Research article

Differentiation of regions with atypical oligonucleotide composition in bacterial genomes

Oleg N Reva*1,2 and Burkhard Tümmler1

Address: 1Klinische Forschergruppe, OE6711, Medizinische Hochschule Hannover, Carl-Neuberg-Strasse 1, D-30625 Hannover, Germany and 2Danylo Zabolotny Institute of Microbiology and Virology of the National Academy of Science of Ukraine, Dep. of Antibiotics, 154 Zabolotnogo Str., D03680, Kyiv GSP, Ukraine

Email: Oleg N Reva* - reva@serv.imv.kiev.ua; Burkhard Tümmler - tuemmler.burkhard@mh-hannover.de

* Corresponding author

Abstract

Background: Complete sequencing of bacterial genomes has become a common technique of present day microbiology. Thereafter, data mining in the complete sequence is an essential step. New in silico methods are needed that rapidly identify the major features of genome organization and facilitate the prediction of the functional class of ORFs. We tested the usefulness of local oligonucleotide usage (OU) patterns to recognize and differentiate types of atypical oligonucleotide composition in DNA sequences of bacterial genomes.

Results: A total of 163 bacterial genomes of eubacteria and archaea published in the NCBI database were analyzed. Local OU patterns exhibit substantial intrachromosomal variation in bacteria. Loci with alternative OU patterns were parts of horizontally acquired gene islands or ancient regions such as genes for ribosomal proteins and RNAs. OU statistical parameters, such as local pattern deviation (D), pattern skew (PS) and OU variance (OUV) enabled the detection and visualization of gene islands of different functional classes.

Conclusion: A set of approaches has been designed for the statistical analysis of nucleotide sequences of bacterial genomes. These methods are useful for the visualization and differentiation of regions with atypical oligonucleotide composition prior to or accompanying gene annotation.

Background

The number of sequenced prokaryotic genomes increases rapidly each year. Their comprehensive analysis requires the development of new high-throughput computational methods. The analysis of oligonucleotide usage biases has been recognized to be practical for the recognition of pathogenicity islands [1,2] and elucidation of origins of orphan sequences [3-5]. Recently we have developed methods for the global analysis of oligonucleotide usage (OU) in complete sequences of bacterial chromosomes, plasmids and phages [6]. The patterns of deviations of oligonucleotide frequencies from expectations were shown to be genome signatures reflecting to some extent the phylogenetic links between microorganisms [3,4,7,8].

The usage of oligonucleotides in bacterial sequences is not random. Frequencies of the oligonucleotide words (further – words) depend strongly on their physicochemical properties such as base stacking energy, propeller twist angle, bendability, position preference and protein deformability [6]. Oligonucleotide usage in bacterial genomes is strongly influenced by codon usage [9].
OU constraints in bacterial DNA

OUV values of OU patterns from n0_7 mer to n6_7 mer were calculated for the complete genome sequences of Bacillus subtilis 168, Escherichia coli K12 and Pseudomonas putida KT2440 (Fig. 1). OUV of n0_7 mer patterns depends strongly on GC-content getting minima in genomes with a GC content of about 50% such as in E. coli (Fig. 1) and maxima in AT-rich and, especially, GC-rich organisms, probably because OU is more strongly biased in GC-rich sequences [6,11]. Normalization of OU by mononucleotide frequency significantly removes this bias caused by GC-content (Fig. 1 and see ref. [6]). OUV n1_7 mer, however, is still high (Fig. 1). OUV decreases continuously with increase of the word length of internal normalization getting close to zero for n5 and n6 normalization of heptanucleotide usage (Fig. 1). This observation suggests that most OU constraints are caused by mononucleotide frequency and di-, tri- and tetranucleotide combinations while biases in frequencies of longer oligonucleotide words are probably just an extension of constraints of shorter component words.

Local variations of OU patterns

To analyze local variations of OU in bacterial genomes, the sliding window approach was used. 163 bacterial chromosomes of eubacteria and archaea published in the NCBI database were analyzed. Local OU patterns were calculated for 8 kb genome fragments with 2 kb sliding windows [6]. Fig. 2 shows the distances D of local n0_4 mer patterns in three selected bacterial genomes: E. coli K12, P. putida KT2440 and B. subtilis 168 chromosomes. Genomic regions termed the 'core sequences' were characterized by OU patterns being similar to the global pattern of the chromosome. However, multiple genomic loci with alternative OU patterns that can make up more than 10% of the whole genome [11] were also detected in the three tested bacterial genomes (Fig. 2). Locally deviant OU patterns were found to comprise of 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.

These functionally and evolutionarily unrelated subsets of atypical genomic loci were differentiated by the other OU statistical parameters: OUV and PS. These parameters often exhibited extreme values in detected atypical regions, however, their profiles were not congruent to each other. For example, consider the two adjacent gene islands in the P. putida KT2440 genome from 160 kbp to 240 kbp (Fig. 3). The first region (coordinates 170,815 – 180,000 bp) comprises of two tandem operons for ribosomal RNAs (rrnA-rrnA') [12], while the second 26,045 bp sequence covers the largest P. putida gene PP0168 encoding the surface adhesion protein [11]. Both regions were...
recognized by alternative OU patterns (maximal D:n0_4 mer were 59% and 37.5%, respectively, see Figs. 2 and 3). Notably, OUV:n1_4 mer has its genomic minimum (0.08) in the first region but its genomic maximum (0.88) in the second region, whereas PS:n0_4 mer is maximal (74.7%) in the first region and it is closer to the average level (47.5%) in the second region. This example illustrates that the combination of several OU pattern parameters may be useful for the differentiation of unrelated gene subsets.

