum of the log-likelihoods of the COG families (g_i) each represented by a gene tree.

$$\begin{aligned} \ell(AB|CD) &= \sum_{i=1}^{780} \ell_{g_i}(AB|CD) \\ \ell(AC|BD) &= \sum_{i=1}^{780} \ell_{g_i}(AC|BD) \\ \ell(AD|BC) &= \sum_{i=1}^{780} \ell_{g_i}(AD|BC). \end{aligned} \tag{5}$$

All three log-likelihood values ℓ_{g_i} are set to 0 if at least one of the species A, B, C, or D does not occur in the corresponding COG family g_i.

Afterwards, we used TREE-PUZZLE (Schmidt *et al.*, 2002) to construct the species tree

topology of the log-likelihood values of all $\sum_{i=1}^{780} \binom{|g_i|}{4}$=184,521,526 quartet topologies ($|g_i|$ is the number of taxa represented by the COG family g_i).

To assign branch lengths to this topology, we performed a clock test (Felsenstein, 1988) for all 780 protein families. The result contained 443 clocklike and 337 non-clocklike COG families. Only three of all families occurred in all 44 species (COG0013: Alanyl-tRNA synthetase, COG0092: Ribosomal protein S3, COG0541: Signal recognition particle GTPase Ffh), but none of them evolved clocklike. Therefore, we had to use an appropriate set of gene trees which covers all 44 species. For each clocklike evolving COG family with taxa set X, we reconstructed the corresponding subtree $\tau(S_{\mathrm{COG}}|X)$ with a total branch length measured in numbers of substitutions per site. Furthermore, we identified a set G of subtrees fulfilling the following conditions: (a) G covers the species tree completely and (b) each branching point is determined by at least one subtree. Such a coverage was found for the three clocklike evolving families: COG0419 (ATPase involved in DNA repair), COG0173 (Aspartyl-tRNA synthetase) and COG1242 (uncharacterized FeS oxidoreductases). As some of the splitting times are given by two or three of the named families and each of them evolved with a different rate, we computed the ratio of these rates to estimate the splitting times relative to one protein family, in this case COG0419.

Eventually, the reconstructed species tree $T(S_{\mathrm{COG}})$ was used to simulate distributions of tree topologies for different numbers of HGT events.

Comparing Trees

To compare the most frequent gene tree with the species tree, we extracted all quartet topologies from the 780 gene trees and added up the information in a descending sorted list representing each topology by the number of its occurrence. Afterwards, we built a quartet set that finally consisted of 35.7 % of the initially extracted quartet topologies. Starting with the most frequent topology, we put each one successively in the set if the corresponding quartet tree does not contradict a quartet tree already in the current set. For the final quartet set, we reconstructed a tree using TREE-PUZZLE. To compare the obtained tree topology with the COG species tree, we built a consensus tree using the

program CONSENSE of the PHYLIP package (Felsenstein, 1989).

Results

Quality Check

To obtain reliable simulation and estimation results, we repeated the procedure for different parameter settings. Hence, we used a program that simulates an HGT rate λ on a clocklike species tree. The corresponding number of HGT events was drawn from the Poisson distribution. This kind of simulation generates a new data set, which is comparable to the 780 gene trees of the COG data. As we know the true HGT rate λ, we can check the reliability of the estimation procedure.

First of all, we estimated the probability $P^*(\tau(X) \mid \text{HGT} = h)$ to get the tree topology $\tau(X)$ if exactly h HGT events happened on the species tree $T(S)$ with 44 taxa for the COG data. Thus, we simulated N times h events on $T(S)$ and assumed that $P^*(\tau(X) \mid \text{HGT} = h)$ is the relative occurrence of the topology $\tau(X)$.

To analyze the influence of the size of the quartet set, we generated 1,000 gene trees for several HGT rates λ. We extracted all quartet topologies and used a randomly chosen subset of these topologies to estimate the HGT rate λ. Repeating this for the quartet set sizes 100, 1,000, and 10,000, we got the results visualized in figure 2 where the true HGT rate λ to generate gene trees is plotted against the estimated rate. It turned out that a set of 10,000 topologies was large enough to get reliable estimation results.

