Published online 4 February 2011

Nucleic Acids Research, 2011, Vol. 39, No. 9 ***e56***
doi:10.1093/nar/gkr059

Towards more robust methods of alien gene detection

Rajeev K. Azad and Jeffrey G. Lawrence*

Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA

Received September 2, 2010; Revised January 7, 2011; Accepted January 20, 2011

ABSTRACT

Because the properties of horizontally-transferred genes will reflect the mutational proclivities of their donor genomes, they often show atypical compositional properties relative to native genes. Parametric methods use these discrepancies to identify bacterial genes recently acquired by horizontal transfer. However, compositional patterns of native genes vary stochastically, leaving no clear boundary between typical and atypical genes. As a result, while strongly atypical genes are readily identified as alien, genes of ambiguous character are poorly classified when a single threshold separates typical and atypical genes. This limitation affects all parametric methods that examine genes independently, and escaping it requires the use of additional genomic information. We propose that the performance of all parametric methods can be improved by using a multiple-threshold approach. First, strongly atypical alien genes and strongly typical native genes would be identified using conservative thresholds. Genes with ambiguous compositional features would then be classified by examining gene context, including the class (native or alien) of flanking genes. By including additional genomic information in a multiple-threshold framework, we observed a remarkable improvement in the performance of several popular, but algorithmically distinct, methods for alien gene detection.

INTRODUCTION

In recent years, tremendous effort has been directed toward understanding the evolutionary dynamics of bacterial genomes. Among their many remarkable features, chimerism arising from the acquisition of genes from unrelated organisms has evoked intense debate (1,2). This phenomenon, termed horizontal or lateral gene transfer (LGT), is now considered a potent force driving bacterial genome evolution (3), and the accumulation of whole genome sequences has allowed its scope to be evaluated with increasing precision. Because change in gene inventory is an historical process, determining genes' evolutionary history depends on indirect evidence imbedded in their sequences. A number of disparate approaches to identify horizontally acquired genes have been proposed, falling mainly into two classes (4,5): phylogenetic methods are based on comparative study of many genomes to find genes with unusually taxonomic distributions, while parametric methods explore a single genome to find genes that are atypical with respect to the majority of genes. Approaches combining these classes are most successful (6).

Parametric methods exploit the unusual compositional features of acquired genes to identify them; while native genes have evolved together, the properties of recently acquired genes will reflect the mutational proclivities of their donor genomes. Thus, alien genes can be identified by measuring their atypicality against the recipient genome background. As a proof of concept, Lawrence and Ochman (7) examined the G+C content of protein-coding genes at their first and third codon positions; if they differed by two standard deviations from their respective genomic means, the gene was deemed likely to be alien. While phylogenetic analysis showed that the majority of putative alien genes were indeed absent from the sister *Salmonella* lineage (8), there were many false negatives and false positives. Karlin suggested dinucleotide composition (9) or overall codon usage patterns (10) could provide more effective statistical determinants, thereby improving performance of alien gene detection algorithms. Here, atypicality was assessed through an odds ratio or difference in codon frequencies, once again comparing a gene's composition to the genomic average. Next-generation methods (Table 1) use increasingly more complex measures, e.g. octamer frequencies (11), to refine the distinction between typical and atypical genes. Multiple-class methods—e.g. the *k*-mean clustering algorithm of Hayes and Borodovsky (12) or the AIC or Jensen–Shannon entropic divergence methods of Azad and Lawrence (13,14)—are even more sophisticated, identifying more than one class of atypical gene in the context of a native gene background by clustering genes in *n*-dimensional parametric space.

*To whom correspondence should be addressed. Tel: +1 412 624 4204; Fax: +1 412 624 4759; Email: jlawrenc@pitt.edu

Table 1. Parametric approaches to alien gene detection (in order of introduction)

