Comparative Genomics of *Pseudomonas aeruginosa* PAO1 and *Pseudomonas putida* KT2440: Orthologs, Codon Usage, Repetitive Extragenic Palindromic Elements, and Oligonucleotide Motif Signatures

Christian Weinel,1,* David W. Ussery,2 Hakan Ohlsson,2 Thomas Sicheritz-Ponten,2 Claudia Kiewitz,1 and Burkhard Tümmler1

1 Klinische Forschergruppe, OE 6711, Medizinische Hochschule Hannover, D-30625 Hannover, Germany
2 Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Lyngby, Denmark

ABSTRACT: *Pseudomonas aeruginosa* PAO1 and *Pseudomonas putida* KT2440 were compared for global genome features. About half of the coding sequences share more than 50% amino acid sequence identity. The alignment of orthologs shows an X-type pattern, with large sections of conserved gene order indicative of symmetric inversions around the origin and terminus of replication since the two taxa diverged from a common ancestor. However, codon usage patterns are different in the two type I pseudomonads, as indicated by unrelated genome codon index and codon adaptation index profiles. The *P. aeruginosa* and *P. putida* chromosomes share similar fractions of global repeats and contain several hundred repetitive extragenic palindromic (REP) elements, part of which are organized in regularly spaced clusters as bacterial interspersed mosaic elements. The frequency and map position of all 2-meric to 14-meric oligonucleotides were determined in the two genomes. Among the short oligonucleotides the universally rare TA, CTA, TAG, and CTAG and stretches of oligoG and oligoC are strongly counterselected in the GC-rich pseudomonads. Frequent words encode peptides with strong codon bias, such as the tripeptide LLL.

Keywords: Codon Adaptation Index, Genome Codon Index, Genome Evolution, Synteny, Oligonucleotide, Ortholog, *Pseudomonas aeruginosa*, *Pseudomonas putida*, Repetitive Extragenic Palindromic Sequence, Bacterial Interspersed Mosaic Elements.

1 INTRODUCTION

Pseudomonads are ubiquitous, metabolically versatile Gram-negative bacteria [1]. *Pseudomonas putida* is able to colonize soil, water, and the plant rhizosphere [2]. This nonpathogenic saprophytic species is considered to be innocuous for the environment, and hence *P. putida* strain KT2440 is the first Gram-negative soil bacterium to be certified as a safety strain by the Recombinant DNA Advisory Committee [3] and is the preferred host for cloning and gene expression for Gram-negative soil bacteria [4]. The type species of the genus *Pseudomonas aeruginosa* is also extremely versatile metabolically and is found in numerous aquatic and soil habitats in the environment [5], but in contrast to *P. putida* it is also an important opportunistic pathogen for plants, animals, and humans [5]. *P. aeruginosa* causes disease in cystic fibrosis, burn, cancer, and intensive-care patients associated with substantial morbidity and mortality [5].

The 6.2-Mb large genomes of the reference strains *P. putida* KT2440 [6] and *P. aeruginosa* PAO1 [7] were recently completely sequenced. The two strains share homology for 77% of the predicted coding regions [6]. Nelson et al. [6] performed a comparative metabolic analysis
between the two strains. In this paper, the first comparison of global features of the PAO1 and KT2440 genomes is presented. We analyzed DNA structural parameters [8], codon usage expressed by codon adaptation (CAI) and genome codon indices (GCI) [9], and the genome localization and sequence similarity of homologous genes.

Another major objective was the identification of repeats and of the most abundant and least frequent oligonucleotides (di- to tetradeanucleotides). These oligonucleotides constitute a library of the rarest and the most common “words” in the genome. We wanted to determine the extent to which these oligonucleotide motifs are shared by the two phylogenetically related pseudomonads and whether they are a measure of the genomic signature.

Highly repetitive oligonucleotide motifs are known in other taxa to be often associated with regulatory or structural elements [10]. For example, chi sites and REP elements (repeated extragenic palindrome of unknown function) are abundant in enterobacteria [11]. REP sequences had already been detected in the P. putida KT2440 genome [12], and hence an analogous search for REP elements was undertaken in the P. aeruginosa PAO1 genome.

2. MATERIALS AND METHODS

2.1. Sequences

The complete genome sequences and the annotation tables of P. aeruginosa PAO1 and P. putida KT2440 were obtained from the Pathogenesis (http://www.pseudomonas.com) and TIGR (http://www.tigr.org) web sites, respectively. The annotated DNA sequence of Escherichia coli K-12 genome was downloaded from the National Center for Biotechnology Information ftp server (ftp://ncbi.nlm.nih.gov).

2.2. Analyses

2.2.1. Structural Parameters. Three different types of structural characteristics were predicted by applying algorithms extracted from the literature as described before [8]: DNA curvature, DNA stacking energy, and DNA flexibility. Briefly, the structural values are assigned to every nucleotide in a DNA sequence by looking up the values for corresponding di- or trinucleotides in a table; in the case of curvature, the value for a 21-bp window was calculated. The level of each parameter was depicted on the circles of DNA Atlases [13] as differences in color and color intensity.