For a second test, we used 10,000 quartet topologies and varied the value of h_{max} (the maximal number of simulated transfers on the species tree) while N (number of simulations for a fixed h_{max}) was 100,000. Figure 3 displays the estimation based on $h_{max} \in \{20, 30, 40, 60\}$. One can see that for each h_{max} exists a maximum rate which can be estimated reliably while rates above get underestimated.

In another analysis (data not shown), we also checked that $N = 100,000$ is a feasible value for the number of simulations.

The Most Frequent Gene Tree

To determine whether the most frequent gene tree is similar to the reconstructed species tree, we compared both trees. We computed a quartet set of all quartet topologies of the 780 COG gene trees which only consisted of those quartet topologies that did not contradict each other. A comparison of this quartet set with the species tree $T(S_{\mathrm{COG}})$ comprising 44 taxa led to the consensus tree depicted in figure 4. Both trees support all bifurcations except for two nodes indicated by multifurcations in the consensus tree. We can conclude that $T(S_{\mathrm{COG}})$ and the tree reconstructed of the quartet set – consisting of not contradicting quartet topologies of the gene trees – are nearly equal. As the latter quartet set represents the most frequent quartet topologies, we can also deduce that the most common gene tree is very similar to $T(S_{\mathrm{COG}})$.

Estimating the HGT Rate λ for the COG Data

The quality tests described above have shown that an HGT rate λ of 0.7 can be estimated reliably if we extract 10,000 quartet topologies of the 780 COG gene trees and set the parameters $N = 100,000$ and $h_{max} = 60$.

We applied this procedure to the COG data, repeated the estimation for 50 sets of quartet topologies, and obtained results for $\hat{\lambda}$ between 0.43 and 0.48 and for $\hat{\Lambda}$ between 12.86 and 14.35 presented in figure 5A. As Λ is the parameter of the Poisson distribution, which describes the occurrence of HGT events in time, $\hat{\Lambda}$ is the expected value for the number of HGT events that happened on $T(S)$, here $T(S_{\mathrm{COG}})$. The estimated HGT rate $\hat{\lambda}$ is relative to the number of substitutions in COG0419 (ATPase involved in DNA repair) which were used to assign branch lengths to the species tree.

To test the reliability of the results, we checked if the estimated HGT rates differ from those estimated of quartet sets randomly chosen from all quartet topologies of the 44 species tree organisms. Figure 5A shows the estimation results of quartet topologies which could be found in the 780 gene trees and figure 5B represents estimations over all $\binom{44}{4} \cdot 3$ quartet topologies. The graph indicates an estimated HGT rate $\hat{\lambda}$, which is about 10 times higher, namely between 4.66 and 4.7. As these rates must be higher because the

set consists of many quartet topologies (7 %) which are not part of the gene trees, so that many more HGT events are necessary to get the distribution, the estimated rates for the COG data seems quite acceptable.

Rate Correction Taking into Account the Inaccuracies of Gene Tree Reconstructions

We performed a further analysis taking into account the inaccuracies and uncertainties of gene tree reconstructions. For each protein family g_i, representing a taxa set X_{g_i}, we restricted $T(S_{\mathrm{COG}})$ to X_{g_i} denoted by $T(S_{\mathrm{COG}}|X_{g_i})$ and assigned branch lengths to all of these tree topologies using TREE-PUZZLE (Schmidt *et al.*, 2002). Afterwards, we simulated protein sequences of the same size than the corresponding COG sequences with SEQ-GEN (Rambaut and Grassly, 1997) along the calculated trees, using the Dayhoff substitution model (Dayhoff *et al.*, 1978). We repeated this step five times, then calculated the corresponding gene trees, and repeated the estimation procedure. As the newly acquired gene trees are based on trees which are subtrees of the species tree $T(S_{\mathrm{COG}})$, we expected to estimate an HGT rate $\hat{\lambda}$ of about zero.

After the estimation of ten randomly chosen quartet sets for each of the five simulated data sets, we got the distribution which is shown in the stacked histogram in figure 6. Each of the five colors represents one data set. The estimation results are nearly constant, at about 0.1 (0.1 ± 0.01). This result could be interpreted as a kind of background noise due to inaccuracies in the applied gene tree reconstruction procedure because in the setting there should be no difference in the branching patterns of gene and species tree and therefore, it leads to the conclusion that the estimated average HGT rate $\hat{\lambda}$ of about 0.46 per gene and unit time is about 22 % too high. This would decrease the total amount of HGT events which is necessary to transform the species tree topology into one gene tree from 14 to 11 events per gene.

