View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Genomics 95 (2010) 7–15

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Review

Bacterial genomic G + C composition-eliciting environmental adaptation

Scott Mann [a], Yi-Ping Phoebe Chen [a,b,*]

[a] *Faculty of Science and Technology, Deakin University, Australia*
[b] *ARC Centre of Excellence in Bioinformatics, Australia*

ARTICLE INFO

Article history:
Received 13 May 2009
Accepted 1 September 2009
Available online 9 September 2009

Keywords:
Genome profiling
G + C content profiling
Bacteria
Pathogenicity
Genome reduction

ABSTRACT

Bacterial genomes reflect their adaptation strategies through nucleotide usage trends found in their chromosome composition. Bacteria, unlike eukaryotes contain a wide range of genomic G + C. This wide variability may be viewed as a response to environmental adaptation. Two overarching trends are observed across bacterial genomes, the first, correlates genomic G + C to environmental niches and lifestyle, while the other utilizees intra-genomic G + C incongruence to delineate horizontally transferred material. In this review, we focus on the influence of several properties including biochemical, genetic flows, selection biases, and the biochemical-energetic properties shaping genome composition. Outcomes indicate a trend toward high G + C and larger genomes in free-living organisms, as a result of more complex and varied environments (higher chance for horizontal gene transfer). Conversely, nutrient limiting and nutrient poor environments dictate smaller genomes of low GC in attempts to conserve replication expense. Varied processes including translesion repair mechanisms, phage insertion and cytosine degradation has been shown to introduce higher AT in genomic sequences. We conclude the review with an analysis of current bioinformatics tools seeking to elicit compositional variances and highlight the practical implications when using such techniques.

© 2009 Elsevier Inc. All rights reserved.

Contents

Introduction

G + C content describes the guanine and cytosine content of a biological sequence and has historically been reported to range between 25% and 75% for bacterial genomes [1] and more recently as low as 20% in the Carsonella genome [2]. The wide G + C variation of bacterial genomes poses many interesting questions for the researcher. Such questions center on the role evolution plays toward shaping genome content and the mechanisms that abruptly alter such processes. Selectionist theory implies that changes in G + C content over evolutionary time are in response to environmental conditions in order to confer advantage. For example, free-living bacteria have average G + C content higher than obligatory pathogens and symbionts, as evidenced across several taxonomic branches [3]. In this review we explore the implications of lifestyle on genomic G + C content and the selection for energetic, genetic and

[*] Corresponding author. Faculty of Science and Technology, Deakin University, Australia.
E-mail address: phoebe.chen@deakin.edu.au (Y.-P.P. Chen).

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.09.002

biochemical properties toward the characteristics we see in today's sequenced bacterial genomes. We end with a brief overview of the computational tools serving to elicit G + C based information.

Biological context—energetic

Energy considerations in the context of genomic G + C composition center upon the cost of nucleotide synthesis and maintenance. As shown by [3], selection against G + C is attributed to the increased energy cost in GTP and CTP synthesis, consequently, the mutational bias toward A/T may therefore confer selective advantage and impact coding sequence length. The unavailability under stress of cytosine, suggests short coding sequences with longer sequences devoid of cytosine to be a feature of prokaryote genomes [4]. Such outcomes in the context of well-established correlations between genome length and G + C content [5–7] raise somewhat conflicting outcomes. The overall trend witnessed in larger genomes e.g. >3 Mb, show relative G + C content greater than the smaller genomes. Data plots of G + C vs. genome size markedly show this trend [8]. Given the clear shift toward low G + C genomes in bacterial symbionts and other organisms occupying nutrient limiting environments, energy conservation could indeed be seen as a property to be selected for, above neutral fixation. Distinction between coding sequence and whole genome composition cannot be drawn for bacterial genomes, which compose mostly coding sequence.

Biological context—genetic

Acquisition of "non-self" genetic material can be seen as a mechanism for diversification and environmental adaptation. The heterogeneous environments of free-living organisms potentially impact their genomic G + C content via lateral DNA transfer [9,10]. Several underlying processes contribute toward lateral transfers, including transduction, transformation and conjugation of bacterial DNA. Transduction is the process by which genetic material is transferred to a host via bacteriophage invasion and replication [11]. Two forms of transduction commonly referred to in the literature include "generalized" and "specialized" transduction. Generalized transduction involves integration of non-specific sequences from the donor, while in specialized transduction, regions bordering phage attachment sequences are incorporated by the host. The length of genetic material and ultimately the presence of phage DNA in the host is dependent upon the virus recognizing the bacterium and the payload capacity (content of genetic material) of the viral capsid [12]. Conjugation, the third method for lateral transfer, involves direct contact between cells of donor and recipient, transmission of material is typically facilitated by plasmids. Several recognizable features in the literature are associated with at least one of these three processes, namely prophages, pathogenicity islands, integrons and insertion sequences. The trend observed upon integration (of the above elements) into the recipient genome, is the abrupt differential G+C content of the introduced sequence vs. the background recipient genome [13]. Caveats exist however and center around amelioration, the process by which insert DNA is modified to match the background nucleotide usage of the recipient genome. The potential effect of amelioration-based obfuscation increases with the age of the insert. Of considerable interest are pathogenicity islands (PAI), tracts of DNA conferring specific pathogenic potential to the host genome. The "island" represents a cluster of genes contributing pathogenic potential to the recipient either via integration into the chromosome or plasmid DNA [14]. Typical presentations arise as either integrated plasmids, bacteriophages or conjugative transposons, or part thereof [15]. Methods for the detection of PAIs and genomic islands (GI's) in prokaryote genomes have been proposed and studied by many research works [16–24], where differential G + C content has been used as a significant indicator of non-self genomic content. To enable survival in variable environmental niches, the acquisition of such abilities has led to colonization of environments not typically associated with other strains of the same species [15]. The contribution of horizontal transfer in *Escherichia coli* has been estimated to be 10–50 times greater than vertical processes for a given single nucleotide difference [25]. The commonality between *Escherichia* genomes extends to their G + C genomic content, between 50.4% and 50.8%, a very narrow range. With G + C genomic content of ~50% and chromosome sizes ranging from 4639675 bp (K12) to 5528445 bp (O157:H7 EDL), the *Escherichia* represent organisms that have adapted to free-living environments, including intra and extra body, soil and aquatic niches [26]. The *E. coli* K12 genome has been reported to contain 12.8% foreign DNA [12]. There exist several excellent reviews of pathogenicity islands and their role in genome evolution [15,27] and references therein. PAIs as suggested in the literature, represent cases of macro and micro scale evolution, their enhanced fitness to niche environments rather than their detrimental effects to their hosts are selected for in the genome [15]. The acquisition of cell surface binding proteins is one such example found in the uropathogenic *E. coli* (UPEC) organisms [28–30] and determined by our analysis in Staphylococci (see computational analysis section). Additionally, the contribution PAIs play toward colonization of environmental niches was soundly reviewed based on the *E. coli* CFT073 strain and other UPEC strains [31]. To elucidate the differences between genomes of varying G + C content, a division based on compositional G + C was sought. Two datasets were formed and included genomes with chromosomal G + C content less than 35% (101 genomes) forming the "low G + C set" while chromosomes containing greater than 65% G + C (108 genomes) formed the "high GC set." Genomic data was sourced from the NCBI Genomes resource (ftp://ncbi.nlm.nih.gov/genomes/Bacteria). Sequences forming each set can be found in Supplementary data. In summary, the high G + C set includes members of the Actinobacteria, Acidobacteria, Betaproteobacteria, and the Deinococcus-Thermus taxa. The lower G + C set was comprised predominantly of Epsilonproteobacteria, Firmicutes, Fusobacteria and Spirochetes. For each dataset, keywords were extracted from the Genbank coding sequence "product" entries for each aberrant G + C island. In comparison to the low %G + C set, the high %G + C set contained a vastly greater occurrence count for each product except ribosomes. Since high %G + C genomes generally constitute large bacterial genomes of free-living organisms, the scope and capability to acquire laterally transferred DNA is increased. Results obtained in this study indicate the low %G + C set formed 519 aberrant G + C island predictions containing 20856 products while the high %G + C set formed 1448 G + C island predictions containing 58962 products. It can be stated that large genomes contain a greater volume of anomalous G + C regions, hence product numbers. As the study incorporated relatively proportionate numbers of genomes in the high and low G + C sets (108/101 respectively), prediction count and product number was strongly affected by the G + C content of the genome. Of interest are the proteins with the highest occurrence as detailed in Table 1.