2.2.2. Codon Usage. Codon usage patterns of P. putida KT2440 and P. aeruginosa PAO1 genes were analyzed with the in-house program NiceCUS and the program CodonW (written by John Peden and available from ftp://molbiol.ox.ac.uk/cu). The relative synonymous codon usage (RSCU) was determined for each gene, which is the observed frequency of a particular codon divided by its expected frequency under the assumption of equal usage of the synonymous codons for an amino acid. The GCI provides a quantitative measure of the synonymous codon bias of a particular gene compared with the average codon bias in all chromosomal genes [9]. The GCI was defined in analogy to the CAI [14], which indicates the relative adaptiveness of the codon usage of a particular gene to a set of highly expressed genes. For P. putida this set was established according to that of E. coli and includes 16 genes of ribosomal proteins (rpsL, rpsE, rpsB, rpsA, rpsL, rplL, rplB, rplQ, rplM, rplA, rplX, rplT, rplO, rplC), three genes of elongation factors (tuf-1, tuf-2, efp), and the chaperonin gene groEL. For definition of the reference gene set of highly expressed genes of P. aeruginosa see Kiewitz et al. [9]. GCI and CAI values were calculated for each gene of the P. putida and the P. aeruginosa chromosomes. A correspondence analysis [15] of relative synonymous codon usage (RSCU) values was carried out with CodonW to identify a set of genes showing the highest bias in codon usage. Correspondence analysis plots all genes according to their codon usage in a 59-dimensional space and identifies the most important trends in codon usage as those axes through this multidimensional hyperspace that account for the largest variation among genes. The 5% most highly biased genes positioned at one extreme of the principal axis were employed as the reference gene set to calculate CAI values for each gene, in analogy to Sharp and Li [14].

2.2.3. Absolute Oligonucleotide Frequencies in the Whole Genome. Absolute frequencies (observed values) of all 2-mers to 14-mers were calculated for each genome. Based on the oligonucleotides one nucleotide shorter, the expectation values were calculated with the use of an n-1 Markov chain model. For each oligonucleotide, the ratio between observed and expected values was subjected to a \( \chi^2 \) test, and then the oligonucleotides were sorted by decreasing \( \chi^2 \) values. Over- and underrepresented oligonucleotides with a level of significance of \( \chi^2 > 10,000 \) were selected for further analysis.

2.2.4. Oligonucleotide Frequencies in Genome Sections. Dinucleotide bias and variance of tetradeanucleotide frequency were calculated in sliding windows of 4 kb with a step size of 1 kb in the complete genome sequences of the KT2440 and PAO1 strains. Dinucleotide relative abundances (dinucleotide bias) were calculated as odds ratios \( p_{XY} \) of the frequencies of the dinucleotide XY and its intrinsic mononucleotides X and Y. To measure the distance between the dinucleotide biases of a certain region and the average value of the genome, we calculated the differential dinucleotide relative abundances as \( \delta_{di} = \frac{1}{16} \sum |p_{XY}(\text{sequence window}) - p_{XY}(\text{genome})| \) [16]. The variance of tetradeanucleotide frequencies was calculated for each window as \( \delta_{n} = \frac{1}{1256} \sum |f_{i}(\text{observed}) - f_{i}(\text{expected})| \), where \( f_{i}(\text{observed}) \) denotes the observed
frequency of a tetranucleotide and \( f_T \) (expected) the expected frequency if all tetranucleotide frequencies are equal [17].

2.2.5. Repeat Sequences. Global repeats were defined as sequences of at least 100 bp in length that share at least 80% sequence identity. Global repeats can be located on the same strand (global direct repeat) or opposite strand (global inverted repeat) [8]. For the identification of global repeats the chromosomal DNA sequence was compared against itself [13]. REP elements [11, 12] in the \( P. \) aeruginosa chromosome were reconstructed from overrepresented oligomers that shared the two motifs, 5'-CGCGGATAAC and 5'-GTTATTCGCCCTAC, at a distance of about 10 nucleotides from each other. The approximate length of the repeat sequence was determined by a BLAST [18] search of the whole chromosome, and then the map positions of the extended 40 ± 1 bp sequence (mismatch penalty −2) were determined by a further BLAST search. All hits of > 30 bp were collected, redundant hits were eliminated, and the remaining oligomers were sorted by ClustalX. Finally all oligonucleotides were removed from the list that exhibited more than 10 mismatches in the core of 32 conserved residues. The consensus sequence was displayed by “sequence logos” [19].

Gene-by-gene comparisons were performed by sequence alignment using BLAST [18]. Only best hits were counted with an expectation value of \(< 10^{-5}\) or an amino acid sequence identity of at least 25%. Moreover, homologous stretches had to cover at least 60% of the coding sequence.

3. RESULTS AND DISCUSSION

3.1. Codon Usage

Codon usage in bacteria is driven by genomic and translational constraints. We calculated the GCI [9] and the CAI [14] for all genes of the \( P. \) putida KT2440 and \( P. \) aeruginosa PAO1 chromosomes, whereby in case of the CAI the reference gene set either was given in analogy to \( E. \) coli by an \textit{a priori} defined set of highly expressed genes or was empirically determined by the set of genes with the strongest codon bias. The GCI compares the RSCU values of a gene with the maximum genomic RSCU values, and hence the GCI indicates the relative adaptability of the codon usage of a particular gene to the preferential codon usage in the genome. On the other hand, the CAI indicates the relative adaptability of the codon usage of a particular gene to the codon usage of a set of highly expressed genes or a set of genes with high codon bias.

Figure 1 displays the distribution of GCI and CAI values in the two \textit{Pseudomonas} genomes. The GCI and two CAI indices are consistently high for the \( P. \) aeruginosa PAO1 genes and exhibit largely overlapping skewed Gaussian distributions [9]. In contrast to \( P. \) aeruginosa, the GCI and CAI of \( P. \) putida KT2440 genes are generally not correlated. The three frequency curves have different shapes, means, and variances (Fig. 1).

