Alignment-Free Genome Tree Inference by Learning Group-Specific Distance Metrics

Kaustubh R. Patil[1,3,*] and Alice C. McHardy[1,2]

[1]Max-Planck Research Group for Computational Genomics and Epidemiology, Max-Planck Institute for Informatics, Saarbrücken, Germany

[2]Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Germany

[3]Present address: Affective Brain Lab, University College London, London, United Kingdom

*Corresponding author: E-mail: patil@mpi-inf.mpg.de.

Accepted: July 4, 2013

Abstract

Understanding the evolutionary relationships between organisms is vital for their in-depth study. Gene-based methods are often used to infer such relationships, which are not without drawbacks. One can now attempt to use genome-scale information, because of the ever increasing number of genomes available. This opportunity also presents a challenge in terms of computational efficiency. Two fundamentally different methods are often employed for sequence comparisons, namely alignment-based and alignment-free methods. Alignment-free methods rely on the genome signature concept and provide a computationally efficient way that is also applicable to nonhomologous sequences. The genome signature contains evolutionary signal as it is more similar for closely related organisms than for distantly related ones. We used genome-scale sequence information to infer taxonomic distances between organisms without additional information such as gene annotations. We propose a method to improve genome tree inference by learning specific distance metrics over the genome signature for groups of organisms with similar phylogenetic, genomic, or ecological properties. Specifically, our method learns a Mahalanobis metric for a set of genomes and a reference taxonomy to guide the learning process. By applying this method to more than a thousand prokaryotic genomes, we showed that, indeed, better distance metrics could be learned for most of the 18 groups of organisms tested here. Once a group-specific metric is available, it can be used to estimate the taxonomic distances for other sequenced organisms from the group. This study also presents a large scale comparison between 10 methods—9 alignment-free and 1 alignment-based.

Key words: sequence comparison, alignment, alignment-free, genome signature, taxonomy, genome tree, distance metric learning.

Introduction

We here address the problem of inferring distances between whole genome (genic + nongenic) sequences to recover their evolutionary relationships in the form of a tree that we will refer to as the genome tree. The evolutionary relationships between different organisms, and hence their genomes, are typically represented in the form of a phylogenetic tree. Phylogenies are often inferred from individual gene sequences, such as the highly conserved small subunit ribosomal RNA (Woese and Fox 1977) or from a set of conserved orthologous genes (Ciccarelli et al. 2006; Wu and Eisen 2008). Phylogenies inferred from different genes or gene sets often disagree with each other and only show a plausible evolutionary history for the genes used which is not necessarily the evolutionary history of the analyzed taxa (Hasegawa and Hashimoto 1993; Karlin and Cardon 1994). Furthermore, to apply gene-based methods, one must first identify orthologous genes from different organisms, which can be difficult due to evolutionary processes such as gene loss, duplication, and horizontal transfer (Doolittle 1999). With the availability of a large number of completely sequenced genomes whole-genome based methods were proposed to alleviate the shortcomings of gene based methods which have attracted much attention in recent years. Various properties of the genome such as gene content, gene order, whole genome sequence similarity, and nucleotide composition biases have been used to measure distances between genomes (see Coenye et al. 2005; Delsuc et al. 2005; Snel et al. 2005 for recent reviews). In this work, we focused on the analysis of sequence-based methods for which no additional information, such as gene annotations, is required.

Two fundamentally different methods are commonly employed for sequence comparison; alignment-based and alignment-free. Alignment methods, such as the basic local alignment search tool (BLAST) (Altschul et al. 1990), are used to identify orthologs from different taxa based on sequence similarity, which subsequently can be analyzed with standard phylogenetic inference methods to infer their evolutionary relationships. There are two major shortcomings of alignment-based methods: 1) alignment methods cannot be applied to sequences that are not well-conserved across taxa and thus have no orthologs and 2) they are computationally expensive. Alignment-free methods are therefore employed to address these shortcomings; however, they tend to be less accurate than alignment-based methods in some settings (Vinga and Almeida 2003; Höhl and Ragan 2007; Reinert et al. 2009). Alignment-free methods utilize the "genome signature," the evolutionary signal that is contained in the oligonucleotide composition of microbial genomes (Blaisdell 1986; Karlin and Burge 1995). However, the signal strength varies for different groups of genomes (Mrazek 2009). An important property of the genome signature is that it allows comparison between nonhomologous sequences. For a given species or higher-level clade, it allows an accurate distinction for 1,000 bp or longer segments, with longer segments encoding a stronger signal (Deschavanne et al. 1999; Sandberg et al. 2001; Jernigan and Baran 2002; McHardy and Rigoutsos 2007; Patil et al. 2011). As more whole genome sequences are deposited in public databases, in comparison with alignment-based approaches computationally less expensive alignment-free methods become increasingly attractive for the analysis of large-scale data sets (Höhl et al. 2006; Yang and Zhang 2008). Some limitations of the genome signature have been pointed out, such as a lower correlation with phylogenetic distance, especially for distantly related genomes (Mrazek 2009), as well as the clustering of distantly related genomes with similar GC-content (Coenye and Vandamme 2003; Pride et al. 2003; van Passel et al. 2006; Takahashi et al. 2009).

In alignment-free sequence comparison, most research has focused on the identification of the appropriate length for oligonucleotides (Karlin and Burge 1995; Karlin et al. 1997; Kirzhner et al. 2002; Pride et al. 2003; Wu et al. 2005; Mrazek 2009; Sims et al. 2009; Takahashi et al. 2009), normalization procedures (Hao and Qi 2003; Xu and Hao 2009), and different distance functions (Wu et al. 1997; Kirzhner et al. 2002; Höhl et al. 2006). The genome signature is inherently redundant due to the reverse complementarily of the DNA strands. Under the influence of selection, all oligonucleotides might not be equally important in taxonomic distance calculation, in case they evolve at different rates. These issues have not been given enough attention. Based on the hypothesis that a group of genomes with similar phylogenetic, genomic or ecological attributes might have specific oligonucleotide weights that reflect their importance in distance calculation, we propose a novel method that aims at improving genome signature-based inference of genome trees. Thus, our goal is to enhance the signal for a group by learning group-specific oligonucleotide weights. We propose a supervised distance metric learning method that exploits the structure of a known reference taxonomy to guide the learning process (see Materials and Methods). We use the taxonomy as reference for calculation of phenetic distances, rather than a phylogeny (such as one inferred from the 16S rRNA gene), due to its "polyphasic" nature that takes genotypic and phenotypic aspects into account (Vandamme et al. 1996) and not to bias our analysis toward possible shortcomings of gene-based methods. However, we verified that phenetic distance strongly correlates with phylogenetic distance (see Materials and Methods).

The aim of our method is to identify a diagonal positive semi-definite matrix (supplementary text, Supplementary Material online) parameterizing Mahalanobis distance metric such that it maximizes the Spearman's rank correlation coefficient between the resulting distances and the phenetic distances within the reference taxonomy. This distance metric learning problem is posed as a regularized optimization problem (see Materials and Methods). We identified 18 groups based on phylogenetic, genomic, or ecological factors. Contrary to other genome tree inference methods, our aim is to improve performance for a group of genomes defined by a common factor, such as genome-wide GC-content or habitat, and not to reconstruct the entire tree of life. When the species composition or ecological characteristics of the organisms at hand is approximately known, one can learn a group-specific distance metric using other available reference data. Once a specific distance metric has been learned, it can be employed for the analysis of novel genome sequences from the same group.

Various methods have been proposed for the evolutionary comparisons of entire genomes or large genome segments, including alignment-free methods (Burge et al. 1992; Karlin and Cardon 1994; Kirzhner et al. 2002; Pride et al. 2003; Qi et al. 2004; Sims et al. 2009; Takahashi et al. 2009; Li et al. 2010) and the alignment-based methods, such as the genome blast distance phylogeny (GBDP) (Henz et al. 2005). A direct comparison between genome tree inference methods is lacking, especially with the alignment-based method GBDP. Therefore, in addition to proposing a new method, we also present a large-scale numerical comparison of the performance of 10 genome tree inference methods, including 9 alignment-free methods and 1 alignment-based method.

Materials and Methods

Following the notation used in (Mrazek 2009), each genome signature is denoted by a pattern lknm, where lk denotes an oligonucleotide of length k and nm is the oligonucleotide length m used for normalization. Thus, for example, the tetranucleotide signature normalized using base frequencies is denoted as l4n1. The notation is optionally followed by the alphabet used (e.g., "ry"), if an alphabet other than nucleotide was used.

