2

Linguistic Approaches for Annotation, Visualization and Comparison of Prokaryotic Genomes and Environmental Sequences

Oliver Bezuidt, Hamilton Ganesan, Phillip Labuschange, Warren Emmett, Rian Pierneef and Oleg N. Reva
University of Pretoria, Dep. Biochemistry,
Bioinformatics and Computational Biology Unit, Pretoria,
South Africa

1. Introduction

Sequencing of bacterial genomes has become a common technique of the present day microbiology. Thereafter, data mining in complete genome sequence is an essential step to uncover the uniqueness and evolutionary success of microorganisms. Oligonucleotide usage (OU or k-mer) statistics provides invaluable tools to get insight into genome organization and functionality.

The study of genome OU signatures has a long history dating back to early publications by Karlin et al. 1995, 1997, 1998, who focused mainly on dinucleotide compositional biases and their evolutionary implications. Statistical approaches of OU comparison were further advanced by Deschavanne et al., 1999, who applied chaos game algorithms; and by Pride *et al.*, 2003, who extended the analysis to tetranucleotides using Markov Chain Model simulations. Later, a number of practical tools for phylogenetic comparison of bacterial genomes (Coenye & Vandamme, 2004; van Passel et al., 2006); identification of horizontally transferred genomic islands (Mrázek & Karlin, 1999; Pried & Blaser, 2002; Nakamura et al., 2004; Azad & Lawrence, 2005; Dufraign et al., 2005; Becq et al., 2007) and assignment of unknown genomic sequences (Abe et al., 2003; Teeling et al., 2004) based on OU statistics became publicly available. These approaches exploited the notion that genomic OU composition was less variable within genomes rather than between them, regardless of which genomic regions had been taken into consideration (Jernigan & Baran, 2002). A general belief was that if a significant compositional difference was discovered in genomic fragments relative to the core genome, these loci most likely can be assigned to horizontally transferred genetic elements (transposons, prophages or integrated plasmids). This approach was criticized by several researchers (Koski et al., 2001; Wang 2001), who pointed out that codon bias and base composition are poor indicators of horizontal gene transfer. Therefore, there is a need for more informative parameters which also take into account higher order DNA variation. An overview of the current OU statistical methods based on di-, tetra- and hexanucleotides has been published recently (Bohlin et al., 2008). The conclusion of the review was that all methods were context dependent and, though being efficient and powerful, none of them were superior in all applications. Thus, the major motivation of our work was to develop more flexible and informative algorithms seamlessly integrating di- to heptanucleotides OU analysis for reliable identification of divergent genomic regions.

2. Linguistic approaches for genomics and metagenomics

Genome linguistics is respectively known as the analysis of frequencies of k-mers in genome wide DNA sequences. The basic hypothesis is that biased distribution of oligonucleotides in bacterial genome is genome specific and may serve as a signature. Each OU pattern may be characterized by a number of OU statistical parameters, namely: local pattern deviation (D), pattern skew (PS), relative variance (RV) and several others that will be explained below. The requirements for the OU statistics are as follows: i) distances between patterns of different word length (from di- through to heptanucleotides) calculated for the same sequence must be comparable; i.e. one may use longer word patterns to perform a large scale analysis and then switch to shorter word patterns for a more detailed view; ii) OU patterns calculated for sequences of different lengths must be comparable provided that the length of the sequence is longer than the specified thresholds; iii) alterations of OU patterns may be analyzed by different non-redundant parameters (D, PS and RV with different schemes of normalization by frequencies of shorter constituent words). Superimposition of these OU characteristics allows better discrimination of divergent genomic regions.

2.1 Oligonucleotide usage pattern concept

OU pattern was denoted as a matrix of deviations $\Delta_{[\xi 1 \ldots \xi N]}$ of observed from expected counts for all possible words of length N. Oligonucleotides or words are distributed in sequences logarithmically and deviations of their frequencies from expectations may be found as follows:

$$\Delta_w = \Delta_{|\xi 1 \ldots \xi N|} = 6 \times \frac{\ln\left(\dfrac{C^2_{|\xi 1 \ldots \xi N||obs} \sqrt{C^2_{|\xi 1 \ldots \xi N||e} + C^2_{|\xi 1 \ldots \xi N||0}}}{C^2_{|\xi 1 \ldots \xi N||e} \sqrt{C^2_{|\xi 1 \ldots \xi N||obs} + C^2_{|\xi 1 \ldots \xi N||0}}} \right)}{\sqrt{\ln\left(\left[C^2_{|\xi 1 \ldots \xi N||0} \Big/ C^2_{|\xi 1 \ldots \xi N||e} \right] + 1 \right)}} \tag{1}$$

where ξ_n is any nucleotide A, T, G or C in the N-long word; $C_{[\xi 1 \ldots \xi N] | obs}$ is the observed count of a word $[\xi_1 \ldots \xi_N]$; $C_{[\xi 1 \ldots \xi N] | e}$ is its expected count and $C_{[\xi 1 \ldots \xi N] | 0}$ is a standard count estimated from the assumption of an equal distribution of words in the sequence: ($C_{[\xi 1 \ldots \xi N] | 0} = L_{seq} \times 4^{-N}$).

Expected counts of words $C_{[\xi 1 \ldots \xi N] | e}$ were calculated in accordance to the applied normalization scheme. For instance, $C_{[\xi 1 \ldots \xi N] | e} = C_{[\xi 1 \ldots \xi N] | 0}$ if OU is not normalized, and $C_{[\xi 1 \ldots \xi N] | e} = C_{[\xi 1 \ldots \xi N] | n}$ if OU is normalized by empirical frequencies of shorter constituent words of length n. The expected count of a word $C_{[\xi 1 \ldots \xi N] | e}$ of the length N in a L_{seq} long sequence normalized by frequencies of n-mers ($n < N$) is calculated as follows:

$$C_{[\xi_1 \ldots \xi_{l_w}]\|n} = L_{seq} \times F_{[\xi_1 \ldots \xi_n]} \times \prod_{i=2}^{N-n+1} \left(\frac{F_{[\xi_i \ldots \xi_{i+n-1}]\xi_{i+n}}}{\sum\limits_{\xi}^{A,T,G,C} F_{[\xi_i \ldots \xi_{i+n-1}]\xi}} \right) \tag{2}$$

Where the $F_{[\xi 1 \ldots \xi n]}$ values are the observed frequencies of a particular word of length n in the sequence and ξ is any nucleotide A, T, G or C. For instance, the expected count of a word ATGC in a sequence of L_{seq} nucleotides normalized by frequencies of trinucleotides would be determined as follows:

$$C_{ATGC} = L_{seq} \times F_{ATG} \times \frac{F_{TGC}}{F_{TGA} + F_{TGT} + F_{TGG} + F_{TGC}} \tag{3}$$

Two approaches of normalization have been exploited where the F values are calculated for the complete genome (generalized normalization) or for a given sliding window (local normalization).

The distance D between two patterns was calculated as the sum of absolute distances between ranks of identical words (w, in a total 4^N different words) after ordering of words by $\Delta_{[\xi 1 \ldots \xi N]}$ values (equation 1) in patterns i and j as follows:

$$D(\%) = 100 \times \frac{\sum_{w}^{4^N} \left| rank_{w,i} - rank_{w,j} \right| - D_{\min}}{D_{\max} - D_{\min}} \tag{4}$$

Application of ranks instead of relative oligonucleotide frequency statistics made the comparison of OU patterns less biased to the sequence length provided that the sequences are longer than the limits of 0.3, 1.2, 5, 18.5, 74 and 295 kbp for di-, tri-, tetra-, penta-, hexa- and heptanucleotides, respectively (Reva and Tümmler, 2004).

PS is a particular case of D where patterns i and j are calculated for the same DNA but for direct and reversed strands, respectively. $D_{max} = 4^N \times (4^N - 1)/2$ and $D_{min} = 0$ when calculating a D, or, in a case of PS calculation, $D_{min} = 4^N$ if N is an odd number, or $D_{min} = 4^N - 2^N$ if N is an even number due to the presence of palindromic words. Normalization of D-values by D_{max} ensures that the distances between two sequences are comparable regardless of the word length.

Relative variance of an OU pattern was calculated by the following equation:

$$RV = \frac{\sum_{w}^{4^N} \Delta_w^2}{\left(4^N - 1\right) \times \sigma_0} \tag{5}$$

where N is word length; Δ^2_w is the square of a word w count deviation (see equation 1); and σ_0 is the expected standard deviation of the word distribution in a randomly generated sequence which depends on the sequence length (L_{seq}) and the word length (N):

$$\sigma_0 = \sqrt{0.02 + \frac{4^N}{L_{seq}}} \tag{6}$$

2.2 Compositional polymorphism of bacterial genomes

Biased distribution of k-mers may be explained by selective forces of DNA reparation enzymes of microorganisms, which may sense stereochemical properties of DNA fragments. A strong correlation was discovered between frequencies of oligonucleotides and their physicochemical properties such as base stacking energy, propeller twist angle, bendability, protein deformability and position preference in the DNA helical repeats (Fig. 1) calculated by the additive scale approach proposed by Baldi & Baisnée, 2000. It looks plausible that proteins of the replication-reparation system may sense the stereochemical properties of the DNA molecule and allow higher mutation rates in atypical regions; however, it has not yet been proved experimentally. The latter may explain the pervasive properties of genomic signatures that are reported for bacterial genomes (Jernigan & Baran, 2002). Despite a significant conservation of the OU pattern in genomic core DNA sequences, every bacterial genome contains loci of DNA which differ significantly from the core sequence. These loci usually contain gene clusters for ribosomal RNA and ribosomal proteins, horizontally transferred genomic islands, DNA fragments with multiple repeats and some other features. Superimposition of different OU parameters allows discrimination of divergent genomic regions. Briefly: rRNA operons are characterized by extremely high PS and low RV; giant genes with multiple repeated elements have high or moderate PS and high RV; horizontally transferred genetic elements are characterized by increased divergence between RV and GRV accompanied by high D; and genes for ribosomal proteins show a moderate increase of D, PS and RV above genomic averages. In the examples given above D denotes the distance between a local pattern calculated for a sliding window and the global pattern determined for the complete genome; PS is local pattern skew; and RV and GRV are variances of local OU patterns normalized by GC-content of the sliding window and the complete genome, respectively.

A Web-based applet SeqWord Genome Browser (SWGB) was developed and available on-line at www.bi.up.ac.za/SeqWord/ to visualize DNA compositional variations in pre-calculated bacterial and viral genomes. The SWGB is basically comprised of two views, denoted by the 'Gene Map' and 'Diagram' tabs. The 'Gene Map' tab offers a simple view of an entire genome at a glance and gives users access to a number of important pre-calculated OU statistics superimposed on the gene map (Fig. 2). The 'Diagram' tab allows flexible filtering of the underlying data based on the criteria chosen by users. Although the underlying data is pre-calculated, the user may, by simply changing selected parameters, generate many alternative plots, which give different insights into the natural genomic variation. On the dot-plot diagram, each genomic fragment selected by the sliding window is represented by a dot with X and Y coordinates, which correspond to values of OU parameters chosen from X and Y drop-down lists, respectively. The Z axis parameter may be set as well. In this case, the dots are coloured by values of OU parameters selected for the Z axis, and the colour range is displayed on the vertical colour bar on the left of the plot area (Fig. 3).

Several routines have been developed to identify horizontally transferred genomic islands, genes for ribosomal RNA and proteins, non-functional pseudogenes and genes of other functional categories. All these routines are described in detail with illustrations in supplementary web-pages (use the 'Help' link in the applet window). Take for example the genome of *Pseudomonas putida* KT2440, a known mosaic genome with 105 genomic islands above 4000 bp in length (Weinel et al., 2002). Many of these features can be visualized at a glance using the SWGB without any in depth analysis (see Fig. 2). On the 'Diagram' view the parameters n1_4mer:RV, n1_4mer:GRV and n0_4mer:D were selected for the X, Y and Z axes, respectively, as we showed previously (see Fig. 3).