Table 1
Keyword analysis of gene products derived from genomes of disparate G + C.

Product keyword	Occurrence in low genomic %G + C dataset	Occurrence in high genomic %G + C dataset
Restriction enzyme	14	50
Transposase/integrase	224	2002
Drug/resistance	111	395
Phage	96	645
Ribosome	1713	1708
ABC-transporter	372	1096
Pilin	2	48
Sugar	182	475
Kinase	347	756

Counts indicate the occurrence of each product keyword in the dataset.

The coding sequences identified in Table 1, represent those typically found in pathogenicity islands [15].

Biological context—biochemical and molecular biasing

Codon usage and G + C nucleotide distribution at a genomic level has revealed several interesting trends. Studies have shown A + T richness to be a feature surrounding the replication terminus [32]. Closely associated with A + T biasing near the replication terminus is the translesion repair mechanism [33,34], a less accurate repair mechanism favoring adenine incorporation. Under the translesion scheme, sequences located toward the replication terminus have hindered repair potential caused by their single copy number in the cycle. The late replicating nature results in less opportunity for other repair mechanisms (including homologous recombination) to be effective. Interestingly, genes associated with the region proximally preceding the replication terminus evolve at an accelerated rate [35]. The induced mis-incorporation of A-T nucleotides impacts both gene evolution and has implications for mis-identification of laterally transferred sequences based on codon usage measures. Nucleotide stability and the tendency to mutate offer at the chemical level, an explanation as to the shift in G + C content. Observed degradation of cytosine to uracil or methylated cytosine to thymine has been studied [36,37] and constitute the most common degradation profile [4]. Studies in *E. coli* have elicited the effect nucleoside diphosphate kinase (NDK) absence has toward increasing polymerase errors [38,39]. The role of regulating dNTP pools are unbalanced by *ndk* disruption, resulting in increased dCTP and dGTP, hence shifting nucleotide composition due to mispairing, A:T → G:C. Interestingly, the replication terminus is site of increased recombination potential [40–42], contributing factors include site specific recombination (*dif* locus), site specific replication pausing and sister chromosome homologous recombination. Experimentation has shown no inherent sequence products to support this assertion, therefore attention has focused upon nucleoid rebuilding effecting the catenation of sister chromosomes [41]. Sister chromosome linking and the proximity at which these sequences are arranged promote genetic exchange. Recently, the role of DNA polymerase III, specifically its α subunit and its impacts on genomic G + C has been demonstrated [43]. The work showed a correlation between α subunit isoforms and the genomic G + C content of the host genome, thus suggesting these subunits effect genomic G + C content. The tRNA, additionally has an important role in shaping genomic G + C [44]. The most abundant tRNA species dictates the shaping of codon usage patterns, particularly in highly expressed genes. The second feature arising from Table 1 is the identification of ribosomes. The detection pattern is different to the other product categories as the high and low %G + C sets return near identical results. The conclusion drawn from this finding would indicate ribosome coding sequences are of conserved G + C content. In terms of defining core chromosomal coding sequences whose nucleotide content remains conserved, ribosomal sequences are strong candidates. The nucleotide level preservation in the context of selectionist theory would suggest that the crucial importance of rRNA sequences would thus limit mutations from being propagated. A study of the G + C content of bacterial rRNA relative to the genomic G + C found rRNA content to vary within a limited range of ~50–53% against a wide genomic G + C range ~40–60% [45]. Evidence however has been elicited regarding rRNA and their ability to be laterally transferred [46–48]. Posed with two potential explanations, whether the G + C aberration of rRNA is due to its conserved function [45] or seen as a factor explained by horizontal transfers, literature would suggest the former in light of rRNA as markers for phylogenetic studies. Our analysis on the available *Buchnera aphidicola* strains has identified regions of G + C composition deviating from the parent sequence (strains APS and Sg). The genomic region spanning nucleotides 527000...567090 in the *Schizaphis graminum* strain genome and the genomic region spanning nucleotides 527000...571042 in the *Acyrthosiphon pisum* strain genome correspond to rRNA encompassing sequences. The maximal G + C deviating regions within the aforementioned bounds consist of 43.18 and 43.96 %G + C within chromosome background G + C of 26.3% (APS strain) and 25.3% (Sg strain), respectively. Such stark deviances suggest the reductive forces acting upon the genome have little impact toward shaping rRNA content. The rRNA represents an interesting compositional element whose overall G + C composition remains relatively intact across a wide spectrum of organisms. A hypothesis to explain such patterns relates to the number and complexity of the interactions undertaken by the gene or protein [9].

Biological context—selective pressure and environmental

Biased mutation rates among the four nucleotides are in large part a response to silencing deleterious changes [49]. The fixation of functionally neutral changes would thus indicate genomic non-coding regions would evolve at a faster rate than coding sequences. The resultant impact of biased mutation impacts codon usage, with stark conclusions drawn to the third position in the codon [49] often denoted as "G + C3." Recently, environmental considerations toward a relationship between optimal growth temperature and G + C genome content have been proposed [50], and then questioned [51] and rebutted [52]. Implication as to DNA stability within extreme temperature environments originally suggested GC bonding patterns as representing a selectable more thermostable complex in such genomes as *Thermus thermophilus* [53]. In addition to the debate over G + C vs. optimal growth temperatures, the environmental division between aerobic and anaerobic prokaryotes have shown increased G + C content for aerobes [54], as shown in Table 2.

The environmental division between free-living and obligate organisms impacts genome content and size greatly. Obligate and non–free-living organisms generally present shorter, A + T rich genomes as highlighted by the currently sequenced endosymbiont genomes, refer Table 3.

Genomes of such organisms tend to be small via a process of reductive evolution [13], in addition to chromosomal deletions. Selective pressure to maintain gene structure is lessened by analogous host gene expression, hence gene loss is not critical. The homogenous host environment additionally may impose nutrient limitations, hence the requirement for lessened energy utilization is preferential. As previously stated, A/T synthesis is energetically less demanding, therefore the genome of the bacterial symbionts are A + T rich. This property has been shown in Table 3, however there exist two notable exceptions including *Serratia proteamaculans* 568 (5.4 Mb chromosome, G + C 0.55) and *Sodalis glossinidius str. morsitans* (4.2 Mb chromosome, G + C 0.55). The former, contributes positive qualities to its host, including plant growth, anti-fungal properties and insect

Table 2
Number summaries of aerobic/anaerobic organism chromosome GC content.

