Pseudomonas

Genomics and Molecular Biology

Edited by Pierre Cornelis

Caister Academic Press

RoxAnn R. Karkhoff-Schweizer
Department of Microbiology
Immunology and Pathology
Colorado State University
Fort Collins, CO
USA

roxannkarkoff-schweizer@colostate.edu

Jens Klockgether
Klinische Forschergruppe
Medizinische Hochschule Hannover
Hannover
Germany

klockgether.jens@mh-hannover.de

Ayush Kumar
Department of Microbiology
Immunology and Pathology
Colorado State University
Fort Collins, CO
USA

ayush.kumar@colostate.edu

Joseph S. Lam
Dept. of Molecular and Cellular Biology
University of Guelph
Guelph, ON
Canada

jlam@uoguelph.ca

José L. Martínez
Dept. Biotecnología Microbiana
Centro Nacional de Biotecnología (CSIC)
Campus UAM
Cantoblanco
Madrid
Spain

jlmtnez@cnb.uam.es

Sandra Matthijs
Laboratory of Microbial Interactions
Department of Molecular and Cellular
Interactions
Flanders Institute for Biotechnology
Vrije Universiteit Brussel
Brussels
Belgium

slmatthi@vub.ac.be

Wayne L. Miller
Department of Biochemistry
McGill University
Montreal, Quebec
Canada

waynelmiller@gmail.com

Takehiko Mima
Department of Microbiology
Immunology and Pathology
Colorado State University
Fort Collins, CO
USA

takehiko.mima@colostate.edu

Norberto J. Palleroni
Department of Biochemistry and
Microbiology
Rutgers University
Cook College
New Brunswick, NJ
USA

palleroni@aesop.rutgers.edu

Kristin F. Picardo
Department of Biology
St. John Fisher College
Rochester, NY
USA

kpicardo@sjfc.edu

Oleg Reva
Bioinformatics and Computational Biology
Unit
University of Pretoria
Pretoria
South Africa

oleg.reva@up.ac.za

Herbert P. Schweizer
Department of Microbiology
Immunology and Pathology
Colorado State University
Fort Collins, CO
USA

Herbert.Schweizer@colostate.edu

Lily A. Trunck
Department of Microbiology
Immunology and Pathology
Colorado State University
Fort Collins, CO
USA

lily.trunk@colostate.edu

Burkhard Tümmler
Klinische Forschergruppe
Medizinische Hochschule Hannover
Hannover
Germany

tuemmler.burkhard@mh-hannover.de

David W. Ussery
Center for Biological Sequence Analysis
Technical University of Denmark
Lyngby
Denmark

dave@cbs.dtu.dk

Victoria E. Wagner
Department of Microbiology and
Immunology
University of Rochester School of Medicir
and Dentistry
Rochester, NY
USA

victoria_wagner@urmc.rochester.edu

Lutz Wiehlmann
Klinische Forschergruppe
Medizinische Hochschule Hannover
Hannover
Germany

wiehlmann.lutz@mh-hannover.de

Dieco Würdemann
Klinische Forschergruppe
Medizinische Hochschule Hannover
Hannover
Germany

wuerdemann.dieco@mh-hannover.de

Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Garber, R.L., Goltry, L., Tolentino, E., Westbrock-Wadman, S.,Yuan, Y., Brody, L.L., Coulter, S.N., Folger, K.R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G.K., Wu, Z., Paulsen, I.T., Reizer, J., Saier, M.H., Hancock, R.E., Lory, S., and Olson, M.V. (2000). Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. Nature *406*, 959–964.

Tenover, F.C., Arbeit, R.D., Goering, R.V., Mickelsen, P.A., Murray, B.E., Persing, D.H., and Swaminathan, B. (1995). Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J. Clin. Microbiol. *33*, 2233–2239.

Tillier, E.R., and Collins, R.A. (2000). The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J. Mol. Evol. *50*, 249–257.

van der Meer, J.R., and Sentchilo, V. (2003). Genomic islands and the evolution of catabolic pathways in bacteria. Curr. Opin. Biotechnol. *14*, 248–254.

Visca, P., Imperi, F., and Lamont, I.L. (2007). Pyoverdine siderophores: from biogenesis to biosignificance. Trends Microbiol. *15*, 22–30.

Weinel, C., Ussery, D.W., Ohlsson, H., Sicheritz-Ponten, T., Kiewitz, C., and Tümmler B. (2003). Comparative genomics of *Pseudomonas aeruginosa* PAO1 and *Pseudomonas putida* KT2440: orthologs, codon usage, repetitive extragenic palindromic elements, and oligonucleotide motif signatures. Genome Lett. *1*, 175–187.

Wiehlmann, L., Wagner, G., Cramer, N., Siebert, B., Gudowius, P., Morales, G., Köhler, T., van Delden, C., Weinel, C., Slickers, P., and Tümmler, B. (2007). Population structure of *Pseudomonas aeruginosa*. Proc. Natl Acad. Sci. *104*, 8101–8106.

Willenbrock, H., Petersen, A., Sekse, C., Kiil, K., Wasteson, Y., and Ussery, D.W. (2006). Design of a 7 genomes *Escherichia coli* microarray for comparative genomic profiling. J. Bacteriol. *188*, 7713–7721.

Winstanley, C., Coulson, M.A., Wepner, B., Morgan, J.A., and Hart, C.A. (1996). Flagellin gene and protein variation amongst clinical isolates of *Pseudomonas aeruginosa*. Microbiology *142*, 2145–2151.

Wolfgang, M.C., Kulasekara, B.R., Liang, X., Boyd, D., Wu, K., Yang, Q., Miyada, C.G., and Lory, S. (2003). Conservation of genome content and virulence determinants among clinical and environmental isolates of *Pseudomonas aeruginosa*. Proc. Natl Acad. Sci. USA *100*, 8484–8489.

Yoon, S.S., Coakley, R., Lau, G.W., Lymar, S.V., Gaston, B., Karabulut, A.C., Hennigan, R.F., Hwang, S.H., Buettner, G., Schurr, M.J., Mortensen, J.E., Burns, J.L., Speert, D., Boucher, R.C., and Hassett, D.J. (2006). Anaerobic killing of mucoid *Pseudomonas aeruginosa* by acidified nitrite derivatives under cystic fibrosis airway conditions. J. Clin. Invest. *116*, 436–446.

3 Oligonucleotide Usage Signatures of the *Pseudomonas putida* KT2440 Genome

Oleg Reva and Burkhard Tümmler

Abstract

Di- to pentanucleotide usage and the list of the most abundant octa- to tetradecanucleotides are useful measures of the bacterial genomic signature. The *Pseudomonas putida* KT2440 chromosome is characterized by strand symmetry and intra-strand parity of complementary oligonucleotides. Each tetranucleotide occurs with similar frequency on the two strands. Tetranucleotide usage is biased by G+C content and physicochemical constraints such as base stacking energy, dinucleotide propeller twist angle or trinucleotide bendability. The 105 regions with atypical oligonucleotide composition can be differentiated by their patterns of oligonucleotide usage into categories of horizontally acquired gene islands, multidomain genes or ancient regions such as genes for ribosomal proteins and RNAs. A species-specific extragenic palindromic sequence is the most common repeat in the genome that can be exploited for the typing of *P. putida* strains. In the coding sequence of *P. putida* LLL is the most abundant tripeptide.

Pseudomonas putida KT2440

Pseudomonas putida strains are rapidly growing bacteria frequently isolated from most temperate soils and waters, particularly polluted soils (Timmis, 2002). They are nutritional opportunists *par excellence* and a paradigm of metabolically versatile microorganisms that recycle organic wastes in aerobic and microaerophilic compartments of the environment, and that play a key role in the maintenance of environmental quality. *P. putida* strain KT2440 (Nelson *et al.*, 2002) is probably the best characterized saprophytic laboratory Pseudomonad that has retained its ability to survive and function in the environment. The bacterium is a plasmid-free derivative of a toluene-degrading bacterium, originally designated *Pseudomonas arvilla* strain mt-2 and subsequently reclassified as *P. putida* mt-2 (Nakazawa, 2002). It is the first Gram-negative soil bacterium to be certified by the Recombinant DNA Advisory Committee (RAC) of the United States National Institutes of Health as the host strain of a host-vector biosafety (HV1) system for gene cloning in Gram-negative soil bacteria (Federal Register, 1982). An extensive spectrum of versatile genetic tools, in particular mini-transposons and tools based on these, have been

new catabolic pathways for pollutants (e.g. Erb *et al.*, 1997; Ramos *et al.*, 1986; Rojo *et al.*, 1987), the production by biocatalysis of intermediates, including chiral synthons for chemical syntheses (Williams *et al.*, 1976), and quality improvement of fossil fuels, for example by desulphurization (Galan *et al.*, 2000). KT2440 is also able to colonize the rhizosphere of a variety of crop plants, such as corn, wheat, strawberry, sugar cane and spinach (Espinosa-Urgel *et al.*, 2000), and is being used to develop new biopesticides and plant growth promoters that function in the plant rhizosphere.