Plotting local relative OU variance (RV) against global relative variance (GRV) basically shows the effect of normalization by global mononucleotide content. The core genome is then represented on the dot plot as the positive linear correlation line where RV ≈ GRV (Fig. 3). In other words, these fragments exhibit such compositional closeness to the core genome that normalizing by local mononucleotide content does not have any effect compared to normalizing by the global content. These genomic fragments also exhibit compositional similarity to the genomic average; and are therefore coloured blue. Scattered dots lying peripheral to the expected strong linear correlation do not belong to the core genome and also have a higher distance from the genomic average and are hence coloured green and red.

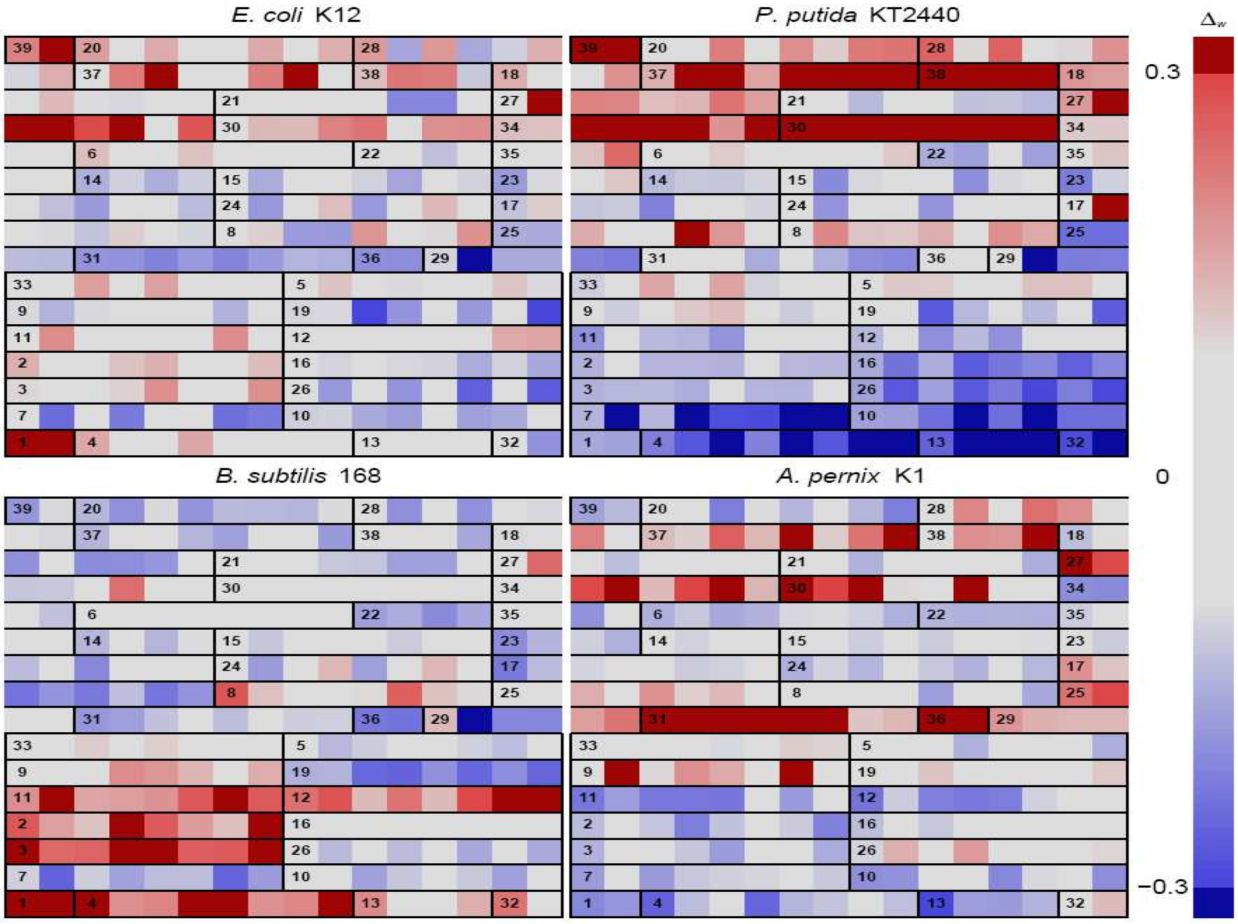
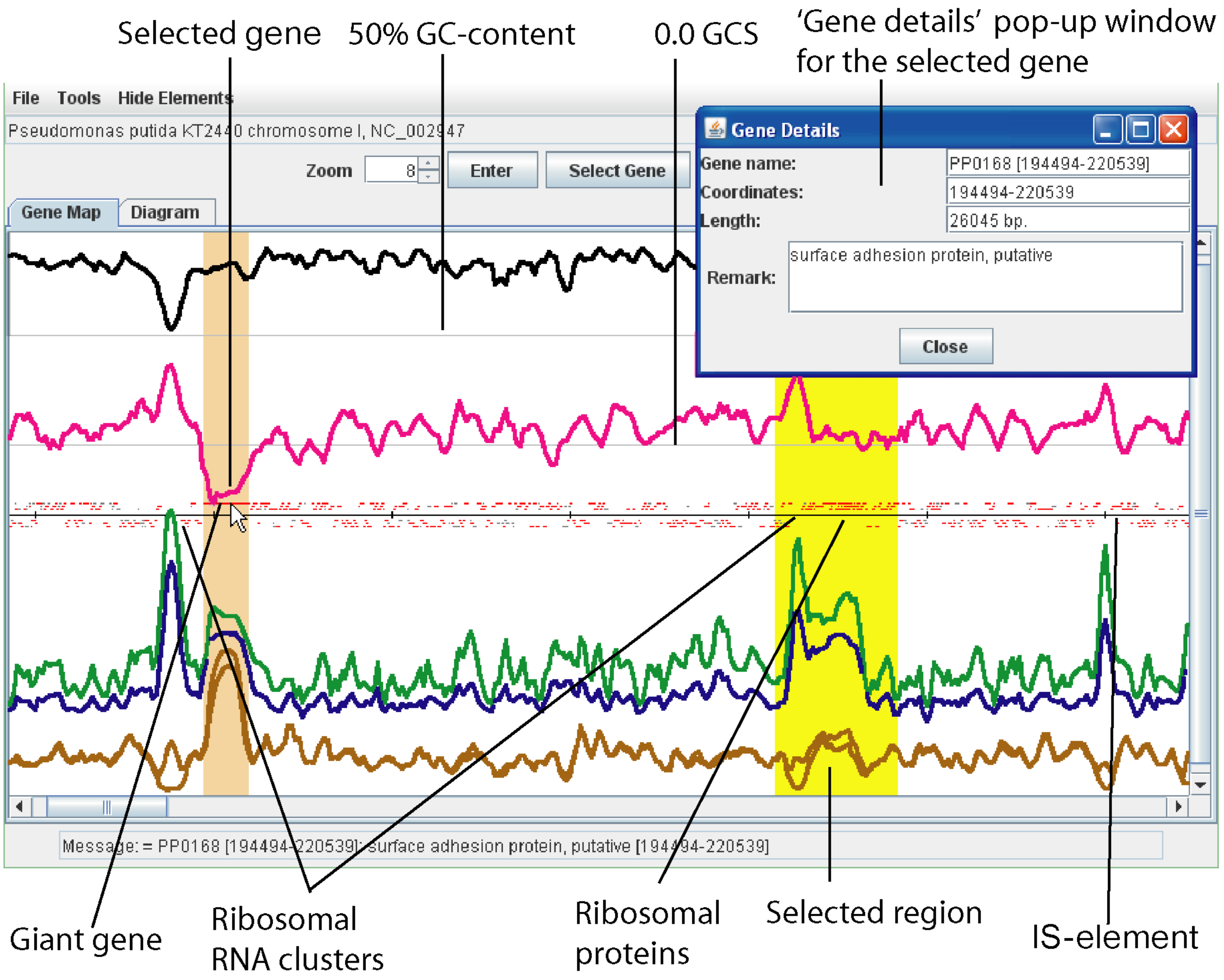


Fig. 1. Tetranucleotide usage patterns calculated for genomes of four different organisms. The deviations Δ_w of observed from expected counts are shown for all 256 tetranucleotide permutations (16×16 cells) by a colour code (right bar) depicting overrepresented (red) and rare (blue) words. The words are grouped into 39 equivalence classes and ordered by decreasing base stacking energy row-by-row starting from the upper corner (class 39).



Pseudomonas putida KT2440 chromosome I, NC_002947

Zoom 8

Gene name: PP0168 [194494-220539]

Coordinates: 194494-220539

Length: 26045 bp.

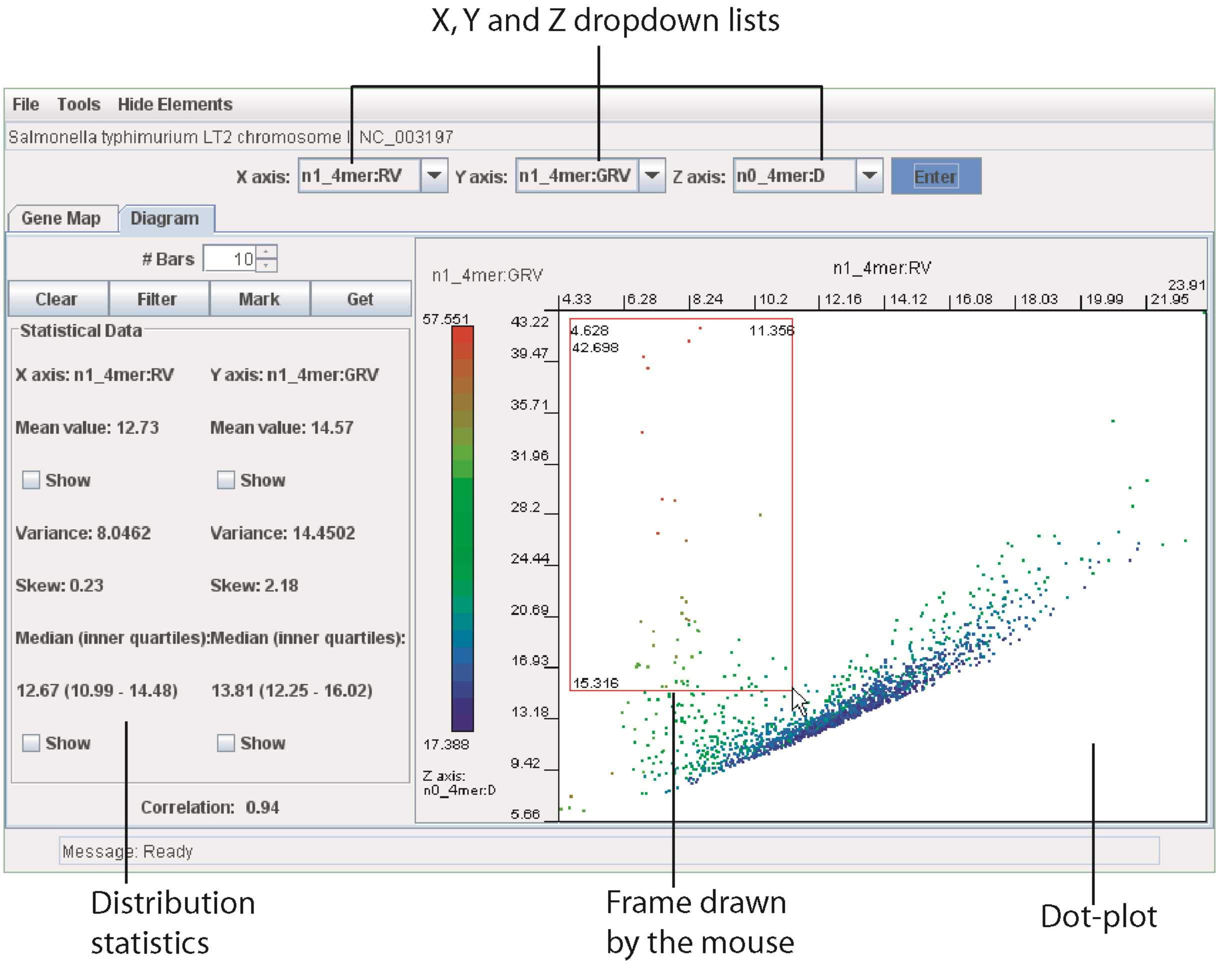
Remark: surface adhesion protein, putative

Message: = PP0168 [194494-220539]: surface adhesion protein, putative [194494-220539]

Fig. 2. Identification of divergent genomic regions on the 'Gene Map' view. Superimposition of different OU parameters such as GC (black line), GCS (pink), PS (green), D (blue), GRV (upper brown line) and RV (lower brown line) allows discrimination of divergent genomic regions. In this example a part of the chromosome of *Pseudomonas putida* KT2440 (127-774 kbp) is displayed in the applet window. A genomic fragment was highlighted using the function 'Select region' and a giant gene, PP0168, was selected by 'Select gene'. A pop-up window 'Gene Details' was opened by double-clicking the gene on the map. Genes are indicated by red and grey (for hypothetical) bars. The black horizontal line separates genes by their direction of translation.