Method/software	Discriminant criterion	Measure	Classes	References
GC bias	G+C content	Deviations in G+C content	2	(7,30)
Karlin's dinucleotide	Dinucleotide composition	Difference in dinucleotide relative abundances	2	(9)
Karlin's codon bias	Codon usage bias	Difference in codon frequencies	2	(10)
k-means clustering	Codon usage bias	Kullback–Leibler divergence	2 and 3	(12)
Naïve Bayesian classifier	Oligonucleotide bias	Maximum *a posteriori* probability	Unspecified	(31)
3:1 Genomic signature	Dinucleotide composition at 3:1 codon positions	T^2 distance	2	(32)
Z curve	Biases in GC content, codon usage and amino acid usage	Abrupt variations in cumulative GC profile, deviations in codon and amino acid usage pattern	2	(33)
Horizontal transfer index	Hexamer frequencies	*A posteriori* probability	2	(21)
SIGI	Codon usage bias	Log likelihood ratio	2	(34)
Wn	*k*-mer (k = 6–8) frequencies	Covariance	2	(22)
Wn-SVM	*k*-mer (k = 6–8) frequencies	One-class support vector machines	2	(37)
AIC clustering	Many	Maximum likelihood and Akaike Information Criterion	Many	(13)
Chaos game representation	Tetranucleotide composition	Euclidian distance	2	(35)
IVOM/Alien Hunter	Interpolated octamer frequencies	Kullback–Leibler divergence	2	(11)
JSD clustering	Many	Jensen–Shannon divergence	Many	(14)
Design-Island	Tetranucleotide composition	Difference in tetranucleotide frequencies	2	(36)
MJSD	Dinucleotide and trinucleotide composition	Markovian Jensen–Shannon divergence	2	(6)

Despite these improvements in assessing sequence diversity, the classification of native and foreign genes by parametric measures remains notoriously error-prone (15). The reason these methods fail to achieve high accuracy is related more to the genes' compositional continuum than to the core principles underlying their approaches. The compositional features of acquired and native genes often overlap significantly, so that a simple boundary between atypical and typical genes does not exist (Figure 1A). Despite the development of increasingly sophisticated methods for quantifying atypical character (Table 1), the critical issue of classifying genes with ambiguous compositional features has not been addressed satisfactorily. This limitation reflects the common strategy of parametric methods in balancing type I (false positive) and type II (false negative) classification error within a single-threshold framework. An optimal threshold minimizing both type I and II error is impossible to achieve as the two error parameters share a reciprocal relationship. More conservative thresholds decrease the number of false positives at the expense of increased numbers of false negatives, while relaxed criteria increase the number true positives at the expense of increased false ones. No single-threshold approach can eliminate this trade-off.

We assert that this problem cannot be solved by examining only the compositional characteristics of individual genes. Existing methods treat genes as independent data objects, abandoning potentially useful biological information that may influence their composition such as the strand of transcription (leading or lagging), position relative to the replication origin or, more importantly, position within operons or gene clusters. Because alien genes often arrive as genomic islands (GIs), introducing multiple potentially atypical genes in a single event (16), a weakly atypical gene lying within a cluster of moderately- or strongly-atypical genes is likely to be of foreign origin, whereas a weakly atypical gene embedded within an otherwise unremarkable operon is likely to be native. We posit that gene context and operon structural information can resolve the origin of many compositionally ambiguous genes, as suggested by the results of our (14) and others' (17–19) research. Our results here show a remarkable improvement in the performance of popular parametric methods for alien gene detection when implementing this approach, thus strongly advocating for the use of additional biological information in the development of novel parametric methods.

METHODS

Chimeric artificial genomes

Artificial genomes were modeled on the properties of genuine genomes; sequences were downloaded from NCBI and genes were extracted using the existing annotation. To quantify native variability within genuine genomes, native core genes were extracted from each genome using a gene clustering algorithm based on Akaike Information Criterion (13,20); this process eliminated unusual genes that were acquired by LGT. A *k*-means clustering algorithm was then used to segregate the core genes into distinct classes representing the variability among the core genes. Artificial genomes were generated by generalized hidden Markov models with parameters learned from both these distinct gene classes and from the non-coding sequences (13); the length distribution of intergenic spacers was modeled explicitly. Chimeric artificial genomes were constructed by simulating transfer of one or more contiguous genes from several donor