The *P. putida* KT2440 genome has been sequenced by a German US consortium some years ago (Nelson *et al.*, 2002). Taking the 6.2 Mbp large chromosome of KT2440 as an instructive example, this review illustrates the usefulness of oligonucleotide usage (OU) analysis to characterize global features of bacterial genomes.

Introduction into OU analysis of bacterial genomes

The local G+C content and oligonucleotide frequencies are measures of variability in bacterial genomes (Karlin *et al.*, 1997; Abe *et al.*, 2003; Pride *et al.*, 2003; Teeling *et al.*, 2004). The order of nucleotides is governed not only by the encoded information, but also by physical and biological constraints. With the exception of long-range physical forces on DNA structure, all sections of the genome should be exposed to the same constraints and consequently should have the same fingerprints of oligonucleotide frequencies, i.e. frequencies being consistently either low or high for the same oligonucleotide (Weinel *et al.*, 2002). Figure 3.1 shows the frequency of the 256 tetranucleotides throughout the *P. putida* KT2440 genome. The frequency of each tetranucleotide is indeed approximately the same throughout the genome. Only a few regions exhibit an atypical oligonucleotide composition indicating that this DNA has been exposed to particular constraints other than those seen in the bulk of the genome. In most cases the regions with atypical OU are genomic islands that were integrated into the genome by horizontal transfer. 105 islands with atypical OU of a size of 4 kb or more were identified in the KT2440 chromosome (Table 3.1) (Weinel *et al.*, 2002).

Figure 3.1 visualizes the uniform frequency of each tetranucleotide throughout most regions of the KT2440 chromosome. DNA structural features have been related from theoretical calculations and empirical measurements to di- or trinucleotide scales such as base stacking energy (Ornstein *et al.*, 1978), propeller twist angle (Hassan and Calladine, 1996), protein deformability (Olson *et al.*, 1998), position preference (Pedersen *et al.*, 1998) and bendability (Brukner *et al.*, 1995). By permutation analysis one can assign the 256 tetranucleotides to 39 classes each of which characterized by the same values for the five scales mentioned above (Baldi and Baisnee, 2000; Baisnee *et al.*, 2001).

Figure 3.2 shows the deviations Δ_w of observed from expected counts of all 256 tetranucleotides (see equation 2 below) on one clockwise strand of the *P. putida* KT2440 chromosome in comparison to those of the archaeon *Aeropyrum pernix* K1 and the other two major genetic reference and certified bacterial safety strains with a G+C content that is different from that of the G+C rich *P. putida*, that is the A+T rich *Bacillus subtilis* and *Escherichia coli* K12 with a G+C content of about 50%. Tetranucleotides are sorted by

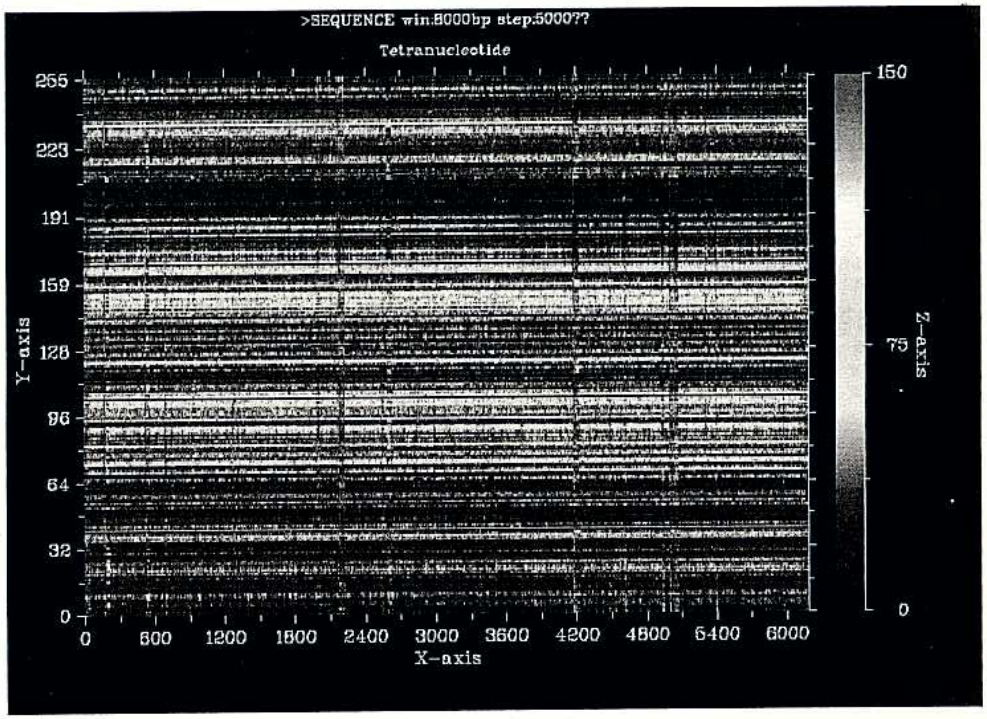


Figure 3.1 Tetranucleotide frequencies in the *P. putida* KT2440 genome. The abscissa indicates the coordinates of the genome. The 256 tetranucleotides are arranged in alphabetical order along the ordinate. The frequency of each tetranucleotide is given for 8 kbp sliding windows in steps of 5 kbp by the color code defined in the right bar. Genomic islands show up as vertical bars with atypical tetranucleotide frequencies. See also Plate 3.1.

a taxon-specific feature. For example, whereas the GC-rich *P. putida* KT2440 prefers the extreme with high stacking energy, the selection of tetranucleotides in the AT-rich *Bacillus subtilis* 168 is biased towards low stacking energy (Figure 3.2). 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. Importantly, a tetranucleotide and its reverse complement always belong to the same class. This tendency towards intrastrand parity of complementary tetranucleotides will consequently lead to a symmetric distribution of tetranucleotides on both strands (Baisnee *et al.*, 2002). In other words, these physicochemical features drive strand symmetry as it has been predicted by Chargaff more than 50 years ago in his second parity rule (Chargaff, 1951). This universal rule of strand symmetry for short oligonucleotides has meanwhile been empirically verified for more than 95% of sequenced bacterial chromosomes (Reva and Tümmler, 2004).

Oligonucleotide bias is a signature of a microbial genome and carries a phylogenetic

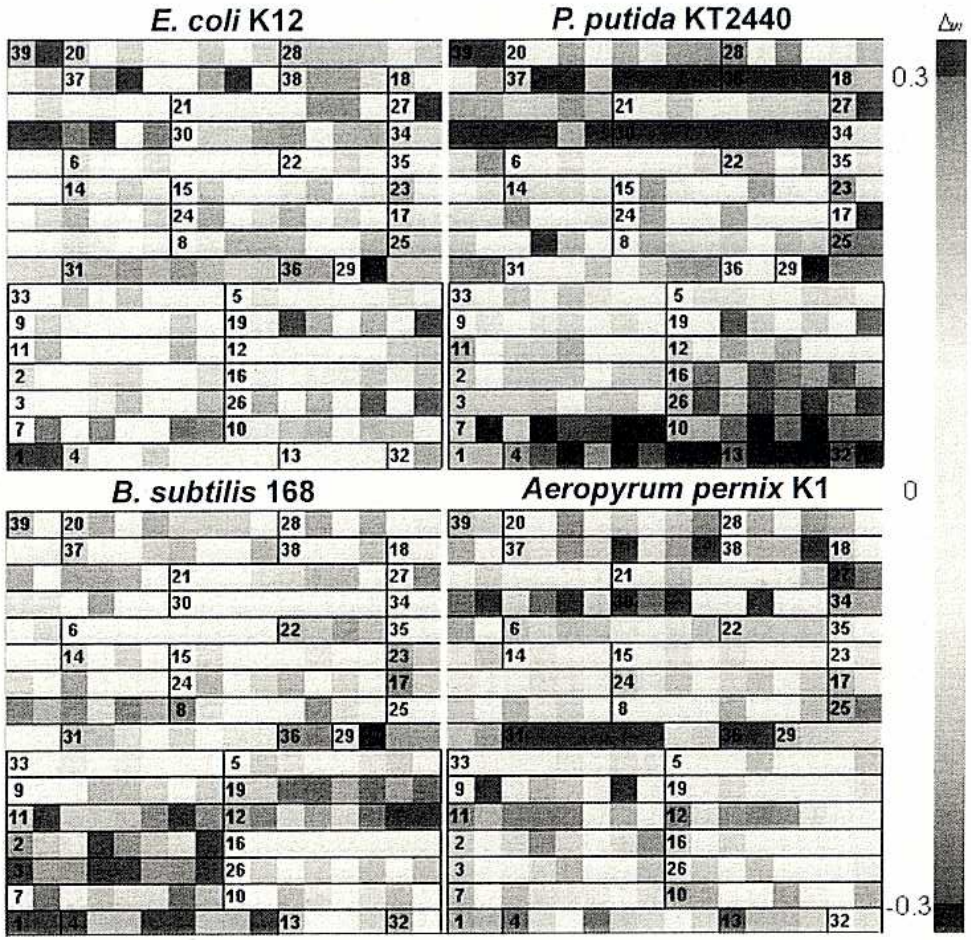


Figure 3.2 Tetranucleotide usage patterns of *E. coli* K12, *P. putida* KT2440, *B. subtilis* 168 and *A. pernix* K1. The deviation Δ_w of observed from expected counts defined by eqn 3.2 is shown for all 256 tetranucleotides (16 × 16 cells) by color code (right bar). Tetranucleotides are grouped into 39 classes of equivalent structural features (Baisnee *et al.*, 2001) and sorted by decreasing base stacking energy row-by-row starting at the upper left corner (class 39). Members within a class are sorted alphabetically. See also Plate 3.2.