Measure (% G + C)	Anaerobic	Aerobic
Mean	44.03	53.74
Standard error	1.30	0.98
Median	43.10	58.65
Standard deviation	11.80	13.99
Sample variance	139.30	195.68
Skewness	0.28	-0.40
Range	46.30	47.40
Minimum	27.20	26.80
Maximum	73.50	74.20
Sum	3654.70	10963.90
Count	83	204
Largest (1)	73.50	74.20
Smallest (1)	27.20	26.80

Chromosome data collected from NCBI fully sequenced bacterial genomes.

Table 3
Comparison of pathogen vs. endosymbiont genome GC and length.

Organism	Genome size (Mb)	Genome %GC	Environment
Cross environment			
Streptococcus suis 05ZYH33	2.1	41.1	Pig
Burkholderia ambifaria MC40-6	7.6	66.4	Soil
Yersinia pestis Antiqua	4.88	47.7	Soil
Delftia acidovorans SPH-1	6.8	66.5	Soil
Yersinia pestis Nepal516	4.61	47.6	Soil
Campylobacter jejuni subsp. jejuni 81–176	1.68	30.5	Animal feces
Campylobacter jejuni subsp. jejuni NCTC 11168	1.6	30.5	Animal feces
Burkholderia cenocepacia MC0-3	7.9	66.6	Soil
Pseudomonas putida KT2440	6.18	61.5	Soil/water
Campylobacter jejuni subsp. doylei 269.97	1.8	30.6	Animal feces
Pseudomonas fluorescens PfO-1	6.44	60.5	Soil
Ralstonia solanacearum GMI1000	5.8	67	Soil
Gluconobacter oxydans 621H	2.92	60.8	Flowers/fruits
Escherichia coli SMS-3-5	5.25	50.5	Soil
Mesorhizobium loti MAFF303099	7.6	62.5	Soil/root nodules
Sinorhizobium meliloti 1021	6.74	62.2	Soil/root nodules
Listeria monocytogenes EGD-e	2.94	38	Soil/water/organic matter
Methylococcus capsulatus str. Bath	3.3	63.6	Soil
Streptococcus pneumoniae TIGR4	2.2	39.7	Animal nasopharyngea
Streptococcus pneumoniae R6	2.04	39.7	Animal nasopharyngea
Agrobacterium tumefaciens str. C58	5.65	59	Soil/plant roots
Oceanobacillus iheyensis HTE831	3.63	35.7	Deep sea mud
Klebsiella pneumoniae subsp. *pneumoniae* MGH 78578	5.69	57.1	Soil/water/plant surface
Yersinia pestis KIM	4.7	47.7	Animal vectors
Campylobacter jejuni subsp. jejuni 81116	1.6	30.5	Animal feces
Yersinia pestis CO92	4.88	47.6	Animal Vectors
Lactobacillus fermentum IFO 3956	2.1	51.5	Plant/animal material
Finegoldia magna ATCC 29328	1.99	32.1	Human skin
Coxiella burnetii RSA 493	2.03	42.6	Soil
Roseobacter denitrificans OCh 114	4.3	58.9	Marine sediment
Campylobacter jejuni RM1221	1.8	30.3	Animal feces
Shewanella woodyi ATCC 51908	5.9	43.7	Sea water
Average values	4.21	49.07	
Endosymbioints			
Baumannia cicadellinicola str. Hc	0.69	33.2	*Homalodisca coagulata*
B. aphidicola str. APS	0.66	26.4	*A. pisum*
B. aphidicola str. Bp	0.62	25.3	*Baizongia pistaciae*
B. aphidicola str. Cc	0.42	20.1	*Cinara cedri*
B. aphidicola str. Sg	0.64	25.3	*S. graminum*
Candidatus Vesicomyosocius okutanii HA	1	31.6	*Calyptogena okutanii*
S. glossinidius str. *morsitans*	4.29	54.5	*Glossinia morsitans*
Wigglesworthia glossinidia	0.7	22.5	*Glossina brevipalpis*
Wolbachia	1.27	35.2	*Drosophila melanogaster*
Average	1.14	30.46	

pathogenicity, see [55] and references within. The nature of its soil-based environment would thus pose an amenable environment for lateral transfers, c.f. the environments of the pathogens in Table 3. *S. glossinidius* (an endosymbiont of tsetse flies), showed an inordinate number of pseudogenes for a bacterial genome (972) and provided indications of recent divergence from a free-living ancestor [56]. Bacterial symbionts thus retain a rich AT genome character based on the current body of sequenced symbiont genomes, a trend to be investigated by further sequencing projects. In addition to ribosomes, essential house keeping sequences, cell surface adhesin proteins were detected in the datasets mentioned previously. The presence of the cell wall binding proteins in a G + C scan was very significant. Firstly, their presence would suggest a level of conservation, secondly a pointer to their origin, since they deviate compositionally from the background genome and thirdly, a correlation with the pathogenic binding ability to specific cell types in the host organism. Subsequent BLASTN [57] analysis of these regions against the "refseq_genomic" database revealed two important findings. The first finding indicated the cell wall protein of *S. epidermidis* RP62A (2317000..2319555) showed homology to Streptococcal genomes. Utilizing BLASTN analysis, resulting hits showed similarity to other strains within the *Staphylococci* (scoring 1.18e + 04 to 52), however organisms outside this genus show *Streptococcal* alignments with scores ranging between 319 and 44.6. The second finding was the resolution of the cell wall component in *S. saprophyticus* subsp. ATCC 15305 (153000..157915), the uniqueness of this protein has direct implication for its binding ability, a uterine pathogen c.f. the nasopharyngeal/integumentary Staphylococci. This isolated protein performing specific cellular attachment with the host tissue type was not found in any other organism, related or non-related following subsequence (discontiguous and mega) BLASTN analysis. Positive selection of these proteins may indicate a strong case for natural selection. The larger size of the *Staphylococci* and their high G + C genomic content compared to the other chromosomes in the "low GC" dataset (consisting of low G + C and short genome endosymbionts) would suggest a greater potential for lateral transfer.

The simplest analogue is presented by influenza A-hemagglutination binding to sialic-acid-containing cell-surface receptors has been well studied [58]. The binding ability displayed in viral infection has analogous presentations in many higher order taxonomic classifications. Owing to the ability for viral DNA to be inserted into host genomes, the transfer of generic material is one explanation for the presence and divergence of progenitor cell wall attachment proteins in bacteria. The compositional persistence of the cellular attachment sequence would thus indicate a potential lateral origin and its subsequent importance in colonization and survival in an atypical environment.

From the G + C analysis perspective, we theorize, profiling of low % G + C genomes for aberrant regions would highlight core regions critical to survival (in an already reduced genome). The environment of their host heavily limits symbionts and as such, evolutionary pressures adapt to mold genomic priorities, nutrient availability, energy costing and the lack of background heterogeneity limits lateral uptakes. The genome reduction found in the *Buchnera* strains suggest a close association with its aphid host to the extent of true symbiotic mutual metabolite transfer [59], showing no evidence of horizontal transfers [60]. Interestingly, the *Buchnera* genome includes genes for metabolites not encoded by the host genome and vice-versa, suggesting a very ancient association with the host [5,59].

Computational context

From a historical perspective, computational effort toward genomic segmentation of compositional features has centered around usage and distribution patterns of nucleotides. As summarized by Karlin [61], compositional variances can be analysed through G + C window calculations, dinucleotide bias (often termed genomic signature) and discernable differences in codon and amino acid usage. The previously mentioned techniques, provide at an overview level, a reasonable level of description for a genomic sequence. Detractions from these techniques readily occur when compositional deviance levels are small in magnitude (with respect to the parent background genome) and resolution concerns become apparent through the use of sliding windows of unknown size to define features not fitting the *a priori* consensus length.