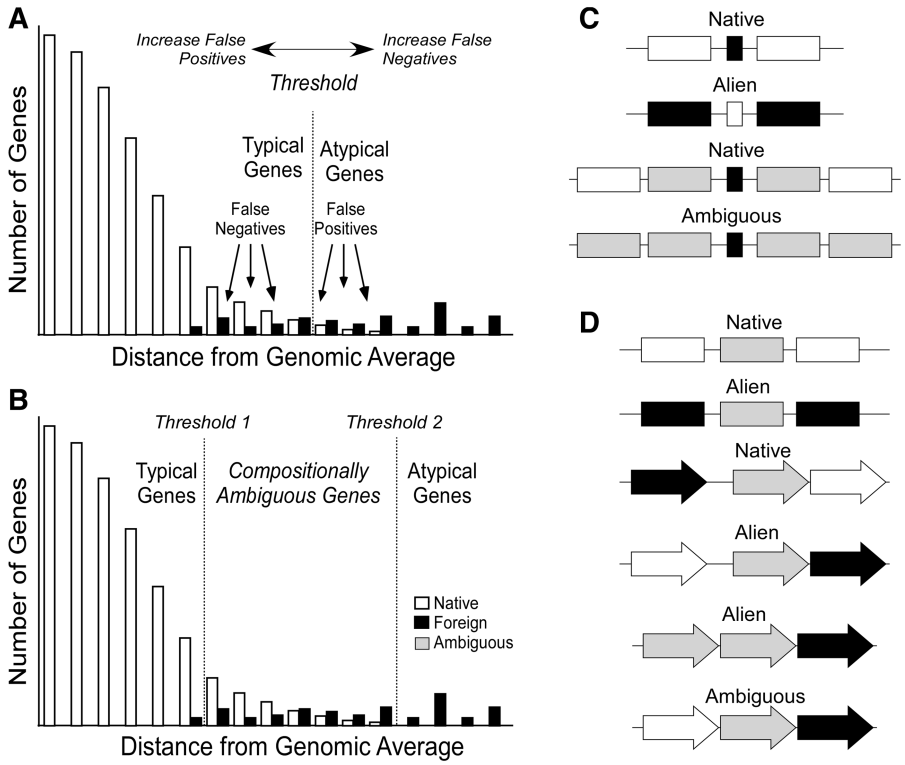

Figure 1. Solving the intrinsic problems with single-threshold approaches. (**A**) In single-threshold approaches, genes are sorted into native and foreign classes according to degree of atypicality. A trade-off between type I and II error results when the threshold is determined because compositional features between native and foreign genes overlap. (**B**) In multiple-threshold approaches, compositionally ambiguous genes are classified as native or foreign based on genomic context. (**C**) Reassignment of short-length genes based on genomic context. (**D**) Assignment of ambiguous genes based on genomic context.

genomes into a recipient genome. We chose an artificial *Escherichia coli* genome as the recipient genome, and acquired genes (~15% of all genes) were provided by 10 donor genomes modeled on *Archaeoglobus fulgidus* (1%), *Bacillus subtilis* (1%), *Deinococcus radiodurans* (2%), *Haemophilus influenzae* Rd (2%), *Methanocaldococcus jannaschii* (1%), *Neisseria gonorrhoeae* (1%), *Ralstonia solanacearum* (2%), *Sinorhizobium meliloti* (2%), *Synechocystis* PCC6803 (1%) and *Thermotoga maritima* (2%).

Single-threshold methods

Discrimination by atypical G+C content was implemented as suggested by Lawrence and Ochman (7); if the G+C content of a gene's first and third codon positions deviated significantly from their respective genomic means, the gene was deemed alien. Dinucleotide bias (Karlin's dinucleotide) was assessed through an odds ratio comparing the frequencies of each gene's dinucleotides to the genomic averages (9). If the deviation exceeded an established threshold, the gene was deemed sufficiently atypical to be classified as alien. Codon usage bias (Karlin's codon bias) was similarly assessed as described (10,14); if the codon usage bias of a gene was significantly different from the bias averaged over a genome, the gene was classified as alien. The horizontal transfer index (HTI) uses fifth-order Markov models to assess the biases in hexamer frequencies in a Bayesian framework (21). A 96-bp window was moved along a genome with in 12-bp steps and its *a posteriori* probability to be part of protein-coding region was computed for the six reading frames. The foreign origin of a gene was inferred by averaging the scores of successive in-frame windows that lie within the gene and are in same coding frame as the gene. If the *a posteriori* probability for a gene to be protein coding according to the Markov model of protein-coding sequences was less than a threshold, the gene was deemed alien. Heptamer frequency bias was assessed by the Wn method (22) using a covariance measure to assess the atypicality of a gene against the genome average.

Gene clustering methods

Azad and Lawrence (14) used Jensen–Shannon divergence (JSD) to measure the compositional difference between two sequences. Gene clustering was accomplished in a hierarchical agglomerative framework. Genes that are most similar (smallest JSD) are grouped first, provided this grouping is deemed statistically significant. The algorithm proceeds recursively, adding genes that are most similar to existing genes and gene clusters until the distinction between resulting gene classes becomes significantly large (clusters are too different to be merged). Thus this method generates multiple native classes (representing stochastic variability) and alien classes (representing distinct

gene donors) using any discriminant criterion as the basis for clustering.