Codon usage is homogeneous in \( P. \) aeruginosa, but heterogeneous in \( P. \) putida. Codon usage in the latter is more related to the phylogenetically more distant enterobacteria [20] than to the type species of the genus \textit{Pseudomonas}. This unexpected finding can be explained at least partly by two differentially operating genomic constraints in the two related taxa. First, the mean \( G+C \) content in \( P. \) aeruginosa is 67.2% [7], whereas it is 61.6% in \( P. \) putida [21]. The \( G+C \) content was calculated to account for 73% of codon usage variation in \( P. \) aeruginosa [9], and correspondingly GC mutational bias is the major determinant of GCI and CAI in this taxon. In the less \( G+C \)-rich \( P. \) putida, other constraints such as translational demands should be more important than the GC bias. Second, \( P. \) aeruginosa has a more homogeneous genome architecture. The uniform
pattern of consistently high GCI values in PAO1 is interrupted by only 15 gene islands with a *Pseudomonas* atypical codon usage [9]. In contrast, 105 gene islands with atypical G+C content and/or oligonucleotide signature are scattered throughout the *P. putida* KT2440 chromosome [21]. Hence, differential G+C bias and mosaicism make it plausible that the codon usage pattern would be so different in the two related taxa.

### 3.2. Orthology and Synteny: Common Heritage of the Two Genomes

After determining that codon usage had evolved differently in *P. putida* and *P. aeruginosa*, we became suspicious about the extent to which gene order and coding information have been retained by the two species since their divergence from a common ancestor. First, we searched by BLASTP for orthologs in the two genomes defined by an amino acid sequence identity of at least 25%, a sequence overlap in the alignment of at least 60%, and a cutoff expectancy value of the alignment below $10^{-5}$. According to these criteria 3,351 orthologs were identified among the 5,570 PAO1 and 5,420 KT2440 genes.

Large sections of conserved gene order (synteny) were detected by comparing the genome locations of orthologs between *P. putida* and *P. aeruginosa* (Fig. 2). The alignment of these two genomes shows an X-type pattern indicative of symmetric inversions around the origin and terminus of replication, as has been observed before for several pairs of related genomes [22]. The lack of synteny near the terminus of replication suggests that constraints on gene order are stronger near the origin. Conserved gene order around the origin, but not around the terminus of replication, has already been inferred from genetic mapping of auxotrophic markers [23, 24] in *P. aeruginosa* and *P. putida* and is now confirmed by complete genome sequence data.

The genetic mapping data from the 1960s and 1970s [23, 24], moreover, provided evidence that genetic material that is not homologous with *P. aeruginosa* DNA clusters in the auxotroph-poor region around the terminus of replication of the *P. putida* chromosome. The genome sequence corroborates this suspicion from the early days of *Pseudomonas* genetics (Fig. 3). Whereas the nonhomologous genes that are absent from the genome of *P. putida* are evenly distributed on the *P. aeruginosa* PAO1 chromosome, the *P. putida* genes that are absent from the genome of *P. aeruginosa* are overrepresented around the terminus of replication (Fig. 3).

Figure 4 provides an overview of the amino acid sequence identity of homologous genes in the two genomes. About half of all genes in each genome share a sequence identity of at least 50% with at least one gene in the other taxon. Despite the fact that codon usage is different between *P. putida* and *P. aeruginosa*, most proteins...
are highly homologous. Thus the sequences preferentially diverged at synonymous sites, with the coding information being retained. The orthologous genes constitute the group of highest sequence homology (Fig. 4). The more than 2,000 orthologs with an amino acid sequence identity of 70% or more and the still detectable synteny of genes provide compelling evidence for the descent from a common ancestor and the lasting close phylogenetic relatedness of today’s *P. aeruginosa* and *P. putida*.

### 3.3. Compositional Di- and Tetranucleotide Bias

The compositional bias of the di- and tetranucleotide content is known to be a robust approach to the characterization of a bacterial genome [25]. The differential dinucleotide relative abundance values and the variance of tetranucleotide frequencies were calculated along the PAO1 and KT2440 chromosomes. The two latter parameters are complementary measures of oligonucleotide frequencies [17, 25, 26]. The “differential dinucleotide relative abundance value” (dinucleotide bias) indicates the difference between the odds ratios of the dinucleotide frequencies of two sequences (here between 4 kb windows and the complete genome sequence), whereas in the case of the “variance of tetranucleotide frequency” the variance from an equal frequency of all tetranucleotides is calculated for each of the 256 tetranucleotides in a window (see the Materials and Methods section for the exact algorithms). Sequence analyses of various microbial genomes have demonstrated that the “di- and tetranucleotide biases” of successive segments of a genome as calculated by these two algorithms are more similar to each other than to those of sequences from more distant organisms and hence visualize the genomic signature of a taxospecies [17, 25].

Figure 5 shows the frequency distribution curves of the two parameters in the *P. putida* KT2440 and *P. aeruginosa* PAO1 chromosomes. The distribution of dinucleotide relative abundance values is virtually identical in the two genomes, with 95% of data within the range between 0 and 0.22 (Fig. 5A). The two genomes share an identical dinucleotide bias. However, the spectrum of the variance of the tetranucleotide frequencies is different. The variance of tetranucleotide frequencies is higher in the PAO1 genome and shows an almost symmetrical Gaussian distribution (Fig. 5B). The KT2440 genome is characterized by a skewed distribution of genome sections with low (minor fraction) or high (major fraction) variance (Fig. 5B). The G + C content and the presence of gene islands with atypical tetranucleotide composition are the two major factors that shape the variance of tetranucleotide frequencies [21]. Hence, the lower average G + C content and the mosaic genome architecture with its large number of gene islands account for the bimodal Gaussian distributions and the comparatively lower variance in *P. putida*.