G + C frequency profiling

Computational models and algorithms generally seek to segment sequences based on identifiable shifts in G + C composition. Several techniques exist beyond simple compositional measures for genome segmentation with variances based on prior threshold setting and resolution capability. The ultimate measure of effectiveness of such algorithms relies on their ability to handle segmentation with the best judgment/refinement/optimization of input parameters. The heterogeneity of target sequence length e.g. PAIs whose length range can extend between 10 and 200 kb [27,62], poses significant challenges for techniques examining low-level defined features such as codon and individual nucleotide G + C counts. To begin the background study of more advanced techniques, the reader is advised to keep foremost in their minds, the degree to which decision criteria are optimized, ultimately produces the most desirable outcome. One should also take a view to the generalization of such techniques for profiling features covering gene (1–2 kb) to large lateral tract length scales (~200 kb). A segmentation algorithm referred to as "entropic segmentation" [63] has been proposed to identify regions of abnormal composition for isochore determination at the eukaryote sequence level. Isochores represent long stretches of defined G + C content DNA in eukaryote genomes. The central aspect of the approach focuses on the use of the Jensen–Shannon divergence measure [64], a technique rooted in information theory. Using this process, nucleotide divergence relative to parent sequences can be determined. Since the statistical foundations call for a confidence level, this key criteria has the potential to limit effectiveness. The aim of the procedure is to identify homogeneity within a region (isochore) against a heterogeneous background (genome). The entropic segmentation method consisted of three main components: entropic decision making, stop criteria determination and filtering. Reported outcomes include a more refined boundary for human major histocompatibility complex isochores. Practical implications of entropic decision making for bacterial genome segmentation arise through their use of enhancements (statistical significance) to cut site selection. The iterative approach terminated by stop criteria offer another practical consideration, the *a priori* determination of feature length, in this case isochores, for bacterial genomes, pathogenicity islands. Toward targeting pathogenicity islands and lateral transfer regions, a wavelet shrinkage approach has been effectively applied to smooth insignificant G + C variations [65,66]. Wavelet analysis is the process in which analysis is proportional to scale, a desirable attribute when G + C regions are of unknown size. Given G + C as a continuous property over the length of the genome, wavelet techniques seek to form components (GC compositional features in this context) with boundaries appropriate to the scale of the feature. The multiple scale representation of wavelets extends their usability to modeling features of variable lengths. Showing similarity with the optimizations made in the entropic segmentation method, wavelet coefficients are eliminated "shrinkage" with the use of thresholds, effectively removing noise from the data. Threshold values approaching 1.0 were recognized as providing the most useful level of smoothing while capturing features of the data. Two applications of wavelet procedures were employed to smooth and subsequently identify significant profile component characteristics. Outcomes for the method suggest a technique for modeling both small and large features; reported findings included the identification of two putative pathogenicity islands in N. mengingitidis [65]. Hidden Markov Models (HMM) have been applied to G + C tract identification [67], with segmentation (coding vs. non-coding) ability in the yeast genome. The HMM represents a model widely used for many bioinformatics applications. Toward G + C genome segmentation, the approach concentrated on determining the optimal number of states, a process designed to optimize model effectiveness analogous to the parameter optimization witnessed in the previous two techniques. Outcomes indicated a four state model as ideal for describing genomic G + C patterns.

Apart from the inability to determine resolution that often leads to sub-standard outcomes, other limitations of the window approach have been detailed in [22]. Primary among these limitations include the inability to accurately define feature boundaries down to the single nucleotide scale and the prior knowledge required to estimate window sizing. To overcome some of these limitations, a cumulative G + C profiling technique [22], a windowless approach (Z-curve), has been proposed toward identifying regions of G + C abnormality. The idea of this G + C profiling technique is founded on the independent distribution of purine/pyrimidine, amino/keto and weak/strong hydrogen bonds, the latter describing GC/AT distribution. Outcomes include a clear demarcation of the genomic island where G + C abnormality arises. Simplistic G + C window calculations often prove useful toward gaining an overview level of the genomic nucleotide distribution. G + C window calculations consist of dividing a genome into fixed windows of a defined length *l* and sampled at frequency *f*. Windows can be discrete or overlapping, calculations for each window are evaluated as (G + C)/(A + T + G + C) * 100% and represented in the literature as "%G + C". G + C detection methods outlined previously constitute an important technique for compositional profiling; however as we discuss next, more subtle factors effecting nucleotide distribution can be used.

Constrained nucleotide

Using the umbrella term "constrained nucleotide", techniques including GC skews (C-G/C + G), dinucleotide biasing and aberrant nucleotide-amino acid usage, can be employed to elicit evidence supporting horizontal transfers [61]. Common to these processes is the use of windows in which frequency calculations are made relative to other windows and to the broader genome. The genome signature method [68] and the technique proposed by [61], both detect G + C abnormalities through the shift in expected dinucleotide usage relative to a random null model. Such calculations are made according to $p_{XY,} = fr_{XY} / fr_x fr_y$ whereby fr_x represents the frequency of nucleotide X whereas fr_{XY} represents the dinucleotide frequency. The genomic signature difference between two sequences δ^* is thus the summed difference between all dinucleotides of both sequences, similarly differences in codon and amino acid usage can be calculated, see [61]. In the genome signature method, windows can be of an arbitrary length. Practical outcomes obtained via codon or amino acid usage comparisons should encompass gene(s) in clusters of selected length. Codon usage and the biases introduced via several biological processes including lateral transfers, tRNA species biasing and translational efficiency, transcription effectiveness and transcript stability constitute detectable variances, see [69] for a review. Nucleotide substitution rates "Q" pose an interesting approach for compositional profiling [70]. Central to this approach is the principle of genome specific mutational bias and the resulting nucleotide shift it imparts. Horizontal transfers are thus hypothesized to transfer genes from a donor of differing rate matrix into an acceptor whose rate matrix is detectibly different. The key qualification of this approach is the discrimination ability to gauge incongruence between rate matrices. Using orthologous gene displacement as a case study (replacement or supplementation by the inserted gene), such models mandate orthologous genes had shared a common ancestor and the inserted sequence evolved according to a Q different from the acceptor genome under analysis. Simulated experimentation has shown a 10-fold reduction in error rates utilizing a multitude of predictor statistics encapsulating phylogenetic tree and compositional (codon 3rd position) data.