We used 1,076 complete microbial genome sequences available from NCBI in April 2010 for this study. This corresponds to 578,350 pairs of taxa to compare in terms of their taxonomic and genomic distances. To compute pairwise distances between species, nine alignment-free methods for computing pairwise genome distances were tested; the Euclidean distance based on the l4n1 genome signature, the Euclidean distance based on the l4n1 signature after dimensionality reduction with principal component analysis (PCA), the Euclidean distance based on the l6n1 signature, CVTree with the l6n5,4 signature (Hao and Qi 2003), the compositional spectrum based on the l10r2 signature and $n = 200$ (Kirzhner et al. 2002) and the feature frequency profile based on the RY alphabet with $l = 10$ (Sims et al. 2009). In addition, we also evaluated the GBDP method based on BLAST alignments (Henz et al. 2005), for which we aligned all pairs of genomes. Pairwise alignments between nucleotide sequences were generated using the "bl2seq" program (version 2.2.18) with default parameters.

The genomes were subsequently classified into 18 groups according to the following five factors: Phylum membership (4 groups), genomic GC-content (3 groups), habitat (5 groups), temperature range (3 groups), and oxygen requirement (3 groups). For each of these factors, the groups were exclusive (supplementary table S1, Supplementary Material online).

Genomes, Taxonomy, and Ecological Information

Genome sequences were obtained from GenBank (http://www.ncbi.nlm.nih.gov/genome, last accessed July 29, 2013). The taxonomy from the NCBI taxonomy database (http://www.ncbi.nlm.nih.gov/Taxonomy/, last accessed July 29, 2013) and the ecological information was obtained from the NCBI "lproks" service (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, last accessed July 29, 2013) (Sayers et al. 2009).

Genome Signature

The dinucleotide genome signature (Josse et al. 1961) was extensively studied by Karlin and colleagues (Karlin and Burge 1995; Karlin et al. 1997). It is defined as relative abundance of dinucleotides over long stretches of DNA, typically 50 kb covering the whole genome. The dinucleotide signature tends to be preserved throughout the genome and to be more similar among closely related organisms than among distantly related organisms. The genome signature concept was subsequently extended to incorporate longer oligonucleotides (Pride et al. 2003) and different normalization strategies.

The genome signature represents a sequence as a point in a multidimensional metric space. The dimensionality of the space is defined by the size of the alphabet and the length of oligonucleotides. In our case, the alphabet comprises four nucleotides (A, T, G, and C) and the oligonucleotide length considered is four, which gives rise to a 4^4 dimensional space. The vector representation of sequences allows application of distance metric functions to these points to uncover their interrelationships.

The elements of a tetranucleotide signature vector normalized based on mononucleotide frequencies for a genome G are defined as:

$$\alpha^{\mathrm{l4n1}}_{abcd|\mathrm{G}} = \frac{f(abcd)}{f(a)f(b)f(c)f(d)}. \quad (1)$$

Here, f denotes the frequency of an oligonucleotide. Thus, a tetranucleotide signature contains 256 elements (4^4), each corresponding to a possible tetranucleotide. To take the double stranded nature of the DNA into account, we averaged the values of the elements and their corresponding reverse complements (rev_comp).

$$\alpha*^{\mathrm{l4n1}}_{abcd|\mathrm{G}} = \frac{\alpha^{\mathrm{l4n1}}_{abcd|\mathrm{G}} + \alpha^{\mathrm{l4n1}}_{\mathrm{rev\_comp}(abcd)|\mathrm{G}}}{2}. \quad (2)$$

The hexanucleotide signature l6n1 was calculated in a similar fashion.

Phenetic Distances between Pairs of Taxa in the Reference Taxonomy

As our target variable, or reference distance, we used the phenetic distance between taxa in the NCBI taxonomy. The phenetic distance between a pair of taxa was defined as the maximum number of edges in the path between one of the taxa in the pair and their lowest common ancestor. Seven major taxonomic ranks; species, genus, family, order, class, phylum, and superkingdom, were used to calculate the phenetic distances. Note that the number of edges to the lowest common ancestor can differ in the NCBI Taxonomy for two taxa at a given rank, due to missing internal nodes on the path from these taxa to their lowest common ancestor. The matrix containing pairwise phenetic distances will be denoted as $\mathrm{D_{TAX}}$.

To compare the phenetic distances with phylogenetic distances, aligned 16S rRNA gene sequences were obtained from the greengenes database (DeSantis et al. 2006) (http://greengenes.lbl.gov, last accessed July 29, 2013). When multiple genes were available for an organism only the first was chosen. In total, genes for 887 organisms were identified. Pairwise distances between the aligned genes were calculated with the "dist.seqs" function emulating the DNADIST distance in the Mothur package (Schloss et al. 2009). The phenetic distances showed a strong correlation with the phylogenetic distances (Pearson's $R = 0.84$ and Spearman's $\rho = 0.81$, $P = 0.001$ based on 999 permutations). This suggests that our results should be valid if 16S rRNA distances were used as reference instead of phenetic distances.

Comparing Trees Based on Cophenetic Correlation

The correlation between two tree path metrics has been used to compare tree topologies (Pazos and Valencia 2001; Kuramae et al. 2007). We here used a similar approach to search for a distance metric which best approximates the phenetic distances between pairs of taxa in a given reference tree. As we were interested in the topology of the trees and not branch lengths, we used Spearman's rank correlation coefficient to quantify the agreement between the phenetic distances in the reference topology and pairwise distances between genome sequences. Although commonly used, Pearson correlation between distance matrices does not always imply better topology recovery (Lapointe and Legendre 1992). Spearman's rank correlation is furthermore more appropriate when outliers are present and there is a nonlinear relationship between the variables. As we are calculating correlation between two symmetric matrices, they are first vectorized using either the upper or lower half triangle. Spearman's ρ is calculated on the ranks of elements $\mathbf{x}_i$ and $\mathbf{y}_i$ in the vectorized distance matrices according to the following equation:

$$\rho(\mathbf{x},\mathbf{y}) = \frac{\sum_i (\mathbf{x}_i - \bar{x})(\mathbf{y}_i - \bar{y})}{\sqrt{\sum_i (\mathbf{x}_i - \bar{x})^2 \sum_i (\mathbf{y}_i - \bar{y})^2}}. \tag{3}$$

The correlation between a data-derived matrix of pairwise distances and a phenetic distance matrix is also known as the cophenetic correlation coefficient (CPCC) (Sokal and Rohlf 1962). The CPCC has been used for assessing how well tree topologies inferred with different hierarchical clustering methods agree with a matrix of pairwise distances inferred from the data. Here, we use it to evaluate how well different data-derived distance metrics agree with phenetic distances between pairs of taxa in reference taxonomy. Although typically Pearson correlation is used to calculate CPCC, the use of rank based correlation has been proposed before (Johnson 1967; Mrazek 2009).

Topological Distance between Trees

As the cophenetic correlation might not directly correspond to topological similarity (Farris 1969), we also calculated topological distances between trees. The topological distances between trees were calculated using the normalized quartet distance, as implemented in the program "QDist" (Nielsen et al. 2011) version 2.0, downloaded from http://birc.au.dk/software/qdist/.

Note that an increase in congruence between tree topologies results in an increase in the CPCC and a decrease in the quartet distance. The cophenetic correlation was used also as the optimization criterion (discussed later).

Distance Metric Learning

The Euclidean distance metric is often used to calculate dissimilarities for data that can be represented as points in a multidimensional metric space. However, it may not be ideal to infer taxonomic distance between pairs of genome signatures. This is particularly true when some of the variables are more important than others or when some dimensions are correlated and/or have different scales, for instance, some different genomic features could be subject to different evolutionary constraints and evolve at different rates. In such cases, a more suitable distance metric than the Euclidean metric can be learned from data. Originally, distance metric learning was proposed for clustering applications where side *information* such as similarity and dissimilarity constraints is available (Xing et al. 2002). The information available in our case is the phenetic distances between pairs of taxa in the reference taxonomy.