### 3.4. Under- and Overrepresented Oligonucleotides

The frequency and localization of individual oligonucleotides provide the most detailed information about the global features of a genome. We determined the frequency and map position of all 2-mers to 14-mers in the *P. putida* KT2440 and *P. aeruginosa* PAO1 chromosomes. Penta-, tetra-, tri-, and dinucleotides occur in from a thousand to up to several hundred thousand copies in a 6.2-Mb genome. The frequency of these short oligonucleotides is mainly determined by G + C content, chromosome structure and dynamics, discrimination between self and foreign DNA, and replication and repair, as well as by codon usage and transcriptional and translational demands; in other words, their frequency distribution constitutes the global genomic signature. Hexa-, hepta-, octa-, and nonanucleotides are found from a few to up to several thousand copies. These oligonucleotides may literally be considered the “common words” of the genome book. Deca-, undeca-, dodeca, trideca-, and tetradecamers constitute the category of “rare words”; just 2.3% of all possible 14-mers are expected to be found in a 6.2-Mb genome.

We focused our analysis on the extreme outliers of over- and underrepresented oligonucleotides defined by a cutoff value in $\chi^2$ tests of at least 10,000. This stringent criterion
Figure 6. Number of underrepresented (A) and overrepresented (B and C) di- to tetradecanucleotides in the *P. putida* KT2440 (orange) and *P. aeruginosa* PAO1 (green) genomes. Only those oligonucleotides were counted whose observed frequency differed from the expected value by a $\chi^2$ value of more than 10,000.

is satisfied for 717 (*P. putida*) and 444 (*P. aeruginosa*) dimers to tetradecamers. Figure 6 provides an overview of the distribution of these oligonucleotides by size. Significantly underrepresented oligomers were of course only detected in the class of short oligonucleotides, because the expected frequencies for individual n-mers decrease by $4^n$-fold with increasing length *n*.

Tables 1 and 2 list the under- and overrepresented di-, tri-, tetra-, and pentanucleotides in the two genomes. Twelve of the 22 and 24 underrepresented oligonucleotides in *P. putida* and *P. aeruginosa* are common to both genomes, of which TA, CTA, TAG, and CTAG are known to be universally rare in eu-bacteria [27, 28]. Interestingly, oligo G and oligo C stretches are strongly counterselcted in both GC-rich type I pseudomonads. In *P. aeruginosa*, $G_{2-5}$, $C_{2-5}$, $C_{3-4}G$, $G_{3-4}C$, $G_{3-4}C$, $C_{3-4}G$, $T_{C_{3-4}}$,

<table>
<thead>
<tr>
<th>Table 1. Underrepresented di-, tri-, tetra-, and pentanucleotides in the <em>P. aeruginosa</em> PAO1 and <em>P. putida</em> KT2440 genomes.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P. aeruginosa</strong></td>
</tr>
<tr>
<td>CC</td>
</tr>
<tr>
<td>CCCC</td>
</tr>
<tr>
<td>CCC</td>
</tr>
<tr>
<td>GGG</td>
</tr>
<tr>
<td>TA</td>
</tr>
<tr>
<td>CCCC</td>
</tr>
<tr>
<td>GGG</td>
</tr>
<tr>
<td>TA</td>
</tr>
<tr>
<td>GCCC</td>
</tr>
<tr>
<td>TAG</td>
</tr>
<tr>
<td>CTAG</td>
</tr>
<tr>
<td>TAA</td>
</tr>
<tr>
<td>TTA</td>
</tr>
<tr>
<td>GCCC</td>
</tr>
<tr>
<td>GGGC</td>
</tr>
<tr>
<td>TCCC</td>
</tr>
<tr>
<td>CCCG</td>
</tr>
<tr>
<td>CCCGG</td>
</tr>
<tr>
<td>CGGG</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Overrepresented di-, tri-, tetra-, and pentanucleotides in the <em>P. aeruginosa</em> PAO1 and <em>P. putida</em> KT2440 genomes.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P. aeruginosa</strong></td>
</tr>
<tr>
<td>TCGA</td>
</tr>
<tr>
<td>TCG</td>
</tr>
<tr>
<td>CGA</td>
</tr>
<tr>
<td>CAG</td>
</tr>
<tr>
<td>CCA</td>
</tr>
<tr>
<td>CAG</td>
</tr>
<tr>
<td>CAG</td>
</tr>
<tr>
<td>GAG</td>
</tr>
<tr>
<td>TAG</td>
</tr>
<tr>
<td>TAG</td>
</tr>
<tr>
<td>CGG</td>
</tr>
<tr>
<td>CT</td>
</tr>
<tr>
<td>AGC</td>
</tr>
<tr>
<td>TCA</td>
</tr>
<tr>
<td>GCC</td>
</tr>
<tr>
<td>GCT</td>
</tr>
<tr>
<td>CTC</td>
</tr>
<tr>
<td>CGC</td>
</tr>
<tr>
<td>CGT</td>
</tr>
<tr>
<td>CGT</td>
</tr>
</tbody>
</table>