More recently and more widely used computational methods include the window-based approach [71] and the windowless approach [22] for G + C content profiling. Window based schemes rely on taking selective portions of the genome and calculating compositional properties, exemplar properties include those detailed above, and in addition, G + C frequency usage data. Two variables are commonly associated with the window approach: window size and sampling frequency. Window size is a variable governing resolution; a small window is susceptible to statistical abnormalities, while a large window may not resolve the region. Secondly, the period of sampling i.e. the frequency at which windows are sampled, is closely related to window size. Infrequent sampling may negate the ability to resolve features and are compensated by large windows (inherently low resolution), while the inverse argument is applicable. Several variations of the window-based approach for G + C content profiling can be found in [72]. Other forms of compositional modeling centering on codon usage include correspondence analysis (CA) [73] and the codon adaptation index (CAI) [74]. Correspondence analysis is the statistical analysis of two and multi-way tabular data (contingency tables). The use of CA toward sequence composition has been varied over an extended period and include, HGT detection [20], coding region detection [75] and amino acid composition trends [76], moreover the technique is used to highlight the relationship between variables. Typical associations could entail modeling codon usage vs. genes, as in [74]. Factors influencing the effectiveness of the CA approach are exemplified by their inherent tendency to over fit, in addition, relative measures such as relative synonymous codon usage (RSCU) may mask trends observed using simple codon counts. The univariate measure CAI [77] aims to show the direction of synonymous codon bias. As before, tabular calculations of codon usage for a set of genes serve as a starting point with CAI methodology. Interpretation of results indicates the fitness of the gene with respect to the tRNA pool of the genome. High CAI values may indicate a greater level of expression. Such comparisons of CAI vs. a reference set of genes, may highlight the deviance encountered by laterally transferred DNA [78]. The "Wn" method [79], based on templates of size n (n>2) offers the ability to rate gene typicality without the knowledge of codon boundaries. The procedure utilized higher order nucleotide templates independent of individual di- and trinucleotide boundaries to aid discrimination. The resultant performance increase over CAI and G + C was significant, exemplified by relative improvement increases of between 11% and 44% over CAI. In subsequent work [80], support vector machines were used as an enhanced similarity measure, thus gaining an average 10% improvement and adapting the method to the analysis of short genomes.

Phylogenetics schemes

Taking the phylogenetics approach, genes are mapped according to their similarities with respect to related and distant genomes. The clustering of a gene to distantly related genomes instead of those within close taxonomic classifications can potentially indicate HGT. Several complications exist when seeking HGT events via phylogenetics techniques. The first detraction is centered on data limitation, the lack of candidate similar and distant genomes effect the meaningfulness of the predictions. Other detractions concentrate over genetic processes that could simulate the outcomes of HGT, this can include lineage sorting arising from random genetic drift [81]. Further complications are caused by gene duplications, gene loss, and variable rates of sequence evolution. Algorithms to address these shortcomings center on the underlying basis causing the incongruence between gene and species trees [82–85] and reviewed in [86]. Strengths of the phylogenetic technique over "sequence based" (surrogate) techniques rely upon modeling evolutionary information [87]. Tree comparisons are conducted via algorithms such as efficient evaluation of edit paths (EEEP) [85], whereby the subtree prune-and-regraft distance [88] is seen as a suitable metric for determining HGT. The basis relies upon pruning of the recipient tree to reattach in line with donor (reference) tree topology, a process mirroring HGT. The minimal path length between reference and test tree is an indication of the number of HGT events undergone by the gene in the test tree. Tree comparisons of this form have enabled determination of HGT patterns between closely and distantly related lineages [89]. Further discussion of tree comparison methods and their associated tools have been provided recently [90], with references therein. Of significant importance is the choice of reference tree construction, as differing methods (16S rRNA or genome common proteins) have the potential to cloud results based on phylogenetic assumptions arising from mutational bias.

Database schemes

Lateral transfers are commonly predicted to transfer extensive regions of clustered functionality [15]. Characteristic proteins found in laterally transferred DNA include the encoded products of transposases, toxins, secretion and transport systems. This functional set of genes provides a reference set against which similarity alignment measures can be compared. Several databases exist to aid similarity searching against putative [20] and known genomic islands [91] and more specifically virulence factors [92] and pathogenicity islands [93]. The strategy of comparing putative transferred regions vs. those of known HGT regions via BLAST has been criticized [94]. The key contention arises from the simplicity by which conclusions are drawn from alignment results. Phylogenetic reconstruction is clouded by convoluting factors including compositional biases in thermophilic bacteria/archaea and the sweeping conclusions drawn across taxonomic domains based on limited analysis and truncated datasets. The effect of gene loss in studies limited by the scope of the comparative clade pool may lead to cursory findings without the appropriate knowledge to draw more meaningful conclusions. Practical uses of BLAST toward eliciting HGT events can be found in various studies and include typical use-cases of HGT boundary exploration [95], cross-

Table 4
Comparison of computational techniques for genomic segmentation and island characterization.

Attribute	Method entropic	HMM	Wavelet	Z-curve	Nucl. distributions	Phylogenetics	Database
Training dependency	POS	NEG model architecture based on training data	POS	POS	POS	NEG requires a defined set of comparative sequences	NEG requires a defined set of comparative sequences
Explicit parameter Setting	NEG stop parameters	NEG model construction limits borders	NEG smoothness threshold	POS	NEG parameters	MAR tree parameters	NEG thresholds common to gauge similarity
Fine resolution	NEG relative error of boundaries 0.15–0.05% for sequences ~300 kb	NEG difficult to gauge based on "best" model	POS	POS	NEG multitude of analysis techniques may not converge	N.A.	MAR matches the quality of the alignment
Scalability	MAR tuned for specific application	NEG models specific to feature resolved	POS	POS	NEG potentially limited by resolution	MAR dependent on the number of sequences in the comparative set	MAR dependent on the number of sequences in the comparative set

Attributes for each method are defined as emergent properties common to G + C profiling algorithms. Three acronyms indicate positive (POS), marginal (MAR) and negative (NEG) aspects. Attributes are defined as: Training dependency: reliance on training, Explicit parameter setting: requirement for pre-defined thresholds and other criteria, Fine resolution: the nucleotide offset expected per prediction, Scalability: determination if the model can be applied outside of the target application to whole genome analysis. N.A: Not Applicable.

species and higher taxon presence [96], protein characterization and homolog determination [97].

Practical guidance

Bacterial genomic content profiling is a powerful method for eliciting macro and micro scale properties of genomes. The choice of technique, whether G + C frequency, constrained nucleotide, phylogenetic or database, largely falls upon the choice of target. The biological examples presented in this manuscript highlight several such case studies that can be answered through compositional profiling. Preliminary genomic scans to reveal large regions of HGT are best matched with G + C frequency and constrained nucleotide analysis. These techniques operate at the nucleotide level and function independently of any assumed or comparative knowledge employed by other techniques. With "islands" identified further investigation of the highlighted regions can be undertaken via constrained nucleotide and database techniques. G + C frequency analysis can also be used at this level, sensitive implementations including wavelet and z-curves refine accurately HGT segment boundaries. Window based methods, as previously stated, lack the required resolution to accurately determine borders of laterally transferred DNA. With putative HGT regions identified, database techniques can be employed to determine their occurrence in other organisms. Database similarity searching for the purpose of identifying origins however is highly dependent on the organism under analysis and the body of prior sequences upon which to perform comparative analysis. To perform more sensitive "source of origin" studies, phylogenetic techniques are required.

A comparison of techniques highlights the issues facing genomic segmentation, see Table 4. The attributes listed in Table 4, follow a general trend whereby reliance on training data limits scalability.

To draw conclusions between all the methods presented in this review, with the exception of the Z-curve, all techniques are restricted to varying degrees by dependency on optimal initial parameter estimation. Table 5, demonstrates a comparison of different techniques in non-specific nucleotide aberrant region detection available from published studies. All techniques identified known PAIs, however, differentiation occurred in the detection of informational genes, in particular AlienHunter [98].

The approaches mentioned thus far have been non-specific with respect to identifying regions of G + C aberrance. There however exist derived techniques for the identification of PAIs (PAI-DB [93]), phages (PhageFinder [99]) and genomic islands (Islander [91], AlienHunter [98]). Many of these algorithms are based upon BLAST alignment or modeling characteristics (insertion sites, gene sequences) of these defined units of HGT.

