Phylogenetic Signals in DNA Composition: Limitations and Prospects

Jan Mrázek

Department of Microbiology and Institute of Bioinformatics, University of Georgia

The concept of genome signature allows sequence comparisons without alignment. It relies on the premise that oligonucleotide compositions of DNA segments from the same or closely related genomes tend to be more similar than those from distantly related genomes. This concept has been used in detection of lateral gene transfer, phylogenetic classification of metagenome sequences (binning), and in studies of evolution of viruses and plasmids. The goal of this work is to explore limitations of genome signature in phylogenetic classification of DNA sequences and to identify formal representations of genome signature that expose best the phylogenetic relationships among prokaryotes. We found that genome signatures that best represent phylogenetic relationships are those normalized to factor out differences in G + C content and utilizing the standard A-C-G-T alphabet or the degenerate R-Y (purine–pyrimidine) alphabet. The main limitation of all genome signature representations tested is lack of divergence among some distantly related species. "Crowding" of the genome signature space and absence of molecular clock likely contribute to this phenomenon. We introduce "periodicity signatures"—formal representations of periodic sequence patterns related to DNA curvature—which can discriminate between bacterial and archaeal DNA sequences. Interestingly, archaea of the order Halobacteriaceae have periodic signatures similar to bacteria, possibly due to their early divergence from other archaea, extensive lateral gene transfer, or due to their adaptation to high salt environments. Our results have practical implications for development and application of genome signature–based methods for analysis and classification of DNA sequences.

Introduction

Comparing oligonucleotide compositions of DNA sequences was proposed as an alternative method for sequence comparison, which does not require sequence alignment (Blaisdell 1986; Brendel et al. 1986; Pietrokovski et al. 1990). The concept of "genome signature" (Karlin and Burge 1995) refers to a relative intragenomic invariance of oligonucleotide composition—typically represented as arrays of oligonucleotide frequencies, often normalized to factor out frequencies of the embedded shorter oligonucleotides. The exact mechanisms that generate and maintain the genome signature remain unknown but most likely involve differences in species-specific properties of replication and repair machineries (Karlin and Burge 1995; Karlin et al. 1997). A recent analysis determined that DNA repair enzymes in proteobacteria coevolve with the genome signature. By contrast, metabolic enzymes showed only a weak relationship to genome signature (Paz et al. 2006). A recent study also indicated that oligonucleotide compositions of genomes can be affected by environmental factors, namely, temperature and presence of oxygen (Kirzhner et al. 2007).

The concept of genome signature was recently applied in detection of lateral gene transfer events (Karlin 2001; Merkl 2004; Dufraigne et al. 2005; Tsirigos and Rigoutsos 2005), analyses of plasmid, phage, and viral genomes (Blaisdell et al. 1996; Campbell et al. 1999; Robins et al. 2005; Pride et al. 2006; Mrázek and Karlin 2007; Suzuki et al. 2008), and in phylogenetic classification (binning) of metagenome sequences (Teeling et al. 2004; Garcia Martin et al. 2006; Mavromatis et al. 2007; McHardy and Rigoutsos 2007; Warnecke et al. 2007; Chan et al. 2008; Zhou et al. 2008). Most of these methods use genome signatures derived from di- or tetranucleotide frequencies in combination with various metrics, clustering algorithms, or supervised machine learning methods. Some methods employ signatures derived from longer oligonucleotides binned into groups to reduce the dimensionality of the representation (Kirzhner et al. 2003, 2007) or specific words of variable length that are characteristically underrepresented or overrepresented in different genomes (Robins et al. 2005). All genome signature applications are based on the observation that genome signature tends to be relatively homogeneous within a genome, and most applications also imply that closely related genomes have more similar signature than distantly related ones (Pietrokovski et al. 1990; Karlin and Cardon 1994; Karlin et al. 1998). In this work, we investigate how different forms of genome signature satisfy the latter assumption by comparing the signature-based distances among prokaryotic genomes with distances derived from the alignment of their 16S rRNA sequences, and we identify the signatures that carry the strongest phylogenetic signal. We also introduce "periodicity signatures" based on periodic signals in nucleotide sequences related to DNA curvature (Herzel et al. 1999).

Key words: genome signature, periodicity signature, DNA curvature, metagenome binning, lateral gene transfer, halophilic archaea.

E-mail: mrazek@uga.edu.

Mol. Biol. Evol. 26(5):1163–1169. 2009
doi:10.1093/molbev/msp032
Advance Access publication February 20, 2009

Methods

DNA Sequences

Complete genomic DNA sequences of 130 prokaryotes (both archaea and bacteria) were downloaded from the NCBI ftp server (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). The 130 genomes were selected to cover all diverse taxa for which complete genomes were available, and all ranges of phylogenetic relatedness. Generally, no more than one species of the same genus was included. Exceptions were made for *Vibrio*, *Bacillus*, and *Mycoplasma*, where multiple species were included in the data set. The complete list of the 130 genomes is provided in the supplementary table S1, Supplementary Material online.