Discussion

In the previous paragraph, we have described some results based on a new approach to estimate an overall rate of HGT with the help of a likelihood framework. We are able to estimate such a rate under the assumptions that all differences between a gene tree and the corresponding species tree are caused by HGT and that the HGT rate is homogeneous over the whole tree. Note that we did not make any statement about the probability if a gene is transferred at all, but how many events have occurred within one COG family on average. Thus, we are assuming that every gene is transferred with the same probability.

The extent to which HGT has shaped the individual genomes is controversially discussed (Kurland *et al.*, 2003). Although some research groups support the opinion that HGT plays an important role in evolution and appears very frequently (Doolittle, 1999; Eisen, 2000; Garcia *et al.*, 2000), others think that the impact of HGT is overestimated due to problems in the various inferring procedures (Brown, 2003). A recent publication by Ge *et al.* (2005) also analyzed the COG data and detected HGT in 33 out of 297 protein families. To do so, they used a novel test statistic based on tree topology comparisons. Unfortunately, they did not say anything about how many HGT events happen in each of the 33 detected COGs, which would be interesting in order to compare their results with ours.

There are several other approaches trying to estimate an HGT rate. For example, Huelsenbeck *et al.* (2000) developed a bayesian framework for the analysis of cospeciation, which could also be used to estimate rates of genetic transfer. Suchard (2005) published two stochastic models serving the same purpose. The first model, developed by Suchard (2005), is based on subtree prune and regraft operations and is applicable if the number of taxa under consideration is small, whereas the other approach is a random walk over complete graphs and offers a solution for an increasing number of taxa. In both publications, the fact that the corresponding framework can deal with gene and species tree topologies which are not known without error is highlighted. But on the other hand, both HGT models require that all gene and species trees are based on the same set of present-day species. In contrast, the new approach, which we have introduced

here, can incorporate gene families which are incomplete by using quartet subtrees and we can estimate reliable rates even if the gene tree taxa sets are only subsets of the whole species tree taxa set. As an example, the COG data set only comprises three gene families which represent genes for all 44 taxa (see Supplementary Material). Furthermore, this new framework also takes into account the inaccuracies in the gene tree reconstruction method.

As genomes are not only shaped by HGT, but also by processes like hybridization, gene loss, duplication, genesis and fusion/fission (Snel *et al.*, 2002), it becomes clear that the estimated rate of about 11 events per gene and unit time is a kind of upper bound because we assume that all conflicts in the gene tree topologies are caused by HGT. However, it remains as yet unclear, how the rate estimate changes if multi-copy genes were included in the analysis. Nevertheless, the estimate seems to be quite high. This can be explained by the fact that a lot of HGT events will not change the tree topology, for example events between two nodes that share parents. This seems to be important because 71 % of the total branch length of the COG species tree can be involved in HGT events which do not change the branching pattern. As it is most likely that the majority of HGT events in nature take place between closely related taxa it becomes clear that the number of these events would be underestimated by just counting visible incongruences between two given trees. Moreover, if one gene is transferred back and forth between two lineages, these events will not be detected either. The importance to take unobservable HGT events into account is supported by the fact that the topologies of 264 (34 %) of the 780 COG gene trees are equal to the species tree, restricted to the corresponding gene tree taxa. This means that the gene tree topology can be explained without any single HGT event. As the number of taxa of these 264 trees differs widely, and even gene trees with up to 36 taxa are equal to the corresponding species tree restriction, we can assume that HGT events happened during the evolution of the corresponding gene although we cannot see any of them. This is also supported by the fact that it is still not proven if a core of non-transferable genes exists (Nesbo *et al.*, 2001). Summing up, the importance of simulating HGT events on a given species tree, instead of just counting visible differences between a

species and gene tree, becomes obvious and distinguishes our approach from some previous work on estimating an HGT rate. To get an impression of the probability that an HGT event does not change the tree topology, we counted the simulated trees which are equal to the COG species tree. The result indicates that this probability is 9 % (0.9 %) for the simulated trees after one (two) transfer(s). As our approach includes simulations on the species tree which gave us a distribution of trees after different numbers of HGT events, we automatically include unobservable HGT events and therefore, the estimated rate is higher than in other approaches. But this high rate also indicates that HGT influences the tree topologies strongly, as described by Doolittle (Doolittle, 1999).

