Think big – giant genes in bacteria

Oleg Reva[1] and Burkhard Tümmler[2] *

[1] Biochemistry Department, University of Pretoria, Lynnwood Road, Hillcrest, 0002 Pretoria, South Africa.

[2] Klinische Forschergruppe, OE6711, Medizinische Hochschule Hannover, Carl-Neuberg-Strasse 1, D-30625 Hannover, Germany.

Correspondence to *E-mail tuemmler.burkhard@mh-hannover.de;

Abstract

Long genes should be rare in archaea and eubacteria because of the demanding costs of time and resources for protein production. The search in 580 sequenced prokaryotic genomes, however, revealed 0.2% of all genes to be longer than 5 kb (absolute number: 3732 genes). Eighty giant bacterial genes of more than 20 kb in length were identified in 47 taxa that belong to the phyla Thermotogae (1), Chlorobi (3), Planctomycetes (1), Cyanobacteria (2), Firmicutes (7), Actinobacteria (9), Proteobacteria (23) or Euryarchaeota (1) (number of taxa in brackets). Giant genes are strain-specific, differ in their tetranucleotide usage from the bulk genome and occur preferentially in non-pathogenic environmental bacteria. The two longest bacterial genes known to date were detected in the green sulfur bacterium Chlorobium chlorochromatii CaD3 encoding proteins of 36 806 and 20 647 amino acids, being surpassed in length only by the human titin coding sequence. More than 90% of bacterial giant genes either encode a surface protein or a polyketide/non-ribosomal peptide synthetase. Most surface proteins are acidic, threonine-rich, lack cystein and harbour multiple amino acid repeats. Giant proteins increase bacterial fitness by the production of either weapons towards or shields against animate competitors or hostile environments.

Introduction

Human curiosity is often driven by the search for the extremes. Within the field of microbiology, one common theme has been the exploration of extreme habitats. Microbial ecosystems were investigated in hot Saharan (Chanal et al., 2006) or cold Antarctican deserts (Wierzchos et al., 2005), oceanic deep subsurface sediments (Newberry et al., 2004), ultradeep natural caves (Northup et al., 2003) or mines (Onstott et al., 2003), hot springs (Roeselers et al., 2007) or deep-sea hydrothermal vents (McCliment et al., 2006) and highly acidic biofilms (pH 0–1) (Macalady et al., 2007) or hypersaline brines (Bolhuis et al., 2004), to quote just some spectacular examples of microbial communities at the limits of life.

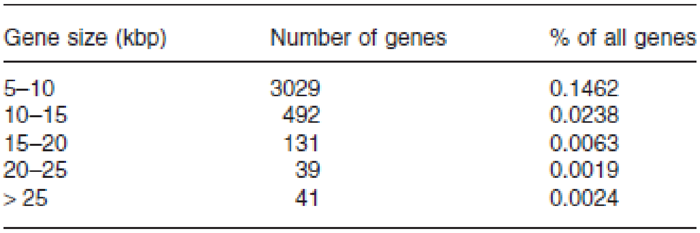
Here we report on the outcome of a further expedition to an extreme aspect of microbial life which was not accomplished by an outdoors adventure but by computer-based data mining of 580 totally sequenced prokaryotic genomes. We searched for the longest bacterial genes known to date. Close to 4000 genes of more than 5 kb in size were detected. Giant genes of more than 15 kb in length were subjected to in silico analysis. The majority of giant genes was annotated to either encode a cell surface protein or a non-ribosomal peptide or polyketide synthetase (NRPS, PKS). The giant cell surface proteins are typically acidic, serine- and threonine-rich and devoid of cystein, whereas NRPS and PKS do not substantially differ in their amino acid utilization from regular size proteins. Common features of the giant genes are repeats and an anomalous tetranucleotide usage that set them apart from the bulk of the genome. Although the current bacterial genome database is biased towards mammalian pathogens, most giant genes were found in non-pathogenic environmental bacteria.

Results and discussion

Distribution and frequency of long genes in bacterial genomes

The search in 580 totally sequenced prokaryotic genomes (533 bacterial genomes, 47 archaeal genomes) revealed 3732 annotated open reading frames (ORFs) longer than 5 kb in size corresponding to a frequency of 0.2% of 2071 329 ORFs (Table 1). In other words, one out of 500 ORFs is at least fivefold longer than the average protein encoding ORF with a match to a protein databases (Skovgaard et al., 2001). The reader may note that the absolute number of long ORFs could be slightly higher because many commonly used gene finders (such as Glimmer) tend to choose several smaller ORFs as compared with one extremely long ORF.

Table 1. Frequency of long genes in bacterial genomes.



Gene size (kbp)	Number of genes	% of all genes
5–10	3029	0.1462
10–15	492	0.0238
15–20	131	0.0063
20–25	39	0.0019
> 25	41	0.0024

Source: 580 totally sequenced microbial genomes deposited in the NCBI database; 1 October 2007.

Table S1 lists the distribution of long genes in 62 genomes that harbour at least one giant gene of more than 20 kb in length. The taxa Staphylococcus aureus, Staphylococcus epidermidis, Legionella pneumophila, Pseudomonas putida and Mycobacterium avium are represented by more than one strain. Giant genes were identified in 47 taxa that belong to the phyla Thermotogae (1), Chlorobi (3), Planctomycetes (1), Cyanobacteria (2), Firmicutes (7), Actinobacteria (9), Proteobacteria (23) or Euryarchaeota (1) (number of taxa in brackets). Of the 47 bacterial species, there are six human pathogens, one fish pathogen, one insect pathogen and one plant pathogen. Five of these nine species preferentially inhabit soil or aquatic habitats, only the pathogens S. aureus, S. epidermidis, Streptococcus pyogenes and Pseudomonas entomophila are typically associated with an animate host. Considering the current bias towards mammalian pathogens in the databases, it is evident that giant genes are not a typical feature of pathogens, but of environmental bacteria.