The application of this procedure to a whole genome is shown in Fig. 4 for the cases of P. putida KT2440 and Mycobacterium leprae TN. Dots corresponding to the genome fragments were plotted in accordance with their D:n0_4 mer, OUV:n1_4 mer and PS:n0_4 mer values. The majority of fragments that represent the core genome clusters in one area. Three outlier groups detected in P. putida KT2440 and in the majority of other tested genomes were termed sections (Fig. 4A). Section I is heterogeneous and includes long intergenic regions, clusters of short hypothetical genes, laterally transferred elements and genes for ribosomal RNAs. The OU patterns of section I are characterized by low OUV and high PS. The operons for ribosomal RNAs exhibited the highest PS values (depicted by red dots, see Fig. 4). Genes for ribosomal proteins are localized in section II. This separation of ribosomal protein genes from the bulk genome was observed in most analyzed bacterial chromosomes but in some slow-growing microorganisms such as M. leprae these genes were not distinct from the core sequence (Fig. 4B). This observation is consistent with the notion that the codon usage in genes encoding ribosomal proteins is separate from the rest of genes in fast-growing bacteria but indistinguishable in slow-growing bacteria [13]. The differential codon usage of fast-growing bacteria has the consequence that ribosomal protein mRNA transcripts utilize other tRNA pools than the other mRNA species for the most abundant amino acids and hence the synthesis of the translational machinery is uncoupled from all other translational demands of the cell [14].

Section III encompasses the regions with outermost OUV (approximately 3 to 15 standard deviations of genomic OUV) and locus-specific OU patterns (large D values). The genetic repertoire covered by these loci is represented
in Table 1. These regions typically comprise of one or more large multidomain genes of over 4 kbp in length or non-coding sequences with multiple tandem repeats. Examples are genes coding for surface proteins (\textit{P. putida} KT2440, \textit{Staphylococcus aureus} N315, \textit{Xylella fastidiosa} Temecula 1), hemagglutinins and hemolysins (\textit{Acinetobacter} sp., \textit{Bordetella bronchiseptica} RB50, \textit{Pseudomonas aeruginosa} PA01, \textit{Pseudomonas syringae} DC3000, \textit{X. fastidiosa} Temecula 1 and \textit{Yersinia pestis} KIM), fatty-acid synthetases (\textit{Corynebacterium efficiens} YS-314) and genes for proteins with an overrepresentation of a few amino acids (\textit{Mycobacterium tuberculosis} H37Rv, \textit{Streptomyces coelicolor} A3(2)).

\textbf{Figure 2}
\textit{Distances D between local n0_4 mer patterns and the global n0_4 mer patterns in the A)E. coli K12; B)P. putida KT2440 and C)B. subtilis 168 chromosomes}. Local patterns were calculated for the sequence fragments of 8 kbp with sliding windows of 2 kbp. The 90% confidence interval of D values is depicted by horizontal lines. The loci with D-values exceeding the genomic confidence interval are considered as gene islands. The abscissa indicates the coordinates of the bacterial chromosomes as they were published in the NCBI database [27].
Many bacterial chromosomes lack these genetic elements. It seems that these genes or multidomain regions are species specific. For example, consider the *M. leprae* genome lacking such genetic elements (Fig. 4B) in comparison with the closely related *M. tuberculosis* H37Rv (Table 1). The genetic elements of section III were not observed in the following tested genomes: *Aeropyrum pernix* K1, Agrobacterium tumefaciens C58, *Aquilarchaeum fulgidus* DSM4304, *Azorarcus* sp. EbN1, *Bacillus anthracis* Ames, *B. subtilis* 168, *Bdellovibrio bacteriovorus* HD100, *Borrelia burgdorferi* B31, *Campylobacter jejuni* NCTC 11168, *E. coli* K12, *Enterococcus faecalis* V583, *Francisella tularensis* Schu 4, *Haemophilus influenzae* KW20, *Halobacterium* sp. NRC1, *Helicobacter pylori* J99, *Lactococcus lactis* IL1403, *Mesorhizobium loti* MAFF303099, *Prochlorococcus marinus* CCMP1375, *Pyrococcus furiosus* DSM 3638, *Salmonella enterica* Ty2, *Shigella flexneri* 2457T, *Streptococcus pneumoniae* R6, *S. pyogenes* MGAS8232, *Treponema pallidum* Nichols. Section I is heterogeneous. The genes for ribosomal RNAs are discerned from the other genes in section I by their extremely high PS of 60 – 70% that are usually the highest values in the genome. For further differentiation of the gene classes in section I, the next chapter describes the strategy to apply further OU statistical parameters to identify the subgroup of horizontally acquired elements.

