BMC Bioinformatics

BioMed Central

Research article

Open Access

Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns

Oleg N Reva* and Burkhard Tümmler

Address: Klinische Forschergruppe, OE6711, Medizinische Hochschule Hannover, Carl-Neuberg-Strasse 1, D-30625 Hanover, Germany

Email: Oleg N Reva* - reva.oleg@mh-hannover.de; Burkhard Tümmler - tuemmler.burkhard@mh-hannover.de

* Corresponding author

Published: 07 July 2004

BMC Bioinformatics 2004, **5**:90 doi:10.1186/1471-2105-5-90

Received: 28 June 2004

Accepted: 07 July 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/90

Abstract

Background: Oligonucleotide frequencies were shown to be conserved signatures for bacterial genomes, however, the underlying constraints have yet not been resolved in detail. In this paper we analyzed oligonucleotide usage (OU) biases in a comprehensive collection of 155 completely sequenced bacterial chromosomes, 316 plasmids and 104 phages.

Results: Two global features were analyzed: pattern skew (PS) and variance of OU deviations normalized by mononucleotide content of the sequence (OUV). OUV reflects the strength of OU biases and taxonomic signals. PS denotes asymmetry of OU in direct and reverse DNA strands. A trend towards minimal PS was observed for almost all complete sequences of bacterial chromosomes and plasmids, however, PS was substantially higher in separate genomic loci and several types of plasmids and phages characterized by long stretches of non-coding DNA and/or asymmetric gene distribution on the two DNA strands. Five of the 155 bacterial chromosomes have anomalously high PS, of which the chromosomes of *Xylella fastidiosa* 9a5c and *Prochlorococcus marinus* MIT9313 exhibit extreme PS values suggesting an intermediate unstable state of these two genomes.

Conclusions: Strand symmetry as indicated by minimal PS is a universally conserved feature of complete bacterial genomes that results from the matching mutual compensation of local OU biases on both replichors while OUV is more a taxon specific feature. Local events such as inversions or the incorporation of genome islands are balanced by global changes in genome organization to minimize PS that may represent one of the leading evolutionary forces driving bacterial genome diversification.

Background

The analysis of oligonucleotide usage (OU) biases is a useful approach to study bacterial genome organization [1-5]. A number of computational approaches have been designed to visualize pathogenicity regions [2,6-8] or to separate sequences of diverse origins by differential oligonucleotide composition [1,3,5,9,10]. These methods were based on analysis of OU variances and departures from expectations. Di-, tri- and tetranucleotide frequencies were shown to be conserved signatures for individual genomes [1,4,5,11,12] albeit the underlying biological and physicochemical constraints that lead to the over- and underrepresentation of oligonucleotides are only partially known and understood. Mutagenesis and repair, restriction-modification systems, amino acid frequency, codon bias and structural parameters, to name just a few, should

all influence the frequency and sequence of oligonucleotides. For example, DNA flexibility and curvature has been related to the nucleotide sequence by trinucleotide bendability, dinucleotide base stacking energy and propeller twist angle [13].

OU biases are characteristic features of a bacterial genome [4,11], but according to Chargaff's first and second parity rule [14] the complementary strands are believed to be symmetric. However, with more and more complete bacterial genome sequences at hand, the AT and GC composition are meanwhile known to be skewed between leading and lagging strand [15,16] and the extrapolation of the first-order parity rule to higher orders of oligonucleotide composition has yet not been investigated in sufficient depth to draw any valid conclusions. The only published evidence that strand symmetry extends from base composition (first order) to oligonucleotides (higher orders) has been provided by numerical analysis of the sequence of human chromosome 22 [12]. According to this data strand symmetry does not result from a single cause, but rather seems to emerge from numerous mechanisms [12]. We were curious to know whether or not strand symmetry of higher order is a universal feature in the microbial world. Physicochemical constraints operating at the oligonucleotide level could lead to higher order strand symmetry at the local and/or global level. Hence we determined strand symmetry and its relationship to oligonucleotide composition in a comprehensive collection of completely sequenced bacterial chromosomes, plasmids and phages.

Results and Discussion

DNA structure and oligonucleotide usage

First, we wanted to know whether an association exists between structural features of DNA and the frequency of particular tetranucleotide words in bacterial genomes. DNA structural features have been related from theoretical calculations and empirical measurements to di- or trinucleotide scales such as base stacking energy [17], propeller twist angle [18], protein deformability [19], position preference [20] and bendability [21]. By permutation analysis the 256 tetranucleotides were assigned to 39 equivalence classes each of which characterized by the same values for the five scales mentioned above [13].

The deviations Δ_w of observed from expected counts (see eq. 6 in 'Methods' section) were determined for all 256 tetranucleotides on one clockwise strand of four bacterial genomes and then sorted by decreasing stacking energy with class 39 (CGCG, GCGC) and 32 (ATAT, TATA) having the highest and lowest energy, respectively. The preference of tetranucleotides for low, intermediate or high base stacking energy was found to be a taxon-specific feature. For example, whereas the GC-rich *Pseudomonas putida* KT2440 prefers the extremes with low and high stacking energy, the selection of tetranucleotides in the AT-rich *Bacillus subtilis* 168 is biased towards low stacking energy (Fig. 1). All analyzed genomes, however, shared the feature that virtually all tetranucleotide words within the same class were distributed with similar taxon-specific frequency. In other words, tetranucleotides with matching structural features apparently occur with comparable frequency in a bacterial genome. The high-order parity of complementary DNA strands has been postulated previously, implicitly or explicitly, as the consequence of first-order symmetry [5,12,22,23], but was not substantiated by comprehensive analysis of real sequence data. The reader should note that an oligonucleotide and its reverse complement always belong to the same equivalence class. This tendency towards intrastrand parity of complementary oligonucleotides will consequently lead to a symmetric distribution of oligonucleotides on both strands. To verify the universality of this phenomenon, we systematically analyzed a comprehensive selection of 155 bacterial chromosomes, 104 phages and 316 plasmids by the calculation of pattern skew (PS), a global genome parameter of intrastrand disparity. The values determined for all analyzed genomes are presented in additional data files [see additional data files 1,2 and 3 for chromosomes, plasmids and phages, respectively].

OU pattern skew

The PS values of n0_4mer patterns (definition given in the 'Methods' section) determined for all sequenced bacterial chromosomes were in the range of 1 to 8% (Fig. 2A) except for the 3 strains *Clostridium tetani* E88, *Nitrosomonas europaea* ATCC 19718 and *Haemophilus ducreyi* 35000 HP with PS values of about 9% and the two strains *Prochlorococcus marinus* MIT9313 and *Xylella fastidiosa* 9a5c with extreme values of 15.97% and 24.27%, respectively. No reliable links between the taxonomic position of the organism and its PS were detected. The strains *X. fastidiosa* Temecula, *P. marinus* ssp.*marinus* CCMP1375 and *P. marinus* ssp.*pastoris* CCMP1378 exhibit symmetric patterns in contrast to their close relatives *X. fastidiosa* 9a5c and *P. marinus* MIT9313 [see additional data file 1].

