Environmental Microbiology (2008) **10**(3), 614–625 doi:10.1111/j.1462-2920.2007.01483.x

Global features of the *Alcanivorax borkumensis* SK2 genome

Oleg N. Reva,[1,3] Peter F. Hallin,[2] Hanni Willenbrock,[2] Thomas Sicheritz-Ponten,[2] Burkhard Tümmler[1] and David W. Ussery[2]

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

[2]*Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark.*

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

Summary

The global feature of the completely sequenced *Alcanivorax borkumensis* SK2 type strain chromosome is its symmetry and homogeneity. The origin and terminus of replication are located opposite to each other in the chromosome and are discerned with high signal to noise ratios by maximal oligonucleotide usage biases on the leading and lagging strand. Genomic DNA structure is rather uniform throughout the chromosome with respect to intrinsic curvature, position preference or base stacking energy. The orthologs and paralogs of *A. borkumensis* genes with the highest sequence homology were found in most cases among γ-Proteobacteria, with *Acinetobacter* and *P. aeruginosa* as closest relatives. *A. borkumensis* shares a similar oligonucleotide usage and promoter structure with the *Pseudomonadales.* A comparatively low number of only 18 genome islands with atypical oligonucleotide usage was detected in the *A. borkumensis* chromosome. The gene clusters that confer the assimilation of aliphatic hydrocarbons, are localized in two genome islands which were probably acquired from an ancestor of the *Yersinia* lineage, whereas the *alk* genes of *Pseudomonas putida* still exhibit the typical *Alcanivorax* oligonucleotide signature indicating a complex evolution of this major hydrocarbonoclastic trait.

Received 8 August, 2007; accepted 26 September, 2007. *For correspondence. E-mail tuemmler.burkhard@mh-hannover.de; Tel. (+49) 511 5322920; Fax (+49) 511 5326723.

Introduction

Alcanivorax borkumensis strain SK2 is a cosmopolitan oil-degrading oligotrophic marine γ-proteobacterium (Yakimov *et al*., 1998). The SK2 strain is the paradigm for hydrocarbonoclastic bacteria that are specialized for hydrocarbon degradation but have an otherwise highly restricted substrate spectrum, being capable of utilizing only a few organic acids such as pyruvate, but not simple sugars, for growth (Yakimov *et al*., 1998; Sabirova *et al*., 2006). *A. borkumensis* is present in low abundance in unpolluted environments, but it rapidly becomes the dominant bacterium in oil-polluted open ocean and coastal waters, where it can constitute 80–90% of the oil-degrading microbial community (Harayama *et al*., 1999; Kasai *et al*., 2001; 2002; Syutsubo *et al*., 2001; Röling *et al*., 2002; Hara *et al*., 2003; McKew *et al*., 2007a,b).

The genome of *A. borkumensis* was recently sequenced and annotated (Schneiker *et al*., 2006). In this paper, we perform a genome wide comparative genomics analysis and a detailed characterization of the global features of the *A. borkumensis* strain SK2 genome. This work on *A. borkumensis* strain SK2 aimed to visualize the prospective potential of genome linguistic approaches for functional and comparative analysis of bacterial genomes.

Results and discussion

DNA structure and highly expressed genes

The genome atlas (Fig. 1) shows a combination of some general informative properties of the chromosome. These are structural features (intrinsic curvature, stacking energy and position preference), repeat properties (global direct and inverted repeats) and the main base composition features (GC skew and percent AT). Stacking energy measures helix rigidity and position preference is a flexibility measure (Jensen *et al*., 1999; Pedersen *et al*., 2000). Regions that exhibit low position preference correlate with an enrichment of highly expressed genes (Dlakic *et al*., 2004; Willenbrock and Ussery, 2007). Examples in *A. borkumensis* are the *rrn* operons, the genes encoding ribosomal proteins and the gene cluster labelled *rpoC* on the atlas which among others encodes RNA polymerase subunits. Low position preference was found to correlate with high codon adaptation indices as the common

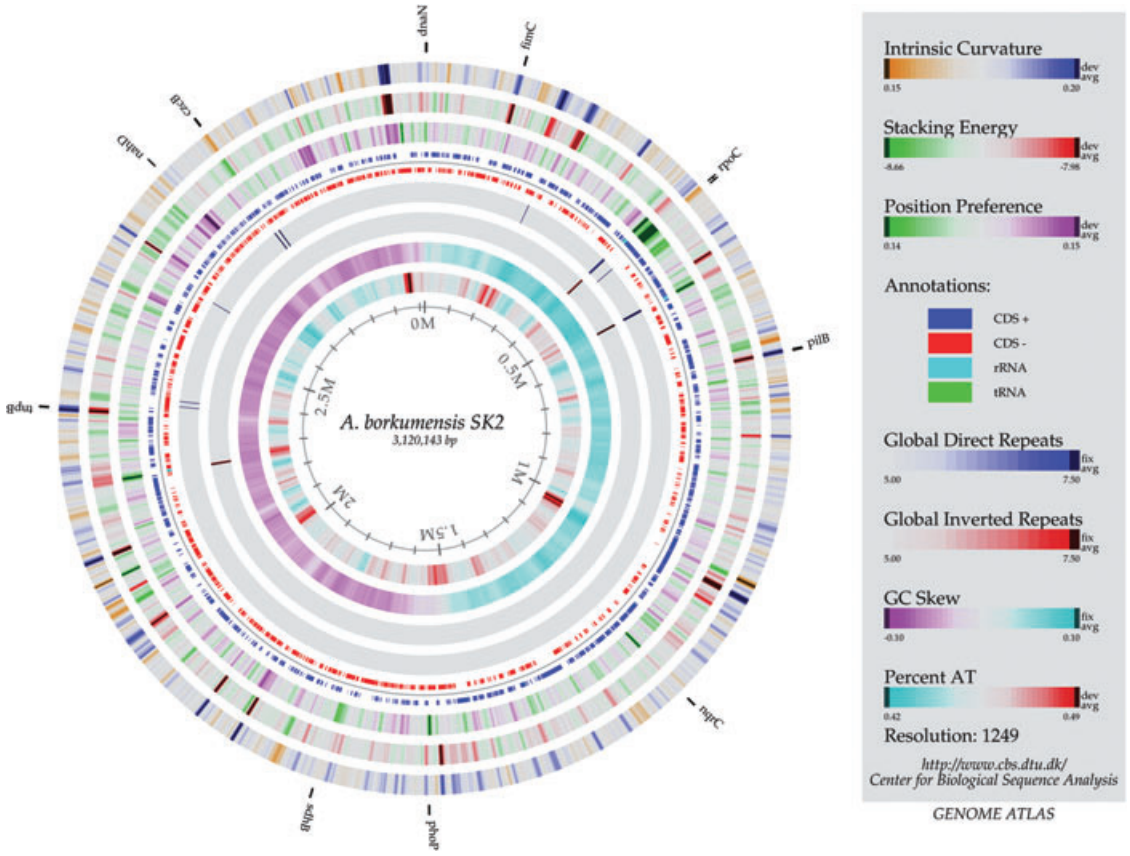

Fig. 1. Genome Atlas of *A. borkumensis* SK2 showing different structural parameters and the distribution of global repeats, GC skew and A + T contents. Colour intensity increases with the deviation from the average. Values close to the average are shaded very light grey; values with more than 3 standard deviations from the average are most strongly coloured.

measure for highly expressed genes (Willenbrock *et al.*, 2006) indicating that the local DNA structure is an important determinant of codon usage and gene expression. Moreover, intrinsic curvature is often encountered upstream of highly expressed genes (Skovgaard *et al.*, 2002) which correlates well with the fact that promoter DNA tends to be more curved than DNA in coding regions (Pedersen *et al.*, 2000).