Distance metric learning can be viewed as a transformation of the input space into another (possibly lower dimensional) space, in which the Euclidean distance between the points represent as accurately as possible the target relationships (Jain et al. 2012). Practically, this can be achieved by using the Mahalanobis distance function. The Mahalanobis distance is a distance metric, parameterized by a positive semi-definite matrix $\mathbf{M} \in \Re^{p \times p}$. The Mahalanobis distance between two vectors $\mathbf{x}, \mathbf{y} \in \Re^p$ is defined as;

$$\mathrm{Mahal}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{M}(\mathbf{x}-\mathbf{y})}. \tag{4}$$

We propose learning a diagonal matrix $\mathbf{M}$ with nonnegative entries that maximizes the performance criterion; that is the Spearman's correlation coefficient between the resulting $n \times (n-1)/2$ pairwise Mahalanobis distances for n analyzed genome signatures with the corresponding target phenetic distances. The entries in the target distance matrix, $\mathbf{D}_{\mathbf{TAX}}$, were defined as described earlier. The diagonal elements of the matrix $\mathbf{M}$ represent the relative weights for the corresponding oligonucleotides. The Euclidean distance is a special case of the Mahalanobis distance, when it is parameterized by an identity matrix. The Mahalanobis distance corresponds to a weighted Euclidean distance, when it is parameterized with a diagonal matrix. We will denote a function that returns all pairwise Mahalanobis distances between a set of vectors S given a parameterizing matrix $\mathbf{M}$ as $\mathrm{D}_{\mathrm{Mahal}}$.

Even though a learned metric works well for a given set of signatures (training data), it might not provide improvement for novel signatures (test data). Such over-fitting is not desirable and hence we pose the learning problem as a regularized optimization problem with the following objective function;

$$\min_{\mathbf{M}} (1 - \rho(\mathrm{D}_{\mathrm{Mahal}}(\mathbf{S},\mathbf{M}), \mathbf{D}_{\mathbf{TAX}})) + \lambda \frac{\sum_{i=1}^{p} \mathbf{M}_{ii}}{p} \tag{5}$$

$$\text{s.t.} \quad 0 \le \mathbf{M}_{ii} \le 1 \quad i \in 1 \ldots p\,.$$

Here, p is the number of oligonucleotides and $\mathbf{S}$ is a matrix with each row representing a genome signature, while first term in the objective function maximizes correlation, the second term is a regularizer that controls complexity of the solutions in terms of the L1-norm of the diagonal entries of $\mathbf{M}$. Thus, higher values of λ ($\lambda \geq 0$) will lead to sparse diagonal entries. As only the relative contributions of the oligonucleotides and not their absolute magnitudes are important, the diagonal entries of $\mathbf{M}$ were constrained to values within the interval [0, 1], to allow comparisons between solutions for different experiments. The parameter λ was varied in the grid {0, 0.1, 1, 10}. For each value in the grid, a 3-fold cross-validation procedure was performed on randomly partitioned training data as follows; three metrics were learned separately by excluding each of the three partitions and the generalization performance was assessed with the Spearman's correlation between the target distances and the distances with the learned metric on the excluded partition. The resulting three correlations for each λ value were averaged to get an estimate of the generalization performance. The value with the highest generalization performance was chosen to learn a metric on the complete training data. The aim of the regularizer here is obtaining generalizable solutions and not to enforce sparse solutions. Thus, if a less sparse solution yields a higher generalization performance (as estimated by cross-validation) than a more sparse solution, then the less sparse solution is selected. Note that although it is possible to formulate the optimization problem we describe here with a weight vector instead of the matrix $\mathbf{M}$, the more general formulation clarifies that this method is easily adaptable for learning a full matrix.

We used the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) (Hansen et al. 2003) as the optimization procedure. Any other global optimization procedure can be used. As this optimization problem is nonlinear and nonconvex, gradient-based optimization techniques are not appropriate. The python code for CMA-ES was obtained from the website http://www.lri.fr/~hansen/cmaes_inmatlab.html. The tolerance for solution improvement was set to 1e−3 and the number of iterations was set to 500 during cross-validation and 1,000 for learning the metric with a selected λ. Only the diagonal of the covariance matrix was adapted to reduce the computational complexity. The population size for CMA-ES was set to 20 and the step-size to 0.5.

Distance Metrics

The distance metrics used for comparison are described later. The metrics were chosen to reflect the diversity of the popular metrics found in the literature, in terms of oligonucleotide lengths, normalization strategies and distance metrics. In the following p denotes the length of the genome signature vectors.

Group Specific

The group-specific distance between two signatures of genomes from a group is given by the following:

$$\mathrm{Specific}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{p} (\mathbf{x}_i - \mathbf{y}_i)^2 \mathbf{M}_{ii}}, \qquad (6)$$

where $\mathbf{M}$ is a diagonal matrix learned by maximizing the estimated generalization performance with training data from the same group (as $\mathbf{x}$ and $\mathbf{y}$). For simplicity, the group-specific distance metrics will be referred to as specific distance metrics.

Random Learned

The random distance between two signatures calculated for a pair of genomes from a group is given by the following equation:

$$\mathrm{Rand}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{p} (\mathbf{x}_i - \mathbf{y}_i)^2 \mathbf{M}_{ii}}, \qquad (7)$$

where $\mathbf{M}$ is a diagonal matrix learned by maximizing estimated generalization performance using randomly selected training data. For simplicity, this metric will be referred to as the random metric.

Euclidean

The Euclidean distance between two signatures is defined as

$$\mathrm{Eucl}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{p} (\mathbf{x}_i - \mathbf{y}_i)^2}. \qquad (8)$$

This distance was used with the l4n1 and l6n1 signatures.

Euclidean PCA

This distance was calculated similarly to the Euclidean distance, but in a lower dimensional space after application of PCA to retain either the principal components explaining at least one original variable, that is the principal components with eigenvalue ≥ 1 or three principal components, whichever is larger. This distance metric was used with the l4n1 signature.

Delta

The delta distance (Mrazek 2009) between two signatures is defined as following:

$$\mathrm{Delta}(\mathbf{x},\mathbf{y}) = \frac{1}{p}\sum_{i=1}^{p} |\mathbf{x}_i - \mathbf{y}_i|. \qquad (9)$$

The delta distance between two genomes G1 and G2 was calculated using all pairs of nonoverlapping 50 kb segments. If n_1 and n_2 are number of nonoverlapping segments X and Y in genomes G1 and G2, respectively, then the delta distance between the genomes was calculated as follows:

$$\mathrm{Delta\ 50kb}(G1, G2) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathrm{Delta}(X_i, Y_j). \qquad (10)$$

This distance was used with the l4n1 signature.

CVTree

The CVTree signature was calculated using oligonucleotides of length 6 normalized by its constituent 4- and 5-mers (Gao et al. 2007), that is the l6n5,4 signature. The sequences were appended with their reverse complement for calculating this signature. The expected frequency of a hexanucleotide *abcdef* was calculated as;

$$f^0(abcdef) = \frac{f(abcde)f(bcdef)}{f(bcde)} \times \frac{(L-k+1)(L-k+3)}{(L-k+2)^2}.$$

Then, the normalized elements of the signature vector were calculated as follows:

$$\begin{aligned}\alpha(abcdef) &= \frac{f(abcdef) - f^0(abcdef)}{f^0(abcdef)} && \text{if } f^0 \neq 0\\ \alpha(abcdef) &= 0 && \text{if } f^0 = 0\end{aligned}$$

The distances between the resulting vectors was calculated using the cosine similarity as follows;

$$\mathrm{CVTree}(\mathbf{x}, \mathbf{y}) = \frac{1 - \mathrm{cosine}(\mathbf{x}, \mathbf{y})}{2}. \qquad (11)$$

Compositional Spectrum

Compositional spectrum (CompSpec) distances over the DNA alphabet {A, T, C, G} were calculated using the parameter settings as in (Kirzhner et al. 2007). We first generated 200 random words of length 10 and then counted their imperfect occurrences of up to 2 mismatches (l10r2 signature) over the complete genomes. The distances between the resulting 200 dimensional vectors were calculated using Spearman's rank correlation coefficient ρ as follows:

$$\mathrm{CompSpec}(\mathbf{x}, \mathbf{y}) = 1 - \rho(\mathbf{x}, \mathbf{y}) \qquad (12)$$

Feature Frequency Profile

The feature profile frequency (FFP) distances were calculated using the program ffp version 3.19, downloaded from http://ffp-phylogeny.sourceforge.net/. The default settings of two-letter RY alphabet was used with the lengths of l-mers set of 10 (l10ry signature). The distance between the normalized feature frequency profile vectors $\mathbf{x}$ and $\mathbf{y}$ were calculated using the Jensen-Shannon divergence as follows:

$$\mathrm{FFP}(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\mathrm{KL}(\mathbf{x}, \mathbf{z}) + \frac{1}{2}\mathrm{KL}(\mathbf{y}, \mathbf{z}). \qquad (13)$$

Here, $\mathbf{z}_i = (\mathbf{x}_i + \mathbf{y}_i)/2$ and KL is the Kullback–Liebler divergence.