Catalogs of horizontally transferred genes

High-confidence GIs, regions of horizontally-transferred genes that confer specific functions (16), were extracted from both the Islander and tRNAcc databases (23,24). In addition, 453 genes unique to *Salmonella enterica* Typhi CT18 genome were identified as those not found in the genomes of related enteric bacteria including *E. coli* CFT073, *E. coli* W3110, *E. fergusonii* ATCC 35469, *C. koseri* ATCC BAA-895 and *K. pneumoniae* 342 (6). Genes <400 bp in length were not considered.

RESULTS AND DISCUSSION

A generalized, multiple-threshold approach

We took a two-pronged approach to solve the problem of trade-offs between types I and II measurement error. To begin, we abandoned the use of a single-threshold between typical and atypical genes. Rather, genes were classified using two conservative thresholds, each set to minimize either type I or II error. The first threshold was used to identify strongly typical native genes (those with scores less than threshold one in Figure 1B), while the second was used to identify strongly atypical alien genes. As a result, compositionally ambiguous genes lying between these two thresholds were not initially classified as either native or alien, but were reassigned to either the foreign or native class by invoking gene context and operon structural information (Figure 1C and D). This approach can be applied to any metric which is used to assess the atypicality of genes, and thus can be used to refine any existing method for detecting potentially alien, compositionally atypical genes. Genes were classified in seven steps.

(1) The strongly typical native and strongly atypical alien genes were first identified using conservative thresholds. Atypicality was assessed by comparing genes against a reference set of all genes, which served as a surrogate for strictly native genes.
(2) The reference set of all genes was replaced with the set of strongly typical native genes identified above. This set was iteratively refined until convergence.
(3) Before assessing compositionally ambiguous genes, the classes of native and alien genes were refined. Short native genes (<300 bp in length) are often incorrectly assigned to the alien class; here, their apparent atypical character simply reflects stochastic variation. This problem can be resolved by reassigning short, atypical genes to the native class if one or more of their flanking genes are in the native class and no flanking gene is in the foreign class (Figure 1C). If both flanking genes are in ambiguous class, one may examine the next flanking genes sequentially. Similarly, if a short gene in the native class is flanked on both sides by strongly atypical genes, it is moved to the alien class; the logic here is that strongly atypical gene insertions are unlikely to occur on both sides of a single native gene. Otherwise, if none of the neighboring genes (typically 4 and 5) is in the native class, the gene is moved to the ambiguous class.
(4) Next, genes in the ambiguous class can be assigned to either the native or alien classes using the classification of their flanking genes (Figure 1D). Unlike single-threshold approaches, we are essentially ignoring the atypicality score and are relying instead on the potentially more informative contextual data. Ambiguous genes were classified in two steps: (i) if both the flanking genes were in either native or foreign class, the gene was moved to that class; and (ii) if the flanking genes were in different classes, the orientation and intergenic distance between this gene and the flanking genes were examined to determine if it formed an operon with one of its flanking genes; if so, this gene was moved to that class. Here, we are using the presence of an operon as a likely indicator of ancestry, either alien or native. If all three genes formed a likely operon, ambiguity was resolved only if one of the flanking genes was also in the ambiguous class; the gene in question was then moved to the 'non-ambiguous' class of the other flanking gene. If genes flanking an ambiguous gene were both ambiguous and all three genes formed an operon, the adjacent flanking genes were investigated for being part of this operon in either direction; if this search encountered a gene that is a member of one of the high-confidence gene classes, the entire operon was moved to that class.
(5) The short genes in the alien class were examined again. If both flanking genes belonged to the native class, these genes were reassigned to the native class. If only one flanking gene belonged to the native class, the gene was reassigned to the native class only if the other flanking gene was in ambiguous class. If both flanking genes belonged to ambiguous class we examined the genes on both sides (typically up to 10 genes on either side) sequentially; if a native gene was found without encountering a foreign gene, the gene in question was moved to the native class.
(6) Further refinement was achieved by averaging the scores of consecutive genes. Here, one relies not on the weak atypical character of a single gene but on the mean compositional character of consecutive genes. Only if the region in question is of foreign origin, one would expect many consecutive atypical genes.
(7) Finally, the remaining ambiguous genes were assigned to the class, either native or alien, whose class average for the metric was closest to the gene being analyzed. Thus, a solely metric-based approach (assigning the gene to a class based on its score alone) was used only for those genes where genomic context was not informative.