Many other approaches (e.g. see Hao and Golding, 2006; Dagan and Martin, 2007) exist which also estimate an HGT rate. All those methods are quite different from one another and it is difficult to compare their results with ours. The two mentioned publications are based on gene present and absent patterns, whereas the method that we have introduced here, uses the information of reconstructed gene trees to calculate an HGT rate. Dagan and Martin (2007) have presented a method in which they inferred a conservative lower bound estimate of about 1.1 HGT events per gene family and gene family lifespan considering the genome size of present day species. As already explained above, the estimates represented here are a kind of upper bound and therefore, they are much higher. As both methods (Hao and Golding, 2006; Dagan and Martin, 2007) are tested on different data sets, it would be interesting to see how much the results really differ when both are applied to the same data set.

Certainly, this newly developed method to estimate a rate of HGT is based on a number of key assumptions (as described in the Materials and Methods paragraph) and we are aware of the fact that probably some of them are not reasonable from a more biological point of view. Nevertheless, we started with such a simplified model to get a first impression about the overall rate of HGT for a set of present-day species and we intend to consider more biological relevant aspects of HGT in the future. For example, like Suchard (2005), we would like to include heterogeneous HGT rates in the analysis because such rates are important to take into account that genes belonging to different

functional categories have different transferabilities (Nakamura *et al.*, 2004). Another interesting and important extension for the simulation would be to include uncertainties of the species tree branch lengths. So far, we assume that these lengths are exactly known.

Furthermore, the species tree represents the evolutionary history of all 780 examined COG families and is therefore slightly different from a tree obtained from 16/18 S rDNA sequences which are often used to reconstruct a species tree for a set of given taxa (Woese, 2000). Probably we would even estimate a higher HGT rate if we used such a species tree because that one would not consider the information of the COG data, which would lead to more differences between species and gene trees. Therefore, the estimation of an HGT rate with the help of a different species tree would be a further interesting task.

Currently, the newly developed method only deals with trees representing a single gene copy per species. Since phylogenies often present several distantly related copies for a given organism, the HGT estimates based on orthologs only could be too low. We intend to include multi-copy genes into the analysis to overcome this problem in the future.

Supplementary Material

Supplementary tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe. oxfordjournals.org/). The first table S1 is a summary of all 780 COG gene trees that were used in the described analysis. For each gene tree, the corresponding COG, the maximum likelihood value of the reconstructed gene tree, the number of taxa, whether or not the gene evolved clocklike, and the species names are given. As we used abbreviations for the exact species names, table S2 provides a translation table assigning each species name such an abbreviation. A software package to estimate an HGT rate is available online (http://www.cibiv.at/software/hgt/). It consists of several C/C++ programs, Perl scripts, and a short user manual.

Acknowledgments

We wish to thank the Bioinformatics Institute at the University of Düsseldorf and the Center for Integrative Bioinformatics in Vienna (CIBIV). For coding some extensions to the TREE-PUZZLE program, we would like to thank Heiko Schmidt. We thank Roland Fleißner, Ingo Paulsen, Ricardo de Matos Simões, and two anonymous referees for useful comments on an earlier version of the manuscript and Mareike Fischer and Ramona Schmid for helpful discussions and critical proofreading. This work was supported by DFG grant SFB-TR1 (Deutsche Forschungsgemeinschaft). We also thank the Vienna Science and Technology Fund (WWTF) for financial support.

Literature Cited

Andersson, JO. 2005. Lateral gene transfer in eukaryotes. Cell Mol Life Sci 62:1182–1197.

Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau ME, Nesbø CL, Case RJ, Doolittle WF. 2003. Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet 37:283–328.

Brown JR. 2003. Ancient horizontal gene transfer. Nat Rev Genet 4:121–132.

Bushman F. 2002. Lateral DNA Transfer: Mechanisms and Consequences. Cold Spring Harbor Laboratory Press.

