International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:10, 2009

SeqWord Gene Island Sniffer: a Program to Study the Lateral Genetic Exchange among Bacteria

Bezuidt O., Lima-Mendez G. and Reva O. N.

***Abstract*—**SeqWord Gene Island Sniffer, a new program for the identification of mobile genetic elements in sequences of bacterial chromosomes is presented. This program is based on the analysis of oligonucleotide usage variations in DNA sequences. 3,518 mobile genetic elements were identified in 637 bacterial genomes and further analyzed by sequence similarity and the functionality of encoded proteins. The results of this study are stored in an open database (http://anjie.bi.up.ac.za/geidb/geidb-home.php). The developed computer program and the database provide the information valuable for further investigation of the distribution of mobile genetic elements and virulence factors among bacteria. The program is available for download at www.bi.up.ac.za/SeqWord/sniffer/index.html.

***Keywords*—**mobile genetic elements, virulence, bacterial genomes.

I. Introduction

IDENTIFICATION and distribution of horizontally transferred mobile genetic elements (MGE) in bacterial communities and also tracing their evolutionary origins have always been the greatest challenge in computational genomics. MGE are subjected to high mutation rates and recombination; and current techniques for the identification of horizontal transfer events suffer from precise predictions of borders of MGE inserts within a genome. Hence our ignorance of MGE behavioral measures conceals the functional and evolutionary importance of the environmental bacterial community. MGE are recognized as atypical genomic entities in prokaryotic genomes that influence the dissemination of genes that contribute to bacterial antibiotic resistance, diversity and virulence [1]. Virulence associated genomic elements were initially detected in human pathogenic microorganisms. These virulence determinants have since been detected in numerous environmental species that put mankind in jeopardy, as there still is an emergence of pathogens that harbor genes of yet unknown function. Besides pathogenicity, MGE may confer other traits such as: fitness, metabolic versatility, adaptability, symbiosis, commensalism, and speciation [2].

Genomes of every bacterial species may be characterized by a unique oligonucleotide usage pattern (OUP) designated as a genomic signature. OUP, also referred to as a bias in frequencies of short oligonucleotides of 2-7 bp, serve as a prevalent characteristic that distinguish between different organisms while being invariant along the large part of the genome [3, 4]. However, explorations of genomic sequences made it evident that DNA composition may vary significantly in bacterial chromosomes. At least partly these variations result from the exchange of genetic components between bacterial species by a mechanism of horizontal transfer [5].

Horizontally transferred genomic regions often possess DNA compositional characteristics which are distinct from the rest of the chromosome. It has previously been shown that divergent genomic segments can be detected based on their OUP [3, 6]. Yet, it was proved that oligonucleotide frequencies serve as phylogenetic signals, and variations in signature patterns could be revealed by studying distributions of words as small as dinucleotides [7]. Distributions of longer words were later studied suggesting that their usages may be efficient in the classification of species since shorter words are poorly species specific. Tetranucleotide frequencies were shown to be highly specific and could therefore be used to discriminate between bacterial species [8, 9]. These observations motivated the development of SeqWord Gene Island Sniffer (SWGIS), a novel tool that examines variations in frequencies of oligonucleotides and traces down the distributions of mobile genomic elements across genomes by analyzing the patterns of 4-bases long words. The method that is implemented in the tool is based on the concepts that have been introduced previously [5, 10].

II. Results and Discussion

SWGIS is a new computational tool for an automated identification of MGE in bacterial and plasmid DNA sequences. It is available for download at the site

B. O. Author is with the University of Pretoria, Department of Biochemistry, Bioinformatics and Computational Biology Unit, Lynnwood Rd., Hillcrest, Pretoria, South Africa (e-mail: bezuidt@gmail.com).

L.-M. G. Author is with the Laboratoire de Bioinformatique des Génomes et des Réseaux, Université Libre de Bruxelles, 1050 Bruxelles, Belgium (e-mail: gipsi@scmbb.ulb.ac.be).

