A Machine-Learning Approach Reveals That Alignment Properties Alone Can Accurately Predict Inference of Lateral Gene Transfer from Discordant Phylogenies

Mayo Roettger, William Martin, and Tal Dagan

Institut für Botanik III, Heinrich-Heine Universität Düsseldorf, Germany

Among the methods currently used in phylogenomic practice to detect the presence of lateral gene transfer (LGT), one of the most frequently employed is the comparison of gene tree topologies for different genes. In cases where the phylogenies for different genes are incompatible, or discordant, for well-supported branches there are three simple interpretations for the result: 1) gene duplications (paralogy) followed by many independent gene losses have occurred, 2) LGT has occurred, or 3) the phylogeny is well supported but for reasons unknown is nonetheless incorrect. Here, we focus on the third possibility by examining the properties of 22,437 published multiple sequence alignments, the Bayesian maximum likelihood trees for which either do or do not suggest the occurrence of LGT by the criterion of discordant branches. The alignments that produce discordant phylogenies differ significantly in several salient alignment properties from those that do not. Using a support vector machine, we were able to predict the inference of discordant tree topologies with up to 80% accuracy from alignment properties alone.

Introduction

The phylogenetic approach for lateral gene transfer (LGT) inference from the frequency of incongruent branching patterns in gene trees has so far delivered widely conflicting results, ranging from estimates that as few as 2% (Ge et al. 2005) to possibly 14% of all genes in prokaryote genomes are affected by LGT (Beiko, Harlow, and Ragan 2005). Such divergent estimates using phylogenetic tree comparisons can, in principle, be attributed to many factors including the obvious, such as lineage sampling, the inherent uncertainties of various approaches to phylogenetic reconstruction (Penny et al. 1992; Hillis 1995; Lopez et al. 2002) and the threshold levels of support set to score the presence of genuinely conflicting topologies. But phylogenetic trees of molecular sequences are always inferred from multiple sequence alignments. Nei et al. (1995) and Nei (1996) pointed out early on that alignment of highly diverged sequences may result in erroneous phylogenetic reconstruction. Interest in this aspect of phylogeny has renewed with several reports investigating the alignment step itself as it specifically relates to phylogenetic inference (Landan and Graur 2007; Deusch et al. 2008; Löytynoja and Goldman 2008; Wong et al. 2008)

Here, we wished to examine the extent to which LGT inference by the phylogenetic method might be sensitive to the properties of alignments themselves. For this purpose, we investigated the comprehensive data set compiled and carefully analyzed by Beiko, Harlow, and Ragan (2005), who kindly made their data available. Their data set is highly suitable for the present study 1) because it consists of 22,437 carefully assembled gene families of prokaryotic orthologs, in which paralogs have been sorted out by using a conservative similarity cutoff (Beiko, Harlow, and Ragan 2005, Supplementary Material online), 2) because they used a widely employed filter, Gblocks (Castresana 2000), to exclude poorly aligned regions from their analysis prior to phylogenetic reconstruction, and 3) because they used a very stringent (conservative) threshold for the scoring of discordant phylogenies. In brief, Beiko, Harlow, and Ragan (2005) constructed a consensus supertree for the proteins encoded in 144 prokaryotic genomes and constructed from the same data 22,437 individual phylogenetic trees containing from 4 to 144 sequences each using a Bayesian approach. They inferred LGT only from highly significant (posterior probability ≥ 0.95) discordant tree topologies in comparison to the consensus supertree topology (Beiko, Harlow, and Ragan 2005). For 5,822 of those trees, one or more LGT was inferred on the basis of discordance to the consensus topology, we designate those amino acid sequence alignments as “LGT positive” or LGT for short. The remaining 16,615 of the alignments investigated by Beiko, Harlow, and Ragan (2005) did not produce branches (bipartitions) that were discordant (conflicting) with the consensus supertree topology and are considered here as “vertical gene inheritance” or VGI alignments. We examined the properties of the LGT alignments in comparison to the properties of the VGI alignments.

Methods

For the analysis, we used a data set of 22,437 protein families from 144 prokaryotes for which LGT) was inferred using the phylogenetic method (Beiko, Harlow, and Ragan 2005). The data for each protein family include a multiple sequence alignment yielding the highest score according to the word-oriented objective function (Beiko, Chan, and Ragan 2005) from a set of alignments reconstructed by several different algorithms: ClustalW (Thompson et al. 1994), T-coffee (Notredame et al. 2000), MAFFT (Katoh et al. 2002), POA (Grasso and Lee 2004), and PRRP (Gotoh 1996), a partial alignment of relatively conserved regions constructed with Gblocks (Castresana 2000), and a phylogenetic tree inferred with MrBayes (Huelsenbeck and Ronquist 2001). Bipartitions

Key words: lateral gene transfer, molecular phylogeny, discordant tree topologies, support vector machine, principal component analysis.

E-mail: mayo.roettger@uni-duesseldorf.de.

Mol. Biol. Evol. 26(9):1931–1939. 2009
doi:10.1093/molbev/msp105
Advance Access publication May 14, 2009

in the phylogenetic tree were considered as concordant if they overlap with the reference supertree, or discordant otherwise, which were interpreted as LGT events (Beiko, Harlow, and Ragan 2005).