*Oligonucleotides are sorted by decreasing $\chi^2$ values with a lower cutoff of 10,000.*
and G,A constitute the species-specific set of underrepresented short oligonucleotides. GC base pairs made up of homooligomers are thermodynamically stable and predispose to slippage during replication. Moreover, the subgroup of d(G,Cn) tracts tends to stack in a A-DNA like fashion, even within a B-DNA framework. An uncommon double helix of DNA is formed in which A-like base stacking is combined with B-type orientation of the deoxribose rings [29, 30]. The high GC-rich P. aeruginosa seems to avoid this scenario. G35 and C35 are also strongly counterselected in P. putida. In addition, the complementary tetranucleotide pairs GGAG/CTCC, CGGA/TCCG are underrepresented in P. putida. The counterselection of oligo G and oligo C is an exception rather than the rule. In general, most bacteria have a bias toward overrepresentation of A-DNA predisposing purine stretches (data not shown, but see http://www.cbs.dtu.dk/services/GenomeAtlas/A_DNA/index_Bacteria_PercentADNA.html) [31]. However, free-living soil bacteria like the type I pseudomonads that are capable of colonizing the rhizosphere tend to have less than half the expected frequency. The rhizosphere is an environment of high ionic strength. A-DNA conformations are favored with increasing salt concentrations [29, 30]. The counterselection of oligo G and oligo C of bacteria like the type I pseudomonads that inhabitate the rhizosphere could hence reflect the adaptation to a comparatively high salt environment.

Forty-three and 68 short oligonucleotides are overrepresented in the P. putida and P. aeruginosa chromosomes. Nineteen oligomers are common to both genomes, of which eight contain the trinucleotide CTG and six the trinucleotide CAG (reverse of CTG), which represent the most frequently used codons for leucine and glutamine, whereby, moreover, leucine is the most abundant amino acid [9] (see below). When all mononucleotides were counted in these frequent oligomers, interestingly equimolar amounts were found for G and C (35% each in P. putida and 32% each in P. aeruginosa) and for A and T (15% each and 18% each), respectively. Three palindromes are overrepresented: GATC (both), GGCC (P. putida), and GCGC (P. aeruginosa). Only two purine stretches (GAA, GAAG) and six alternating pyrimidine/purine stretches (GC, CA, TG, GCA, TGC, GCGC) are present among the 111 frequent di- to pentamers. This data demonstrate that oligomers with these above-mentioned symmetry motifs are counterselected among the most common short oligonucleotides.

The most frequent short oligonucleotides in P. putida can be represented as parts of three consensus sequences:

\[ 5'\text{-TGTCA-3'} \] (1)
\[ 5'\text{-TG GCC-3'} \] (2)
\[ 5'\text{-GC C-3'} \] (3)

All three consensus sequences can be translated in one frame into the most frequently used codons of amino acids that encode for peptides such as (L or I or V or M)-L-X (1) or (L or I or V or M or C or R or S or G)-L-A (2). The reader may note that sequence (3) is partially the reverse of (2) and that parts of the consensus sequences (2) and (3) match with the most frequent octamers in the genome (see the following).

In P. aeruginosa the combination of short oligonucleotides led to somewhat more complex consensus sequences:

\[ 5'\text{-T GATC A-3'} \] (4)
\[ 5'\text{-C G-3'} \] (5)
\[ 5'\text{-G(C or G)CGAAG} \] (6)
\[ 5'\text{-GC/}T \text{ or C-CTG/(CT or A or GC)} \] (7)
\[ 5'\text{-G/TC or GC or AG-CAG-} \] (CG or CA or GC or GT) (8A)
\[ 5'\text{-G(T or G)/TCG/GCGC) (8B) } \]

Consensus sequences (4), (6), and (7) are related to the corresponding sequences (1), (2) and (3) of P. putida. Sequence (5) can be translated into the peptides (A or G)-Q-X or X-(P or A)-R, and sequence (8) shows numerous combination with a serine-encoding codon. Hence, at least in one frame the overrepresented short oligonucleotides are frequently used codons [9], and, correspondingly, the short oligonucleotides more likely reflect the strong codon bias in P. aeruginosa [9] and some frequent peptide motifs more likely than structural constraints on the double helix. However, the most important observation is that the most common short oligonucleotides are overlapping because otherwise the oligomers could not have been combined to consensus sequences. In other words, the overrepresented short oligonucleotides are very often part of larger “words.”

3.5. Frequent “Words”

Figure 7 displays the map position of overrepresented 8-mers to 14-mers in P. putida KT2440 and P. aeruginosa PAO1. The intra- or intergenic localization of these frequent words is differentiated by blue and red colors. If numerous “words” map to the same position, they are part of a larger oligonucleotide, and if the same pattern is observed at numerous localizations, this larger oligonucleotide occurs in several copies. To start with two easy-to-follow examples: the large blue stretch close to the origin in the P. putida chromosome represents the multiple repeats of the 8862 amino acid large surface adhesion protein PP0168. The N-terminal repeat consists of nine units, each 100 amino acids in length, and the C-terminal repeat...
consists of 29 units, each 219 amino acids in length [21]. Moreover, the red-colored four **rrn** operons in *P. aeruginosa* [7] and seven **rrn** operons in *P. putida* (two of which occur in tandem) [6] can be clearly discerned by matching patterns at several map positions (Fig. 7).