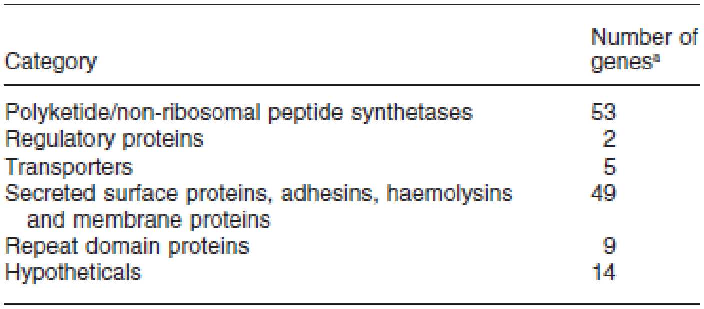
By the time of writing, the phyla Thermotogae, Chlorobi, Planctomycetes and Cyanobacteria were represented by only few sequenced strains in the database. Compared with the low number of sequenced genomes, the number of giant genes is high. The Planctomycetes strain Rhodopirellula baltica SH 1 T (Glöckner et al., 2003) and the Chlorobi strain Chlorobium chlorochromatii CaD3 each carry four giant genes in their chromosomes. Within each strain, the four genes are more homologous to themselves than to any other gene in the database, indicating that the genes encode functions not yet characterized in any other phyla of the pro- and eukaryotic kingdoms.

The two largest genes of the green sulfur bacterium C. chlorochromatii are the two longest bacterial genes known to date (Table S2). Their coding sequences encode proteins of 36 806 and 20 647 amino acids respectively. The coding sequence of these giant genes is only surpassed by the human titin gene whose 363 exons together code for 38 138 amino acid residues (4200 kDa) (Bang et al., 2001).

Functional categories of giant bacterial genes

Table S2 lists the genome coordinates of genes longer than 15 kb with a consistent annotation. Two functional categories with a similar number of entries dominate within the giant genes: surface proteins and NRPS/PKS (Table 2).

Table 2. Functional categories of giant proteins.



Category	Number of genes[a]
Polyketide/non-ribosomal peptide synthetases	53
Regulatory proteins	2
Transporters	5
Secreted surface proteins, adhesins, haemolysins and membrane proteins	49
Repeat domain proteins	9
Hypotheticals	14

a. A number of genes was assigned to more than one category. Data refers to entries in Table S2.

Non-ribosomal peptide synthetase and PKS are multi-enzymes with a typical signature of modules and domains (Fischbach and Walsh, 2006; Haynes and Challis, 2007). The sequence of the polypeptide or the structure of the polyketide can still only be roughly predicted from the gene sequence (Wilkinson and Micklefield, 2007), but the assembly of domains, modules and enzymatic sites is characteristic and allows an unequivocal identification of PKS and NRPS genes.

Typically, the synthesis of a peptide is accomplished by three to six genes within an operon (average length of a single NRPS gene: 5.9 kb, n = 494 genes), but in case of the giant PKS and NRPS genes listed in Table S2 the whole task is executed by a single multidomain protein. The giant genes probably evolved by gene fusion. In other words, a giant PKS/NRPS gene is not longer than a typical PKS or NRPS operon and the gene products synthesize peptides of the common length (n = 5–8). The largest known NRPS genes of more than 40 kb are encoded by the Nocardia farcinia, Myxococcus xanthus and Pseudomonas syringae B728a genomes (Table S2). The other large group of genes was annotated to encode giant extracellular surface proteins, cell surface receptors, haemolysins and membrane proteins (Table 2). The prediction of the individual protein topology was based on the presence of secretion signals or of transmembrane segment(s). The annotation is strongly supported by the features of the few functionally characterized gene products (see below) and the abundance of encoded acidic and polar amino acids found in all 49 genes of this category (see section Amino acid utilization in 'Results and discussion').

Wet lab data exist for only few of the giant genes discovered by bacterial genome sequencing projects. Of the 145 genes listed in Table S2, 16 encoded products have yet been characterized to some extent including the two orthologues of Rtx in two sequenced L. pneumophila strains, nine orthologues of the surface protein Ebh in nine sequenced S. aureus strains and three ORFs in a polyketide synthase operon. Thus in total an experimental analysis was performed on five independent entities, a polyketide synthase and four secreted proteins.

Streptomyces avermitilis produces the antiparasitic agent avermectin, a glycosylated pentacyclic macrolactone (Yoon et al., 2004). The large avermectin polyketide synthase genes aveA1–aveA2 and aveA3–aveA4 (Table S2) encode 12 sets of enzyme activities (modules) and a total of 55 active-site domains to synthesize the initial aglycon. In each reaction cycle, a beta-keto group is formed that is subsequently reduced and dehydrated to a double bond.

The secreted proteins have been characterized by comparison of phenotypes between wild-type strain and isogenic transposon or deletion mutants. Utilizing these genetic approaches the Rtx protein was shown to assist pathogenic activities of L. pneumophila to protozoa and mammalian monocytes such as adherence, entry, cytotoxicity, pore formation and intracellular growth (Cirillo et al., 2000; 2001; 2002). The large adhesion protein LapA of P. putida KT2440 is part of an ABC transporter operon. Transposon mutants in lapA were strongly impaired in the attachment to corn seeds and abiotic surfaces (polystyrene, polypropylene, borosilicate) (Espinosa-Urgel et al., 2000), suggesting that LapA is the major adhesin of P. putida KT2440. Ebh of S. aureus is also an adhesin and can bind to human fibronectin (Clarke et al., 2002).

The giant square halophilic archaeon Haloquadratum walsbyi (Bolhuis et al., 2004) dominates in NaCl-saturated aquatic ecosystems in which the salinity increases up to about 10 times the average seawater concentration. Haloquadratum walsbyi encodes the 9195 amino acids large halomucin (Hmu1) which is similar in amino acid sequence and domain organization to animal mucins (Bolhuis et al., 2006). Halomucin may confer an aqueous shield so that H. walsbyi can survive in low water activity environments. Hmu1 is yet the only giant bacterial gene that has experimentally been demonstrated to be transcribed in full length (Bolhuis et al., 2006).

Oligonucleotide usage

Tetranucleotide frequencies are measures of variability in bacterial genomes and carry a phylogenetic signal (Karlin et al., 1997; Abe et al., 2003; Pride et al., 2003; Teeling et al., 2004). Tetranucleotide frequencies are typically similar throughout the genome (Reva and Tümmler, 2004). Only a few regions exhibit an atypical oligonucleotide composition, indicating that this DNA has exposed to particular constraints other than those seen in the bulk of the genome (Reva and Tümmler, 2005).

Tetranucleotide usage of large genes (red dots in Fig. 1) was compared with that of the whole genome for the six species C. chlorochromatii, Synechococcus, N. farcinia, S. aureus, Photorabdus luminescens and R. baltica that were chosen as representatives for the phyla Chlorobi, Cyanobacteria, Actinobacteria, Firmucutes, Proteobacteria and Planctomycetes respectively. Values for a 8 kb sliding window in steps of 2 kb were compared with the global tetranucleotide usage of the whole chromosomes. Common measures of tetranucleotide usage are the variance of tetranucleotide frequencies OUV and the distance D. OUV is the variance between the empiric frequency and the null hypothesis of an equal frequency of all 256 tetranucleotides (Reva and Tümmler, 2004; 2005). OUV is primarily shaped by the local G+C content, and hence we calculated OUV:n1_4mer normalized for mononucleotide frequencies (Reva and Tümmler, 2004). The parameter distance D compares the rank order of tetranucleotide frequencies in two patterns (Reva and Tümmler, 2004), i.e. in this case the rank order in a 8 kb window compared with that of the whole genome (see Experimental procedures).