In shorter sequences both the PS mean values and standard deviation increase exponentially due to size limitation of the sequences. Regression equations were calculated from the empirical data to simulate this dependence for average PS (PS_{avr}):

$$PS_{avr} = 100 - 95.24 \times EXP(-2796.6/ L_{sec}) \quad (1)$$

and for the standard deviation σ_{PS}

$$\sigma_{PS} = 100 - 98.66 \times EXP(-652.27/ L_{sec}) \quad (2)$$

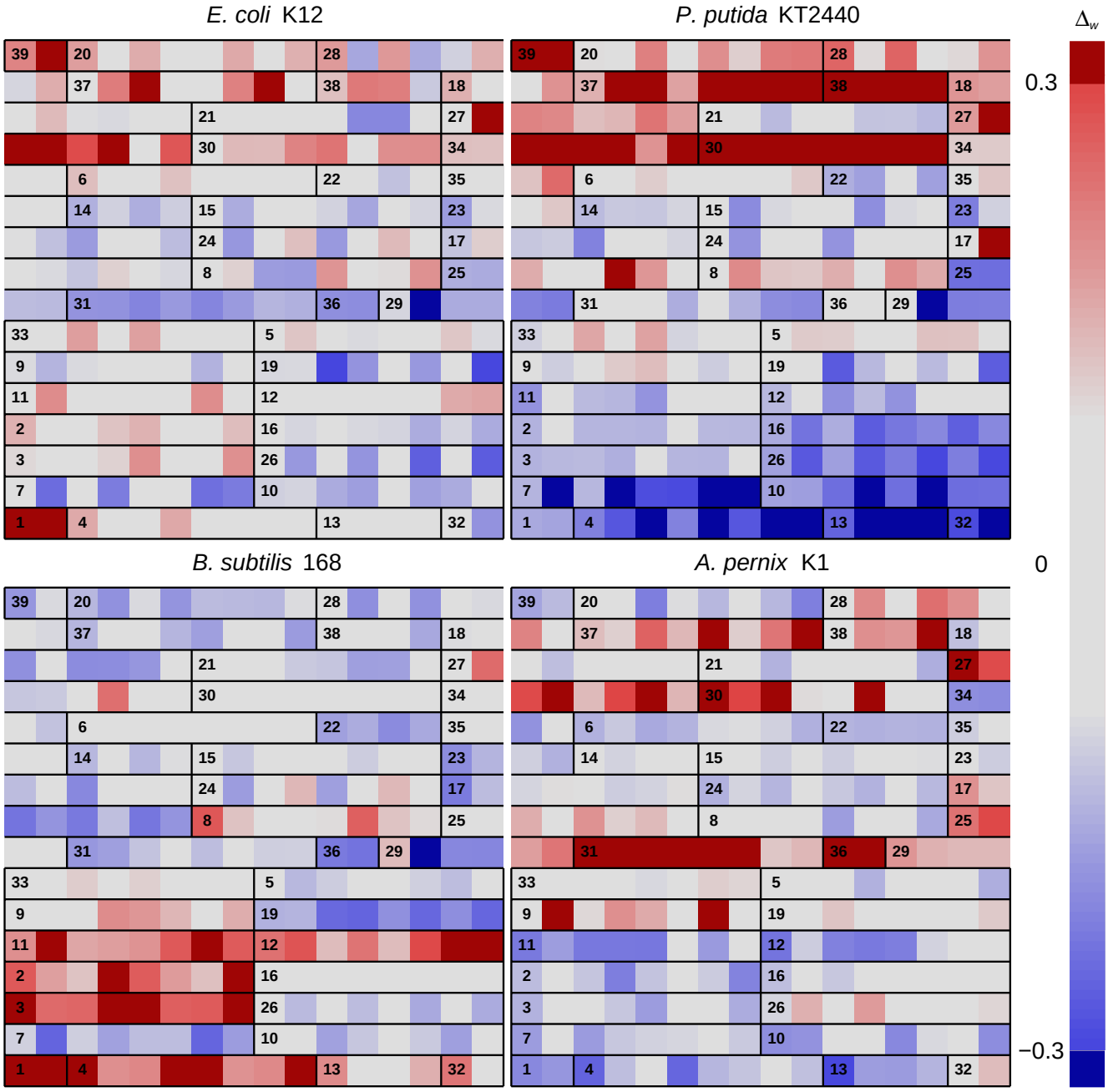

Figure 1

Tetranucleotide usage patterns of *E. coli* K12, *P. putida* KT2440, *B. subtilis* 168 and *A. pernix* K1. The deviation Δ_w of observed from expected counts defined by eq. 6 is shown for all 256 tetranucleotides (16 × 16 cells) by color code (right bar). Tetranucleotides are grouped into 39 classes of equivalent structural features [13] and sorted by decreasing base stacking energy row-by-row starting at the upper left corner (class 39). Within a class members are sorted alphabetically.

The 95% confidential intervals of PS values (grey shaded area) in bacterial chromosomes and plasmids in dependence of sequence length are shown on Fig. 2A.

The increase of PS with decreasing genome size is not a biological feature, but just reflects the increase of local stochastic fluctuations in oligonucleotide usage. The smaller