**Identification of horizontally transferred elements**

Identification of laterally acquired elements in chromosomal sequences is of great importance because genomic islands often comprise pathogenicity and catabolic versatility determinants [15,16]. Two types of normalization of local OU patterns, – internal and global (see above), – were applied to visualize horizontally transferred gene islands within a genome sequence. The reason for introduction of these additional parameters was to improve the discrimination of foreign inserts in genome sequences. In core sequences, where the mononucleotide

---

**Figure 3**

Curves of D:n0_4 mer, PS:n0_4 mer and OUV:n1_4 mer in a locus of the *P. putida* KT2440 genome covering two regions with atypical OU: rRNA-rRNA* gene cluster and a long multidomain gene PP0168 encoding the surface adhesion protein. Local OU patterns were analyzed in 5 kbp sliding windows with steps of 1 kbp. Curves are specified by a color code: blue for D, green for PS and brown for OUV. Protein coding genes are shown by red bars and genes for ribosomal RNAs are shown in black. The abscissa indicates the coordinates of the locus in the chromosome. The upper horizontal line shows the upper boundary of the 95% confidence interval of intragenomic deviation of D values. The lower horizontal line separates genes by their direction of transcription.
content is virtually the same as in the complete genome, results of internal and global normalization are identical in contrast to the laterally transferred loci characterized by an alternative mononucleotide content (in terms of GC-content, G/C-skew and A/T-skew). Correspondingly, values of OUV:n\textsubscript{1} \_4 mer and OUV:n\textsubscript{1} \_g \_4 mer should merge in core sequences but widely diverge in gene islands (Fig. 5A). This concept was proven for genomes with known gene islands: SKIN element in \textit{Bacillus subtilis} 168 [17], phage related gene islands in \textit{P. putida} KT2440 [11] and in \textit{Salmonella enterica} Ty2 [18], pathogenicity island LEE in \textit{E. coli} O157:H7 [19], IS-elements, pathogenicity and prophage islands in \textit{Shigella flexneri} 2457T [20], IS\textit{Ftu}1 element in \textit{Francisella tularensis} Schu4 [21], \textit{cag} pathogenicity island in \textit{Helicobacter pylori} 26695 [2] and 67 kbp gene island in \textit{X. fastidiosa} 9a5c [22]. All mentioned gene islands were successfully localized from the comparison of local with global OU patterns, however, no large foreign regions were observed in sequences of \textit{Bradyrhizobium japonicum} and \textit{Mesorhizobium loti} chromosomes [23,24]. It looks as if these gene islands had been acquired a long time ago and hence their OU patterns adapted to the host genome OU signatures by genome amelioration [4,25].

An example for the identification of a laterally acquired gene island is shown in Fig. 5. The island in the chromosome of \textit{P. putida} KT2440 has significantly divergent OUV:n\textsubscript{1} \_4 mer and OUV:n\textsubscript{1} \_g \_4 mer values and D:n\textsubscript{0} \_4 mer values beyond the 95% confidence interval of the complete chromosome (Fig. 5A). Since OUV:n\textsubscript{1} \_n\textsubscript{mer} and OUV:n\textsubscript{1} \_g \_n\textsubscript{mer} in local patterns and the difference thereof are automatically calculated by the program, the method may be used for the high-throughput identification of horizontally transferred elements in bacterial genomes. Whereas OUV:n\textsubscript{1} \_4 mer and OUV:n\textsubscript{1} \_g \_4 mer values are strongly correlated in the bulk \textit{P. putida} genome, all islands show up by high OUV:n\textsubscript{1} \_g \_4 mer and low OUV:n\textsubscript{1} \_4 mer values (Fig. 5B).

\textbf{Informative assignments of the OU statistical parameters}

The objective of our work was to analyze the informative assignment and applicability of different statistical parameters of OU. Di-, tri- and tetranucleotide usage patterns are charged with most information content (see Fig.
The optimal word length will provide maximal information about the question of interest. First, one has to consider the minimal sequence length that gives reliable OU statistics. The threshold values of the minimum length of sequence were calculated to be 0.3, 1.2, 5 and 20 kbp for di-, tri-tetra- and pentanucleotides, respectively [6]. However, to be informative, the window should of course be not too long, because otherwise short range

**Table 1: Genetic repertoire of loci characterized by atypical tetranucleotide usage patterns and extreme OUV (section III in Fig. 4) identified in bacterial chromosomes**