Multiple Alignment Properties

For each protein family (alignment), we calculated alignment properties as follows: Number of operational taxonomic units (OTUs) is the number of orthologs in the family. Proportion of gaps is the proportion of gap characters in the Gblocks output alignment. Entropy was calculated for each Gblocks output alignment as the average entropy of its sites. For the calculation, we used the Shannon information content normalized by the number of OTUs in the alignment (Valdar 2002):

$$\text{Entropy} = \frac{\sum_{\text{col}=1}^{N_{\text{sites}}} \left(-\lambda_t \sum_{i=1}^{K} p_{i,\text{col}} \log_2 p_{i,\text{col}}\right)}{N_{\text{sites}}},$$

where N_{sites} is the number of alignment sites, K is the alphabet size, which is 21 in this case (20 amino acids plus one gap symbol), and $p_{i,\text{col}}$ the probability of observing the ith character in alignment column col. The λ_t factor is used to scale the entropy into a [0,1] range by the number of OTUs (N_{seq}) and the alphabet size K.

$$\lambda_t = [\log_2(\min(N_{\text{seq}}, K))]^{-1}.$$

Invariant sites are positions in the Gblocks alignment where all sequences contain the same amino acid, and informative sites are defined as alignment columns containing at least two different amino acids, each one observed in at least two sequences at the position. Average pairwise identity is calculated as the proportion of amino acid identities between all sequence pairs and averaged for the protein family as follows:

$$\text{API} = \frac{2}{N_{\text{seq}}(N_{\text{seq}} - 1)N_{\text{sites}}} \sum_{\text{col}=1}^{N_{\text{sites}}} \sum_{i=1}^{N_{\text{seq}}-1} \sum_{j=i+1}^{N_{\text{seq}}} \left\{ \begin{array}{lcl} 1 : a_{i,\text{col}} & = & a_{j,\text{col}} \\ 0 : a_{i,\text{col}} & \neq & a_{j,\text{col}} \end{array} \right\},$$

where $a_{i,\text{col}}$ is the observed amino acid in Gblocks-alignment sequence i at position col.

In addition, we tested for alignment reliability using the Heads-or-Tails (HoT) method (Landan and Graur 2007). Tails alignments were obtained by aligning the reversed sequences of each protein family using exactly the same alignment procedure that was used for the original (Heads) alignments. Column score (CS) is calculated as the proportion of identical columns between Heads and Tails alignments, and sum of pairs score (SPS) was calculated as the proportion of identical residue position pairs between Heads and Tails alignments.

Each alignment comprises protein sequences from different species, each sequence named by Beiko et al. with a unique pipeline id. Additional files contained information about the original gi number in the RefSeq database (Pruitt et al. 2005) together with the current gi number in the database and the genome id of the sequence used by the National Center for Biotechnology Information (NCBI). Two hundred and twenty-three proteins in 183 alignments had no information about the sequence except the original gi number in the database. These database entries were replaced or removed from the database. In the calculation of the number of different phyla and the classification of the sequence into the kingdom groups, we discarded these sequences from the alignment. We obtained taxonomical classification information for each sequence from NCBI and counted the number of sequences being classified as different phyla for each cluster. We used the term archaea for clusters that contain only sequences classified as of archaebacterial origin, the term eubacteria for clusters that contain only sequences classified as of bacterial origin, and the term universal for clusters that contain sequences of both kingdoms.

For the comparison of the property distributions between LGT and VGI alignments, the Wilcoxon nonparametric test was used.

Orthologs Pairwise Distances

Protein pairwise distances between orthologs from LGT and VGI families were calculated for several genome pairs that were selected for their high frequency in the data: 1) *Vibrio vulnificus* versus *Yersinia pestis* (1,406 protein pairs where both species were present in the respective protein families), 2) *Brucella suis* versus *Mesorhizobium loti* (1,697 protein pairs), 3) *Agrobacterium tumefaciens* versus *M. loti* (2,205 protein pairs), and 4) *Bradyrhizobium japonicum* versus *M. loti* (1,794 protein pairs), 5) *Staphylococcus aureus* versus *Bacillus cereus* (924 protein pairs), 6) *Nostoc* sp. versus *Pyrococcus furiosus* (114 protein pairs), and 7) *Bacteroides thetaiotaomicron* versus *Sulfolobus solfataricus* (62 protein pairs). Pairwise protein distances were extracted from the distance matrix calculated from the multiple sequence alignments with PROTDIST (Felsenstein 1996) using Jonen-Taylor-Thornton substitution matrix (Jones et al. 1992). In addition, we calculated pairwise distances with the same method after realigning the orthologous sequences using MUSCLE (Edgar 2004).

Classification Procedure

Prediction of LGT (discordant tree bipartition) from alignment properties entailed a support vector machine (SVM) classifier (Christiani and Shawe-Taylor 2000). For the SVM training and classifying procedures, we used the svmtrain and svmclassify functions from the MATLAB 7.6 bioinformatics toolbox with the following parameters: Radial basis function (RBF) kernel, RBFSigmaValue = 1, Mlp_ParamsValue = [1,−1], MethodValue = SMO, BoxConstraintValue = 1, and AutoscaleValue = true. In order to obtain significance levels for the SVM performance, we applied 10-fold crossvalidation in each step using the small 1/10 subset for training and the 9/10 for testing. The LGT/VGI ratio in the training set was adjusted by randomly selecting different numbers of LGT and VGI samples from the preliminary training set to form an equal-sized training set for each validation step.