R. O. N. Author is with the University of Pretoria, Department of Biochemistry, Bioinformatics and Computational Biology Unit, Lynnwood Rd., Hillcrest, Pretoria, South Africa (corresponding author, phone: +2712-420-5810; fax: +2712-420-5800; e-mail: oleg.reva@up.ac.za).

www.bi.up.ac.za/SeqWord/sniffer/index.html.The approach is based on the analysis of compositional biases in the genome-wide distribution of tetranucleotides as it was described previously [5, 11]. In SWGIS the ability to identify precise insertion borders has significantly been improved as compared to the method used in the previously published SeqWord Genome Browser [11]. In contrast to the latter method, SWGIS allows a fully automated processing of multiple genomes per single run. A global search of lateral inserts throughout 637 complete bacterial genomes with SWGIS retrieved 3,518 putative MGE. The obtained dataset of genomic islands was compared with the MGE identified by a method independent from a compositional based approach to examine how well the results fitted together. Comparison of the results obtained from SWGIS with those obtained earlier by Prophinder — a tool that predicts prophages on the basis of gene annotation and similarity searches of conserved DNA pairs using BLASTP [12], showed consistency in many cases. However, SWGIS failed to identify short and ancient ameliorated MGE that were efficiently detected by Prophinder, whereas Prophinder was deficient in the identification of horizontally transferred gene cassettes and truncated MGE that were detected by SWIGS, likely because they do not harbor any phage-specific genes. Thus, a combination and synergistic usage of the both latter methods may be recommended as they could increase the efficiency of MGE detection.

Early sequence comparisons of MGE revealed that they are genetic mosaics, where regions with sequence similarity alternate with unrelated regions [13]. To study the evolutionary parameters and networks of MGE sharing similar DNA sequences, the BLASTN algorithm was used, where every MGE sequences were compared against one another in a pairwise alignment fashion. The BLASTN pairwise alignment regions were computationally processed in a way that the overlapping or adjacent segments longer than 100 bp were fused into longer regions and stored in the database along with the information of all counterpart DNA fragments obtained from other MGE that share sequence similarity. A total of 7,302 BLAST matching regions were found in 1,570 of the 3,517 identified MGE. Up to 8 (in average 2) independent BLAST matches were obtained per MGE that is in consistence with the MGE mosaic structure hypothesis. The similarity search was conducted in order to group genomic islands according to the compositional features that they have in common and to also allow evaluations and inferences on homology between MGE obtained from varying bacterial lineages. The database of the identified MGE and their homologous fragments is freely accessible at http://anjie.bi.up.ac.za/geidb/GEI-neighbours.php (Fig. 1).

Following the all against all MGE BLASTN search, a total of 1,328 groups were created and stored in Genbank GBFF files available for download from ftp://milliways.bi.up.ac.za/SeqWord/MGE/. These files were ordered according to the total number of MGE homologous fragments that each group entails. Annotation of the top 10 groups of MGE sharing the BLASTN matching regions is presented in Table 1.

The largest group #1 comprises of 2,648 MGE BLAST matches from 31 genera of microorganisms. The MGE in the group share similarities in genes that are involved in a variety of entities, such as transposable elements and proteins involved in lateral transfer of genetic elements and bacterial virulence. The other sets of genes shared in the group are functionally uncharacterized due to poor annotation. Multitude of such genes appears to be either hypothetical or putative. Some of the most occurring ones are: putative inner and outer membrane proteins, transposases and IS elements. Sets of such genes are the most shared within the group, yet the fact that they are uncharacterized deprives the understanding of the functional characteristics that are conserved in and between MGE. However, those with the predicted functionality are essential in bacterial virulence and the synthesis of surface polysaccharide O-antigens.