Terms of OU statistics

Overlapping oligonucleotide words of a certain length l_w are counted in the sequence of L_{seq} nucleotides by shifting the window in steps of 1 nucleotide. The total word number (W_{total}) is $L_{seq} - l_w$ 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) are compared with the expected counts of words (C_e). Assuming the same distribution frequency for all words of a common length l_w irrespective of their composition and sequence, C_e matches the standard count number C_{n0}

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

$$C_e = C_{n1}$$

The deviation Δ_w of observed from expected counts is given by

$$\Delta_w = (C_0 - C_e) \times C_{n0}^{-1} \tag{3.2}$$

OU patterns are differentiated by the following abbreviation: 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.

The variance OUV of word deviations is calculated as follows:

$$OUV = \frac{\sum_w \Delta_w^2}{N_w - 1} \tag{3.3}$$

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

$$D(\%) = 100 \times \frac{\sum_w \left| rank_{w,i} - rank_{w,j} \right| - D_{\min}}{D_{\max} - D_{\min}} \tag{3.4}$$

whereby

$$D_{\max} = \frac{N_w (N_w - 1)}{2} \tag{3.5}$$

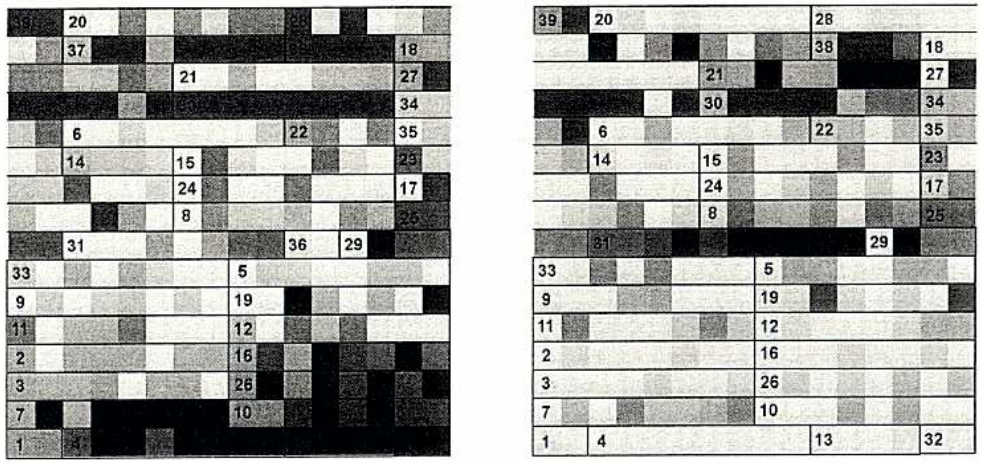
$D_{\max}$ is the maximal distance that is theoretically possible between two patterns of l_w long words (equation 3.5). $D_{\min}$ is the minimal distance between two patterns. The minimal distance is zero for two independent sequences, but has a positive value for the two complementary strands of the same DNA sequence that equals to 4^{lw} (if l_w is an odd number) or $4^{lw} - 2^{lw}$ (if l_w is an even number) because of the mutual dependence of the frequencies of words and their reverse complements in both strands. This distance be-

OU statistics of the *P. putida* KT2440 chromosome

G+C content is a major determinant of OU. However, if one normalizes the deviation Δ_{ij} of observed from expected counts by mononucleotide frequency, the preferential selection of oligonucleotide words becomes apparent that is not driven by the (trivial) constraints of mononucleotide frequency. The comparison of the right and left panel in Figure 3.3 demonstrates that the normalization by mononucleotide frequency unravels a strong signal of G+C independent bias of OU in the KT2440 genome. The chromosome prefers three related classes of tetranucleotides of similar physicochemical characteristics with a rather high stacking energy and strongly counter-selects three other adjacent classes with intermediate stacking energy.

Most bacterial chromosomes obey the rule of strand symmetry including the KT2440 strain (Reva and Tümmler, 2004). The pattern skew PS of n0_4mer of more than 95% of sequenced bacterial chromosomes are in the range of 1 to 8%. The extreme outlier is the *Xylella fastidiosa* 9a5c chromosome with a value of 24%.

The genomes of *P. putida* KT2440 and *X. fastidiosa* 9a5c were compared in their OU patterns consistency in order to test whether strand asymmetry reflects asymmetric genome topology and/or multiple inserts of foreign DNA. Both genomes are characterized by multiple genomic islands and unequal lengths of the clockwise and counterclockwise replichors (Figure 3.4). 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 behavior of the clockwise and counterclockwise replichors (Figure 3.4). The distances D (see eqn 3.4) of n0_4mer patterns were calculated between the clockwise and counterclockwise replichors and between the clockwise and the reverse complement of the counterclockwise replichors to be 44.09% and 9.03%, respectively. Compared to *X. fastidiosa* 9a5c, the OU pattern of the *P. putida* KT2440 genome is



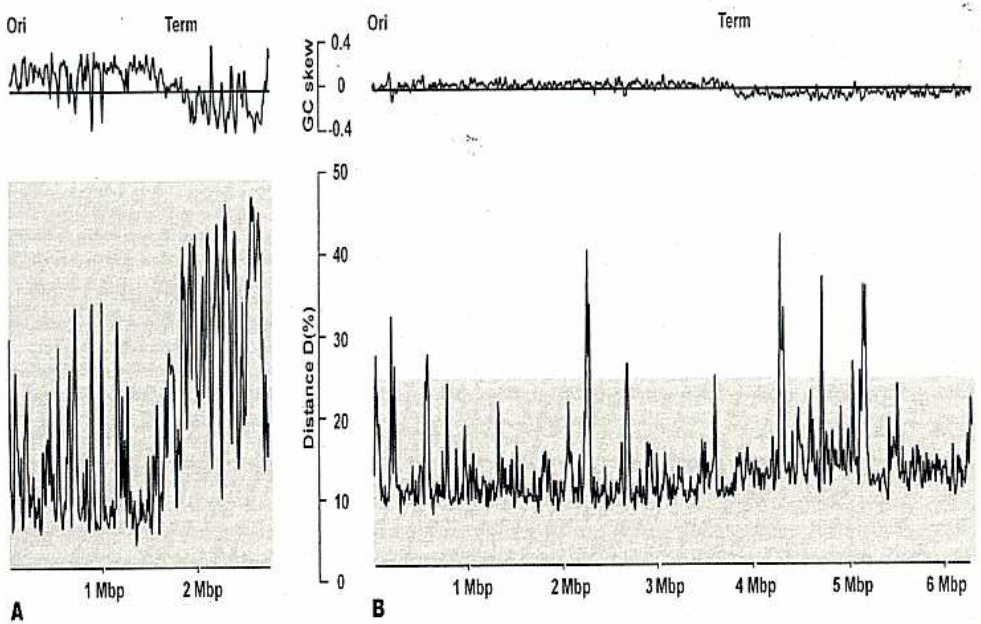


Figure 3.4 Local deviations of oligonucleotide usage patterns in sliding windows of two bacterial chromosomes. Lower panel: Distances (eqn 3.4) between n0_4mer patterns calculated for local regions of the leading strand and the standard patterns determined for the clockwise replichor of the two bacterial 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 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.

smoother and more uniform in spite of many genome islands with atypical OU patterns scattered throughout the chromosome (Figure 3.4). OU patterns 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 just 1.96%. Notably, despite the length difference, the shorter counterclockwise replichor manages to compensate the mirror OU skew in the clockwise strand due to its greater OU variance. Local OU differs largely from the bulk KT2440 chromosome in many regions, but these deviations sum up to approximately the same value on both replichors thus compensating each other to values close to zero. In other words, there are local regions with strand asymmetry in the KT2440 chromosome, but the whole chromosome exhibits close-to-perfect strand symmetry. The approximately equal share of PS on both replichors is not observed in the highly anomalous *X. fastidiosa* 9a5c chromosome so that this genome exhibits considerable strand asymmetry.

termed the "core sequences" are characterized by OU patterns being similar to the global pattern of the chromosome. However, many loci with alternative OU patterns contribute to in total more than 10% of the whole genome (Weinel *et al.*, 2002). These loci with atypical OU patterns comprise heterogeneous subsets of parasitic and recent foreign DNA, ancient genes for ribosomal constituents (RNAs and proteins), multidomain genes and non-coding sequences with multiple tandem repeats (Reva and Tümmler, 2005).