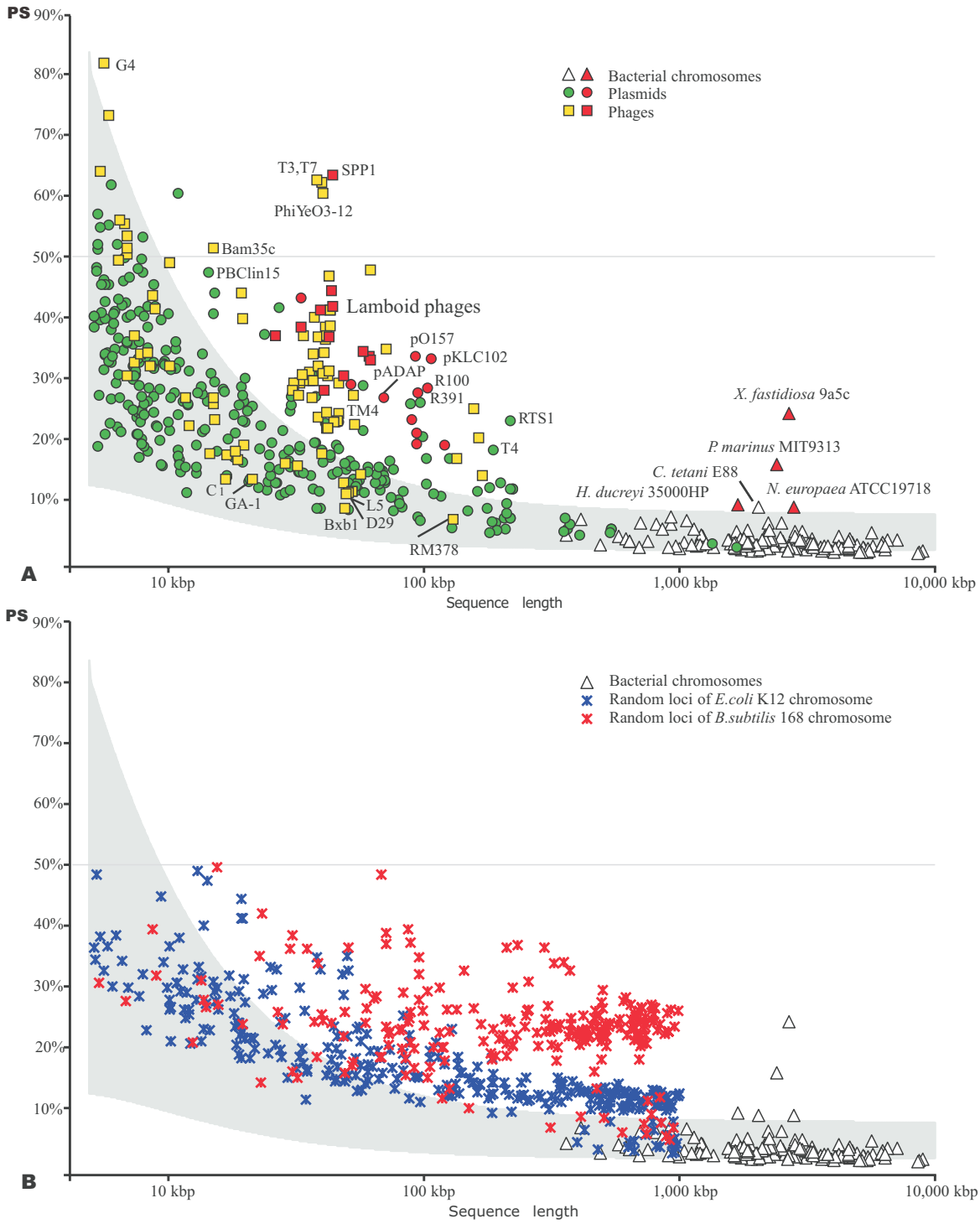

Figure 2

Pattern skew of DNA sequences of different length. (A) Pattern skew (n0_4mer PS) values determined for a comprehensive collection of sequences of bacterial chromosomes, plasmids and phages are plotted against the logarithmic scale of sequence lengths. The grey shaded area depicts the 95% confidence intervals of variation of n0_4mer PS values in the complete chromosomal and plasmid sequences. Accession names of genomes been discussed in the text are presented. The genomes where the n0_4mer PS values exceeded the corresponding n1_4mer PS values by more than 2.5 σ_{PS} (see equation 2) are shown in red. (B) The n0_4mer PS values determined for arbitrary loci randomly cut out of the *E. coli* K12 (blue) and *B. subtilis* 168 (red) chromosome sequences are compared with PS values determined for complete bacterial chromosomes.

the genome and correspondingly the smaller the frequency of the individual oligonucleotide, the larger the variance. The portion of noise increases exponentially in shorter sequences making the oligonucleotide distribution parameters closer to the characteristics of a random sequence [24].

Strand symmetry was found to be a global feature of bacterial chromosomes, but this did not apply to all plasmids and phages: PS values varied over a large range (Fig. 2A). PS values of circular and linear plasmid and phage genomes overlapped (data not shown). Consider some examples of symmetric and asymmetric genomes. The majority of large plasmids have low PS values like complete bacterial chromosomes [see additional data file 2]. However, some large plasmids including Rts1 of *Proteus vulgaris* exhibit unexpectedly high PS. Among the 300 ORFs of this plasmid, 253 are oriented in the same direction [25].

Higher PS was found in so-called conjugative genome islands such as pKLC102 [26], R100, R391 [27] and virulence plasmid pO157 [28] and pADAP [29] which share a mosaic genome structure of plasmid and phage origins. Another example of this transition zone of phages and plasmids are the plasmid pBClin15 of *Bacillus cereus* and the phage Bam35c of *Bacillus thuringiensis*. Both shares an almost identical gene repertoire, all genes are leftward oriented [30] and in both cases PS values are beyond the 95% confidence limit.

The majority of phage genomes exhibit higher PS values than plasmids or chromosomes [see additional data file 3]. Most head-and-tail bacteriophages have a genome sequence length of around 50 kbp that fits to capsid size [31] (Fig. 2A). Short tail *Podoviridae*, such as streptococcal phage C_1 [32] and bacillar phage GA-1, have symmetric OU, while PS increases in long tail lamboid *Siphoviridae* up to extreme values of 60–64% in SPP1, T3, T7, phiYe03-12 phages. A large portion of the genome of long tail phages is non-coding and serves to anchor the DNA within the capsular head and tail [33,34]. Thus in SPP1 phage only 32 kbp out of 47 kbp are coding sequences and all genes are transcribed from one strand [35]. The highest PS value of 82% was observed in the 5,415 bp single strand DNA enterobacterial phage G4.

Symmetric OU is characteristic for *Myoviridae* T4 and RM 378 and for mycobacterial *Siphoviridaea* D29, L5 and Bxb1. The genomes of the former have a symmetric structure of two arms of leftward and rightward transcribed genes. However, in the close relative mycobacteriophage TM4 all genes are transcribed in the rightward direction [36] and its PS value for the n0_4mer pattern is threefold larger than that of D29, L5 and Bxb1. In conclusion, asymmetric genome topology and/or the presence of large inserts of less conserved non-coding DNA stretches apparently cause strand specific OU asymmetry in some plasmids and most phages.

Next, we wanted to test whether the association of PS values being higher for non-coding than for coding sequences does not only apply to phages, but also to bacterial chromosomes. PS values of non-coding regions of tested bacterial chromosomes were higher than PS values of coding sequences (Table 1), however, this higher PS is just caused to major extent by its shorter total length in the genome. The comparatively higher PS values are within expectation for a shorter sequence for most genomes (Fig. 2A). Exceptions were the two outliers *X. fastidiosa* 9a5c and *P. marinus* MIT9313 (see above) and chromosomes of *Campylobacter jejuni* NCTC 11168, *Corynebacterium glutamicum* ATCC 13032 and *Xylella fastidiosa* Temecula1. In other words, the same underlying principles shape oligonucleotide distributions in coding and non-coding sequences in bacterial genomes; i.e. codon usage is not a major determinant for global PS. A bias of coding sequences to one strand as seen in phages, however, leads to increased PS according to their other lifestyle and evolution.