The chromosome is rather homogeneous in all analysed structural features. The number of repeats is low, and the terminus of replication is opposite to the origin of replication as indicated by GC skew (Ussery *et al.*, 2002). The three rRNA operons organized in the order 16S–23S–5S are located in three areas with low position preference (green marks in the 3rd circle) and possible upstream regions with high intrinsic curvature (blue in the 1st circle) near 0.4 Mb – 0.5 Mbases (two regions) and 2.25 Mbases (one region).

Phylogenomics by sequence homology

The genome of *A. borkumensis* was compared with existing sequence information in other Proteobacteria by constructing phylogenetic trees for each amino acid sequence and organisms for which a similar gene existed. By extracting the phylogenomic information of the resulting 1919 phylogenetic trees a phylome atlas could be constructed (Fig. 2). In most cases the orthologs and paralogs with the highest sequence homology were found among γ-Proteobacteria. A substantial proportion of *A. borkumensis* genes had their closest homologues in α- and β-Proteobacteria, but no closest homologue was detected in δ- and ε-Proteobacteria. Inspection of the collected phylogenetic connections revealed that the most closely related organisms are *Acinetobacter* sp. and *Pseudomonas aeruginosa*, although in trees where both *Pseudomonas* and *Acinetobacter* are present, *A. borkumensis* tends to cluster more often with the latter one. No obvious horizontal gene transfers seem to have taken place. Regions around 350.000 and 450.000 are very 'pure' γ-proteobacteria regions.

Genome analysis of oligonucleotide usage

Oligonucleotide usage (OU) has been shown to be a genome specific signature (Pride *et al.*, 2003; Reva and

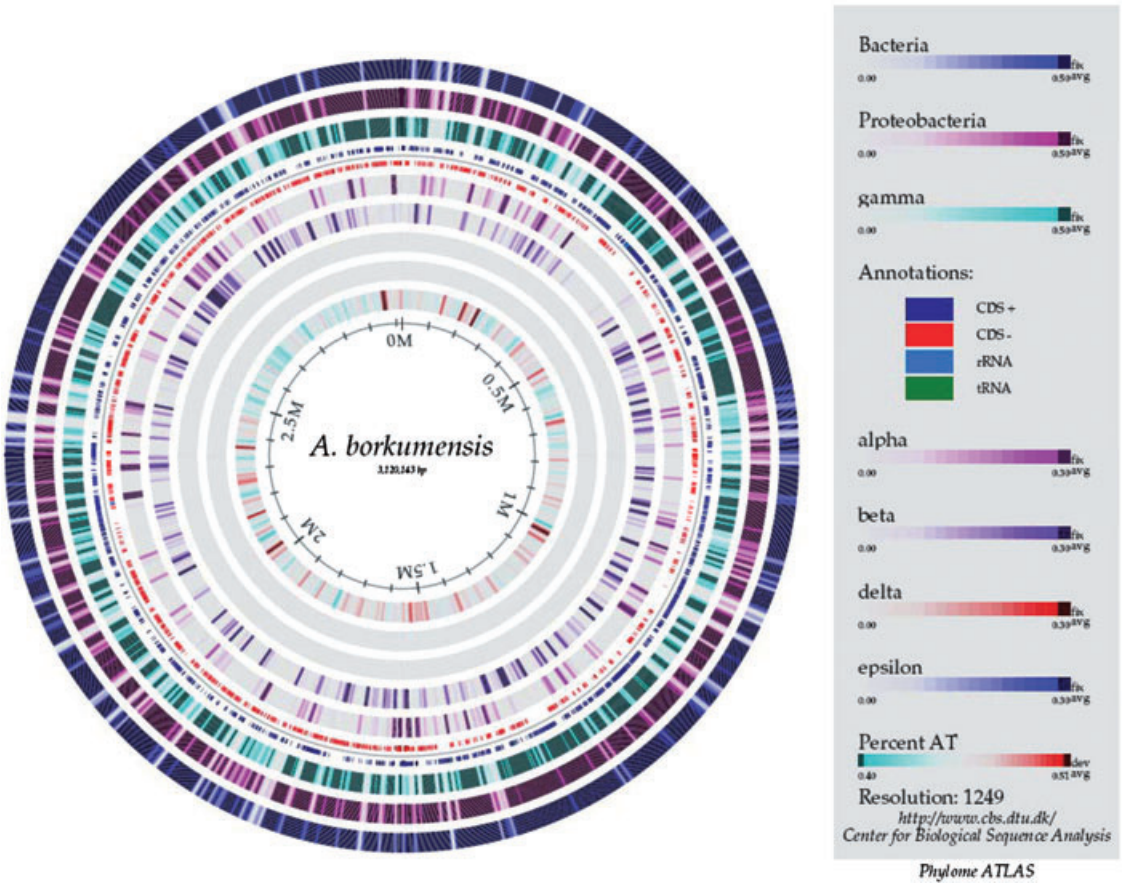

Fig. 2. Phylome Atlas of *A. borkumensis* SK2 genes indicating their closest bacterial homologues. Each of the concentric circles represents a taxonomic group as described in the figure legend on the right, with the outermost circle corresponding to the top-most feature, and the innermost circle corresponding to the bottom-most feature. Light bands indicate *A. borkumensis* SK2 genes with no homologue in the respective taxonomic group.

Tümmler, 2004). Genomic regions termed the 'core sequences' are characterized by OU patterns being similar to the global pattern of the chromosome. However, many loci with alternative OU patterns typically contribute to in total more than 10% of a bacterial genome. These loci with atypical OU patterns comprise heterogeneous subsets of parasitic and recent foreign DNA, ancient genes for ribosomal constituents (RNAs and proteins), multidomain genes and non-coding sequences with multiple tandem repeats (Reva and Tümmler, 2005). Hence laterally transferred gene islands can be reliably identified in complete genomes by their atypical oligonucleotide usage (Reva and Tümmler, 2005; Chen *et al.*, 2007; Klockgether *et al.*, 2007). Here, we focused on tetranucleotide usage (TU) parameters because the 256 different tetranucleotide words are optimal to differentiate bacterial genome sequences by the frequency and informativeness of the individual element. TU patterns represent the deviations of tetranucleotide word counts in a given sequence from an equiprobable distribution. Selection and counterselection of the oligonucleotide words are driven by their stereochemical properties such as base stacking energy, propeller twist angle, protein deformability, bendability and position preference (Reva and Tümmler, 2004). By permutation analysis, the 256 tetranucleotides were assigned to 39 equivalence classes each of which characterized by the same values for the five properties mentioned above (Baldi and Baisnee, 2000). Words of the same equivalence class tend to occur at similar frequencies in a nucleotide sequence (Reva and Tümmler, 2004). Oligonucleotide usage conservation reflects to some extent the phylogeny of microorganisms (Pride *et al.*, 2003; Teeling *et al.*, 2004).