Genome BLAST

The whole-genome BLAST distances between two genomes were calculated with alignments inferred with the bl2seq program of the NCBI BLAST executables (version 2.2.18) with default parameters. The resulting tabular report was then parsed using BioRuby (version 1.4.1) and the high scoring pairs were converted into a similarity score using the greedy version of the GBDP algorithm without trimming (Henz et al. 2005). Because of computational restrictions, we used only one directional alignment instead of averaging over both directions.

$$\mathrm{GBDP}(G1, G2) = -\log\frac{|G1_{match}| + |G2_{match}|}{2 \times \min(|G1|, |G2|)} \qquad (14)$$

Measures of Group Phylogenetic Structure

We calculated two metrics of group phylogenetic structure. The metrics; net relatedness index (NRI) and nearest taxon index (NTI) correspondingly quantify the distribution of the taxa relative to a phylogeny (Webb et al. 2002). Although both NRI and NTI increase with increasing clustering, they become negative with increasing dispersal of taxa. Clustering at the terminal nodes causes more increase in NTI relative to NRI. We calculated both measures with respect to the reference taxonomy for each of the 18 groups using 999 randomizations. The corresponding methods were implemented in the R statistical environment (version 2.11.1, http://www.r-project.org/) (see supplementary text [Supplementary Material online] for details).

Other Methods

The distances were subsequently used to construct ultrametric trees, which were inferred using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm implemented in the "phangorn" package for the R statistical environment. The resulting tree topologies were compared with the reference tree topology based on the quartet distance. PCA was performed in R (version 2.11.1) with the "princomp" function. These data were centered and scaled to unit variance before performing PCA.

Data Availability

The data used in this study and the learned metrics can be obtained from http://algbio.cs.uni-duesseldorf.de/webapps/wa-download/index.php.

Results

Setup

The tetranucleotide signature corrected for bias in base frequencies (l4n1), that is, normalized using the zero-order Markov criterion, was chosen to learn the metrics, as it is has been previously shown to contain a strong phylogenetic signal (Pride et al. 2003; van Passel et al. 2006; Mrazek 2009). The Euclidean distance for the l4n1 signature was used as the baseline for comparison. We used two measures to quantify the performance of the methods: The first is the CPCC (Sokal and Rohlf 1962) using Spearman's rank correlation, which is also a part of the optimization function used to learn the specific metrics (see Materials and Methods). We also calculated the normalized quartet distance (Nielsen et al. 2011) (referred to as quartet distance hereafter) between two trees built with UPGMA; one using the phenetic distances and the other using the genome-based distances (see Materials and Methods). We say that a metric performs better only if it shows improvement on both measures; that is, a higher CPCC and a lower quartet distance. We show results for 18 groups defined by five different attributes: taxonomy, genomic GC-content, habitat, growth temperature, and oxygen requirement (supplementary table S1, Supplementary Material online).

For the proposed metric learning method to be of practical value, it has to be able to learn a generalizable distance metric, meaning a metric that works well on novel genomes not used for learning, from a limited amount of data. Therefore, our experimental setup consisted of randomly sampling genomes of 30 species (one genome per species) from a group and then learning a Mahalanobis metric from the corresponding l4n1 signatures, guided by the target phenetic distances, such that the estimated generalization performance is maximized (see Materials and Methods). A Mahalanobis metric learned using signatures from one group is referred to as a group-specific metric. The performance of a learned metric was quantified on the test genomes, that is, the genomes from the same group not used for learning the metric. For a set of test genomes, distances were then computed with the learned metric and compared with the corresponding phenetic distances. At the same time, the performance of the other methods was also quantified on the test genomes by comparing their distances with the phenetic distances. To quantify the variability of the learned metrics, this procedure was repeated 30 times for each of the 18 groups by using different random training samples. This resulted in 30 performance measurements for the CPCC and quartet distances for each group and each method, except for the Actinobacteria, for which only 28 metrics were learned. The statistical significance of an observed improvement in the 30 repetitions was tested using a one sided Wilcoxon rank sum test. Although for CPCC, the alternative hypothesis was that a metric produces higher CPCC values than the baseline metric, for the QD, the alternative hypothesis was that a metric results in a lower quartet distances than the baseline metric. For simplicity, both tests will be referred to as Wilcoxon tests.

Furthermore, we used the Hotelling–Williams test to test whether a learned metric resulted in a significantly different CPCC from the baseline (Steiger 1980). Specifically, we tested whether the CPCC of one metric was significantly different from the CPCC of another metric (supplementary text, Supplementary Material online).

Phylum

We begin by showing that the taxonomic signal of the l4n1 genome signature can be improved with specifically learned metrics for phylogenetic groups at the phylum level. Four extensively sequenced phyla, the Proteobacteria, Firmicutes, Actinobacteria, and Euryarchaeota, were chosen for this analysis (supplementary table S1, Supplementary Material online). Our results show that better distance metrics, that is higher cophenetic correlation and lower quartet distance on the test genomes when compared with the baseline, could be learned for the phylogenetic groups except for Euryarchaeota, where the learned metrics did not show improvement over the Euclidean metric (fig. 1; supplementary table S3, Supplementary Material online). The Proteobacteria metrics showed only a marginal, but significant ($P < 0.05$, Wilcoxon test) improvement, which might be due to its diverse and nonmonophyletic nature (Garrity 2005). A disagreement of the inferred tree with the reference taxonomy was also observed with the Proteobacterial CVTree based on translated protein products (Li et al. 2010). The best performance improvement due to specific metrics was observed for the phylum Actinobacteria, where the average cophenetic correlation significantly increased from 0.39 to 0.64 ($P = 8.23\text{e}{-10}$, Wilcoxon test), whereas the average quartet distance decreased from 0.53 to 0.43 ($P = 2.73\text{e}{-13}$, Wilcoxon test). More than 25 (out of the 30) learned metrics showed significantly different correlation coefficients for the Proteobacteria, Firmicutes and Actinobacteria (Hotelling–Williams test, $P < 0.05$) (supplementary fig. S1, Supplementary Material online). The other l4n1 based distances, the Euclidean distances after applying PCA and the delta distances averaged over 50 kb segments, performed either similar or only slightly better than the baseline. The metrics learned from randomly sampled species over the entire taxonomy performed worse than the baseline except for a slight performance improvement for the Actinobacteria.

The phyla-specific metrics also performed better than the l6n1 signature-based Euclidean distances. This shows the advantage of learning specific metrics in comparison with signatures based on longer oligonucleotides. The Euclidean distances based on the l4n1 and l6n1 signatures performed similarly, except for the Actinobacteria, where the l6n1 signature performed better. CVTree with the l6n5,4 signature showed overall better performance than the l6n1 Euclidean

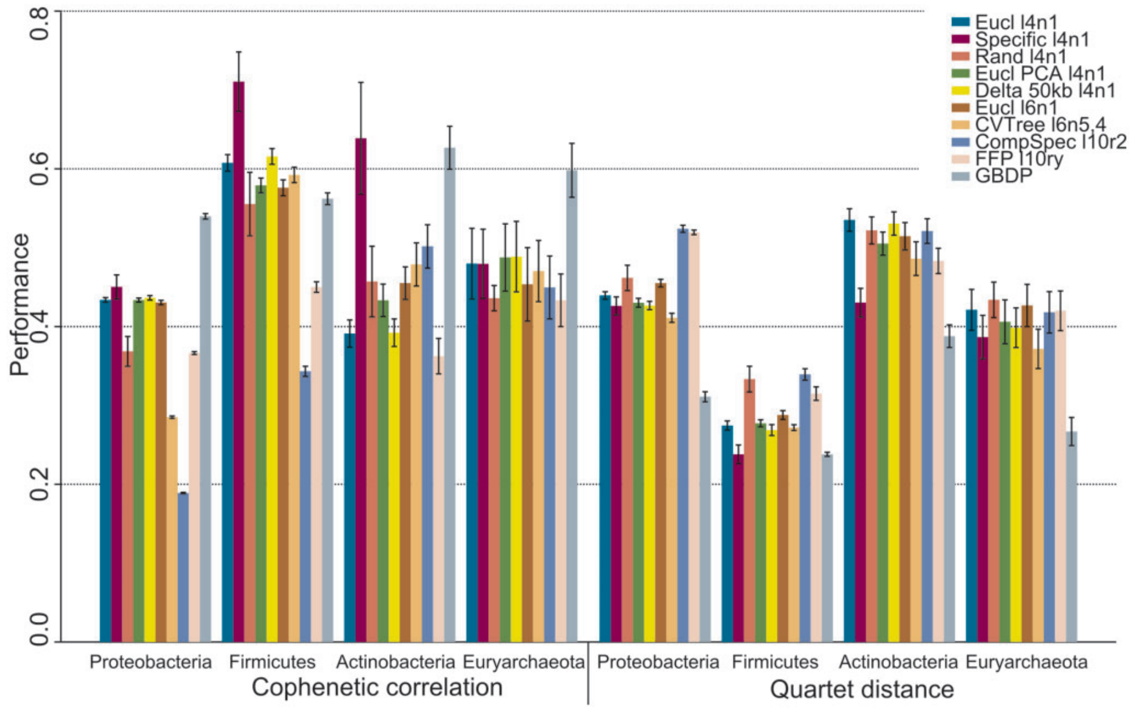

Fig. 1.—Performance on the phylogenetic groups. Each bar shows a performance measure along with error bars showing SD.

distances, the compositional spectrum and FFP distances performed less well in comparison. Interestingly, all signature-based distances with long oligonucleotides (Euclidean l6n1, CVTree l6n5,4, CompSpec l10r2, and FFP l10ry) with lower overall cophenetic correlation, except for FFP, performed better for the Actinobacteria than the baseline ($P < 0.05$, Wilcoxon test). This might be due to the close relatedness of the genomes in the phylum Actinobacteria and their characteristically high GC-content, making longer oligonucleotides more informative. For all groups except Firmicutes, the alignment-based method GBDP performed better than the alignment-free methods, however, this performance comes at a considerable computational cost.