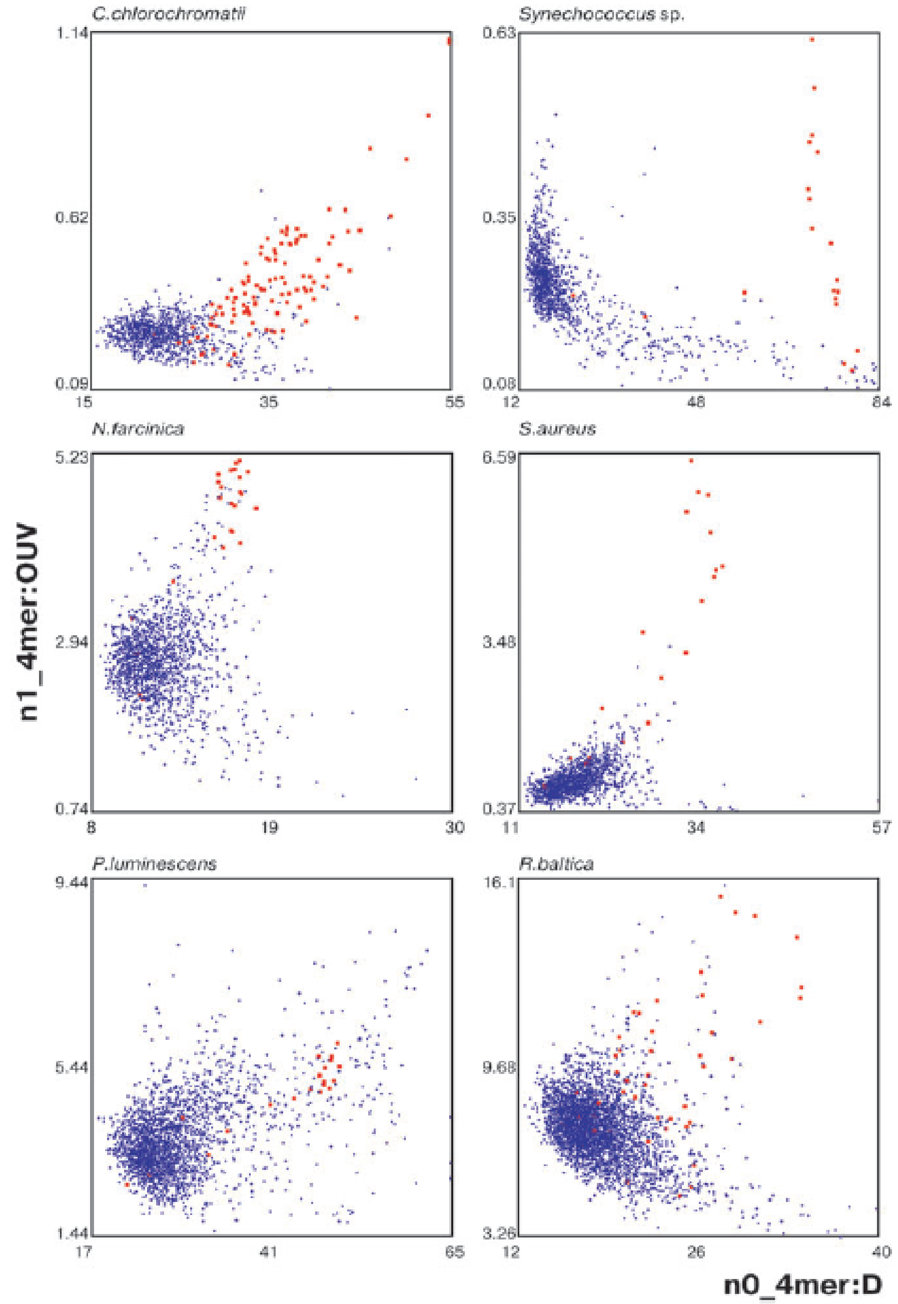


Fig. 1. Dot-plot presentation of 8 kb genomic fragments of the C. chlorochromatii CaD3, Syneochoccus sp., N. farcinia IFM10152, S. aureus MRSA252, P. luminescens TTO1 and R. baltica SH 1 T genomes. Fragments of 8 kbp were generated with sliding window steps of 2 kbp. Each dot represents the D:n0_4mer (%) and OUV:n1_4mer values of one fragment. The red dots visualize genomic fragments within giant genes (> 20 kb).

Figure 1 demonstrates that all large genes are endowed with an atypical tetranucleotide signature. The individual patterns, however, differ from species to species.

In C. chlorochromatii CaD3 more than 80% of the genomic segments with the most anomalous tetranucleotide usage can be attributed to the four largest genes of the genome. The individual n0_4mer:D and n1_4mer:OUV values of the 8 kb windows in the large genes show a linear positive correlation. In other words, a globally uncommon tetranucleotide usage is coupled with the selection for individual tetranucleotides, i.e. the more atypical the global usage is within a 8 kb segment, the stronger is the bias for a restricted set of tetranucleotides.

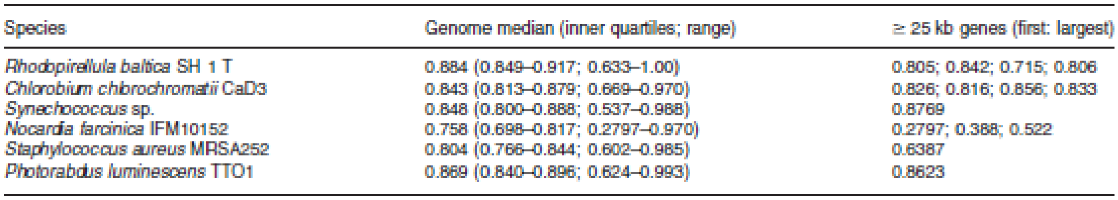
In the Synechococcus genome most 8 kb segments of the giant 32 kb gene share a similar anomalously high n0_4mer:D-value of about 75% that set them all apart from the bulk genome, but the individual n1_4mer:OUV values differ from window to window over the maximal range from 0.1 to 0.6. Each segment has an individual pattern of over- and under-represented tetranucleotides that is neither seen in any other gene nor within the 32 kb gene itself upstream or downstream of the 8 kb window. In other words, a gradient of tetranucleotide usage exists within the largest Synechococcus gene.