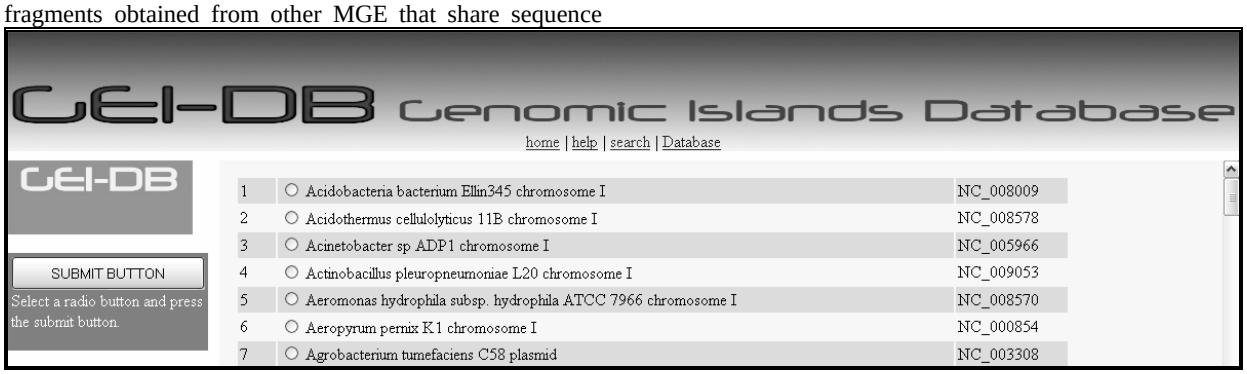

Fig. 1. GUI of the database of MGE identified in bacterial genomes. (The current database interface was tested for compatibility only with Mozilla Firefox).

TABLE I
ANNOTATION OF THE TOP 10 GROUPS OF MGE SHARING THE BLASTN MATCHING REGIONS

Gr	Size	Annotation of the most common genes (ordered by frequencies of occurrence)	Genera
1	2648	Hypothetical protein, IS elements, transposases, integrases, fimbrial-like adhesin proteins, outer membrane proteins, pathogenicity island proteins, glucose-1-phosphate thymidylyltransferases, dtdp-glucose 4,6-dehydratases, cell invasion proteins, arac-type regulatory proteins, dna-binding transcriptional regulators, hydroxyethylthiazole kinases, dtdp-4-dehydrorhamnose 3,5-epimerases, gdp-mannose 4,6-dehydratases, regulator of length of O-antigen components, 6-phosphogluconate dehydrogenases, bacteriophage proteins, inner membrane proteins, rhs family proteins, tRNA-ser, dtdp-6-deoxy-l-mannose-dehydrogenases, transport system permeases, acetyl-coa acetyltransferases, AMP nucleosidases, d-serine dehydratases, invasion plasmid antigens, multidrug resistance proteins, lipoproteins, 5-keto-4-deoxyuronate isomerases, arsenate reductases, isocitrate dehydrogenases, type-1 fimbrial proteins, type 1 fimbriae regulatory proteins.	*Escherichia, Shigella, Salmonella, Shewanella, Pseudomonas, Sodalis, Erwinia, Polaromonas, Brucella, Aeromonas, Methylobacillus, Alcanivorax, Marinobacter, Dechloromonas, Yersinia, Idiomarina, Desulfovibrio, Neisseria, Herminiimonas, Burkholderia, Actinobacillus, Saccharophagus, Psychrobacter, Chlorobium, Agrobacterium, Hahella, Haemophilus, Pelodictyon, Rhodoferax, Bacteroides, Photorhabdus.*
2	180	Hypothetical protein, frpc operon proteins, adhesins, is1016c2 transposases, putative bacteriocin resistance proteins, mafb-related proteins.	*Neisseria*
3	98	Lipopolysaccharide core biosynthesis proteins, heptosyl transferases, adp-heptose--lps heptosyltransferases, glucosyltransferases, adp-l-glycero-d-mannoheptose-6-epimerases, hypothetical proteins, lipopolysaccharide core biosynthetic proteins.	*Salmonella, Shigella, Escherichia, Erwinia*
4	83	Transposases, hypothetical proteins, zinc-containing alcohol dehydrogenase.	*Pseudomonas*
5	74	Hypothetical protein, phage infection proteins, hyaluronoglucosaminidases.	*Streptococcus*
6	66	Hypothetical proteins, putative secreted proteins, transposases, dtdp-glucose 4,6-dehydratases, putative capsular polysaccharide biosynthesis proteins.	*Corynebacterium, Bacillus*
7	46	Phosphoribosylglycinamide formyltransferases, phosphoribosylamine--glycine ligases, phosphoribosylaminoimidazole carboxylases, phosphoribosylaminoimidazole synthetases.	*Streptococcus*
8	45	Transposases, hypothetical proteins.	*Pseudomonas*
9	40	Hypothetical proteins, transposases	*Bacillus*
10	39	Diguanylate cyclase/phosphodiesterases, hypothetical proteins, peptidase m16 domain proteins, sensory box proteins, fatty acid oxidation complex proteins.	*Shewanella*