Next, the hypothesis was tested whether global strand symmetry of bacterial chromosomes also extends to local regions. The local PS values were calculated for 15 selected chromosomes (Table 1). PS values of n0_4mer patterns were determined for 100 randomly generated arbitrary loci of 5 to 1,000 kbp in size. These local patterns exhibited significantly higher PS than complete genomes of the same size for 12 of 15 analyzed chromosomes ($P < 10^{-6}$ in all 12 cases, χ^2 tests, P values corrected for multiple testing; the genome of *Mycoplasms pulmonis* UAB CTIP was not applicable for this test due to its small genome size). Figure 2B displays the local patterns randomly generated from the *Escherichia coli* K12 and *Bacillus subtilis* 168 chromosomes. The majority of PS values of chromosomal segments were higher than the 95% confidence interval predicted by equations 1 and 2 for genome sequences. In *Aquifex aeolicus*, *Bradyrhizobium japonicum*, and *Streptomyces coelicolor*, however, local PS was only slightly higher than in the whole chromosome (Table 1) indicating strand symmetry throughout the chromosome. In all other selected species pronounced strand asymmetry was determined to differential genome-specific extent (see *X. fastidiosa* TemeculaI as an extreme example, Table 1), but these local skews of tetranucleotide usage are mutually compensated by other regions so that PS skew of the whole chromosome is reduced to 1–8%.

Normalization of OU frequencies by mononucleotide content typically had either no effect or increased PS,

Table 1: Local and global PS of bacterial chromosomes

Bacterial chromosome	Length (bp) of sequence			Global n0_4mer PS (%)			Local n0_4mer PS (%)*
	Total	Coding	Non-coding	Total	Coding	Non-coding	Median (inner quartiles, range)
Aeropyrum pernix K1	1,669,695	1,490,824	178,871	4.92	4.46	7.81	9.61 (6.19 – 13.63, 4.37 – 41.24)
Aquifex aeolicus VF5	1,551,335	1,448,950	102,385	3.40	3.49	8.18	5.37 (4.15 – 7.36, 2.89 – 32.50)
Bacillus subtilis 168	4,214,814	3,684,952	529,862	2.50	2.40	5.27	23.71 (21.50 – 26.81, 4.36 – 47.16)
Bradyrhizobium japonicum USDA 110	8,619,960	7,515,107	1,104,853	1.27	1.38	3.01	5.89 (5.18 – 7.09, 2.18 – 23.18)
Campylobacter jejuni NCTC 11168	1,641,481	1,555,799	85,682	2.26	2.49	15.10	15.86 (11.46 – 20.79, 1.80 – 28.85)
Corynebacterium glutamicum ATCC 13032	3,309,401	2,867,342	442,059	3.71	3.49	9.68	17.84 (16.45 – 21.99, 6.26 – 56.73)
Escherichia coli K-12	4,639,221	4,096,745	542,476	2.15	2.44	5.85	15.01 (12.04 – 23.11, 2.86 – 49.02)
Mycoplasma pulmonis UAB CTIP	963,879	869,493	94,386	2.45	2.52	6.37	Not applicable
Prochlorococcus marinus MIT9313	2,410,873	1,982,808	428,065	15.97	15.75	14.45	31.40 (19.54 – 35.60, 3.46 – 45.97)
Prochlorococcus marinus ssp. *marinus* CCMP1375	1,751,080	1,566,066	185,014	1.82	1.81	4.85	9.99 (6.89 – 12.90, 1.58 – 37.52)
Pseudomonas putida KT2440	6,181,863	5,439,657	742,206	2.75	2.22	5.85	10.95 (9.67 – 12.91, 2.14 – 22.30)
Rhodopirellula baltica 1	7,145,576	6,817,640	327,936	3.45	3.45	6.65	14.83 (11.77 – 17.05, 3.69 – 26.42)
Staphylococcus aureus N315	2,814,816	2,357,692	457,124	2.12	2.04	3.78	22.21 (19.99 – 23.74, 7.27 – 38.55)
Streptomyces coelicolor A3(2)	8,667,507	7,379,401	1,288,106	1.48	1.42	1.77	4.70 (3.51 – 6.52, 1.91 – 20.42)
Xylella fastidiosa 9a5c	2,679,306	2,244,990	434,316	24,27	21.00	36.74	50.64 (39.86 – 57.52, 7.41 – 68.29)
Xylella fastidiosa Temecula1	2,519,802	1,967,507	552,295	6.38	5.02	13.69	53.17 (41.73 – 59.23, 8.80 – 71.71)

*PS values of n0_4mer patterns were calculated for 100 arbitrary loci (200 for *E. coli* and *B. subtilis*) of 5 to 1,000 kbp in size (median 289,752 bp, inner quartiles 114,406 – 477,801 bp).

whereby in case of the latter the OU variance decreased. For instance, the mean values of PS of n0_4mer and n1_4mer patterns of the analyzed bacterial chromosomes [see additional data file 1] were 3.4 ± 3.0% and 3.7 ± 2.0%, respectively. In some cases, however, the normalization significantly decreased PS. On Fig. 2A the sequences are marked where the n0_4mer PS values exceeded the corresponding n1_4mer PS values by more than 2.5 σ_{PS} (see eq.2). All these genomes exhibit high PS values, amongst them are some lamboid phages, conjugative genome islands and four bacterial chromosomes: *X. fastidiosa*9a5c, *P. marinus* MIT9313, *N. europaea* ATCC 19718 and *H. ducreyi* 35000 HP.

Next, we wanted to know whether PS changes with the length of oligomers. PS values increased with the length of oligonucleotide words in all examined bacterial chromosomes to an extent that is consistent with the increase of numbers of n-mers and concomitant random fluctuations between expected and observed counts of words; i.e. strand symmetry is maintained to similar extent for di-, tri-, tetra- and pentanucleotides. However, in the two outliers *X. fastidiosa* 9a5c and *P. marinus* MIT9313 with the largest n0_4mer PS values (additional data file 1, Fig. 2A) PS decreased with increasing word length (Fig. 3). This behaviour could be attributed in both strains to extreme dinucleotide skew that is diluted in longer words. In case of the three other outliers *C. tetani* E88, *N. europaea* ATCC 19718 and *H. ducreyi* 35000 HP di-, tri- and tetranucleotide skews all contribute to the pronounced pattern skew. Since we did not observe any trend towards common oligonucleotides in all five strains that account for the skew but rather individual patterns that are characteristic for the peculiar strain, we conclude that high pattern skew is a strain specific feature.