The six very long NRPS genes of N. farcinia cluster in a segment with high n0_4mer:D and maximal n1_4mer:OUV values implying a highly biased gene-specific selection of tetranucleotides not seen in any other segment of the genome. The extreme outliers of tetranucleotide usage within the sequenced S. aureus genomes reside all within the ebh gene. In case of P. luminescens and R. baltica, however, the large genes constitute only a part of the regions with atypical tetranucleotide usage. The giant genes of R. baltica SH 1 T cover a broad range from the most extreme outlier to inconspicuous n0_4mer:D and n1_4mer:OUV values. The tetranucleotide usage of the NRPS genes of P. luminescens clusters within an area that is close to the typical values of the host genome. In summary, the largest genes of bacteria exhibit an anomalous gene-specific signature of tetranucleotide usage.

Codon usage

Table 3 compares the codon usage of the giant genes (> 25 kb) with that of whole genomes. We selected the same six species that were exemplarily analysed in their oligonucleotide usage (OU, see above). No apparent trend can be seen for the codon usage of large genes. The huge genes of C. chlorochromatii, Syneochococcus and P. luminescens select the same codons as the average gene in their genomes. The four largest genes of R. baltica SH 1 T prefer less frequently used, but still rather common codons. In contrast, codon usage of ebh of S. aureus and that of the three longest genes of N. farcinica is strikingly different from that of the genome average. The 43 kb NRPS gene is the most extreme outlier with the minimal genomic codon index of 0.2797 of all genes in the N. farcinica genome, indicating that it was probably recently acquired by horizontal transfer and retained its preference for other tRNAs than the bulk of the N. farcinica genes.

Table 3. Genome codon index of giant genes compared with the complete host genome.



Species	Genome median (inner quartiles; range)	≥ 25 kb genes (first: largest)
Rhodopirellula baltica SH 1 T	0.884 (0.849–0.917; 0.633–1.00)	0.805; 0.842; 0.715; 0.806
Chlorobium chlorochromatii CaD3	0.843 (0.813–0.879; 0.669–0.970)	0.826; 0.816; 0.856; 0.833
Synechococcus sp.	0.848 (0.800–0.888; 0.537–0.988)	0.8769
Nocardia farcinica IFM10152	0.758 (0.698–0.817; 0.2797–0.970)	0.2797; 0.388; 0.522
Staphylococcus aureus MRSA252	0.804 (0.766–0.844; 0.602–0.985)	0.6387
Photorabdus luminescens TTO1	0.869 (0.840–0.896; 0.624–0.993)	0.8623

Amino acid utilization

Huge bacterial proteins with their average length of 10 000 amino acids have an amino acid utilization (Fig. 2, Table 4) that is different from the average bacterial protein with its mean length of 300 amino acids (Skovgaard et al., 2001).

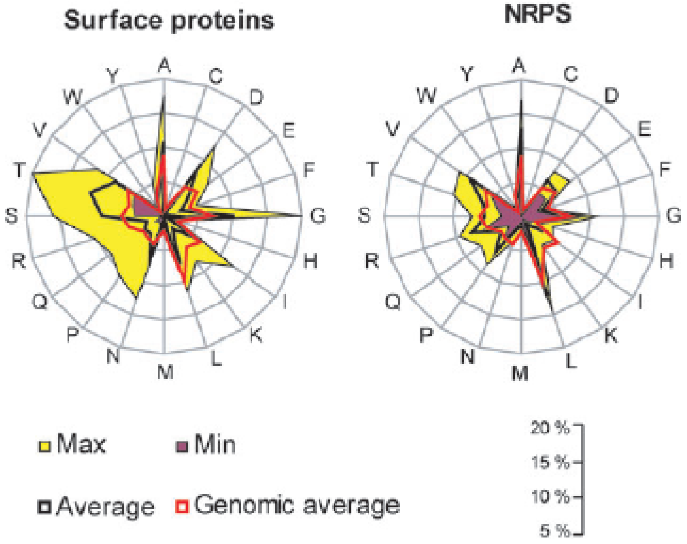
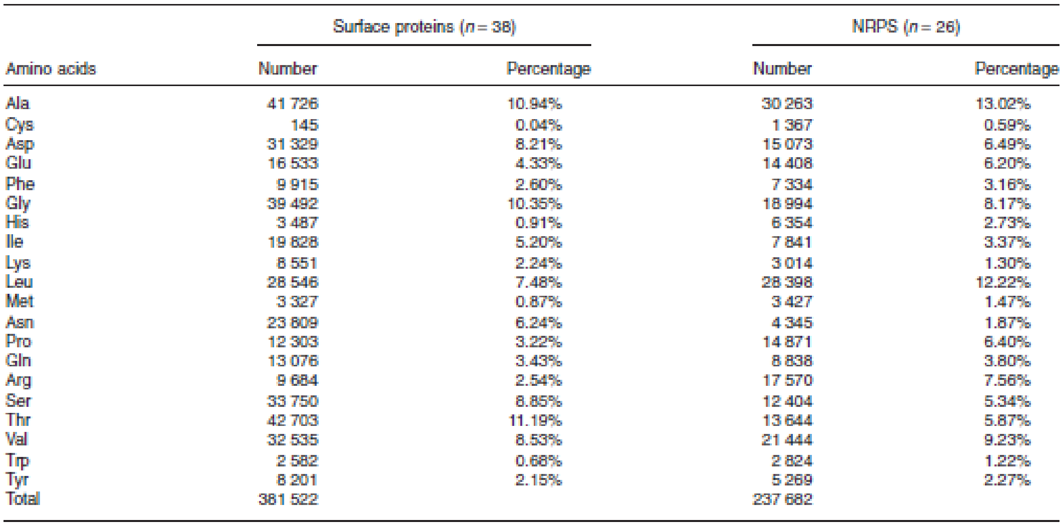


Fig. 2. Amino acid utilization of 49 giant putative surface proteins, adhesins, haemolysins and membrane proteins (left) and of 53 giant polyketide/non-ribosomal peptide synthetases (right) (classification by functional category according to Table 2). Frequencies of individual amino acids were counted for each gene and normalized. The black line indicates the average amino acid utilization within the groups of surface proteins and NRPS respectively. The pink and yellow values indicate the minimal and maximal amino acid frequencies. In other words, the yellow area shows the full range of values. The red line indicates the average usage of individual amino acids in 580 totally sequenced reference genomes.