| Genome                  | Genes and the encoded protein                                                                 | Start† | Length (bp) | \(\Delta D\)† | \(\Delta OUV\)‡ |
|-------------------------|-----------------------------------------------------------------------------------------------|--------|-------------|---------------|----------------|
| *Acinetobacter sp.*     | putative hemagglutinin/hemolysin-related protein                                               | 923,008| 11,136      | 3.11          | 4.13           |
|                         | non-coding multiple repeats TTTAGAAA                                                           | 2,448,000| 5,600       | 2.24          | 17.33          |
| *Bordetella bronchiseptica* RB50 | Bbl186: putative hemolysin                                                                 | 1,268,967| 10,041      | 5.13          | 4.12           |
| *Bradyrhizobium japonicum* USDA110 | bll325: unknown                                                                           | 3,592,327| 17,058      | 3.17          | 4.65           |
|                         | bll356: unknown                                                                                | 3,930,196| 10,326      | 6.23          | 5.02           |
|                         | bll371: unknown                                                                                | 4,106,955| 12,387      | 4.39          | 4.95           |
|                         | bll347: unknown                                                                                | 6,017,600| 12,633      | 5.04          | 6.16           |
| *Corynebacterium efficiens* YS-314 | fasA: fatty-acid synthase I                                                               | 962,188| 8,919       | 2.85          | 3.85           |
|                         | fasB: fatty-acid synthase II                                                                  | 2,541,750| 9,069       | 2.88          | 5.42           |
| *Deinococcus radiodurans* R1 chromosome 1 | non-coding tandem repeats CCGGCC                                                         | 1,465,188| 10,000      | 2.19          | 8.27           |
| *E. coli* O157:H7       | Z0609, Z0615: RTX family exoproteins                                                          | 581,356| 20,160      | 1.82          | 9.43           |
| *Mycobacterium tuberculosis* H37Rv | Rv0272c-Rv0279c: hypothetical Gly-, Ala-rich proteins                                       | 328,573| 10,499      | 1.52          | 9.15           |
|                         | Rv0297-Rv0304c: hypothetical Gly-, Ala-, Asn-rich proteins                                     | 361,332| 11,431      | 8.79          | 7.91           |
|                         | Rv0355c: Asn-rich protein                                                                     | 424,775| 9,903       | 8.31          | 10.91          |
|                         | Rv0573c-Rv0578c: hypothetical Gly-rich proteins                                               | 665,849| 10,066      | 0.60          | 4.72           |
|                         | Rv0742-Rv0747: hypothetical Gly-rich proteins                                                 | 832,979| 7,876       | 1.24          | 3.97           |
|                         | Rv1060-Rv1068c: hypothetical Gly-, Ala-rich proteins                                           | 1,183,506| 8,641       | 1.04          | 5.54           |
|                         | Rv1084-Rv1092c: hypothetical proteins                                                         | 1,207,634| 11,395      | 2.19          | 6.44           |
|                         | multiple repeats CCGCCGCC                                                                       | 1,630,636| 7,592       | 2.33          | 8.84           |
| *Pseudomonas aeruginosa* PA01 | PA1874: hypothetical protein                                                                  | 2,036,441| 7,407       | 2.61          | 5.61           |
| *P. putida* KT2440      | PA168: Thr-rich surface adhesion protein                                                       | 194,494| 26,046      | 2.58          | 6.97           |
| *P. syringae* DC3000 | PSPTO3229: filamentous hemagglutinin                                                           | 3,629,677| 18,825      | 2.34          | 7.87           |
| *Rhodopirellula baltica* 1 | RB3077: putative cyclic nucleotide binding protein                                             | 1,588,083| 18,024      | 1.62          | 6.19           |
|                         | RB4375: large polymorphic membrane protein, probable extracellular nucleic acid               | 2,242,933| 9,171       | 3.23          | 7.09           |
|                         | RB17769: probable aggregation factor core protein MAFp3                                       | 6,335,006| 24,522      | 5.25          | 6.31           |
| *Rhodopseudomonas palustris* CGA009 | conserved hypothetical protein                                                                 | 1,459,664| 8,991       | 2.61          | 3.38           |
|                         | conserved hypothetical protein                                                                | 1,475,303| 13,008      | 2.89          | 4.18           |
| *Sulfolobus solfataricus* P2 | non-coding tandem repeats GAATTGAAAG                                                            | 1,228,221| 12,238      | 1.94          | 15.25          |
| *Staphylococcus aureus* N315 | ebbA – ebbB: large surface anchored proteins                                                      | 1,437,928| 20,142      | 4.04          | 10.07          |
|                         | SA2447: similar to streptococcal hemagglutinin                                                  | 2,755,253| 6,816       | 3.03          | 9.29           |
| *Streptomyces coelicolor* A3(2) | SCBF4.01c: Ala/Glu-rich protein                                                                | 586,509| 3,981       | 2.16          | 5.40           |
|                         | SC2H4.02: hypothetical protein                                                                | 6,836,057| 6,552       | 2.86          | 4.80           |
| *Xanthomonas campestris* ATCC33913 | yopH: putative autotransporter adhesin                                                            | 2,374,740| 11,886      | 3.22          | 6.61           |
| *Xylella fastidiosa* Temecula 1 | non-coding sequence, multiple                                                 | 1,183,606| 11,095      | 1.31          | 9.81           |
|                         | repeats (GTT), pspA1: hemagglutinin                                                            | 1,447,312| 11,139      | 1.37          | 10.91          |
|                         | pspA2: hemagglutinin                                                                          | 2,082,143| 10,134      | 1.06          | 9.78           |
| *Yersinia pestis* KIM | sipA-2: yersiniabactin peptide/polyketide synthetase; yopH: putative autotransporter adhesin | 2,654,642| 15,867      | 4.27          | 6.05           |
|                         | y3579: putative filamentous hemagglutinin                                                       | 3,747,888| 11,133      | 2.66          | 8.60           |
|                         | ypspA: putative filamentous hemagglutinin                                                       | 3,961,333| 9,888       | 3.31          | 4.32           |