Asymmetric OU in bacterial chromosomes may reflect asymmetric genome topology and/or multiple inserts of foreign DNA with high PS. To test this hypothesis, the genomes of *X. fastidiosa* 9a5c and *P. putida* KT2440 were selected for a comparative study of OU pattern consistency. Both genomes are characterized by multiple genome islands and unequal lengths of the clockwise and counterclockwise replichors [37,38]. The standard n0_4mer OU patterns were defined for the leading strands of both chromosomes. Next, the local n0_4mer patterns were determined in 15 kbp sliding windows in steps of 7.5 kbp and compared against the standard genome-wide patterns. The *X. fastidiosa* 9a5c chromosome is characterized by large local variations of OU patterns and a qualitatively different behaviour of the clockwise and counterclockwise replichors (Fig. 4). We calculated the distances *D* (see eq.8 in 'Methods' section) of n0_4mer patterns between the clockwise and counterclockwise replichor and between the clockwise and the reverse complement of the counterclockwise replichor to be 44.09% and 9.03%, respectively. Compared to *X. fastidiosa* 9a5c, the OU pattern of the *P. putida* KT2440 genome is smoother and more uniform in spite of many genome islands with atypical OU patterns scattered throughout the chromosome (Fig. 4). OU pat-

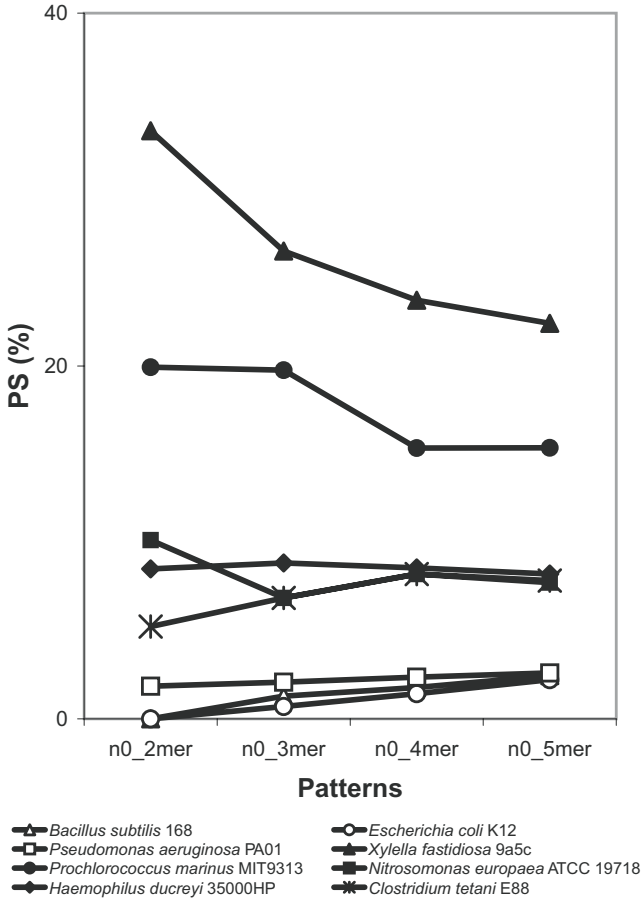

Figure 3
Relation between PS and the length of oligonucleotide words. OU patterns were determined for typical bacterial genomes represented by sequences of *B. subtilis* 168, *E. coli* K12 and *P.aeruginosa* PA01 chromosomes, and 5 chromosomal sequences with anomalously high PS values.

terns of the clockwise and counterclockwise replichors were different but complementary to each other. Thus, in *P. putida* KT2440 the distance of n0_4mer patterns between the clockwise and counterclockwise replichors was 10.43%, while that between the clockwise and the reverse-complement of the counterclockwise replichors was only 1.96%. Notably, despite the length difference, the counterclockwise replichor manages to compensate the mirror OU skew in the clockwise strand due to its greater OU variance in the shorter replichor.

OU variance

OUV is strongly dependent on GC-content [7,39]. If the data from sequenced bacterial chromosomes are superimposed, a parabolic curve is generated with the minimum of n0_4mer OUV at 50 mol% GC [39]. Normalization of OU patterns by mononucleotide content significantly reduced this bias, however, did not remove it completely. OUV values determined for n1_4mer patterns of 155 bacterial chromosomes varied from 0.05 to 0.5. According to regression analysis average values (OUV_{avr}) and variances (σ_{OUV}) of OUV retained some parabolic dependence on mol% GC (X_{GC}):

$$OUV_{avr} = \left(31.26 \times X_{GC}^2 - 27 \times X_{GC} + 11.46\right)^2 \quad (3)$$

$$\sigma_{OUV} = \left(12.81 \times X_{GC}^2 - 14.6 \times X_{GC} + 5.14\right)^2 \quad (4)$$

OUV_{avr} has its minimum at 43% GC and σ_{OUV} at 57% GC content. In GC-rich and AT-rich sequences OU is strongly biased leading to higher OUV. Moreover, the lower variance of OUV values in sequences of high GC-content (Fig. 5) suggests that the use of GC-rich words is more biased than that of AT-rich words, presumably due to the extreme values of base stacking energy, propeller twist and protein deformability [13].

The OUV values of all studied bacterial chromosomes and 3-sigma limits of their variations are shown in Fig. 5. In general strains of the same species or genera have similar OUV [see additional data file 1]. No links between OUV and PS values were observed. The strains *P. marinus* MIT9313 and *X. fastidiosa* 9a5c with atypical PS have similar OUV as their close relatives.

Conclusions

Two global genome features based on OU statistics were considered in this study: PS and OUV. They provide non-redundant characteristics of the complete sequence of genomes and allow the discrimination of bacterial, plasmid and phage genomes by phylogeny, the arrangement of coding and non-coding sequence and the distribution of islands and islets.

A strong taxonomic signal was observed in genome specific OUV values. Strains belonging to the same species or genus usually have similar OUV. In general, the higher is the OUV, the less random is the sequence. Multiple influences such as DNA structure and topology, codon usage, DNA repair and restriction-modification systems contribute to the surrogate parameter OUV, and hence it is plausible that the OUV is a taxon-specific feature. Future work on the frequency and distribution of individual words should elucidate the biological meaning of the genome specific OUV for the individual taxon (see Weinel *et al.*, 2002 [40] as one of the few published examples).