Standard Phylogenetic Distances

The standard phylogenetic distances to which genome signature distances are compared were derived from 16S rRNA sequences. The aligned 16S rRNA sequences for

the 130 species were downloaded from the GreenGenes database (http://greengenes.lbl.gov/) (DeSantis et al. 2006), and pairwise distances were calculated using the DNADIST program of the PHYLIP package (Felsenstein 1989) implemented at the GreenGenes web server (http://greengenes.lbl.gov/cgi-bin/nph-distance_matrix.cgi).

Genome Signatures Derived from Oligonucleotide Frequencies

The term genome signature refers to a vector representation of oligonucleotide frequencies or some indices derived from oligonucleotide frequencies of a DNA sequence. In this work, we used arrays of k-mer frequencies $\{\rho_i^{(0)}=f_{a_1\ldots a_k}\}$ for $2 \le k \le 6$ (a_j stands for A, C, G, or T and $f_{a_1\ldots a_k}$ is the frequency of the k-mer $a_1 \ldots a_k$), k-mer frequencies normalized to factor out the embedded mononucleotide frequencies $\{\rho_i^{(1)}=f_{a_1\ldots a_k}/f_{a_1}\ldots f_{a_k}\}$, k-mer frequencies normalized to factor out the embedded dinucleotide frequencies using the first-order Markov model $\{\rho_i^{(2)}=f_{a_1\ldots a_k}f_{a_2}f_{a_3}\ldots f_{a_{k-1}}/(f_{a_1a_2}f_{a_2a_3}\ldots f_{a_{k-1}a_k})\}$, and k-mer frequencies normalized to factor out the embedded $k-1$-mers using the maximum order Markov model $\{\rho_i^{(\max)}=f_{a_1\ldots a_k}f_{a_2\ldots a_{k-1}}/(f_{a_1\ldots a_{k-1}}f_{a_2\ldots a_k})\}$. All oligonucleotide frequencies are symmetrized with respect to the two DNA strands by averaging the frequencies for complementary oligonucleotides. When using the degenerate two-letter alphabets R-Y (purine vs. pyrimidine, or [A,G] vs. [C,T]), S-W ("strong" vs. "weak," or [G,C] vs. [A,T]), and M-K ([A,C] vs. [G,T]), the genome signatures were evaluated for up to 10-mers. We refer to different forms of genome signature by a code "<signature>" in the form of the letter l followed by the oligonucleotide length, the letter n followed by the length of the oligonucleotides used in the normalization, and the letters "ry," "sw," or "mk" to specify the alphabet if other than A-C-G-T.

We define a distance $\delta(A,B)$ between two sequences A and B as average absolute difference between values $\rho_i(A)$ and $\rho_i(B)$, that is,

$$\delta_{<\text{signature}>}(A,B) = \frac{1}{m}\sum_{i=1}^{m}|\rho_i(A) - \rho_i(B)|, \qquad (1)$$

where m is the number of variables in the genome signature representation. A distance D between two genomes G_1 and G_2 is defined as average distance between all pairs of nonoverlapping sequence samples A_i and B_j of a given length L from genomes G_1 and G_2, respectively:

$$D_{<\text{signature}>}(G_1,G_2) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\delta_{<\text{signature}>}(A_i,B_j). \qquad (2)$$

n_1 and n_2 are numbers of nonoverlapping sequence samples from genomes G_1 and G_2, respectively.

Periodicity Signatures

DNA sequences generally contain two strong periodic signals. The dominant signal has a period 3 bp and relates to biased codon and amino acid usages in protein-coding genes. The second significant periodic signal has a period close to 10.5 bp (the average length of one helical turn of DNA in the canonical B conformation) and relates to DNA curvature and/or bendability (Sinden 1994). This periodic signal is most pronounced in the distribution of short runs of A or T (Trifonov 1985; Tolstorukov et al. 2005; Mrázek 2006). Interestingly, the predominant period tends to be close to 11 bp in bacterial genomes but near 10 bp in archaeal genomes, possibly indicating different DNA supercoiling propensities in bacterial and archaeal chromosomes (Herzel et al. 1998; Schieg and Herzel 2004). We devised a heuristic measure of distance between DNA sequences based on analysis of sequence periodicity. We refer to these representations as periodicity signatures.