Assessing the multiple-threshold approach

We evaluated our approach by modifying several parametric methods to use multiple-thresholds and assessing

their performance using chimeric artificial genomes wherein the evolutionary 'history' of genes is known. We created a series of genomes with a constant artificial recipient core and alien genes originating from 10 compositionally distinct artificial donor genomes. Alien genes were inserted in clusters of several genes (modeled after the number of contiguous genes on the same strand); critically, the lengths of intergenic spacers were modeled explicitly to allow for operon prediction. Previous studies found intergenic spacer length most informative in predicting operons (25,26); the majority of genes showed spacers ~35 bp in all cases considered here (Supplementary Figures 1–5) and this was used as the threshold for localizing operons. We then assessed atypicality of genes in these chimeric artificial genomes by three widely used approaches, GC bias (nucleotide composition), Karlin's dinucleotide bias and Karlin's codon bias (solid points in Figure 2A). The trade-off between false positives and false negatives was examined by varying the threshold parameters. Significant improvement in the performance of all three parametric methods was observed when the multiple-threshold framework was implemented (open points in Figure 2A). When assessing nucleotide composition, type I error decreases almost 2-fold for a given type II error. Improvements were greater for dinucleotide- and codon bias-based methods, reaching 4- and 6-fold, respectively. These results demonstrate that compositionally ambiguous genes can be placed into alien and native gene classes more accurately when gene context information is considered.

Detecting alien genes in genuine genomes

The use of artificial genomes suggests that multiple-threshold approaches can result in significant improvement in parametric methods for alien gene detection. However, the above results rely on our model for horizontal gene transfer, including the nature of the donors, the number of contiguous genes transferred and the distribution of insertion sites in the recipient genome. To validate these results in genuine genomes, parameters of both the original single-threshold and augmented multiple-threshold algorithms were optimized on artificial genomes before attempting to identify horizontally-transferred genes previously cataloged in four genomes of *E. coli* and *S. enterica*. We cannot report precise type I and II error rates because the evolutionary histories of genes in genuine genomes are not known with certainty. Rather, we assess the relative performance of the single- and multiple-threshold methods in identifying annotated GIs and phylogenetically unique *S. enterica* Typhi genes. Better-performing methods will identify larger numbers of cataloged alien genes using the fewest numbers of predictions of potentially alien genes. This will allow assessment of the performance of the augmented methods without calculation of precise type I and II error rates.

In all cases, the use of multiple thresholds improved the detection of GI-borne genes. For example, 327 GI genes are reported in the *E. coli* O157 genome. When approximately 1725 alien genes were predicted by Karlin's codon bias method, a greater fraction of the GI-borne genes was

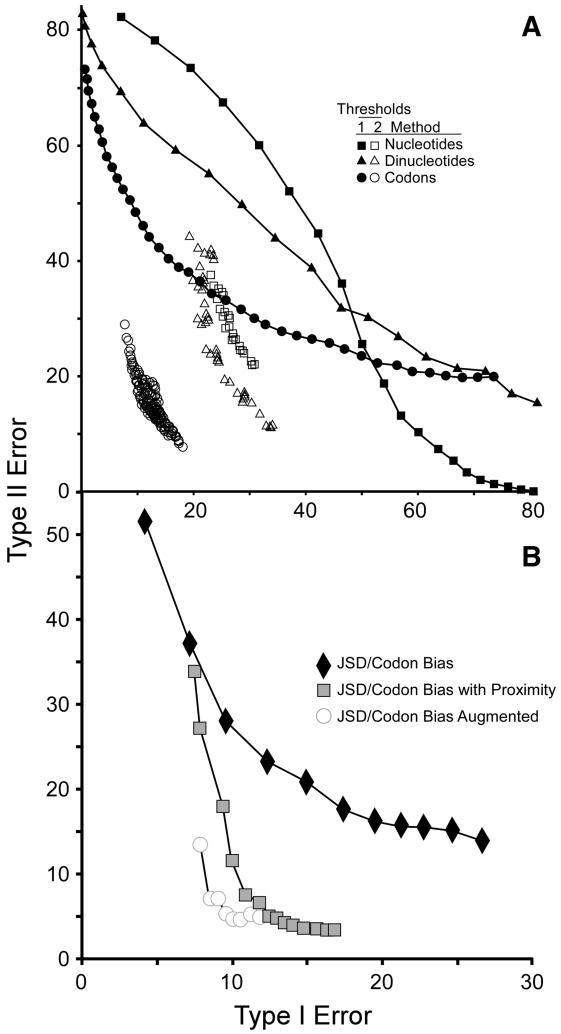

Figure 2. Improvement of threshold methods by including multiple thresholds and positional information. (**A**) Improvement in standard single-threshold methods. Here 'nucleotides', 'dinucleotides' and 'codons' refer to GC bias, Karlin's dinucleotide and Karlin's codon bias method, respectively. (**B**) Improvement in gene clustering methods. The standard Jensen–Shannon divergence (JSD) approach (14) is here annotated 'JSD/codon bias'; the 'proximity' method groups similar genes first in order of their physical distance within a genome, whereas the 'augmented' method uses gene context and operon structure information within a multiple-threshold framework.