The MGE that constituent the group #1 inhabit the genomes of gamma-, alfa- and delta-Proteobacteria, and Chlorobi. Most of these MGE genes encode fimbria that apparently allow adherence of bacterial cells to the host cell surface; invasion proteins, toxin subunits and lipoproteins to confer resistance to bactericidal effects and survival within phagolysosomes [14]. The other most occurring functional genes but for transposases, integrases and IS elements, are udp-glucose/gdp-mannose dehydrogenases and udp-glucose 6-dehydrogenases that are involved in different type of metabolic and biosynthesis pathways including nucleotide sugars metabolism, polysaccharide biosynthesis and synthesis of polyketide toxins and antibiotics [15, 16].

Fig. 2 represents the exchange of genetic materials of the first group of MGE sharing BLAST similarity. This exchange takes place across the borders of species, genera and bacterial classes. Many MGE in the group show a significant large number of sequence matches with *E. coli*, comprising a total of 1,089 hits followed by *Salmonella* with 757 hits that probably resulted from a biased overrepresentation of these organisms among completely sequenced bacteria.

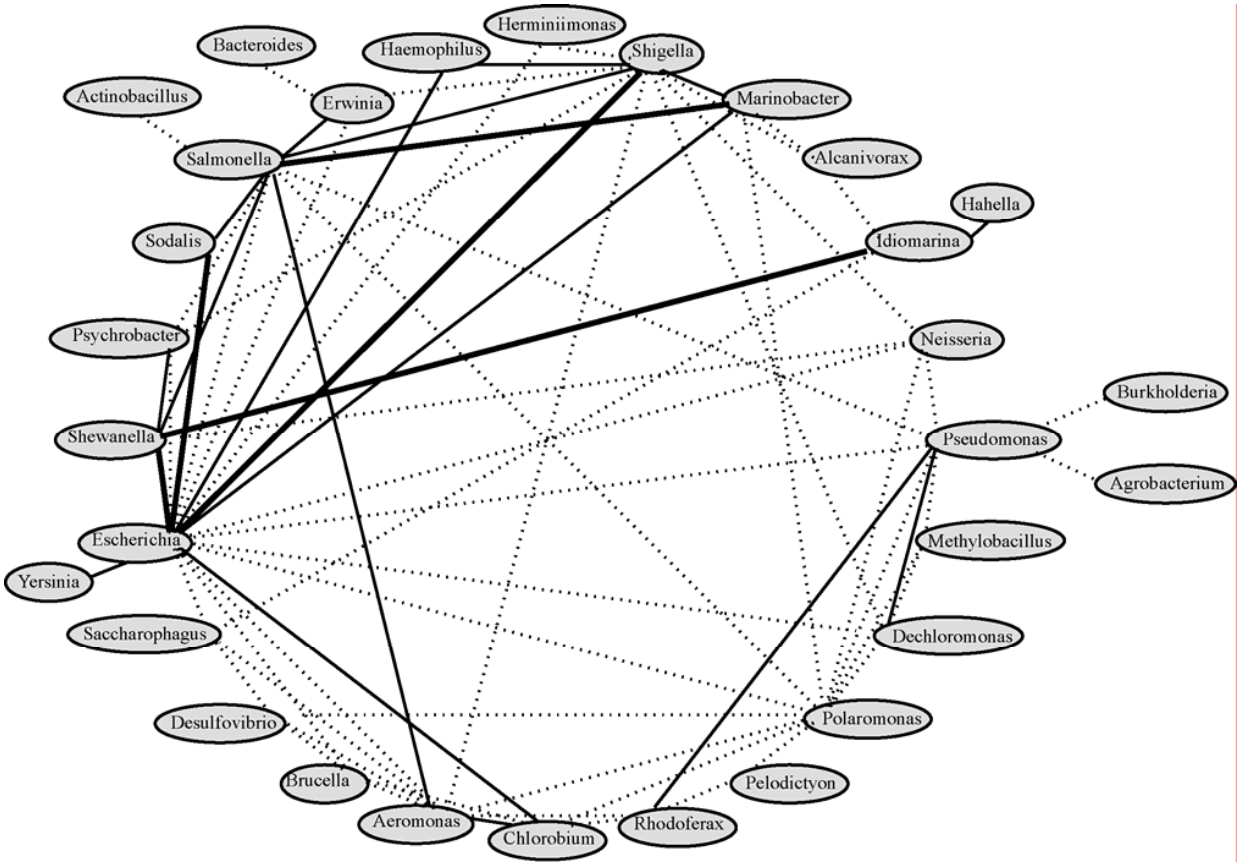

Fig. 2. Exchange of genetic materials between bacterial genera. Genera with the largest number of BLAST matching sequences (over 50 registered hits) are linked by thick solid lines, whereas those with less than 10 matching regions are represented by dotted lines and those sharing from 10 to 50 hits are linked by thing solid lines.

Lateral gene exchange is believed to be facilitated by the sharing of the same habitat. For example, marine bacteria *Shewanella*, *Idomarina*, *Marinobacter*, *Alcanivorax* and *Hahella* share a number of homologous MGE regions, probably because they inhabit the same niche. The habitat enclosure may also explain the fact that the organisms of three Chlorobi genera, — *Chlorobium*, *Bacteroides* and *Pelodictyon*, — exchange genetic materials with much more distant Proteobacteria rather than with each other. The other interesting event observed in this network is that in this network the organisms of the genus *Pseudomonas* exchange MGE entities equally likely with other gamma-Proteobacteria and the organisms belonging to alfa- and beta-Proteobacteria. Although many bacteria share a common pool of MGE, the organisms of such genera as *Yersinia*, *Actinobacillus*, *Bacteroides*, *Hahella*, *Burkholderia*, *Agrobacterium* appear to be blind branches of this network. They show to only accept MGE from specific donors rather than further transferring them to other recipients.

The next approach followed in the analysis was the clustering of MGE based on the hierarchy of proteins that they have in common and functional properties they possess. Measures of pairwise similarities of all the MGE protein sequences were obtained by using BLASTP. Proteins obtained from the search were clustered into families using the Markov clustering algorithm. The analysis was carried out as follows: an all vs all sequence comparison was conducted on all the genomic islands' CDSs and MGE sharing proteins of the same functional families were denoted as MGE classes. The classes that shared functional properties were weighted and assigned significance scores (SIG) designating their probable similarities in biochemical functions. However, assignment of these classes does not necessary imply significant sequence similarity within the class [17]. A class was denoted if SIG value for the group was bigger than 1.

In total 2,316 functional MGE classes were found to mostly be represented by 1 or 2 individual gene islands. Annotation of the top 5 largest classes is shown in Table 2. The largest classes were found to comprise of genes that encode polysaccharides and O-antigen biosynthesis enzymes and transport proteins (220 MGE), outer membrane proteins (173 MGE), and ABC iron transporters (19 MGE). For every identified MGE an oligonucleotide usage pattern (OUP) was calculated and searched for similarity against a reference database of OUP that were calculated for all the completely sequenced plasmids, bacteriophages and bacterial chromosomes. The latter approach allowed the identification

of MGE putative origins, and to also tracing down the ways of distributions of MGE throughout the donor-recipient chains.