GC-Content

We performed similar experiments with the genomes divided into three groups according to their genome-wide GC-content (≤30%, >30–≤50%, and >50–≤70%; supplementary table S1, Supplementary Material online). It has been previously noted that GC-content affects oligonucleotide-based trees grouping similar GC-content genomes together irrespective of their phylogenetic relationships, while tetra- to octanucleotide frequency based trees of genomes with similar GC-content show high congruence with gene based trees at genus and family level (Takahashi et al. 2009). Therefore, we expected that improved distance metrics could be learned for groups of genomes with similar GC-content. The GC-specific metrics, we inferred improved in cophenetic correlation over the baseline for all three GC-content groups. There was also a decrease in the quartet distance for the genomes with 30% or less GC-content and for genomes with GC-content between 50% and 70%. Most metrics for the individual groups had significantly different correlation coefficients from the baseline method ($P < 0.05$, Hotelling–Williams test) (supplementary fig. S1, Supplementary Material online). In general, while a strong signal was observed for all the alignment-free methods on the low GC-content group, a weaker signal was observed on the moderate GC-content genomes (supplementary table S3, Supplementary Material online). Of the other alignment-free methods, only CVTree consistently and significantly ($P < 8.2e{-}6$, Wilcoxon test) performed better than the baseline. The compositional spectrum and FFP methods performed well only on the genomes with GC-content of 30% or less. GBDP performed better than the baseline in all the groups and performed worse than the learned l4n1 metrics on the ≤30% GC-content group (fig. 2).

Ecological Attributes

Next, we investigated whether specific metrics for ecological groups show an improvement over the baseline. This is a challenging task as ecological groups might contain distantly related genomes, a scenario in which alignment-free methods can face difficulties (Mrazek 2009). Three ecological factors were chosen to define groups: habitat (5 groups), temperature range (3 groups), and oxygen requirement (3 groups; supplementary table S1, Supplementary Material online).

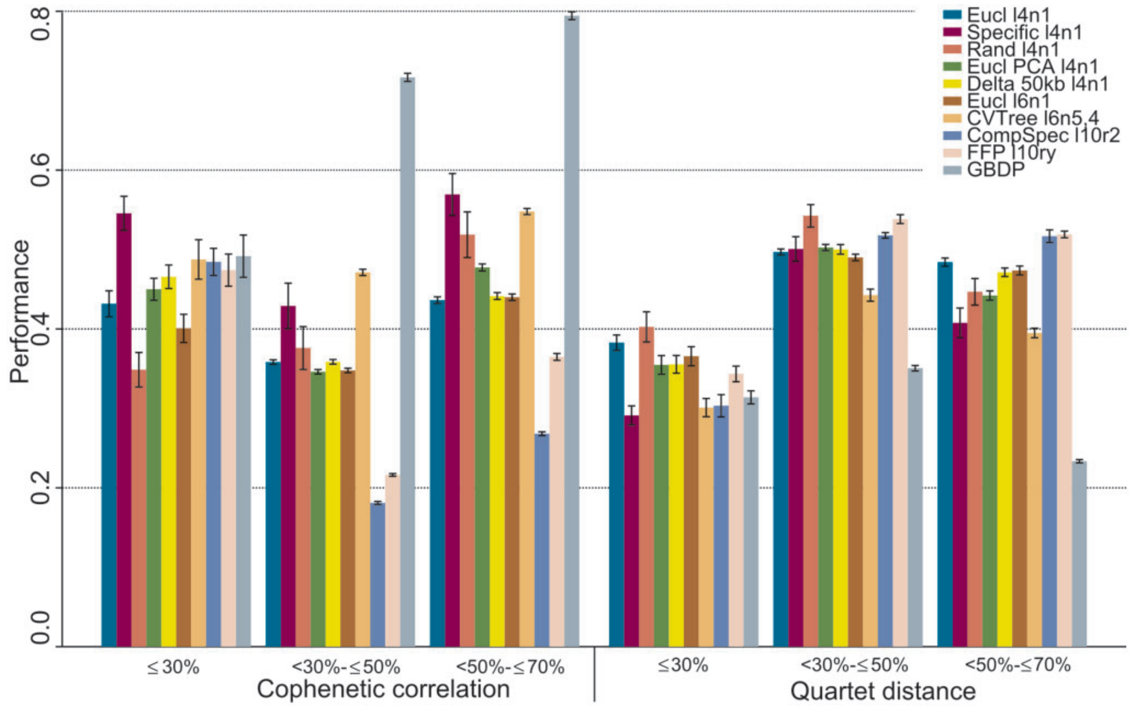

Fig. 2.—Performance on the GC-content groups. Each bar shows a performance measure along with error bars showing SD.

The habitat-specific l4n1 metrics showed an improvement over the baseline both in terms of the CPCC and the quartet distance for all five groups. Only the improvement of the quartet distance for the host-associated metrics was not significant (fig. 3; supplementary table S3, Supplementary Material online). Although CVTree showed an increase in the CPCC for all five habitat groups and an increased quartet distance for the aquatic and specialized groups, FFP showed an improvement over the baseline only for the multiple habitat genomes ($P < 7.74e{-}15$, Wilcoxon test).

In computation of the taxonomic distances and genome trees for genomes from all three temperature range groups, the learned l4n1 metrics performed better than the baseline ($P < 7e{-}3$, Wilcoxon test), except for an increase in the quartet distance for the mesophiles group. Interestingly, for the mesophiles group 19 specific metrics did show a significant change in correlation (supplementary fig. S1, Supplementary Material online). CVTree performed well for all groups except a decrease in the CPCC for hyperthermophiles, while FFP showed improvement only for the hyperthermophiles group ($P < 1.3e{-}3$, Wilcoxon test).

We also observed an improvement for the learned l4n1 metrics for all oxygen-requirement types (aerobe, anaerobe, and facultative anaerobes) ($P < 1.2e{-}6$, Wilcoxon test), except for a performance reduction in term of an increase in the quartet distance for the facultative anaerobes. CVTree, as before, showed improvement for the anerobes and facultative groups ($P < 3.15e{-}15$, Wilcoxon test) and performed similarly to GBDP for the genomes of the facultative anaerobes. Although the Euclidean metric with the l4n1 signature after performing PCA showed a marginal but significant improvement for aerobes and anaerobes, the Delta50kb and Euclidean metric with the l6n1 signature showed significant improvements for the anaerobe and facultative anaerobe groups, respectively. The other methods did not show a consistent performance pattern.

Overall, for all 11 ecological groups 23 or more metrics showed a significant change in the correlation coefficients with the phenetic metric of the reference taxonomy in comparison with the baseline ($P < 0.05$, Hotelling–Williams test). For three habitats—aquatic, host-associated, and specialized—as well as the mesophilic and aerobic groups, all 30 metrics differed significantly (supplementary fig. S1, Supplementary Material online). GBDP performed best for all groups defined by the three ecological attributes ($P < 1.46e{-}9$, Wilcoxon test).