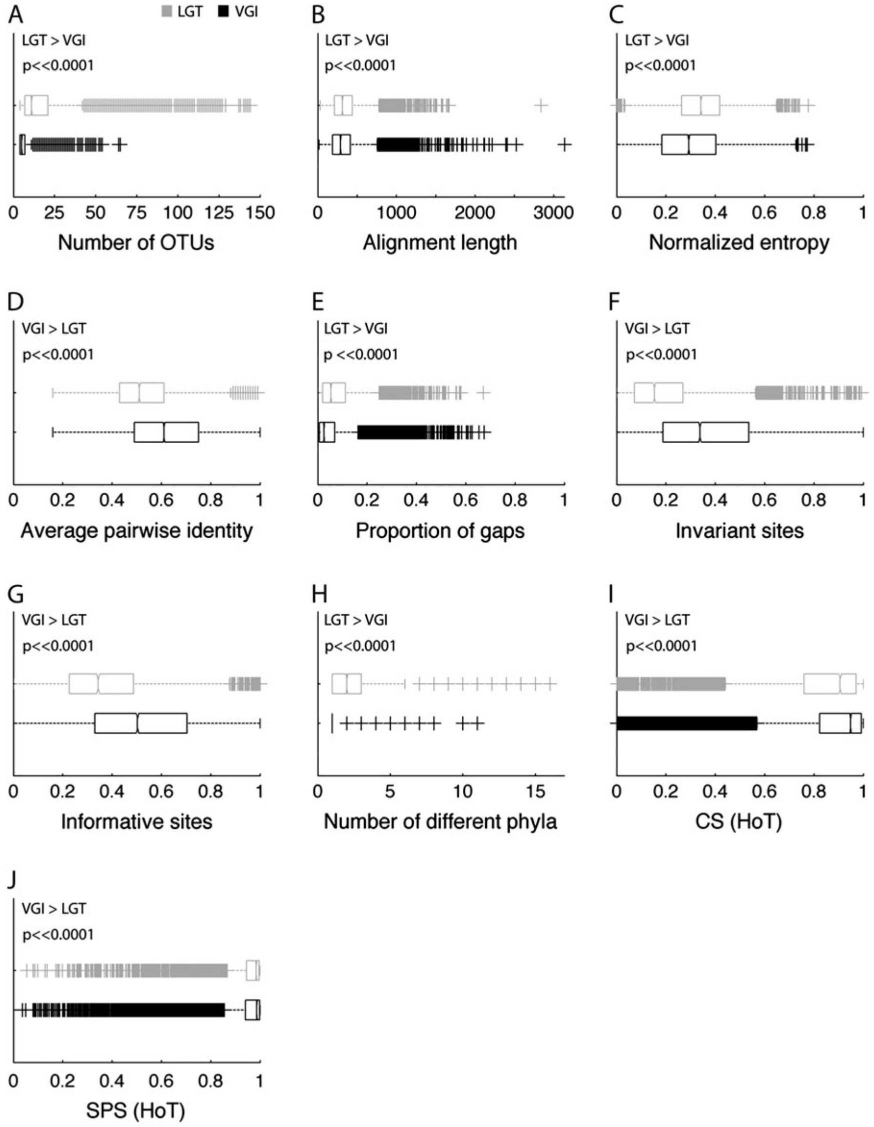

FIG. 1.—Distributions of alignment properties in the LGT and VGI groups. Differences in the distributions of the two groups were tested by the Wilcoxon nonparametric test (*P* values presented at the top of each graph).

SVM performance was evaluated by "accuracy," that is, the proportion of alignments correctly classified as LGT or VGI, "sensitivity," which is the true positive rate or the number of true positives (LGT alignments that are classified as such) divided by the sum of true positives plus false negatives (LGT alignments classified as VGI), and "specificity," as the true negative rate or the number of true negatives (VGI alignments that are classified as such) divided by the sum of true negatives plus false positives (VGI alignments classified as LGT).

To test the performance of the classifier under different LGT/VGI proportions in the training set and in the test set, we used LGT proportions ranging from 25% to 75% while including all 11 alignment properties.

To explore the contribution of the different features and their combinations to the classification performance, we tested all possible 2,047 combinations of the 11 alignment features analyzed in this study using a training set with equal proportions of LGT and VGI alignments.

Multivariate Analysis

We performed principal component analysis (PCA) using the princomp function of MATLAB 7.6. The data for each alignment property were normalized before the analysis, so that all properties had only values ranging from zero to one.

Results and Discussion

It is known that the probability of obtaining incorrect trees increases with the number of sequences (OTUs) analyzed (Nei 1996). The LGT alignments investigated here contained significantly larger numbers of OTUs ($P \ll 0.0001$) than the VGI alignments (fig. 1*A*). No VGI alignment in the present sample contains more than 65 sequences, whereas 5% of the LGT alignments contain $\geq$65 sequences. It is also known that for a given level of sequence divergence, the probability of obtaining incorrect trees is

Table 1
General Statistics of Protein Family Alignment Properties Grouped by LGT and VGI Categories

	Range (Min–Max)		Mean ± SD		Median	
MSA Parameter	VGI	LGT	VGI	LGT	VGI	LGT
Normalized Shannon entropy	0.000–0.772	0.002–0.777	0.294 ± 0.151	0.343 ± 0.117	0.292	0.341
Average pairwise identity	0.160–1.000	0.160–0.990	0.621 ± 0.176	0.523 ± 0.132	0.610	0.510
Proportion of gaps	0.000–0.675	0.000–0.671	0.052 ± 0.072	0.080 ± 0.084	0.025	0.053
Number of OTUs	4–65	4–144	6.6 ± 4.2	19.0 ± 21.7	5.0	11.0
Alignment length	14–3,135	31–2,837	325.0 ± 203.9	352.0 ± 212.3	287.0	311.0
Proportion of invariant sites	0.000–1.000	0.000–0.992	0.381 ± 0.247	0.191 ± 0.160	0.337	0.153
Proportion of informative sites	0.000–1.000	0.000–1.000	0.524 ± 0.250	0.366 ± 0.188	0.502	0.343
CS (HoT)	0.000–1.000	0.000–1.000	0.856 ± 0.212	0.823 ± 0.213	0.948	0.905
SPS (HoT)	0.037–1.000	0.053–1.000	0.939 ± 0.118	0.943 ± 0.111	0.986	0.984
Number of different phyla	1–11	1–16	1.323 ± 0.708	2.651 ± 2.475	1.0	2.0

NOTE.—The data set contains 22,437 protein family alignments, 5,822 of which are LGT and 16,615 are VGI.

higher when short sequences are analyzed than when longer sequences are analyzed (Nei 1996). However, the LGT protein families investigated here contain sequences that are significantly ($P < 0.0001$) longer than VGI protein families (supplementary fig. S1, Supplementary Material online), producing also longer alignments (mean = 352; table 1) than the VGI alignments (mean = 325; table 1) (fig. 1*B*), suggesting that if incorrect trees are involved in LGT inference in the present data, then short sequences are not the cause.