* left coordinate of the locus in the chromosomal sequence;
† deviation of the D:n0.4 mer value calculated for the locus from the mean genomic D:n0.4 mer in standard deviations;
‡ deviation of the OUV:n1.4 mer value calculated for the locus from the mean genomic OUV:n1.4 mer in standard deviations;
Figure 5
Gene islands in the *P. putida* KT2440 genome identified by discordant OUV:n1₄mer and OUV:n1₄mer values A) in a local gene map and B) globally in the complete genome. Genome fragments of 8 kbp were generated with a sliding window in step of 2 kbp. Red bars in figure A indicate protein coding genes and black bars-hypothetical genes. The horizontal line in the part A separates genes by direction of transcription. The yellow-shaded 8 kbp long fragment in A corresponds to the red dot indicated by an arrow in B.
fluctuations of OU will vanish. We recommend that the window should not be longer than 10-fold of its minimal length. Tetranucleotide (and, sometimes, pentanucleotide) usage patterns are more appropriate for the global analysis of sequences. A long sliding window silences signals from the local repeats and structural biases at the level of individual genes so that the characteristics of whole operons and gene islands become apparent. For a more detailed analysis of chromosomal loci or short genomes of bacterial plasmids and phages, tri- and dinucleotide usage patterns may be more appropriate. For example, in Fig. 6 the mosaic structure of the plasmid pKLC102 was recovered by investigation of local trinucleotide usage patterns (genomic fragments were segregated by 1.2 kbp sliding windows in steps of 200 bp). Three peaks of high D values depict recombination sites of the plasmid where additional genetic elements (transposons, integrons and gene cassettes) may be inserted [26]. A region with extremely high OUV:n1_3 mer corresponds to the putative replication origin of the plasmid [26].

To check whether the local fluctuations of OU parameters are statistically valid, a sequence of 100 kbp of mononucleotide content similar to pKLC102 was randomly generated. The ranges of 3-sigma fluctuation of D:n0_3 mer and OUV:n1_3 mer in the random sequence are depicted in Fig. 6 by vertical grey bars along the corresponding D and OUV axes. In the real sequences these values vary over a significantly larger range with the mean value of D smaller and the mean OUV higher than in the randomly generated sequence. (The plasmid pKLC102 sequence and the randomly generated sequence are included in the additional files as examples of source data files pKLC102.fts and random.fts, respectively.)

Normalization of OU by the internal component words changes the information assignment of OU biases. The three parameters D, PS and OUV were calculated for n0_4 mer, n1_4 mer, n2_4 mer and n3_4 mer local patterns for the pKLC102 genome and a part of the E. coli K12 chromosome from 1 Mbp to 2 Mbp. The former one is an example of a mosaic genome, and the latter one represents a regular bacterial chromosome. Correlation coefficients were calculated for respective OU statistical parameters determined for non-normalized and normalized local OU patterns. The correlation coefficients varied between 0.10 and 0.89 for pKLC102 and between 0.46 and 0.94 for E. coli (Table 2). This data demonstrates that n0, n1, n2 and n3 of 4 mer local patterns measure different characteristics of a sequence. In other words, the statistical parameters with different types of normalization provide non-redundant information that can be exploited for a refined anal-
ysis of genome organization. In case of tetranucleotide usage analysis four types of patterns exist: n0_4 mer, n1_4 mer, n2_4 mer and n3_4 mer. Each pattern type can be characterized by three parameters, D, PS and OUV that provide in total a comprehensive set of 12 non-redundant parameters for the nucleotide sequence analysis. Moreover, two subtypes of normalized OU patterns were introduced above, – with internal and global normalization, – that results in a total set of 21 non-redundant tetranucleotide usage statistical parameters each suitable for the refinement of functional gene classes in a raw nucleotide sequence.

**Conclusion**

Bacterial genomes are not homogeneous but contain polymorphic blocks including horizontally transferred gene islands, non-coding sequences, long multidomain genes and ancient conserved gene clusters. The structural polymorphism of bacterial genomes may be effectively analyzed by local OU pattern signatures. A set of statistical approaches has been designed to perform this structural analysis of nucleotide sequences. Moreover, two subtypes of normalized OU patterns were introduced above, – with internal and global normalization, – that results in a total set of 21 non-redundant tetranucleotide usage statistical parameters each suitable for the refinement of functional gene classes in a raw nucleotide sequence.