Group-Specific Metrics Notably Improved Tree Inference for Group Members

One could argue that a learned metric performs well for a group by chance and not because it inferred specifics of evolutionary rates for different tetranucleotides for the group. To investigate this question, we learned 30 metrics from 30 randomly selected species each (the "random metrics") and compared their performance with the performance of the 30 group-specific learned metrics for each of the 18 groups

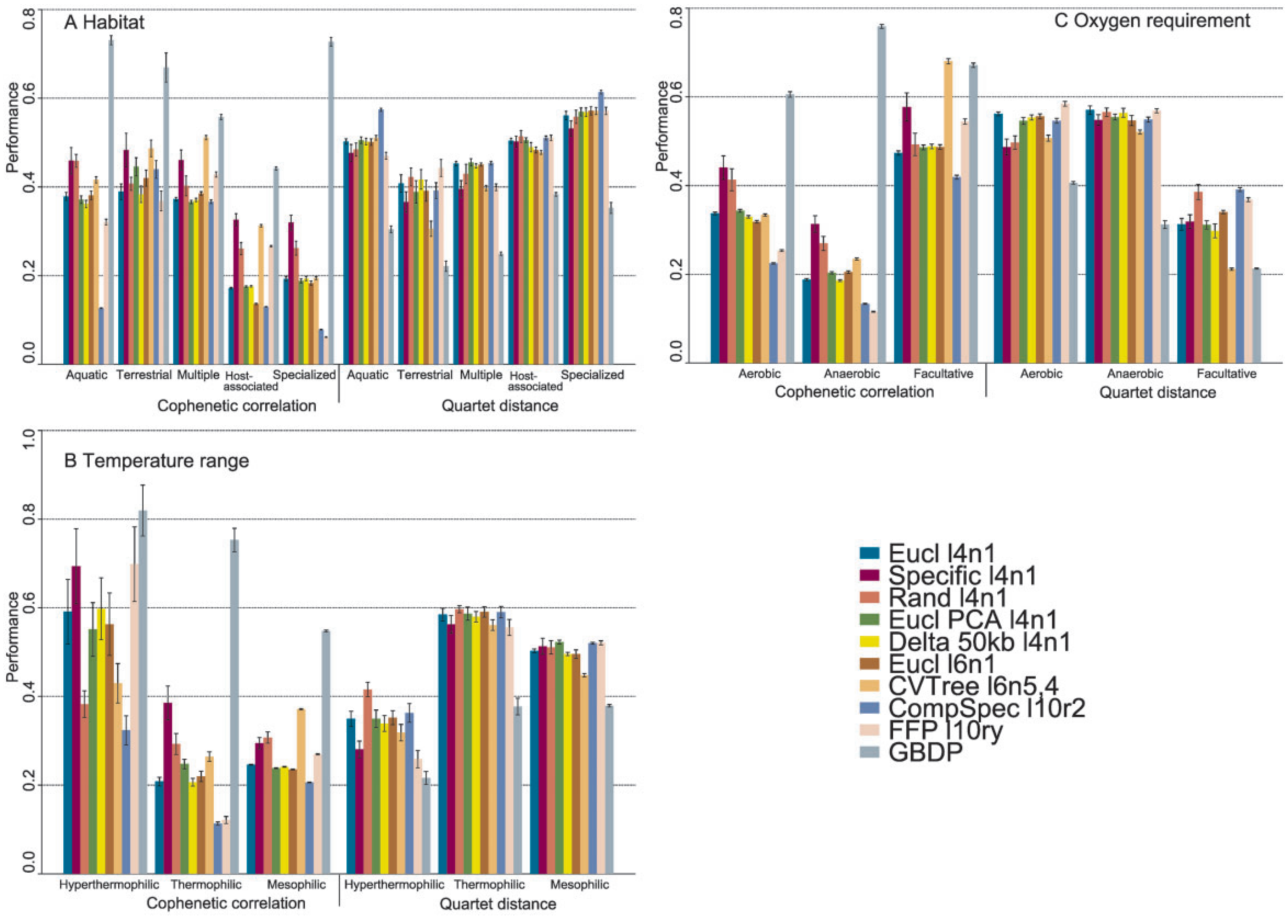

Fig. 3.—Performance on the ecological groups from three attributes. The bars show the performance measures and the error bars indicate SD.

using the one-sided Wilcoxon rank sum test. We tested whether the group-specific metrics had a higher CPCC and lower quartet distance than the random metrics to the reference phenetic distances. Note that the random metrics showed a significantly better performance than the baseline metric for the Actinobacteria, GC content between 50% and 70%, aquatic and aerobic groups ($P < 3.61\text{e}{-}2$, Wilcoxon test) (supplementary table S3, Supplementary Material online).

For all the groups, except the aquatic, mesophiles and aerobes, the group-specific metrics performed significantly better than the random metrics ($P < 3.86\text{e}{-}2$, Wilcoxon test) (table 1). This implies that the group-specific metrics perform better than the ones learned on randomly sampled genomes and that group-specific aspects of tetranucleotide usage allow an improved inference of the taxonomic relationships for the respective organisms. The lack of improvement for aquatic species, mesophiles, and aerobes might be in part caused by abundance of these groups among the genomes (supplementary table S1, Supplementary Material online). This may have resulted in some of the learned metrics from randomly selected species to partially represent specific properties of these groups.

Dimensionality Reduction Resulted in Marginal Improvement

Unsupervised dimensionality reduction techniques, such as PCA, have been used for noise reduction and visualization of genome signatures (Sandberg et al. 2001; Mrazek 2009). PCA embeds the input space into a potentially lower dimensional space defined by orthogonal basis vectors. We calculated cophenetic correlations and quartet distances for all the groups individually using the original and PCA-transformed l4n1 distances (table 2). The dimensionality of the reduced space was selected to be the dimensions explaining at least one original variable, that is, dimensions with eigenvalues of at least one. Interestingly, approximately 20 dimensions (18–25) were retained for all the groups, capturing 93–98% of variance. Although PCA resulted in a marginal nonsignificant improvement, it performed less well than the group-specific metrics (figs. 1–3; supplementary figs. S2–S6, Supplementary Material online). Similarly, when PCA was applied to the l6n1 signature with the Euclidean distance metric, a large reduction in the dimensionality was observed (38–114 principal components explaining 97.81–99.96% variance), with no significant

Table 1

P Values from the One-Sided Wilcoxon Rank Test, Testing the Specificity of the Learned Metrics for the Respective Groups

Attribute	Group	CPCC	QD
Phylum	Proteobacteria	**0.0000**	**0.0001**
	Firmicutes	**0.0000**	**0.0000**
	Actinobacteria	**0.0000**	**0.0000**
	Euryarchaeota	**0.0032**	**0.0029**
GC-content	≤30%	**0.0000**	**0.0000**
	>30–≤50%	**0.0014**	**0.0000**
	>50–≤70%	**0.0000**	**0.0013**
Habitat	Aquatic	0.5957	0.3762
	Terrestrial	**0.0000**	**0.0005**
	Multiple	**0.0000**	**0.0057**
	Host-associated	**0.0000**	**0.0386**
	Specialized	**0.0001**	**0.0006**
Temperature range	Hyperthermophilic	**0.0000**	**0.0000**
	Thermophilic	**0.0001**	**0.0000**
	Mesophilic	0.8850	0.6349
Oxygen requirement	Aerobic	**0.0154**	0.1150
	Anaerobic	**0.0030**	**0.0011**
	Facultative	**0.0000**	**0.0000**

Note.—While for the CPCC, the alternative hypothesis was that the group-specific metrics produce higher CPCC values than randomly learned metrics, for the QD, the alternative hypothesis was that the group-specific metrics result in lower quartet distances than the randomly learned metrics. Significant results (<0.05) are shown in bold.

performance improvement (supplementary table S4, Supplementary Material online).

Trends Across Groups

We investigated whether the genomic and taxonomic composition of the groups are relevant for the improvement obtained by the specific metrics over the baseline. The aim of this analysis was to get a better understanding of when application of the proposed method might be most relevant. We calculated nine statistics for the groups (number of genomes, number of species, mean genome size, standard deviation (SD) of genome sizes, mean GC-content, SD of GC-content, NRI, and NTI) and correlated them with the change in the mean cophenetic correlation of the specific metrics relative to the baseline (table 3; supplementary table S2, Supplementary Material online) across the groups. A positive correlation here means that an increase in the statistic corresponds to an improvement in the CPCC on average and vice versa. The Actinobacteria and Euryarchaeota groups were removed from this analysis because they behaved like an outlier with respect to change in the CPCC, above the 99th percentile and below the 1st percentile, respectively.