The major finding was that words and their reverse complements occur with similar frequencies in the complete sequence, because they share the same structural features

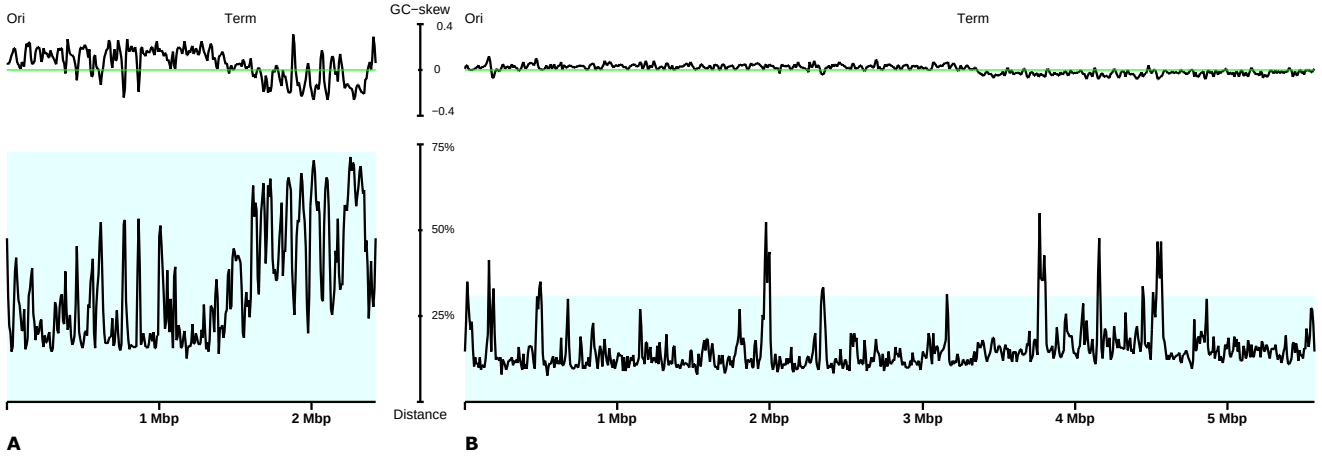

Figure 4

Deviations of oligonucleotide usage patterns in local loci of two bacterial chromosomes. Lower panel: Distances (eq.8) between n0_4mer patterns calculated for local regions of the leading strand and the standard patterns determined for the clockwise replichor of the two bacterial chromosomes: A) *X. fastidiosa* 9a5c; B) *P. putida* KT2440. Local patterns were determined in 15 kbp sliding windows in steps of 7.5 kbp. The 95% confidence interval of distance values is depicted as the turquoise shaded area. The abscissa indicates the coordinates of the chromosomes starting from the putative replication origins (Ori). Positions of the putative chromosomal replication termini are depicted by Term. Upper panel: GC-skew between leading and lagging strands of the (A) *X. fastidiosa* 9a5c and (B) *P. putida* KT2440 chromosomes.

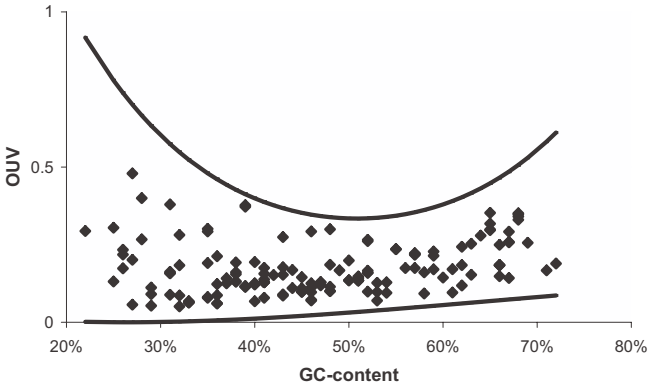

Figure 5

The OUV values defined for n1_4mer patterns of 155 bacterial chromosomes and plotted against the mol% GC content. Curve lines depict the boundaries of the 95% confidence interval of OUV variation determined by eq.3 and 4.

in terms of DNA flexibility and curvature thus generating strand symmetry. However, in local regions PS is substantially more variable. Local PS values, for example, were up to 70% in the *X. fastidiosa* Temecula1 genome (see Table 1), that means the complementary words were deliberately counterselected in these sequences, but even in this extreme case the global PS of the chromosome remained low. Local asymmetric OU patterns of leading and lagging strands of DNA mutually compensate each other restoring in that way the OU symmetry of the complete sequence. Notably, OU symmetry was retained in sequences with replichors of disproportionate length and multiple horizontally acquired genome islands like in the *P. putida* KT2440 genome. Low PS is a feature of complete genomes. This conclusion is drawn from the fact that the PS values of plasmid genomes and whole chromosomes are smaller than those of arbitrarily selected chromosomal loci of the same size (see Fig. 2B, Table 1).

High PS was predominant for phages and conjugative genome islands that corresponds to their dual lifestyle to either exist in the episomal state or to be integrated into the chromosome. In addition, the high PS correlates with the known high mutation rates of these genetic elements and an intensive horizontal exchange that is an important component of their evolution [26,41-43]. Among the analyzed bacteria only the chromosomes of *X. fastidiosa* 9a5c and *P. marinus* MIT9313 show high PS which in both cases is caused by dinucleotide skew of four complementary pairs (TT/AA, GG/CC, GT/AC and TG/CA). This dinucleotide skew led to large fluctuations of local OU patterns, high global PS that is not compensated by the replichors (Fig. 4) and loss of intrastrand parity which should give rise to an uncommon, probably unstable chromosome structure. Future experimental work may

show whether or not such a status of the genome is associated with reduced fitness and/or more rapid evolution.

In summary, referring to the comparative analysis of PS in local regions and complete genomes shown in Fig. 2, the proportional usage of words and their reverse complements, i.e. strand symmetry is important for genome stability, and there could exist yet unknown housekeeping mechanisms to control compensating matching of OU patterns of different loci of a sequence. Local events such as inversions or the incorporation of a genome island are balanced by global changes in genome organization to minimize pattern skew. This response of the genome to local perturbations may represent one of the leading evolutionary forces that drive bacterial genome diversification and speciation.

Methods

Overall DNA properties

Intragenomic GC-content and GC-skew variations were determined as quantities of (G+C) and (G-C)/(G+C), respectively, averaged over a sliding window of certain length.