Conclusion

We have seen the breadth of knowledge that can be elicited from bacterial genomes through compositional profiling. The role of

Table 5
Comparison of aberrant nucleotide detection algorithms.

Genome	Predictions				Unique predictions				Unique products			
	AlienHunter	Wavelet	Z-curve	Genome signature	AlienHunter	Wavelet	Z-curve	Genome signature	AlienHunter	Wavelet	Z-Curve	Genome signature
D. radiodurans R1 chr I	48	7	–	–	25	1	–	–	Informational genes	Unknown	–	–
H. pylori 26695	44	5	–	–	40	0	-	-	Informational Genes	0	–	–
H. pylori J99	47	3	–	–	39	1	–	–	Informational Genes	Type II DNA modification	–	–
N. meningitidis Z2491	36	8	–	–	18	0	–	–	Informational genes	0	–	–
N. meningitidis MC58	34	10	–	–	18	1	–	–	Informational genes	Unknown	–	–
C. glutamicum	47	–	1	–	40	–	0	–	Informational genes	–	0	–
V. vulnificus CMCP6 chr. I	50	–	3	10	37	–	0	0	Informational genes	–	0	0
V. vulnificus CMCP6 chr. II	58	–	–	11	44	–	–	1	Informational genes	–	–	GI

Informational genes indicate rRNA, DNA polymerases. GI indicates a genomic island.

environmental adaptation and its outcomes in terms of genomic G + C has been presented via factors that increase (HGT, aerobiosis) and decrease (translesion, phage insertion, cytosine degradation) G + C. Computational techniques to aid the bioinformatician have formed the basis for discovering such knowledge, however much has still to be determined regarding algorithmic biological relevancy. The techniques presented in this review constitute in large part, specialized approaches to target *a priori* desired features. To date, many of these efforts are constrained by thresholding and other parameters needed in order to adapt the techniques for practical outcomes. In conclusion, current algorithmic limitations mirror those of the traditional sliding window schemes, thus *a priori* knowledge is still seen as a determining factor to the success of computational genomic profiling algorithms.

Acknowledgments

The work in this paper was partially supported by Australian Research Council Grants LP0349235 and LX0560616.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2009.09.002.