This diverse batch of atypical genomic loci can be differentiated into rather homogeneous subtypes by their profiles of D, OUV and PS. The three OU parameters often exhibit extreme values in the atypical regions, but the individual profiles are in most cases not congruent among themselves. An instructive example is shown in Figure 3.5. Two gene islands are located back-to-back at coordinates 160 kbp to 240 kbp of the *P. putida* KT2440 genome (islands 2 and 3 in Table 3.1). Island 2 comprises two tandem operons for ribosomal RNAs (*rrnA-rrnA'*), while island 3 contains the largest *P. putida* gene, PP0168, encoding a threonine-rich surface adhesion protein. Both islands have an atypical oligonucleotide composition, but they differ from each other in their OU signature. Figure 3.5 shows that OUV:n1_4mer has its genomic minimum (0.08) in island 2 but its genomic maximum (0.88) in island 3. Moreover, PS:n0_4mer is maximal (74.7%) in island 2 and is closer to the average level (47.5%) in island 3. This example illustrates that the combination of several OU parameters can be diagnostic for the categorization of genomic islands.

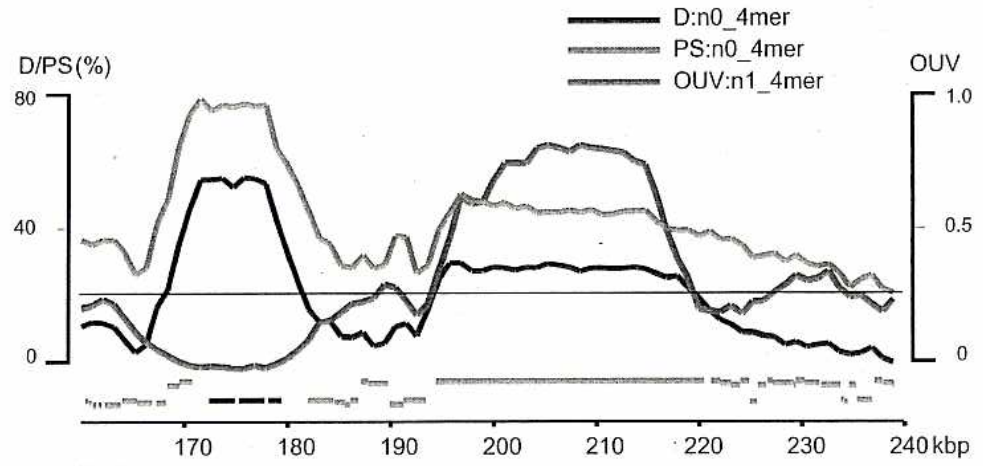


Figure 3.5 Curves of D:n0_4mer, PS:n0_4mer and OUV:n1_4mer in a locus of the *P. putida* KT2440 genome covering two regions with atypical OU: *rrnA-rrnA'* gene cluster and a long multidomain gene PP0168 encoding the surface adhesion protein (islands 2 and 3 in Table 3.1). Local OU patterns were analyzed in 5 kbp sliding windows in 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

Table 3.1 Gene islands in the *P. putida* KT2440 genome

Island	From	To	Size	orfs	Annotation
1	16000	58000	42000	36–37	Transposase, heavy metal efflux transporters, sugar transferases, porin P
2	171000	182000	11000		Ribosomal operons *rrnA, A'*
3	194000	218000	24000	1	Surface adhesion protein
4	238000	244000	6000	10	AlgP, FklB, AIQ
5	287000	291000	4000	2–3	Porin E, antioxidant protein LsfA
6	332000	336000	4000	5–6	Cluster of phage-related genes, large intergenic regions
7	360000	367000	7000	6–7	Gene cluster incl. metallopeptidase, dehydrogenase
8	396000	402000	6000	6	Transposase, acyltransferase, hypothetical proteins
9	404000	409000	5000	3	Sensory box protein, pyruvate dehydrogenase components
10	480000	484000	4000	3	Cluster of conserved hypothetical proteins
11	524000	530000	6000		Ribosomal operon *rrnB*
12	530000	566000	36000	44–46	Cluster of ribosomal proteins
13	620000	624000	4000	3–4	TonB-dependent outer membrane receptor
14	696000	703000	7000		Ribosomal operon *rrnC*
15	742000	757000	15000	10–13	GroupII intron, hypothetical proteins
16	804000	811000	7000	5	Cluster of conserved hypothetical genes
17	844000	851000	7000	6	Cluster of metabolic enzymes
18	897000	902000	5000	5–6	CysZ, thioredoxin, hypothetical proteins
19	922000	945000	23000	4	Surface adhesion protein + type I secretion system
20	1178000	1182000	4000	2	Sulfatase domain protein, Cu-dependent Mn^{2+} oxidase
21	1195000	1206000	11000	11	Type II secretion operon (XcpPQRSTUVWYZ, GspN)
22	1274000	1284000	10000	5	Invertase, recombinase, large intergenic regions
23	1296000	1300000	4000	1–2	Transposase
24	1325000	1330000	5000		Ribosomal operon *rrnD*
25	1397000	1401000	4000	4	Part of an transport operon: TolA, B, OprL

Table 3.1 continued

Island	From	To	Size	orfs	Annotation
27	1502000	1506000	4000	2–5	Ribosomal proteins
28	1748000	1777000	29000	11	Part of a 40 kb phage
29	1867000	1876000	9000	10	Cobalamin biosynthesis operon
30	1884000	1889000	5000	3–4	Sodium-solute transporter/sensor-kinase/ regulator fusion protein
31	1979000	2040000	61000	39	Cluster of LPS biosynthesis genes
32	2114000	2128000	14000	5	Cluster of hypothetical proteins (weak homology to *rhs* genes)
33	2147000	2151000	4000	1	Ribonuclease E
34	2162000	2224000	62000	45	Phage integrase, arsenate resistance operon, cluster of biosynthesis genes
35	2295000	2301000	6000	2–3	Cluster of exonuclease genes
36	2303000	2310000	7000	7	Cluster of conserved hypotheticals
37	2343000	2347000	4000	4–5	Regulator, hypothetical proteins, intergenic region
38	2381000	2385000	4000	3	Enzymes, transporter
39	2461000	2465000	4000	3	Lipoprotein releasing system
40	2535000	2539000	4000	3	Multicomponent potassium transporter, subunits ABCD
41	2547000	2554000	7000		Ribsomal operon *rrnE*
42	2586000	2626000	40000	37	Bacteriophage
43	2691000	2696000	5000	6–7	Chaperone usher pathway operon
44	2809000	2814000	5000	5	Cluster of ribosomal proteins, tRNA-synthases and IF-3
45	2817000	2822000	5000	3	Glutathion S-transferase, regulator, intergenic regions
46	2831000	2862000	31000	32	Cluster of oxidoreductases, DNA repair protein RadC, regulators and phage protein, phage-like integrase, cyclohydrolase
47	2911000	2922000	11000	2	Secreted hemolysin-type calcium-binding domain protein, antibiotic biosynthesis protein
48	2961000	2968000	7000	6	Putative siderophore secretion and uptake transporter operon
49	2983000	3011000	28000	20	VgrG homolog, *Salmonella* SciNOPBCIHS