Changing of the set of parameters as shown on Fig. 4 allows separation of core housekeeping genes from clusters of genes encoding ribosomal proteins and ribosomal RNA, vestigial regions with pseudogenes and giant genes with multiple repeats.

SWGB is linked to a database of pre-calculated OU patterns of bacterial genomes (2243 complete sequences, including bacterial chromosomes, plasmids and some viruses were available at the time of writing of this chapter and new sequences are regularly being added). The SWGB allows tentative annotation of the various divergent regions and provides overviews for use in comparative genomics. Users may download the command line version of the OligoWords program to analyze locally their own sequences. A packaged version of the SWGB allows users to view and manipulate their OligoWords results locally using a compatible web-browser.



Salmonella typhimurium LT2 chromosome I NC_003197

X axis: n1_4mer:RV Y axis: n1_4mer:GRV Z axis: n0_4mer:D

Bars 10

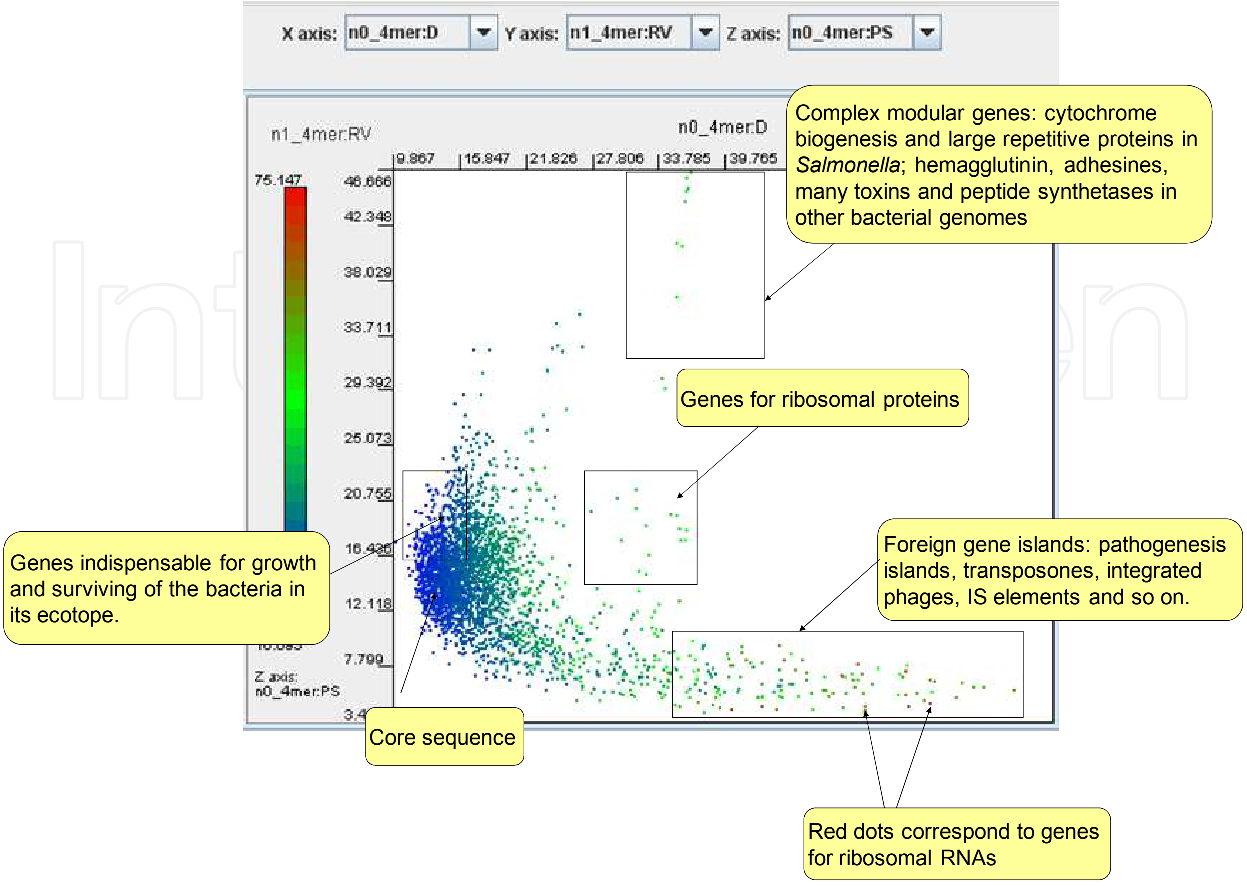
Statistical Data

X axis: n1_4mer:RV	Y axis: n1_4mer:GRV
Mean value: 12.73	Mean value: 14.57
□ Show	□ Show
Variance: 8.0462	Variance: 14.4502
Skew: 0.23	Skew: 2.18
Median (inner quartiles):	Median (inner quartiles):
12.67 (10.99 - 14.48)	13.81 (12.25 - 16.02)
□ Show	□ Show

Correlation: 0.94

Message: Ready

Fig. 3. The ‘Diagram’ view. In this example n1_4mer:RV, n1_4mer:GRV and n0_4mer:D were selected for the X, Y and Z axes, respectively. Every dot on the dot-plot corresponds to a genomic fragment selected by the sliding window. Dots are spread and coloured in accordance with their values of the selected statistical OU parameters. Information for each dot may be found by one of the following methods: i) information for a dot pointed by the mouse is shown in the ‘Message’ bar; ii) double clicking a dot returns us to the ‘Gene map’ tab with the corresponding genomic fragment highlighted; iii) framing the dots and clicking the ‘Get’ button opens a new applet window with the information about all selected regions. In this example the genomic regions of *Salmonella typhimurium* LT2 (NC_003197) which correspond to horizontally transferred genetic elements were selected.



X axis: n0_4mer:D Y axis: n1_4mer:RV Z axis: n0_4mer:PS

Complex modular genes: cytochrome biogenesis and large repetitive proteins in *Salmonella*; hemagglutinin, adhesines, many toxins and peptide synthetases in other bacterial genomes

Genes for ribosomal proteins

Genes indispensable for growth and surviving of the bacteria in its ecotope.

Foreign gene islands: pathogenesis islands, transposones, integrated phages, IS elements and so on.

Core sequence

Red dots correspond to genes for ribosomal RNAs

Fig. 4. The ‘Diagram’ view. In this example n0_4mer:D, n1_4mer:RV and n0_4mer:PS were selected for the X, Y and Z axes to identify genomic areas of interest.

2.3 Signature words and identification of environmental sequences

In genomic and metagenomic literature the occurrences of 2 to 7 bp oligonucleotides have been studied extensively. Patterns of short oligomers (words) have been used successfully for DNA read clustering (Chatterji et al., 2008; Kislyuk et al., 2009; Saeed & Halgamuge, 2009). However, short oligonucleotide patterns usually do not provide enough information for binning DNA reads to bacterial species or higher taxonomic units. Longer words of 8 to 14 nucleotides generally are more specific. Nevertheless, it was illustrated that the approach based on the analysis of frequencies of all the permutations of oligonucleotides of a given length such as discussed above is not effective for analysis of 8 to 14 letter words (Bohlin et al., 2008). Furthermore, an analysis of all possible permutations of 8 to 14 bp words would be computationally expensive because the total number of possible permutations of words is 4^L where L is the word length. For words of length 8 to 14 bp, this quantity becomes very large. Additionally, the random changes in the frequencies of such a large number of words obscure the genome specific information present in a few signature words. According to Kirzhner et al., 2005, less than 1% of 10-mers are informative in a large-scale comparison of bacterial genomes.

A first investigation into exploiting the information present in 8 to 14 letter words in 13 strains from the genus *Pseudomonas* was made by Davenport et al., 2009. It has been shown as well that certain profiles of signature words may help to distinguish DNA fragments that originated from different genomes (Saeed & Halgamuge, 2009).

In this work an attempt was made to standardize the linguistic approaches of binning metagenomic DNA reads by creating a database of signature words of Eubacteria and Archaebacteria represented in GenBank. The first step was to develop methods for summarizing the large amounts of data associated with these words. To avoid using the words that do not provide any taxonomic information, the most divergent words were selected and stored in the database. Currently, the frequencies of 172,636 signature words calculated in 768 bacterial chromosomes are stored in a binary database file available for download from the SeqWord project Web-site at www.bi.up.ac.za/SeqWord/oligodb/. Furthermore, scoring functions were designed, which measure the likelihood that a given DNA fragment originated from a given taxonomic group.

This tool also may be used to identify the origin of DNA sequences or whole clusters of DNA sequences. There are a number of programs such as LikelyBin (Kislyuk et al., 2009), CompostBin (Chatterji et al., 2008) and some others that cluster DNA sequences, but there is no default methodology for inferring the taxonomic affinity of these clusters. Typically, BLAST is used to compare these clusters to the databases of DNA sequences. Frequently these clusters consist of several short sequences that cannot be easily assembled, which makes using BLAST complicated. TETRA identifies long unknown DNA sequences by comparison of the whole patterns of frequencies of tetranucleotides (Teeling et al., 2004). A tool based on occurrences of 8 to 14 letter words is expected to work equally well both on clusters of sequences and single long sequences.

2.3.1 Statistical background of selection of signature words

To identify prospective signature words, the distribution score coefficients were calculated for each 8 to 14-mer permutation as follows:

$$DS = \frac{100000}{\sqrt{\mu^2 + \sigma^2}} \tag{7}$$

where μ indicates the average length of spans (in base pairs) between the repeated words in the sequence (i.e., μ = sequence_length/number_of_words); and σ is the standard deviation of the span lengths. The DS for a word increases in value when there is a high frequency of occurrence (μ is minimal) and the words are evenly distributed (σ tends to 0). The DS coefficient assigns low scores to infrequent words and to local repeats while giving higher scores to words occurring frequently and evenly distributed throughout the genome. The words that have DS above the threshold value of 0.3 in at least one genome were included in the template of signature words. Furthermore, their frequencies were recalculated for all genomes. The threshold value was empirically determined to ensure that the template contained similar numbers of words for each different word length and that an appropriate ratio between template size and word specificity is obtained. The final template contains 172,636 signature words; that is approximately 0.1% of the total number of all possible permutations of 8 bp to 14 bp oligonucleotides. Note that in this work each oligonucleotide and its reverse complement were considered as the same word so that the two different strands of the DNA molecule will be assigned identical scores.