Two words in *P. putida* and the very same two and another four words in *P. aeruginosa* are equally distributed in coding sequences throughout the chromosomes and are almost exclusively found in the same reading frame. These 8-mers are coding for short peptides. All of these words are overrepresented by more than an order of magnitude over what is expected.

The two words found both in *putida* and *aeruginosa* are the 8-mers 5'-TGCTGCTG and its reverse, 5'-CAGCAGCA. These 8-mers are almost exclusively found in a single reading frame encoding the tripeptides MLL, VLL, or LLL. Leucine is the most abundant amino acid in *P. aeruginosa* [9] and *P. putida* proteins. Codon usage is highly biased and CUG is the most frequently used codon (RSCU = 4.07 in *P. aeruginosa*) [9]. We suppose that leucinyl-leucinyl-leucine is a characteristic tripeptide signature of the genus *Pseudomonas*, because first, in both pseudomonads the LLL tripeptide was predominantly identified in the core genome outside of gene islands, and second, no overrepresentation of LLL was observed in the genomes of phylogenetically related enterobacteria such as *E. coli* and *Vibrio cholerae*. LLL is hydrophobic, and correspondingly its frequency was significantly 2-fold higher in transporters and membrane proteins than in other metabolic categories.

In *P. aeruginosa*, the next two common octamers in coding sequences read 5'-TCGGCCTG and its reverse, 5'-CAGGCCGA. As in the first case, a single reading frame is overrepresented that encodes the tripeptides IGL or LGL or FGL or VGL. Leucine, glycine, and valine are the first, third, and fifth most common amino acids in *P. aeruginosa* proteins [9]. The overrepresented octanucleotide contains common codons for isoleucine (RSCU 2.78), glycine (RSCU 3.02), leucine (RSCU 4.07), phenylalanine (RSCU 1.93), and valine (RSCU 1.69) [9]. The octanucleotide is overrepresented in genes that encode transporters and two-component systems and underrepresented in genes that are associated with horizontal gene transfer and genome mobility (phages, transposons).

The third octanucleotide pair reads 5'-CCTGCTGG and its reverse, 5'-CCAGCAGG. The encoded tripeptide PAG is overrepresented in transporters. However, in contrast to the other octanucleotides described above, in this case less biased codons are utilized to encode this tripeptide.

Further frequent “words” in the *P. aeruginosa* genome are the oligonucleotides 5'-CTTGGTATGGTLATCATCA and 5'-TAGACCTTGGTTCTTG, and 5'-GATCACATCTTCCAG found in multiple copies in adhesin genes; the oligonucleotides 5'-GTGAAATTTGAAGTGAATTGCA and 5'-GCCAAACAAGTATTCCA located adjacent to transposase genes; and the oligonucleotide 5'-GGACAGATTATTTCCCTA found in a rhs-like element [32].

Prominent examples in the *P. putida* genome for frequent words are simple repeats of the hexamer (AAGATC)ₙ (6 ≤ n ≤ 15) located in hypothetical genes adjacent to transposase genes and the *P. putida*-specific oligonucleotides 5'-TACCTGTGAGGGG, 5'-GGAGCAGGCCCTTTGTGCGGCA, and 5'-TGTAG(G) CCGGCCCTTTCCG, found in both coding sequences and intergenic regions. The two latter “words” are related to a 35-bp species-specific repetitive extragenic palindromic (REP) sequence [12].

### 3.6. REP Elements

REP elements were first described in *E. coli* as 35-bp sequences composed of a highly inverted repeat with the potential of forming a stem-loop structure [11]. The REP sequence consists of a central palindromic motif and characteristic residues that define the head and the tail. The
**P. aeruginosa** REPs (class 1 and 2) occur as single units or pairs. Two hundred twenty-five REP sequences are isolated and 372 REP elements occur in tandem on opposite strands. Clusters of three, four, or five REP sequences are found in 36, 12, and one case, respectively [12]. We searched for REP elements in the *P. aeruginosa* PAO1 genome by the strategy outlined in the Materials and Methods section. Two classes of 33-35-bp related REP sequences were identified:

- REP class1: cGGCGGATAaCCGC(N)_{1-3}
  \[\text{gCGGTTAT}_{1}CGCCCTaCg\]
- REP class2: cGGCGGATAaCGCc(N)_{1-3}
  \[\text{gCGG}_{1}CGCCCTaCg\]

(small letters: 50–89% conserved residue; capital letters: 90–100% conserved residue; palindromic region underlined). Each REP element consists of two imperfect palindromes separated by one to three nonconserved residues. Figure 9 displays the 35-bp consensus sequence of the REP element. The PAO1 genome contains 109 copies of the REP class1 sequence and 111 copies of the class2 sequence. The class1 and class2 REP elements typically occur in tandem on opposite strands at an interval of 45–57 bp or 65–85 bp apart from each other. Half of the REPs are also organized as BIMEs made up of three to 13 REPs (Fig. 10). The REPs are exclusively located in extragenic regions. Figure 11 shows an example of a complex architecture of REPs flanking the two putative transposase genes PA2319 and PA3434. The reader may note the regular spacing of class1-class2 pairs in the BIMEs.

The sequence logo in Figure 8 displays the consensus sequence of the KT2440 REP element with all common variations. The consensus sequence reads: 5’-ccggccctcTTCCGGGtaaGCCCGCttcaggg (small letters: 50–89% conserved residue; capital letters: 90–100% conserved residue; palindromic region underlined). In contrast to *E. coli*, where most REP elements are organized into complex “bacterial interspersed mosaic elements” (BIMEs) [33], most REP elements in *P. putida* occur as single units or pairs. Two hundred twenty-five REP sequences are isolated and 372 REP elements occur in tandem on opposite strands. Clusters of three, four, or five REP sequences are found in 36, 12, and one case, respectively [12].