Phylogenomics by tetranucleotide usage analysis

TU patterns were calculated for all sequenced genomes of γ-Proteobacteria. Four examples of TU patterns determined for *A. borkumensis* SK2, *Pseudomonas putida* KT2440, *Escherichia coli* K-12 and *Shewanella oneidensis* MR-1 are shown in Fig. 3. Tetranucleotide words were grouped by the equivalence classes and sorted in order of

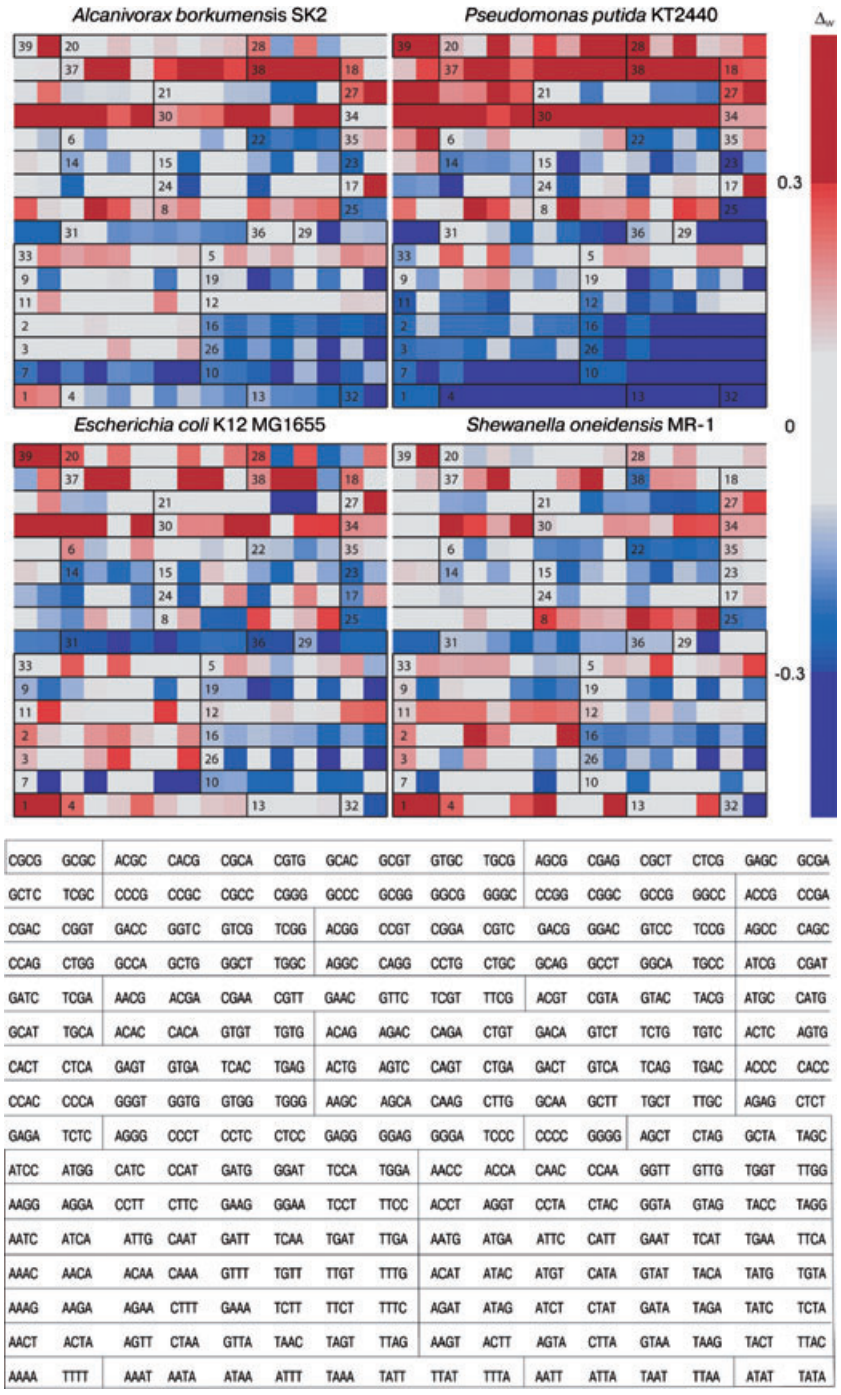

Fig. 3. Tetranucleotide usage patterns of *A. borkumensis* SK2, *P. putida* KT2440, *E. coli* K12 MG1655 and *S. oneidensis* MR-1. The deviation Δ_w of observed from expected counts is shown for all 256 tetranucleotide words (16 × 16 cells) by colour code (right bar). Tetranucleotides are grouped into 39 classes of equivalent structural features (Baldi and Baisnee, 2000) and sorted by decreasing base stacking energy row-by-row starting at the upper left corner (class 39). The words corresponding to the cells in colour plots are shown in the table in lower part of the figure.

decrease of the base stacking energy. Figure 4 visualizes the phylogenetic relationships differentiated by TU patterns of 29 γ-Proteobacterial taxa each of which represented by not more than a single sequenced strain.

A. borkumensis forms a cluster with *Pseudomonas*, *Methylococcus*, *Xanthomonas* and *Xylella* (Fig. 4). Despite the variation in GC-content, from 52 to 54% in *Xylella* and *Alcanivorax* to more than 65% in *Xanthomonas* and *Pseudomonas*, the TU patterns of these microorganisms are similar and separated from other γ-Proteobacteria. There is an abundance of GC-rich tetranucleotides with high base stacking energy in the sequence of *A. borkumensis* SK2 (words belonging to equivalence classes 37–39, 30 and 27) that is similar to the TU pattern of *P. putida* KT2440 (Fig. 3). Words of the AT-rich classes 7, 10, 13 and 32 are significantly underrepresented in both species. The major difference between TU patterns is the abundance of poly A and poly

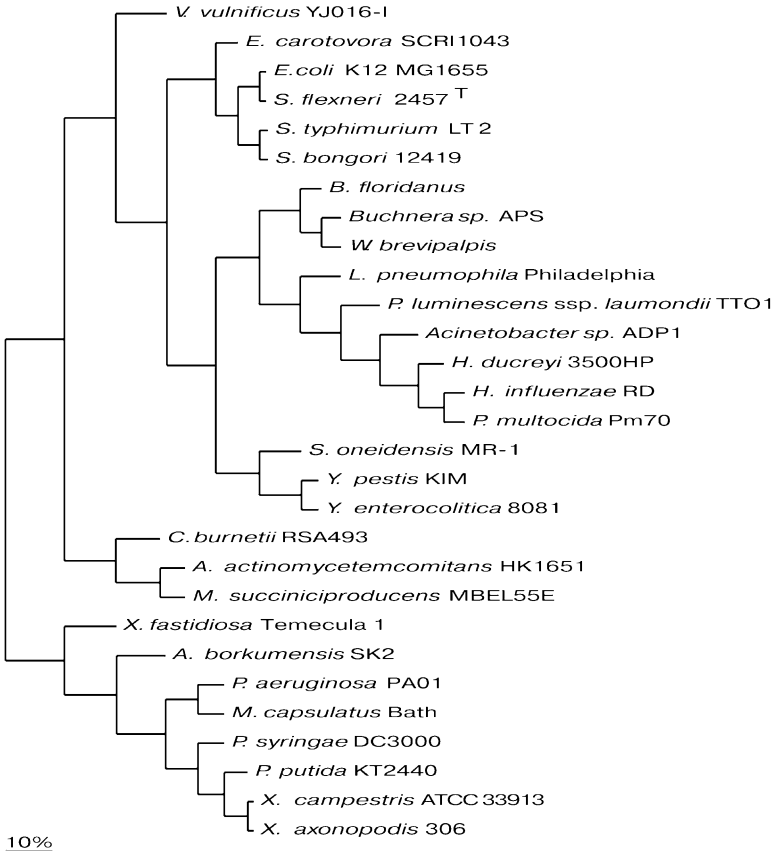

Fig. 4. Tree of the similarity of TU patterns of completely sequenced γ-*Proteobacteria* strains. Distance *D*-values (see Experimental procedures) between two TU patterns were calculated, and the tree was constructed from the distance matrix of all *D*-values by the minimum evolution neighbour-joining method (Saitou and Nei, 1987).

T stretches (words of class 1) in *A. borkumensis* in correspondence with its lower GC-content of 54.7%. Although *E. coli* and *S. oneidensis* share a similar GC contents with *A. borkumensis*, their tetranucleotides usage is different from *Alcanivorax*. The parity of GC with AT in the genome correlates with a balanced use of GC-rich and AT-rich words with high and low base stacking energy. In contrast, words with intermediate values of the base stacking energy (classes 25, 31, 36 and 29) are mostly underrepresented (Fig. 3). The data suggests that oligonucleotide usage drives GC-content and not vice versa. To give another example: the GC-rich words of class 21 are rare in all γ-Proteobacteria irrespectively of their GC-content (Fig. 3), but these words are overrepresented in α-Proteobacteria (*Agrobacterium*, *Bordetella*, *Caulobacter*, *Rhizobium*).