To improve maintenance and operational flexibility of the database, the numeric frequencies of words may be replaced by percentile values without any significant loss of information. The empirical cumulative distribution of the frequency of occurrence of the words in the template was studied and the following non-linear regression model was fitted to the data:

$$f = \frac{\exp(3p + 9)}{L^{4.5}} \tag{8}$$

where f is the frequency of a word per 100 Kbp, L is the word length and p is the probability that the word occurs at a frequency less than or equal to f. For example, according to equation 8 for 50% of words of the length 8 bp ($p = 0.5$; $L = 8$) the frequency f is in the range from 0 to 3.13 words per 100 Kbp of the given sequence; and 90% ($p = 0.9$) of 8-mers have frequencies from 0 to 10.41. Four categories were designated for rare ($p < 0.1$), common ($0.1 \le p < 0.5$), frequent ($0.5 \le p < 0.9$) and abundant ($0.9 \le p$) words. The borders of the percentile categories calculated by equation 8 are shown in Table 1.

The performance of a signature word to separate DNA reads of different origins or to bin a cluster of reads to a taxonomic unit depends on the set of taxonomic units to be differentiated and the task formulation. Several scoring algorithms were used in this study. All the scores were normalized to a range from 0 to 10. The scores were used to order the words in the database and to select the ones with the highest scores.

Word divergence is scored by the variance of percentile values (see Table 1) in the selected genomes normalized by the maximum possible variance. The most diverse word would be rare in one half of the selected genomes and abundant in the other half of the genomes.

To select the words, which are rare or abundant in all selected genomes, the following score was used:

$$\text{Score} = 10 \times (\text{Av} - 0.05)/0.9 \tag{9}$$

where *Av* is the average of the percentile values calculated for a word in selected genomes. To select rare words (*10 – Score*) was used.

The perfect word to distinguish between taxa is one that is similarly distributed in genomes belonging to the same taxon but is differently distributed in different taxa. The scores were assigned in the spirit of ANOVA by computing the ratio of the sums of square deviations over the average values between taxa and within every taxon.

Another practical task may consist in distinguishing one taxon (outgroup) from a number of other taxa (counterparts) by diverse, abundant or rare words. In our study this approach was termed as *confronted comparison*. Three scoring algorithms were used:

$$DiversityScore = \left|Av_0 - Av_g\right| \Big/ \sqrt{1 + Var_0/n} \tag{10}$$

$$AbundanceScore = \left(10 + Av_0 - Av_g\right) \Big/ 2\sqrt{1 + Var_0/n} \tag{11}$$

$$ScarceScore = \left(10 + Av_g - Av_0\right) \Big/ 2\sqrt{1 + Var_0/n} \tag{12}$$

where Av_0 and Av_g are average frequencies of the word in genomes of the outgroup and counterpart taxonomic units, correspondingly; Var_0 is the variance of the word frequencies in the outgroup genomes and n is the number of genomes in the outgroup taxonomic unit.

Computer simulation of metagenomic datasets was done by the MetaSim program (Richter et al., 2008). DNA reads were clustered by the LikelyBin algorithm (Kislyuk et al., 2009). The database of signature words and the OligoDBViewer program are available for download from www.bi.up.ac.za/SeqWord/oligodb/.

Word length	Percentiles			
	Rare **– (0.0)**[*]	**Common** **+ (0.25)**	**Frequent** **++ (0.75)**	**Abundant** **+++ (1.0)**
8 bp	< 0.94†	≥ 0.94 and < 3.13	≥ 3.13 and < 10.4	≥ 10.4
9 bp	< 0.56	≥ 0.56 and < 1.85	≥ 1.85 and < 6.13	≥ 6.13
10 bp	< 0.35	≥ 0.35 and < 1.15	≥ 1.15 and < 3.81	≥ 3.81
11 bp	< 0.23	≥ 0.23 and < 0.75	≥ 0.75 and < 2.48	≥ 2.48
12 bp	< 0.15	≥ 0.15 and < 0.51	≥ 0.51 and < 1.68	≥ 1.68
13 bp	< 0.11	≥ 0.11 and < 0.35	≥ 0.35 and < 1.17	≥ 1.17
14 bp	< 0.08	≥ 0.08 and < 0.25	≥ 0.25 and < 0.84	≥ 0.84

[*] Rare, common, frequent and abundant words are marked as –, +, ++ and +++, respectively. The numeric values representing each percentile category are used for score calculations.

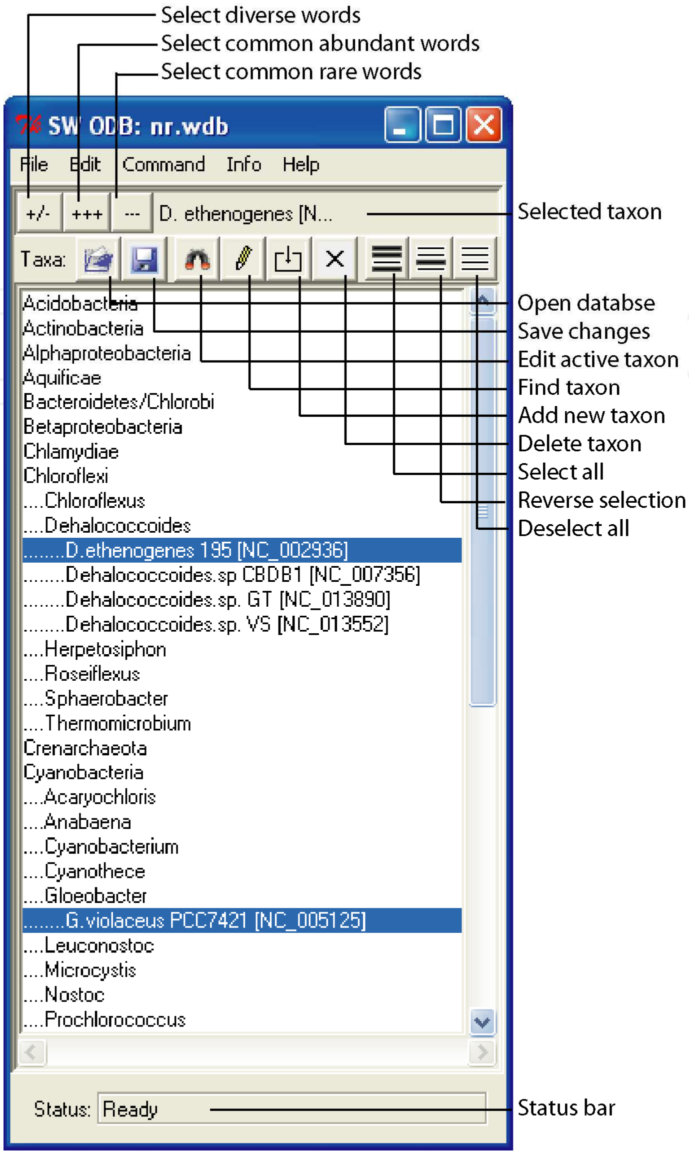
[†] These are *f*-values calculated by equation 8 for the cumulated likelihoods (*p*) 0.1, 0.5 and 0.9, respectively.

Table 1. The percentile border frequencies calculated for the words of different length.

2.3.2 OligoDBViewer and the database of signature words

The main window of OligoDBViewer is shown in Fig. 5. The functionality of the OligoDBViewer is described in detail on the project Web site. The program allows selecting genomes or taxonomic units from the list and searching for the best discriminative words by using the program functions accessible through the toolbar and the 'Command' menu. The resulting list of ordered words will be shown in the pop-up panel on the right hand side as in Fig. 6. In this example the diverse words were searched with the goal of separating the genomes of *Mycobacterium avium* k10 [NC_002944] and *M. tuberculosis* F11 [NC_009565]. The divergence scores are shown in column 'C'. The number of times an oligomer falls into the categories of rare, common, frequent or abundant words is shown respectively from left to right in column 'Stat'.

The list of the words returned by OligoDBViewer may be highly redundant. For example, words that only differ by a single nucleotide may be expected to have similar distributions in genomes. To reduce the redundancy of the selected words, several filter options may be set. The filter settings in Fig. 6 removes all the words that differ from the words with the highest scores by less than 30 % similarity (another option is available to set the minimal number of mismatches) as well as the words that are left or right shifted sub-words or shorter constituents of longer words that have higher scores. Then the list is cut off so that only the top 10 words remain. Additionally, the list of selected words may be filtered by the word length and the score threshold. All word filtering settings is reversible.



Acidobacteria
Actinobacteria
Alphaproteobacteria
Aquificae
Bacteroidetes/Chlorobi
Betaproteobacteria
Chlamydiae
Chloroflexi
....Chloroflexus
....Dehalococcoides
........D.ethenogenes 195 [NC_002936]
........Dehalococcoides.sp CBDB1 [NC_007356]
........Dehalococcoides.sp. GT [NC_013890]
........Dehalococcoides.sp. VS [NC_013552]
....Herpetosiphon
....Roseiflexus
....Sphaerobacter
....Thermomicrobium
Crenarchaeota
Cyanobacteria
....Acaryochloris
....Anabaena
....Cyanobacterium
....Cyanothece
....Gloeobacter
........G.violaceus PCC7421 [NC_005125]
....Leuconostoc
....Microcystis
....Nostoc
....Prochlorococcus

Fig. 5. The main window of OligoDBViewer.

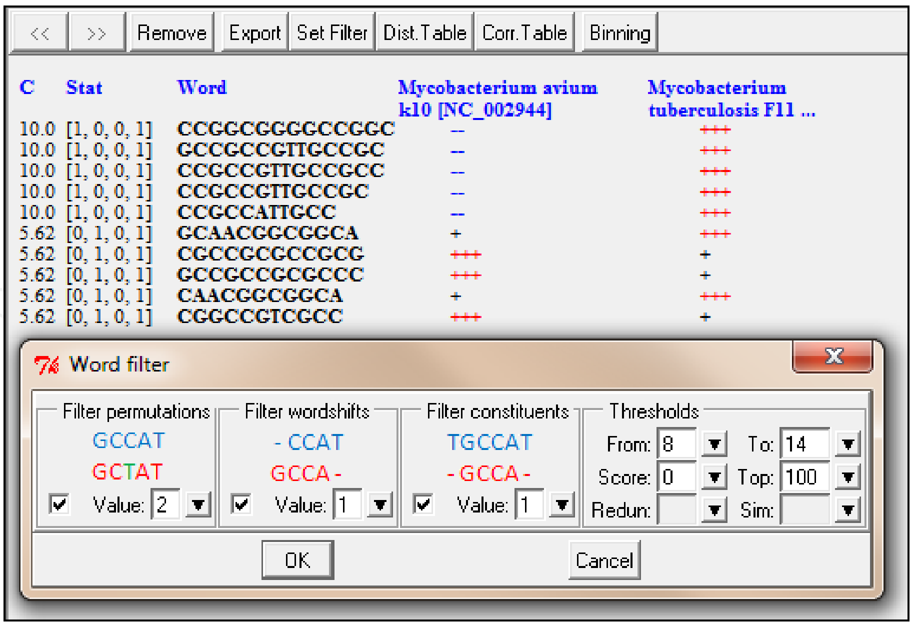
To facilitate large scale calculations and the database updates on remote servers, several command line utilities are available for download. They are fully described on the project Web site.

2.3.3 Algorithms of binning of clusters of DNA reads to taxonomic units

To estimate the similarity of a cluster of DNA reads to bacterial taxonomic units the percentile values were used (Table 1). All DNA reads of the cluster were concatenated in an artificial sequence and the frequency of the words normalized per 100 Kbp were counted (*f*-value). Then the *f*-values were converted to percentiles:

$$p = \frac{\ln(f) + 4.5\ln(L) - 9}{3} \qquad (13)$$

Note that equation 13 is the inverse of the equation 8. The meanings of the coefficients were explained above.