The strongest and significant negative correlation (Pearson's $R = -0.54$, $P = 0.03$) was with the phylogenetic community measure NRI (Webb et al. 2002). NRI measures the phylogenetic clustering of taxa and becomes negative

Table 2

The CPCC and Quartet Distance before (CPCC and QD) and after (CPCC_PCA and QD_PCA) PCA based on the l4n1 Signature

Attribute	Group	CPCC	CPCC_PCA	QD	QD_PCA	Dimension	Variance (%)
Phylum	Proteobacteria	0.42	0.43	0.45	0.43	21	94.46
	Firmicutes	0.57	0.54	0.32	0.29	20	96.25
	Actinobacteria	0.39	0.44	0.55	0.50	19	96.32
	Euryarchaeota	0.46	0.45	0.47	0.43	20	97.20
GC-content	≤30%	0.30	0.34	0.43	0.40	19	96.73
	>30–≤50%	0.36	0.34	0.51	0.51	25	94.27
	>50–≤70%	0.44	0.48	0.48	0.43	22	94.49
Habitat	Aquatic	0.39	0.38	0.51	0.51	24	94.78
	Terrestrial	0.39	0.45	0.39	0.38	18	96.43
	Multiple	0.37	0.36	0.46	0.45	21	95.17
	Host-associated	0.17	0.18	0.51	0.51	21	94.65
	Specialized	0.20	0.19	0.57	0.57	23	95.28
Temperature range	Hyperthermophilic	0.46	0.41	0.43	0.46	18	97.93
	Thermophilic	0.19	0.24	0.59	0.58	22	96.03
	Mesophilic	0.25	0.24	0.51	0.52	22	93.49
Oxygen requirement	Aerobic	0.34	0.34	0.56	0.56	22	94.65
	Anaerobic	0.19	0.20	0.58	0.55	24	94.71
	Facultative	0.46	0.47	0.30	0.35	23	95.32
Average		0.35	0.36	0.48	0.47	21.33	95.45

Note.—The dimension and variance columns show the number of dimensions and variance retained, respectively. No significant improvement was observed after applying PCA either for the CPCC or the QD ($P > 0.3$, one-sided Wilcoxon rank sum test, see text for details).

Table 3

Correlation of the Mean Change in the CPCC with Different Statistics across the Groups

Correlation	Value	No. of Genomes	No. of Species	Genome Size (Mean)	Genome Size (SD)	GC-Content (Mean)	GC-Content (SD)	NRI	NTI
Pearson's	R	**−0.54**	−0.17	−0.34	−0.33	0.03	0.02	**−0.54**	−0.35
	P value	**0.03**	0.52	0.19	0.22	0.92	0.95	**0.03**	0.19
Spearman's	ρ	−0.46	−0.13	−0.44	−0.44	0.06	0.03	−0.40	−0.26
	P value	0.07	0.63	0.09	0.09	0.81	0.93	0.12	0.32

Note.—The Actinobacteria and Euryarchaeota groups were removed for this analysis, as they behaved like outliers (see text for details). Significant results ($P < 0.05$) are shown in bold.

with their increasing dispersion; therefore this negative correlation suggests that as the taxa become more clustered on the taxonomy, the specific metrics provide less improvement. This result was expected, as for closely related taxa the baseline (the l4n1 signature with the Euclidean distance) is expected to perform well (Mrazek 2009). A lower, but also negative correlation was observed for the nearest taxa index (NTI) which increases as taxa cluster at the terminal nodes (Webb et al. 2002).

The overall number of genomes in a group also showed a significant negative correlation with the mean change of the cophenetic correlation, suggesting that our method provides a larger improvement in the CPCC for smaller groups For the negative correlation with genome sizes we speculate that larger genomes may exhibit a noisier genome signature, for example due to presence of phages and plasmids (Suzuki et al. 2010), the specific metrics might provide an improvement by learning appropriate weights for oligonucleotides, such that the noise is reduced.

Interestingly, no significant correlation was observed with either the mean or the SD of the GC-content for each group ($P > 0.8$), suggesting that the improvement provided by the specific metric does not depend on the group GC-content, except for the Actinobacteria.

Group-Specific Metrics Generalized across Larger Taxonomic Distances

To investigate the effect of the genome relatedness on learning group-specific metrics, we removed genomes of the same species and order as the ones used for learning independently for each group-specific metric and recomputed the performance measures. These experiments were performed on the 1,951 genomes obtained from NCBI GenBank in June 2012. We observed similar trends as before (supplementary figs. S2–S6, Supplementary Material online), suggesting that metric learning is advantageous even when closely related genomes are not available. However, in many cases the performance of all tested methods degraded after this removal, indicating that signature-based methods perform better at lower taxonomic distances.

Discussion

In this work, we proposed a method to learn taxonomic distance metrics from genome signatures and the corresponding phenetic distances between them. Our aim was to improve genome signature-based genome tree inference for groups of genomes where the groups were defined by phylogenetic, genomic, or ecological attributes. Our empirical analyses showed that genome trees inferred from genome signature can be improved by learning group-specific taxonomic distance metrics. As expected, metrics learned for different phyla and GC-content groups showed significant improvement in the quality of inferred genome trees (for three groups out of four and two groups out of three, respectively). Working with the hypothesis that environmental selective forces shape the nucleotide composition of genomes, that is, that different niches drive the oligonucleotide composition in different directions, we learned specific metrics for different ecological groups. The ecological group-specific metrics showed performance improvements for 8 out of 11 ecological groups.

The performance improvement shown by specific metrics for phylogenetic and GC-content groups of species was relatively higher and generalized better for distant genomes than for the ecological groups. Nevertheless, also for the ecological groups, the learned metrics in most cases showed a performance improvement. The ecological groups in particular contain genomes of species only distantly related to each other, where the alignment-free methods are known to be less accurate. Of the other alignment-free methods evaluated here only CVTree showed a consistent improvement over the baseline. The better performance of CVTree compared with the l6n1 signature might be due to a more appropriate normalization.

An important property, in our opinion, of the CompSpec metric is that it only covers a subset of the whole compositional space. For instance, the employed parameters account for 9,200 ($200 \times [1 + {}^{10}C_2]$) words out of 1,048,576 (4^{10}) possible words amounting less than 1%. We speculate that the information loss due to this low coverage is, at least partly, responsible for lower performance we observed with the CompSpec distances. Although multiple samples of 200 words are used to build a number of trees, which are then aggregated into a final tree using a consensus method (Kirzhner et al. 2007), it is not straightforward to compare the resulting distances and resulting trees in this way. Therefore, we here used a single sample of 200 words was used in this study. For the FFP metric, we also computed distances between randomly sampled 50 kb continuous segments from the genomes, to investigate whether different genome sizes might be confounding the distance calculation. The results were similar (data not shown). We did not implement the block-FFP and optimal range finding algorithms (Sims et al. 2009) and it will be interesting to see whether those may lead to performance improvement. Furthermore, our experiments showed that dimensionality reduction with PCA resulted only in a marginal or no performance improvement.

Another observation from our analysis was that the BLAST alignment-based genome dissimilarity metric (GBDP) was the overall best performing method, both in terms of the cophenetic correlation and quartet distance. The good performance of GBDP implies that the information necessary for tree inference can be uncovered using genome-wide alignments. The comparatively lower performance of the alignment-free methods suggest that the distances calculated from the genome signatures do not represent universal taxonomic relationships with the same accuracy. The good performance of GBDP might also partly be due to the use of an evolutionary model in sequence alignments. At the same time, the lower performance of alignment-free methods might result from a loss of information, when encoding a longer sequence by means of shorter oligonucleotides. Further research is needed to pin point the advantages and shortcomings of the different methods.

However, performing alignments is computationally expensive and hence difficult to scale to a large number of genomes. The group-specific metrics we introduced can be learned from a small number of genomes, that is, 30 different species, and knowledge of the target phenetic distances between them in reference taxonomy. Therefore, to save computational cost, in case a resolved taxonomy for a group of genomes is not available, one could first infer a partial taxonomy from a subset of the genomes with an accurate method like GDBP and then use the taxonomy to learn a signature based distance metric that in turn could be applied to infer taxonomic distances for the remaining genomes.

In summary, our findings suggest that different groups of organisms have specific distance metrics over the genome signature and that these can be uncovered by considering their ecological, genomic or phylogenetic attributes. Our new method performed significantly better than the baseline technique for 13 out of 18 groups, indicating that group-specific aspects define the genome signature and that their consideration can improve the inference of taxonomic relationships. The existence of ecology-specific metrics strengthens the hypothesis that environmental factors affect the oligonucleotide usage of genomes. We also stress the need for more fine grained terms to describe specific environments and sample source information in public repositories, as provided by the environmental ontology (Hirschman et al. 2008). With the rapid advance in sequencing technologies large number of genome from microorganisms, even the ones not cultivable with traditional sequencing methods, will become available in the near future. Accurate and efficient methods are necessary to analyze this large-scale data. Our proposed method is a step towards this goal.