Following is the algorithm for determining the periodicity signature of a DNA sequence: First, the sequence at hand is converted into a histogram of spacings between the A/T nucleotides or short runs of A or T, and the dominant 3-bp period is removed with a 3-bp sliding window average (fig. 1*a*). In order to enhance the periodic signal, a parabolic least-square regression is subtracted from the histogram (fig. 1*b*). Finally, this modified histogram is converted into a power spectrum using the Fourier transform (fig. 1*c*). Several values of the power spectrum about the period 10.5 bp (e.g., the 21 values corresponding to periods between 9.5 and 11.5 bp taken at the 0.1-bp interval) are selected to represent the periodicity signature of the analyzed sequence. We refer to these values as the array $\{\tau_i^{<\text{signature}>}\}$, where <signature> signifies the parameters of the algorithm described above, including the letter l followed by the length of the A/T runs (between 1 and 4 bp), the letter h followed by the section of the histogram in figure 1*b*, which is converted into the power spectrum (ranges 11–60, 31–130, and 31–200 bp were tested), and the letter r followed by the section of the power spectrum (fig. 1*c*) centered at 10.5 bp, which is covered by the values τ_i at 0.1-bp intervals.

The distance s between two sequences A and B is defined as

$$s_{<\text{signature}>}(A,B) = 1 - r(\{\tau_i(A)\},\{\tau_i(B)\}), \qquad (3)$$

where $r(\{\tau_i(A)\},\{\tau_i(B)\})$ is the standard Pearson correlation coefficient between the arrays $\{\tau_i(A)\}$ and $\{\tau_i(B)\}$ derived from sequences A and B, respectively. The distance between two genomes G_1 and G_2 is defined as the average distance between all pairs of nonoverlapping sequence samples, analogous to Formula (2):

$$D_{<\text{signaure}>}(G_1,G_2) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}s_{<\text{signaure}>}(A_i,B_j). \qquad (4)$$

The algorithm for determining the periodicity signature and the metric for genome comparisons are the results of extensive experimentation with alternative forms, including autocorrelation function, representations derived directly from the spacing histogram (fig. 1*a* and *b*), and Euclidian and Manhattan distances (data not shown). These alternative definitions of periodic signature provided poorer distinction between bacteria and archaea.

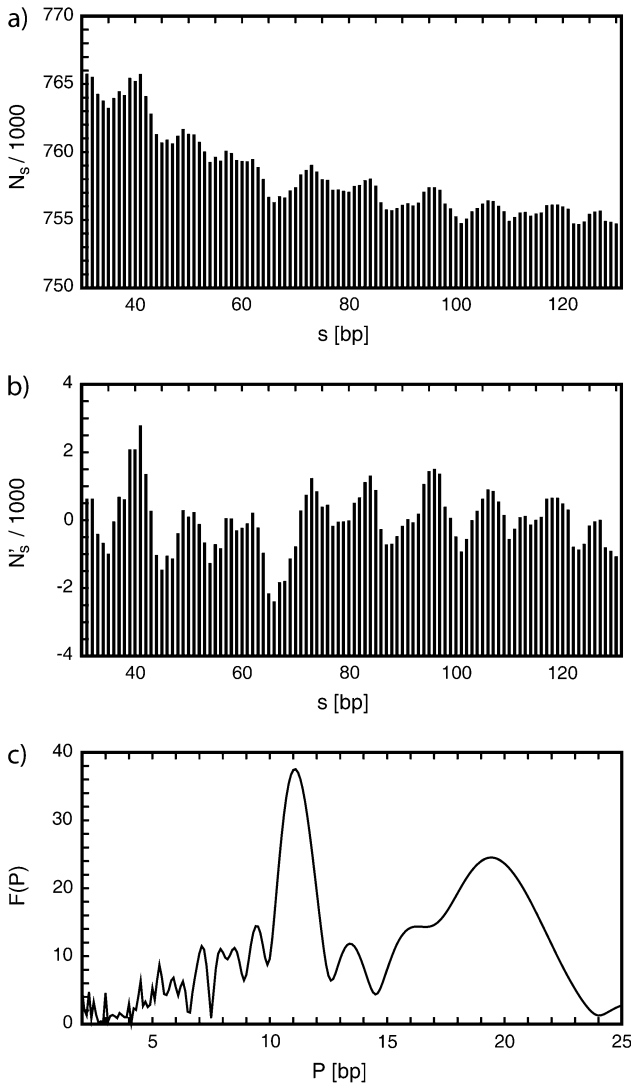

FIG. 1.—(*a*) Histogram of spacings between A/T nucleotides in chromosome 1 of *Burkholderia pseudomallei* K96243. The abscissa shows a distance *s* between a pair of A/T nucleotides in the DNA sequence and the ordinate shows how many times A/T nucleotides are found at this particular distance from each other. (*b*) The same histogram after subtracting a parabolic least-square regression. (*c*) Power spectrum generated by Fourier transform from the normalized histogram. The ordinate displays a measure of strength of a periodic signal corresponding to the period *P* shown by the abscissa. See Methods for details.

Assessing Performance of Genome Signatures

For the purpose of this work, we define "performance" of a genome signature strictly as its ability to reproduce the "standard" evolutionary distances between species (16S rRNA distances). The complete set of 130 genomes used in this study embodies 8,385 pairwise comparisons. The agreement between the 16S rRNA distance and a signature-based distance is measured in terms of a correlation coefficient across all 8,385 pairwise comparisons. We use the standard Pearson correlation coefficient as well as the nonparametric Kendall–tau rank correlation coefficient (Noether 1981).