We searched for REP elements in the *P. aeruginosa* PAO1 genome by the strategy outlined in the Materials and Methods section. Two classes of 33-35-bp related REP sequences were identified:

- REP class1: cGGCGGATAaCCGC(N)_{1-3}
  \[\text{gCGGTTAT}_{1}CGCCCTaCg\]
- REP class2: cGGCGGATAaCGCc(N)_{1-3}
  \[\text{gCGG}_{1}CGCCCTaCg\]

(small letters: 50–89% conserved residue; capital letters: 90–100% conserved residue; palindromic region underlined). Each REP element consists of two imperfect palindromes separated by one to three nonconserved residues. Figure 9 displays the 35-bp consensus sequence of the REP element. The PAO1 genome contains 109 copies of the REP class1 sequence and 111 copies of the class2 sequence. The class1 and class2 REP elements typically occur in tandem on opposite strands at an interval of 45–57 bp or 65–85 bp apart from each other. Half of the REPs are also organized as BIMEs made up of three to 13 REPs (Fig. 10). The REPs are exclusively located in extragenic regions. Figure 11 shows an example of a complex architecture of REPs flanking the two putative transposase genes PA2319 and PA3434. The reader may note the regular spacing of class1-class2 pairs in the BIMEs.

The sequence of this REP element was not found in any species other than *P. aeruginosa*. Moreover, the REP elements were identified exclusively in regions of the core genome defined by the typical genomic signature of differential dinucleotide relative abundance values and variance of tetrancleotide frequencies. The two features suggest that this type of REP element is a species-specific feature of *P. aeruginosa*. Interestingly, the *P. aeruginosa* REP (Fig. 9) is more closely related to the *E. coli* REP than to the *P. putida* REP (Fig. 8), suggesting that these two classes of REPs perform different biological functions. In both *P. putida* and *P. aeruginosa* the REP elements are species-specific features that could be exploited for the identification of the two taxospecies [34, 35]. Figure 12 depicts the distribution of the distance between adjacent

**Figure 8.** Sequence logo [19] of the consensus sequence of the 35-bp REP element of *P. putida* KT2440. The height of each letter is proportional to its frequency, and the letters are sorted so that the common one is on top. The height of the entire stack is adjusted to the information content of the sequence at that position given in bits.

**Figure 9.** Sequence logo [19] of the consensus sequence of the 35-bp REP element of *P. aeruginosa* PAO1.

**Figure 10.** Distribution of the organization of *P. aeruginosa* PAO1 (green) and *P. putida* KT2440 (orange) REPs in single units, pairs, and BIMEs.
REP sequences in \(P.\) \textit{putida} and \(P.\) \textit{aeruginosa}. The peaks below 100 bp visualize the BIMEs in the two genomes, whereas the other peak is compatible with a random distribution given by the ratio of genome size to the total number of REPs not organized in BIMEs.

### 3.7. Structural Parameters and Repeats

Although the cumulative presentation in Figure 12 is suggestive for an even distribution of REP sequences on both the \(P.\) \textit{putida} and \(P.\) \textit{aeruginosa} chromosomes, a higher-resolution view clearly demonstrates some systematic deviation from randomness. The genome atlases in Figures 13 and 14 visualize the association between structural parameters and the map localization of global repeats and REP elements. We observed with interest that in both genomes all regions with maximum intrinsic DNA curvature and minimal stacking energy are devoid of REP elements. REP sequences in enterobacteria are known to be recognized and cleaved by DNA gyrase, which could be a mechanism for controlling the level of negative supercoiling [36]. Hence, the action of gyrase on REP elements could influence the expression of adjacent genes. Regions of high curvature and low stacking energy might represent topological boundaries; hence this regulatory control would not be applicable to these regions [8]. Moreover, these regions cannot accommodate large palindromic motifs because stem-loop structures formed by the inverted repeats would destroy a highly curved DNA conformation. In summary, both structural and functional arguments make it plausible that REP sequences are absent from regions with extreme curvature and low stacking energy.

The three outermost circles in Figures 13 and 14 show the position preference, base-stacking energy, and intrinsic curvature in the two genomes. Whereas the range of position preference as a measure of DNA flexibility is comparable in \(P.\) \textit{aeruginosa} and \(P.\) \textit{putida}, the range of the intrinsic DNA curvature is broader in \textit{putida} than in \textit{aeruginosa}. Moreover, the values of the base pair stacking energy are lower in \textit{putida}. The two latter findings are compatible with the lower G + C content and the higher degree of mosaicism in the \textit{putida} genome.

Segments with low position preference, high flexibility, and comparatively low stacking energy are identified in both genomes adjacent to the \textit{rrn} operons and in the 4.1–5.2-Mb region in \(P.\) \textit{putida} and in the 0.75–1.5-Mb region in \(P.\) \textit{aeruginosa}. We have preliminary evidence from Genechip hybridization experiments that these regions contain highly expressed genes (unpublished data). Such a correlation between the Travers flexibility measure (“position preference”) and highly expressed genes has already been observed in the \textit{E. coli} chromosome [8, 37].