detected when multiple thresholds were used (Table 2, lines 1 and 2). Only when stringency was relaxed to predict an additional 520 alien genes (predicting 2245 alien genes) was this level of sensitivity achieved without the use of multiple thresholds (Table 2, line 3), no doubt resulting in far more false positives. Even more dramatic improvements were seen when dinucleotide frequencies were used to detect alien genes (Karlin's dinucleotide); here, the multiple-threshold method detected 83% of the island-borne genes as alien while the single-threshold method could detect only 59% for a comparable number of putatively alien genes. Only when nearly twice as many alien gene predictions were made—amounting to more than half of the genome being classified as alien—did the single-threshold approach identify as many GI-borne

Table 2. Improved performance of position-augmented parametric methods in detecting genomic islands in genuine genomes

Method for detection[a]	*Escherichia coli* O157 Sakai[b]			*Escherichia coli* O157 EDL933[c]			*Escherichia coli* CFT073[d]			*Salmonella enterica* Typhi CT18[e]		
	Predicted	Detected	Percent	Predicted	Detected	Percent	Predicted	Detected	Percent	Predicted	Detected	Percent
Karlin's codon bias	1724	214	65	1715	460	65	1655	451	53	1194	444	49
Karlin's codon bias augmented	1726	246	75	1712	532	75	1645	556	65	1194	574	64
Karlin's codon bias	2245	246	75	2308	532	75	2202	556	65	1681	574	64
Karlin's dinucleotide	1654	184	56	1581	387	55	1671	416	48	1112	373	41
Karlin's dinucleotide augmented	1653	270	83	1580	539	76	1670	623	73	1112	552	61
Karlin's dinucleotide	3106	272	83	2893	544	77	2694	626	73	1858	553	61
HTI/hexamer	1912	238	73	1921	523	74	2163	650	76	1537	605	67
HTI/hexamer augmented	1912	279	85	1920	572	81	2165	716	83	1536	678	75
HTI/hexamer	2725	279	85	2299	572	81	2570	715	83	1901	677	75
Wn/heptamer	1851	203	62	1736	444	63	1857	544	63	1593	621	69
Wn/heptamer augmented	1851	225	69	1735	486	68	1854	608	71	1594	701	78
Wn/heptamer	2176	225	69	2117	486	68	2233	608	71	2127	701	78
JSD/codon bias	1966	190	58	1938	531	74	1599	449	52	1189	457	51
JSD/codon bias augmented	1958	316	96	1928	667	93	1592	745	86	1162	650	72
JSD/codon bias	4050	311	95	3438	659	92	3677	741	86	1902	653	72

[a]Augmented methods use multiple thresholds.
[b]Predicted: total number of putative alien genes predicted. Detected: number of the 327 genes from the Islander database that were among the total number of predicted. Percent: fraction of the database-archived alien genes detected.
[c]Seven hundred and ten genes from the tRNAcc database.
[d]Eight hundred and fifty-nine genes from the tRNAcc database.
[e]Nine hundred and three genes as reported by Vernikos and Parkhill (27).

genes as the multiple-threshold one. Similar results were seen in the three other genomes examined, and when the more sensitive tRNAcc database is used to supply target GIs for identification (Table 2). Therefore, we conclude that the improvement in alien gene detection quantified using artificial genomes remains when the algorithms are applied to genuine genomes.

One could argue that the improvement afforded by the use of positional information is restricted to a more robust identification of large GIs. Therefore, we also created a dataset of 453 genes phylogenetically unique to *S. enterica* Typhi regardless of their residency within a GI. All methods showed improvement when positional information was included (Table 3). The improvement was most pronounced for Karlin's first-generation methods; for example, more than twice as many alien genes were detected by aberrant dinucleotide frequencies when multiple thresholds and positional information were considered (see Supplementary Tables S8 and S9 for the threshold configurations for all methods).