**Table 2: Correlation coefficients between D, PS and OUV of n0_4 mer local patterns with those of the corresponding n1, n2 and n3 normalized patterns**

| Parameters       | n1_4 mer |                  | n2_4 mer |                  | n3_4 mer |
|------------------|----------|-----------------|----------|-----------------|----------|
|                  |          | Normalization type |          |                  |          |
| plasmid pKLC102, window 5,000 bp, step 2,500 bp |          |                  |          |                  |          |
| D:n0_4 mer       | 0.85*    | 0.82            | 0.40     |                  |          |
| PS:n0_4 mer      | 0.40     | 0.60            | 0.10     |                  |          |
| OUV:n0_4 mer     | 0.89     | 0.83            | 0.39     |                  |          |
| 1 Mbp-2 Mbp locus of E. coli K12 chromosome, window 10,000 bp, step 5,000 bp |          |                  |          |                  |          |
| D:n0_4 mer       | 0.94     | 0.84            | 0.63     |                  |          |
| PS:n0_4 mer      | 0.88     | 0.75            | 0.53     |                  |          |
| OUV:n0_4 mer     | 0.61     | 0.46            | 0.35     |                  |          |

*Values in the cells of the table indicate the correlation coefficients between respective OU statistical parameters D, PS and OUV determined for n0 patterns and the normalized patterns n1, n2 and n3. For example, 0.85 is the correlation coefficient between series of values D:n0_4 mer and D:n1_4 mer determined for overlapping 5 kbp fragments of pKLC102.

A command line version of the Python program to apply the OU statistics methods mentioned above is available as additional file. To run the program, first the Python interpreted language program must be downloaded from the Web-site [http://www.python.org/download/](http://www.python.org/download/) and installed on the computer. The source DNA sequence (or sequences) should be saved in FASTA format in text file(s) with .FST file name extensions. Users may choose the OU statistical parameters to be calculated and the parameters of the sliding window by setting corresponding command line arguments. Many different OU parameters may be determined by a single run of the program and all FST files in the target folder will be processed continuously in a batch. For each source data file an output file in TXT format will be saved in the same folder. The full list of arguments and description of how to use the program are documented in the readme.doc file provided in the additional files. The program is fast enough to calculate all set of OU parameters mentioned in this paper for a complete bacterial genome of average length in 10–20 min depending on the computer performance.

Several general conclusions about OU in bacteria can be drawn from this report. First, most OU constraints are hidden in di-, tri- and tetranucleotide combinations that vanish with increasing word length (see Fig. 1). For example, in case of a hexamer the four possible heptamer words will have the same likelihood to occur next in the sequence. Hence, i) the analysis of the oligonucleotide distribution of up to 4-mers is sufficient to uncover all OU constraints in the sequence; and ii) neighbor effects are limited to dipeptides so that protein evolution is not skewed by oligonucleotide biases. Second, D and PS values are correlated in local patterns (see the examples for D:n0_4 mer and PS:n0_4 mer in Fig. 3 and 4). This observation is in accordance with the general trend in bacterial sequences to keep parity of frequencies of words and their reverse complements, in other words- a trend towards minimal PS [6]. OU parity is most pronounced for the OU pattern of the whole chromosome, whereas fluctuations of OU in local patterns lead to an increased PS. The exceptions are the laterally transferred elements with their...
island-specific OU signature. In this case, large D values of the local OU patterns may be associated with low PS (see blue and green dots in section I in Fig. 4).

Methods
Sequences of 163 bacterial chromosomes including eubacterial and archaean genomes published in the NCBI database [27] were analyzed in this study.

The OU statistical parameters-variance of word deviations (OUV); distances between patterns (D); pattern skew between lagging and lagging strand (PS) were calculated by applying the algorithms described previously [6]. In a sequence of \( L_{seq} \) nucleotides we calculated numbers of occurrence of overlapping \( N \)-long oligonucleotide words. There are \( 4^N \) possible combinations of nucleotides and the total number of words in a sequence corresponds to the sequence length \( L_{seq} \). OU pattern was denoted as a matrix of deviations \( \Delta f_{[\xi_1, ... \xi_N]} \) of observed from expected counts for all possible words of the length \( N \):

\[
\Delta f_{[\xi_1, ... \xi_N]} = (C_{[\xi_1, ... \xi_N]} \text{obs} - C_{[\xi_1, ... \xi_N]} \text{e}) / C_{[\xi_1, ... \xi_N]} \text{e} \\
\text{where } \xi \text{ is any nucleotide A, T, G or C at the position 1, } 2, 3, ... N \text{ in the } N\text{-long word;} \ C_{[\xi_1, ... \xi_N]} \text{obs} \text{ is the observed count of the word, } [\xi_1, ... \xi_N]; \ C_{[\xi_1, ... \xi_N]} \text{e} \text{ is the expected count and } C_{[\xi_1, ... \xi_N]} \text{e} \text{ is a standard count estimated from the assumption of an equal distribution of words in the sequence: } (C_{[\xi_1, ... \xi_N]} \text{e} = L_{seq} \times 4^{-N}).
\]