The probability of obtaining incorrect trees increases when sequence divergence becomes too great (Nei 1996). Several alignment properties can address the issue of sequence divergence. Normalized Shannon entropy provides an estimate for the average number of different amino acids that occur per site in an alignment (Valdar 2002). The mean normalized Shannon entropy of LGT alignments is about 17% higher than for the VGI alignments in the present data (fig. 1*C*), a highly significant difference ($P < 0.0001$). Average sequence identity across all pairwise comparisons is a very simple and robust measure of sequence variability in an alignment. The average pairwise identity of the VGI alignments (mean = 0.621; table 1) is significantly higher ($P < 0.0001$) than in the LGT alignments (mean = 0.523) (fig. 1*D*). In addition, LGT alignments contain on average 50% more gaps than VGI alignments (fig. 1*E*; table 1). Another proxy for sequence divergence in an alignment is the proportion of invariant sites, the mean of which is 2-fold higher ($P < 0.0001$) in the VGI alignments than in the LGT alignments (fig. 1*F*). Furthermore, the proportion of informative sites, defined here as alignment columns containing at least two different amino acids each observed in at least two sequences at the position, is significantly lower ($P < 0.0001$) in the LGT alignments (mean = 0.366; table 1) than in the VGI alignments (mean = 0.524; table 1) (fig. 1*G*).

Thus, several alignment parameters that are known to increase the probability of obtaining incorrect trees—higher numbers of OTUs, sequence divergence exceeding 50% differences on average, and low numbers of informative sites—are significantly different in the LGT and the VGI alignments, and in all cases, LGT alignments are skewed toward the value that increases the probability of obtaining an incorrect tree. This does not directly indicate that the LGT alignments have produced branches that are highly supported but nonetheless incorrect (Delsuc et al. 2003), yet the tendency is consistent.

The proportion of invariant sites, the proportion of informative sites, and average pairwise identity show an inverted trend to the Shannon entropy in the PCA of the total data set (fig. 2). These three measures correlate negatively with Shannon entropy ($r = -0.84$, $r = -0.82$, and $r = -0.97$, $P < 0.0001$, respectively, supplementary fig. S2*A*–*C*, Supplementary Material online). This means that the less variable alignments may lack phylogenetic information due to high proportions of invariable sites, where the proportion of informative sites in these alignments will still be high. Yet these correlation coefficients are weaker in the LGT alignments ($r = -0.70$ and $r = -0.77$ and $r = -0.95$, $P < 0.0001$, respectively, supplementary fig. S2*D*–*F*, Supplementary Material online), than in the VGI alignments ($r = -0.87$ and $r = -0.84$ and $r = -0.98$, $P < 0.0001$, respectively; supplementary fig. S2*G*–*I*, Supplementary Material online) so that even though the LGT alignments are more variable than the VGI alignments, they generally contain not only fewer invariant sites but also fewer informative sites.

High alignment variability in the LGT alignments could be also the result of large numbers of sequences per alignment, as is the case for the LGT group alignments (fig. 1*A*). However, we found no correlation between number of OTUs and normalized entropy ($r = 0.01$, $P = 0.27$), and only weak correlation between number of OTUs and average pairwise identities ($r =-0.11$, $P < 0.0001$), or the proportion of gaps ($r = 0.23$, $P < 0.0001$; supplementary fig. S3, Supplementary Material online). Also, the number of phyla represented in the alignment, another possible source for higher alignment variability, is higher in the LGT groups than in the VGI group (fig. 1*H*). But this measure as well shows no significant correlation with any of the variability measures (supplementary fig. S4, Supplementary Material online). Hence, the high variability of the LGT alignments is not explained by the large number of sequences or the large number of phyla represented in these families.

The more variable the sequences in an alignment are, the more difficult they are to align and the more likely it is that the alignment procedures themselves can produce collections of site patterns that induce topological effects at the tree-building stage (Landan and Graur 2007; Wong et al. 2008). Thus, the LGT alignments, which are more variable than those in the VGI group, might be more error prone at the alignment step than the VGI alignments. To estimate

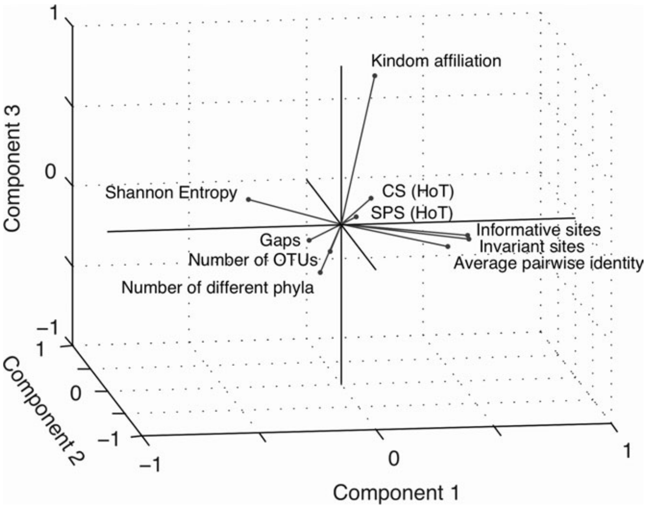

FIG. 2.—Principal component analysis of alignment properties. The axes represent the first three components, explaining 85% of the variability in the data (see supplementary tables S3 and S4, Supplementary Material online for details). Alignment properties are represented as vectors of their principal component coefficients. Alignment length is omitted due to its marginal contribution to the first three principal components. Two-dimensional views of every two respective components can be found in supplementary fig. S7, Supplementary Material online.