Anomalous local TU patterns in the A. borkumensis *genome*

A. borkumensis shares a common taxonomic group with *Pseudomonas*, *Methylococcus*, *Xanthomonas* and *Xylella*. Although the TU patterns are genome specific signatures, the oligonucleotide usage may vary locally in segments made up by horizontally acquired elements, phylogenetically ancient genes such as rRNAs or genes with peculiar codon usage (Reva and Tümmler, 2004; 2005). In other words, anomalous local TU patterns can be expected for the most recent and the most ancient genes. Local TU patterns were calculated in 8 kbp long overlapping sliding windows in steps of 2 kbp. Distances D between local and global TU patterns are shown in Fig. 5. The 18 regions with *D*-values above the 95% confidence interval are listed in Table 1.

Three clusters with anomalous *D*-values encode ribosomal RNAs that belong to the most ancient and conserved elements of all bacterial genomes. All the other 15 regions with atypical TU most likely were recently acquired, three of which contain transposase genes. In total 11 transposases were annotated in the *A. borkumensis* SK2 genome but for five of them no significant deviations of the local TU patterns were detected in adjacent regions. If inserted mobile elements had lost their mobility due to disruptive mutations, they undergo an amelioration process smoothing the differences in oligonucleotide usage between inserts and the host genome and thus cannot be detected by anomalous TU patterns anymore (Pride *et al*., 2003).

Five regions with high *D*-values (Fig. 5) only encode hypothetical proteins (Table 1). One further region contains genes of the type II secretion system and two regions encode type IV pili biogenesis proteins the latter

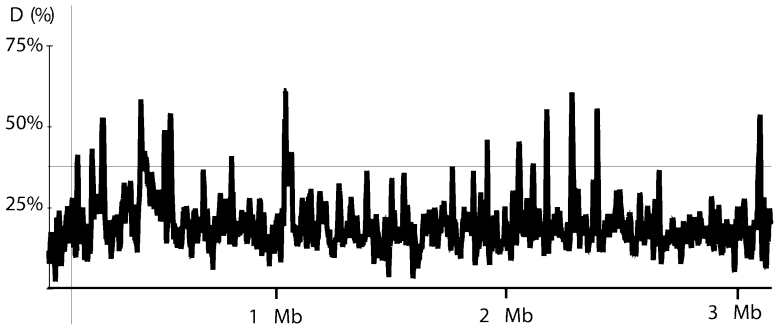

Fig. 5. Deviations of TU patterns in local regions of *A. borkumensis* SK2 chromosome. Local TU patterns were determined in 8 kbp sliding window in steps of 2 kbp. D, the distance betweeen local and chromosomal tetranucleotide patterns as defined in Experimental procedures, is plotted versus the coordinates of the chromosome starting from the putative replication origin.The upper border of the 95% confidence interval of *D*-values is shown by the horizontal line.

of which are known to have spread among proteobacteria by horizontal transfer with the original codon usage and GC content being retained (Spangenberg *et al*., 1997).

The most extended region with high *D*-values encodes a cluster of genes for glycosyltransferases and polysaccharide biosynthesis proteins (Abo_858-Abo_880: 1 018 000–1 060 000 bp) characterized by the second largest *D*-value and low GC-content (minimum 45% GC). The region terminates abruptly after Abo_880 at an Asn-tRNA gene. The TU pattern of the locus was compared with those of 177 sequenced bacterial chromosomes, 316 plasmids and 104 phages (Reva and Tümmler, 2004). The pattern was distant from all analysed sequences. The best hit of D = 34.9% was observed for the 5833 bp large bacteriophage Pf3 that infects *P. aeruginosa* harbouring the RP1 plasmid (Luiten *et al*., 1985). A stretch of 1550 bp upstream of the tRNA gene is 48% identical in nucleotide sequence with the Pf3 sequence (2344–4078 bp). According to this *in silico* finding we propose that this gene island was captured from a phage that typically target the 3′-end of a tRNA gene (Dobrindt *et al*., 2004).

The *alkB* genes encoding the degradation of alkanes which is the prominent name-giving feature of the taxon *Alcanivorax*, are located in two islands (Schneiker *et al*., 2006) with anomalous TU patterns (Table 1). Very close homologues were identified in marine bacteria and *Pseudomonas* species (Schneiker *et al*., 2006). The alkane hydroxylase gene cluster is widely distributed among hydrocarbon-utilizing γ-*Proteobacteria* due to its possible horizontal transfer (van Beilen *et al*., 2001; 2004). The role of these genes in the degradation of

Table 1. Chromosomal regions of *A. borkumensis* with atypical TU patterns.

Coordinates			
Left	Right	D[a] (%)	Annotation
126 000	140 000	42.20	Abo_114–120: *lysR* transcriptional regulator, haloacid dehalogenase hydrolase, *amiC* amidase, *gntR* transcriptional regulator, *alkB2* alkane monooxygenase, type I pili biogenesis proteins
190 000	198 000	40.47	Abo_172–178: *ilvD*-1 dihydroxy-acid dehydratase, conserved hypothetical proteins, long-chain-fatty-acid-CoA ligase, acyl-CoA dehydrogenases
234 000	245 000	47.95	Abo_209–214: conserved hypothetical proteins, transposase, type II secretion system proteins
400 000	408 000	49.42	first operon for rRNAs
502 000	510 000	46.26	Abo_439–446: *ispA* lipoprotein signal peptidase, *fkpB* peptidyl-prolyl *cis-trans* isomerase, *ispH* hydroxymethylbutenyl pyrophosphate reductase, type IV pili biogenesis proteins, conserved hypothetical proteins
526 000	534 000	43.41	second operon for rRNAs
670 000	678 000	40.29	Abo_581–583: type IV pili biogenesis proteins
792 000	800 000	43.00	Abo_2680–2681: hypothetical proteins
1 020 000	1 056 000	50.43	Abo_859–878: polysaccharide biosynthesis proteins
1 742 000	1 750 000	40.88	Abo_1439: periplasmic binding domain/transglycosylase SLTdomain fusion
1 892 000	1 900 000	46.32	Abo_2841–2847: hypothetical proteins
2 026 000	2 034 000	41.90	Abo_1668–1671: conserved hypothetical proteins, 3 transposases, siderophore biosynthesis protein, glycosyl transferase
2 088 000	2 096 000	40.65	Abo_ 1707–1708: conserved hypothetical proteins
2 146 000	2 154 000	47.05	Abo_2897–2905: *iscA* iron-binding protein *IscA*, metal-sulfur cluster biosynthetic enzyme, *sufE* Fe-S metabolism associated domain protein, *iscS* cysteine desulfurase, *rrf2* family protein, hypothetical proteins, SIR2-like transcriptional silencer
2 254 000	2 262 000	49.71	third operon for rRNAs
2 364 000	2 372 000	52.56	Abo_1942: penicillin-binding protein, hypothetical proteins, 2 transposases
2 632 000	2 640 000	40.17	Abo_2979–2984: hypothetical proteins
3 060 000	3 076 000	42.94	Abo_2516–3066: Na+/H+ antiporter, *alkS* alkB1GHJ regulator, *alkB1* alkane monooxygenase, *alkG* rubredoxin, *aldH* aldehyde dehydrogenase, hypothetical proteins

a. D, distance betweeen local and chromosomal TU patterns as defined in Experimental procedures.