OU parameters of words of length \( N \) were normalized by shorter words \( n \) (\( 0 \leq n < N \)) as follows:

\[
\bar{C}_{[\xi_1, ... \xi_N]} \text{e} = C_{[\xi_1, ... \xi_N]} \text{e} / C_{[\xi_1, ... \xi_N]} \text{e} \\
\text{and } \bar{C}_{[\xi_1, ... \xi_N]} \text{e} = C_{[\xi_1, ... \xi_N]} \text{e} / C_{[\xi_1, ... \xi_N]} \text{e} \text{ if OU is normalized by empirical frequencies of all shorter words of the length } n. \text{ The normalization was performed as follows. First at all, we calculated observed frequencies } \bar{f}_{[\xi_1, ... \xi_N]} \text{ of } n\text{-long words in the sequence. Each word of length } N \text{ can be represented as a consecutive set of } N - n + 1 \text{ overlapping component words of length } n. \text{ For example, a pentamer ATGGC can be expressed as a set of 4 overlapping dimers: AT, TG, GG and GC. In a general case of a } N\text{-long word, a component word } [\xi_1, ... \xi_n] \text{ reduces the set of available options for the next word in the sequence to 4 possible oligonucleotides: } [\xi_2, ... \xi_n, A], [\xi_2, ... \xi_n, T], [\xi_2, ... \xi_n, G] \text{ and } [\xi_2, ... \xi_n, C]. \text{ The relative frequencies of these words are:}
\]

\[
\bar{f}_{[\xi_2, ... \xi_n, A]} \times ([\bar{f}_{[\xi_2, ... \xi_n, A]} + \bar{f}_{[\xi_2, ... \xi_n, T]} + \bar{f}_{[\xi_2, ... \xi_n, G]} + \bar{f}_{[\xi_2, ... \xi_n, C]})]^{-1}
\]

whereby the \( F \) values are the observed frequencies of the particular word of length \( n \) in the complete sequence and \( \xi \) is any nucleotide A, T, G or C. The expected count of a word \([\xi_1, ... \xi_n] \text{ of length } N \text{ in a } L_{seq} \text{ long sequence normalized by frequencies of } n\text{-mers } (n < N) \text{ was calculated as follows:}

\[
C_{[\xi_1, ... \xi_n]} \text{e} = L_{seq} \times \bar{f}_{[\xi_1, ... \xi_n]} \prod_{i=2}^{N-n+1} \left( \sum_X \frac{\bar{f}_{[\xi_1, ... \xi_n, x]} \bar{f}_{[\xi_1, ... \xi_n, x]}}{A,T,G,C} \right) \\
\text{For further processing of OU statistics, the words were sorted by their } \Delta f_{[\xi_1, ... \xi_n]} \text{ and the ranks of words instead the real values of deviations of observed from expected counts were used. The rank values (from 1 to 256 in the case of tetranucleotide analysis) were assigned to the words in accordance with their } \Delta f_{[\xi_1, ... \xi_n]} \text{ values by ordering the words from the most overrepresented one (the greatest } \Delta f_{[\xi_1, ... \xi_n]} \text{ to the least represented one (the lowest } \Delta f_{[\xi_1, ... \xi_n]} \text{). This approach made the OU statistical parameters free from any dependence on the sequence length, provided that the sequence has a minimum length } L_{min} \text{ so that in a random sequence of the same length } L_{min} 95\% \text{ of all words of length } N \text{ occur at least ten times (see above and [6]). Hence, local OU patterns that meet these requirements could be compared with the global pattern.}

The distance } D \text{ between two patterns was calculated as the sum of absolute distances between ranks of identical words } (w) \text{ in a total } 4^N \text{ different words } (w) \text{ in patterns } i \text{ and } j \text{ as follows:}

\[
D(\%) = 100 \times \frac{\sum_{w} |\text{rank}_{w,i} - \text{rank}_{w,j}|}{D_{max} - D_{min}} - D_{min}
\]

PS is a particular case of } D \text{ where patterns } i \text{ and } j \text{ were calculated for the same DNA but for direct and reversed strands, respectively. } D_{max} = 4^N(4^N - 1)/2 \text{ and } D_{min} = 0 \text{ when calculating a } D, \text{ or, in a case of PS calculation, } D_{min} = 4^N \text{ if } N \text{ is an odd number or } D_{min} = 4^N - 2N \text{ if } N \text{ is an even number [6].}

The definition of OUV was provided in our previous paper [6].

The random sequence was generated by a in-house program using the Python randomizer [28].
List of abbreviations
OU – oligonucleotide usage;
OUV – oligonucleotide usage variance;
PS – pattern skew;
D – distance between two OU patterns of an identical type.

Authors’ contributions
ONR did Python programming. Both authors contributed equally to all other presented data.

Additional File material

Additional File 1
There is an additional ZIP archive file OligoWords for BMC Bioinf.zip comprising following documents: OligoWords1.1.exe.py - a command line version of the program implemented in Python2.2 [28]. readme.doc - description of the project in Word97 format. plKC102.fst - sequence of the plasmid pKLC102 [26] in FASTA format that may be used as a source data file for the program OligoWords1.1.exe.py (see readme.doc). random.fst - a randomly generated sequence comparable with one of the plasmid pKLC102 by length and mononucleotide content. The file is in FASTA format that may be used as a source data file for the program OligoWords1.1.exe.py (see readme.doc). Click here for file. (http://www.biomedcentral.com/content-supplementary/1471-2055-6-251-S1.zip)

Acknowledgements
This work was supported by the DFG-sponsored Europäisches Graduiertenkolleg 653.