References

[1] N. Sueoka, On the genetic basis of variation and heterogeneity of DNA base composition, Proc. Natl. Acad. Sci. U. S. A. 48 (1962) 582–592.
[2] A. Nakabachi, A. Yamashita, H. Toh, H. Ishikawa, H.E. Dunbar, N.A. Moran, M. Hattori, The 160-kilobase genome of the bacterial endosymbiont Carsonella, Science 314 (5797) (2006) 267.
[3] E.P.C. Rocha, A. Danchin, Base composition bias might result from competition for metabolic resources, Trends Genet. 18 (6) (2002) 291–294.
[4] X. Xia, H. Wang, Z. Xie, M. Carullo, H. Huang, D. Hickey, Cytosine usage modulates the correlation between CDS length and CG content in prokaryotic genomes, Mol. Biol. Evol. 23 (7) (2006) 1450–1454.
[5] N.A. Moran, Microbial minimalism: genome reduction in bacterial pathogens, Cell 108 (5) (2002) 583–586.
[6] D. Mitchell, GC content and genome length in Chargaff compliant genomes, Biochem. Biophys. Res. Commun. 353 (1) (2007) 207–210.
[7] K.U. Foerstner, C. von Mering, S.D. Hooper, P. Bork, Environments shape the nucleotide composition of genomes, EMBO Rep. 6 (12) (2005) 1208–1213.
[8] S.D. Bentley, J. Parkhill, Comparative genomic structure of prokaryotes, Annu. Rev. Genet. 38 (2004) 771–792.
[9] R. Jain, M.C. Rivera, J.A. Lake, Horizontal gene transfer among genomes: the complexity hypothesis, PNAS 96 (7) (1999) 3801–3806.
[10] L.S. Frost, R. Leplae, A.O. Summers, A. Toussaint, Mobile genetic elements: the agents of open source evolution, Nat. Rev. Microbiol. 3 (9) (2005) 722–732.
[11] C. Canchaya, G. Fournous, H. Brussow, The impact of prophages on bacterial chromosomes, Mol. Microbiol. 53 (1) (2004) 9–18.
[12] H. Ochman, J.G. Lawrence, E.A. Groisman, Lateral gene transfer and the nature of bacterial innovation, Nature 405 (6784) (2000) 299–304.
[13] J. Wixon, Featured organism: reductive evolution in bacteria: Buchnera sp., Rickettsia Prowazekii and Mycobacterium Leprae, Compar. Funct. Genom. 2 (1) (2001) 44–88.
[14] B. Alberts, Molecular biology of the cell, 4th edGarland Science, New York, 2002.
[15] J. Hacker, J.B. Kaper, Pathogenicity islands and the evolution of microbes, Annu. Rev. Microbiol. 54 (2000) 641–679.
[16] B. Middendorf, B. Hochhut, K. Leipold, U. Dobrindt, G. Blum-Oehler, J. Hacker, Instability of pathogenicity islands in uropathogenic *Escherichia coli* 536, J. Bacteriol. 186 (10) (2004) 3086–3096.
[17] U. Dobrindt, G. Blum-Oehler, G. Nagy, G. Schneider, A. Johann, G. Gottschalk, J. Hacker, Genetic structure and distribution of four pathogenicity islands (PAI I (536) to PAI IV(536)) of uropathogenic *Escherichia coli* strain 536, Infect. Immun. 70 (11) (2002) 6365–6372.
[18] G. Schneider, U. Dobrindt, H. Bruggemann, G. Nagy, B. Janke, G. Blum-Oehler, C. Buchrieser, G. Gottschalk, L. Emody, J. Hacker, The pathogenicity island-associated K15 capsule determinant exhibits a novel genetic structure and correlates with virulence in uropathogenic *Escherichia coli* strain 536, Infect. Immun. 72 (10) (2004) 5993–6001.
[19] S.H. Yoon, C.G. Hur, H.Y. Kang, Y.H. Kim, T.K. Oh, J.F. Kim, A computational approach for identifying pathogenicity islands in prokaryotic genomes, BMC Bioinformatics. 6 (2005) 184.
[20] S. Garcia-Vallve, E. Guzman, M.A. Montero, A. Romeu, HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes, Nucleic Acids Res. 31 (1) (2003) 187–189.
[21] R. Romeu, C.T. Zhang, Genomic islands in the *Corynebacterium efficiens* genome, Appl. Environ. Microbiol. 71 (6) (2005) 3126–3130.
[22] R. Zhang, C.T. Zhang, A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I, Bioinformatics 20 (5) (2004) 612–622.
[23] J. Hacker, G. Blum-Oehler, I. Muhldorfer, H. Tschape, Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution, Mol. Microbiol. 23 (6) (1997) 1089–1097.
[24] S. Garcia-Vallve, A. Romeu, J. Palau, Horizontal gene transfer in bacterial and archaeal complete genomes, Genome Res. 10 (11) (2000) 1719–1725.
[25] D.S. Guttman, D.E. Dykhuizen, Clonal divergence in *Escherichia coli* as a result of recombination, not mutation, Science 266 (5189) (1994) 1380–1383.
[26] S.L. Chen, C.S. Hung, J. Xu, C.S. Reigstad, V. Magrini, A. Sabo, D. Blasiar, T. Bieri, R.R. Meyer, P. Ozersky, et al., Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach, Proc. Natl. Acad. Sci. U. S. A. 103 (15) (2006) 5977–5982.
[27] H. Schmidt, M. Hensel, Pathogenicity islands in bacterial pathogenesis, Clin. Microbiol. Rev. 17 (1) (2004) 14–56.
[28] L. Hagberg, U. Jodal, T.K. Korhonen, G. Lidin-Janson, U. Lindberg, C. Svanborg Eden, Adhesion, hemagglutination, and virulence of *Escherichia coli* causing urinary tract infections, Infect. Immun. 31 (2) (1981) 564–570.
[29] H. Connell, M. Hedlund, W. Agace, C. Svanborg, Bacterial attachment to uro-epithelial cells: mechanisms and consequences, Adv. Dent. Res. 11 (1) (1997) 50–58.
[30] C.S. Eden, L.A. Hanson, U. Jodal, U. Lindberg, A.S. Akerlund, Variable adherence to normal human urinary-tract epithelial cells of *Escherichia coli* strains associated with various forms of urinary-tract infection, Lancet 1 (7984) (1976) 490–492.
[31] N.J. Parham, S.J. Pollard, R.R. Chaudhuri, S.A. Beatson, M. Desvaux, M.A. Russell, J. Ruiz, A. Fivian, J. Vila, I.R. Henderson, Prevalence of pathogenicity island IICFT073 genes among extraintestinal clinical isolates of *Escherichia coli*, J. Clin. Microbiol. 43 (5) (2005) 2425–2434.
[32] V. Daubin, G. Perriere, G+C3 structuring along the genome: a common feature in prokaryotes, Mol. Biol. Evol. 20 (4) (2003) 471–483.
[33] P. Deschavanne, J. Filipski, Correlation of GC content with replication timing and repair mechanisms in weakly expressed *E. coli* genes, Nucleic Acids Res. 23 (8) (1995) 1350–1353.
[34] B.A. Bridges, Error-prone DNA repair and translesion DNA synthesis. II: The inducible SOS hypothesis, DNA Repair (Amst.) 4 (6) (2005) 725–726, 739.
[35] P.M. Sharp, D.C. Shields, K.H. Wolfe, W.H. Li, Chromosomal location and evolutionary rate variation in enterobacterial genes, Science 246 (4931) (1989) 808–810.
[36] J.I. Glass, E.J. Lefkowitz, J.S. Glass, C.R. Heiner, E.Y. Chen, G.H. Cassell, The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*, Nature 407 (6805) (2000) 757–762.
[37] T. Lindahl, Instability and decay of the primary structure of DNA, Nature 362 (6422) (1993) 709–715.
[38] J.H. Miller, P. Funchain, W. Clendenin, T. Huang, A. Nguyen, E. Wolff, A. Yeung, J.H. Chiang, L. Garibyan, M.M. Slupska, et al., *Escherichia coli* strains (ndk) lacking nucleoside diphosphate kinase are powerful mutators for base substitutions and frameshifts in mismatch-repair-deficient strains, Genetics 162 (1) (2002) 5–13.
[39] Q. Lu, X. Zhang, N. Almaula, C.K. Mathews, M. Inouye, The gene for nucleoside diphosphate kinase functions as a mutator gene in *Escherichia coli*, J. Mol. Biol. 254 (3) (1995) 337–341.
[40] R. Rothstein, B. Michel, S. Gangloff, Replication fork pausing and recombination or "gimme a break", Genes. Dev. 14 (1) (2000) 1–10.
[41] J. Louarn, F. Cornet, V. Francois, J. Patte, J.M. Louarn, Hyperrecombination in the terminus region of the *Escherichia coli* chromosome: possible relation to nucleoid organization, J. Bacteriol. 176 (24) (1994) 7524–7531.
[42] N.R. Leslie, D.J. Sherratt, Site-specific recombination in the replication terminus region of *Escherichia coli*: functional replacement of dif, EMBO. J. 14 (7) (1995) 1561–1570.
[43] X. Zhao, Z. Zhang, J. Yan, J. Yu, GC content variability of eubacteria is governed by the pol III alpha subunit, Biochem. Biophys. Res. Commun. 356 (1) (2007) 20–25.
[44] T. Ikemura, Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system, J. Mol. Biol. 151 (3) (1981) 389–409.
[45] N.R. Pace, Structure and synthesis of the ribosomal ribonucleic acid of prokaryotes, Bacteriol. Rev. 37 (4) (1973) 562–603.
[46] W.H. Yap, Z. Zhang, Y. Wang, Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon, J. Bacteriol. 181 (17) (1999) 5201–5209.
[47] Y. Wang, Z. Zhang, Comparative sequence analyses reveal frequent occurrence of short segments containing an abnormally high number of non-random base variations in bacterial rRNA genes, Microbiology 146 (Pt. 11) (2000) 2845–2854.
[48] L.M. Schouls, C.S. Schot, J.A. Jacobs, Horizontal transfer of segments of the 16S rRNA genes between species of the *Streptococcus anginosus* group, J. Bacteriol. 185 (24) (2003) 7241–7246.
[49] A. Muto, S. Osawa, The guanine and cytosine content of genomic DNA and bacterial evolution, Proc. Natl. Acad. Sci. U. S. A. 84 (1) (1987) 166–169.
[50] H. Musto, H. Naya, A. Zavala, H. Romero, F. Alvarez-Valin, G. Bernardi, Correlations between genomic GC levels and optimal growth temperatures in prokaryotes, FEBS Lett. 573 (1-3) (2004) 73–77.
[51] H.C. Wang, E. Susko, A.J. Roger, On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors, Biochem. Biophys. Res. Commun. 342 (3) (2006) 681–684.

[52] H. Musto, H. Naya, A. Zavala, H. Romero, F. Alvarez-Valin, G. Bernardi, Genomic GC level, optimal growth temperature, and genome size in prokaryotes, Biochem. Biophys. Res. Commun. 347 (1) (2006) 1–3.

[53] Y. Kagawa, H. Nojima, N. Nukiwa, M. Ishizuka, T. Nakajima, T. Yasuhara, T. Tanaka, T. Oshima, High guanine plus cytosine content in the third letter of codons of an extreme thermophile. DNA sequence of the isopropylmalate dehydrogenase of *Thermus thermophilus*, J. Biol. Chem. 259 (5) (1984) 2956–2960.

[54] H. Naya, H. Romero, A. Zavala, B. Alvarez, H. Musto, Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes, J. Mol. Evol. 55 (3) (2002) 260–264.

[55] K.E. Ashelford, J.C. Fry, M.J. Bailey, M.J. Day, Characterization of Serratia isolates from soil, ecological implications and transfer of *Serratia proteamaculans* subsp. quinovora Grimont et al. 1983 to Serratia quinivorans corrig., sp. nov, Int. J. Syst. Evol. Microbiol. 52 (Pt. 6) (2002) 2281–2289.

[56] H. Toh, B.L. Weiss, S.A. Perkin, A. Yamashita, K. Oshima, M. Hattori, S. Aksoy, Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host, Genome Res. 16 (2) (2006) 149–156.

[57] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.

[58] A. Varki, National Center for Biotechnology Information (U.S.): Essentials of glycobiology, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1999.

[59] S. Shigenobu, H. Watanabe, M. Hattori, Y. Sakaki, H. Ishikawa, Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS, Nature 407 (6800) (2000) 81–86.

[60] R.C. van Ham, J. Kamerbeek, C. Palacios, C. Rausell, F. Abascal, U. Bastolla, J.M. Fernandez, L. Jimenez, M. Postigo, F.J. Silva, et al., Reductive genome evolution in *Buchnera aphidicola*, Proc. Natl. Acad. Sci. U. S. A. 100 (2) (2003) 581–586.