short-chain *n*-alkanes by *A. borkumensis* SK2 and AP1 was experimentally proven (Smits *et al*., 2002; Hara *et al*., 2004; Sabirova *et al*., 2006). Interestingly, the two regions comprising of *alkS*, *alkB1*, *alkG* and *aldH* alkane-degradation genes and of *alkB2* and transcriptional regulators, respectively (Table 1), are as similar to each other in their TU patterns (D = 34.3%) as each of them is to *Yersinia pestis* (D = 32.2% for *alkB1*, D = 33.4% for *alkB2*), *Yersinia enterocolitica* (D = 29.5% for *alkB1*, D = 34.4% for *alkB2*) and *Shewanella oneidensis* MR-1 (D = 32.5% for *alkB1*, D = 42.4% for *alkB2*). This data suggests that the *alkB1* and *alkB2* genes were delivered to *A. borkumensis* from an ancestor of the *Yersinia* lineage. The AlkB1 amino acid sequences of *A. borkumensis* strains AP1 and SK2 are highly homologous to that of *P. putida* strains P1 and GPO1 (van Beilen *et al*., 2001; 2004; Smits *et al*., 2002; Hara *et al*., 2004), but their TU patterns are not that similar (D = 37.1). Surprisingly, the TU pattern of the *alkB* cluster of *P. putida* is significantly more similar with the global TU pattern of the whole *A. borkumensis* chromosome (16.7%, strain GPO1, 19%, strain P1), but more distant from the *P. putida* KT2440 chromosome (30.1% and 30.3%). *D*-values of 17 or 19% are within the first quartile (0–26%) far below the median value of 28.4% for local TU patterns of the *A. borkumensis* chromosome (Fig. 5) indicating that. the *P. putida alkB* gene behaves as if it were part of the *Alcanivorax* core genome. We note the striking phenomenon that there was converging evolution of the coding sequence of the catabolic *alk* transposon in *Alkanivorax* and *Pseudomonas*, but that the genes retained the oligonucleotide signature of their donors, most likely *Alkanivorax* for *Pseudomonas* and *Yersinia*-like organisms for *Alkanivorax.*

Origin of replication

The GC skew plotted in the seventh circle of the genome atlas (Fig. 1) reflects a general bias of purines towards the leading strand of DNA replication, however, it has almost no correlation to the structural properties of DNA (Skovgaard *et al*., 2002). The GC skew is often useful when locating the origin and terminus of replication (Jensen *et al*., 1999).

The circle is blue on the right side and purple on the left side. The two big gaps of colours in the top and in the bottom of the circle may be the origin and the terminus of replication. This may also be visualized more clearly in the origin plot (Fig. 6) (Worning *et al*., 2006). Here, the difference between hypothetical leading and lagging strand is plotted (red) for various positions on the chromosome. The peaks indicating maximal oligonucleotide skew correspond to origin and terminus. The terminus was identified as the peaks showing low G/C weighted strand bias at 1 502 000 bp position. The origin was identified as the other peak at 3 118 000 bp position. The signal to noise of 14.0 was among the top 10% of sequenced Proteobacteria, indicating a big difference between leading and lagging strand making the prediction of origin very confident.

Structural analysis of promoter regions

Structural features of the genomic DNA may indicate promoter regions, as promoters normally have high curvature, melt easily and are more rigid. The DNA structural parameters mentioned earlier (position preference, stacking energy, and intrinsic curvature) together with AT content and DNAse sensitivity (Brukner *et al*., 1995) were

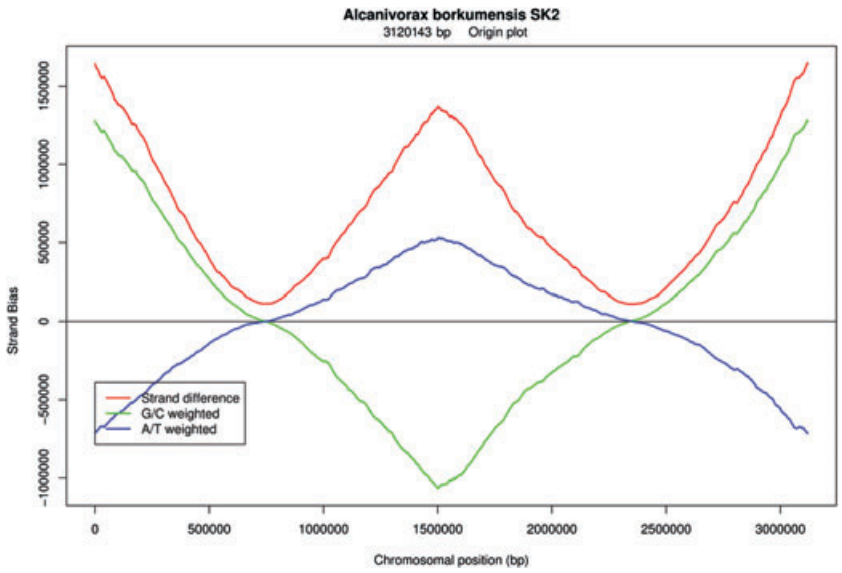

Fig. 6. Localization of the origin and the terminus of replication in the *A. borkumensis* SK2 chromosome derived from strand bias curves: the median oligonucleotide skew curve (red), the GC weighted median (green) and the AT weighted median (blue) (Worning *et al*., 2006).

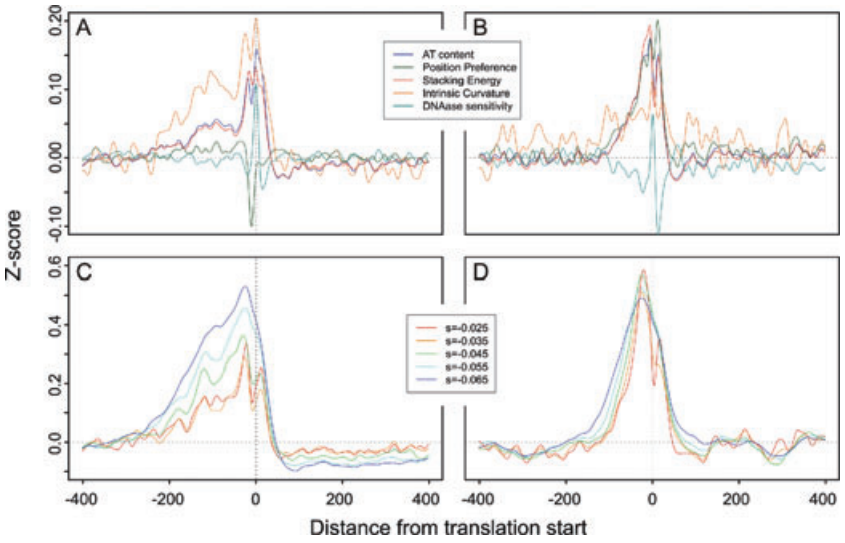

Fig. 7. Profile of structural properties of promoter regions (A and B) and probabilities of opening during stress-induced DNA duplex destabilization at various super-helical densities (C and D) in the *A. borkumensis* SK2 (A and C) and *Candidatus Pelagibacter ubique* HTCC1062 (B and D) chromosomes. Each annotated gene was aligned at the translation start site and the average values for the SIDD probabilities, AT-content, position preference, stacking energy, intrinsic curvature and DNase sensitivity were calculated at each position in the alignment. The values were subsequently converted into z-scores, using the average and standard deviation of the entire chromosome. Values are smoothed over a 5 bp window.

compiled into a structural profile of all upstream regions of *A. borkumensis* (see section Experimental procedures). The profile uses z-scores to measure how the average value of the properties vary from minus 400 bp to 400 bp around the translation start (Fig. 7). *A. borkumensis* has only a coding density of 87% causing a wider spacer of the intergenic region and this appears to give rise to a larger and wider peak of curvature, stacking energy and AT content (Fig. 7A). For comparison we also analysed the promoter profile of another ocean bacterium, *Candidatus Pelagibacter ubique* HTCC1062 (Giovannoni *et al*., 2005), an example of a highly streamlined genome with a coding density of 96%. Here we observed a much weaker curvature signal, and the distribution of stacking energy and AT content was more narrow and had higher maxima (Fig. 7B).