References
1. Noble PA, Citek RW, Ogunseitan OA: Tetranucleotide frequencies in microbial genomes. Electrophoresis 1998, 19:528-535.
2. 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.
3. Abe T, Kansaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T: Informatics for unveiling hidden genome signatures. Genome Res 2003, 13:693-702.
4. Pride DT, Meinersmann RJ, Wassenaa TRM, Blaser MJ: Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Res 2003, 13:145-155.
5. Teeling H, Waldmann J, Lombardot T, Bauer M, Göckner FO: TETRA: a web service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics 2004, 5:163.
6. Reva ON, Tümmel B: Global features of sequences of bacterial genomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. BMC Bioinformatics 2004, 5:90.
7. Karlin S: Global dinucleotide signatures and analysis of genomic heterogeneity. Curr Opin Microbiol 1998, 1:598-610.
8. Karlin S, Mrázek J, Campbell A: Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 1997, 179:3899-3913.
9. Gorban AN, Popova TG, Zinovyev AY: Four basic symmetry types in the 7-cluster structure of microbial genomic sequences. In Silico Biol 2005, 5:0025.
10. Weinel C, Ussery DW, Ohlsson H, Sicheritz-Ponten T, Kiewitz C, Tümmel B: Comparative genomics of Pseudomonas aeruginosa PA01 and Pseudomonas putida KT2440: orthologs, codon usage, REP elements and oligonucleotide motif signatures. Genome Letters 2002, 1:175-187.
11. Weinel C, Nelson KE, Tümmel B: Global features of the Pseudomonas putida KT2440 genome sequence. Environ Microbiol 2002, 4:809-818.
12. Weinel C, Tümmel B, Hilbert H, Nelson KE, Kiewitz C: General method of rapid Smith/Birnstiel mapping adding for gap closure in shotgun microbial genome sequencing projects: application to Pseudomonas putida KT2440. Nucleic Acids Res 2001, 29:El10.
13. Carbon E, Zinovyev A, Képès C: Codon adaptation index as a measure of dominating codon bias. Bioinformatics 2003, 19:2005-2015.
14. Kiewitz C, Weinel C, Tümmel B: Codon usage index of Pseudomonas aeruginosa: a codon index that utilizes whole genome sequence data. Genome Letters 2002, 1:61-70.
15. Hacker J, Kaper JB: Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol 2000, 54:641-679.
16. van der Meer JR, Sentchilo V: Genomic islands and the evolution of catabolic pathways in bacteria. Curr Opin Biotechnol 2003, 14:248-254.
17. Sato T, Kobayashi Y: The ars operon in the skin element of Bacillus subtilis confers resistance to arsenate and arsenite. J Bacteriol 1998, 180:1655-1661.
18. Deng W, Liu SR, Plunkett 3rd, Mayhew GF, Rose DJ, Burland V, Kodiyani V, Schwartz DC, Blattner FR: Comparative genomics of Salmonella enterica serovar Typhi strains Ty2 and CT18. J Bacteriol 2003, 185:2330-2337.
19. Perna NT, Mayhew GF, Posfai G, Elliott S, Donnenberg MS, Kaper JB, Blattner FR: Molecular evolution of a pathogenicity island from enterohemorrhagic Escherichia coli O157:H7. Infect Immun 1998, 66:3810-3817.
20. Wei J, Goldberg MB, Burland V, Venkatesan MM, Deng W, Fournier G, Mayhew GF, Plunkett 3rd, Donnenberg MS, Kaper JB: Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T. Infect Immun 2003, 71:2775-2786.
21. Larsson P, Oyston PC, Chain P, Chu MC, Duffield M, Fuxelius HH, Garcia E, Halltorp G, Johansson D, Isherwood KE, et al.: The complete genome sequence of Francisella tularensis, the causative agent of tularemia. Nat Genet 2005, 37:153-159.
22. Simpson AJ, Reinach FC, Arruda P, Abreu FA, Acencio M, Alvarenga R, Alves LP, Araya JE, Baia GS, Baptista CS, et al.: The complete genome sequence of Francisella tularensis H1R5, the causative agent of tularemia. J Bacteriol 2003, 185:1758-1765.
23. Kaneko T, Nakamura Y, Sato S, Asamizu E, Kato T, Sasamoto S, Watanabe A, Idesawa K, Iriguchi M, Kawashima K, Watanabe A, Iida A, Ishikawa T, Sasamoto S, Watanabe A, Idesawa K, Kawashima K, et al.: The complete genome sequence of Francisella tularensis, the causative agent of tularemia. J Bacteriol 2003, 185:1758-1765.
24. Kaneko T, Nakamura Y, Sato S, Minamisawa K, Uchiumi T, Sasamoto S, Watanabe A, Idesawa K, Kawashima K, et al.: The complete genome sequence of Francisella tularensis, the causative agent of tularemia. J Bacteriol 2003, 185:1758-1765.
25. Lawrence JG, Ochman H: Comparison of microbial genome sequences reveals unexpected diversity of DNA replication systems. J Bacteriol 2003, 185:1758-1765.
26. Klockgether J, Reva O, Larbig K, Tümmler B: Global differences in the 7-cluster structure of microbial genomic sequences. In Silico Biol 2005, 5:0025.