C	Stat	Word	Mycobacterium avium k10 [NC_002944]	Mycobacterium tuberculosis F11 ...
10.0	[1, 0, 0, 1]	CCGGCGGGGCCGGC	---	+++
10.0	[1, 0, 0, 1]	GCCGCCGTTGCCGC	---	+++
10.0	[1, 0, 0, 1]	CCGCCGTTGCCGCC	---	+++
10.0	[1, 0, 0, 1]	CCGCCGTTGCCGC	---	+++
10.0	[1, 0, 0, 1]	CCGCCATTGCC	---	+++
5.62	[0, 1, 0, 1]	GCAACGGCGGCA	+	+++
5.62	[0, 1, 0, 1]	CGCCGCGCCGCG	+++	+
5.62	[0, 1, 0, 1]	GCCGCCGCGCCC	+++	+
5.62	[0, 1, 0, 1]	CAACGGCGGCA	+	+++
5.62	[0, 1, 0, 1]	CGGCCGTCGCC	+++	+

Filter permutations: GCCAT / GCTAT — ☑ Value: 2
Filter wordshifts: - CCAT / GCCA - — ☑ Value: 1
Filter constituents: TGCCAT / - GCCA - — ☑ Value: 1
Thresholds: From: 8 To: 14 Score: 0 Top: 100 Redun: Sim:

Fig. 6. The top 10 most diverse words which separate *M. tuberculosis* from *M. avium*. The filter was set to reduce the redundancy of the selected words.

Next, the distance values D between an unknown sequence and the taxonomic units were calculated as follows:

$$D = 10 \times \sqrt{\frac{(p_i' - p_i)^2 \times \sum \left(1 + \frac{p_i' - p_i}{2 - p_i}\right)}{(m_i - p_i)^2 \times \sum \left(1 + \frac{m_i - p_i}{2 - p_i}\right)}} \qquad (14)$$

where p_i is the percentile value from the OligoDB database for the word i in a bacterial genome, or the average of the percentile values for the genomes of a taxonomic unit; p_i' is the percentile value of this word in the query sequence as calculated by equation 13; and m_i is an indicator variable equal to 0 if $p_i \geq 0.5$ and equal to 1 if $p_i < 0.5$. Thus, the denominator is the maximum possible distance. D values fall in the range from 0 to 10.

Consider the following example: let the 10-mer TTAAAGAAAA be distributed in the concatenated cluster sequence with the frequency 2.81 words per 100 Kbp and let the 8-mer TCTTTTAA occur 6.35 times per 100 Kbp. According to equation 13, the percentile value of the word TTAAAGAAAA is:

$$p = \frac{\ln(2.81) + 4.5\ln(10) - 9}{3} = .79 \qquad (15)$$

and for the word TCTTTTAA the percentile value is:

$$p = \frac{\ln(6.35) + 4.5\ln(8) - 9}{3} = .74 \qquad (16)$$

Next D is calculated by equation 14. The motivation for using D values rather than Euclidian distances is based on the fact that the observed frequency of occurrence of words in the clustered reads is frequently lower than in the original genome due to asymmetric distribution of word frequencies. Another factor that contributes to this observation is that the clusters of metagenomic DNA reads often contain fragments from more than one organism. This leads to false similarity of metagenomic sequences to taxa where the signature words are uncommon. To remove this bias, the equation 14 was constructed so that the difference between p_i' and p_i is given less weight if p_i' is smaller than p_i than if p_i' is larger than p_i.

To evaluate the discriminative power of the algorithms, several simulated metagenomic datasets were prepared using MetaSim. Then DNA reads were clustered by LikelyBin.

The first set was a simple random selection of 50 DNA fragments of the chromosome of *Bacillus subtilis* 168 [NC_000964]. The total length of all the fragments was 691 Kbp. The OligoDB program was used to compare the compositional similarity of these randomly selected sequences with the original chromosome and the closely related organisms of the genus *Bacillus* and class Firmicutes. The obtained distances are shown in Table 2.

Genome	D*
B. subtilis [NC_000964]	1.74
B. amyloliquefaciens [NC_009725]	2.22
B. licheniformis [NC_006270]	2.23
B. pumilus [NC_009848]	3.43
B. clausii [NC_006582]	3.70
Lactobacillus brevis [NC_008497]	3.83
B. halodurans [NC_002570]	4.05
B. pseudofirmus [NC_013791]	4.49
B. anthracis [NC_003997]	5.60
B. cereus [NC_004722]	5.68

*In this and the following tables the filter settings for the program were as follows: only the top 100 words of 8 to 12-mers with the sequence similarity ≤ 30 % were considered. Further, one nucleotide shifted words and one nucleotide shorter constituent words were filtered out as in the filter setting window in Fig. 6.

Table 2. Identification of DNA fragments generated from the *B. subtilis* chromosome.

For the next test, two quite distant organisms were selected: *Burkholderia cenocepacia* AU1054 [NC_008062] and *Psychrobacter arcticus* 273-4 [NC_007204]. 552 genomic fragments with an average length of 500 bp were generated randomly by MetaSim from the *B. cenocepacia* chromosome and 448 fragments of the same average length were obtained from the *P. arcticus* chromosome. All these fragments were mixed together and used as the input for LikelyBin. These randomly generated genomic fragments were then grouped by DNA composition similarity into 13 clusters. The two biggest clusters contained DNA fragments that were generated exclusively from one origin: 347 of the fragments generated from *B. cenocepacia* were grouped into cluster A and 437 of the fragments from the *P. arcticus* chromosome were in cluster B. Now, the OligoDB algorithm was used to identify the organisms most similar to the cluster. For the comparative analysis several representatives of β- and γ-Proteobacteria were selected (Table 3).

Cluster A (172 Kbp)	D	Cluster B (220 Kbp)	D
B. cenocepacia [NC_008062]	2.07	*P. arcticus* [NC_007204]	1.59
B.ambifaria [NC_010557]	3.05	*P. cryohalolentis* [NC_007969]	1.66
B. mallei [NC_006348]	3.51	*P. haloplanktis* [NC_007481]	1.92
B. phymatum [NC_010622]	6.69	*P. atlantica* [NC_008228]	2.25
B. xenovorans [NC_007952]	6.73	*P. ingrahamii* [NC_008709]	2.28
R. solanacearum [NC_003295]	7.67	*S. baltica* [NC_009052]	2.72
R. eutropha [NC_007347]	7.94	*S. enterica* [NC_003198]	7.06
C. metallidurans [NC_007974]	8.59	*E. pyrifoliae* [NC_012214]	7.72
P. arcticus [NC_007204]	8.59	*B. cenocepacia* [NC_008062]	8.67
R. pickettii [NC_010682]	8.66	*P. putida* [NC_002947]	8.90

Table 3. Identification of DNA fragments generated from *B. cenocepacia* (cluster A) and *P. arcticus* (cluster B).

Cluster A (87 Kbp)	D	Cluster B (99 Kbp)	D
P. haloplanktis [NC_007481]	3.00	*S. enterica* [NC_003198]	2.66
P. cryohalolentis [NC_007969]	3.06	*E. pyrifoliae* [NC_012214]	3.21
P. ingrahamii [NC_008709]	3.06	*P. putida* [NC_002947]	3.50
P. mirabilis [NC_010554]	3.10	*S. baltica* [NC_009052]	7.00
P. arcticus [NC_007204]	3.29	*P. atlantica* [NC_008228]	7.57
P. atlantica [NC_008228]	4.25	*P. arcticus* [NC_007204]	7.92
S. baltica [NC_009052]	4.80	*P. cryohalolentis* [NC_007969]	7.99
S. enterica [NC_003198]	7.49	*P. ingrahamii* [NC_008709]	8.03
E. pyrifoliae [NC_012214]	7.57	*P. mirabilis* [NC_010554]	8.09
P. putida [NC_002947]	8.04	*P. haloplanktis* [NC_007481]	8.17

Table 4. Identification of a chimerical cluster A that contains DNA fragments from *P. haloplanktis* and *S. enterica*, and a monophyletic cluster B containing fragments of the *S. enterica* genome.

The clusters were identified correctly; however, the separation of genomic fragments of *P. arcticus* (Cluster B) from other close relative organisms of genera *Psychrobacter*, *Psychromonas* and *Pseudoalteromonas* was not reliable. An additional round of identification is needed where the signature words are selected specifically to distinguish between these organisms.

The next set of DNA fragments was generated from two genomes of γ-Proteobacteria: *Pseudoalteromonas haloplanktis* TAC125 [NC_007481] and *Salmonella enterica* CT18 [NC_003198]. LikelyBin clustered the fragments into 49 clusters. Half of the clusters contained sequences generated from both chromosomes. The two biggest clusters were selected for analysis by the OligoDB algorithm. Cluster A contains 166 fragments of the *P. haloplanktis* chromosome and 9 sequences originating from *S. enterica*. Cluster B contains 195 DNA fragments generated from *S. enterica* only. Results of the identification are shown in Table 4.

Both of these clusters were identified correctly. The mix of two organisms in Cluster A yields D values that are higher than all other examples in this paper. D values calculated for more complex chimerical clusters were around 5; indicating that it will be difficult to associate such a set of sequences with a specific taxonomic unit.

2.4 Stratigraphic analysis of bacterial genomes

DNA molecules encoding functional enzymes, transcriptional regulators and virulence factors are fluxing through the bacterial taxonomic walls. They endow environmental and clinical strains of bacteria with new unexpected properties. Lateral genetic exchange, particularly of drug tolerance genes has been recognized for a long time; however the ontology of genomic islands and their donor-recipient relations remain generally obscure because of methodological problems. Horizontally transferred genes are highly mutable and the mobilome entities having been inserted into host chromosomes undergo multiple events of fragmentation, partial duplications and deletions. Even prediction of insertion sites in host chromosomes remained to be a challenge.

Genome linguistics methods are applicable to study and visualize intrinsic relationships between mobile genetic elements in bacterial genomes. *Mycobacterium tuberculosis*, a bacterial pathogen which is a leading cause of human death worldwide, was selected as a subject for this study. Emergence and evolution of this deadly pathogen are still ambiguous and not fully understood even after having done the sequencing and comparative studies on multiple strains of this genus.

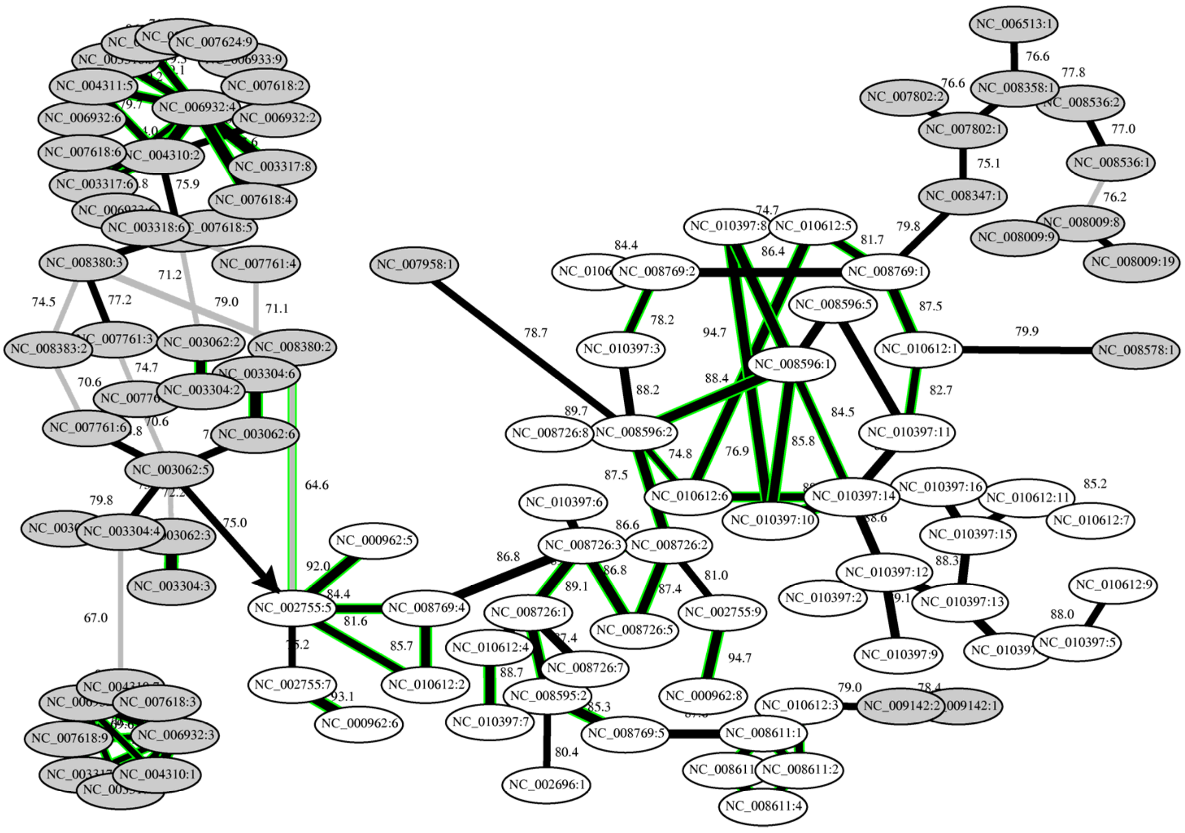


Fig. 7. GIs identified in *Mycobacterium* genomes and other organisms share compositional similarity. GIs identified in Mycobacteria are represented by white nodes and species of other genera by grey nodes. Each node represents one GI tagged by NC number of the host organism as in NCBI followed by the reference number of GIs as in GEI-DB. The edges depicted by green halo link GIs sharing similar DNA sequences longer than 100 bp identified by blast2seq. The layout was created by an in-house Python program that incorporates executable files of Graphviz 2.26.3 for Windows.