I. Conclusion

The biggest functional class of MGE (Table 2) only partly overlaps with the biggest group #1 of MGE sharing significant nucleotide sequence similarity (Table 1). Grouping of mobile genetic elements by protein functionality does not imply common ancestry of these MGE but evolutionary approved combination of genes encoding specific proteins in a single mobile genetic segment. The most abundant functional classes of MGE comprise those genes that easily mobilize (transposases and integrases are obvious) and those that are more likely utilized in foreign bacteria as they tend to make them evolutionary advanced either immune tolerant. Table 2 shows that easily mobilized bacterial enzymes are those that are involved in lipopolysaccharides and O-antigen biosynthesis, ABC-transporters, regulatory proteins and restriction-modification system proteins. MGE can overcome borders between bacterial genera, families and classes. However, the exchange of genetic materials in bacteria is not a random event. Fig. 1 shows the existence of stable pairs of donors and recipients that often share the same habitat. Thriving of the bacteria in the same econiche cannot be the only condition necessary for the MGE exchange. The factors that benefit or limit the exchange of genetic materials between bacteria remain generally obscure and require future studies that may be grounded and facilitated by the SWGIS program and the database of MGE presented in this work.

Acknowledgements

This work was funded by the National Bioinformatics Network of South Africa.

TABLE II
Annotation of the top 5 largest MGE functional classes

Cl	Size	Annotation	Genera
1	220	Polysaccharide biosynthesis and transport proteins, ABC transporters, capsular polysaccharide biosynthesis proteins (CDP-6-deoxy-delta-3,4-glucoseen reductases, DegT/DnrJ/EryC1/StrS aminotransferases, dTDP-rhamnosyl transferases, gluconate-6-phosphate dehydrogenases, glycosyl transferases, kinases that phosphorylates core heptose of lipopolysaccharides), lipid A-core surface polymer ligases; lipopolysaccharide core biosynthesis proteins, O-antigen export system permeases; O-antigen ligases, O-antigen polymerases, regulator of length of O-antigen components of lipopolysaccharide chains.	*Acidobacteria, Actinobacillus, Aeromonas, Aeropyrum, Agrobacterium, Alcanivorax, Archaeoglobus, Arthrobacter, Bacillus, Bacteroides, Bdellovibrio, Bifidobacterium, Brucella, Carboxydothermus, Chlorobium, Chromobacterium, Clostridium, Corynebacterium, Coxiella, Dechloromonas, Deinococcus, Desulfitobacterium, Erwinia, Escherichia, Geobacillus, Geobacter, Gloeobacter, Gluconobacter, Hahella, Haloarcula, Herminiimonas, Idiomarina, Lactobacillus, Leptospira, Magnetococcus, Mannheimia, Marinobacter, Methanocorpusculum, Methanoculleus, Methanosaeta, Methanosarcina, Methanospirillum, Methanothermobacter, Methylobacillus, Moorella, Mycobacterium, Neisseria, Nitrosospira, Oceanobacillus, Oenococcus, Pasteurella, Pediococcus, Pelobacter, Pelodictyon, Photobacterium, Picrophilus, Polaromonas, Porphyromonas, Prochlorococcus, Pseudoalteromonas, Pseudomonas, Psychrobacter, Pyrobaculum, Ralstonia, Rhizobium, Rhizobium, Rhodoferax, Rhodopirellula, Saccharophagus, Salmonella, Shewanella, Shigella, Sodalis*
2	173	Membrane transporters, outer membrane and catabolic proteins, nickel/cobalt efflux protein, outer membrane proteases, isocitrate dehydrogenases, AraC family transcriptions regulators, IS-elements, transposases.	*Erwinia, Escherichia, Hahella, Pseudomonas, Psychrobacter, Salmonella, Shigella, Silicibacter*
3	19	ABC Mn+2/Fe+2 transporters	*Bacillus, Escherichia, Shigella, Sodalis*
4	12	Restriction-modification system proteins type I and III.	*Bacillus, Bdellovibrio, Bifidobacterium, Geobacillus, Lactobacillus, Neisseria, Picrophilus, Rhodopirellula*
5	11	ABC transport system proteins.	*Lactobacillus, Salmonella, Shigella.*

REFERENCES

[1] M. Juhas, J. R. van der Meer, M. Gaillard, R. M. Harding, D. W. Hood and D. W. Hood. "Genomic islands: tools of bacterial horizontal gene transfer and evolution," *FEMS Microbiol Rev*, 2009, 33, pp. 376–393.