Results

Performance of Oligonucleotide Genome Signatures with 50-kb Sequence Samples

Table 1 shows Pearson correlation coefficients between the 16S rRNA distances and the best-performing oligonucleotide genome signature distances for the complete data set of 130 genomes and several smaller, less diverse data sets. Complete data are shown in supplementary table S2, Supplementary Material online. The best overall performers include signatures using the standard A-C-G-T alphabet and the degenerate R-Y alphabet. Signatures based on S-W and M-K alphabets performed poorly, although some S-W and M-K signatures performed well on data sets restricted to archaea and to proteobacteria (supplemental table S2, Supplementary Material online). Graphical representations for two of the best-performing signatures (fig. 2 and supplementary figure S3a, Supplementary Material online) show that closely related genomes (those characterized by low 16S rRNA distances at the left side of the plot) tend to have similar signatures, but distant organisms may have both similar and dissimilar signatures. This principal limitation applies to all best-performing genome signatures although some signatures are less affected than others.

We used the Unweighted Pair Group Method with Arithmetic Mean algorithm to hierarchically cluster the genome signatures according to the similarity of their results on the 130-species data set (supplementary figure S3b, Supplementary Material online). Signatures using direct oligonucleotide frequencies $\{\rho_i^{(0)}\}$ in both A-C-G-T and S-W alphabets produced very similar results, indicating that direct oligonucleotide frequencies are dominated by differences in G + C content. These signatures performed worse than signatures normalized to factor out differences in mononucleotide composition (table 1 and supplementary table S2, Supplementary Material online). Interestingly, the best-performing signatures using the A-C-G-T alphabet were of the type $\{\rho_i^{(1)}\}$ (i.e., factoring out mononucleotide frequencies), whereas for the R-Y alphabet the best performers were of the type $\{\rho_i^{(\max)}\}$ (i.e., factoring out frequencies of k–1-mers). Dinucleotide relative abundances (Karlin and Burge 1995) perform about equally well or better than high-dimensional signature representations derived from frequencies of longer oligonucleotides (table 1), suggesting that the biased nearest-neighbor associations carry most of the relevant information.

Periodicity Signatures

The concept of periodicity signature is motivated by the work of Herzel and coworkers, who noted differences between bacterial and archaeal genomes in predominant periodic sequence patterns, presumably related to different DNA supercoiling propensities (Herzel et al. 1998, 1999). We used periodicity signatures to help differentiate between archaeal and bacterial DNA sequences, an area where genome signatures perform poorly (fig. 2 and supplementary figure S3a, Supplementary Material online). Performance of periodicity signatures with several different sets of parameters is summarized in supplementary table S4, Supplementary Material online. The signature 11h31–130r2

Table 1
Best-Performing Genome Signatures with 50-kb Sequence Samples

Signature Code	Complete Data Set	Bacteria	Archaea	Proteo	γ-Proteo	Nonproteo	Meso	Thermo
l2n1	0.35	0.28	0.39	0.41	0.38	0.23	0.29	0.18
l3n1	0.36	0.34	0.43	**0.47**	0.36	0.31	0.29	0.24
l3n2	0.21	0.29	**0.49**	0.47	0.19	0.28	0.16	**0.27**
l4n1	0.34	0.32	0.47	0.47	0.35	0.31	0.28	0.25
l4n2	0.18	0.27	**0.49**	**0.47**	0.22	0.27	0.14	0.24
l5n1	0.33	0.32	0.48	0.46	0.32	0.32	0.26	0.26
l6n1	0.31	0.33	**0.48**	0.43	0.29	0.34	0.25	0.26
l4n2ry	0.20	0.26	0.39	0.38	0.14	0.20	0.17	**0.28**
l5n4ry	0.14	0.21	0.05	**0.48**	0.34	0.06	0.08	0.01
l6n5ry	0.30	0.32	0.31	0.30	**0.49**	0.27	**0.40**	0.16
l7n6ry	**0.42**	0.35	0.16	0.36	0.18	**0.35**	**0.44**	**0.27**
l8n1ry	0.22	**0.36**	0.29	0.38	0.25	0.29	0.22	0.17
l8n7ry	0.34	0.35	0.24	0.33	**0.43**	0.29	0.29	0.09
l9n8ry	**0.37**	**0.39**	0.05	0.27	0.11	**0.36**	0.26	0.11
l10n9ry	**0.44**	**0.41**	0.06	0.29	0.12	**0.37**	**0.36**	0.15
l6n5mk	0.11	0.22	0.08	0.30	**0.42**	0.23	0.20	0.05

The table shows correlations of pairwise distances derived from genome signature indicated by the code in the leftmost column with the 16S rRNA distances measured by the Pearson correlation coefficient. The correlations were evaluated in the complete data set of 130 genomes and less diverse subsets restricted to bacteria, archaea, proteobacteria (Proteo), γ-proteobacteria (γ-proteo), all bacteria excluding proteobacteria (Nonproteo), mesophilic bacteria and archaea (Meso), and thermophilic and hyperthermophilic bacteria and archaea (Thermo). Three highest values for each data set are shown in bold face.