2.4.1 Identification and grouping of mycobacterial genomic islands

Linguistic methods were applied to study the distribution of genomic islands (GIs) in complete genome sequences of *Mycobacterium*. GIs were identified by SeqWord Gene Island Sniffer (SWGIS available at www.bi.up.ac.za/SeqWord/sniffer/). The identified GIs were grouped by compositional similarity of oligonucleotide usage (OU) patterns (Fig. 7). They were further pair-wise compared by blast2seq and the proteins encoded by GIs' genes were searched by BLASTp through the local databases of bacterial, plasmid and phage proteins. The latter analysis was performed to check if the GIs that cluster together share syntenic genes and to also deduce the types of genes that are most frequently transferred horizontally across species and genus borders.

In genomes of virulent and environmental *Mycobacterium* multiple genomic islands were identified which share both sequence and OU similarity (Fig. 7). An exception is *M. leprae* which genomic islands were unrelated to GIs of other Mycobacteria (data not shown but check http://anjie.bi.up.ac.za/geidb/geidb-home.php). In Fig. 7 GIs identified in *M. tuberculosis*, *M. bovis*, *M. marinum*, *M. vanbaalenii*, *M. abscessus* and *M. smegmatis* are represented by white nodes and those of species of other genera by grey nodes. Each node represents one GI tagged by NC accession number of the host organism as in NCBI followed by colons and reference numbers of GIs as in GEI-DB (http://anjie.bi.up.ac.za/geidb/geidb-home.php). Furthermore, six GIs identified in *M. tuberculosis*, *M. bovis* and *M. marinum* (framed in Fig. 2) share similarity in both DNA sequence and OU with GIs distributed among α-Proteobacteria, particularly to those of *Rhizobium* and *Agrobacterium*.

2.4.2 Stratigraphic analysis of genomic inserts

To determine the relative time of GI insertions, the similarity in OU patterns of GIs and corresponding host chromosomes was calculated for all organisms. The results are depicted by grey gradient colors in Fig. 8. GIs that significantly deviate from their hosts (recent inserts) are shown dark grey; and those that already underwent genomic amelioration (Lawrence & Ochman, 1997) are shown light grey. Most mycobacterial GIs revealed to be ancient inserts that is in consistence with the fact that they are shared by different species. Few of the GIs that showed to be in possession of OU patterns similar to GIs of *Rhizobium* and *Agrobacterium* are relatively recent acquisitions. Comparison of the patterns of the GIs and host genomes was revised in order to determine donor-recipient relationships between these organisms (Fig. 9). The analysis revealed that these mycobacterial GIs are compositionally more similar to the chromosomes and mobilomes of *Agrobacterium* and that they are most likely originated from this source as indicated in Fig. 9.

43 Mycobacterial GIs (unframed in Fig. 2) contain 910 annotated genes among which 386 were hypothetical or unknown. Functional genes are listed in Table 5. Predominance of phage related genes suggests that these GIs are mostly prophages. Genes that are harboured by the GIs of the α-Proteobacteria origin (framed in Fig. 2) encode several transferases, esterases, mmcH proteins and hypothetical proteins organized into operon structures (Fig. 10), which may be involved in the biosynthesis of some yet unknown compounds. Shaded areas in Fig. 10 link regions sharing DNA sequence similarity determined by blast2seq. The compared genomes are NC_000962 (*M. tuberculosis* H37Rv); NC_002755 (*M. tuberculosis* CDC1551); NC_008769 (*M. bovis* BCG str. Pasteur 1173P2); and NC_010612 (*M. marinum* M). Lengths of GIs are also indicated.

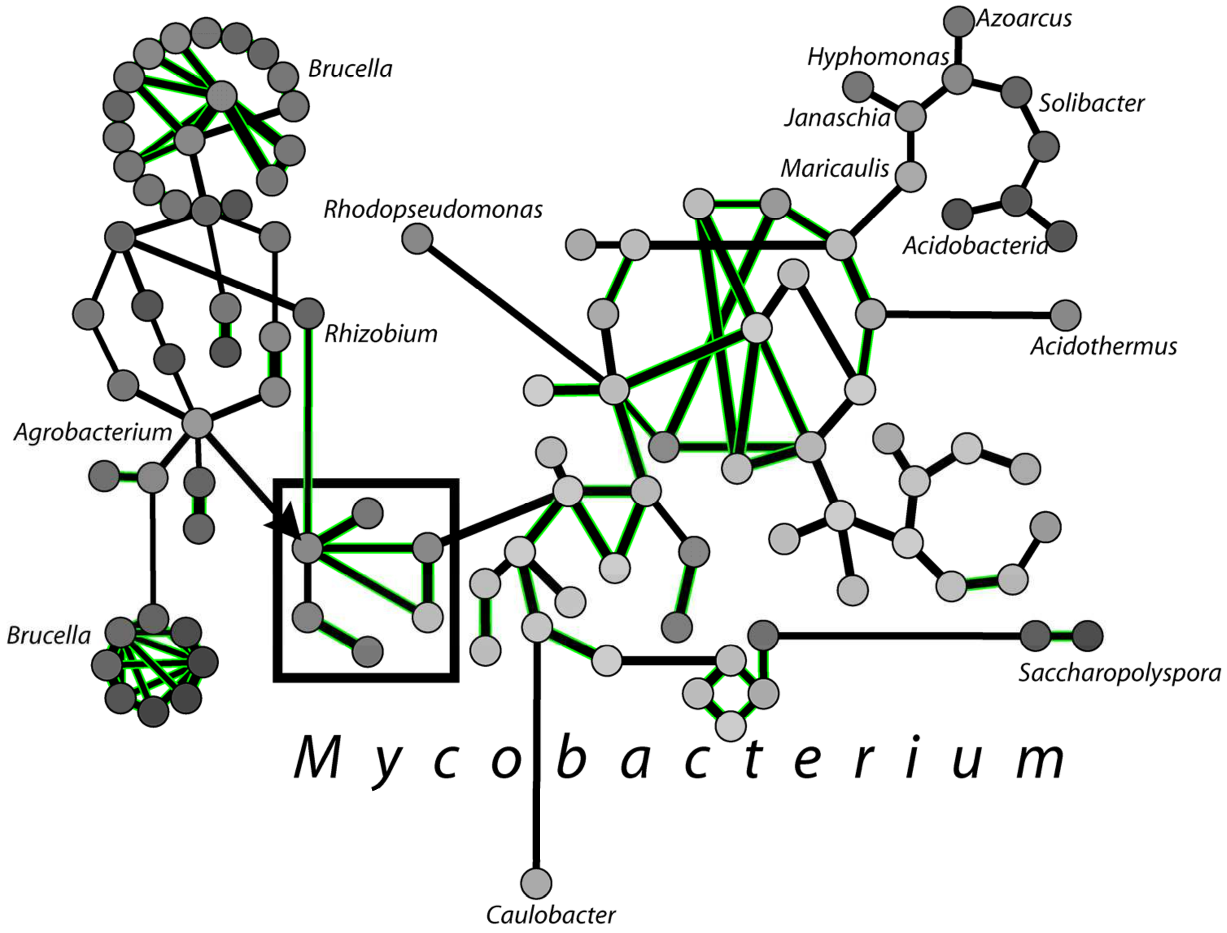


Fig. 8. Stratigraphic analysis of GIs. The edges depicted by green halo link GIs sharing similar DNA sequences longer than 100 bp identified by blast2seq. The layout of nodes is the same as in Fig. 7.

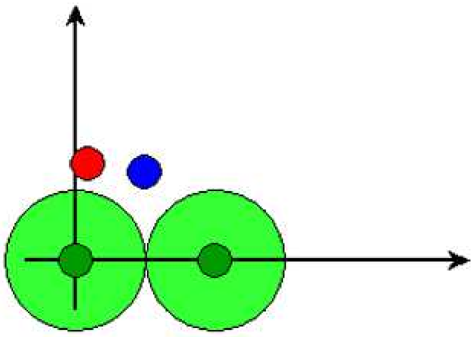


Fig. 9. Donor-recipient relationships between GIs and host organisms of *Agrobacterium* and *Mycobacterium*. Dark green circles indicate OU patterns of the host organisms. Light green shaded areas represent half-distances between chromosomal OU patterns. OU patterns of genomic islands of *M. tuberculosis* NC_002755 and *A. tumefaciens* NC_003062 (blue and red circles respectively) were plotted according to the calculated distances between them and OU patterns of the chromosomes. Plotting was done by an in-house Python program.

Gene categories	Number of genes
Phage related proteins, integrases and transposases	91
Dehydrogenases	31
Transcriptional regulator	23
Peptide synthetase and polyketide	13
Membrane proteins	23
Monooxygenase	11
Glycosyl transferases	11
Oxidoreductase	10
Dioxygenase	9
PE-PGRS proteins	7
Esterases	5

Table 5. Proteins encoded by genes in ancient GIs of *Mycobacterium*.

Protein BLAST analysis of Mycobacterial GIs retrieved similarities in proteins shared with a great variety of bacterial plasmids and phages, particularly in the plasmid pSOL1 from *Clostridium acetobutylicum* ATCC 824. Acquisition of genetic materials from intracellular parasitic and symbiotic species of α-Proteobacteria by an ancestral strain of *Mycobacterium* may be an event that had triggered the evolution of former saprophytic organisms towards the parasitic lifestyle.