Table 3.1 continued

Island	From	To	Size	orfs	Annotation
51	3057000	3062000	5000	4	Two component system
52	3088000	3092000	4000	4	Methionine synthase, reductase, conserved hypothetical proteins
53	3226000	3231000	5000	4	Dehydrogenase, regulator, transporter
54	3283000	3287000	4000	5–6	Transporter, regulators, cytochrome B561, catalase
55	3341000	3383000	42000	22	Several transposases and integrases, regulators, oxidoreductases, peroxidase, Mg^{2+}/Co^{2+} transporter
56	3401000	3415000	14000	10–19	Phage recombinase, LysE type translocators, aa efflux transporter, regulators, DNA-methylase, enzymes
57	3418000	3428000	10000	14	Start of pyocin R2 cluster
58	3447000	3452000	5000	4	End of pyocin R2 cluster
59	3493000	3525000	32000	4	Part of a gene cluster homolog to Salmonella SciHI_CBGNOPS genes, VrgS/ VrgG homolog, Rhs-related protein
60	3526000	3530000	4000	3	LexA, cell division inhibitor SulA, putative UV resistance factor, DNA polymerase III (SOS system)
61	3578000	3582000	4000	2	Large intergenic region, BenR, hypothetical protein (probably inserted in benzoate operon)
62	3645000	3649000	4000	3	Biosynthetic enzymes, outer membrane protein
63	3725000	3733000	8000	9	Cluster of stress response genes
64	3749000	3754000	5000	7	Cluster of heat shock proteins
65	3764000	3768000	4000	1	Ring-cleaving dioxygenase
66	3781000	3785000	4000	4	Nickel ABC transporter
67	3847000	3853000	6000	5	Curli/fimbrin nucleator-like protein
68	3907000	3912000	5000	4–5	Putative bacteriophage receptors
69	3935000	3941000	6000	5	Transposase, curli assembly protein-like proteins
70	4007000	4012000	5000	3–4	4-hydroxybenzoate hydroxylase, regulators
71	4068000	4072000	4000	4	RND efflux transporter
72	4073000	4077000	4000	2	Transposase, thiol peroxidase

Table 3.1 continued

Island	From	To	Size	orfs	Annotation
75	4152000	4156000	4000	4	Putative nitrobenzoate uptake and metabolizing operon
76	4174000	4232000	58000	37	Transposase, mismatch repair protein, DNA helicase, phosphatase, regulator, hypothetical proteins, intergenic regions
77	4259000	4264000	5000	4	ABC transporter
78	4298000	4324000	26000	38	Phage-related repressor, cluster of enzymes, operon for biosynthesis of a secondary metabolite, transposase
79	4363000	4428000	65000	85	Bacteriophage
80	4472000	4501000	29000	28	Integrase, two-component system, several enzymes
81	4534000	4540000	6000	7–9	Transposase, conserved hypothetical proteins
82	4566000	4573000	7000	3	Part of the amylase operon
83	4614000	4631000	17000	29	Cluster of Rhs core protein, hypothetical proteins
84	4737000	4742000	5000	6	Succinate dehydrogenase genes, citrate synthase
85	4799000	4803000	4000	5–6	Short hypothetical proteins
86	4812000	4818000	6000	3	Homolog of microcin b17 activating protein
87	4821000	4828000	7000	1	Pyoverdine synthetase
88	4862000	4867000	5000	3	Cell division proteins
89	4936000	4944000	8000	6	Putative amino acid metabolizing operon
90	4946000	4950000	4000	6	Part of flagellar operon (FliMNOPQR)
91	4965000	4971000	6000	6	Part of flagellar operon (genes between a conserved hypothetical and 3-oxoacyl-(acyl-carrier-protein) synthase III gene)
92	5000000	5076000	76000	60	Transposases, integrase, amino acid uptake and metabolizing gene cluster, opine uptake and metabolizing gene cluster, regulators, enzymes, hypothetical proteins
93	5139000	5143000	4000	3	Transporter, regulator
94	5149000	5153000	4000	2	Hypoth. genes, intergenic space
95	5308000	5313000	5000		*rrnF*

Table 3.1 continued

Island	From	To	Size	orfs	Annotation
97	5368000	5373000	5000	4–5	Carbamoylphosphate synthetase gene cluster, reductase
98	5391000	5402000	11000	3–4	Type-I restriction-modification operon
99	5552000	5556000	4000	3–4	*Salmonella* SciM homolog, hypothetical proteins
100	5601000	5608000	7000	4	Secreted serine protease + secretion system
101	5674000	5688000	14000	3	Part of pilin operon (PilL/ChpA, PilJ)
102	5851000	5855000	4000	3	Transporter, conserved hypotheticals
103	5971000	5976000	5000	1–2	*Salmonella* SciM homolog
104	6128000	6169000	41000	31	Transposase, copper resistance genes, heavy metal efflux transporter, transcription regulator, genes found on *Pseudomonas* plasmids
105	6175000	6179000	4000	3–4	Part of ATP synthase operon

The application of this procedure to the whole KT2440 genome is shown in Figure 3.6 (Reva and Tümmler, 2005). Dots corresponding to the genome fragments are plotted in accordance with their D:n0_4mer, OUV:n1_4mer and PS:n0_4mer values. The majority of fragments that represent the core genome clusters in one area. Three non-overlapping outlier groups were termed sections.

Section I is heterogeneous and includes long intergenic regions, clusters of short hypothetical genes, laterally transferred elements and genes for ribosomal RNAs. The local 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 Figure 3.6).

Genes for ribosomal proteins are localized in section II. 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. 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 (Kiewitz *et al.*, 2002).

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). Examples are the two large multidomain surface adhesion proteins PP0168 (26 046 bp) and PP0806 (18 930 bp) that contain numerous large repeats.

Section I is heterogeneous. The genes for ribosomal RNAs are discerned from the

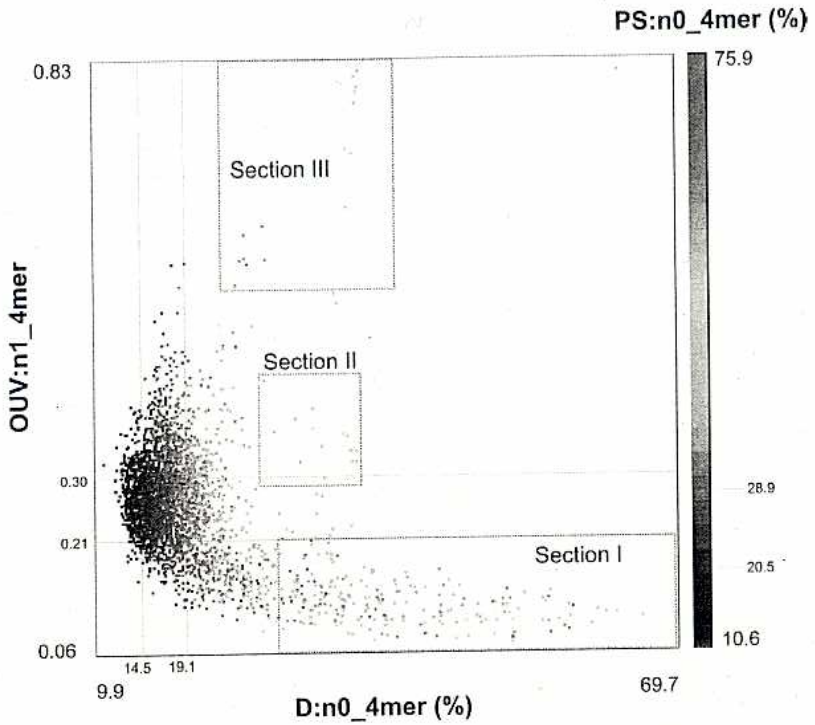


Figure 3.6 Dot-plot presentation of 8 kb genomic fragments of the *P. putida* KT2440 chromosome. Fragments of 8 kbp were generated with a sliding window in steps of 2 kbp. Each dot represents the D:n0_4mer, OUV:n1_4mer and PS:n0_4mer values of one fragment. The latter parameter is depicted by a color code represented by the bar in the right part of the figure. The grey lines indicate borders of the inner quartiles of values for the corresponding OU statistical parameters. See also Plate 3.6.

frequencies of oligonucleotides in the genomic fragment of interest (internal normalization, *i*) or normalization by frequencies of oligonucleotides in the complete sequence of the genome (global normalization, *g*). For example, internal and global OUV determined for a local n1_4mer pattern are designated as OUV:n1$_i$_4mer and OUV:n1$_g$_4mer, respectively. 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 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:n1$_i$_4mer and OUV:n1$_g$_4mer merge in core sequences but widely diverge in gene islands (Figure 3.7B).