[2] U. Dobrindt, B. Hochhut, U. Hentschel and J. Hacker. "Genomic islands in pathogenic and environmental organisms," *Nat Rev Microbiol*, 2004, 2, pp. 414–424.

[3] C. Dufraigne, B. Fertil, S. Lespinats, A. Giron and P. Deschavanne. "Detection and characterization of horizontal transfers in prokaryotes using genomic signatures," *NAR*, 2005, 33.

[4] J. Li and K. Sayood. "A genome signature based on Markov modeling," in *Proc IEEE Eng Med Biol Soc*, 2005, pp. 2832–2835.

[5] O. N. Reva and B. Tümmler. "Differentiation of regions with atypical oligonucleotide composition in bacterial genomes," *BMC Bioinformatics*, 2005, 6, pp. 251.

[6] S. Karlin. "Global dinucleotide signatures and analysis of genomic heterogeneity," *Curr Opinion Microbiol*, 1998, 1, pp. 598–610.

[7] S. Karlin, J. Mrázek and A. M. Campbell. "Compositional biases of bacterial genomes and evolutionary implications," *J Bacteriol*, 1997, 179, pp. 3899–3913.

[8] P. A. Noble, R. W. Citek and O. A. Oqunseitan. "Tetranucleotide frequencies in microbial genomes," *Electrophoresis*, 1998, 19, 528–535.

[9] D. T. Pride, R. J. Meinersmann, T. M. Wassenaar and M. J. Blaser. "Evolutionary implications of microbial genome tetranucleotide frequency biases," *Genome Res*, 2003, 13, pp. 145–158.

[10] O. N. Reva and B. Tümmler. "Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns," *BMC Bioinformatics*, 2004, 5, pp. 90.

[11] H. Ganesan, A. S. Rakitianskaia, C. F. Davenport, B. Tümmler and O. Reva. "The SeqWord genome browser: an online tool for the identification and visualization of atypical regions of bacterial genomes," *BMC Bioinformatics*, 2008, 9, pp. 333.

[12] G. Lima-Mendez, J. van Helden, A. Toussaint and R. Leplae. "Prophinder: a computational tool for prophage prediction in prokaryotic genomes," *Bioinformatics*, 2008, 24, pp. 863–865.

[13] W. S. Jermyn and E. F. Boyd. "Molecular evolution of *Vibrio* pathogenicity island-2 (VPI-2): mosaic structure among *Vibrio cholera* and *Vibrio mimicus* natural isolates," *Microbiology*, 2005, 151, pp. 311–322.

[14] K. Jann and B. Jann. "Assembly of cellular surface structures," in *Biology of the Prokaryotes*, ed. J. W. Lengeler, G. Drews and H. G. Schlegel, Blackwell Science, Oxford, 1999, pp. 555–570.

[15] M. Kanehisa. "From genomics to chemical genomics: new developments in KEGG," *NAR*, 2006, 34, pp. D354–D357.

[16] C. F. Snook, P. A. Tipton and L. J. Beamer. "Crystal structure of GDP-mannose dehydrogenase: a key enzyme of alginate biosynthesis in *P. aeruginosa*," *Biochemistry*, 2003, 42, 4658–4668.

[17] G. Lima-Mendez, J. van Helden, A. Toussaint and R. Leplae. "Reticulate representation of evolutionary and functional relationships between phage genomes," *Mol Biol Evol*, 2008, 25, pp. 762–777.