(a measure of periodic spacing of A and T mononucleotides over the spacing range 31–130 bp and periodicity range 9.5–11.5 bp; see Methods for details) yielded the highest correlations with the 16S rRNA distances (Pearson correlation coefficient 0.28 and Kendall–tau 0.18). The spacing range 31–130 bp emphasizes signals related to DNA curvature: The 31 bp minimum distance eliminates most of the periodic signal from α-helices in proteins (Herzel et al. 1998), whereas the maximum 130 bp corresponds approximately to the length of bent DNA segments (Tolstorukov et al. 2005). The plot in figure 3 reveals that most pairwise genome comparisons show no significant similarity or dissimilarity (distances close to 1), which signifies no consistent positive or negative correlations between sequence samples taken from the two genomes. However, for those comparisons that yield distances significantly different from 1, most distances <1 involve comparisons between two bacterial or two archaeal genomes, whereas distances >1 generally involve comparisons between a bacterium and an archaeon. Interestingly, the largest deviations from this pattern involved comparisons of *Halobacterium* NRC-1 (an archaeon) with other genomes. Indeed, the power spectrum derived from A/T spacings in the *Halobacterium* genome shows a strong peak centered at period close to 11 bp (supplementary figure S3c, Supplementary Material online), which is common in bacteria but uncharacteristic of archaeal genomes.

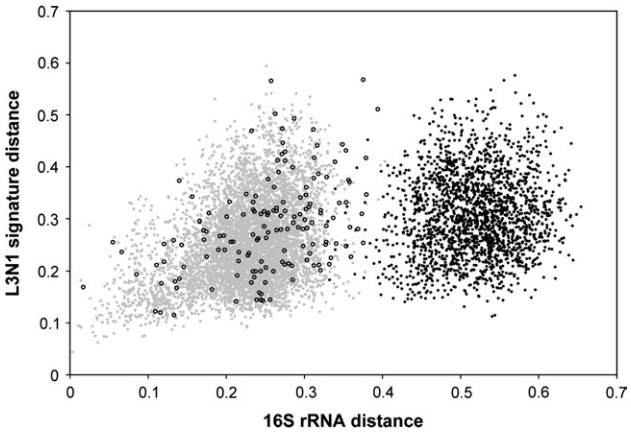

FIG. 2.—Comparison of the l3n1 genome signature distances (trinucleotide frequencies factoring out the embedded mononucleotide frequencies) and 16S rRNA distances among prokaryotic genomes. Each point represents a pair of genomes. Gray dots signify bacterium–bacterium distances, open circles archaeon–archaeon distances, and black dots bacterium–archaeon distances. Some sample 16S rRNA distances are given next to help readers appraise the scale of the horizontal axis: *Anabaena variabilis* versus *Nostoc* PCC7120, 0.003 (the smallest 16S rRNA distance in the data set); *Escherichia coli* versus *Salmonella typhimurium*, 0.014; *E. coli* versus *Haemophilus influenzae*, 0.109; *E. coli* versus *Xanthomonas campestris*, 0.156; *E. coli* versus *Sinorhizobium meliloti*, 0.195; *E. coli* versus *Bacillus subtilis*, 0.230; *E. coli* versus *Aquifex aeolicus*, 0.403; *E. coli* versus *Methanococcus maripaludis*, 0.508; *Ureaplasma parvum* versus *Nanoarchaeum equitans*, 0.654 (the largest 16S rRNA distance in the data set).

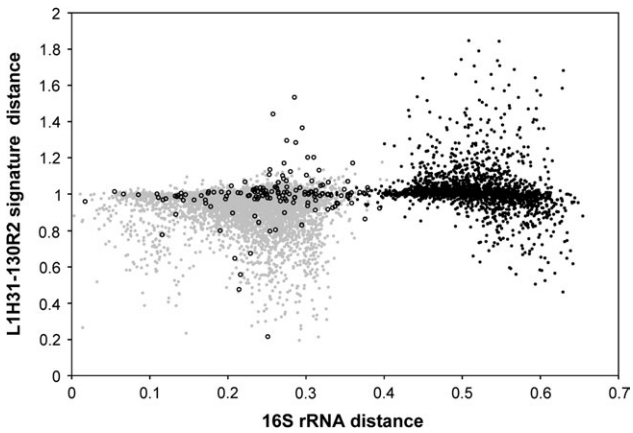

FIG. 3.—Comparison of the l11h31–130r2 periodic signature distances (see Methods for details) and 16S rRNA distances among prokaryotic genomes. See legend to figure 2.