The analysis of the group-specific oligonucleotide weights and whether they provide insights into any evolutionary characteristics or adaptive evolution specific for the group will be an important future research direction. In this work the group-specific metrics were learned only from group-specific data, therefore the learned oligonucleotide weights do not necessarily contain discriminatory information. Furthermore, the limited number of genomes (30) used for metric learning and correlations between the oligonucleotide frequencies can lead to divergent metrics for a group, where weights can be distributed across different correlated oligonucleotides to obtain the same result. This prevented the interpretation of a biological or evolutionary meaning of the learned weights with the method described here.

The current work was confined to learning linear distance metrics. This can be extended to learning nonlinear distance metrics in the future, which may lead to further performance improvements. It will be also interesting to investigate whether learning a full matrix instead of a diagonal matrix for weighting oligonucleotides would be beneficial. We here used the cophenetic correlation with Spearman's rank correlation coefficient as a proxy objective function for tree similarity. Although the increase in the cophenetic correlation was correlated with the decrease in the quartet distance (all groups combined Pearson's $R = 0.46$, $P < 2.2\mathrm{e}{-16}$), further research is necessary to identify other more suitable optimality criteria. Furthermore, distance metric learning has the potential to be extended to unsupervised binning of metagenome data (McHardy and Rigoutsos 2007) to improve performance on a particular ecological niche, for example, the marine environment and the human gut.

Supplementary Material

Supplementary text, tables S1–S4, and figures S1–S6 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Lars Steinbrück (MPI Informatik, HHU) for critical reading of the manuscript and useful comments. K.R.P. thanks Ingolf Sommer, Nico Pfeifer (MPI Informatik), and Alfonso Caiazzo (Weistrass Institute, Berlin) for insightful discussions. This work was supported by the Max-Planck society funding to K.R.P. and A.C.M. and Heinrich-Heine University to A.C.M. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Blaisdell BE. 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci U S A. 83: 5155–5159.

Burge C, Campbell AM, Karlin S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci U S A. 89: 1358–1362.

Ciccarelli FD, et al. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287.

Coenye T, Gevers D, Van de Peer Y, Vandamme P, Swings J. 2005. Towards a prokaryotic genomic taxonomy. FEMS Microbiol Rev. 29: 147–167.

Coenye T, Vandamme P. 2003. Extracting phylogenetic information from whole-genome sequencing projects: the lactic acid bacteria as a test case. Microbiology 149:3507–3517.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6:361–375.

DeSantis TZ, et al. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 72:5069–5072.

Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B. 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol Biol Evol. 16:1391–1399.

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

Farris JS. 1969. On the cophenetic correlation coefficient. Syst Zool. 18: 279–285.

Gao L, Qi J, Sun J, Hao B. 2007. Prokaryote phylogeny meets taxonomy: an exhaustive comparison of composition vector trees with systematic bacteriology. Sci China C Life Sci. 50:587–599.

Garrity G. 2005. Bergey's manual of systematic bacteriology. The proteobacteria. Introductory essays, Part 1, Vol. 2. Berlin (Germany): Springer.

Hansen N, Muller SD, Koumoutsakos P. 2003. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). Evol Comput. 11:1–18.

Hao B, Qi J. 2003. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. Proc 2003 IEEE Bioinformatics Conf. 2:375–384.

Hasegawa M, Hashimoto T. 1993. Ribosomal RNA trees misleading. Nature 361:23.

Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC. 2005. Whole-genome prokaryotic phylogeny. Bioinformatics 21:2329–2335.

Hirschman L, et al. 2008. Habitat-Lite: A GSC case study based on free text terms for environmental metadata. OMICS 12:129–136.

Höhl M, Ragan MA. 2007. Is multiple-sequence alignment required for accurate inference of phylogeny? Syst Biol. 56:206–221.

Höhl M, Rigoutsos I, Ragan MA. 2006. Pattern-based phylogenetic distance estimation and tree reconstruction. Evol Bioinform Online. 2: 359–375.

Jain P, Kulis B, Davis J V., Dhillon IS. 2012. Metric and Kernel learning using a linear transformation. J Machine Learn Res. 13:519−547.

Jernigan RW, Baran RH. 2002. Pervasive properties of the genomic signature. BMC Genomics 3:23.

Johnson SC. 1967. Hierarchical clustering schemes. Psychometrika 32: 241–254.

Josse J, Kaiser AD, Kornberg A. 1961. Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. J Biol Chem. 236:864–875.

Karlin S, Burge C. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11:283–290.

Karlin S, Cardon LR. 1994. Computational DNA-sequence analysis. Annu Rev Microbiol. 48:619–654.

Karlin S, Mrazek J, Campbell AM. 1997. Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol. 179: 3899–3913.

Kirzhner V, Korol A, Bolshoy A, Nevo E. 2002. Compositional spectrum—revealing patterns for genomic sequence characterization and comparison. Physica A. 312:447–457.

Kirzhner V, Paz A, Volkovich Z, Nevo E, Korol A. 2007. Different clustering of genomes across life using the A-T-C-G and degenerate R-Y alphabets: early and late signaling on genome evolution? J Mol Evol. 64: 448–456.

Kuramae EE, Robert V, Echavarri-Erasun C, Boekhout T. 2007. Cophenetic correlation analysis as a strategy to select phylogenetically informative proteins: an example from the fungal kingdom. BMC Evol Biol. 7:134.

Lapointe FJ, Legendre P. 1992. Statistical significance of the matrix correlation-coefficient for comparing independent phylogenetic trees. Syst Biol. 41:378–384.

Li Q, Xu Z, Hao B. 2010. Composition vector approach to whole-genome-based prokaryotic phylogeny: success and foundations. J Biotechnol. 149:115–119.

McHardy AC, Rigoutsos I. 2007. What's in the mix: phylogenetic classification of metagenome sequence samples. Curr Opin Microbiol. 10: 499–503.

Mrazek J. 2009. Phylogenetic signals in DNA composition: limitations and prospects. Mol Biol Evol. 26:1163–1169.

Nielsen J, Kristensen AK, Mailund T, Pedersen CNS. 2011. A sub-cubic time algorithm for computing the quartet distance between two general trees. Algorithms Mol Biol. 6:15.

Patil KR, et al. 2011. Taxonomic metagenome sequence assignment with structured output models. Nat Methods. 8:191–192.

Pazos F, Valencia A. 2001. Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng. 14:609–614.

Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Res. 13:145–158.

Qi J, Wang B, Hao B. 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J Mol Evol. 58:1–11.

Reinert G, Chew D, Sun F, Waterman MS. 2009. Alignment-free sequence comparison (I): statistics and power. J Comput Biol. 16:1615–1634.

Sandberg R, et al. 2001. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. Genome Res. 11: 1404–1409.

Sayers EW, et al. 2009. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 37:D5–15.

Schloss PD, et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 75:7537–7541.

Sims GE, Jun SR, Wu GA, Kim SH. 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Natl Acad Sci U S A. 106:2677–2682.

Snel B, Huynen MA, Dutilh BE. 2005. Genome trees and the nature of genome evolution. Annu Rev Microbiol. 59:191–209.

Sokal R, Rohlf J. 1962. The comparison of dendrograms by objective methods. Taxon 11:33–40.

Steiger JH. 1980. Tests for comparing elements of a correlation matrix. Psychol Bull. 87:245–251.

Suzuki H, Yano H, Brown CJ, Top EM. 2010. Predicting plasmid promiscuity based on genomic signature. J Bacteriol. 192:6045–6055.

Van Passel MW, Kuramae EE, Luyf AC, Bart A, Boekhout T. 2006. The reach of the genome signature in prokaryotes. BMC Evol Biol. 6:84.

Takahashi M, Kryukov K, Saitou N. 2009. Estimation of bacterial species phylogeny through oligonucleotide frequency distances. Genomics 93:525–533.

Vandamme P, et al. 1996. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev. 60:407–438.

Vinga S, Almeida J. 2003. Alignment-free sequence comparison-a review. Bioinformatics 19:513–523.

Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. 2002. Phylogenies and community ecology. Annu Rev Ecol Syst. 33:475–505.

Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A. 74:5088–5090.

Wu M, Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 9:R151.

Wu T-J, Burke JP, Davison DB. 1997. A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. Biometrics 53:1431–1439.

Wu TJ, Huang YH, Li LA. 2005. Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. Bioinformatics 21:4125–4132.

Xing E, Ng A, Jordan M, Russell S. 2002. Distance metric learning, with application to clustering with side-information. Adv Neural Info Process Syst. 15:505–512.

Xu Z, Hao B. 2009. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res. 37:W174–W178.

Yang K, Zhang LQ. 2008. Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction. Nucleic Acids Res. 36:e33.

Associate editor: José Pereira-Leal.