this effect, we compared alignment reliability of LGT and VGI alignments using the HoT method (Landan and Graur 2007). For these HoT comparisons, we realigned the original sequences (i.e., before filtering with Gblocks) as kindly provided by Beiko, Harlow, and Ragan (2005) in the C-to-N-direction to form the Tails alignments and compared them with the original (Heads) alignments. Both HoT parameters show an inverted trend to the number of OTUs, number of different phyla, and the proportion of gaps in the PCA analysis (fig. 2). Moreover, we found that the LGT alignments have a significantly ($P < 0.0001$) lower CS, which is the proportion of site columns reconstructed identically in the Heads and Tails alignment (fig. 1*I*), and a slightly but significantly ($P \ll 0.0001$) lower SPS, which is the proportion of identically reconstructed site pairs (fig. 1*J*), than the VGI alignments. Hence, LGT alignments contain significantly more alignment artifacts that are introduced by the sequence alignment process alone, independent of subsequent tree-building procedures. The bias in alignment quality within the LGT set is unlikely to be related with the erroneous guide tree used for the alignment because alignment errors are only marginally affected by the guide-tree quality (Landan and Graur 2008). Beiko, Harlow, and Ragan (2005) used a very conservative rule for inclusion in the LGT set that comprises only trees having at least one highly significant ("posterior probability" $\geq$ 0.95) discordant branch, whereas all other trees are considered as VGI. This results in abias toward highly supported (though not necessarily true) trees in the LGT set, where the proportion of highly significant branches per tree is 47 ± 28% versus 48 ± 40% (median 41% vs. 33%) in the VGI set. To test if this bias is related to the differences we found in the alignment properties, we deleted from the VGI set those alignments yielding trees with no highly significant branches, leaving 13,811 alignments yielding trees having at least one highly significant (posterior probability $\geq$ 0.95) concordant branch. This resulted in a set of trees, designated here VGI95, having a much higher proportion of highly significant branches per tree (57 ± 37%, median 50%). A comparison of alignment properties between the LGT and VGI95 sets resulted in identical conclusions to those detailed above for the comparison between the LGT and VGI sets (supplementary fig. S5, Supplementary Material online), so that the bias toward highly resolved trees in the LGT set has no relation to the bias in multiple alignment properties.

The comparison of alignment properties between VGI and LGT alignments summarized so far (fig. 1; table 1) shows that the LGT alignments are more variable than the VGI alignments. It is thus possible that the laterally transferred protein-coding sequences are inherently more variable than vertically inherited ones. To test this possibility, we compared pairwise protein distances between genomes to see if there were differences between LGT and VGI sequences with respect to overall sequence conservation. If so, then orthologous sequence pairs from VGI alignments should have smaller protein distances (i.e., should be more conserved) than orthologous pairs from LGT alignments. We tested that hypothesis for seven frequent genome pairs having proteins in both groups. The contrary was observed: Orthologous pairs from LGT alignments are more conserved (i.e., have smaller protein distances) than orthologous pairs from VGI alignments (fig. 3; supplementary fig. S6, Supplementary Material online). Hence, the higher variability observed in LGT alignments cannot be explained by a systematic bias in protein conservation among inherited versus laterally transferred proteins. This conclusion seems to contradict the bias toward variable multiple sequence alignments in the LGT set. A possible reconciliation between these two findings may be found in a study by Elhaik et al. (2006) showing that conserved proteins have higher probability of being detected by a similarity search, which leads to the composition of larger protein families, hence alignments with more OTUs that are probably more difficult to align. However, we found no correlation between the number of OTUs and the different alignment variability measures in our data set (supplementary fig. S3, Supplementary Material online).

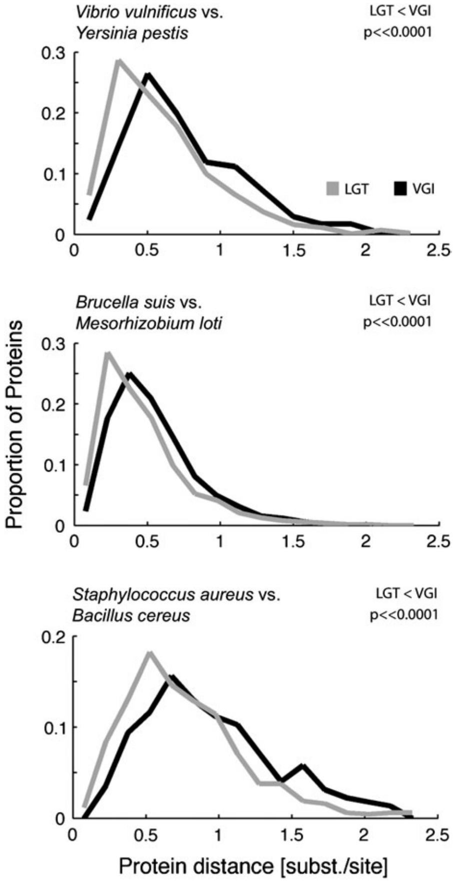

FIG. 3.—Comparison of protein pairwise distances between genome pairs found in LGT and VGI families. Distance distributions were compared using the Wilcoxon nonparametric test (*P* values presented at the top right of the graph).