Table 4. Amino acid utilization of the largest NRPS and surface proteins.



	Surface proteins (*n* = 38)		NRPS (*n* = 26)	
Amino acids	Number	Percentage	Number	Percentage
Ala	41 726	10.94%	30 263	13.02%
Cys	145	0.04%	1 367	0.59%
Asp	31 329	8.21%	15 073	6.49%
Glu	16 533	4.33%	14 408	6.20%
Phe	9 915	2.60%	7 334	3.16%
Gly	39 492	10.35%	18 994	8.17%
His	3 487	0.91%	6 354	2.73%
Ile	19 828	5.20%	7 841	3.37%
Lys	8 551	2.24%	3 014	1.30%
Leu	28 546	7.48%	28 398	12.22%
Met	3 327	0.87%	3 427	1.47%
Asn	23 809	6.24%	4 345	1.87%
Pro	12 303	3.22%	14 871	6.40%
Gln	13 076	3.43%	8 838	3.80%
Arg	9 684	2.54%	17 570	7.56%
Ser	33 750	8.85%	12 404	5.34%
Thr	42 703	11.19%	13 644	5.87%
Val	32 535	8.53%	21 444	9.23%
Trp	2 582	0.68%	2 824	1.22%
Tyr	8 201	2.15%	5 269	2.27%
Total	381 522		237 682	

The whole class of secreted cell surface proteins exhibits a characteristic signature of amino acid usage (Fig. 2, left panel). These proteins are rich in the polar aliphatic amino acids aspartate, glutamate, threonine, serine and asparagine, poor in lysine and arginine and devoid of cystein. Nine of the 49 proteins do not contain any cystein and 14 proteins harbour one to three cysteins. The surface proteins are acidic, hydrophilic and lack cystein. Hence they should avidly bind cations and water and should be highly flexible owing to the lack of any constraints by covalent disulfide bridges.

The utilization of amino acids by the NRPS and PKS is not so divergent from that of the bulk genome as it is seen for the cell surface proteins (Fig. 2, right panel). Non-ribosomal peptide synthetase and PKS are rather poor in lysine and rich in arginine and valine. The helix-breaking proline is abundant, indicating that long stretches of secondary structure are counterselected in these complex factories of multi-enzyme arrays.

Repeats

One common feature of most giant genes listed in Table S2 is the abundance of repeats. In case of the PKS and NRPS genes the repeat structure reflects the cycle of reactions catalysed by the encoded enzyme. For each cycle of non-ribosomal peptide synthesis, the attachment of an amino acid to a growing peptide chain requires modules for condensation, amino acid adenylation/ligation, thiolation and optionally modification such as methylation or epimerization followed by a thioesterase (Fischbach and Walsh, 2006; Haynes and Challis, 2007). The minimal module comprises six domains for substrate recognition and enzymatic activities.

Figure 3 shows the localization of repeats in genes other than PKS or NRPS, the majority of which encode cell surface proteins. Long repeats are most abundant in the longest genes of the P. putida, L. pneumophila and Psychrobacter arcticus genomes. The rtxA gene of L. pneumophila contains 26 copies of a direct 522 nucleotides long repeat that exhibit amino acid variations at only 10 of the 174 positions. LapA of P. putida KT2440 encodes two large threonine-rich repeats. The N-terminal repeat is made up of nine units each 100 amino acids in length. The C-terminal repeat consists of 29 units each 219 amino acids in length that are divided into two subgroups (no. 1–

7 and no. 8–29). The repeat region of the 20 kb gene of P. arcticus is assembled by two highly similar 102 and 101 amino acid repeats that only differ in the C-terminal sequence from each other (95-IPSVTTAD-102 versus 95-ASIVIAD-101). The sequence of repeats is organized as follows: N-terminal sequence – 101–$(102)_7$–101–$(102)_5$–101–$(102)_{11}$–101–102–101–$(102)_4$–101–102–101–$(102)_{10}$–101–$(102)_4$–101–102–101 (102, n = 44; 101, n = 10). The smwB gene of Synechococcus sp. contains 28 copies of imperfect repeats of 357 nt with no overlapping gaps and three copies of imperfect repeats of 666 nt. The Ebh protein of S. aureus harbours in its central region 44 imperfect repeats of 126 amino acids. The symbiotic betaproteobacterium Verminephrobacter eiseniae colonizes the excretory organs of earthworms. Its huge gene ORF1974 encodes a putative outer membrane protein that harbours 54 copies of an imperfect 111 amino acid repeat. In this context it is interesting to note that V. eiseniae harbours the largest locus of a clustered regularly interspaced short palindromic repeat consisting of 245 repeats on one side and 45 repeats on the other side of an IS element (Grissa et al., 2007). The direct repeat is 28 bp long and the average spacer length is 32 bp.