Table 3. Improved performance of position-augmented parametric methods in detecting phylogenetically unique genes in *S. enterica* Typhi CT18 genome

Method for detection[a]	Predicted[b]	Detected[b]	Percent[b]
Karlin's codon bias	1194	210	46
Karlin's codon bias augmented	1194	303	67
Karlin's codon bias	1956	303	67
Karlin's dinucleotide	1112	120	26
Karlin's dinucleotide augmented	1112	264	58
Karlin's dinucleotide	2059	264	58
HTI/hexamer	1537	321	71
HTI/hexamer augmented	1536	359	79
HTI/hexamer	1829	359	79
Wn/heptamer	1593	367	81
Wn/heptamer augmented	1594	389	86
Wn/heptamer	1930	389	86
JSD/codon bias	1189	274	60
JSD/codon bias augmented	1162	320	71
JSD/codon bias	1501	322	71

[a]Augmented methods use multiple thresholds.
[b]Predicted: total number of alien gene predicted. Detected: number of the 453 unique CT18 genes (those not found in the genomes of related enteric bacteria including *E. coli* CFT073, *E. coli* W3110, *E. fergusonii* ATCC 35469, *C. koseri* ATCC BAA-895 and *K. pneumoniae* 342) that were among the total number of predicted. Percent: fraction of the database-archived alien genes detected.

Assessing the stepwise approach to alien gene detection

As outlined above, gene-context information was assessed in seven steps. We assessed the differential contributions from each step in improving Karlin's dinucleotide method (Tables 4 and 5). Gene-context information (steps 4a and 6) was found most effective, contributing over 16% of total ~26% improvement in alien gene detection in *E. coli* O157 (Table 4). The remaining 10% improvement came from the application of other steps including operon structural information (3%, step 4b), short gene corrections (3%, steps 3 and 5) and metric-based ambiguous gene assignment (2%, step 7). Similar trend was observed in *E. coli* O157 EDL933 and *E. coli* CFT073.

Table 4. Relative performance of the Karlin's dinucleotide method versus its augmented version following the seven steps used in augmenting its classification ability

Step	Method[a]	*Escherichia coli* O157 Sakai[b]					*Escherichia coli* O157 EDL933[c]					*Escherichia coli* CFT073[d]					*Salmonella enterica* Typhi CT18[e]				
		Amb.	Native	Alien	TP	SN	Amb.	Native	Alien	TP	SN	Amb.	Native	Alien	TP	SN	Amb.	Native	Alien	TP	SN
1	Augmented	2881	2061	418	59	18.0	2800	2053	455	128	18.0	2738	2018	622	142	16.5	2315	1740	338	119	13.1
	Standard	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	Augmented	2749	2135	476	67	20.4	2699	2108	501	151	21.2	2641	2072	665	169	19.6	2208	1785	400	152	16.8
	Standard	–	4883	477	65	19.8	–	4807	501	138	19.4	–	4713	665	155	18.0	–	3993	400	137	15.1
3	Augmented	2884	2132	344	64	19.5	2826	2124	358	141	19.8	2784	2147	447	153	17.8	2318	1789	286	134	14.8
	Standard	–	5014	346	53	16.2	–	4949	359	100	14.0	–	4931	447	91	10.5	–	4106	287	97	10.7
4a	Augmented	2297	2649	414	85	25.9	2226	2654	428	187	26.3	2208	2665	505	181	21.0	1797	2242	354	177	19.6
	Standard	–	4945	415	59	18.0	–	4879	429	114	16.0	–	4872	506	104	12.1	–	4038	355	123	13.6
4b	Augmented	2442	2409	509	85	25.9	2419	2411	478	191	26.9	2420	2423	535	188	21.8	1958	2041	394	208	23.0
	Standard	–	4850	510	69	21.1	–	4839	478	131	18.4	–	4843	535	116	13.5	–	3999	394	136	15.0
4a and b	Augmented	1855	2926	579	106	32.4	1819	2941	548	237	33.3	1844	2941	593	216	25.1	1437	2494	462	251	27.7
	Standard	–	4781	579	73	22.3	–	4760	548	150	21.1	–	4785	593	131	15.2	–	3930	463	162	17.9
5	Augmented	1855	2928	577	106	32.4	1819	2943	546	237	33.3	1844	2946	588	216	25.1	1437	2497	459	250	27.6
	Standard	–	4783	577	73	22.3	–	4762	546	150	21.1	–	4790	588	131	15.2	–	3934	459	161	17.8
6	Augmented	1447	2928	985	185	56.5	1506	2943	859	365	51.4	1313	2946	1119	453	52.7	1225	2497	671	384	42.5
	Standard	–	4373	987	113	34.5	–	4448	860	216	30.4	–	4259	1119	275	32.0	–	3722	671	236	26.1
7	Augmented	0	3707	1653	270	82.5	0	3728	1580	539	75.9	0	3708	1670	623	72.5	0	3281	1112	552	61.1
	Standard	–	3706	1654	184	56.2	–	3727	1581	387	54.5	–	3707	1671	416	48.4	–	3281	1112	373	41.3

[a]Augmented methods use multiple thresholds.
[b]Predicted: total number of putative alien genes predicted. Detected: number of the 327 genes from the Islander database that were among the total number of predicted. Percent: fraction of the database-archived alien genes detected.
[c]Seven hundred and ten genes from the tRNAcc database.
[d]Eight hundred and fifty-nine genes from the tRNAcc database.
[e]Nine hundred and three genes as reported by Vernikos and Parkhill (27).
Amb., ambiguous.