Next, the probability of opening during stress-induced DNA duplex destabilization was computed by using the program SIDD (Wang *et al*., 2004), covering five different values of the super-helical density s = {–0.025, –0.035, –0.045, –0.055, –0.065}. As super-coiling is being pushed, the probability of opening increases at lower super-helical densities in *A. borkumensis* (Fig. 7C). In contrast, a narrower SIDD profile that exhibits only minor dependence on super-helical density (Fig. 7D), was calculated for the *Candidatus Pelagibacter ubique* HTCC1062 genome.

The structural profile for the promoter regions of *A. borkumensis* was compared with that of closely related species as found above (see Fig. 4). Generally, it looked more like the promoter profile of members of the *Pseudomonadales* than the general comparison organism, *E. coli*. Moreover, the promoter profile was very different compared with the promoter profile of *X. fastidiosa* strains, even though they where very similar with regard to their TU profile (see Fig. 4). The promoter profiles for the above mentioned organisms may be found at our website (http://www.cbs.dtu.dk/services/GenomeAtlas/).

Amino acid and codon usage

We have examined the codon and amino acid usage of *A. borkumensis* and compared this with both the usage of bacteria in general and of 16 oceanic bacteria (Entrez project IDs 230, 10 645, 12 530, 13 233, 13 239, 13 282, 13 642, 13 643, 13 654, 13 655, 13 902, 13 906, 13 910, 13 911, 13 989, 15 660) Willenbrock *et al*., 2006). In Fig. 8, the codon usage plot of *A. borkumensis* is superimposed on the cumulative plot of all completely sequenced bacteria in public databases ($N = 518$, Fig. 8A) or of that of 16 oceanic bacteria (Fig. 8B). A few codons are differentially utilized in *A. borkumensis* (GUC, CUG), but all values are within the range of three standard deviations. In other words, codon usage of *A. borkumensis* resides within the typical range of eubacteria.

Interestingly, the sequenced oceanic bacteria share a very similar amino acid usage (Fig. 8D), whereas broad variations thereof were noted amongst all sequenced bacteria that represent the whole spectrum of habitats (Fig. 8C). *A. borkumensis* roughly follows the profile of the oceanic bacteria, although cysteine, tryptophan, leucine, proline, arginine, serine are under-utilized, and glutamic acid, lysine, phenylalanine, histidine, methionine, and tyrosine are over-utilized – all exceeding the three-standard deviation boundaries.

Conclusion

Inspection of the collected phylogenetic connections revealed that the most closely related organisms are *Acinetobacter* sp. and *Pseudomonas aeruginosa*,

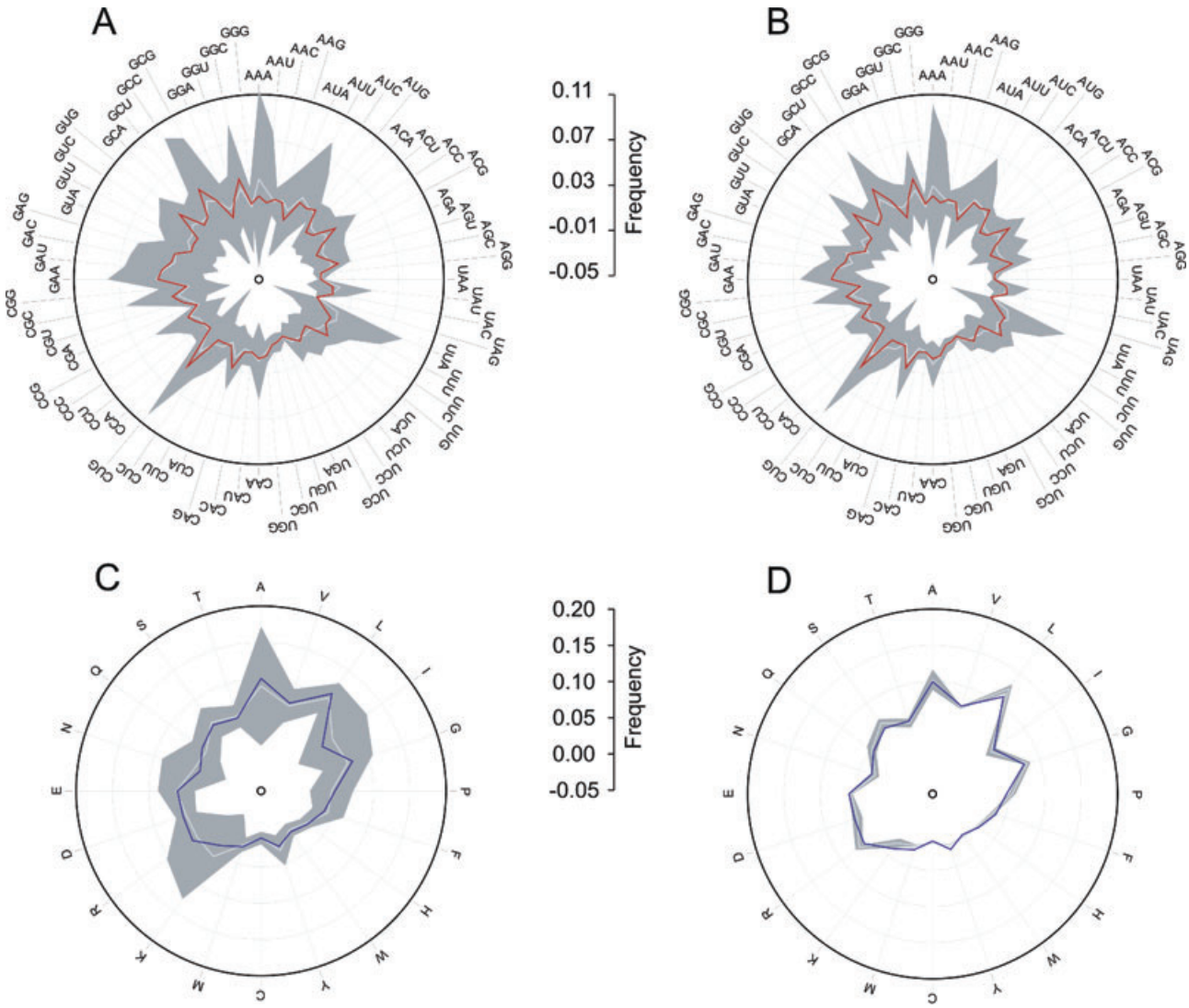

Fig. 8. Codon usage (A and B) and amino acid usage (C and D) of *A. borkumensis* SK2 compared with those of 518 completely sequenced bacteria (A and C) or compared with those of 16 sequenced oceanic bacteria. Frequencies of amino acids and codons were counted for each genome and normalized. Mean value (grey line) and three standard deviations (grey solid area) represent the global usage of individual codons (A and B) and amino acids (C and D) in the 518 (A and C) or 16 (B and D) reference genomes. The red line (A and B) shows the codon usage and the blue line (C and D) shows the amino acid usage of *A. borkumensis*.

although in trees where both *Pseudomonas* and *Acinetobacter* are present, *A. borkumensis* tends to cluster more often with the latter one.

The major structural feature of the *A. borkumensis* chromosome is its symmetry and homogeneity. The genome contains only very few regions with extraordinarily low or high curvature, position preference or base stacking energy. The chromosomal frame is symmetric: The origin and the terminus of replication are located opposite to each other in the chromosome and are clearly discerned by maxima of oligonucleotide usage biases between leading and lagging strand.