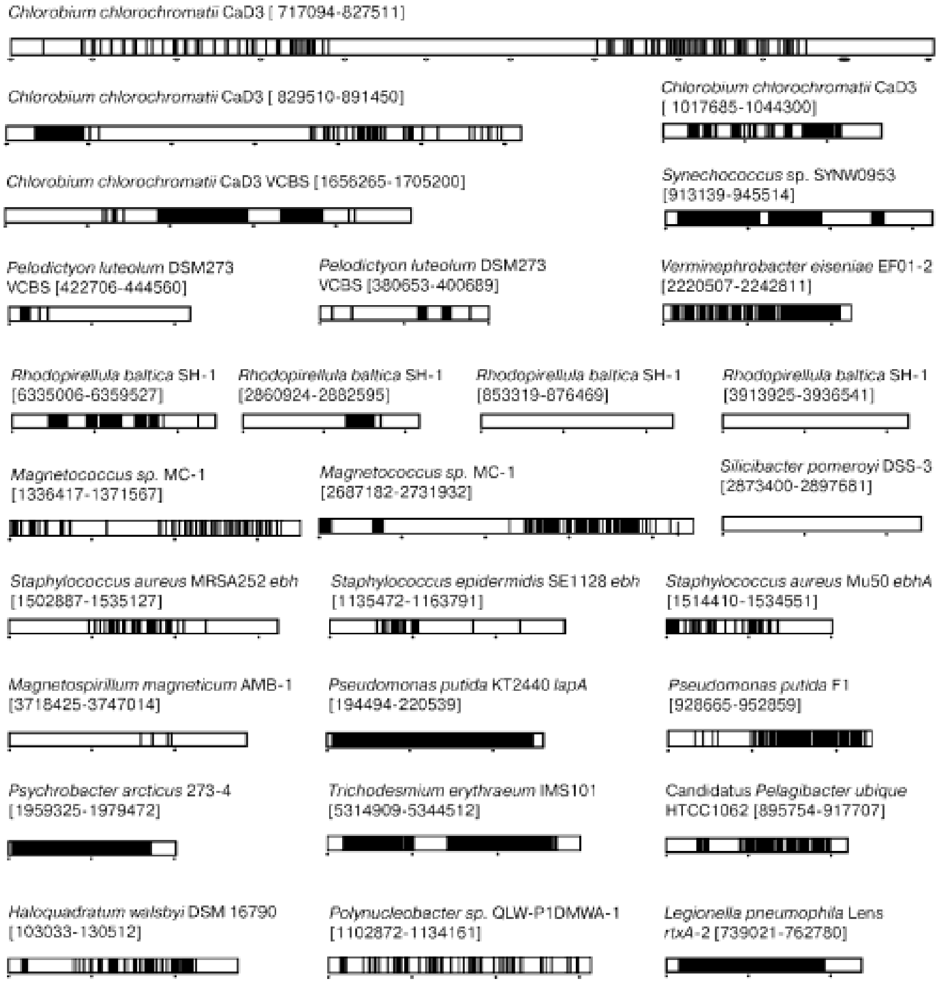

Chlorobium chlorochromatii CaD3 [717094-827511]
Chlorobium chlorochromatii CaD3 [829510-891450]
Chlorobium chlorochromatii CaD3 [1017685-1044300]
Chlorobium chlorochromatii CaD3 VCBS [1656265-1705200]
Synechococcus sp. SYNW0953 [913139-945514]
Pelodictyon luteolum DSM273 VCBS [422706-444560]
Pelodictyon luteolum DSM273 VCBS [380653-400689]
Verminephrobacter eiseniae EF01-2 [2220507-2242811]
Rhodopirellula baltica SH-1 [6335006-6359527]
Rhodopirellula baltica SH-1 [2860924-2882595]
Rhodopirellula baltica SH-1 [853319-876469]
Rhodopirellula baltica SH-1 [3913925-3936541]
Magnetococcus sp. MC-1 [1336417-1371567]
Magnetococcus sp. MC-1 [2687182-2731932]
Silicibacter pomeroyi DSS-3 [2873400-2897681]
Staphylococcus aureus MRSA252 *ebh* [1502887-1535127]
Staphylococcus epidermidis SE1128 *ebh* [1135472-1163791]
Staphylococcus aureus Mu50 *ebhA* [1514410-1534551]
Magnetospirillum magneticum AMB-1 [3718425-3747014]
Pseudomonas putida KT2440 *lapA* [194494-220539]
Pseudomonas putida F1 [928665-952859]
Psychrobacter arcticus 273-4 [1959325-1979472]
Trichodesmium erythraeum IMS101 [5314909-5344512]
Candidatus *Pelagibacter ubique* HTCC1062 [895754-917707]
Haloquadratum walsbyi DSM 16790 [103033-130512]
Polynucleobacter sp. QLW-P1DMWA-1 [1102872-1134161]
Legionella pneumophila Lens *rtxA*-2 [739021-762780]

Fig. 3. Repeat regions of giant bacterial genes (NRPS and PKS genes excluded). The filled bars indicate regions that carry repeats of more than 24 nt in length.

Long repeats in low copy number are present in the largest gene of the Candidatus Pelagibacter ubique genome. Eight sequences are duplicated (length: 97, 124, 167, 177, 250, 340, 378, 431 amino acids) and two sequences occur four times (length: 74, 75 amino acids). All repeats are rich in alanine, aspartate, serine and threonine (sum > 50%) and lack any cystein. In the largest bacterial gene known to date, C. chlorochromatii CaD3 (717094–827511), an 84 amino acids repeat occurs six times.

Another type of organization of repeats is realized in ORF 3470 of the genome of the marine filamentous cyanobacterium Trichodesmium erythraeum IMS101. This putative cell surface protein is composed to large extent of repetitive sequence motifs. It contains 27 different 8 to 20 amino acids long repeats that follow one after the other. An alternating sequence of two or three repeats was seen for five and one cases respectively. The 27 different amino acid sequence motifs designated in alphabetical order A to AA appear in the protein in the order 19 times motif A, 19 times motif B, 10 C, 20 D, 21 E, 19 F, 3 G, 16 H, 4 I, 28 J, 4 K, 2 L (23 M, 8 N) (12 O,19 P) (6 Q, 18 R), 26 S (18 T, 6 U) (9 V, 7 W, 7 X), 20 Y (eight times motif Z, eight times motif AA). The repeats differ substantially in length and hydrophobicity from each other. In summary, the repeat regions of ORF 3470 are composed of 27 sequential direct repeat modules.

All other repeats in the giant genes known to date are substantially smaller (< 100 nt). The repeat region of P. luteolum (422706–444560), for example, carries four copies each of 50, 55, 58 and 68 nt repeats. The giant gene in Polynucleobacter sp. harbours numerous short repeats, the longest of which are 21 copies of an imperfect 99 nt repeat. Halomucin of H. walsbyi, the largest archaeal protein known to date, contains 68 copies of a short VGGL peptide repeat.

In summary, different types of repeat organization are materialized in huge bacterial genes: multiple copies of few long repeats (rtxA, lapA, smwB), few copies of many long repeats [P. ubique (895754–917707)], sequential arrangement of numerous short direct repeats (T. erythraeum, ORF 3470) or many copies of a short repeat (halomucin).

Synopsis and conclusions

The search for giant bacterial genes in the database of completely sequenced genomes revealed that more than 90% of genes can be assigned to two functional categories: NRPS/PKS and surface proteins. Both classes of proteins can be easily distinguished from each other and other classes of proteins by characteristic structural properties. NRPS and PKS are highly organized modular multi-enzymes which combine the invariant sequential arrangement of modular blocks of enzymatic reaction centres with high substrate specificity (reviewed by Fischbach and Walsh, 2006; Haynes and Challis, 2007). The giant surface proteins are acidic, rich in threonine and other polar amino acids and contain no or few cysteins. Long hydrophilic amino acid repeats are common. Like mucins and collectins in mammals, these features endow a flexible protein structure and the abundant binding of water, ions and other substrates. Although giant surface proteins have yet not been analysed in their structural and biochemical properties in the wet lab, the evidence gained from their sequence features is compelling that their global function is to generate a micromilieu around the cell.

The modular organization of NRPS/PKS and surface proteins strongly suggests that both categories of giant genes evolved by repetitive gene duplications and/or gene fusions. However, as shown in Fig. 3, there are a few exceptions of giant genes that are devoid of repeats.