The inner circles of the genome atlases show the map localization of global direct repeats and global inverted repeats in the two pseudomonads. Examples are the multicopy \textit{rrn} operons and IS elements [6]. Other interesting cases in \textit{putida} are the nine copies of a group II intron-encoding maturation and the seven copies of a ISFpu10 transposase. All seven ISFpu10 transposase genes share 100% identical nucleotide sequence. ISFpu10 transposase is a member of the novel Piv/MooV family of transposases and site-specific recombinases, which mediate both site-specific DNA recombination and insertion sequence transposition in bacteria [38]. The gene must encode a functional product because it is essential for the attachment of \(P.\) \textit{putida} to seeds [39]. The highest homology is displayed with the pilin gene inverting proteins of \textit{Neisseria}, although none of the seven ISFpu10 transposase genes in \textit{putida} is physically linked with pilin biosynthesis genes. The nine full-length group II introns are another fascinating feature of the \textit{putida} genome. Group II introns are self-splicing RNAs [40]. Experiments in other taxa have shown that group II introns are capable of inserting at intron-specific locations in intronless alleles (homing) or at novel (ectopic) sites at
Figure 13. Genome Atlas of *P. aeruginosa* PAO1 showing different structural parameters and the distribution of REP elements and global repeats. Color intensity increases with the deviation from the average. Values close to the average are shaded very light gray; values with more than 3 standard deviations from the average are most strongly colored.

low frequency (transposition) [40]. Hence the nine group II introns could contribute to the mobility of the *P. putida* genome.

Global direct repeats make up 4.1% and global inverted repeats encompass 2.9% of the *P. aeruginosa* chromosome. The corresponding values for the *P. putida* chromosome are 5.1% and 3.1%. These values are close to the average found in sequenced bacterial genomes [41] (see also http://www.cbs.dtu.dk/services/GenomeAtlas/repeats/index_Bacteria_BlastDpercent.html).

The GC skew is depicted in the innermost circles of the genome atlases. GC skew values are similar in *putida* and *aeruginosa*. In many prokaryotic genomes, the leading strand is rich in G and the lagging strand is rich in C, and, correspondingly, the origin (ori) and the terminus of replication (ter) are indicated by the change in the GC skew along the chromosome [42]. In both the PAO1 and KT2440 genomes the ter locus is asymmetrically located with respect to the ori (Figs. 13 and 14).

4. CONCLUSIONS

In this study *P. putida* and *P. aeruginosa* were compared in numerous global genome parameters. The outcome is mixed: approximately the same number of concordant and discordant parameters were identified. Table 3 highlights the major findings. The large proportion of orthologs, the partially retained synteny of genes, the overrepresentation of the LLL tripeptide the and the presence of REP elements and BIMEs are striking observations that point to the close relationship of the two taxa. However, the correspondence between the two type I pseudomonads is low in codon usage and codon indices and the variance of tetranucleotide frequencies. The repertoire of the most over- and underrepresented words is distinct, albeit overlapping. In summary, the two genomes are still strongly related in the organization and content of the coding sequences, but they have considerably diverged in features that are strongly determined by G + C content and nucleotide substitution at synonymous sites.

Acknowledgments: This study was supported by grants provided by the Bundesministerium für Forschung und Technologie (project 0311758) and the Fifth Framework Programme of the European Union (contract no.QLK3-CT-2000-00170). D.U. was funded by a grant from the Danish Research Foundation.
**Figure 14.** Genome atlas of *P. putida* KT2440, showing different structural parameters and the distribution of REP elements and global repeats. Color intensity increases with the deviation from the average. Values close to the average are shaded very light gray; values with more than 3 standard deviations from the average are most strongly colored.

**Table 3.** Concordant and discordant global features of the *P. aeruginosa* PAO1 and *P. putida* KT2440 genomes.

<table>
<thead>
<tr>
<th>Concordant feature</th>
<th>Discordant feature</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genes</strong></td>
<td></td>
</tr>
<tr>
<td>Abundant number of orthologous genes</td>
<td>Genome codon index</td>
</tr>
<tr>
<td>Synteny of genes</td>
<td>Codon adaptation index</td>
</tr>
<tr>
<td></td>
<td>Fraction of gene islands, mosaicism</td>
</tr>
<tr>
<td><strong>Peptide motif and protein sequences</strong></td>
<td></td>
</tr>
<tr>
<td>Overrepresented tripeptide LLL</td>
<td></td>
</tr>
<tr>
<td>&gt;50% amino acid sequence identity in half of all genes</td>
<td></td>
</tr>
<tr>
<td><strong>Structural features</strong></td>
<td></td>
</tr>
<tr>
<td>Range of position preference</td>
<td>Range of intrinsic curvature</td>
</tr>
<tr>
<td>GC skew</td>
<td>Range of stacking energy</td>
</tr>
<tr>
<td><strong>Repeats</strong></td>
<td></td>
</tr>
<tr>
<td>Fraction of local and global repeats</td>
<td></td>
</tr>
<tr>
<td>Presence of REP elements and BIMEs</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Oligonucleotides</strong></td>
<td></td>
</tr>
<tr>
<td>Differential dinucleotide relative abundance</td>
<td></td>
</tr>
<tr>
<td>50% of underrepresented short oligomers</td>
<td></td>
</tr>
<tr>
<td>50% of overrepresented short oligomers</td>
<td></td>
</tr>
</tbody>
</table>

**Annotations:**
- REP +
- REP -

**Global Direct Repeats**

**Global Inverted Repeats**

**GC Skew**

**Resolution:** 2473
References and Notes