The genetic repertoire of *A. borkumensis* is most similar to that of *Acinetobacter* and *P. aeruginosa.* Moreover, *A. borkumensis* shares a similar oligonucleotide usage with the *Xanthomonadales* and *Pseudomonadales* indicating close phylogenetic relationships with these orders in accordance with 16S rDNA sequence relatedness (Schneiker *et al*., 2006). Amongst this subgroup of completely sequenced genomes, the *A. borkumensis* chromosome harbours the relatively lowest number of genome islands with atypical tetranucleotide usage. *P. putida* KT2440, for example, carries threefold more islands per Megabase in its chromosome (Weinel *et al*., 2002). Interestingly, one of the three enzyme systems that are upregulated in alkane-grown cells (Sabirova *et al*., 2006), the well-known *alkB1* cluster, is encoded by genome islands. The molecular evolution of the *alk* genes that are

encoded by a catabolic transposon (van Beilen *et al*., 2001) is remarkable: the *Alcanivorax* genes were probably acquired from the *Yersinia* lineage, whereas the *P. putida* genes exhibit the typical *Alcanivorax* tetranucleotide signature. Horizontal gene transfer was relevant to confer the – probably – most important metabolic trait to *A. borkumensis*, but otherwise the stable seawater habitat apparently did not favour the shuffling and exchange of genes with other taxa. Instead a symmetric and structurally homogeneous chromosome evolved that lacks numerous metabolic traits (Yakimov *et al*., 1998; Schneiker *et al*., 2006) found in their versatile *Pseudomonas* relatives which are endowed with twofold larger chromosomes (Stover *et al*., 2000; Nelson *et al*., 2002).

Experimental procedures

Genomic sequence

The comparative genomics analyses were based on the genomic sequence of *A. borkumensis* SK2 (Golyshin *et al*., 2003) and its annotation (Schneiker *et al*., 2006).

Atlas visualization

Atlases, developed in house, make it possible to visualize correlations between position dependent information contained within a chromosome. Circular graphical representations of the entire *A. borkumensis* genome were created using the atlas visualization tool, GeneWiz. Each feature, such as AT content is represented by a separate circle in the atlas. Typically, mean values are pictured in grey and extreme values are highlighted in a user defined colour (Pedersen *et al*., 2000).

Phylome atlas. For each amino acid sequence, phylogenetic trees were automatically constructed as described in Sicheritz-Ponten and Andersson (2001). The phylogenomic information of the resulting 1919 phylogenetic trees was extracted and analysed in the PyPhy system.

Genome atlas. The genome atlas is a combination of some general informative properties. These are some structural features (intrinsic curvature, stacking energy and position preference), some repeat properties (global direct and inverted repeats) and the main base composition features (GC skew and percent AT).

Intrinsic curvature was calculated using the CURVATURE software (Shpigelman *et al*., 1993). Stacking energy of a DNA segment was determined by the method of Ornstein and colleagues (1978). Position preference was based on a trinucleotide model that estimates the helix flexibility (Satchwell *et al*., 1986). Base composition is generally divided into AT content and GC skews. Both were calculated from the nucleotide sequence. Global direct and inverted repeats were found using variations of an algorithm that finds the highest degree of homology for a 15 bp repeat within a window of length 100 bp (Jensen *et al*., 1999).

Codon and amino acid usage

Codon and amino acid usage were calculated from all coding regions in the genome as annotated in the GenBank entries. The relative synonymous codon usage was calculated by comparing the codon distribution from a set of highly expressed genes with a background distribution estimated from the codon usage of all coding regions in the genome (Willenbrock *et al*., 2006). In order to identify a set of constitutively highly expressed genes in *A. borkumensis*, the reference set of 27 very highly expressed *Escherichia coli* genes originally compiled by Sharp and Li (1986) was aligned at the protein level against all genes annotated in the GenBank entry using BLASTP version 2.2.9 (Altschul *et al*., 1997). For each of these very highly expressed genes, the gene with the best alignment was added to a set of very highly expressed genes if it had an *E*-value below 10^{-6}.

TU patterns

Overlapping tetranucleotide words were counted in the bacterial nucleotide sequences by shifting the window in steps of 1 nucleotide. The total word number in a circular sequence equals to the sequence length. The observed counts of words (C_o) were compared with the expected counts of words (C_e). Assuming the same distribution frequency for all words irrespective of their composition and sequence mononucleotide content, C_e matches the ratio of the sequence length to the number of different tetranucleotide words N_w (256 for tetranucleotides).

The deviation Δ_w of observed from expected counts is given by

$$\Delta_w = (C_o - C_e) \times C_o^{-1}$$

For the comparison of sequences by TU patterns, the words in each sequence were ranked by Δ_w values. Rank numbers instead of word counts were used to simplify pattern comparison and to remove sequence length bias.

The distance D between two patterns was calculated as the sum of absolute distances between ranks of identical words in patterns i and j as follows and expressed as a percent of the possible maximal distance:

$$D(\%) = 100 \times \frac{\sum_w |\text{rank}_{w,i} - \text{rank}_{w,i}|}{D_{max}}$$

where

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

D_{max} is the maximal distance that is theoretically possible between two patterns. For TU patterns N_w is 256. For more information about methods of oligonucleotide usage statistics see Reva and Tümmler (2004; 2005).

Origin plot

The origin plot was constructed as described in Worning and colleagues (2006). In brief, the difference between a hypothetical leading and lagging strand is plotted for various positions on the chromosome. The frequencies of all

oligonucleotides from 2-mers to 8-mers on the leading and lagging strands in a 60% window are counted and the information content was calculated and summarized over all oligos for every putative origin. The G/C and A/T weighted strand bias were included to distinguish between origin and terminus.

Structural profile of the promoter region

Each annotated gene was aligned at the translation start site and the average values for five DNA structural features (AT content, position preference, stacking energy, intrinsic curvature, DNase sensitivity; see chapter on Genome Atlas) were calculated at each position in the alignment. The values was subsequently centered and scaled and smoothed within a 5 bp window using Gaussian smoothing.

Acknowledgements

The analysis has been performed within the frame of the 'Task Force Genome Linguistics' of the competence network 'Genome Research on Bacteria Relevant for Agriculture, Environment and Biotechnology' funded by the Federal Ministry of Education and Research (BMBF), Germany (Contracts 031U213D and 031U113D). We thank Peter Golyshin, Vitor Martins dos Santos and Kenneth N. Timmis, Helmhotz Center for Infection Research, Braunschweig, for stimulating discussions during the initiation of the study and Olaf Kaiser, Lehrstuhl für Genetik, Universität Bielefeld, for the provision of sequence data at an early stage of the sequencing project. O.R. has been a recipient of a postdoctoral stipend of the DFG-sponsored International Training Group 'Pseudomonas: Pathogenicity and Biotechnology'.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Baldi, P., and Baisnee, P.F. (2000) Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics* **16:** 865–889.

van Beilen, J.B., Panke, S., Lucchini, S., Franchini, A.G., Rothlisberger, M., and Witholt, B. (2001) Analysis of *Pseudomonas putida* alkane-degradation gene clusters and flanking insertion sequences: evolution and regulation of the *alk* genes. *Microbiology* **147:** 1621–1630.

van Beilen, J.B., Marin, M.M., Smits, T.H.M., Röthlisberger, M., Franchini, A.G., Witholt, B., and Rojo, F. (2004) Characterization of two alkane hydroxylase genes from the marine hydrocarbonoclastic bacterium *Alcanivorax borkumensis*. *Environ Microbiol* **6:** 264–273.

Brukner, I., Sanchez, R., Suck, D., and Pongor, S. (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J* **14:** 1812–1818.

Chen, X.-H., Koumoutsi, A., Scholz, R., Eisenreich, A., Schneider, K., Schneider, I., *et al.* (2007) Comparative analysis of the complete genome sequence of the plant growth promoting *Bacillus amyloliquefaciens* FZB42. *Nat Biotechnol* **25:** 1007–1014.