An example for the identification of a laterally acquired gene island is shown in Figure 3.7. The island in the chromosome of *P. putida* KT2440 has significantly divergent OUV:n1$_i$_4mer and OUV:n1$_g$_4mer values and D:n0_4mer values beyond the 95% confidence

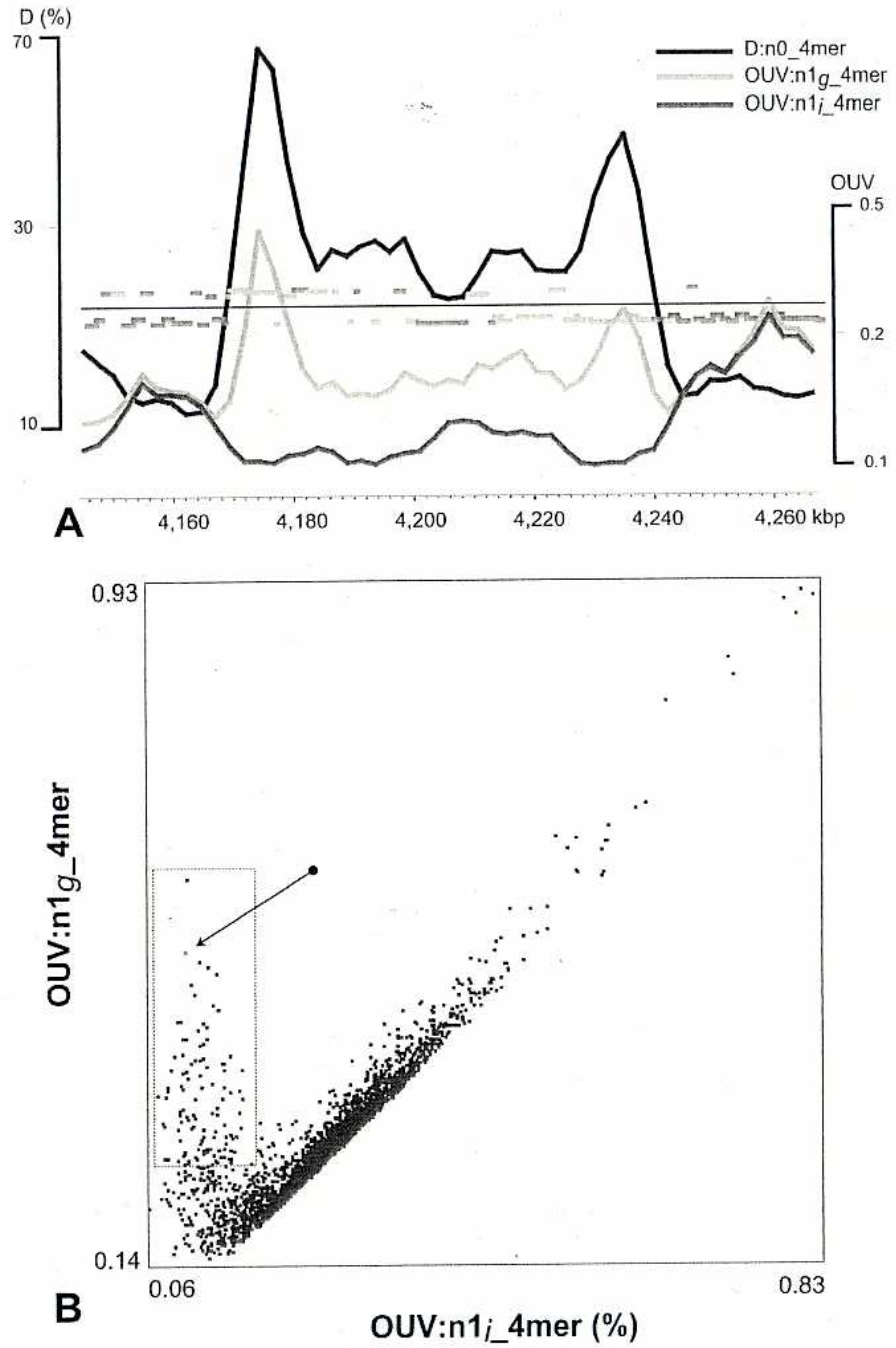


Figure 3.7 Gene islands in the *P. putida* KT2440 genome identified by discordant OUV:n1$_i$_4mer and OUV:n1$_g$_4mer values A) in a local gene map and B) globally in the complete genome. Genome fragments of 8 kbp were generated with a sliding window in step of 2 kbp. Red bars

In summary, the *P. putida* KT2440 genome is not homogeneous but contains polymorphic blocks including horizontally transferred gene islands, non-coding sequences, long multidomain genes and ancient conserved gene clusters. This structural polymorphism is visualized effectively by local OU pattern signatures.

Di-, tri- and tetranucleotide usage patterns are charged with most information content. The optimal word length will provide maximal information about the question of interest. Considering the minimal sequence length that gives reliable OU statistics, the threshold values of the minimum length of sequence were determined to be 0.3, 1.2, 5 and 20 kbp for di-, tri- tetra- and pentanucleotides, respectively (Reva and Tümmler, 2004). However, to be informative, the window should of course be not too long, because otherwise short range fluctuations of OU will vanish. Hence, 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.

Gene islands in the *P. putida* KT2440 genome detected by OU statistics

The 105 gene islands are rather homogeneously distributed along the chromosome, albeit the highest density was observed between 3.3 and 5.1 Mb. Table 3.1 provides information for each gene island about its genome coordinates, number of ORFs and annotation. Ten islands are related to phage, transposons, IS-elements and group II introns, the typical representatives of alien gene islands *sensu stricto*. Four islands are flanked at least on one side by a tRNA gene and contain an integrase gene like a phage-derived "pathogenicity island" (Hacker and Kaper, 2002). Two islands are embedded by complete tRNA genes. As in *P. aeruginosa* the genes for LPS biosynthesis reside in a large island. A further six islands are made up by the seven ribosomal operons. In 11 of the 13 GC-rich islands with the highest oligonucleotide bias, the database search identified genes in the phylogenetically related *P. aeruginosa* as the closest homologues. The encoded proteins are typical for *P. aeruginosa*, many of which are likely involved in virulence traits of this opportunistic pathogen: surface adhesion proteins, regulators of pilin synthesis, elements of the type I and type II secretion systems, the enzymes for cobalamin or pyoverdine biosynthesis and elements for the degradation or repair of DNA (Croft *et al.*, 2000; Stover *et al.*, 2000, and references therein).

An atypically low G+C content segregated with an extraordinarily low variance of tetranucleotide frequency in 32 islands. The majority of islands adds to host defense and protection or to the metabolic capacity of *P. putida*. The eight largest islands of 10 kb or more encode the singular restriction modification system of the KT2440 genome, homologues of Sci proteins and of the RhsG accessory genetic element VgrG (Croft *et al.*, 2000), pyocins and/or are related to mobile genetic elements.

Table 3.2 summarizes the content of the 105 islands in terms of the functional categories defined for the annotation of the *P. aeruginosa* PAO1 genome (Stover *et al.*, 2000). The

Table 3.2 Functional categories encoded by the 105 gene islands of *P. putida* KT2440 identified by OU statistics

Functional category	Number of islands
Adaptation, protection	7
Biosynthesis of cofactors, prosthetic groups	11
Carbon compound catabolism	13
Cell wall/LPS	7
DNA modification and repair	8
Protein secretion/export apparatus	6
Putative enzymes	20
Regulatory systems	18
Related to phage, transposon, or plasmid	29
Secreted surface proteins	4
Translation	11
Toxin production and resistance	10
Transport of small molecules	24
Hypothetical, unclassified, unknown	11
Intergenic regions	6

chemicals, ion transport and the synthesis and secretion of secondary metabolites. Other islands endow *P. putida* with determinants of resistance and defense or with constituents and appendages of the cell wall. Twenty-nine islands carry the signature of mobile elements indicating the recent acquisition by horizontal gene transfer.

Housekeeping genes were not detected in any island in accordance with expectation that central metabolic pathways should be encoded by the core genome. Besides the citrate synthase and succinate dehydrogenase genes in island 84 the only exceptions were genes of the translational apparatus, most of them located in section II (Figure 3.6). These evolutionarily highly conserved ribosomal genes with their lower G+C content have experienced a less stringent bias towards GC-rich codons and hence prefer codons other than the typical *P. putida* genes. The differential codon usage of ribosomal proteins has the consequence that they do not compete with typical *P. putida* proteins for the same tRNA species during translation.

Over- and underrepresented oligonucleotide words in the *P. putida* KT2440 genome

Di- to pentanucleotides are optimal to scan the genome for structural polymorphisms. Alternatively, one may search for repeats and the most abundant oligonucleotides (octa- to tetradecanucleotides). These oligonucleotides constitute a library of the most common

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 6.1 Mb genome.