Evaluation of oligonucleotide patterns

Overlapping oligonucleotide words of a certain length l_w were 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 in a linear sequence or W_{total}=L_{seq} in a circular sequence. Since $L_{seq} >> l_w$, $W_{total} \cong L_{seq}$ in all cases. For a given word length l_w, $N_w = 4^{lw}$ different words are possible for a sequence of four letters A, T, G and C. 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 of a common length l_w irrespective of their composition and sequence, C_e matches the standard count number C_{n0}

$$C_e = C_{n0} = W_{total} \times N_w^{-1} \qquad (5)$$

Correspondingly, if we normalize oligonucleotide usage (OU) by mononucleotide content using zero-order Markov method [44], C_e becomes

$$C_e = C_{n1}$$

The deviation Δ_w of observed from expected counts is given by

$$\Delta_w = \left(C_0 - C_e \right) \times C_{n0}^{-1} \qquad (6)$$

In the present work we used the following abbreviations for the different types of patterns: type_$l_w$mer. Types are called 'n0', if they are not normalized by mononucleotide frequency, or 'n1', if they are normalized by the zero-order Markov method. For example, the non-normalized trinucleotide usage pattern is a n0_3mer type, the normalized pentanucleotide usage pattern is a n1_5mer type.

Variance OUV of word deviations were determined as following:

$$OUV = \frac{\sum_w \Delta_w^2}{N_w - 1} \qquad (7)$$

Pattern comparison and pattern skew

For the comparison of sequences by OU patterns of the same type, the words in each sequence were ranked by Δ_w values according to equation 6. Rank numbers instead of word counts were used to simplify pattern comparison and to remove sequence length bias. Assuming that 95% of all words should occur at least ten times in a random sequence, the threshold for the minimum length of the sequence was chosen to be 0.3, 1.2, 5 and 20 kb for di-, tri-, tetra- and pentanucleotides, respectively.

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:

$$D(\%) = 100 \times \frac{\sum_w | rank_{w,i} - rank_{w,j} | - D_{\min}}{D_{\max} - D_{\min}} \qquad (8)$$

where

$$D_{\max} = \frac{N_w \left(N_w - 1 \right)}{2} \qquad (9)$$

$D_{\max}$ is the maximal distance that is theoretically possible between two patterns of l_w long words (equation 9). $D_{\min}$ is the minimal distance between two patterns. The minimal distance is zero for two independent sequences, but has a positive value for the two complementary strands of the same DNA sequence, because the OU patterns designed for both strands of the same DNA molecule cannot be identical. The pattern skew (PS) describes this distance between opposite strands of the same DNA and is a measure of OU asymmetry. The minimal theoretical distance between two patterns of opposite strands is realized if the words and their reverse complements are distributed with similar frequencies in the sequence and it is

$D_{\min} = 4^{lw}$, if l_w is an odd number (10a)

but

$D_{\min} = 4^{lw} - 2^{lw}$, if l_w is an even number (10b).

because palindromes, which occur in both strands with the same frequency, only exist in words with an even number of nucleotides and the total number of all possible palindromes is 2^{lw}.

Data mining and storage

A computational program for determining OU patterns, their comparative analysis and storage in a database was written on Python 2.2 [45]. Sequences of 155 bacterial chromosomes including eubacterial, archaeal and cyanobacterial genomes, 316 plasmids and 104 phages published in NCBI database [46] and the plasmid genome database [47] were analyzed in this study. Regression analysis has been done using DataFit7.1.44 software. Coding and non-coding sequences were selected by gene coordinates provided in the NCBI database [46]. Intergenic spacer regions shorter or equal 50 bp were included in the coding sequence of a genome, while longer spacer regions were concatenated into the non-coding sequence.

List of abbreviations

OU- oligonucleotide usage;

OUV- oligonucleotide usage variance;

PS- pattern skew

Authors' contribution

ONR did programming on Python. Both authors contributed equally to all other presented data.

Additional material

Additional File 1

Global features of all studied bacterial chromosomes *bacterial chromosome, length, mol% GC, OUV, n0_4mer PS(%), n1_4mer PS(%).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-90-S1.xls]

Additional File 2

Global features of sequences of bacterial plasmids *plasmid, length, mol% GC, type, n0_4mer PS(%), n1_4mer PS(%).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-90-S2.xls]

Additional File 3

Global features of sequences of bacteriophages *phage, length, mol% GC, type, n0_4mer PS(%), n1_4mer PS(%).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-90-S3.xls]

Acknowledgements

This work was supported by the DFG-sponsored Europäisches Graduiertenkolleg 653. The analysis was executed within the 'Task Force for Microbial Genome Linguistics' that is part of the BMBF-sponsored Competence Network 'Genome Research on Bacteria Relevant for Agriculture, Environment and Biotechnology'.

References

1. Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T: **Informatics for unveiling hidden genome signatures.** *Genome Res* 2003, **13:**693-702.
2. Karlin S: **Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes.** *Trends Microbiol* 2001, **9:**335-343.
3. Karlin S, Cardon LR: **Computational DNA sequence analysis.** *Annu Rev Microbiol* 1994, **48:**619-654.
4. Karlin S, Mrazek J, Campbell A: **Compositional biases of bacterial genomes and evolutionary implications.** *J Bacteriol* 1997, **179:**3899-3913.
5. Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ: **Evolutionary implications of microbial genome tetranucleotide frequency biases.** *Genome Res* 2003, **13:**145-155.
6. Friis C, Jensen LJ, Ussery DW: **Visualization of pathogenicity regions in bacteria.** *Genetica* 2000, **108:**47-51.
7. Noble PA, Citek RW, Ogunseitan OA: **Tetranucleotide frequencies in microbial genomes.** *Electrophoresis* 1998, **19:**528-535.
8. Pride DT, Blaser MJ: **Identification of horizontally acquired elements in *Helicobacter pylori* and other prokaryotes using oligonucleotide difference analysis.** *Genome Lett* 2002, **1:**2-15.
9. Freeman JM, Plasterer TN, Berry A, Paton J: **Patterns of genome organization in bacteria.** *Science* 1998, **279:**1827.
10. Ussery DW, Larsen TS, Wilkes KT, Friis C, Worning P, Krogh A, Brunak S: **Genome organisation and chromatin structure in *Escherichia coli*.** *Biochimie* 2001, **83:**201-212.
11. Karlin S: **Global dinucleotide signatures and analysis of genomic heterogeneity.** *Curr Opin Microbiol* 1998, **1:**598-610.
12. Baisne P-F, Hampson S, Baldi P: **Why are complementary DNA strands symmetric?** *Bioinformatics* 2002, **18:**1021-1033.
13. Baldi P, Baisne P-F: **Sequence analysis by additive scales: DNA structure for sequences and repeats of all length.** *Bioinformatics* 2000, **16:**865-889.
14. Chargaff E: **Structure and function of nucleic acids as cell constituents.** *Fed Proc* 1951, **10:**344-360.
15. Baran RH, Ko H, Jernigan RW: **Methods for comparing sources of strand compositional asymmetry in microbial chromosomes.** *DNA Res* 2003, **30:**85-95.
16. Tillier ER, Collins RA: **The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes.** *J Mol Evol* 2000, **50:**249-257.
17. Ornstein RL, Rein R, Breen DL, MacElroy RD: **An optimized potential function for the calculation of nucleic acid interaction energies. I. Base stacking.** *Biopolymers* 1978, **17:**2341-2360.
18. Hassan MAE, Calladine CR: **Propeller twist of base-pairs and the conformational mobility of dinucleotide steps in DNA.** *J Mol Biol* 1996, **259:**95-103.
19. Olson WK, Gorin AA, Lu X, Hock LM, Zhurkin VB: **DNA sequence-dependent deformability deduced from protein-DNA crystal complexes.** *Proc Natl Acad Sci USA* 1998, **95:**11163-11168.
20. Pedersen AG, Baldi P, Brunak S, Chauvin Y: **DNA structure in human RNA polymerase II promoters.** *J Mol Biol* 1998, **281:**663-673.
21. Brukner I, Sánchez R, Suck D, Pongor S: **Sequence-dependent bending propensity of DNA as reveald by DNase I: parameters for trinucleotides.** *EMBO J* 1995, **14:**1812-1818.
22. Fickett JW, Torney DC, Wolf DR: **Base compositional structure of genomes.** *Genomics* 1992, **13:**1056-1064.
23. Forsdyke DR: **Relative role of primary sequence and (G+C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species.** *J Mol Evol* 1995, **41:**573-581.
24. Reinert G, Schbath S, Waterman MS: **Probabilistic and statistical properties of words: an overview.** *J Comp Biol* 2000, **7:**1-46.