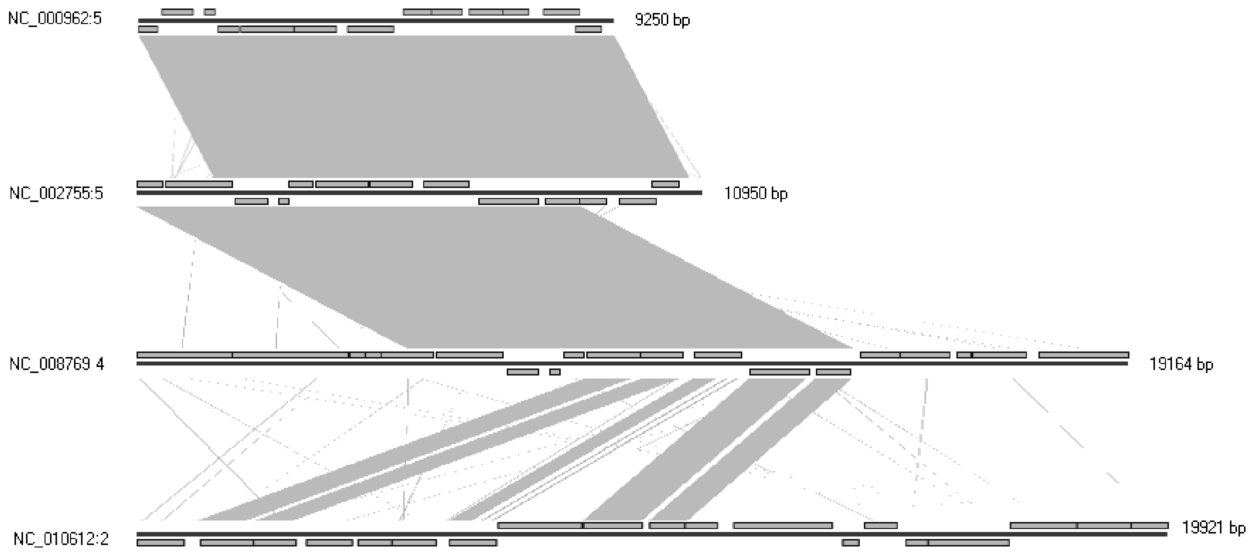


Fig. 10. Homologous genes and operons in GIs shared by *Mycobacterium*. GIs are referred by NC number of the host organism in the NCBI database followed by the reference number of GIs in GEI-DB.

2.4.3 Overview of the horizontal gene transfer in the bacterial world

The exchange of genetic material was found to have occurred in different domains of life: Archaea, Bacteria, and Eukarya (Choi and Kim, 2007). Horizontal gene transfer, defined as a mechanism that promotes the transfer of foreign genomic segments between lineages was found to be relatively common in prokaryotes and less common in higher-order organisms. The transfer of operational genes is a continual process and is far more important in prokaryotic diversity of different sources (Jain et al., 1999; Ochman, 2000). For horizontal gene transfer to become a success, the acquisition of foreign DNA segments must be counterbalanced by DNA loss. Acquired DNA providing functions that are beneficial to the host may be maintained, while DNA providing less beneficial functions may be lost (Lawrence, 1999). Mobile genetic elements possess genes that contribute to bacterial speciation and adaptation to different niches, but also carry with them factors that contribute to the bacteria's fitness traits, secondary metabolism, antibiotic resistance and symbiotic interactions (Dobrindt et al., 2004; Mantri & Williams, 2004) that are of medical and agricultural importance.

The transfer of GIs occurs through three mechanisms: transformation, conjugation and transduction. These mechanisms mediate the movement and transfer of DNA segments intercellularly. Conjugation and transduction are the common players in genetic transfer. They require mobile elements such as plasmids and bacteriophages to transfer genetic elements along with the sequence features of their donor to recipient cells (Hacker & Carniel, 2001). Upon transfer, these genetic elements get established into the recipient cell either as self replicating elements or by getting integrated into the chromosome either by homologous or illegitimate recombination techniques (Dutta and Pan, 2002; Beiko et al., 2005). Transformation, unlike conjugation and transduction does not require any form of a vector to transport genomic elements between bacteria. It is mediated by the uptake of a naked DNA in the environment. The uptake usually takes place upon the release of DNA from decomposing and disrupted cell, or viral particles, or even excretions from living cells (Thomas & Nielsen, 2005).

DNA composition comparisons between lineages have uncovered that genes acquired by the above mechanisms display features that are distinct from those of their recipient genomes (Hacker and Carniel, 2001; van Passel et al., 2006). Genes acquired by horizontal transfer can often display atypical sequence characteristics and a restricted phylogenetic distribution among related strains, thereby producing a scattered phylogenetic distribution (Ochman et al., 2000; Dutta and Pan, 2002). Bacterial species are variable in their overall GC content but the genes in genomes of particular species are fairly uniform with respect to their base composition patterns and frequencies of oligonucleotides (Ochman et al., 2000). The phylogenetic aspect of similarity in base composition among closely related species arises from their common origin. Similarity is also influenced by genome specific mutational pressures that act upon their genes to promote the maintenance of composition stability. Native or core genes in a given organism exhibit homogeneous OU content and codon usage, while foreign genes display atypical characteristic features shared with their mobilomes (phages and conjugative plasmids) or previous host organisms for the genetic segments which were mobilized and integrated by mobilomes (Davenport et al., 2009). Compositional specificity of GIs allows their precise identification by the SWGIS program (see above in this chapter). In this work SWGIS was used to search each prokaryotic genome for foreign inserts based on the comparisons of tetranucleotide usage patterns, whereby the frequencies of particular tetramers are compared with expected occurrences of the same tetramers throughout the whole genome. Identified GIs were stored to GEI-DB (http://anjie.bi.up.ac.za/geidb/geidb-home.php) that contains a set of 3518 precalculated GIs identified in 637 prokaryotic genomes. All these GIs were clustered by the compositional OU pattern similarity that is believed to represent their common ancestry. Similarity between GIs was calculated as 100 – D(%), where D(%) was found by the equation 4. GIs which share more than 75% of similarity were grouped together. Groups of GIs and their distribution among bacteria are shown in Fig. 11.

GIs were identified in all bacterial classes. There are more GIs from *E. coli* and other Enterobacteria and γ-Proteobacteria that partially may be explained by a biased overrepresentation of these microorganisms among other sequenced genomes in the GenBank database. *E. coli*, *Shigella* and *Salmonella* share GIs of one common origin but GIs found in other species often showed to have originated from several different origins. For example, GIs from *Pseudomonas* form several separate clusters associated with either other γ-Proteobacteria or α-Proteobacteria. GIs of α-Proteobacteria and Firmicutes show extreme diversity. *Brucella*, *Agrobacterium* and *Rhizobium* share several unrelated pools of their mobilomes. Relations which were found between GIs of *Mycobacterium* and those of *Agrobacterium* and *Rhizobium* have been discussed above in detail. GIs of *Prochlorococcus* and *Nostoc* cyanobacteria most likely originated from marine γ-Proteobacteria, but GIs of *Synechococcus* are very specific and share no similarity with any other microorganisms.

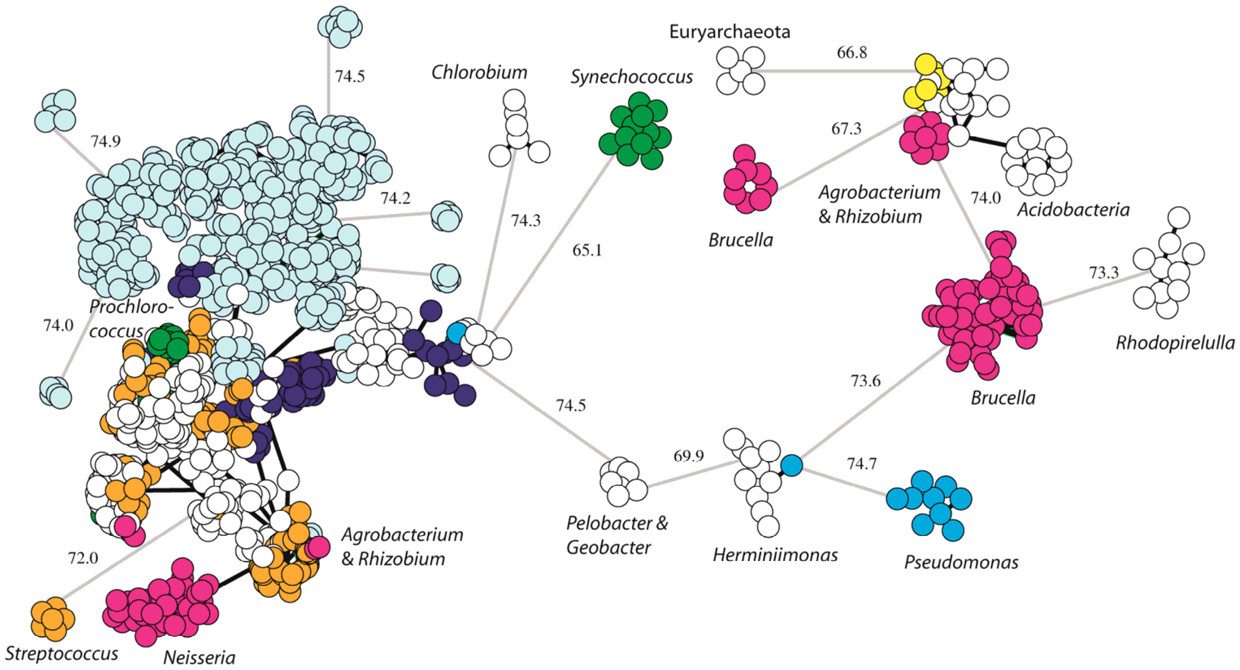


Fig. 11. Groups of GIs joined by compositional OU pattern similarity. Each node represents one GI. Genera of γ-Proteobacteria are shown in light blue (enterobacteria *Escherichia*, *Shigella* and *Salmonella*), cyan (*Pseudomonas*) and dark blue (marine bacteria *Shewanella*, *Hahella*, *Pseudoalteromonas* and *Alcanivorax*); α-Proteobacteria *Agrobacterium*, *Brucella*, *Neisseria*, *Rhizobium* and *Synorhizobium* are depicted by magenta nodes; Firmicutes (*Bacillus*, *Clostridium*, *Geobacillus* and *Streptococcus*) – orange; Actinobacteria (*Corynebacterium* and *Mycobacterium*) – yellow; Cyanobacteria (*Prochlorococcus*, *Nostoc* and *Synechococcus*) – green. Nodes representing other organisms are white. Black edges link nodes which share strong OU similarity above 75% and grey edges represent weaker similarity below 75%.

It may be concluded that GIs indeed may flux through the bacterial taxonomic walls but not in a random fashion. Several species and genera share pools of horizontally transferred genetic elements, which include pathogenicity, antibiotic resistance, O-antigen synthesis and catabolic GIs, whereas the genetic exchange between other groups of microorganisms is seemed very unlikely. Detailed analysis of gene exchange pathways among microorganisms will shed light on the roles played by the horizontal gene transfer in the evolution and pathogenicity of bacteria.

26732 proteins encoded by GIs' genes used in this study were pair-wise compared by BLASTp. The bit-score results were used to produce clusters of proteins by Markov clustering algorithm (MCL) (Vlasblom & Wodak, 2009). MCL with an inflation parameter of 1.8 produced 10837 clusters, however, many of them were of a single hypothetical protein. Due to the large amount of hypothetical and unknown genes in the database not all of these clusters would present biologically significant data. Top 24 clusters containing more than 50 proteins were chosen as significant to represent categories of proteins which are most often mobilized and transferred horizontally among bacteria. Besides phage related proteins which are in a majority, the most frequently bacteria acquire ABC-transporters, transcriptional regulators including GGDEF diguanylates, polysaccharide and O-antigen biosynthesis proteins, dehydrogenases and outer membrane proteins (Table 6).

Functional group	Nr of proteins identified in 3518 GIs
Phage related proteins, IS-elements, transposases;	792
Transcriptional regulators;	599
Polysaccharide and O-antigen biosynthesis proteins;	352
ABC-transporter;	252
Outer membrane proteins;	241
Dehydrogenases;	67
RHS-family proteins;	64

Table 6. Predominant categories of horizontally transferred proteins.

3. Conclusion

Comparative genomics exploits the methods of two major categories based on the analysis of composition and sequence similarity. Having been developed at the beginning of the genomics era, sequence similarity comparison by BLAST (Altschul, 1990) and FASTA (Pearson, 1995), and sequence composition simulation by Markov Chain Models (Schbath, 2000) remain the algorithms of first choice. The algorithms for sequence similarity comparison are widely used because of speed, more straightforward statistics and a clearer biological relevance of sequence alignment based considerations. However, a number of practical tools based on OU statistics have become publicly available. Several novel OU analytical tools of the SeqWord project for genome visualization, genomic island detection and identification of unknown sequences have been presented in this chapter.