Clustering of conserved orthologous proteins results not only in bigger families, but also in families having proteins from many taxomonic groups (supplementary fig. S4A, Supplementary Material online). However, the number of phyla alone is unlikely to reflect the relatedness among the sequences in the protein family because a protein family including sequences from eubacteria and archaebacteria is expected to contain more variability than a protein family including sequences from eubacteria only. Therefore, we divided the multiple alignments into those that comprise 1) eubacterial proteins only, 2) archaeal proteins only, and 3) "universal" alignments including proteins from both groups. A comparison of alignment properties among these three categories shows that universal families are much bigger than either eubacterial or archaeal families (fig. 4*A*). Moreover, all variability measures show that the universal alignments are more variable than alignments in the other two categories: Their entropy is higher (fig. 4*C*), their mean pairwise distance is higher (fig. 4*D*), and they contain more gaps (fig. 4*E*) and less invariant sites (fig. 4*F*). Universal alignments also seem to be of lower quality, in that they contain fewer informative sites and their alignments are less reliable (fig. 4*G–I*).

Finally, we tested for dependency between the taxonomical composition of the alignments and their classification as VGI or LGT and found these two properties are significantly dependent ($P < 0.001$, using χ^2 test). LGT alignments comprise about 30% of the archaeal or eubacterial alignments (as in the total data), but they are overrepresented in the universal group where they comprise 57% of the multiple alignments (fig. 4*J*). This leads us to conclude that the clustering of conserved sequences resulted in protein families that are not only large (as predicted by Elhaik et al. 2006) but also have a universal taxonomic distribution that covers much more diverse sequences and that seems to be the reason for their variability.

Our results so far suggest that alignments possessing properties that are known to increase the probability of obtaining incorrect branches are more frequent in the LGT group than in the VGI group. We then asked a slightly heretical question: Can we predict whether an alignment is likely to generate a tree with a strongly supported discordant branch on the basis of alignment properties alone? For this, we used an SVM classifier (Christiani and Shawe-Taylor 2000). In brief, a SVM is an algorithm that, provided with a learning set of features that might or might not correlate to a classificatory decision of the type "yes" or "no," gains experience with the learning set, and then is asked to classify objects, correctly if possible, on the basis of features alone. In the present case, the features correspond to alignment parameters as summarized in figure 1 and table 1, and the desired classification is the proper assortment of the alignment into LGT or VGI groups as predetermined by phylogenetic analysis. The classification performance is evaluated by its accuracy, sensitivity, and specificity (see Methods).

The SVM algorithm was thus trained and queried using the present alignments. In order to calculate SVM performance and standard deviations (SDs), we performed a 10-fold crossvalidation using 1/10 of the data in each step for training and the rest for testing. Accuracy, sensitivity, and specificity of the classifying process are to a vast extent influenced by the ratio of LGT/VGI in the training set. Accuracy and sensitivity are maximal when the proportion of LGT in the training set is equal to the total data (25%) and they decrease when higher LGT proportions are used. The specificity of the SVM classification is minimal at 25% LGT alignments in the training set and increases when higher proportions are used. When LGT proportion in the training set is fixed to 50%, all SVM performance measures are found in equilibrium (fig. 5*A*). The ratio of LGT/VGI alignments in the test set has no influence on the performance of the SVM classifier (fig. 5*B*). In our SVM classification procedure, we used training sets having an LGT/VGI ratio = 1 (see Methods). Performance of the classifying process was evaluated by trying all possible 2,047 combinations of the 11 properties to explore if there is a set of features that, if omitted from the training process, will deteriorate the results, or if there are some features that tend to impair the performance when included in the analysis.

Table 2 shows the combination of features that yielded the top performance values of accuracy, sensitivity, and specificity. We cannot really decide which is the best combination of feature vectors to be included in the training process because widely different combinations of features induce consistent results in the classification performance. But it seems that for equally high performance values for