25. Murata T, Ohnishi M, Ara T, Kaneko J, Han CG, Li YF, Takashima K, Nojima H, Nakayama K, Kaji A, *et al.*: **Complete Nucleotide Sequence of Plasmid Rts, 1: Implications for Evolution of Large Plasmid Genomes.** *J Bacteriol* 2002, **184:**3194-3202.
26. Klockgether J, Reva O, Larbig K, Tümmler B: **Sequence analysis of the mobile genome island pKLC102 of *Pseudomonas aeruginosa* C.** *J Bacteriol* 2004, **186:**518-534.
27. Böltner D, Osborn AM: **Structural comparison of the integrative and conjugative elements R391, pMERPH, R997, and SXT.** *Plasmid* 2004, **51:**12-23.
28. Burland V, Shao Y, Perna NT, Plunkett G, Sofia HJ, Blattner FR: **The complete DNA sequence and analysis of the large virulence plasmid of *Escherichia coli* O, 157:H7.** *Nucleic Acids Res* 1998, **26:**4196-4204.
29. Hurst MRH, Glare TR, Jackson TA, Ronson CW: **Plasmid-located pathogenicity determinants of *Serratia entomophila*, the causal agent of amber disease of grass grub, show similarity to the insecticidal toxins of *Photorhabdus luminescens*.** *J Bacteriol* 2000, **182:**5127-5138.
30. Strömsten NJ, Benson SD, Burnett RM, Bamford DH, Bamford JKH: **The *Bacillus thuringiensis* linear double-stranded DNA phage Bam35, which is highly similar to the *Bacillus cereus* linear plasmid pBClin15, has a prophage state.** *J Bacteriol* 2003, **185:**6985-6989.
31. Ackermann HW: **Tailed bacteriophages: the order *Caudovirales*.** *Adv Virus Res* 1998, **51:**135-201.
32. Nelson D, Schuch R, Zhu S, Tscherne DM, Fischetti VA: **Genomic sequence of C_1, the first streptococcal phage.** *J Bacteriol* 2003, **185:**3325-3332.
33. Ackermann HW: **Bacteriophage observations and evolution.** *Res Microbiol* 2003, **154:**245-251.
34. Lurz R, Orlova EV, Gunther D, Dube P, Droge A, Weise F, van Heel M, Tavares P: **Structural organisation of the head-to-tail interface of a bacterial virus.** *J Mol Biol* 2001, **310:**1027-1037.
35. Alonso JC, Luder G, Stiege AC, Chai S, Weise F, Trautner TA: **The complete nucleotide sequence and functional organization of *Bacillus subtilis* bacteriophage SPP1.** *Gene* 1997, **204:**201-212.
36. Mediavilla J, Jain S, Kriakov J, Michael E, Ford ME, Duda RL, Jacobs WR Jr, Hendrix RW, Hatfull GF: **Genome organization and characterization of mycobacteriophage Bxb1.** *Mol Microbiol* 2000, **38:**955-970.
37. Nelson KE, Weinel C, Paulsen IT, Dodson RJ, Hilbert H, Martins dos Santos VAP, Fouts DE, Gill SR, Pop M, Holmes M, *et al.*: **Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440.** *Envir Microbiol* 2002, **4:**799-808.
38. Van Sluys MA, de Oliveira MC, Monteiro-Vitorello CB, Miyaki CY, Furlan LR, Camargo LEA, da Silva ACR, Moon DH, Takita MA, Lemos EGM, *et al.*: **Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of *Xylella fastidiosa*.** *J Bacteriol* 2003, **185:**1018-1026.
39. Weinel C, Nelson KE, Tümmler B: **Global features of the *Pseudomonas putida* KT2440 genome sequence.** *Envir Microbiol* 2002, **4:**809-818.
40. Weinel C, Ussery DW, Ohlsson H, Sicheritz-Ponten T, Kiewitz C, Tümmler B: **Comparative genomics of *Pseudomonas aeruginosa* PAO1 and *Pseudomonas putida* KT2, 440: orthologs, codon usage, repetitive extragenic palindromic elements, and oligonucleotide motif signatures.** *Genome Lett* 2002, **1:**175-187.
41. Böltner D, MacMahon C, Pembroke JT, Strike P, Osborn AM: **R, 391: a conjugative integrating mosaic comprised of phage, plasmid, and transposon elements.** *J Bacteriol* 2002, **184:**5158-5169.
42. Hendrix RW: **Bacteriophage genomics.** *Curr Opin Microbiol* 2003, **6:**506-511.
43. Hendrix RW, Hatfull GF, Smith MC: **Bacteriophages with tails: chasing their origins and evolution.** *Res Microbiol* 2003, **154:**253-257.
44. Almagor H: **A Markov analysis of DNA sequences.** *J Theor Biol* 1983, **104:**633-645.
45. **Python home page** [http://www.python.org/]
46. **NCBI home page** [http://www.ncbi.nlm.nih.gov/]
47. **Plasmid genome database** [http://www.genomics.ceh.ac.uk/plasmiddb/]