The synthesis of a giant protein is demanding in terms of energy, time and substrates. In this context it is worth mentioning that giant genes typically do not belong to the repertoire of the core genome, but rather are strain- or clone-specific features. The two large adhesins of P. putida KT2440, for example, have no homologues in P. putida F1, the giant Ebh proteins of six sequenced S. aureus strains exhibit strong sequence diversity, indicating diversifying selection and the PKS genes present in M. avium 104 are absent in strain k10 (Table S2). The latter observation is consistent with common knowledge that NRPS and PKS are strain-specific properties (Fischbach and Walsh, 2006; Haynes and Challis, 2007).

The list in Table S2 demonstrates that a bacterial strain harbours either giant NRPS/PKS or giant cell surface genes. The demands for energy and amino acid precursors seem to be so high that a cell can only afford one functional category of giant genes, either adhesins or NRPS/PKS. The P. putida F1 and P. entomophila chromosomes are the only exceptions which harbour genes of both functional categories. The mutual exclusion makes sense in light of the role of the giant proteins. NRPS and PKS produce secondary metabolites which confer antimicrobial, antifungal or antiparasitic activities. These compounds endow the bacterial host with weaponry against competitors for the same niche. Cell surface proteins have an opposite role: they are shields. They build a micromilieu around the cell to protect from hostile threats, allow adhesion to animate or inanimate surfaces and sense signals from the environment. In summary, NRPS/PKS and surface proteins are two sides of the same coin. They either allow attack against or confer protection from competitors.

Under optimal conditions a eubacterial cell can synthesize a chain of maximal 40 amino acids per second (Watson et al., 1987). Hence, the production of the largest encoded bacterial protein known to date, ORF (717094–827511) of C. clorochromatii CaD3, will require at least 15 min. Generation times for a bacterium may vary from 15 min to days or weeks. Consequently huge proteins will most likely not be synthesized during periods of fast growth, but rather during phases of slow or no growth. Free living environmental organisms typically have slow reproduction cycles and live in cycles of feast and famine situations. They have to deal with long periods of no or slow growth. These conditions may favour the production of huge proteins that increase the fitness of the bacteria to persist in their niche. The gain of fitness must be substantial so that the cell synthesizes protein at its limit of translational ability and then consumes further energy for the non-ribosomal synthesis of secondary metabolites or for the translocation of the protein to the extracellular space. The materialization of the information embedded in giant genes is associated with huge costs for the cell. This extreme demand for cellular resources puts the subject of 'Giant Genes' in line with the investigations of microbial life in extreme habitats that are at the heart of 'environmental microbiology' and the curiosity of its scientific community.

Experimental procedures

Search for long genes

Complete archaeal and bacterial genome sequences and the annotation tables were retrieved from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria) and stored in the local MySQL database. Open reading frames were queried by length.

Oligonucleotide usage (Reva and Tümmler, 2004; 2005)

Overlapping oligonucleotide words of a certain length l_w are counted in the sequence of L_{seq} nucleotides by shifting the window in steps of 1 nucleotide. The total word number (W_{total}) is $L_{seq} - l_w + 1$ in a linear sequence or $W_{total} = l_{seq}$ in a circular sequence. As $L_{seq} >> l_w$, $W_{total} \approx L_{seq}$ in all cases. For a given word length l_w, $N_w = 4^{l_w}$ different words are possible for a sequence of four letters A, T, G and C. The observed counts of words (C_0) are compared with the expected counts of words (C_e). Assuming the same distribution frequency for all words of a common length l_w irrespective of their composition and sequence, C_e matches the standard count number C_{n0}

$$C_e = C_{n0} = W_{total} \times N_w^{-1} \tag{1}$$

Correspondingly, if we normalize oligonucleotide usage (OU) by mononucleotide content using the zero-order Markov method (Almagor, 1983), C_e becomes C_{n1}.

The deviation Δ_w of observed from expected counts is given by

$$\Delta_w = (C_0 - C_e) \times C_{n0}^{-1} \tag{2}$$

In this study we used two types of tetranucleotide usage patterns; the non-normalized pattern n0_4mer and the pattern n1_4mer normalized by the zero-order Markov method.

The variance OUV of word deviations is calculated as follows:

$$OUV = \frac{\sum\limits_{w} \Delta_w^2}{N_w - 1} \tag{3}$$

For the comparison of sequences by OU patterns of the same type, the words in each sequence are ranked by Δ_w values calculated by applying Eq. 2. Rank numbers instead of word counts are used to simplify pattern comparison. The distance D between two patterns is calculated as the sum of absolute distances between ranks of identical words in patterns i and j as follows:

$$D(\%) = 100 \times \frac{\sum\limits_{w} |rank_{w,i} - rank_{w,j}|}{D_{\max}} \tag{4}$$

whereby

$$D_{\max} = \frac{N_w(N_w - 1)}{2} \tag{5}$$

$D_{\max}$ is the maximal distance that is theoretically possible between two patterns of l_w long words (Eq. 5). D is measured in percentage and may vary (theoretically) from 0% to 100%.

Genome codon index

The genome codon index (GCI) provides a quantitative measure to assess the synonymous codon bias of a particular gene compared with the average codon usage in the genome (Kiewitz et al., 2002). It is defined as the geometric mean of the RSCU values corresponding to each of the codons used in that gene, divided by the maximum possible GCI for a gene of the same amino acid composition

$$\mathrm{GCI} = \frac{\mathrm{GCI}_{\mathrm{obs}}}{\mathrm{GCI}_{\mathrm{max}}}$$

$$\mathrm{GCI}_{obs} = \left(\prod_{k=1}^{L} \mathrm{RSCU}_k \right)^{1/L}$$

$$\mathrm{GCI}_{\mathrm{max}} = \left(\prod_{k=1}^{L} \mathrm{RSCU}_{\mathrm{kgenome}} \right)^{1/L}$$

where RSCU_k is the RSCU value for the k^{th} codon in the gene, $\mathrm{RSCU}_{\mathrm{kgenome}}$ is the maximal genomic RSCU value for the amino acid encoded by the k^{th} codon in the gene, and L is the number of codons in the gene. The GCI was defined in analogy to the codon adaptation index (CAI) (Sharp and Li, 1987). In case of the CAI the RSCU values refer to a reference gene set, i.e. the CAI value indicates the relative adaptiveness of the codon usage of a particular gene to a set of highly expressed genes.

Amino acid utilization

Amino acid usage was calculated for all giant genes (> 20 kb) and all coding regions in completely sequenced bacterial genomes as annotated in the GenBank entries.