The performance of periodicity signatures deteriorates more significantly with decreasing sequence sample sizes than that of signatures derived from oligonucleotide frequencies (supplementary table S6, Supplementary Material online). This is expected considering the character of the periodic signal in prokaryotic DNA sequences. In *E. coli* and probably most other prokaryotic genomes, strong periodic signals are generally restricted to short (~100 to 150 bp) segments of intrinsically curved DNA distributed throughout the chromosome (Tolstorukov et al. 2005). Hence, only the sequence samples containing the curved segments will exhibit a strong periodic signal.

We attempted to further improve the correlations between the signature-based distances and the 16S rRNA distances by designing several ad hoc combinations of different types of best-performing signatures. In particular, combining the best-performing genome signatures with periodicity signatures could provide a better resolution over the complete range of phylogenetic distances. When combining different types of oligonucleotide genome signatures, we define the combined distance as an arithmetic average of the individual distances, for example, $D_{l10n9ry \cdot l9n8ry}(G_1, G_2) = \frac{1}{2}[D_{l10n9ry}(G_1, G_2) + D_{l9n8ry}(G_1, G_2)]$. Due to the character of distances derived from periodicity signatures, the combinations involving a periodicity signature are defined as a product, for example, $D_{l2n1 \cdot l1h31-130r2}(G_1, G_2) = D_{l2n1}(G_1, G_2) \cdot D_{l1h31-130r2}(G_1, G_2)$. Although not optimized for performance, some of these ad hoc combinations performed better than individual signatures (supplementary table S5, Supplementary Material online). The highest correlation (Pearson 0.49, Kendall-tau 0.34) was achieved by combining three R-Y alphabet oligonucleotide signatures l8n7ry, l7n6ry, and l6n5ry with the periodicity signature l1h31–130r2 (supplementary fig. S3d, Supplementary Material online).

Discussion

Implications for Development of Computational Tools Utilizing the Genome Signature Concept

The data presented here show that all forms of oligonucleotide genome signatures (using different oligonucleotide lengths, normalizations, and degenerate alphabets) have significant limitations when used as phylogenetic markers. The primary limitation is absence of divergence of oligonucleotide composition in some phylogenetically distant species (fig. 2 and supplementary fig. S3a, Supplementary Material online). This applies to all oligonucleotide signatures tested in this work as well as compositional spectra derived from subsets of longer oligonucleotides (10–20 bp) (Kirzhner et al. 2007). The lack of divergence could result from absence of molecular clock: As long as mutational biases and selective constraints remain in equilibrium oligonucleotide composition of a genome can remain approximately constant even as mutations accumulate. Similar genome signatures among some phylogenetically distant species can also arise form "coincidental convergence" (i.e., not selection driven) due to "crowding" of the genome signature space. This is perhaps expected for low-dimensional signature representations (e.g., derived from dinucleotide frequencies) that may not capture sufficient information to differentiate among many distinct phyla. It is less intuitive that the same phenomenon applies to high-dimensional representations. However, principal component analysis applied to some of the high-dimensional signature representations shows that they can be reduced to few dimensions with little loss of information (table 2), suggesting that even high-dimensional signature representations are dominated by a small number of major trends.

Table 2
Principal Component Analysis for Selected Genome Signatures

	% Variance Explained by Top *n* Principal Components		
n	l3n1 Signature	l4n1 Signature	l7n6ry Signature
1	38.5	34.2	46.1
2	57.6	51.7	62.8
3	67.2	60.5	73.8
4	73.7	66.3	79.4
5	78.0	77.9	83.4
7	85.7	73.9	88.8
10	91.8	84.8	93.2
15	96.3	90.7	97.4

The table shows the percentage of variance of genome signature among the 130 genomes "explained" by the number of principal components specified in the leftmost column. The percentage was calculated as the sum of *n* highest eigenvalues of the correlation matrix divided by the sum of all eigenvalues. These values can be interpreted as the amount of information that will be retained if a high-dimensional representation of the data is reduced to *n* variables.

Despite the above-stated limitations, genome signature has been used successfully in binning of metagenomic sequences (Teeling et al. 2004; Mavromatis et al. 2007; McHardy et al. 2007; McHardy and Rigoutsos 2007; Chan et al. 2008; Zhou et al. 2008). However, the accuracy of the binning methods deteriorates rapidly with an increasing phylogenetic diversity of the sample (see Mavromatis et al. 2007; Zhou et al. 2008 and the data at http://fames.jgi-psf.org/Binning_results.html), which is consistent with the limitations shown here. Moreover, different authors often utilize different formulas when evaluating the accuracy of binning methods, and the reported accuracy rates are not always mutually comparable. We summarize below several ways how binning methods and other applications of genome signature could be further improved in the light of our results.

1. Comparisons among different forms of genome signature show that normalization of oligonucleotide frequencies does matter. For example, many signature-based methods use direct tetranucleotide frequencies (the l4n0 signature), whereas the signature factoring out differences in G + C content (l4n1) correlates significantly better with phylogenetic distances.
2. The principal component analysis (table 2) indicates that selecting the right features to represent the genome signature can reduce the number of variables without significant loss of information, which in turn can improve accuracy. Mahalanobis distance provides a natural way of assigning weights when one sequence sample is compared with a collection of congruent

samples (e.g., from the same genome). This technique was recently applied in comparisons of plasmid and chromosomal sequences and provided better accuracy than alternative metrics in identifying the plasmid hosts (Suzuki et al. 2008). Standardized algorithms for feature selection are sometimes utilized with machine learning methods and allow selecting features that are most informative for the desired classification (Blum and Langley 1997).

3. Binning methods typically use a single signature representation, such as tetranucleotide frequencies. Our data indicate that different signature representations can provide complementary information (e.g., short oligonucleotides in A-C-G-T alphabet with first-order Markov normalization and longer oligonucleotides in R-Y alphabet and maximum order Markov normalization; see supplementary fig. S3b, Supplementary Material online). Combining relevant features of different representations could improve accuracy.
4. Comparisons across genome samples of varying taxonomical composition (table 1) suggest that different genome signatures may provide the best results depending on the composition of the sample. It is likely that performance of signature-based methods could be further improved by fine tuning the signature representation for a particular data set if its approximate species composition is known.
5. Combining features derived from oligonucleotide composition with nontraditional representations of DNA sequences such as the periodicity signatures can improve performance of signature-based techniques.

Bacterial-Like Periodicity Signatures in Halophilic Archaea

The observation that the *Halobacterium* NRC-1 genome exhibits sequence periodicity characteristic of bacteria is intriguing. In fact, the bacterial-like sequence periodicity was previously reported in archaea *Halobacterium* NRC-1 and *Methanopyrus kandleri* (Schieg and Herzel 2004). However, unlike *Halobacterium* NRC-1, the *M. kandleri* periodicity signature is not consistent throughout the genome (data not shown), and its comparisons with other genomes in figure 3 do not stand out as aberrations. We investigated sequence periodicity in three other halophilic archaea with complete genomes available, including *Halobacterium salinarium*, *Haloarcula marismortui*, and *Haloquadratum walsbyi*, all of the order Halobacteriales. *Halobacterium salinarium* and *H. marismortui* have dominant sequence periodicity of about 11.3 bp, similar to that of *Halobacterium* NRC-1 and typically seen in bacteria, whereas *H. walsbyi* features a weaker periodic signal with period about 10.6 bp that extends only over a short range of distances (up to about 60 bp) and could possibly be related to protein α-helices rather than DNA curvature (Herzel et al. 1999). These observations suggest that the bacterial-like DNA periodicity is a general feature of halophilic archaea. Members of the order Halobacteriales typically achieve osmotic balance with the extracellular environment by maintaining high intracellular concentrations of both KCl and NaCl (Kletzin 2007). High sodium concentrations and low hydration promote conformational changes that increase the helical period of the DNA, including transition into A-DNA (Malenkov et al. 1975; Brovchenko et al. 2008). These structural changes might stimulate an adaptation in the DNA sequence and emergence of bacterial-like periodic patterns. Alternatively, the bacterial-like DNA sequence periodicity in halophilic archaea could be due to extensive lateral DNA transfer between bacteria and archaea analogous to *T. maritima* (Worning et al. 2000). Indeed, the *Halobacterium* NRC-1 genome contains a number of genes of apparent bacterial origin (Ng et al. 2000) and whole-genome phylogenetic trees—but not trees based on 16S rRNA or transcription and translation proteins—place *Halobacterium* at the root of archaeal domain (Wolf et al. 2002; Brochier et al. 2004).

Supplementary Material

Supplementary tables S1, S2, S4, S5, and S6 and supplementary figures S3a–S3d are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

I wish to thank Dr. Barny Whitman for his comments on the manuscript and my colleagues in the Department of Microbiology for stimulating discussions in the course of this work.

Literature Cited

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

Blaisdell BE, Campbell AM, Karlin S. 1996. Similarities and dissimilarities of phage genomes. Proc Natl Acad Sci USA. 93:5854–5859.

Blum AL, Langley P. 1997. Selection of relevant features and examples in machine learning. Artif Intellig. 97:245–271.

Brendel V, Beckmann JS, Trifonov EN. 1986. Linguistics of nucleotide sequences: morphology and comparison of vocabularies. J Biomol Struct Dyn. 4:11–21.

Brochier C, Forterre P, Gribaldo S. 2004. Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the *Methanopyrus kandleri* paradox. Genome Biol. 5:R17.

Brovchenko I, Krukau A, Oleinikova A, Mazur AK. 2008. Ion dynamics and water percolation effects in DNA polymorphism. J Am Chem Soc. 130:121–131.

Campbell A, Mrázek J, Karlin S. 1999. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc Natl Acad Sci USA. 96:9184–9189.

Chan CK, Hsu AL, Tang SL, Halgamuge SK. 2008. Using growing self-organising maps to improve the binning process in environmental whole-genome shotgun sequencing. J Biomed Biotechnol. 2008:513701.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 72:5069–5072.

Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P. 2005. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res. 33:e6.
Felsenstein J. 1989. PHYLIP—phylogeny inference package (Version 3.2). Cladistics. 5:164–166.
Garcia Martin H, Ivanova N, Kunin V, et al. (18 co-authors). 2006. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nat Biotechnol. 24:1263–1269.
Herzel H, Weiss O, Trifonov EN. 1998. Sequence periodicity in complete genomes of archaea suggests positive supercoiling. J Biomol Struct Dyn. 16:341–345.
Herzel H, Weiss O, Trifonov EN. 1999. 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics. 15:187–193.
Karlin S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends Microbiol. 9:335–343.
Karlin S, Burge C. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11:283–290.
Karlin S, Campbell AM, Mrázek J. 1998. Comparative DNA analysis across diverse genomes. Annu Rev Genet. 32:185–225.
Karlin S, Cardon LR. 1994. Computational DNA sequence analysis. Annu Rev Microbiol. 48:619–654.
Karlin S, Mrázek J, Campbell AM. 1997. Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol. 179:3899–3913.
Kirzhner V, Nevo E, Korol A, Bolshoy A. 2003. A large-scale comparison of genomic sequences: one promising approach. Acta Biotheor. 51:73–89.
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.
Kletzin A. 2007. General characteristics and important model organisms. In: Cavicchioli R, editor. Archaea: molecular and cellular biology. Washington (DC): ASM Press. p. 14–92.
Malenkov G, Minchenkova L, Minyat E, Schyolkina A, Ivanov V. 1975. The nature of the B-A transition of DNA in solution. FEBS Lett. 51:38–42.
Mavromatis K, Ivanova N, Barry K, et al. (14 co-authors). 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods. 4:495–500.
McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. 2007. Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods. 4:63–72.
McHardy AC, Rigoutsos I. 2007. What's in the mix: phylogenetic classification of metagenome sequence samples. Curr Opin Microbiol. 10:499–503.
Merkl R. 2004. SIGI: score-based identification of genomic islands. BMC Bioinform. 5:22.
Mrázek J. 2006. Analysis of distribution indicates diverse functions of simple sequence repeats in *Mycoplasma* genomes. Mol Biol Evol. 23:1370–1385.
Mrázek J, Karlin S. 2007. Distinctive features of large complex virus genomes and proteomes. Proc Natl Acad Sci USA. 104:5127–5132.
Ng WV, Kennedy SP, Mahairas GG, et al. (42 co-authors). 2000. Genome sequence of *Halobacterium* species NRC-1. Proc Natl Acad Sci USA. 97:12176–12181.
Noether GE. 1981. Why Kendall tau? Teaching Stat. 3:41–43.
Paz A, Kirzhner V, Nevo E, Korol A. 2006. Coevolution of DNA-interacting proteins and genome "dialect". Mol Biol Evol. 23:56–64.
Pietrokovski S, Hirshon J, Trifonov EN. 1990. Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. J Biomol Struct Dyn. 7:1251–1268.
Pride DT, Wassenaar TM, Ghose C, Blaser MJ. 2006. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics. 7:8.
Robins H, Krasnitz M, Barak H, Levine AJ. 2005. A relative-entropy algorithm for genomic fingerprinting captures host-phage similarities. J Bacteriol. 187:8370–8374.
Schieg P, Herzel H. 2004. Periodicities of 10–11bp as indicators of the supercoiled state of genomic DNA. J Mol Biol. 343: 891–901.
Sinden RR. 1994. DNA structure and function. San Diego: Academic Press.
Suzuki H, Sota M, Brown CJ, Top EM. 2008. Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. Nucleic Acids Res. 36:e147.
Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO. 2004. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol. 6: 938–947.
Tolstorukov MY, Virnik KM, Adhya S, Zhurkin VB. 2005. A-tract clusters may facilitate DNA packaging in bacterial nucleoid. Nucleic Acids Res. 33:3907–3918.
Trifonov EN. 1985. Curved DNA. CRC Crit Rev Biochem. 19:89–106.
Tsirigos A, Rigoutsos I. 2005. A new computational method for the detection of horizontal gene transfer events. Nucleic Acids Res. 33:922–933.
Warnecke F, Luginbuhl P, Ivanova N, et al. (38 co-authors). 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature. 450: 560–565.
Wolf YI, Rogozin IB, Grishin NV, Koonin EV. 2002. Genome trees and the tree of life. Trends Genet. 18:472–479.
Worning P, Jensen LJ, Nelson KE, Brunak S, Ussery DW. 2000. Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritima*. Nucleic Acids Res. 28:706–709.
Zhou F, Olman V, Xu Y. 2008. Barcodes for genomes and applications. BMC Bioinform. 9:546.

Edward Holmes, Associate Editor

Accepted February 16, 2009