[61] S. Karlin, Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes, Trends Microbiol. 9 (7) (2001) 335–343.

[62] S. Mann, J. Li, Y.P.P. Chen, A pHMM-ANN based discriminative approach to promoter identification in prokaryote genomic contexts, Nucleic Acids Res. 35 (2007) e121–e127.

[63] J.L. Oliver, P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, Isochore chromosome maps of eukaryotic genomes, Gene 276 (1-2) (2001) 47–56.

[64] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans. Inf. Theory 37 (1) (1991) 145–151.

[65] P. Lio, M. Vannucci, Finding pathogenicity islands and gene transfer events in genome data, Bioinformatics 16 (10) (2000) 932–940.

[66] Y.P.P. Chen, F. Chen, Targets for drug discovery using bioinformatics, Expert Opin. Ther. Targets 12 (4) (2008) 383–389.

[67] L. Peshkin, M.S. Gelfand, Segmentation of yeast DNA using hidden Markov models, Bioinformatics 15 (12) (1999) 980–986.

[68] M.W. van Passel, A. Bart, H.H. Thygesen, A.C. Luyf, A.H. van Kampen, A. van der Ende, An acquisition account of genomic islands based on genome signature comparisons, BMC Genomics 6 (2005) 163.

[69] M.D. Ermolaeva, Synonymous codon usage in bacteria, Curr. Issues. Mol. Biol. 3 (4) (2001) 91–97.

[70] M. Hamady, M.D. Betterton, R. Knight, Using the nucleotide substitution rate matrix to detect horizontal gene transfer, BMC Bioinformatics 7 (2006) 476.

[71] Y. Nishio, Y. Nakamura, Y. Kawarabayasi, Y. Usuda, E. Kimura, S. Sugimoto, K. Matsui, A. Yamagishi, H. Kikuchi, K. Ikeo, et al., Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of Corynebacterium efficiens, Genome. Res. 13 (7) (2003) 1572–1579.

[72] D.G. Lee, J.M. Urbach, G. Wu, N.T. Liberati, R.L. Feinbaum, S. Miyata, L.T. Diggins, J. He, M. Saucier, E. Déziel, et al., Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial, Genome. Biol. 7 (10) (2006) R90.

[73] M.O. Hill, Correspondence analysis: a neglected multivariate method, Appl. Statist. 23 (3) (1974) 340–354.

[74] G. Perriere, J. Thioulouse, Use and misuse of correspondence analysis in codon usage studies, Nucleic Acids Res. 30 (20) (2002) 4548–4555.

[75] G. Fichant, C. Gautier, Statistical method for predicting protein coding regions in nucleic acid sequences, Comput. Appl. Biosci. 3 (4) (1987) 287–295.

[76] J.R. Lobry, C. Gautier, Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes, Nucleic Acids Res. 22 (15) (1994) 3174–3180.

[77] P.M. Sharp, W.H. Li, The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Res. 15 (3) (1987) 1281–1295.

[78] N.T. Perna, G.F. Mayhew, G. Posfai, S. Elliott, M.S. Donnenberg, J.B. Kaper, F.R. Blattner, Molecular evolution of a pathogenicity island from enterohemorrhagic *Escherichia coli* O157:H7, Infect. Immun. 66 (8) (1998) 3810–3817.

[79] A. Tsirigos, I. Rigoutsos, A new computational method for the detection of horizontal gene transfer events, Nucleic Acids Res. 33 (3) (2005) 922–933.

[80] A. Tsirigos, I. Rigoutsos, A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes, Nucleic Acids Res. 33 (12) (2005) 3699–3707.

[81] C. Than, D. Ruths, H. Innan, L. Nakhleh, Identifiability issues in phylogeny-based detection of horizontal gene transfer, Comparative Genomics, vol. 4205, Heidelberg, Springer Berlin, 2006, pp. 215–229.

[82] A. Boc, V. Makarenkov, New efficient algorithm for detection of horizontal gene transfer events, Algorithms in Bioinformatics, Heidelberg, Springer, 2003, pp. 190–201.

[83] M.A. Suchard, Stochastic models for horizontal gene transfer: taking a random walk through tree space, Genetics 170 (1) (2005) 419–431.

[84] J. Hein, Reconstructing evolution of sequences subject to recombination using parsimony, Math. Biosci. 98 (2) (1990) 185–200.

[85] R.G. Beiko, N. Hamilton, Phylogenetic identification of lateral genetic transfer events, BMC Evol. Biol. 6 (2006) 15.

[86] L. Nakhleh, D. Ruths, H. Innan, Gene trees, species trees, and species networks, in: R. Guerra, D. Allison (Eds.), Meta-analysis and Combining Information in Genetics, CRC Press, 2006.

[87] R.G. Beiko, M.A. Ragan, Detecting lateral genetic transfer: a phylogenetic approach, Methods Mol. Biol. 452 (2008) 457–469.

[88] B.L. Allen, M. Steel, Subtree transfer operations and their induced metrics on evolutionary trees, Ann. Comb. 5 (1) (2001) 1–15.

[89] R.G. Beiko, T.J. Harlow, M.A. Ragan, Highways of gene sharing in prokaryotes, Proc. Natl. Acad. Sci. U. S. A. 102 (40) (2005) 14332–14337.

[90] R.G. Beiko, M.A. Ragan, Untangling hybrid phylogenetic signals: horizontal gene transfer and artifacts of phylogenetic reconstruction, Methods Mol. Biol. 532 (2009) 241–256.

[91] Y. Mantri, K.P. Williams, Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities, Nucleic Acids Res. 32 (Database issue) (2004) D55–58.

[92] L. Chen, J. Yang, J. Yu, Z. Yao, L. Sun, Y. Shen, Q. Jin, VFDB: a reference database for bacterial virulence factors, Nucleic Acids Res. 33 (Database issue) (2005) D325–328.

[93] S.H. Yoon, Y.K. Park, S. Lee, D. Choi, T.K. Oh, C.G. Hur, J.F. Kim, Towards pathogenomics: a web-based resource for pathogenicity islands, Nucleic Acids Res. 35 (Database issue) (2007) D395–400.

[94] C.G. Kurland, B. Canback, O.G. Berg, Horizontal gene transfer: a critical view, Proc. Natl. Acad. Sci. U. S. A. 100 (17) (2003) 9658–9662.

[95] D.K. Karaolis, J.A. Johnson, C.C. Bailey, E.C. Boedeker, J.B. Kaper, P.R. Reeves, A *Vibrio cholerae* pathogenicity island associated with epidemic and pandemic strains, Proc. Natl. Acad. Sci. U. S. A. 95 (6) (1998) 3134–3139.

[96] A. Budd, S. Blandin, E.A. Levashina, T.J. Gibson, Bacterial alpha2-macroglobulins: colonization factors acquired by horizontal gene transfer from the metazoan genome? Genome Biol. 5 (6) (2004) R38.

[97] G. Greub, F. Collyn, L. Guy, C.A. Roten, A genomic island present along the bacterial chromosome of the Parachlamydiaceae UWE25, an obligate amoebal endosymbiont, encodes a potentially functional F-like conjugative DNA transfer system, BMC Microbiol. 4 (1) (2004) 48.

[98] G.S. Vernikos, J. Parkhill, Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands, Bioinformatics 22 (18) (2006) 2196–2203.

[99] D.E. Fouts, Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences, Nucleic Acids Res. 34 (20) (2006) 5839–5851.