 doi:10.1111/j.1462-2920.2004.00624.x

Application of tetranucleotide frequencies for the assignment of genomic fragments

Hanno Teeling,[1] Anke Meyerdierks,[2] Margarete Bauer,[1] Rudolf Amann[2] and Frank Oliver Glöckner[1]*

[1]*Department of Molecular Ecology, Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany.*

[2]*Department of Molecular Ecology, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany.*

Summary

A basic problem of the metagenomic approach in microbial ecology is the assignment of genomic fragments to a certain species or taxonomic group, when suitable marker genes are absent. Currently, the (G + C)-content together with phylogenetic information and codon adaptation for functional genes is mostly used to assess the relationship of different fragments. These methods, however, can produce ambiguous results. In order to evaluate sequence-based methods for fragment identification, we extensively compared (G + C)-contents and tetranucleotide usage patterns of 9054 fosmid-sized genomic fragments generated *in silico* from 118 completely sequenced bacterial genomes (40 982 931 fragment pairs were compared in total). The results of this systematic study show that the discriminatory power of correlations of tetranucleotide-derived z-scores is by far superior to that of differences in (G + C)-content and provides reasonable assignment probabilities when applied to metagenome libraries of small diversity. Using six fully sequenced fosmid inserts from a metagenomic analysis of microbial consortia mediating the anaerobic oxidation of methane (AOM), we demonstrate that discrimination based on tetranucleotide-derived z-score correlations was consistent with corresponding data from 16S ribosomal RNA sequence analysis and allowed us to discriminate between fosmid inserts that were indistinguishable with respect to their (G + C)-contents.

Introduction

Until the advent of molecular techniques, studies on the diversity and function of microorganisms were restricted by the need to obtain pure cultures. Although attempts to cultivate bacteria have been conducted over decades, and thousands of isolates are available in culture collections, the vast majority of microorganisms has not yet been characterized. From cultivation-independent techniques it is currently estimated that in many ecosystems less than 1% of the microbial diversity has been seen on agar plates (Amann *et al*., 1995). This fact has dramatically shifted our understanding of microbial communities and their role and function within their natural habitats. To overcome these limitations, methods have been developed to explore the physiological potential of uncultivated microorganisms by extracting and analysing not only genes but large genomic fragments directly from the environment (Béjà *et al*., 2000a; Rondon *et al*., 2000; DeLong, 2002). In this so called metagenomic approach (Schloss and Handelsman, 2003), DNA is directly extracted from environmental samples and cloned into vectors such as cosmids, fosmids or bacterial artificial chromosomes (BACs). These metagenome clone libraries can be screened for inserts carrying specific functions or for sequences of known genes. Employing metagenomics, new enzymes and large genomic fragments of as yet uncultured bacteria have been successfully retrieved from the environment. One of the most prominent examples to date is the discovery of bacteriorhodopsin in marine *Gammaproteobacteria*, a protein that previously was believed to occur exclusively in archaeal *Halobacteria* (Béjà *et al*., 2000b; 2001).

Despite the proven potential of the metagenomic approach to broaden our knowledge about the composition and function of natural microbial communities, several methodological problems remain to be solved. One of the major challenges is the identification of the fragments' organismal origin. Assuming an average genome size of 4 Mb, only about 5–10% of the clones within a fosmid library (40 kb insert size) harbour phylogenetic marker genes like 16S rDNA, *rpoA*, *recA*, etc. and can therefore be assigned to a certain species or taxonomic group. The chance that new and interesting functional genes are located on a fragment that also carries a phylogenetic marker gene is even lower. Hence, there is a clear need for additional tools to provide evidence that, e.g. two fragments belong to the same organism. Identification of unknown fragments is trivial if they overlap for at least one or two kilobases. In this case, fragments can be fused and genome walking techniques can be applied to reconstruct

Received 23 September, 2003; accepted 18 February, 2004. *For correspondence. E-mail fog@mpi-bremen.de; Tel. (+49) 0421-2028938; Fax (+49) 0421 2028580.

as much of the metagenome as possible. However, this is extremely time-consuming, and if a connecting clone cannot be found, the whole procedure is stalled. The sequence-based measure commonly used to assess whether two unlinked fragments belong to the same organism, is their similarity in (G + C)-content. In addition, best BLAST hits as well as codon usage of the genes residing on the fragments can provide valuable hints about their organismal origin. These measures, however, can be misleading. The (G + C)-content of prokaryotes can vary considerably within genomes, and does not carry a strong phylogenetic signal. Regarding gene content, a typical fosmid insert of ~40 kb harbours about 40 genes, from which on average only half yield significant hits when searched against the public databases, e.g. by BLAST. Frequently, these hits do not come from the same phylogenetic taxon and thus provide no hints on the insert's organismal origin. For example, many protein families are phylogenetically unspecific and the phylogenetic significance of others is affected by lateral gene transfer (LGT). Moreover, in many cases databases simply do not contain close relatives to the species under investigation, which affects the metagenome approach in particular, because the discovery of new lineages and functions is one of its key intentions. The codon usage of the genes can also be blurred by LGT, and varies with gene expression level rather than carrying a phylogenetic signal (Karlin and Mrázek, 2000).

In contrast, a rather pronounced phylogenetic signal can be found in the tetranucleotide usage patterns of the inserts' nucleotide sequences (Pride *et al*., 2003). Frequencies of oligonucleotides in genomic sequences are known to carry species-specific signals (Karlin *et al*., 1994; 1998; Karlin and Burge, 1995; Karlin, 1998; Nakashima *et al*., 1998). Using the relative abundances of oligomers up to a length of four nucleotides, this has been shown for prokaryotes (Karlin *et al*., 1994) as well as for eukaryotes (Karlin and Ladunga, 1994; Gentles and Karlin, 2001). Species-specific signals for oligomers of different length have also been detected by means of neuronal networks (Abe *et al*., 2003), using chaos game representations (Goldman, 1993; Deschavanne *et al*., 1999) and naïve Bayesian classifiers (Sandberg *et al*., 2001). The cause for these signals has been attributed – at least for dinucleotides – to species-specific codon usage as well as a selective pressure on stacking energies and base-step conformational preferences, species-specific properties of DNA modification, replication and repair mechanisms, and selection by specific restriction endonucleases (Karlin *et al*., 1998). The evolutionary significance of species-specific patterns that are observed with longer oligonucleotides is unclear so far.

Here we present the application of tetranucleotide-derived z-score correlations as a measure for the relatedness of genomic fragments as well as a comparative assessment of the method's discriminatory power versus the discriminatory power of (G + C)-content differences for fosmid-sized genomic fragments. Furthermore, we demonstrate the successful application of the tetranucleotide method for the assignment of fosmid inserts from metagenome libraries that were constructed from microbial consortia involved in the anaerobic oxidation of methane (AOM).

Results and discussion

In silico *assessment of the discriminatory power of Δ(G + C) versus tetranucleotide-derived z-score correlations*

Pairwise comparisons of 118 bacterial genomes revealed that, using differences in GC-content [Δ(G + C)], artificial fosmid-sized genomic fragments of 40 kb could not be matched correctly to their genomes in 36.0% of 6903 possible comparisons. This can be explained by overlapping ranges of (G + C)-content, which in the worst case can be as extreme as illustrated in Fig. 1A for *Escherichia coli* K-12 and *Neisseria meningitidis* Z2491. For this genome pair, only a (G + C)-content of less than 48% or more than 54% would allow an unambiguous assignment of a given fragment. If a fosmid library was generated from a hypothetical habitat harbouring solely these two bacteria, the combined probability for obtaining two fosmid inserts that could be assigned to their genomes beyond doubt would be only 1%. In contrast, z-scores derived from the respective fragments' tetranucleotide usage patterns exhibit a high correlation within both genomes (>0.69) and a low correlation between them (<0.49). This enables a perfect assignment in all cases (Fig. 1B).

It is obvious that overlapping histograms as presented in Fig. 1A are the more likely the more genomes are present. By this, the Δ(G + C)-method gets saturated quickly and discrimination between fosmids renders problematic, if the number of species in a library exceeds about 10–20 species. The tetranucleotide method is less affected by such saturation, as the possible permutations of species-specific tetranucleotide usage patterns are extremely high. As long as the intragenomic variation within the tetranucleotide usage is low, the addition of more genomes will only slightly decrease the method's discriminatory power, since the correlations of tetranucleotide usage patterns will be high between fragments form the same genome and low between fragments from different genomes. Di- and trinucleotides provide less permutations and hence their discriminatory power is reduced (see Fig. 3 in Sandberg *et al*., 2001).

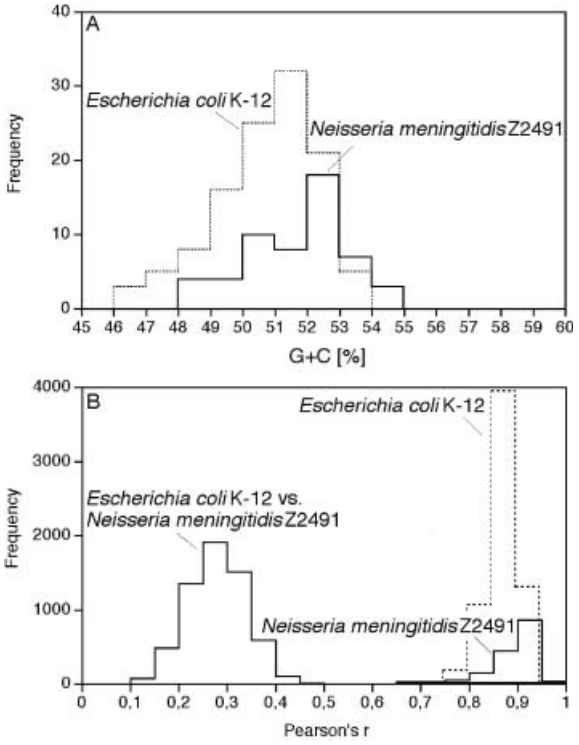

Fig. 1. (G + C)-content and correlation of tetranucleotide-derived z-scores of artificial 40 kb fragments from *Escherichia coli* K-12 and *Neisseria meningitidis* Z2491. In the upper part (A), histograms for the (G + C)-content of all 40 kb fragments from both genomes are shown. Discrimination between the two on the basis of (G + C)-content is impossible in most cases, because both histograms largely overlap. In the lower part (B), histograms of Pearson's correlation coefficients are shown for all possible pairwise comparisons of the fragment's tetranucleotide-derived z-scores. Correlation within both genomes is high (0.69–0.96) while being low between them (0.06–0.49). Discrimination of both genomes on the basis of tetranucleotide usage patterns is possible in all cases.

Our systematic evaluation of the discriminatory power of the Δ(G + C) and tetranucleotide-derived z-score correlations revealed that, in the majority of cases, the discriminatory power of the latter exceeded that of Δ(G + C) (Fig. 2). If one compares the fraction of fragment pairs from two species that can be assigned to their original genomes, then discrimination by the Δ(G + C) was only superior regarding the number of genome pairings with a near to perfect discrimination (less than 4% non-assignable fragment pairs). If one includes, however, genome pairings with a higher percentage of nonassignable fragment pairs, tetranucleotide-derived z-score correlations outperformed Δ(G + C) considerably. For instance, the number of genome pairs with at most 35% nonassignable fragment pairs was 5131 (74.3%) when the Δ(G + C) was used, but 6397 (92.7%) when tetranucleotide-derived z-score correlations were used (Fig. 3, dotted lines). Discrimination was completely impossible for 463 (6.7%) genome pairings when Δ(G + C) was used, but only for 96 (1.4%) genome pairings when tetranucleotide-derived z-score correlations were applied.

In real metagenome projects, the situation is even more complicated. Neither Δ(G + C) values nor tetranucleotide-derived z-score correlations of all possible fragment pairs within and between all genomes are known for real metagenome libraries. Thus, in the absence of marker genes, there is at first no chance to assess whether a given Δ(G + C) or tetranucleotide-derived z-score correlation between two fragments supports or contradicts the assumption that they originate from the same group or species. In order to provide the background for a decision guideline, we investigated how well a given Δ(G + C) or tetranucleotide-derived z-score correlation discriminates on the phylogenetic levels of species, orders, classes, phyla and domains. The results are summarized in Tables 1 and 2. For instance, when two fragments are randomly chosen from the same species, the average probability of obtaining a Δ(G + C) of six per cent or less is 98.0% (Table 1). When two fragments are randomly chosen from different species, the average probability for a Δ(G + C) of six per cent or less is 25.4%. A

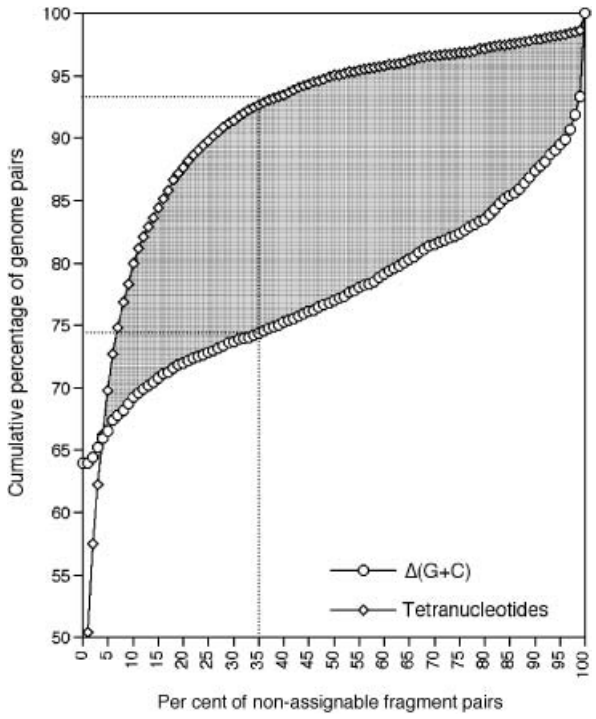

Fig. 2. For all 6903 pairwise comparisons of the 118 bacterial genomes investigated, the percentage of non-assignable fragment pairs was calculated for the Δ(G + C) and for tetranucleotide-derived z-score correlations (abscissa). Both, inter- and intragenome fragment pairs were considered. The number of genome pairs having the indicated or a better (i.e. smaller) percentage of nonassignable fragment pairs was plotted on the ordinate (cumulative plot). The hatched area indicates the region, where tetranucleotide-derived z-score correlations provide a better resolution than the Δ(G + C) and the dotted lines refer to the maximum difference between both methods.

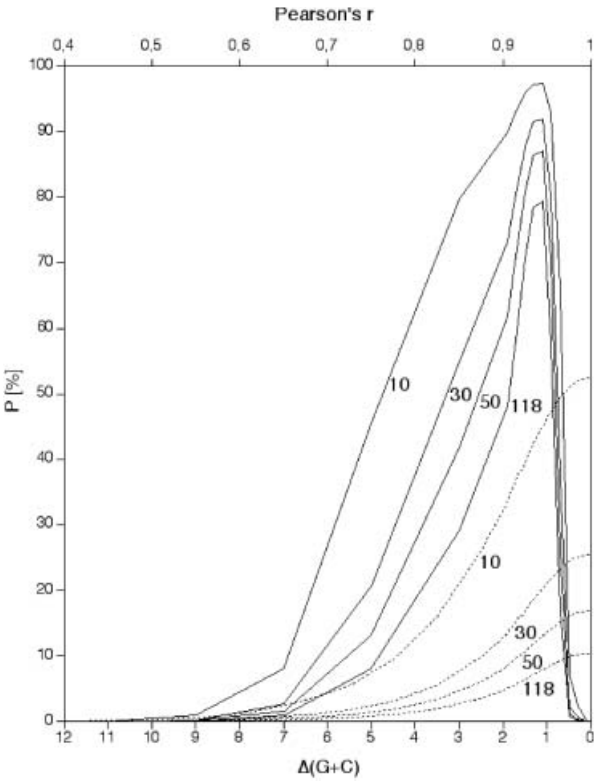

Fig. 3. Probability (P) for two 40 kb fragments to originate from the same species. Values have been calculated for the 118 species investigated in this study and for theoretical libraries containing even amounts of fragments from 10, 30 and 50 species with average-sized genomes of 3.1 Mb. Solid lines represent probabilities obtained for tetranucleotide-derived z-score correlations and dotted lines represent probabilities obtained for the $\Delta(G + C)$.

tetranucleotide-derived z-score correlation of 0.6 corresponds to a $\Delta(G + C)$ of 6% or better, as both yield nearly the same results within species (98.0% versus 98.2%, Table 2). However, the background frequency, i.e. the average probability of a correlation coefficient of 0.6 or better for fragments chosen from an arbitrary pair of genomes, is only 8.3%. This is almost three times less than the corresponding $\Delta(G + C)$ value of 25.4%. This elevated signal-to-noise ratio of tetranucleotide-derived z-score correlations becomes even more pronounced when smaller $\Delta(G + C)$ values and higher correlation coefficients are compared. In general, on each phylogenetic level, tetranucleotide-derived z-score correlations provide a better signal-to-noise ratio than corresponding $\Delta(G + C)$ values as long as the correlation coefficients are better than 0.5. This value can be considered the threshold at which a correlation coefficient begins to discriminate a signal (the relatedness of two fragments) from the noise (an arbitrary relationship).

This can also be seen if one calculates the likelihood indicated by a given $\Delta(G + C)$ or tetranucleotide-derived z-score correlation that a pair of fragments belongs to the

Table 1. Differences in (G + C)-content between 9054 artificial 40 kb fragments from 118 entire genomes (126 chromosomes). The results of all 40 982 931 pairwise comparisons were summarized on the levels of species, orders, classes and phyla (taxonomy according to the List of Bacterial names with Standing in Nomenclature (LBSN) – http://www.bacterio.cict.fr/). Tabulated values are expressed as average percentages of cases (mean ± standard deviation).

$\Delta(G + C)$	Species within	between	Order within	between	Class within	between	Phylum within	between	Domain within	between
cases ≤ 0.5%	23.0 ± 6.1	2.2 ± 0.7	8.7 ± 6.4	1.9 ± 0.7	6.5 ± 6.0	1.7 ± 0.8	4.7 ± 5.0	1.7 ± 0.9	1.2 ± 0.1	1.2 ± 0.1
cases ≤ 1%	43.2 ± 10.3	4.3 ± 1.5	16.8 ± 11.9	3.8 ± 1.5	12.6 ± 11.0	3.6 ± 1.6	9.1 ± 9.0	3.3 ± 1.9	2.3 ± 0.6	2.3 ± 0.7
cases ≤ 2%	71.6 ± 12.4	8.7 ± 2.8	31.9 ± 21.2	7.6 ± 2.9	23.9 ± 19.8	7.2 ± 3.1	17.8 ± 16.7	6.8 ± 3.7	2.8 ± 0.2	2.8 ± 0.2
cases ≤ 3%	86.0 ± 10.0	13.0 ± 4.0	44.8 ± 27.1	11.5 ± 4.0	33.3 ± 25.2	10.9 ± 4.5	25.4 ± 21.3	10.3 ± 5.4	3.5 ± 0.2	3.5 ± 0.3
cases ≤ 4%	92.9 ± 7.2	17.2 ± 5.0	55.3 ± 30.7	15.4 ± 5.1	41.2 ± 28.5	14.7 ± 5.6	32.1 ± 23.8	13.9 ± 7.0	4.4 ± 0.3	4.5 ± 0.3
cases ≤ 5%	96.3 ± 4.9	21.4 ± 6.0	63.7 ± 32.0	19.3 ± 6.1	48.2 ± 30.1	18.4 ± 6.2	38.1 ± 25.2	17.5 ± 8.5	5.6 ± 0.3	5.7 ± 0.4
cases ≤ 6%	98.0 ± 3.3	25.4 ± 7.0	70.2 ± 31.4	23.3 ± 7.1	54.4 ± 30.7	22.2 ± 7.6	44.0 ± 25.9	21.2 ± 9.9	6.8 ± 0.4	7.1 ± 0.4
cases ≤ 7%	98.9 ± 2.2	29.5 ± 7.9	75.2 ± 30.3	27.3 ± 8.0	60.1 ± 30.9	26.1 ± 8.4	49.5 ± 26.6	24.9 ± 11.2	8.1 ± 0.4	8.6 ± 0.4
cases ≤ 8%	99.9 ± 1.4	33.4 ± 8.9	79.0 ± 29.0	31.2 ± 9.0	64.8 ± 30.9	30.0 ± 9.3	54.4 ± 27.3	28.6 ± 12.2	9.5 ± 0.4	10.1 ± 0.5
cases ≤ 9%	99.7 ± 0.9	37.2 ± 9.8	81.9 ± 27.7	35.1 ± 9.9	68.8 ± 30.9	33.8 ± 10.1	58.5 ± 27.4	32.3 ± 13.6	11.0 ± 0.4	11.7 ± 0.5
cases ≤ 10%	99.8 ± 0.6	40.9 ± 10.7	84.2 ± 26.7	38.9 ± 10.7	72.0 ± 30.9	37.6 ± 10.9	62.2 ± 27.9	36.0 ± 14.2	12.5 ± 0.4	13.4 ± 0.5
cases ≤ 11%	99.9 ± 0.4	44.6 ± 11.5	85.9 ± 25.8	42.7 ± 11.5	74.6 ± 30.6	41.4 ± 11.6	65.6 ± 27.9	39.6 ± 15.0	14.0 ± 0.5	15.1 ± 0.5
cases ≤ 12%	100.0 ± 0.3	48.2 ± 12.3	87.4 ± 24.8	46.5 ± 12.3	77.1 ± 29.8	45.2 ± 12.3	69.0 ± 27.6	43.2 ± 15.8	48.6 ± 2.2	52.0 ± 6.7

Table 2. Correlation coefficients of tetranucleotide usage patterns between 9054 artificial 40 kb fragments from 118 entire genomes (126 chromosomes). The results of all 40 982 931 pairwise comparisons were summarized on the levels of species, orders, classes and phyla (taxonomy according to the List of Bacterial names with Standing in Nomenclature (LBSN) – http://www.bacterio.cict.fr/). Tabulated values are expressed as average percentages of cases (mean ± standard deviation).

Pearson's r	Species within	between	Order within	between	Class within	between	Phylum within	between	Domain within	between
cases ≥ 0.95	0.2 ± 0.9	0.0 ± 0.2	0.2 ± 0.6	0.0 ± 0.0	0.1 ± 0.5	0.0 ± 0.0	0.0 ± 0.2	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
cases ≥ 0.9	12.6 ± 21.8	0.1 ± 0.2	2.5 ± 9.3	0.0 ± 0.0	1.5 ± 5.8	0.0 ± 0.0	0.3 ± 0.6	0.0 ± 0.0	0.1 ± 0.3	0.0 ± 0.0
cases ≥ 0.8	65.5 ± 30.4	1.1 ± 2.5	13.9 ± 22.3	0.1 ± 0.3	8.6 ± 14.9	0.1 ± 0.3	3.3 ± 5.9	0.0 ± 0.0	1.3 ± 2.8	0.0 ± 0.1
cases ≥ 0.7	91.1 ± 13.3	3.1 ± 3.9	33.7 ± 26.8	1.1 ± 2.4	21.7 ± 19.9	1.0 ± 2.6	10.4 ± 10.3	0.6 ± 1.5	3.7 ± 4.3	0.3 ± 0.9
cases ≥ 0.6	98.2 ± 3.2	8.3 ± 7.0	55.8 ± 29.6	5.3 ± 5.6	41.3 ± 26.3	4.7 ± 5.8	27.8 ± 19.8	4.0 ± 4.7	10.7 ± 7.3	1.5 ± 2.7
cases ≥ 0.5	99.5 ± 1.3	19.8 ± 12.0	65.5 ± 30.0	17.1 ± 10.4	56.9 ± 29.2	15.8 ± 9.8	41.0 ± 27.4	14.8 ± 9.8	22.3 ± 12.1	6.8 ± 6.4
cases ≥ 0.4	99.8 ± 0.9	40.4 ± 17.4	70.8 ± 28.0	38.5 ± 16.4	69.3 ± 27.8	37.1 ± 16.1	58.4 ± 28.1	35.7 ± 16.6	43.5 ± 17.3	23.5 ± 16.5
cases ≥ 0.3	99.9 ± 0.6	63.9 ± 18.4	75.3 ± 24.8	62.9 ± 18.1	78.1 ± 24.4	61.8 ± 18.1	72.4 ± 25.0	61.0 ± 18.1	66.5 ± 18.4	49.7 ± 22.8
cases ≥ 0.2	99.9 ± 0.3	82.3 ± 14.8	77.2 ± 23.6	81.9 ± 14.6	82.0 ± 21.7	81.4 ± 14.7	81.3 ± 20.5	81.3 ± 14.6	83.8 ± 14.5	74.4 ± 20.9
cases ≥ 0.1	100.0 ± 0.0	93.1 ± 9.1	78.4 ± 20.8	93.0 ± 8.9	85.1 ± 16.4	92.8 ± 9.0	87.2 ± 14.6	92.9 ± 8.8	93.7 ± 8.7	90.4 ± 13.2

same species (Fig. 3). For the 118 genomes investigated, even a $\Delta(G + C)$ of zero translates to a probability of only 10.4%. In other words, the number of intergenome fragment pairs with a $\Delta(G + C)$ of zero is about nine times as high as the intragenome count. A high correlation of tetranucleotide-derived z-scores, however, discriminates far better. For instance, a correlation of 0.94 equals a probability of 79.5% that two fragments originate from the same species. Interestingly, higher correlation coefficients are very rare within genomes. Highly correlated fragment-pairs can always be found by chance between genomes, however, because the number of intergenome fragment pairs is two orders of magnitude higher than the number of intragenome pairs. Therefore, the discriminatory power of tetranucleotide-derived z-score correlations drops dramatically for correlation coefficients above 0.94. It is also important to note that the discriminatory power of both methods decreases with the number of species that are present in the library (Fig. 3), because the number of intraspecies fragment pairs (i.e. the signal) increases linearly with the number of species whereas the number of interspecies fragment pairs (i.e. the noise) increases quadratically. This implies that, in order to achieve a good signal-to-noise ratio for sequence-based fragment assignment, the overall diversity within metagenome libraries should be as low and the abundance of the species of interest should be as high as possible. The 118 genomes used in this study might not be representative for many natural sampling sites, but they indicate that discrimination of more than 100 species that contribute evenly to a library is difficult when using tetranucleotide frequencies and almost impossible when using the $\Delta(G + C)$. When, however, the species of interest clearly dominate a metagenome library, the noise might be low enough to allow for their discrimination even in the presence of a few hundred species of lower abundance. In this regard, optimal sampling sites are microbial consortia, enrichment cultures or extreme habitats where species dominate or the natural diversity is reduced. Based on log-normal distributions, it has been estimated, that the biodiversity of ocean water comprises only 160 different species per millilitre, whereas soil harbours several thousands of species per gram (Curtis *et al*., 2002). Whereas the tetranucleotide method is likely to perform well with ocean water samples, it will fail for the analysis of metagenome libraries from soil samples and other highly diverse habitats. In addition to these limitations, being a sequenced-based measure, the tetranucleotide method is affected by intragenomic fluctuations in base-composition. For example, fragments that exhibit an atypical tetranucleotide usage because they carry a high number of laterally transferred genes will not be assigned correctly. Also, the method is not likely to be able to interrelate fragments

from genomes with a high degree of sequence polymorphism and thus inhomogeneous tetranucleotide usage.

Tetranucleotide usage patterns carry a phylogenetic signal. Therefore, discrimination based on tetranucleotide usage patterns is possible not only on the species-level but also on the level of higher-order taxa, albeit with decreased discriminatory power (Table 2). Pride *et al*. (2003) for example used distances of tetranucleotide frequencies calculated by a zero-order Markov model to construct a phylogenetic tree for 27 bacterial genomes. We found that their results can be improved when whole genome sequences and tetranucleotide-derived z-score correlations from a maximal-order Markov model are used as distance measure (unpublished data – phylogenetic trees available on request). However, despite the evident phylogenetic signal in tetranucleotide frequencies, both methods fail to reconstruct phylogenetic trees for all publicly available genomes, that reflect the standard 16S rRNA based topology. Whereas closely related species are correctly grouped in most cases, more distant ones often are not. In other words, the phylogenetic signal in tetranucleotide usage patterns quickly fades in moving from the species level to the higher order taxa. Both, phylogenetic analysis and the data presented in Table 2 show that distant relationships cannot be resolved on the basis of tetranucleotide usage patterns.

Application of the tetranucleotide method to real fosmid insert sequences

In addition to the *in silico* assessment of tetranucleotide-derived z-score correlations as a measure for the relatedness of genomic fragments, we applied the method to real fosmid insert sequences originating from two metagenome libraries. These libraries were constructed from samples of methane-rich habitats, that exhibit AOM activity and are characterized by high abundances of sulphate-reducing bacteria and *Archaea* of the ANME-2 (Boetius *et al.*, 2000) and the ANME-1 cluster (Michaelis *et al.*, 2002) respectively. Consequently, these libraries were dominated by few species. Phylogenetically, ANME-2 belongs to *Methanosarcinales* whereas ANME-1 is only distantly related to this order. In total, six insert sequences were analysed – two non-overlapping inserts with 16S rRNA genes belonging to ANME-2 (ANME-2a, ANME-2c) and four overlapping inserts belonging to ANME-1 (Fig. 4). Two of the ANME-1 inserts carried identical 16S

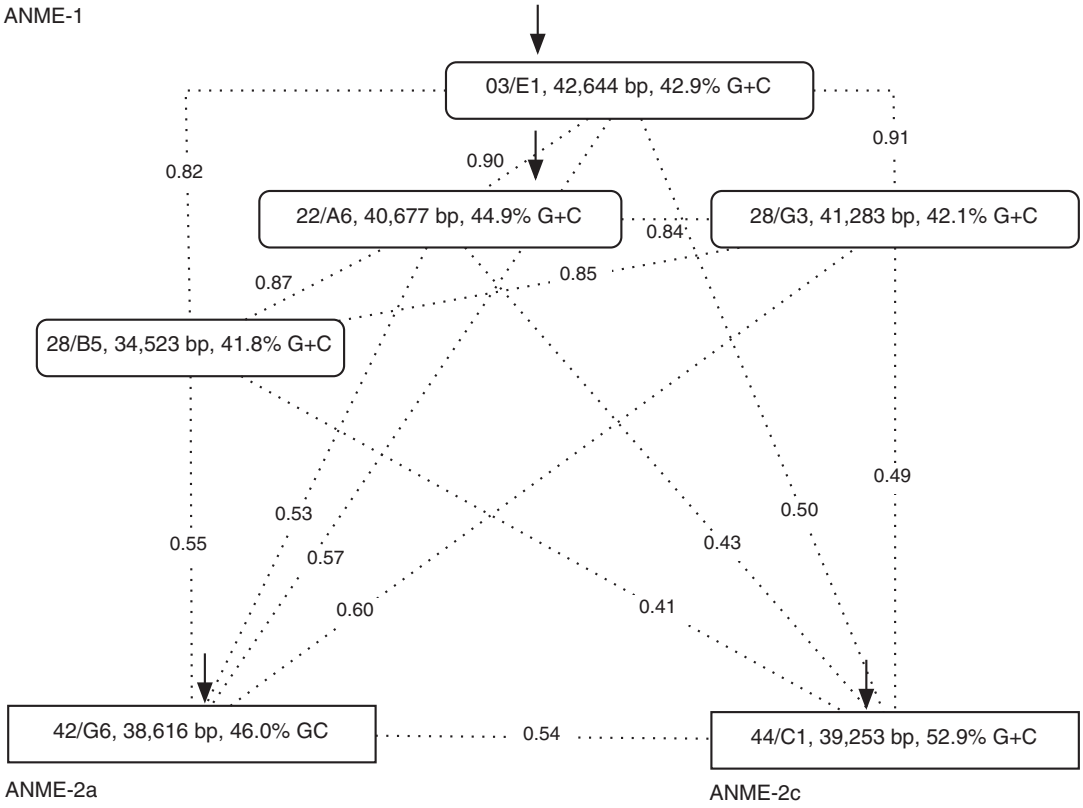

Fig. 4. Correlation coefficients of tetranucleotide-derived z-scores between genomic fragments from bacterial communities mediating AOM. Fragments are represented as rounded boxes (ANME-1) or rectangles (ANME-2). The overlapping regions of the ANME-1 fragments correspond proportion-wise to their observed overlaps. Dotted lines and numbers correspond to the respective correlation coefficients and arrows indicate the positions of 16S rRNA genes.

rRNA genes and overlapped by 21.0 kb, with 47 mismatches in the overlap. This indicates that they belong to closely related but different species. For each of these two inserts, a perfectly overlapping insert without the 16S rRNA gene is available (overlaps 17.5 and 10.4 kb respectively). The ANME-1 sequences have an average (G + C)-content of 42.9% with a maximum Δ(G + C) of 3.1%. By this measure, they are clearly different from ANME-2c (52.9% G + C) but not from ANME-2a (46.0% G + C). Based on the values obtained for a hypothetical library of 10 species (Fig. 3), a Δ(G + C) of 3.1% as observed within the ANME-1 fragments and between the ANME-1 fragments and ANME-2a corresponds to a probability of 21% that these fragments belong to the same species. The Δ(G + C) of 10% that has been found between ANME-1 and ANME-2c corresponds to a probability of less than 1%. Thus, from the Δ(G + C), it is highly unlikely that the ANME-1 and ANME-2c fragments belong to the same species. Discrimination between the ANME-1 and ANME-2a fragments, however, is not possible on the basis of their Δ(G + C), even though their 16S rDNA sequences indicate that they belong to different species (81% 16S rRNA identity). In contrast to this, discrimination results are congruent with the results of the 16S rRNA sequence analysis when tetranucleotide-derived z-score correlations are used. Correlation is high among the ANME-1 fragments (0.82–0.91; identical 16S rRNA) and low between fragments of ANME-1/ANME-2a (≤0.60; 81% 16S rRNA similarity), ANME-1/ANME-2c (≤0.50; 83% 16S rRNA similarity) and ANME-2a/ANME-2c (0.54; 90% 16S rRNA similarity). Referring to the 10-species curve (Fig. 3), a correlation of 0.91–0.82 corresponds to a probability of 75–93% that the ANME-1 fragments originate from the same species, whereas all other correlation coefficients correspond to a probability of less than 4%. Thus, in contrast to the Δ(G + C) method, tetranucleotide-derived z-score correlations are able to discriminate both, the ANME-2c fragment as well as the ANME-2a fragment from the others and in addition strongly suggest that all four ANME-1 fragments belong to the same or at least very closely related species.

We would like to emphasize that in the case of the ANME-1 insert sequences, tetranucleotide-derived z-score correlations were high not only between the overlapping inserts (which is expected, since they share considerable parts of their sequences), but also between the non-overlapping inserts. Therefore, we regard the tetranucleotide method as being well suited to tackle the fragment identification problem in metagenomics.

Implications for metagenomics

Based on tetranucleotide usage patterns, genomic fragments derived for the same (or closely related) species could be assigned with reasonable probabilities even in the absence of suitable marker genes. Thus, together with such widely used identification approaches as marker genes, the Δ(G + C), or the gene's best BLAST hits and codon usage, the analysis of tetranucleotide-derived z-score correlations enhances our capability to classify genomic fragments. Further improvements of the method could include the combination of di-, tri- and tetranucleotide frequencies and the application of the self organizing map variant of neuronal networks. Using these methods, it has been shown that in most cases genomic fragments of 10 kb and sometimes even fragments of 1 kb retain species-specific information (Abe *et al.*, 2003). Even sequences as short as 400 bases can be correctly classified with 85% probability, when the sequences of the genomes they belong to are known and thus a model for signature oligonucleotides can be built. This has been demonstrated using a naïve Bayesian classifier for dimers up to nonamers (Sandberg *et al.*, 2001). These results indicate the large potential of genome linguistic approaches to solve the fragment identification problem in metagenomics. Hence, sequencing does not need to be restricted to genomic fragments carrying phylogenetic markers or functional genes of interest.

When the relatedness of genomic fragments is to be assessed on the basis of skewed oligonucleotide distributions, complete sequencing rather than the cost-effective end-sequencing is recommended since reliability improves with sequence length. We hope that the tetranucleotide method will grow to be a valuable addition to the assignment tools available to scientists in the field of metagenomics.

Experimental procedures

Sequences for in silico *evaluation of (G + C)-content and tetranucleotide usage patterns*

Sequences of 116 publicly available prokaryote genomes were obtained from the National Center for Biotechnology Information (NCBI) website (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/). These were complemented by the recently published genome sequences of Pirellula sp. strain 1 (Glöckner *et al.*, 2003), the as yet unpublished genome sequence of *Desulfotalea psychrophila* from the Real Environmental Genomics project (REGX) consortium (http://www.regx.de) and six insert sequences from metagenome libraries from methane-rich sites (A. Meyerdierks, personal communication). In total, 126 chromosomes from the following 118 genomes were evaluated: *Aeropyrum pernix* K1, *Agrobacterium tumefaciens* strain C58 Cereon, *Agrobacterium tumefaciens* strain C58 UWash, *Aquifex aeolicus* VF5, *Archaeoglobus fulgidus* DSM 4304, *Bacillus cereus* ATCC 14579, *Bacillus halodurans* C-125, *Bacillus subtilis* ssp. *subtilis* 168, *Bacteroides thetaiotaomicron* VPI-5482, *Bifidobacterium longum* NCC2705, *Borrelia burgdorferi* B31, *Bradyrhizobium japonicum* USDA110, *Brucella melitensis*

16M, *Brucella suis* 1330, *Buchnera aphidicola* APS, *Buchnera aphidicola* Sg, *Buchnera aphidicola* Sg, *Campylobacter jejuni* ssp. *jejuni* NCTC 11168, *Caulobacter crescentus* CB15, *Chlamydia muridarum* strain Nigg, *Chlamydia trachomatis* serovar D, *Chlamydophila caviae* GPIC, *Chlamydophila pneumoniae* AR39, *Chlamydophila pneumoniae* CWL029, *Chlamydophila pneumoniae* J138, *Chlorobium tepidum* TLS, *Clostridium acetobutylicum* ATCC 824, *Clostridium perfringens* 13, *Clostridium tetani* E88, *Corynebacterium efficiens* YS-314, *Corynebacterium glutamicum* ATCC 13032, *Coxiella burnetii* RSA 493, *Deinococcus radiodurans* R1, *Desulfotalea psychrophila*, *Enterococcus faecalis* V583, *Escherichia coli* CFT073, *Escherichia coli* K12-MG1655, *Escherichia coli* O157:H7 VT2-Sakai, *Escherichia coli* O157:H7 EDL933, *Fusobacterium nucleatum* subsp *nucleatum* ATCC 25586, *Haemophilus influenzae* Rd, *Halobacterium* sp. NRC-1, *Helicobacter pylori* 26695, *Helicobacter pylori* J99, *Lactococcus lactis* subsp *lactis* IL1403, *Leptospira interrogans* serovar lai strain 56601, *Listeria innocua* CLIP 11262, *Listeria monocytogenes* EGD-e, *Mesorhizobium loti* MAFF303099, *Methanothermobacter thermoautotrophicus* delta H, *Methanocaldococcus jannaschii* DSM2671, *Methanopyrus kandleri* AV19, *Methanosarcina acetivorans* C2A, *Methanosarcina mazei* Goe1, *Mycobacterium leprae* TN, *Mycobacterium tuberculosis* CDC1551, *Mycobacterium tuberculosis* H37Rv, *Mycoplasma genitalium* G37, *Mycoplasma penetrans* HF-2, *Mycoplasma pneumoniae* M129, *Mycoplasma pulmonis* UAB CTIP, *Neisseria meningitidis* serogroup A Z2491, *Neisseria meningitidis* MC58, *Nostoc* sp. PCC 7120, *Oceanobacillus iheyensis* HTE831, *Pasteurella multocida* PM70, *Pirellula* sp. strain 1, *Pseudomonas aeruginosa* PA01, *Pseudomonas putida* KT2440, *Pyrobaculum aerophilum* IM2, *Pyrococcus abyssi* GE5, *Pyrococcus furiosus* DSM 3638, *Pyrococcus horikoshii* shinkaj OT3, *Ralstonia solanacearum* GMI1000, *Rickettsia conorii* Malish 7, *Rickettsia prowazekii* Madrid E, *Salmonella enterica* ssp. enterica serovar Typhi, *Salmonella typhi* CT18, *Salmonella typhimurium* LT2 SGSC1412, *Shewanella oneidensis* MR1, *Shigella flexneri* 2a strain 301, *Shigella flexneri* 2a 2457T, *Sinorhizobium meliloti* 1021, *Staphylococcus aureus* Mu50, *Staphylococcus aureus* ssp. aureus MW2, *Staphylococcus aureus* ssp. aureus N315, *Staphylococcus epidermidis* ATCC 12228, *Streptococcus agalactiae* A909, *Streptococcus agalactiae* 2603 V/R, *Streptococcus agalactiae* NEM316, *Streptococcus mutans* UA159, *Streptococcus pneumoniae* R6, *Streptococcus pneumoniae* TIGR4, *Streptococcus pyogenes* MGAS315, *Streptococcus pyogenes* MGAS8232, *Streptococcus pyogenes* SF370 serotype M1, *Streptomyces coelicolor* A3(2), *Sulfolobus solfataricus* P2, *Sulfolobus tokodaii* strain 7, *Synechocystis* sp. PCC 6803, *Thermoanaerobacter tengcongensis* MB4(T), *Thermoplasma acidophilum* DSM 1728, *Thermoplasma volcanium* GSS1, *Thermosynechococcus elongatus* BP-1, *Thermotoga maritima* MSB8, *Treponema pallidum* Nichols, *Tropheryma whipplei* TW08 27, *Tropheryma whipplei* strain Twist, *Ureaplasma urealyticum* parvum biovar serovar 3, *Vibrio cholerae* El Tor N16961, *Vibrio vulnificus* CMCP6, *Wigglesworthia brevipalpis*, *Xanthomonas axonopodis* pv. citri 306, *Xanthomonas campestris* pv. campestris ATCC 33913, *Xylella fastidiosa* 9a5c, *Xylella fastidiosa* Temecula1, *Yersinia pestis* CO92, *Yersinia pestis* KIM.

Assessment of the resolution power of (G + C)-content and tetranucleotide usage patterns

All 118 genomes were split into artificial fosmid-sized 40 kb fragments (9054 in total). The Δ(G +C) and tetranucleotide-derived z-score correlations were calculated for all 40 982 931 possible fragment pairs (460 084 intra- and 40 522 847 inter-genome pairs). Results were summarized on the levels of species, orders, classes and phyla (Tables 1 and 2).

Calculation of tetranucleotide frequencies and correlation coefficients

In brief, all fragments were extended with their reverse complements. The observed frequencies of all 256 possible tetranucleotides and their corresponding expected frequencies were computed for these sequences. The differences between observed and expected values were transformed into z-scores for each tetranucleotide. The similarity between two fosmids was assessed by calculating the Pearson correlation coefficient for their 256 tetranucleotide-derived z-scores.

In more detail, the expected frequencies and z-scores were computed according to the method published by Schbath *et al.*, (1995): if we denote the observed frequency of a tetranucleotide within a given sequence as N(n1n2n3n4), then the corresponding expected frequency E(n1n2n3n4) can be calculated by means of a maximal-order Markov model:

$$E(n1n2n3n4) = \frac{N(n1n2n3)N(n2n3n4)}{N(n2n3)}$$

The significance of the level of over- or underrepresentation, i.e. the divergence between observed and expected frequencies, can be evaluated using z-scores

$$Z(n1n2n3n4) = \frac{N(n1n2n3n4) - E(n1n2n3n4)}{\sqrt{\mathrm{var}(N(n1n2n3n4))}}$$

whereby the variance var(N(n1n2n3n4)) can be approximated as follows:

$$\mathrm{var}(N(n1n2n3n4)) = E(n1n2n3n4) * \frac{[N(n2n3) - N(n1n2n3)][N(n2n3) - N(n2n3n4)]}{N(n2n3)^2}$$

The question, if two genomic fragments exhibit similar patterns of over- and underrepresented tetranucleotides, can be addressed by calculating the Pearson correlation coefficient for their z-scores. Similar patterns correlate and thus have high correlation coefficients, whereas diverging patterns have low correlation coefficients.

Calculation of percentages of not-assignable fragment pairs

When comparing two genomes, fragment pairs can only be assigned with certainty to their genomes of origin when the respective Δ(G + C) or tetranucleotide-derived z-score correlations do not reside within regions where the values of both genomes overlap (Fig. 1). For each genome pairing, the number of fragment-pairs was determined that fulfill this condition (for both, intra- and intergenome fragment pairs). The

percentage of nonassignable fragment pairs was calculated by relating this number with the total number of possible fragment pairs (Fig. 2). For a pair of genomes (1, 2), with N1 and N2 40 kb fragments, this is the sum of all intra- and intergenome fragment-pairs [N1(N1-1)/2 + N2(N2-1)/2 + N1N2].

Calculation of the probabilities to originate from the same species

The number of fragment pairs having a given $\Delta(G + C)$ or tetranucleotide-derived z-score correlation was determined for all 460 084 intra- and 40 522 847 intergenome pairs. For a hypothetical unbiased fosmid-library containing the 118 genomes investigated, the likelihood for two fragments to originate from the same species corresponds to the fraction that intragenome fragment pairs make up from all fragment pairs having the respective $\Delta(G + C)$ or tetranucleotide-derived z-score correlation (Fig. 3). Assuming that the percentages of intra- and intergenomic values obtained for the 118 genomes are representative for typical bacteria, numbers have also been calculated for hypothetical fosmid libraries of 10, 30 and 50 average-sized genomes respectively [the average size of all publicly available bacterial genomes listed at NCBI to date is 3.1 Mb (http://www.ncbi.nlm.nih.gov/Genomes/)].

Acknowledgements

We would like to thank Tim Frana for cross-checking PERL scripts and results and the Max Planck society for funding this study. Retrieval of metagenome sequences was carried out within the framework of the Competence Network Göttingen 'Genome research on bacteria' (GenoMik) financed by the German Federal Ministry of Education and Research (BMBF).

References

Abe, T., Kanaya, S., Kinouchbi, M., Ichiba, Y., Kozuki, T., and Ikemura, T. (2003) Informatics for unveiling hidden genome signatures. *Genome Res* **13:** 693–705.

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995) Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol Rev* **59:** 143–169.

Béjà, O., Suzuki, M.T., Koonin, E.V., Aravind, L., Hadd, A., Nguyen, L.P., *et al.* (2000a) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* **2:** 516–529.

Béjà, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P. *et al.* (2000b) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289:** 1902–1906.

Béjà, O., Spudich, E.N., Spudich, J.L., Leclerc, M., and DeLong, E.F. (2001) Proteorhodopsin phototrophy in the ocean. *Nature* **411:** 786–789.

Boetius, A., Ravenschlag, K., Schubert, C.J., Rickert, D., Widdel, F., Gieseke, A., *et al.* (2000) A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* **407:** 623–626.

Curtis, T.P., Sloan, W.T., and Scannell, J.W. (2002) Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci USA* **99:** 10494–10499.

DeLong, E.F. (2002) Microbial population genomics and ecology. *Curr Opin Microbiol* **5:** 520–524.

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

Gentles, A.J., and Karlin, S. (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res* **11:** 540–546.

Glöckner, F.O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W., *et al.* (2003) Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc Natl Acad Sci USA* **100:** 8298–82303.

Goldman, N. (1993) Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Res* **21:** 2487–2491.

Karlin, S. (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol* **1:** 598–610.

Karlin, S., and Ladunga, I. (1994) Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci USA* **91:** 12832–12836.

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

Karlin, S., and Mrázek, J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J Bacteriol* **182:** 5238–5250.

Karlin, S., Ladunga, I., and Blaisdell, B.E. (1994) Heterogeneity of genomes: measures and values. *Proc Natl Acad Sci USA* **91:** 12837–12841.

Karlin, S., Campbell, A.M., and Mrázek, J. (1998) Comparative DNA analysis across diverse genomes. *Annu Rev Genet* **32:** 185–225.

Michaelis, W., Seifert, R., Nauhaus, K., Treude, T., Thiel, V., Blumenberg, M., *et al.* (2002) Microbial reefs in the black sea fueled by anaerobic oxidation of methane. *Science* **297:** 1013–1015.

Nakashima, H., Ota, M., Nishikawa, K., and Ooi, T. (1998) Genes from nine genomes are separated into their organisms in the dinucleotide composition space. *DNA Res* **5:** 251–259.

Pride, D.T., Meinersmann, R.J., Wassenaar, T.M., and Blaser, M.J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13:** 145–158.

Rondon, M.R., August, P.R., Bettermann, A.D., Bradly, S.F., Grossman, T.H., Liles, M.R., *et al.* (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorgansims. *Appl Environ Microbiol* **66:** 2541–2547.

Sandberg, R., Winberg, G., Bränden, C.I., Kaske, A., Ernberg, I., and Cöster, J. (2001) Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genome Res* **11:** 1404–1409.

Schbath, S., Prum, B., and de Turckheim. E. (1995) Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J Comput Biol* **2:** 417–437.

Schloss, P.D., and Handelsman, J. (2003) Biotechnological prospects from metagenomics. *Curr Opin Biotechnol* **14:** 303–310.