Dlakic, M., Ussery, D., and Brunak, S. (2004) DNA bendability and nucleosome positioning in transcriptional regulation. In *DNA Conformation and Transcription.* Ohyama, T. (ed.). Austin, TX: Landes Bioscience, pp. 198–211.

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* **2:** 414–424.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., *et al.* (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309:** 1242–1245.

Golyshin, P.N., Martins Dos Santos, V.A., Kaiser, O., Ferrer, M., Sabirova, Y.S., Lunsdorf, H., *et al.* (2003) Genome sequence completed of *Alcanivorax borkumensis*, a hydrocarbon-degrading bacterium that plays a global role in oil removal from marine systems. *J Biotechnol* **106:** 215–220.

Hara, A., Syutsubo, K., and Harayama, S. (2003) *Alcanivorax* which prevails in oil-contaminated seawater exhibits broad substrate specificity for alkane degradation. *Environ Microbiol* **5:** 746–753.

Hara, A., Baik, S.H., Syutsubo, K., Misawa, N., Smits, T.H., van Beilen, J.B., and Harayama, S. (2004) Cloning and functional analysis of *alkB* genes in *Alcanivorax borkumensis* SK2. *Environ Microbiol* **6:** 191–197.

Harayama, S., Kishira, H., Kasai, Y., and Shutsubo, K. (1999) Petroteum biodegradation in marine environments. *J Mol Microbiol Biotechnol* **1:** 63–70.

Jensen, L.J., Friis, C., and Ussery, D.W. (1999) Three views of microbial genomes. *Res Microbiol* **150:** 773–777.

Kasai, Y., Kishira, H., Sasaki, I., Syutsubo, K., Watanabe, K., and Harama, S. (2002) Prodominant growth of Alcanivorax strains in oil-contaminated and nutrient-supplemented sea water. *Environ Microbiol* **4:** 141–147.

Kasai, Y., Kishira, H., Syutsubo, K., and Harayama, S. (2001) Molecular detection of marine bacterial populations on beaches contaminated by the Nakhodka tanker oil-accident. *Environ Microbiol* **3:** 246–255.

Klockgether, J., Würdemann, D., Reva, O., Wiehlmann, L., and Tümmler, B. (2007) Diversity of the abundant pKLC102/PAGI-2 family of genomic islands in Pseudomonas aeruginosa. *J Bacteriol* **189:** 2443–2459.

Luiten, R.G., Putterman, D.G., Schoenmakers, J.G., Konings, R.N., and Day, L.A. (1985) Nucleotide sequence of the genome of Pf3, an IncP-1 plasmid-specific filamentous bacteriophage of *Pseudomonas aeruginosa*. *J Virol* **56:** 268–276.

McKew, B.A., Coulon, F., Osborn, A.M., Timmis, K.N., and McGenity, T.J. (2007a) Determining the identity and roles of oil-metabolizing marine bacteria from the Thames estuary, UK. *Environ Microbiol* **9:** 165–176.

McKew, B.A., Coulon, F., Yakimov, M.M., Denaro, R., Genovese, M., Smith, C.J., *et al.* (2007b) Efficacy of intervention strategies for bioremediation of crude oil in marine

systems and effects on indigenous hydrocarbonoclastic bacteria. *Environ Microbiol* **9:** 1562–1571.

Nelson, K.E., Weinel, C., Paulsen, I.T., Dodson, R.J., Hilbert, H., Martins dos Santos, V.A., *et al.* (2002) Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ Microbiol* **4:** 799–808.

Ornstein, R., Rein, R., Breen, D., and MacElroy, R. (1978) An optimized potential function for the calculation of nucleic acid interaction energies. *Biopolymers* **17:** 2341–2360.

Pedersen, A.G., Jensen, L.J., Brunak, S., Staerfeldt, H.H., and Ussery, D.W. (2000) A DNA structural atlas for *Escherichia coli*. *J Mol Biol* **299:** 907–930.

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.

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.

Röling, W.F., Milner, M.G., Jones, D.M., Lee, K., Daniel, F., Swannell, R.J., *et al.* (2002) Robust hydrocarbon degradation and dynamics of bacterial communities during nutrient – enhanced oil spill bioremediation. *Appl Environ Microbiol* **68:** 5537–5548.

Sabirova, J.S., Ferrer, M., Regenhardt, D., Timmis, K.N., and Golyshin, P.N. (2006) Proteomic insights into metabolic adaptations in *Alcanivorax borkumensis* induced by alkane utilization. *J Bacteriol* **188:** 3763–3773.

Saitou, N., and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406–425.

Satchwell, S.C., Drew, H.R., and Travers, A.A. (1986) Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* **191:** 659–675.

Schneiker, S., Martins dos Santos, V.A., Bartels, D., Bekel, T., Brecht, M., Buhrmester, J., *et al.* (2006) Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium *Alcanivorax borkumensis*. *Nat Biotechnol* **24:** 997–1004.

Sharp, P.M., and Li, W.H. (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res* **14:** 7737–7749.

Shpigelman, E.S., Trifonov, E.N., and Bolshoy, A. (1993) CURVATURE: software for the analysis of curved DNA. *Comput Appl Biosci* **9:** 435–440.

Sicheritz-Ponten, T., and Andersson, S.G. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res* **29:** 545–552.

Skovgaard, M., Jensen, L.J., Friis, C., Stærfeldt, H.H., Worning, P., Brunak, S., and Ussery, D.W. (2002) The atlas visualisation of genome-wide information. In *Methods in Microbiology*. Wren, B., and Dorrell, N. (eds). London, UK: Academic Press, pp. 49–63.

Smits, T.H., Balada, S.B., Witholt, B., and van Beilen, J.B. (2002) Functional analysis of alkane hydroxylases from gram-negative and gram-positive bacteria. *J Bacteriol* **184:** 1733–1742.

Spangenberg, C., Fislage, R., Römling, U., and Tümmler, B. (1997) Disrespectful type IV pilins. *Mol Microbiol* **25:** 203–204.

Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., *et al.* (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406:** 959–964.

Syutsubo, K., Kishira, H., and Harayama, S. (2001) Development of specific oliogonucleotide probes for the identification and in situ defection of hydrocarbon – degrading Alcanivorax strains. *Environ Microbiol* **3:** 371–379.

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., and Glockner, F.O. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* **6:** 938–947.

Ussery, D., Soumpasis, D.M., Brunak, S., Staerfeldt, H.H., Worning, P., and Krogh, A. (2002) Bias of purine stretches in sequenced chromosomes. *Comput Chem* **26:** 531–541.

Wang, H., Noordewier, M., and Benham, C.J. (2004) Stress-Induced DNA Duplex destabilization (SIDD) in the *E. coli* genome: SIDD sites are closely associated with promoters. *Genome Res* **14:** 1575–1584.

Weinel, C., Nelson, K.E., and Tümmler, B. (2002) Global features of the *Pseudomonas putida* KT2440 genome sequence. *Environ Microbiol* **4:** 809–818.

Willenbrock, H., and Ussery, D.W. (2007) Prediction of highly expressed genes in microbes based on chromatin accessibility. *BMC Mol Biol* **8:** 11.

Willenbrock, H., Friis, C., Juncker, A.S., and Ussery, D.W. (2006) An environmental signature for 323 microbial genomes based on codon adaptation indices. *Genome Biol* **7:** R114.

Worning, P., Jensen, L.J., Hallin, P.F., Staerfeldt, H.H., and Ussery, D.W. (2006) Origin of replication in circular prokaryotic chromosomes. *Environ Microbiol* **8:** 353–361.

Yakimov, M.M., Golyshin, P.N., Lang, S., Moore, E.R., Abraham, W.R., Lunsdorf, H., and Timmis, K.N. (1998) *Alcanivorax borkumensis* General nov., sp. nov., a new, hydrocarbon-degrading and surfactant-producing marine bacterium. *Int J Syst Bacteriol* **48:** 339–348.