Dagan T, Martin W. 2007. Ancestral genome sizes specify the minimum rate of lateral gene tranfer during prokaryote evolution. Proc Natl Acad Sci USA 104:870–5.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. In: Dayhoff MO, editor. Atlas of Protein Sequence and Structure. Washington DC: National Biomedical Research Foundation. Vol. 5, p. 345–352.

de la Cruz F, Davies J. 2000. Horizontal gene transfer and the origin of species: lessons from bacteria. Trends Microbiol 8:128–133.

Diruggiero J, Dunn D, Maeder DL, Holley-Shanks R, Chatard J, Horlacher R, Robb FT, Boos W, Weiss RB. 2000. Evidence of recent lateral gene transfer among hyperthermophilic archaea. Mol Microbiol 38:684–693.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. Science 284:2124–2129.

Eisen JA. 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. Curr Opin Genet Dev 10:606–611.

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet 22:521–565.

Felsenstein J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5:164–166.

Garcia-Vallve S, Romeu A, Palau J. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res 10:1719–1725.

Ge F, Wang LS, Kim J. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol 3:e316.

Hao W, Golding GB. 2006. The fate of laterally tranferred genes: Life in the fast lane to adaptiation or death. Genome Res 16:636–43.

Hillis DM, Moritz C, Mable BK. 1996. Molecular Systematics. Sunderland, MA: Sinauer.

Huelsenbeck JP, Rannala B, Larget B. 2000. A Bayesian framework for the analysis of cospeciation. Evolution 54:352–364.

Kurland CG, Canback B, Berg OG. 2003. Horizontal gene transfer: a critical view. Proc Natl Acad Sci USA 100:9658–9662.

Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44:383–397.

Lerat E, Daubin V, Moran NA. 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS Biol 1:e19.

Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol 3:e130.

Nakamura Y, Itoh T, Matsuda H, Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet 36:760–766.

Nei M. 1996. Phylogenetic analysis in molecular evolutionary genetics. Annu Rev Genet 30:371–403.

Nelson KE., Clayton RA, Gill SR, et al. (29 co-authors). 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 399:323–329.

Nesbø CL, Boucher Y, Doolittle WF. 2001. Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. J Mol Evol 53:340–350.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. Mol Biol Evol 5:568–583.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 13:235–238.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504.

Semple C, Steel M. 2003. Phylogenetics. Oxford University Press.

Snel B, Bork P, Huynen MA. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res 12:17–25.

Suchard MA. 2005. Stochastic models for horizontal gene transfer: taking a random walk through tree space. Genetics 170:419–431.

Syvanen M. 1994. Horizontal gene transfer: evidence and possible consequences. Annu Rev Genet 28:237–261.

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22–28.

Woese CR. 2000. Interpreting the universal phylogenetic tree. Proc Natl Acad Sci USA 97:8392–8396.

FIG. 1: Comparison of the species tree (A) and one corresponding gene tree (B) after a single HGT event. The arrow indicates a gene transfer from species A (donor) to species D (acceptor). To check whether the gene tree is a subtree of the species tree, we compute the tree topology $\tau(S|X)$ (C) derived from the species tree.

FIG. 2: Quality of the rate estimation in dependence on the number of quartet topologies: (A) 100, (B) 1,000 and (C) 10,000. $N = 100,000$ and $h_{max} = 60$ are fixed.

FIG. 3: Quality of the rate estimation as a function of the maximum number of simulated HGT events with $N = 100,000$ and $10,000$ quartet topologies. Each displayed value is based on one estimation.

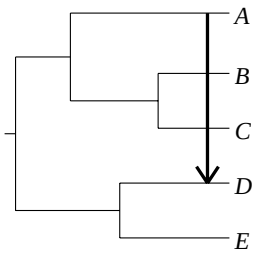
FIG. 4: Consensus tree of the COG species tree and the tree reconstructed of the quartet set which represents the most frequent quartet subtrees of the 780 gene trees. Only two multifurcations exist which indicate incongruity between both input trees. This tree was reconstructed with the strict consensus mode of the CONSENSE program (Felsenstein, 1989).

FIG. 5: Distribution of the estimated HGT rates $\hat{\lambda}$ for the COG data for randomly chosen quartet sets (A) of the 780 gene trees, and (B) over all 44 species tree taxa with $N = 100,000$ and $h_{max} = 60$

FIG. 6: Distribution of the estimated HGT rates $\hat{\lambda}$ for five simulated data sets. Each data set is based on the 780 protein families and their corresponding subtrees in $T(S_{COG})$. For each data set ten randomly chosen quartet sets have been estimated.