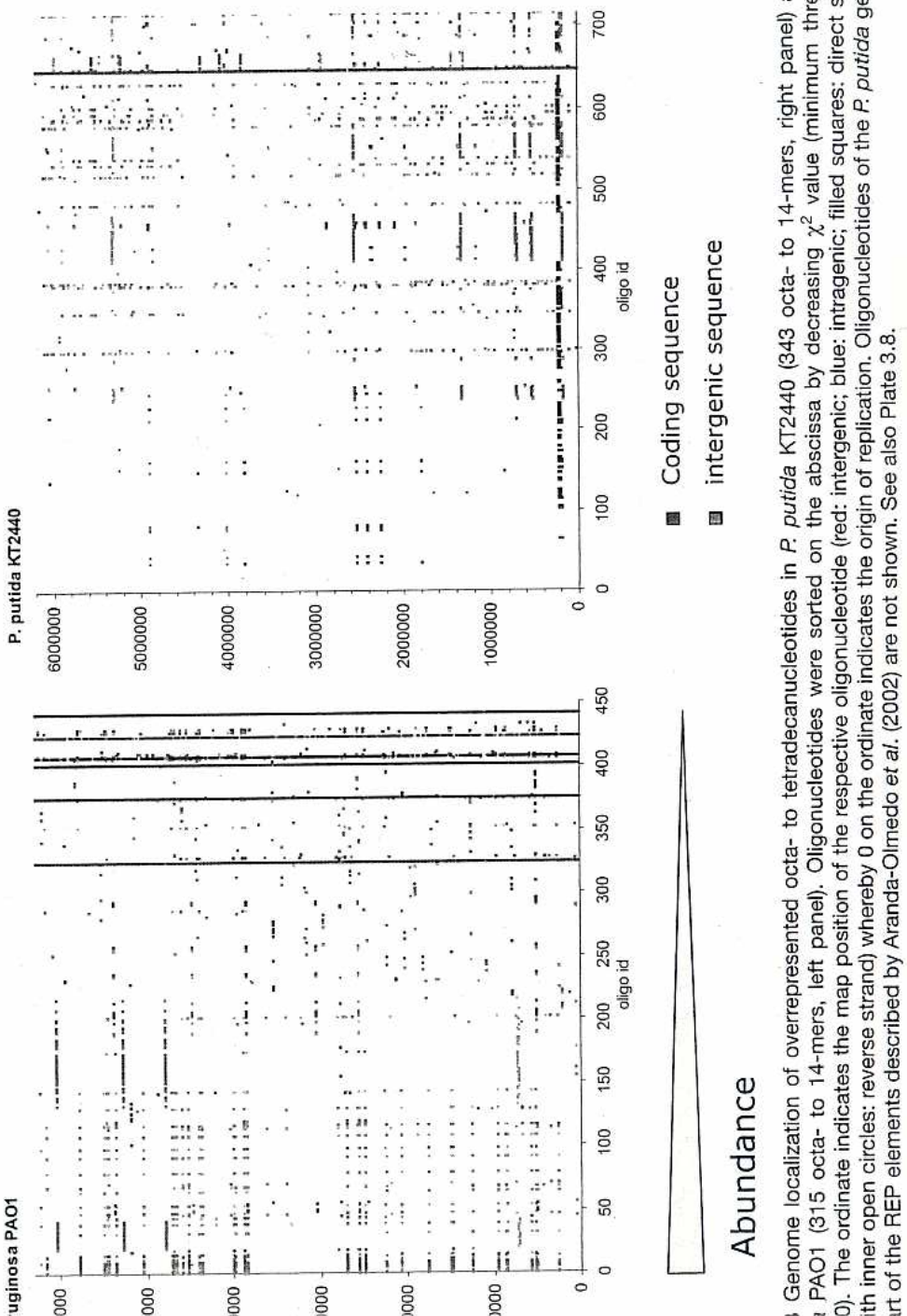
Figure 3.8 displays the map position of overrepresented 8-mers to 14-mers in *P. putida* KT2440 (Nelson *et al.*, 2002) and *P. aeruginosa* PAO1 (Stover *et al.*, 2000). These extreme outliers of overrepresented oligonucleotides were defined by a cut-off value in χ^2 tests of at least 10 000 (Weinel *et al.*, 2003). The intra- or intergenic localization of these frequent words is differentiated by blue and red color. 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 8,862 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. Moreover, the red-colored four *rrn* operons in *P. aeruginosa* and seven *rrn* operons in *P. putida* (two of which occur in tandem) can be clearly discerned by matching patterns at several map positions (Figure 3.8).

Two overrepresented words in *P. putida* and in *P. aeruginosa* are evenly 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. The two words found both in *P. putida* and *P. aeruginosa* are the 8mers 5′-TGCTGCTG and its reverse 5′-CAGCAGCA. These 8-mers are almost exclusively found in one single reading frame encoding the tripeptides MLL, VLL or LLL. Leucine is the most abundant amino acid in *P. aeruginosa* and *P. putida* proteins. Codon usage is highly biased and CUG is the most frequently used codon (RSCU = 4.07 in *P. aeruginosa* (Kiewitz *et al.*, 2002)). 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* (Weinel *et al.*, 2003). LLL is hydrophobic, and correspondingly its frequency was significantly 2-fold higher in transporters and membrane proteins than in other metabolic categories.

Other prominent examples in the *P. putida* genome for frequent words are simple repeats of the hexamer (AAGATC)n ($6 \leq n \leq 15$) located in hypothetical genes adjacent to transposase genes and the *P. putida*-specific oligonucleotides 5′-TACCTGTGGGAGCG, 5′-GGAGCGGCCTTGTGTCGCGAT and 5′-TGT(AG)CCGGCCTCTTCGC found both in coding sequences and intergenic regions. The two latter "words" are related to a 35 bp species-specific repetitive extragenic palindromic (REP) sequence (Aranda-Olmedo *et al.*, 2002; Weinel *et al.*, 2003; Tobes and Ramos, 2005).

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 (Stern *et al.*, 1984).



Genome localization of overrepresented octa- to tetradecanucleotides in *P. putida* KT2440 (343 octa- to 14-mers, right panel) and *P.* PAO1 (315 octa- to 14-mers, left panel). Oligonucleotides were sorted on the abscissa by decreasing χ^2 value (minimum threshold 0). The ordinate indicates the map position of the respective oligonucleotide (red: intergenic; blue: intragenic; filled squares: direct strand; th inner open circles: reverse strand) whereby 0 on the ordinate indicates the origin of replication. Oligonucleotides of the *P. putida* genome rt of the REP elements described by Aranda-Olmedo *et al.* (2002) are not shown. See also Plate 3.8.

complexity of the presentation, Figure 3.8 does not include any oligonucleotide that is part of the REP element in *P. putida*.

The consensus sequence of the KT2440 REP element reads:

5′-ccggcctcTTCGCGGGtaaaCCCGCtcctacaggg

(small letters: 50–89% conserved residue; capital letter: 90–100% conserved residue; palindromic region underlined).

In contrast to *E. coli* where most REP elements are organized in complex "bacterial interspersed mosaic elements," called BIMEs (Bachellier *et al.*, 1994), most REP elements in *P. putida* occur as single units or pairs. 225 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.

In the *P. aeruginosa* PAO1 genome two classes of 33–35 bp related REP sequences were identified (Weinel *et al.*, 2003):

REP class1: cGGCGGATAaCCGC$(N)_{1-3}$gCGGTTATtCGCCCTaCg

REP class2: cGGCGGATAaCGCc$(N)_{1-3}$gGCGttATtCGCCCTaCg

(small letters: 50–89% conserved residue; capital letter: 90–100% conserved residue; palindromic region is underlined).

Each REP element consists of two imperfect palindromes separated by one to three non-conserved residues. 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 thirteen REPs. Interestingly, the *P. aeruginosa* REP is more closely related to the *E. coli* REP than to the *P. putida* REP. In summary, the REP elements of *P. putida* and *P. aeruginosa* are species-specific features which could be exploited for the typing of *P. putida* and *P. aeruginosa* strains.

OU statistics: a universal valuable tool for the analysis of bacterial genomes

Bacterial genomes are not homogeneous but contain polymorphic blocks including horizontally transferred gene islands, non-coding sequences, long multi-domain genes and ancient conserved gene clusters. As shown here for the *P. putida* KT2440 chromosome, the structural polymorphism of bacterial genomes may be effectively analyzed by local OU pattern signatures in terms of OUV, D and PS. These parameters are useful for the visualization of regions with atypical oligonucleotide composition. The combination of the informative parameters that are 21 in case of tetranucleotide usage analysis, facilitates the

Acknowledgments

OR was a member of the European Research Training Group "Pseudomonas: Pathogenicity and Biotechnology" sponsored by the Deutsche Forschungsgemeinschaft when the concepts of OU statistics were developed and applied to the KT2440 genome.

References

Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T., and Ikemura, T. (2003). Informatics for unveiling hidden genome signatures. Genome Res. *13*, 693–702.

Almagor, H. (1983). A Markov analysis of DNA sequences. J. Theor. Biol. *104*, 633–645.

Aranda-Olmedo, I., Tobes, R., Manzanera, M., Ramos, J.L., and Marques, S. (2002). Species-specific repetitive extragenic palindromic (REP) sequences in *Pseudomonas putida*. Nucleic Acids Res. *30*, 1826–1833.

Bachellier, S., Saurin, W., Perrin, D., Hofnung, M., and Gilson, E. (1994). Structural and functional diversity among bacterial interspersed mosaic elements (BIMEs). Mol. Microbiol. *12*, 61–70.

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

Baisnee, P.F., Baldi, P., Brunak, S., and Pedersen, A.G. (2001). Flexibility of the genetic code with respect to DNA structure. Bioinformatics *17*, 237–248.

Baisnee, P.F., Hampson, S., and Baldi, P. (2002). Why are complementary DNA strands symmetric? Bioinformatics *18*, 1021–1033.