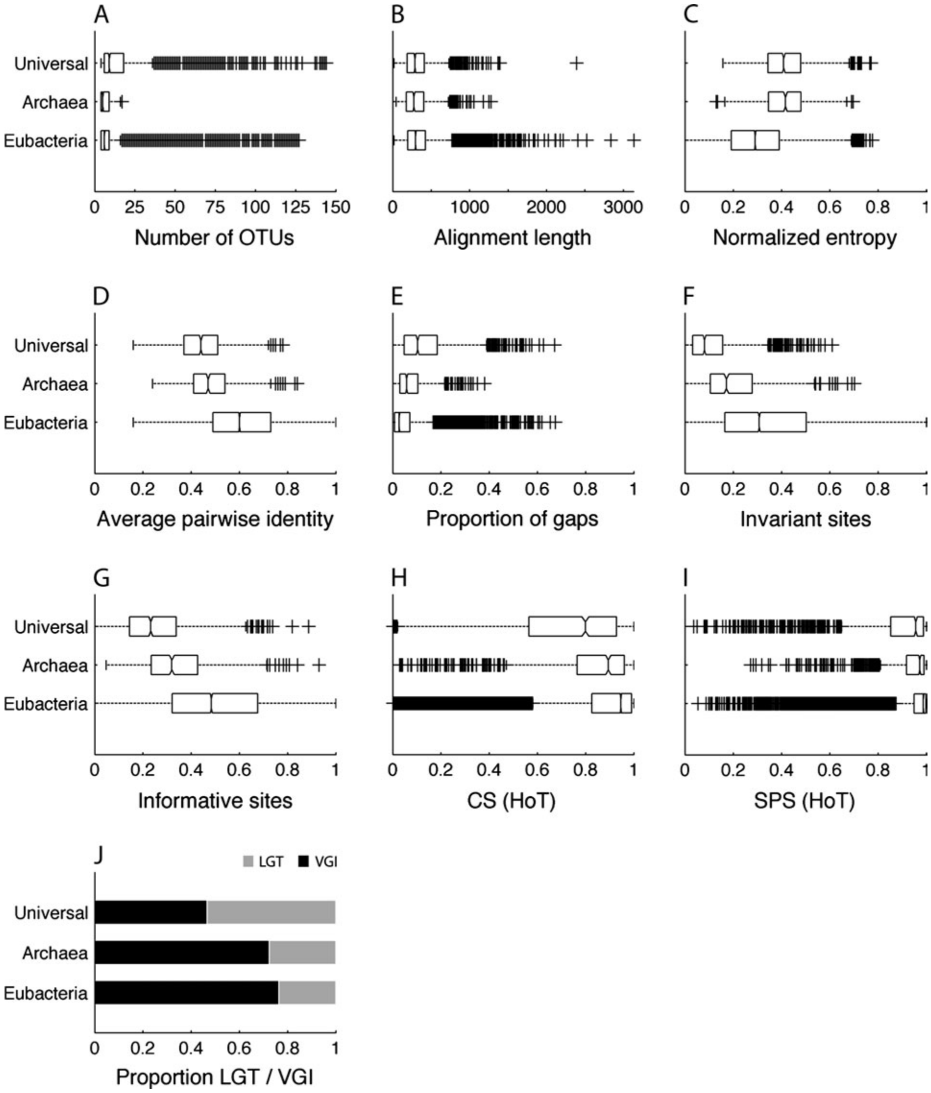

FIG. 4.—Differences in alignment properties for alignments containing only eubacterial sequences, only archaebacterial sequences, or sequences of both kingdoms (universal).

the three parameters (e.g., combinations yielding accuracy = 0.797 or accuracy = 0.796), the number of OTUs, entropy, average pairwise identity, and number of phyla are of particular importance. A complete table with all 2,047 tested combinations of features can be found in supplementary table S1, Supplementary Material online.

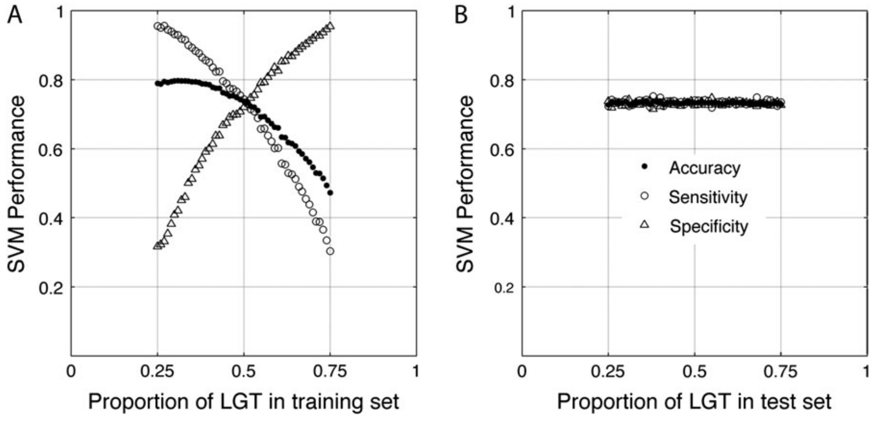

FIG. 5.—Performance of the classifier under different LGT proportions in the training set (*A*) and in the test set (*B*). In (*B*), the LGT/VGI ratio was adjusted to 1.

Table 2
Prediction of LGT/VGI Using an SVM Classifier Trained with Alignment Properties

Combination of Training Parameters											Performance		
Number of OTUs	Shannon Entropy	Average Pairwise Identity	Proportion of Gaps	Proportion of Invariant Sites	Proportion of Informative Sites	Alignment Length	CS (HoT)	SPS (HoT)	Number of Different Phyla	Kingdom Affiliation	Accuracy	Sensitivity	Specificity
v	*v*								*v*		0.797 ± 0.009	0.833 ± 0.020	0.692 ± 0.023
v	*v*							ν	*v*		0.796 ± 0.009	0.835 ± 0.019	0.685 ± 0.021
v	*v*		*v*						*v*		0.796 ± 0.007	0.834 ± 0.015	0.687 ± 0.021
v	*v*				*v*				*v*		0.796 ± 0.009	0.830 ± 0.017	0.699 ± 0.017
v		*v*									0.794 ± 0.009	0.832 ± 0.020	0.688 ± 0.025
v		*v*			*v*				*v*		0.794 ± 0.012	0.828 ± 0.021	0.698 ± 0.018
v	*v*					*v*			*v*		0.794 ± 0.007	0.833 ± 0.015	0.685 ± 0.019
v	*v*	*v*			*v*				*v*		0.794 ± 0.012	0.827 ± 0.025	0.700 ± 0.025
v					*v*				*v*		0.794 ± 0.009	0.831 ± 0.019	0.687 ± 0.021
v		*v*					*v*		*v*		0.793 ± 0.008	0.829 ± 0.016	0.692 ± 0.020
										ν	0.735 ± 0.009	0.931 ± 0.020	0.174 ± 0.022
							v	ν	*v*	ν	0.743 ± 0.008	0.887 ± 0.087	0.331 ± 0.247
								ν		ν	0.711 ± 0.041	0.884 ± 0.081	0.218 ± 0.074
			v					ν	*v*	ν	0.733 ± 0.012	0.873 ± 0.091	0.334 ± 0.229
v							*v*	ν	*v*	ν	0.765 ± 0.019	0.867 ± 0.093	0.474 ± 0.327
v							*v*	ν	*v*		0.784 ± 0.018	0.851 ± 0.061	0.592 ± 0.215
v			*v*				*v*	ν			0.771 ± 0.013	0.849 ± 0.066	0.551 ± 0.216
v			*v*					ν	*v*		0.778 ± 0.015	0.848 ± 0.063	0.576 ± 0.224
			v				*v*	ν	*v*	ν	0.736 ± 0.012	0.848 ± 0.091	0.419 ± 0.235
v			*v*				*v*			ν	0.762 ± 0.014	0.847 ± 0.082	0.519 ± 0.274
		v									0.544 ± 0.015	0.450 ± 0.031	0.812 ± 0.033
		v			*v*			ν			0.598 ± 0.013	0.529 ± 0.025	0.794 ± 0.024
		v						ν			0.576 ± 0.011	0.501 ± 0.023	0.790 ± 0.025
		v			*v*	*v*	*v*				0.586 ± 0.013	0.516 ± 0.027	0.787 ± 0.028
				v	*v*	*v*					0.665 ± 0.019	0.623 ± 0.034	0.786 ± 0.025
		v			*v*			ν		ν	0.601 ± 0.018	0.536 ± 0.037	0.785 ± 0.037
		v			*v*		*v*				0.590 ± 0.012	0.522 ± 0.028	0.785 ± 0.034
		v	*v*		*v*		*v*				0.592 ± 0.019	0.525 ± 0.040	0.783 ± 0.043
		v			*v*						0.582 ± 0.011	0.511 ± 0.024	0.782 ± 0.027
		v					*v*				0.562 ± 0.014	0.484 ± 0.030	0.782 ± 0.029