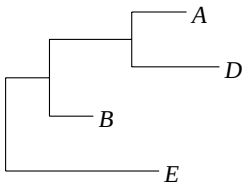
(A) species tree $T(S)$

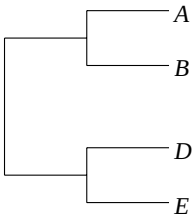
$S=\{A,B,C,D,E\}$



(B) gene tree topology $\tau(X)$

$X=\{A,B,D,E\}$





(C) tree topology $\tau(S|X)$

FIG. 1:

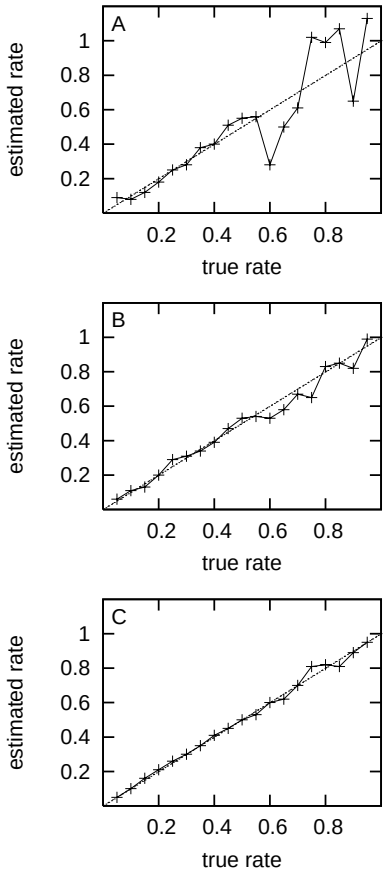


FIG. 2:

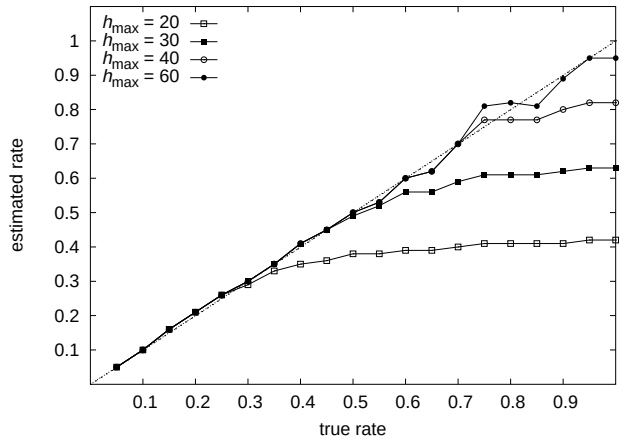

h_{max} = 20
h_{max} = 30
h_{max} = 40
h_{max} = 60
estimated rate
true rate
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

FIG. 3:

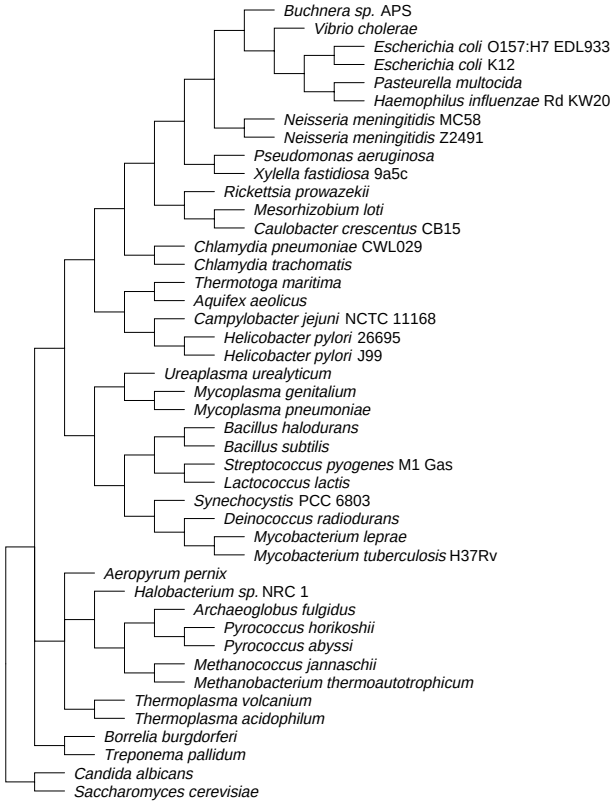

Buchnera sp. APS
Vibrio cholerae
Escherichia coli O157:H7 EDL933
Escherichia coli K12
Pasteurella multocida
Haemophilus influenzae Rd KW20
Neisseria meningitidis MC58
Neisseria meningitidis Z2491
Pseudomonas aeruginosa
Xylella fastidiosa 9a5c
Rickettsia prowazekii
Mesorhizobium loti
Caulobacter crescentus CB15
Chlamydia pneumoniae CWL029
Chlamydia trachomatis
Thermotoga maritima
Aquifex aeolicus
Campylobacter jejuni NCTC 11168
Helicobacter pylori 26695
Helicobacter pylori J99
Ureaplasma urealyticum
Mycoplasma genitalium
Mycoplasma pneumoniae
Bacillus halodurans
Bacillus subtilis
Streptococcus pyogenes M1 Gas
Lactococcus lactis
Synechocystis PCC 6803
Deinococcus radiodurans
Mycobacterium leprae
Mycobacterium tuberculosis H37Rv
Aeropyrum pernix
Halobacterium sp. NRC 1
Archaeoglobus fulgidus
Pyrococcus horikoshii
Pyrococcus abyssi
Methanococcus jannaschii
Methanobacterium thermoautotrophicum
Thermoplasma volcanium
Thermoplasma acidophilum
Borrelia burgdorferi
Treponema pallidum
Candida albicans
Saccharomyces cerevisiae

FIG. 4:

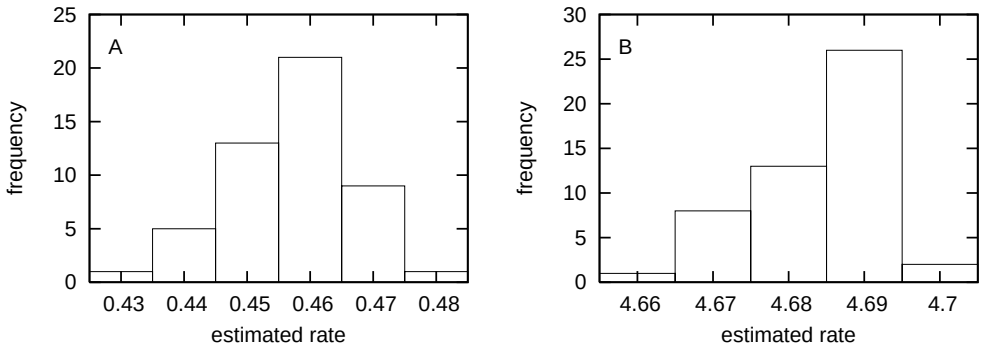

A
frequency
estimated rate
0.43
0.44
0.45
0.46
0.47
0.48
0
5
10
15
20
25
B
frequency
estimated rate
4.66
4.67
4.68
4.69
4.7
0
5
10
15
20
25
30

FIG. 5:

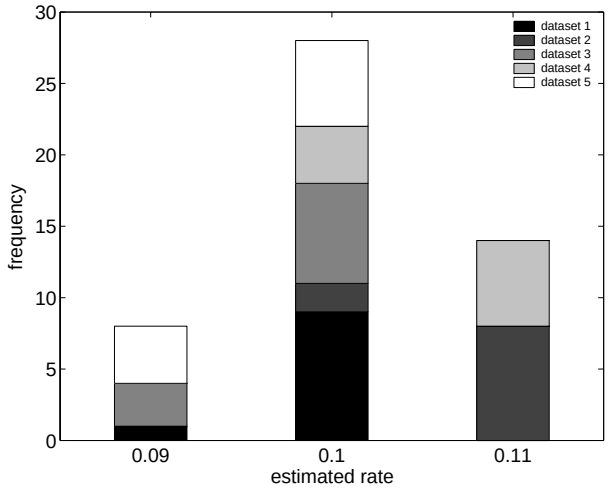

dataset 1
dataset 2
dataset 3
dataset 4
dataset 5
frequency
0
5
10
15
20
25
30
0.09
0.1
0.11
estimated rate

FIG. 6: