Peter A. Noble[1]
Robert W. Citek[2]
Oladele A. Ogunseitan[3]

[1]Belle W. Baruch Institute for Marine Biology and Coastal Research, University of South Carolina, Columbia, SC, USA
[2]Department of Soil and Environmental Science, University of California at Riverside, Riverside, CA, USA
[3]Department of Environmental Analysis and Design, University of California at Irvine, Irvine, CA, USA

Tetranucleotide frequencies in microbial genomes

A computational strategy for determining the variability of long DNA sequences in microbial genomes is described. Composite portraits of bacterial genomes were obtained by computing tetranucleotide frequencies of sections of genomic DNA, converting the frequencies to color images and arranging the images according to their genetic position. The resulting images revealed that the tetranucleotide frequencies of genomic DNA sequences are highly conserved. Sections that were visibly different from those of the rest of the genome contained ribosomal RNA, bacteriophage, or undefined coding regions and had corresponding differences in the variances of tetranucleotide frequencies and GC content. Comparison of nine completely sequenced bacterial genomes showed that there was a nonlinear relationship between variances of the tetranucleotide frequencies and GC content, with the highest variances occurring in DNA sequences with low GC contents (less than 0.30 mol). High variances were also observed in DNA sequences having high GC contents (greater than 0.60 mol), but to a much lesser extent than DNA sequences having low GC contents. Differences in the tetranucleotide frequencies may be due to the mechanisms of intercellular genetic exchange and/or processes involved in maintaining intracellular genetic stability. Identification of sections that were different from those of the rest of the genome may provide information on the evolution and plasticity of bacterial genomes.

1 Introduction

The existing order of nucleotides in prokaryotic chromosomes specifies biological information according to the genetic code. The contiguity of nucleotide sequences is affected by many processes such as restriction enzyme systems that regulate foreign DNA invasion and provide DNA fragments for recombination [1]. The order of nucleotides is also a function of biases introduced during polymerase activities in DNA replication and repair. Such biases include discordance between specificities of the deoxycytosine methylase and the very short patch DNA mismatch repair system [2, 3]. Certain oligonucleotides may be preferred or avoided because they optimize protein binding and codon-mediated regulation of translation [4, 5]. Physical constraints such as dinucleotide stacking energies, curvature and superhelicity of DNA also influence the order of nucleotides [6–8]. For example, the less thermodynamically stable dinucleotide TA is more prevalent at sites involved in untwisting double-strand DNA than other dinucleotides [6]. Presumably, these processes maintain genetic stability by prescribing the order of nucleotides in bacterial genomes. Further imposition on the order of nucleotides in DNA are due to the mechanisms of genetic change, which include deletions, insertions, transpositions, duplications and recombinations of genetic material. These mechanisms alter the genetic composition of bacteria, providing numerous possibilities for variation [4]. Although conventional methods for calculating the similarities of DNA or protein sequences provide information on the evolution of genes, there is a paucity of methods to investigate long DNA sequences (> 2500 bp). Such methods are needed to identify regions of microbial genomes affected by the mechanisms of genetic change and those processes involved in maintaining genetic stability. This information is necessary to understand the evolution of microbial genomes.

In this study, we explore variability in bacterial genomes by computing oligonucleotide frequencies for sections of genomic DNA. The oligonucleotide frequencies will be used for comparing these sections and identifying regions of the genome having similar and dissimilar tetranucleotide frequencies. With the exception of sequences resulting from intracellular genetic exchange, all sections of a given genome may be expected to have similar oligonucleotide frequencies because they have been acted upon by the same mechanisms that maintain genetic stability. Exogenously acquired DNA sequences should have dissimilar oligonucleotide frequencies from those of its host because they have been acted upon by different mechanisms and therefore have different evolutionary histories. Moreover, DNA sequences encoding ribosomal RNA should be evolutionarily conserved because RNA plays an important role in protein synthesis. Since some bacteria exhibit more genetic and physiological diversity than others, variability of genomic DNA should be different among genetically unrelated bacteria, this being a function of dissimilar evolutionary processes.

Here, we describe a computational strategy for examining variability in long DNA sequences. This strategy was used to examine the following bacterial genomes: *Archaeoglobus fulgidus, Mycoplasma genitalium, M. pneumoniae, Methanococcus jannaschii, Haemophilus influenzae, Escherichia coli, Helicobacter pylori, Treponema pallidum and Synechocystis* sp. Composite portraits of bacterial genomes were obtained by computing the tetranucleotide frequencies of sections of genomic DNA, converting the frequencies to color images and arranging

Correspondence: Dr. P. A Noble, Belle W. Baruch Institute for Marine Biology and Coastal Research, University of South Carolina, Columbia, SC 29208, USA (Tel: +803-777-3928; Fax: +803-777-3935; E-mail: noble@biol.sc.edu)

Keywords: Fingerprinting / Visual image / Genomic heterogeneity / Genome organization

0173-0835/98/0404-0528 $17.50+.50/0

the images according to their genetic positions. In addition, we calculated the variance of tetranucleotide frequencies in order to identify sections which were dissimilar from other regions of the genome.

2 Methods

DNA sequences and information pertaining to the location of genes and ribosomal RNA of *H. influenzae, M. jannaschii*, and *M. genitalium* were obtained from http://www.tigr.org/ [9–11]. DNA sequences of *Synechocystis sp.* [12, 13] and *E. coli* were obtained from http://www.kazusa.or.jp/cyano/cyano.html and http://www.genetics.wisc.edu:80/index.html, respectively. Information pertaining to the location of genes and ribosomal RNA of *E. coli* were obtained from Yamamoto *et al.* [14, 15] and Burland *et al.* [16]. Information pertaining to the location of genes and ribosomal RNA of *M. pneumoniae* was obtained from http://www.zmbh.uni-heidelberg.de/M_pneumoniae/Herrmann/Download.html and Himmelreich *et al.* [17, 18]. The complete DNA sequences of *H. pylori, T. pallidum* and *A. fulgidus* were obtained from ftp.tigr.org. Di- and tetra-nucleotides frequencies were sequentially computed for 3000 bp sections of genomic DNA using a C++ program on a Unix computer. The frequencies were compiled into a spreadsheet (Microsoft Excel) and the frequencies of complimentary tetranucleotides (*i.e.*, AAAA and TTTT) were added together. Variance of the tetranucleotide frequencies was computed for each 3000 bp section by using the equation:

$$\text{Variance} = s\hat{}2/X \qquad (1)$$

where s and X are the standard deviation and mean of complimentary tetranucleotide frequencies, respectively. The variance of the tetranucleotide frequencies equals 0 when all possible tetranucleotide frequencies are equal. Composite portraits of bacterial genomes were assembled by converting tetranucleotide frequency data to text files and importing the files to Transform 3.01 software (Spyglass, Inc., Savoy, IL). These data were converted to colors using numerical thresholds preset by the user. For all images, thresholds of 0 (purple) and 50 (red) units were used. Tiff images were converted to Pict format and imported into MacDraw software. Variances of the tetranucleotide frequencies and GC values were graphed by using MS Excel and transferred as PICT images to MacDraw for image manipulation and labeling.

3 Results and discussion

The complete *H. influenzae* genome is depicted in Fig. 1. Comparison of the horizontal bands showed that some tetranucleotides consistently had low or high frequencies. Tetranucleotides with low frequencies (Fig. 1, purple) were entirely composed of cytosine and/or guanine (*e.g.*, CCGG), while tetranucleotides with high frequencies (yellow, orange and red) were composed of adenine and/or thymine (*e.g.*, AAAA and/or TTTT). Examination of the fingerprints showed that some regions of the *H. influenzae* genome have distinctly different tetranucleotide frequencies, as indicated by the colors, than those of other regions (Fig. 1). The most distinctive fingerprints are those of the cryptic Mu-like bacteriophage located in the region between 156 and 159 × 10^4 bp [9]. Differences in the fingerprints were also apparent in regions of the genome encoding ribosomal RNA, located at 12, 24, 63, 66, 77 and 181 × 10^4 bp, and ribosomal proteins, located from 84 to 85 × 10^4 bp (Fig. 1). Composite portraits of the genomes of the other bacteria yielded similar results. Distinctive fingerprints were found in all bacterial species investigated (data not shown). Visual differences in the fingerprints in different regions of the *H. influenzae* genome were not due to extremely high or low tetranucleotide frequencies but rather to changes in the frequencies of many tetranucleotides. Furthermore, these regions had low variances and high GC values when compared to the rest of the genome (Fig. 1), indicating that there might be a relationship among fingerprints, variances of the tetranucleotides and GC content.

3.1 Tetranucleotide frequencies and GC content

To determine the relationship between variance of the tetranucleotide frequencies and GC content, we compared DNA sequences of nine completely sequenced bacterial genomes (Fig. 2). In general, genome sections having low GC values (*i.e.*, less than 0.30 mol) had high variances, indicating that some tetranucleotides, presumably those rich in AT, occurred more frequently in DNA sequences than others (Fig. 2). High variances were also observed in sections having high GC values (*i.e.*, greater than 0.60 mol), but not to the same extent as sections having low GC values. Figure 2 shows that the lowest variances of tetranucleotide frequencies occurred in sections having GC values of approximately 0.50 mol, indicating that the frequencies of all possible tetranucleotides in these sections are more or less similar. Sections with low or high variances had GC values ranging from 0.25 to 0.63 mol, indicating a nonlinear relationship between GC content and variance of tetranucleotide frequencies.

We compared variance and GC values of two genetically related bacteria, *M. genitalium* and *M. pneumoniae*, to determine if any genome sections were similar. Variances and GC values overlapped for several sections of the genomes (Fig. 2). These findings are in agreement with Himmelreich *et al.* [17, 18], who showed that *M. genitalium* and *M. pneumoniae* share many similar coding regions. However, the ranges of GC and variance values for these genomes were dissimilar (Fig. 2). For example, several sections of the *M. pneumoniae* genome had higher GC values and lower variances than sections of the *M. genitalium* genome. Furthermore, sections of the *M. pneumoniae* genome, those having variances and GC values in the range of 4–9, and 0.41–0.44 mol, respectively, were not present in *M. genitalium* (Fig. 2). Since the genome of *M. genitalium* (0.58 Mbp) is smaller than that of *M. pneumoniae* (0.82 Mbp), it is possible that these sections are absent from the *M. genitalium* genome. Alternatively, these sections may be present in the *M. genitalium* genome but at much lower GC values and/or higher variances.

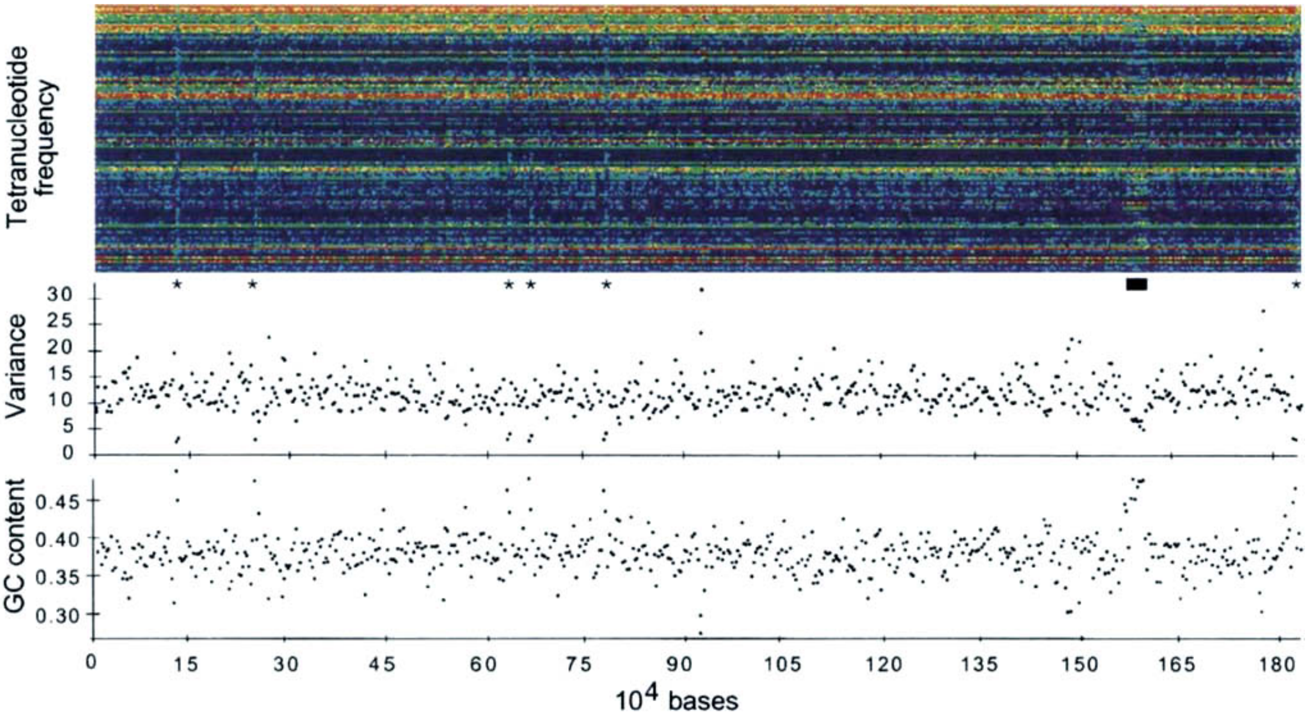

Figure 1. Fingerprints, variances of tetranucleotide frequencies, and GC values of sections of the *Haemophilus influenzae* Rd genome are consecutively ordered from the *Not*I restriction site [9]. Each column of the color image represents the fingerprint obtained from the analysis of one DNA sequence (*i.e.*, a 3000 bp section). Each row represents the frequency of a specific tetranucleotide and its complement. Tetranucleotides are arranged alphabetically on the *y*-axis. Each tetranucleotide is represented by a box, whose color is determined by its frequency, ranging from purple (low) to red (high). A star (*) identifies sections containing ribosomal RNA. The black bar identifies the location of the cryptic Mu-like bacteriophage. The variance and GC values were computed from the analysis of one section.

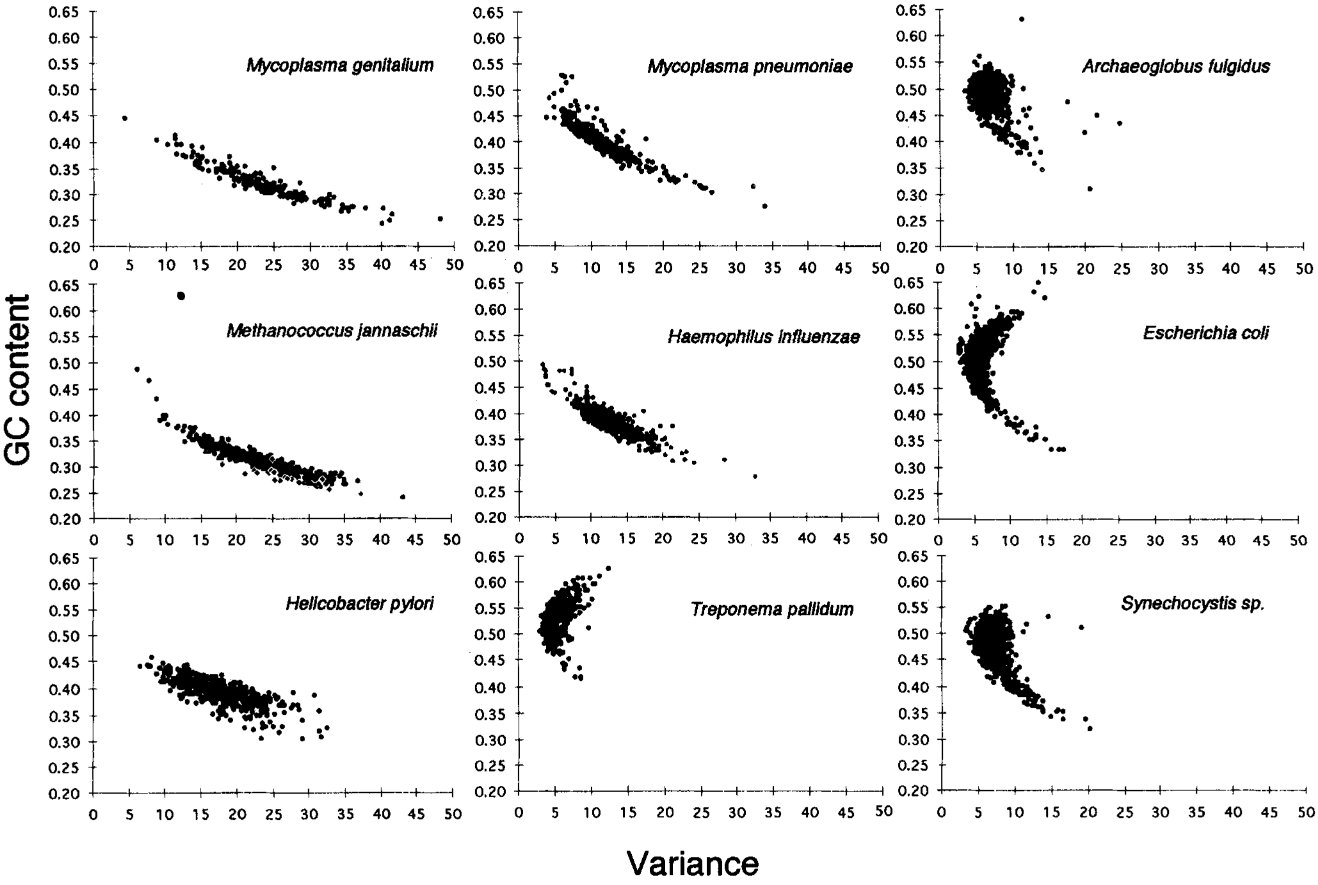

Figure 2. The GC content and variance of tetranucleotide frequencies by genome section. Each datum point represents the GC content and corresponding variance obtained from the analysis of one 3000 bp DNA sequence. Labels: circle, chromosomal DNA; diamond, extrachromosomal DNA 58 kbp of *M. jannaschii*, triangle, 16 kbp DNA of *M. jannaschii*.

Table 1. Subsections of *M.genitalium, M. pneumoniae, H. influenzae, M. jannaschii* and *E. coli* having the lowest variances were sequentially ordered by genetic position. Gene identification are based on Fraser *et al.* [10], Himmelreich *et al.* [17, 18], Fleischmann *et al.* [9] and Bult *et al.* [11], respectively

			Location by base pair:		Number of	
Organism	Variance	GC content	Start	End	Combined Segments	Putative identification:
M. genitalium	4.26-11.27	0.37-0.44	168001	177000	3	*rrn* A
	12.07	0.39	192001	195000	1	ribosomal proteins (*rpL23, rpL22, rpL2, rpS19, rpS3, rpL16*)
	11.04-11.16	0.40-0.41	222001	228000	2	attachment protein (MgPa operon), hypothetical protein
	10.03	0.39	255001	258000	1	pyruvate kinase
	12.23	0.37	330001	333000	1	pyruvate dehydrogenase and dihydrolipoamide acetyltransferase (*pdhABC*)
	11.38	0.39	429001	432000	1	RNA polymerase (*rpoC*)
M. pneumoniae	4.76-6.39	0.49-0.52	15001	27000	4	repetitive DNA sequences REPMP1, REPMP2/3, REPMP4 & REPMP5, hypothetical proteins, ADP1_MYCPN adhesin P1
	6.34	0.46	33001	36000	1	putative lipoprotein, repetitive DNA sequence REPMP2/3, ADP1_MYCPN adhesin P1 presursor homolog
	3.69-6.56	0.43-0.46	84001	93000	4	*rrn* A, repetitive DNA sequences REPMP1 & REPMP5, hypothetical proteins
	6.53	0.46	168001	171000	1	DNA polymerase III (*dnaE*), uracil phosphoribosyltransferase (*upp*), hypothetical protein
	6.18	0.43	183001	186000	1	transport ATP-binding proteins (*msbA* & *pmd1*)
	5.81-6.06	0.45	198001	204000	2	DNA gyrase (*gyrAB*), seryl-tRNA synthetase (*serS*)
	6.03	0.44	267001	270000	1	phosphoglycerate mutase (*pgm*)
	5.90-6.41	0.44-0.46	387001	393000	2	RNA polymerase beta chain (*rpoBC*)
	6.43	0.51	411001	414000	1	repetitive DNA seq REPMP2/3, ADP1_MYCPN adhesin P1 presursor
	4.76-4.86	0.44-0.49	456001	462000	2	repetitive DNA sequence REPMP4 and REPMP5, tRNA, hypothetical proteins, ADP1_MYCPN adhesin P1 precursor, Na(+) translocating ATPase subunit J
	6.17-6.41	0.43-0.45	465001	471000	2	putative lipoprotein, tRNAs, 30 K adhesion-related protein, cytochrome C oxidase polypeptide I & accessory protein (*hmw* 3)
	6.05	0.44	663001	666000	1	tRNAs, pyruvate kinase
	6.31	0.43	744001	747000	1	ribosomal proteins, prolipoprotein diacylglyceryl transferase, elongation factor G
	4.77	0.46	762001	765000	1	excinuclease ABC subunit B (*uvrB*),preprotein translocase (*secA*)
	6.44	0.44	792001	795000	1	ribosomal protein, initiation factor (*infA*), methionine amino peptidase, adenyl kinase, preprotein translocase subunit (*secY*)
H. influenzae	3.05-3.72	0.45-0.49	123001	129000	2	*rrnE*
	3.38	0.48	243001	246000	1	*rrnF*
	6.86	0.43	249001	252000	1	ionsine-5'-monphosphate dehydrogenase and GMP synthetase (*guaAB*)
	6.97	0.41	306001	309000	1	ribonuclease PH (*rph*), tRNA synthetase (*gltX*)
	7.47	0.41	531001	534000	1	No predicted coding regions
	6.23	0.44	561001	564000	1	urease genes (*ureCE*)
	3.474.55	0.44-0.46	624001	630000	2	*rrnA*
	3.18-4.19	0.44-0.48	657001	663000	2	*rrnB*
	7.58	0.41	684001	687000	1	Putative biotin sulfoxide reductase (*bis* C)
	3.43-4.66	0.44-0.46	771001	777000	2	*rrnC*
	6.42	0.42	795001	798000	1	DNA polymerase III (*dnaE*)
	7.47	0.42	840001	843000	1	ribosomal proteins (*rpL23, rpL2, rpS19, rpL22, rpS3, rpL16, rpL29, rpS17*)
	7.40	0.41	1086001	1089000	1	transketolase 1 (*tktA*); hypothetical protein
	5.31-7.49	0.45-0.48	1569001	1590000	7	Mu-like bacteriophage
	3.37-3.53	0.45-0.47	1815001	1821000	2	*rrnD*
M. jannaschii	9.58	0.39	75001	78000	1	methyl coenzyme M reductase II (*mrtABG*)
	6.01-11.96	0.48-0.63	153001	159000	2	*rrnA*
	7.71-12.24	0.43-0.63	636001	645000	3	*rrnB*
	9.85-10.19	0.38-0.39	768001	774000	2	methyl coenzyme M reductase (*mcrABCDG*)
	11.71	0.37	1107001	1110000	1	formylmethanofuran dehydrogenase (*fwdADFG*)
	9.01-9.67	0.39	1131001	1137000	2	methylviologen-reducing hydrogenase(*vhuAG*), polyferredoxin (*mvhB*) formylmethanofuran dehydrogenase (*fwdB*)
E. coli	3.00-3.05	0.49-0.52	222001	231000	1	*rrnH*
	3.25	0.50	924001	927000	1	ATP-dependent Clp protease ATP-binding subunit (*clpA*), initiation factor-(IF-1, *infA*), leucyl/phenylalanyl-tRNA-protein transferase (*aat*)
	3.54	0.49	1863001	1866000	1	hypothetical protein
	2.82-3.22	0.51-0.53	2724001	2730000	2	*rrnG* , heat shock protein (*clpB*)

Table 1. continued

	2.80-2.93	0.53	3420001	3426000	2	*rrnD*
	3.43	0.47	3450001	3453000	1	ribosomal protein (*rpsJ*), putative general secretion pathway (*yheD*) PinO protein (*pinO*)
	2.81-2.82	0.52	3939001	3945000	2	*rrnC*
	2.69-3.01	0.51-0.54	4032001	4038000	2	*rrnA*
	2.80-3.55	0.47-0.53	4161001	4173000	4	*rrnB* , UDP-N-acetylenolpyruvoylglucosamine reductase (*murB*), vitamin B12 receptor (*btuB*), pantothenate kinase (*coaA*), glutamate racemase (*murI*), biotin operon repressor and biotin-[acetyl-CoA carboxylase] synthetase (*birA*), bacteriophage lambda proteins
	2.72-3.24	0.51-0.53	4206001	4212000	2	*rrnE*
	3.51	0.48	4494001	4497000	1	prophage integrase (*intB*), hypothetical proteins, insertion sequence IS2K

Table 2. Subsections of *M. genitalium, M. pneumoniae, H. influenzae, M. jannaschii* and *E. coli* having the highest variances were sequentially ordered by genetic position. Gene identification are based on Fraser *et al.* [10], Himmelreich *et al.* [17, 18], Fleischmann *et al.* [9] and Bult *et al.* [11], respectively

			Location by base pair:		Number of	
Organism	Variance	GC content	Start	End	Combined Segments	Codes for:
M. genitalium	39.80	0.24	1	3000	1	DNA polymerase III (*dnaN*), heat shock protein (*dnaJ*)
	35.16-37.36	0.26-0.27	9001	15000	2	DNA polymerase III (*dnaE, dnaH*), methylene-tetrahydrofolate dehydrogenase (folD), hypothetical proteins, ribosomal protein S6 modification protein (rimK), thiophene and foran oxidizer (*tdhF*)
	35.61	0.27	354001	357000	1	high affinity transport system protein P37, ATP-binding protein P29, transport system permease protein P69
	41.17	0.26	420001	423000	1	nitrogen fixation protein (*nifS*), hypothetical proteins
	34.29	0.27	441001	444000	1	isoleucyl-tRNA synthetase (*ileS*), methylase homolog (*cspR*)
	32.98-47.94	0.25-0.29	459001	465000	2	methionyl-tRNA formyltransferase, hypothetical proteins ribonuclease III
	34.17-35.42	0.27	474001	477000	1	arginyl-tRNA synthetase (*argS*), hypothetical proteins
	39.83	0.27	489001	492000	1	hypothetical proteins, GTP-binding protein (*spg*)
	33.94-40.77	0.24-0.26	516001	522000	2	hypothetical proteins, ribosomal proteins (*rpS9, rpL13*)
	34.22	0.27	528001	531000	1	hypothetical proteins
	34.95	0.27	546001	549000	1	hypothetical GTP-binding protein, ribosomal protein (*rpL19*)
M. pneumoniae	21.41-24.89	0.31-0.32	60001	66000	2	not classified
	24.07-33.86	0.27-0.32	93001	99000	2	restriction enyzme (*hsdS*)
	21.85-32.04	0.31-0.32	135001	141000	2	spermidine/putrescine transport ATP-binding protein, phosphocarrier protein Hpr (*ptsH*)
	24.91-25.63	0.31	243001	249000	2	putative lipoprotein, PTS system mannitol-specific component IIA, mannitol-1-phosphate 5-dehydrogenase
	25.23-26.53	0.30-0.31	606001	612000	2	Type I restriction enzyme (*hsdRS*), 5-formyl tetrahydrofolate cyclo-ligase
	22.85	0.33	684001	687000	1	putative lipoprotein, repetitive DNA sequence REPMP1
H. influenzae	22.96	0.32	264001	267000	1	hypothetical proteins, arsenate reductase (*arsC*)
	23.90-32.48	0.27-0.30	918001	924000	2	hypothetical proteins, *lsg* locus hypothetical protein, glycosyl tranferase (*lgtD*)
	20.88	0.37	1119001	1122000	1	type III restriction-modification enzyme (*ECOP15*), hypothetical proteins
	20.89	0.30	1473001	1476000	1	phosphate regulon sensor and transcriptional regulator proteins (*phoR* & *phoB*)
	22.71	0.30	1479001	1482000	1	periplasmic phosphate-binding protein (*pstS*), ferritin like protein (*rsgA*), hypothetical protein, anthranilate synthase (*trpE*)
	22.25	0.32	1491001	1494000	1	HindIII restriction endonuclease (*hind* IIIR), hypothetical proteins, DNA polymerase III chi subunit (*holC*)
	20.67-28.2	0.33-0.33	1764001	1770000	2	*lsg* locus hypothetical proteins
M. jannaschii	34.75-36.62	0.26-0.27	120001	126000	2	type I restriction-modification enzymes, hypothetical protein, tamdem proteins A2, A3, B2, and B3
	35.06	0.26	798001	801000	1	hypothetical proteins
	43.05	0.24	999001	1002000	1	capsular polysaccharide biosynthesis protein M, hypothetical protein
	34.25-34.85	0.27-0.28	1155001	1161000	2	type I restriction enzyme
E. coli	11.07	0.58	171001	174000	1	ferrichrome-binding periplasmic protein (*fhuD*) ferrichrome transport protein (*fhuB*)
	13.22-13.79	0.63-0.65	282001	288000	2	hypothetical proteins
	10.01	0.59	522001	525000	1	RhsD protein precursor (*rhsD*), hypothetical protein
	10.36	0.37	567001	570000	1	phage proteins
	13.35	0.35	582001	585000	1	transcriptional regulatory protein (*appY*), protease precursor (*ompT*)

Table 2. continued

14.71	0.62	729001	732000	1	rhsC protein precursor (*rhsC*)
11.39	0.38	735001	738000	1	hypothetical protein (*ybfD*)
10.88	0.36	1209001	1212000	1	5-methylcytosine-specific restriction enzyme A (*mcrA*), DNA-invertase (*pin*), hypothetical proteins
10.16	0.38	1542001	1545000	1	nitrite extrusion protein (*narU*), hypothetical proteins (*yddG*)
10.50	0.39	1635001	1638000	1	hypothetical proteins
10.47	0.58	2070001	2073000	1	hypothetical proteins
12.40	0.35	2103001	2106000	1	hypothetical proteins (*yefEHG*), o-antigen polymerase (*rfc*)
10.05	0.56	2214001	2217000	1	hypothetical proteins (*yehWXY*)
11.79	0.36	2466001	2469000	1	hypothetical proteins
12.00	0.36	2781001	2784000	1	hypothetical proteins; cryptic prophage proteins CP4-57
11.11	0.58	2844001	2847000	1	formate hydrogenlyase subunits 3, 4 and 5 (*hycCDE*)
12.93-17.15	0.33-0.35	2985001	2994000	3	hypothetical proteins
14.56	0.35	3264001	3267000	1	catabolic threonine dehydratase (*tdcB*), hypothetical protein (*yhaBC*) tdcABC operon transcriptional activators (*tdcAR*)
10.95	0.37	3579001	3582000	1	hypothetical proteins (*yhhXZ*)
11.52	0.59	3612001	3615000	1	hypothetical proteins, nickel-binding and transport proteins (*nikAB*)
10.45	0.38	3630001	3633000	1	hypothetical proteins (*yhiLM*)
10.87	0.59	3759001	3762000	1	repetitive sequence responsible for duplication withing the chromosome (*rhsA*) hypothetical proteins
13.29-16.43	0.33-0.37	3795001	3804000	3	Lipopolysaccharide synthesis (*rfaCGLKYJIBSP*)
10.34	0.59	4509001	4512000	1	iron(III) dicitrate transport and permease proteins (*fecBCDE*)

Comparing sections from within the same genome provided information on the natural variation of microbial genomes (Fig. 2). In general, each genome consisted of a core of sections having similar GC and variance values. Even sections from the extrachromosomal elements of *M. jannaschii* have similar values to those of the core sections (Fig. 2). In addition, each genome possessed some sections which were notably dissimilar from those of the core, this phenomenon being particularly evident in large genomes, such as those of *E. coli* and *Synechocystis* sp. (Fig. 2). To determine the factors contributing to these dissimilarities, we examined sections of *M. genitalium, M. pneumoniae, H. influenzae, M. jannaschii* and *E. coli* genomes having the lowest and highest variances (Tables 1 and 2). It is possible that variance of the tetranucleotide frequencies is an index to the genetic stability of the section, with low variance sections being more genetically stable than high variance sections because all tetranucleotides are more or less consistently represented. Conversely, high variance sections, which contain strings of repeated oligonucleotides, may be more genetically unstable than low variance sections because these strings provide sites for deletions, insertions, transpositions, duplications and recombination of genetic material. Conserved sequences such as those encoding ribosomal RNA molecules, therefore, should be more genetically stable than other regions of the genome because variations in the structure of RNA may have a direct effect on cell viability. In contrast, sections of the genome which are involved in providing opportunities for genetic variation such as mutational 'hotspots' [19] should have high variances, this being a function of foreign DNA acquisition and/or the generation of new sequence through mutational or recombinational events. Regardless of genetic stability, sections with variances which are significantly different from the rest of the genome may represent DNA acquired from external sources by lateral transfer.

Sections encoding ribosomal RNA in all genomes had low variances of tetranucleotide frequencies (Table 1). Moreover, low variances occurred in sections encoding important proteins. For example, sections coding for ribosomal proteins, pyruvate kinase and dehydrogenase, RNA and DNA polymerase, and DNA repair systems had low variances (Table 1). In *E. coli*, low variances occurred in sections encoding proteins that degrade carbon starvation proteins (*clpAB*), initiate protein synthesis (*infA*), synthesize peptidoglycan (*murIB*) [20, 21], degrade amino-terminal residues and transfer specific amino acids to acceptor proteins (*aat*) [22], synthesize and retain biotin [23], and function as outer member receptors for the adsorption and transport of vitamin B12. However, sections of the *M. genitalium* genome which encode subunits of DNA polymerase III (*dnaE, dnaH* and *dnaN*) and heat shock proteins (*dnaJ*) had high variances (Table 2), indicating that the relationship between variance and functionality of molecules encoded by these sections is not clearly defined.

Genome sections with high variances contained genes encoding a variety of different proteins (Table 2). Interpreting the significance of high variances and functionality of genes encoded by sections of the *M. genitalium, M. pneumoniae, H. influenzae* and *M. jannaschii* genome is difficult, however, since these bacteria have not been as well studied as *E. coli.* Nonetheless, sections having high variances often contained genes coding for restriction/modification enzymes and hypothetical proteins (Table 2). It is possible that a majority of these sections represent horizontally transferred DNA sequences since in *E. coli*, high variance sections often had GC content and gene codon usages which were considerably different from that of the rest of the genome. For example, Rhs elements (*rhsACD*) have GC contents ranging from 0.59 to 0.62 mol, which is the upper limit for the *E. coli* genome (Fig. 2). Rhs elements also have high variances

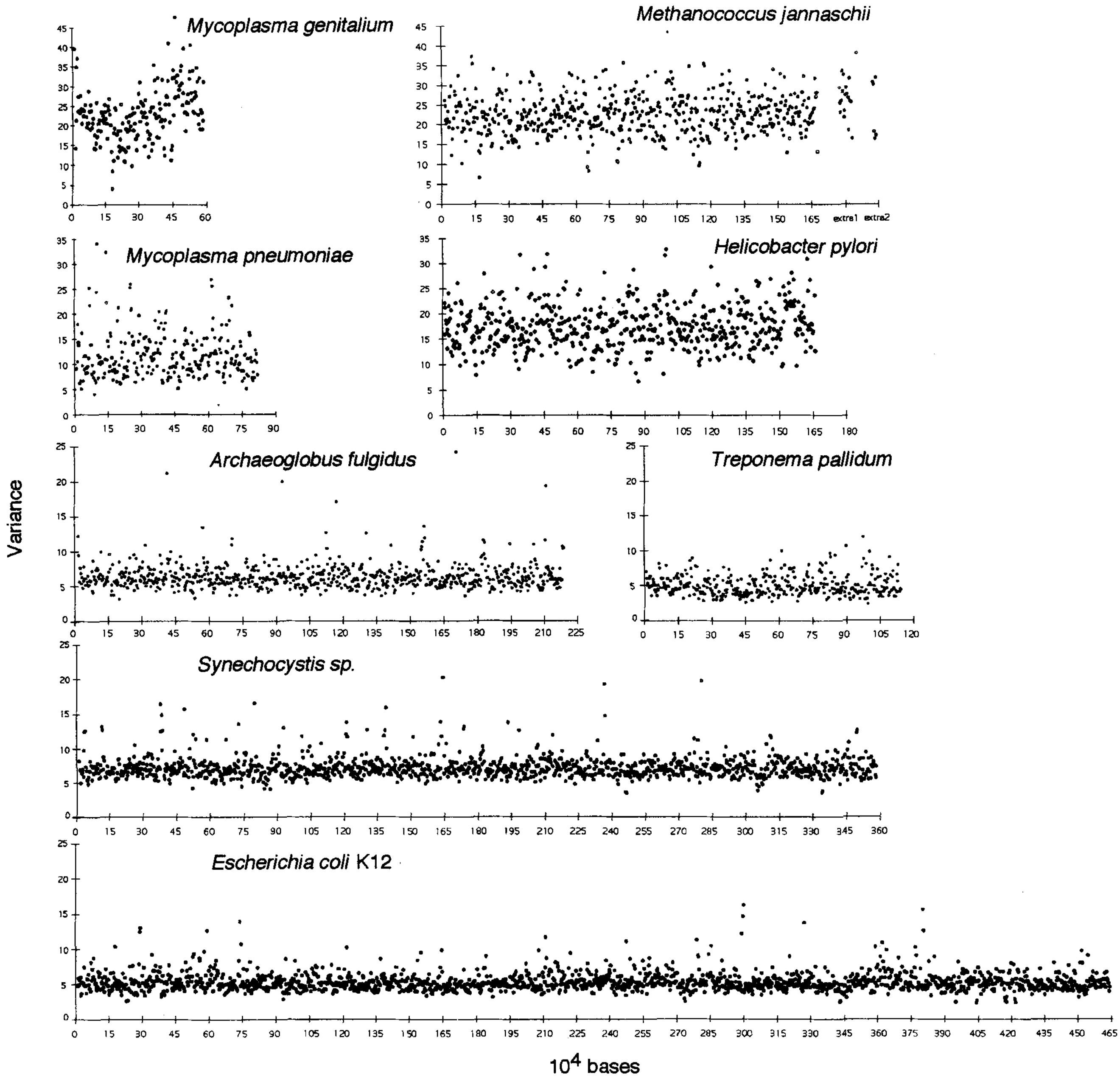

Figure 3. Variance of tetranucleotide frequencies as a function of genetic position. Sections of each genome were ordered by their respective genomic start site. Each datum point represents the variance of tetranucleotide frequencies obtained from a 3000 bp DNA sequence of chromosome or extrachromosomal element (extra 1 and extra 2).

which are considerably different from those of the rest of the *E. coli* genome (Table 2). These data, and data from previous studies [24], suggest that Rhs elements were derived from an organism possessing a high GC content. Furthermore, high variances in these elements may be attributed to genes encoding highly repeated peptide motifs. Other sections having high variances and high GC contents include those encoding ferrichrome (*fhu*)- and nickel (*nik*)-binding and transport proteins, and iron-dicitrate (*fec*) transport and permease proteins. The significance of this finding is difficult to ascertain because *fhu, nik*, and *fec* encode proteins which perform similar functions. Yet, a recent review suggests that these genes arose by gene duplication and divergence events which preceded evolutionary divergence of species [25]. If this is true, it is unlikely that these genes were laterally transferred to *E. coli*. Presumably, high variances in these sections can be attributed to strings of repeated oligonucleotides which are involved in protein structure and/or function.

Many of the high variance sections in the *E. coli* genome have low GC contents (Table 2). For example, a high variance section encoding a transcriptional regulatory protein (*appY*) and an outer membrane protease (*ompT*) (Table 2) has a GC content at the lower limits of the *E. coli* genome (Fig. 2). Previous studies have shown that the coding preferences and GC content of this section corresponds to a remnant lambdoid phage structure [26], and that this phage is responsible for transferring the *appY* gene from an unidentified bacterium to *E. coli* [27]. The *E. coli* section containing the genes methyltransferase (*mcrA*) and DNA invertase (*pin*) also has a high variance and low GC content (Table 2). This section is foreign DNA since these genes reside in the prophage e14 [28]. The *E. coli* sections containing the threonine dehydratase operon (*tdcABC*) is also regarded as foreign DNA since its codon usage and low GC content are different from that of *E. coli* [29, 30]. These examples demonstrate that high variances of the tetranucleotide frequencies in genome sections are often associated with foreign DNA.

3.2 Variance of tetranucleotide frequencies and location

The location of sections with low and high variances are shown in Fig. 3. Sections having low variances were often adjacent and consisted of 1–7 sections (Table 1, Figs. 1 and 3), whereas sections having high variances consisted of 1–3 sections (Table 2, Figs. 1 and 3). Genome regions of extreme variance occurred regularly throughout the microbial genomes, implying that the phenomenon probably occurs in all bacteria. The significance of the distribution and number of adjacent sections with similar variances is presently not clear. Further studies are needed to examine the regularity of the extreme variances and the presence/absence of specific oligonucleotides such as those involved in DNA replication and/or mismatch repair.

4 Concluding remarks

The computation strategy described in this study was employed to develop a method for visualizing sections of microbial genomes. Tetranucleotide frequencies provide information on the architecture of microbial genomes, identifying regions of the genome containing ribosomal RNA, ribosomal proteins, and bacteriophage. Variances of tetranucleotide frequencies can be used as an index to the architecture of microbial genomes since ribosomal RNA, ribosomal proteins, and bacteriophage have variances which are distinct from those of the median. Identification of sections that were different from those of the rest of the genome may provide information on the evolution and the plasticity of bacterial genomes. Differences in the tetranucleotide frequencies may be due to the mechanisms of intercellular genetic exchange and/or processes involved in maintaining intracellular genetic stability.

We thank Dr. Michele Doyle for assistance in providing key information, Anhminh Tran and Hien V. Wu for their computer skills, and Ann Miller, Drs. Robert Dunlap, Paula van Schie, Charles R. Lovell, and Madilyn Fletcher for proofreading the manuscript. Thanks to the anonymous reviewers for their insight. This research was supported in part by the University of California, Irvine through an allocation of computer resources on the convex C3840. Contribution number 1130 of the Belle W. Baruch Institute for Marine Biology and Coastal Research, University of South Carolina.

Received July 17, 1997

5 References

[1] Arber, W., *J. Mol. Evol.* 1995, *40*, 7–12.

[2] Bhagwat, A. S., McClelland, M., *Nucleic Acids Res.* 1992, *20*, 1663–1668.

[3] Merkl, R., Kroeger, M., Rice, P., Fritz, H.-J., *Nucleic Acids Res.* 1992, *20*, 1657–1662.

[4] Krawiec, S., Riley, M., *Microbiol. Rev.* 1990, *54*, 502–539.

[5] Karlin, S., Cardon, L. R., *Annu. Rev. Microbiol.* 1994, *48*, 619–654.

[6] Breslauer, K. J., Frank, R., Bloecker, H., Marky, L. A., *Proc. Natl. Acad. Sci. USA* 1986, *83*, 3746–3750.

[7] Calladine, C. R., Drew, H. R., *Understanding DNA*, Academic Press, San Diego 1992.

[8] Orstein, R., Rein, R., *Biopolymers* 1979, *18*, 1277–1291.

[9] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., *Science* 1995, *269*, 496–512.

[10] Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., *Science* 1995, *270*, 397–403.

[11] Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J.-F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L., Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H.-P., Faser, C. M., Smith, H. O., Woese, C. R., Venter, J. C., *Science* 1996, *273*, 1058–1073.

[12] Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., Tabata, S., *DNA Res.* 1996, *3*, 109–136.

[13] Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., Tabata, S., *DNA Res.* 1996, *3*, 185–209.

[14] Yamamoto, Y., Aiba, H., Baba, T., Hayashi, K., Inada, T., Isono, K., Itoh, T., Ki, *Nucleic Acids Res.* 1995, *23*, 2105–2119.

[15] Yamamoto, Y., Aiba, H., Baba, T., Hayashi, K., Inada, T., Isono, K., Itoh, T., Kimura, S., Kitagawa, M., Makino, K., Miki, T., Mitsuhashi, N., Mizobuchi, K., Mori, H., Nakade, S., Nakamura, Y., Nashimoto, H., Oshima, T., Oyama, S., Saito, N., Sampei, G., Satoh, Y., Sivasundaram, S., Tagami, H., Takahashi, H., Takeda, J., Takemoto, K., Uehara, K., Wada, C., Yamagata, S., Horiuchi, T., *DNA Res.* 1997, *4*, 91–113.

[16] Burland, V., Plunkett III, G., Sofia, H. J., Daniels, D. L., Blattner, F. R., *Nucleic Acids Res.* 1995, *23*, 2105–2119.

[17] Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C., Herrmann, R., *Nucleic Acids Res.* 1996, *24*, 4420–4449.

[18] Himmelreich, R., Plagens, H.,. Hilbert, H., Reiner, B., Herrmann, R., *Nucleic Acids Res.* 1997, *25*, 701–712.

[19] Smith, G. R., *Experientia*, 1994, *50*, 234–241.

[20] Dombrosky, P. M.,. Schmid, M. B., Young, K. D., *Arch. Microbiol.* 1994, *161*, 501–507.

[21] Lathrop, J. T., Wei, B. Y., Touchie, G. A., Kadner, R. J., *J. Bacteriol.* 1995, *177*, 6810–6819.

[22] Abramochkin, G., Shrader, T. E., *J. Biol. Chem.* 1995, *270*, 20621–20628.

[23] Nenortas, E., Beckett, D., *J. Biol. Chem.* 1996, *271*, 7557–7567.

[24] Hill, C. W., Sandt, C. H., Vlazny, D. A., *Mol. Microbiol.* 1994, *12*, 865–871.

[25] Tam, R., Saier, M. H., *Microbiol. Rev.* 1993, *57*, 320–346.

[26] Nakata, N., Tobe, T., Fukuda, I., Suzuki, T., Komatsu, K., Yoshikawa, M., Sasakawa, C., *Mol. Microbiol.* 1991, *9*, 459–468.

[27] Brondsted, L., Atlung, T., *J. Bacteriol.* 1996, *178*, 1556–1564.

[28] Hiom, K., Sedgwick, S. G., *J. Bacteriol.* 1991, *173*, 7368–7373.

[29] Medigue, C., Viari, A., Henaut, A., Danchin, A., *Microbiol. Rev.* 1993, *57*, 623–654.

[30] Riley, M., *Microbiol. Rev.* 1993, *57*, 862–952.