Repeats

Giant genes were searched for the over-representation of short amino sequence motifs (minimum: four amino acids) using an in-house Python program. Long amino acid sequence repeats were then assembled by overlapping nucleotide positions.

References

- Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T., and Ikemura, T. (2003) Informatics for unveiling hidden genome signatures. Genome Res 13: 693–702.
- Almagor, H. (1983) A Markov analysis of DNA sequences. J Theor Biol 104: 633–645.
- Bang, M.L., Centner, T., Fornoff, F., Geach, A.J., Gotthardt, M., McNabb, M., et al. (2001) The complete gene sequence of titin, expression of an unusual approximately 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. Circ Res 89: 1065–1072.
- Bolhuis, H., Poele, E.M., and Rodriguez-Valera, F. (2004) Isolation and cultivation of Walsby's square archaeon. Environ Microbiol 6: 1287–1291.
- Bolhuis, H., Palm, P., Wende, A., Falb, M., Rampp, M., Rodriguez-Valera, F., et al. (2006) The genome of the square archaeon Haloquadratum walsbyi: life at the limits of water activity. BMC Genomics 7: 169.

- Chanal, A., Chapon, V., Benzerara, K., Barakat, M., Christen, R., Achouak, W., et al. (2006) The desert of Tataouine: an extreme environment that hosts a wide diversity of microorganisms and radiotolerant bacteria. Environ Microbiol 8: 514–525.
- Cirillo, S.L., Lum, J., and Cirillo, J.D. (2000) Identification of novel loci involved in entry by Legionella pneumophila. Microbiology 146: 1345–1359.
- Cirillo, S.L., Bermudez, L.E., El-Etr, S.H., Duhamel, G.E., and Cirillo, J.D. (2001) Legionella pneumophila entry gene rtxA is involved in virulence. Infect Immun 69: 508–517.
- Cirillo, S.L., Yan, L., Littman, M., Samrakandi, M.M., and Cirillo, J.D. (2002) Role of the Legionella pneumophila rtxA gene in amoebae. Microbiology 148: 1667–1677.
- Clarke, S.R., Harris, L.G., Richards, R.G., and Foster, S.J. (2002) Analysis of Ebh, a 1.1-megadalton cell wall-associated fibronectin-binding protein of Staphylococcus aureus. Infect Immun 70: 6680–6687.
- Espinosa-Urgel, M., Salido, A., and Ramos, J.L. (2000) Genetic analysis of functions involved in adhesion of Pseudomonas putida to seeds. J Bacteriol 182: 2363–2369.
- Fischbach, M.A., and Walsh, C.T. (2006) Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. Chem Rev 106: 3468–3496.
- 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: 8298–8303.
- Grissa, I., Vergnaud, G., and Pourcel, C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35 (Web Server issue): W52–7.
- Haynes, S.W., and Challis, G.L. (2007) Non-linear enzymatic logic in natural product modular mega-synthases and – synthetases. Curr Opin Drug Discov Devel 10: 203–218.
- Karlin, S., Mrazek, J., and Campbell, A. (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179: 3899–3913.
- Kiewitz, C., Weinel, C., and Tümmler, B. (2002) Genome codon index of Pseudomonas aeruginosa: a codon index that utilizes whole genome sequence data. Genome Lett 2: 61–70.
- Macalady, J.L., Jones, D.S., and Lyon, E.H. (2007) Extremely acidic, pendulous cave wall biofilms from the Frasassi cave system, Italy. Environ Microbiol 9: 1402–1414.
- McCliment, E.A., Voglesonger, K.M., O'Day, P.A., Dunn, E.E., Holloway, J.R., and Cary, S.C. (2006) Colonization of nascent, deep-sea hydrothermal vents by a novel Archaeal and Nanoarchaeal assemblage. Environ Microbiol 8: 114–125.
- Newberry, C.J., Webster, G., Cragg, B.A., Parkes, R.J., Weightman, A.J., and Fry, J.C. (2004) Diversity of prokaryotes and methanogenesis in deep subsurface sediments from the Nankai Trough, Ocean Drilling Program Leg 190. Environ Microbiol 6: 274–287.
- Northup, D.E., Barns, S.M., Yu, L.E., Spilde, M.N., Schelble, R.T., Dano, K.E., et al. (2003) Diverse microbial communities inhabiting ferromanganese deposits in Lechuguilla and Spider Caves. Environ Microbiol 5: 1071–1086.

- Onstott, T.C., Moser, D.P., Pfiffner, S.M., Fredrickson, J.K., Brockman, F.J., Phelps, T.J., et al. (2003) Indigenous and contaminant microbes in ultradeep mines. Environ Microbiol 5: 1168–1191.
- Pride, D.T., Meinersmann, R.J., Wassenaar, T.M., and Blaser, M.J. (2003) Evolutionary implications of microbial genome tetanucleotide frequency biases. Genome Res 13: 145–155.
- Reva, O.N., and Tümmler, B. (2004) Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. BMC Bioinformatics 5: 90.
- Reva, O.N., and Tümmler, B. (2005) Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. BMC Bioinformatics 6: 251.
- Roeselers, G., Norris, T.B., Castenholz, R.W., Rysgaard, S., Glud, R.N., Kuhl, M., et al. (2007) Diversity of phototrophic bacteria in microbial mats from Arctic hot springs (Greenland). Environ Microbiol 9: 26–38.
- Sharp, P.M., and Li, W.H. (1987) The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15: 1281–1295.
- Skovgaard, M., Jensen, L.J., Brunak, S., Ussery, D., and Krogh, A. (2001) On the total number of genes and their length distribution in complete microbial genomes. Trends Genet 17: 425–428.
- Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glöckner, F.O. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics 5: 163.
- Watson, J.D., Hopkins, N.H., Roberts, J.W., Steitz, J.-A., and Weiner, A.M. (1987) Molecular Biology of the Gene, Vol. I, 4th edn. Menlo Park, CA, USA: The Benjamin/Cummings Publishing.
- Wierzchos, J., Sancho, L.G., and Ascaso, C. (2005) Biomineralization of endolithic microbes in rocks from the McMurdo Dry Valleys of Antarctica: implications for microbial fossil formation and their detection. Environ Microbiol 7: 566–575.
- Wilkinson, B., and Micklefield, J. (2007) Mining and engineering natural-product biosynthetic pathways. Nat Chem Biol 3: 379–386.
- Yoon, Y.J., Kim, E.S., Hwang, Y.S., and Choi, C.Y. (2004) Avermectin: biochemical and molecular basis of its biosynthesis and regulation. Appl Microbiol Biotechnol 63: 626–634.