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

Chargaff, E. (1951). Structure and function of nucleic acids as cell constituents. Fed. Proc. *10*, 344–360.

Croft, L., Beatson, S.A., Whitchurch, C.B., Huang, B., Blakeley, R.L., and Mattick, J.S. (2000). An interactive web-based *Pseudomonas aeruginosa* genome database: discovery of new genes, pathways and structures. Microbiology *146*, 2351–2364.

de Lorenzo, V., Herrero, M., Jakubzik, U., and Timmis K.N. (1990). Mini-Tn5 transposon derivatives for insertion mutagenesis, promoter probing, and chromosomal insertion of cloned DNA in gram-negative eubacteria. J. Bacteriol. *172*, 6568–6572.

Erb, R.W., Eichner, C.A., Wagner-Döbler, I, and Timmis, K.N. (1997). Bioprotection of microbial communities from toxic phenol mixtures by a genetically designed pseudomonad. Nat. Biotechnol. *15*, 378–382.

Espinosa-Urgel, M., Salido, A., and Ramos, J.L. (2000). Genetic analysis of functions involved in adhesion of *Pseudomonas putida* to seeds. J. Bacteriol. *182*, 2363–2369.

Federal Register. (1982). Appendix E, Certified host—vector systems. *47*, 17197.

Galan, B., Diaz, E., and Garcia, J.L. (2000). Enhancing desulphurization by engineering a flavin reductase-encoding gene cassette in recombinant biocatalysts. Environ. Microbiol. *2*, 687–694.

Hacker, J., and Kaper, J.B. (2002). Pathogenicity islands and the evolution of pathogenic microbes. Curr. Top. Microbiol. Immunol. *264/I*, 1–211.

Hassan, M.A.E., and Calladine, C.R. (1996). Propeller twist of base-pairs and the conformational mobility of dinucleotide steps in DNA. J. Mol. Biol. *259*, 95–103.

Herrero, M., de Lorenzo, V., and Timmis, K.N. (1990). Transposon vectors containing non-antibiotic resistance selection markers for cloning and stable chromosomal insertion of foreign genes in gram-negative bacteria. J. Bacteriol. *172*, 6557–6567.

Karlin, S., Mrazek, J., and Campbell, A. (1997). Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. *179*, 3899–3913.

Kiewitz, C., Weinel, C., and Tümmler, B. (2002). Genome codon index of *Pseudomonas aeruginosa*: A codon index that utilizes whole genome sequence data. Genome Lett. *1*, 61–70.

Nakazawa, T. (2002). Travels of a *Pseudomonas*, from Japan around the world. Environ. Microbiol. *4*, 782–786.

Nelson, K.E., Weinel, C., Paulsen, I.T., Dodson, R.J., Hilbert, H., Martins dos Santos, V.A.P., Fouts,

Düsterhöft, A., Tümmler, B., and Fraser, C.M. (2002). Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. Environ. Microbiol. *4*, 799–808.
Noble, P.A., Citek, R.W., and Ogunseitan, O.A. (1998). Tetranucleotide frequencies in microbial genomes. Electrophoresis. *19*, 528–535.
Olson, W.K., Gorin, A.A., Lu, X., Hock, L.M., and Zhurkin, V.B (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc. Natl Acad. Sci. USA *95*, 11163–11168.
Ornstein, R.L., Rein, R., Breen, D.L., and MacElroy, R.D. (1978). An optimized potential function for the calculation of nucleic acid interaction energies. I. Base stacking. Biopolymers. *17*, 2341–2360.
Pedersen, A.G., Baldi, P., Brunak, S., and Chauvin, Y. (1998). DNA structure in human RNA polymerase II promoters. J. Mol. Biol. *281*, 663–673.
Pride, D.T., Meinersmann, R.J., Wassenaar, T.M., and Blaser, M.J. (2003). Evolutionary implications of microbial genome tetanucleotide frequency biases. Genome Res. *13*, 145–155.
Ramos, J.L., Stolz, A., Reineke, W, and Timmis K.N. (1986). Altered effector specificities in regulators of gene expression: TOL plasmid *xylS* mutants and their use to engineer expansion of the range of aromatics degraded by bacteria. Proc. Natl Acad. Sci. USA *83*, 8467–8471.
Reva, O.N., and Tümmler, B. (2004). Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. BMC Bioinformatics *5*, 90.
Reva, O.N., and Tümmler, B. (2005). Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. BMC Bioinformatics *6*, 251.
Rojo, F., Pieper, D.H., Engesser, K.H., Knackmuss, H.J., and Timmis, K.N. (1987). Assemblage of ortho cleavage route for simultaneous degradation of chloro- and methylaromatics. Science *238*, 1395–1398.
Stern, M.J., Ames, G.F., Smith, N.H., Robinson, E.C., and Higgins, C.F. (1984). Repetitive extragenic palindromic sequences: a major component of the bacterial genome. Cell. *37*, 1015–1026.
Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Garber, R.L., Goltry, L., Tolentino, E., Westbrock-Wadman, S.,Yuan, Y., Brody, L.L., Coulter, S.N., Folger, K.R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G.K., Wu, Z., Paulsen, I.T., Reizer, J., Saier, M.H., Hancock, R.E., Lory, S., and Olson, M.V. (2000). Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature *406*, 959–964.
Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glöckner, F.O. (2004). TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics *5*, 163.
Timmis, K.N. (2002). *Pseudomonas putida*: a cosmopolitan opportunist *par excellence*. Environ. Microbiol. *4*, 779–781.
Tobes, R., and Ramos, J.L. (2005). REP code: defining bacterial identity in extragenic space. Environ. Microbiol. *7*, 225–228.
Weinel, C., Nelson, K.E., and Tümmler, B. (2002). Global features of the *Pseudomonas putida* KT2440 genome sequence. Environ. Microbiol. *4*, 809–818.
Weinel, C., Ussery, D.W., Ohlsson, H., Sicheritz-Ponten, T., Kiewitz, C., and Tümmler B. (2003). Comparative genomics of *Pseudomonas aeruginosa* PAO1 and *Pseudomonas putida* KT2440: orthologs, codon usage, repetitive extragenic palindromic elements, and oligonucleotide motif signatures. Genome Lett. *1*, 175–187.
Williams, P.A., and Worsey, MJ. (1976). Ubiquity of plasmids in coding for toluene and xylene metabolism in soil bacteria: evidence for the existence of new TOL plasmids. J. Bacteriol. *125*, 818–828.

Genetic Tools for *Pseudomonas*

4

Kyoung-Hee Choi, Lily A. Trunck, Ayush Kumar, Takehiko Mima, RoxAnn R. Karkhoff-Schweizer, and Herbert P. Schweizer

Abstract

Genetic tools are required to take full advantage of the wealth of information generated by genome sequencing efforts, and ensuing global gene and protein expression analyses. Although the development of genetic tools has generally not kept up with the sequencing pace, substantial progress has been made in this arena. PCR- and recombination-based strategies allowed construction of whole genome expression and transposon insertion libraries. Similar strategies combined with improved transformation protocols facilitate high-throughput construction of deletion alleles and development of a broad-host-range mini-Tn*7* chromosome integration system. While to date most of these tools and methods have been developed for and applied in *P. aeruginosa*, they will most likely also be applicable to other *Pseudomonas* with appropriate modifications.

Introduction

Complete genome sequences for several *Pseudomonas* species, including *P. aeruginosa* strains PAO1 (Stover *et al.*, 2000), *P. putida* KT2440 (Nelson *et al.*, 2002), *P. fluorescens* Pf-5 (Paulsen *et al.*, 2005), *P. syringae* pathovar *tomato* DC3000 (Buell *et al.*, 2003), *P. syringae* pathovar *syringae* B728a (Feil *et al.*, 2005), *P. syringae* pathovar *phaseolica* 1448A (Joardar *et al.*, 2005), *P. fluorescens* PfO-1 and *P. entomophila* L48 (www.ncbi.nlm.nih.gov/genomes/proks.cgi) have been determined. Elucidation of these genome sequences produced an immense amount of information about the organization, coding capacity and evolution of these bacteria. While technologies such as DNA microarrays (Ochsner *et al.*, 2002; Schuster *et al.*, 2003; Wagner *et al.*, 2003; Wagner *et al.*, 2004; Salunkhe *et al.*, 2005; Dominguez-Cuevas *et al.*, 2006); and others) and proteomics (Sauer *et al.*, 2002;Nouwens *et al.*, 2003); and others) have been developed and successfully applied to study global gene expression in *Pseudomonas*, there is also an ever increasing demand for development of new and improved genetic tools to take full advantage of the wealth of information thus generated and to elucidate the function of the thousands of unknown genes identified through sequencing. Several recent reviews summarized in detail the arsenal of molecular genetic tools that has steadily grown and improved over the last decade and a half, but also pointed