Composition based methods are termed genome linguistics as they deal with frequencies of words written as chains of given alphabets of nucleotides or amino acids of variable lengths. Genome linguistic approaches may complement or even outperform the sequence similarity comparison in clustering DNA reads (Kislyuk et al., 2009) and detecting inserts of genomic islands (Hsiao et al., 2003). These approaches have also been shown to be instrumental in viral metagenomics (Delwart, 2007). During composition based analysis, longer DNA sequences are rather preferred to shorter ones for the word distribution statistics to be reliable.

DNA similarity vanishes much faster in phylogenetically distant organisms than the OU composition does, especially in highly variable virus, phage, plasmids and genomic islands. Protein similarity may mislead binning or identification of unknown sequences for it mostly reflects the functional conservation of protein domains rather than the taxonomic unity. Another common limitation of the similarity based methods is that the sequence identification is possible only if a homologous DNA or protein sequence is present in the searched database. On the contrary, the genome specific OU pattern is a pervasive property of the whole genome (Jernigan & Baran, 2002) that allows binning of DNA reads to their putative origin even if they do not share any significant sequence similarity.

The advancement in genome sequencing technologies made large scale sequencing affordable for many laboratories. An attractive approach of alignment-independent phylogenetic studies based on the comparison of OU patterns was discussed in several publications and a number of web-based services were proposed (Chapus et al., 2005). We suggest rather a cautious use of these methods as a significant convergence of OU patterns was observed between unrelated organisms. For instance, *Pseudomonas* and *Mycobacterium* share similar OU patterns. Furthermore, a wider application of OU patterns is hindered by the absence of any noteworthy mathematical models simulating the evolutionary changes in OU patterns between organisms in contrast to sequence similarity methods which provide plenty of models of nucleotide and amino acid substitutions. Development and testing of such models is the task that urgently needs to be looked into to advance applicability of genome linguistic approaches.

4. Acknowledgment

Funding for this research was provided by the MetaLingvo grant 71261 of the National Research Foundation (NRF) of South Africa.

5. References

Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. & Ikemura, T. (2003). Informatics for unveiling hidden genome signatures. *Genome Res.*, Vol. 13, No. 4, pp. 693-702.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, Vol. 215, No. 3, pp. 403–410.

Azad, R. K. & Lawrence, J. G. (2005). Use of artificial genomes in assessing methods for atypical gene detection. *PLoS Comput. Biol.*, Vol. 1, No. 6, p. e56.

Baldi, P. & Baisnée, P.-F. (2000). Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics*, Vol. 16, No. 10, pp. 865-889.

Becq, J., Gutierrez, M. C., Rosas-Magallanes, V., Rauzier, J., Gicquel, B., Neyrolles, O. & Deschavanne, P. (2007). Contribution of horizontally acquired genomic islands to the evolution of tubercle bacilli. *Mol. Biol. Evol*. 2007, Vol. 24, No. 8, pp. 1861-1871.

Beiko, R. G., Harlow, T. J. & Ragan, M. A. (2005). Highways of gene sharing in prokaryotes. *Proc. Natl. Acad. Sci. USA*, Vol. 102, No. 40, pp. 14332–14337.

Bohlin, J., Skjerve, E. & Ussery, D. (2008). Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes. *BMC Genomics*, Vol. 9, p. 104.

Chatterji, S., Yamazaki, I., Bai, Z. & Eisen J. A. (2008). CompostBin: a DNA composition based algorithm for binning environmental shortgun reads, In: *RECOMB*, Vingron, M. & Wong, L., pp. 17-28, LNBI 4955.

Coenye, T. & Vandamme, P. (2004). Use of the genomic signatures in bacterial classification and identification. *System. Appl. Microbiol.*, Vol. 27, No. 2, pp. 175-185.

Chapus, C., Dufraigne, C., Edwards, S., Giron, A., Fertil, B. & Deschavanne, P. (2005). Exploration of phylogenetic data using a global sequence analysis method. Vol. 9, No. 5, p. 63.

Choi, I.-G. & Kim, S.-H. (2007). Global extent of horizontal gene transfer. *Proc. Natl. Acad. Sci. USA* Vol. 104, No. 11, pp. 4489–4494.

Davenport, C. F., Wiehlmann, L., Reva, O. N. & Tümmler, B. (2009). Visualization of *Pseudomonas* genomic structure by abundant 8-14mer oligonucleotides. *Environ. Microbiol.*, Vol 11, No. 5, pp. 1092-1104.

Delwart, E. L. (2007). Viral metagenomics. *Rev. Med. Virol.* Vol. 17, No. 2, pp. 115-131.

Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G. & Fertil, B. (1999). Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.*, Vol. 16, No. 10, pp. 1391-1399.

Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms *Nat. Rev. Microbiol.* Vol. 2, No. 5, pp. 414–424.

Dufraigne, C., Fertil, B., Lespinats, S., Giron, A. & Deschavanne, P. (2005). Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic. Acids Res.*, Vol. 33, No. 1, p. e6.

Dutta, C. & Pan, A. (2002). Horizontal gene transfer and bacterial diversity. *J. Biosci.* Vol. 27, No. 1 Suppl. 1, pp. 27-33.

Hacker, J. & Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep.*, Vol. 2, No. 5, pp. 376–381.

Hsiao, W., Wan, I., Jones, S. J. & Brinkman, F. S. L. (2003). IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*, Vol. 19, No. 3, pp. 418-420.

Jain, R., Rivera, M. C. & Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. USA*, Vol. 96, No. 7, pp. 3801–3806.

Jernigan, R. W. & Baran, R. H. (2002). Pervasive properties of the genomic signature. *BMC Genomics*, Vol. 3, No. 1, p. 23.

Karlin, S. & Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature, *Trends Genet.*, Vol. 11, No. 7, pp. 283-290.

Karlin, S. (1998). Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.*, Vol. 1, No. 5, pp. 598-610.

Karlin, S., Mrázek, J. & Campbell, A. (1997). Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.*, Vol. 179, No. 12, pp. 3899-3913.

Kirzhner, V., Bolshoy, A., Volkovich, Z., Korol A. & Nevo E. (2005). Large-scale genome clustering across life based on a linguistic approach. *BioSystems*, Vol. 81, No. 3, pp. 208-222.

Kislyuk, A., Bhatnagar, S., Dushoff, J. & Weitz J. S. (2009). Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinformatics*, Vol. 10, p. 316.

Koski, L. B., Morton, R. A. & Golding, G. B. (2001). Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol. Biol. Evol.*, Vol. 18, No. 3, pp. 404-412.

Lawrence, J. G. & Ochman H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.*, Vol. 44, No. 4, pp. 383-397.

Lawrence, J. G. (1999). Gene transfer, speciation, and the evolution of bacterial genomes. *Curr. Opin. Microbiol.* Vol. 2, No. 5, pp. 519–523.

Mantri, Y. & Williams, K. P. (2004). Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res.* Vol. 32, Database issue, pp. D55–D58.

Mrázek, J. & Karlin, S. (1999). Detecting alien genes in bacterial genomes. *Ann. NY Acad. Sci.*, Vol. 870, pp. 314-329.

Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat. Genet. Vol.* 36, No. 7, pp. 760-766.

Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, Vol. 405, No. 6784, pp. 299–304.

Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Sci.*, Vol. 4, No. 6, pp. 1145-1160.

Pride, D. T. & Blaser, M. J. (2002). Identification of horizontally acquired elements in *Helicobacter pylori* and other prokaryotes using oligonucleotide difference analysis. *Genome Let.*, Vol. 1, No. 1, pp. 2–15.

Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. (2003). Evolutionary implications of microbial genome tetanucleotide frequency biases. *Genome Res.*, Vol. 13, No. 2, pp. 145–155.

Reva, O. N. & Tümmler, B. (2004). Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics*, Vol. 5, p. 90.

Richter, D. C., Ott, F., Auch, A. F., Schmid, R. & Huson D. H. (2008). MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One*, Vol. 3, No. 10, p. e3373.

Saeed, I. & Halgamuge, K. (2009). The oligonucleotide frequency derived error gradient and its application to the binning of metagenome fragments. *BMC Genomics*, Vol. 10, Suppl. 3, p. S10.

Schbath, S. (2000). An overview on the distribution of word counts in Markov chains. *J. Comp. Biol.*, Vol. 7, No. ½, pp. 193-201.

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glöckner, F. O. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.*, Vol. 6, No. 9, pp. 938-947.

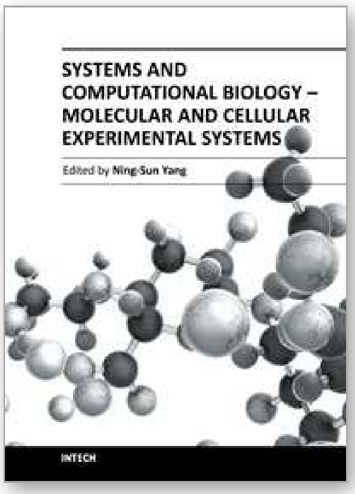
Thomas, C. M. & Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.*, Vol. 3, No. 9, pp. 711–721.

van Passel, M. W., Bart, A., Luyf, A. C., van Kampen, A. H. & van der Ende, A. (2006). The reach of the genome signature in prokaryotes. *BMC Evol. Biol.*, Vol. 6, p. 84.

Vlasblom, J. & Wodak, S. J. (2009). Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics*, Vol. 10, No. 99, p. 1.

Wang, B. (2001). Limitations of compositional approach to identify horizontally transferred genes. *J. Mol. Evol.*, Vol. 53, No. 3, pp. 244-250.

Weinel, C., Ussery, D. W., Ohlsson, H., Sicheritz-Ponten, T., Kiewitz, C. & Tümmler, B. (2002). Comparative genomics of *Pseudomonas aeruginosa* PAO1 and *Pseudomonas putida* KT2440: orthologs, codon usage, repetitive extragenic palindromic elements, and oligonucleotide motif signatures. *Genome Lett.*, Vol. 1, No. 4, pp. 175-187.


SYSTEMS AND COMPUTATIONAL BIOLOGY – MOLECULAR AND CELLULAR EXPERIMENTAL SYSTEMS
Edited by Ning-Sun Yang
INTECH

Systems and Computational Biology - Molecular and Cellular Experimental Systems
Edited by Prof. Ning-Sun Yang

ISBN 978-953-307-280-7
Hard cover, 332 pages
Publisher InTech
Published online 15, September, 2011
Published in print edition September, 2011

Whereas some â€œmicroarrayâ€ or â€œbioinformaticsâ€ scientists among us may have been criticized as doing â€œcataloging researchâ€, the majority of us believe that we are sincerely exploring new scientific and technological systems to benefit human health, human food and animal feed production, and environmental protections. Indeed, we are humbled by the complexity, extent and beauty of cross-talks in various biological systems; on the other hand, we are becoming more educated and are able to start addressing honestly and skillfully the various important issues concerning translational medicine, global agriculture, and the environment. The two volumes of this book presents a series of high-quality research or review articles in a timely fashion to this emerging research field of our scientific community.

Oliver Bezuidt, Hamilton Ganesan, Phillip Labuschange, Warren Emmett, Rian Pierneef and Oleg N. Reva (2011). Linguistic Approaches for Annotation, Visualization and Comparison of Prokaryotic Genomes and Environmental Sequences, Systems and Computational Biology - Molecular and Cellular Experimental Systems, Prof. Ning-Sun Yang (Ed.), ISBN: 978-953-307-280-7, InTech, Available from: http://www.intechopen.com/books/systems-and-computational-biology-molecular-and-cellular-experimental-systems/linguistic-approaches-for-annotation-visualization-and-comparison-of-prokaryotic-genomes-and-environ

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821