NOTE.—Alignment properties included in the training process are marked with *v*. The LGT/VGI ratio in the training set was adjusted to 1. Only combinations yielding the best 10 performance values for accuracy, sensitivity, and specificity are shown, respectively. Definitions of accuracy, sensitivity, and specificity can be found in the text. A table presenting the performance of all possible combinations is presented in the Supplementary Material online.

In other words, the alignment properties of the LGT and VGI groups, although having strongly overlapping distributions for all parameters (fig. 1; table 1), are nonetheless sufficiently different in a consistent manner that we can correctly predict about 78% of the time whether a Bayesian phylogenetic inference will generate a branch from a given alignment that is sufficiently discordant to be scored as an LGT. On the strength of this finding and circumstance that for each alignment parameter the LGT alignments were always skewed toward values that are known from simulation studies to generate incorrect branches (Nei 1996), it is likely that reliable construction of phylogenetic trees is affected and incorrectly reconstructed branches may be a possible source of LGT inference. The correlations are consistent with the view (Landan and Graur 2007) that sequence sets problematic at the level of alignments are likely to be problematic at the level of phylogenetic inference as well.

In principle, one could use our trained SVM on other alignment data sets in order to predict which alignments will result in discordant branches comparing with a reference tree. However, one would still have to distinguish between discordant branches stemming from either genuine LGTs or phylogenetic reconstruction artifacts. The results presented here indicate that the latter are more frequent in problematic alignments; hence, alignment quality has high impact on evolutionary inference from phylogenetic trees. A similar observation was recently presented for phylogenetic inference of ancient LGTs during the endosymbiosis of plastids (Deusch et al. 2008). This indicates that it is important to monitor and assess alignment quality in large-scale phylogenetic analyses, particularly those implementing automated or semiautomated phylogeny pipelines.

Supplementary Material

Supplementary tables S1–S4 and supplementary figures S1–S6 (and additional supporting figures) are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Giddy Landan and David Bryant for discussions. This work was funded by the Deutsche Forschungsgemeinschaft (M.R., W.M.), and the German-Israeli Foundation for scientific research and development (T.D.). We would like to thank the central computing facility of Düsseldorf University.

Literature Cited

Beiko RG, Chan CX, Ragan MA. 2005. A word-oriented approach to alignment validation. Bioinformatics. 21:2230–2239.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci. 102:4332–14337.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Christiani N, Shawe-Taylor J. 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge (MA): Cambridge University Press.

Delsuc F, Phillips MJ, Penny D. 2003. Comment on "Hexapod origins: monophyletic or paraphyletic?". Science. 301:1482d.

Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, Martin W, Dagan T. 2008. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. Mol Biol Evol. 25:748–761.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res. 32:1792–1797.

Elhaik E, Sabath N, Graur D. 2006. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. Mol Biol Evol. 23:1–3.

Felsenstein J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 266:418–427.

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

Gotoh O. 1996. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J Mol Biol. 264:823–838.

Grasso C, Lee C. 2004. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics. 20:1546–1556.

Hillis DM. 1995. Approaches for assessing phylogenetic accuracy. Syst Biol. 44:3–16.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–282.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucl Acids Res. 30:3059–3066.

Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.

Landan G, Graur D. Forthcoming. 2008. Characterization of pairwise and multiple sequence alignment errors. Gene, doi: 10.1016/j.gene.2008.05.016.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol. 19:1–7.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 320:1632–1635.

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

Nei M, Takezaki N, Sitnikova T. 1995. Assessing molecular phylogenies. Science. 267:253–254.

Notredame C, Higgins DG, Heringa J. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302:205–217.

Penny D, Hendy MD, Steel M. 1992. Progress with methods for constructing evolutionary trees. Trends Ecol Evol. 7:73–79.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucl Acids Res. 33:501–504.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res. 22:4673–4680.

Valdar WSJ. 2002. Scoring residue conservation. Proteins. 48:227–241.

Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. Science. 319:473–476.

Dan Graur, Associate Editor

Accepted May 5, 2009