Table 5. Relative performance of the Karlin's dinucleotide method versus its augmented version in detecting phylogenetically unique genes in *S. enterica* Typhi CT18 genome following the seven steps used in augmenting the method's classification ability

Step	Method	Ambiguous	Native	Alien	TP	SN
1	Augmented	2315	1740	338	13	2.8
	Standard	–	–	–	–	–
2	Augmented	2208	1785	400	30	6.6
	Standard	–	3993	400	19	4.1
3	Augmented	2318	1789	286	30	6.6
	Standard	–	4106	287	7	1.5
4a	Augmented	1797	2242	354	52	11.4
	Standard	–	4038	355	14	3.0
4b	Augmented	1958	2041	394	68	15.0
	Standard	–	3999	394	18	3.9
4a and b	Augmented	1437	2494	462	90	19.8
	Standard	–	3930	463	31	6.8
5	Augmented	1437	2497	459	90	19.8
	Standard	–	3934	459	30	6.6
6	Augmented	1225	2497	671	158	34.8
	Standard	–	3722	671	62	13.6
7	Augmented	0	3281	1112	264	58.2
	Standard	–	3281	1112	120	26.4

In *S. enterica* Typhi CT18 where annotated alien genes originate from islands which by definition (27) can have as few as two genes (Table 4) or are independent of the island structure (Table 5), the contribution from operon structural information is somewhat more pronounced. Particularly with the later (Table 5) where 32% improvement was observed in detection of phylogenetically unique genes, the contribution from gene-context information was ~11% while that from operon structural information was ~4.5%. Notably over 10% improvement came from assignment of ambiguous genes based on their distance from native and alien cluster centers (step 7). The improvement from this step was also observed for island originated genes, although less pronounced (4.3% for *E. coli* O157, 0.4% for *E. coli* O157 EDL933, 3.4% for *E. coli* CFT073 and 3.4% for *S. enterica* Typhi CT18). The clusters generated following the preceding steps that incorporate gene context and operon structural information are indeed more helpful in assignment of 'left over' ambiguous genes than the clusters that could be generated using compositional biases alone (i.e. following steps 1 and 2, see Supplementary Table S1). In particular for *E. coli* CFT073 and *S. enterica* Typhi CT18), more alien genes were identified with fewer predictions (Supplementary Table S1). Importantly, variations in conservative thresholds do not impact the augmented method's performance (Supplementary Table S2).

The use of gene-context information also improves moving-window approaches to alien gene detection. Karlin (9) showed that dinucleotide frequencies within successive 50-kb windows represent the genomic signature of an organism and thus can be used for distinguishing alien regions from the native ones. When Karlin's method is used in its moving window formulation, the augmented version continues to outperform the standard approach. In addition to predicting a greater fraction of known alien genes for a given number of total predictions (Supplementary Table S3), the augmented method was far less sensitive to changes in its threshold parameters. For the augmented method, as the total number of alien gene predictions increased, the fraction of known alien genes predicted also increased (Supplementary Tables S4 and S5); this was not true for the non-augmented method, wherein increase in total numbers of alien gene predictions led to unpredictable increase in the detection of known alien genes, apparently an undesirable behavior. While larger windows can help detect longer GIs, they are prone to missing shorter islands. On the other hand, while smaller windows yield better resolution, they meet the same difficulty in reconstructing an island structure as the individual genes (which can be interpreted as 'smaller' windows of variable size). Using strategies similar to one proposed here can help resolve this predicament, reconstructing not just the longer acquisitions but also rendering the detection resolution to as few as one or two alien genes.

Improvements in more advanced algorithms

One may argue that metrics relying on dinucleotide frequencies and codon usage bias alone simply lack sophistication in measuring the compositional differences between genes, and that more advanced techniques would eliminate compositionally ambiguous genes by identifying native and alien genes more robustly. To explore this possibility, we implemented approaches using more advanced algorithms; the HTI method (21) assesses hexamer frequencies and the Wn method (22) can examine oligomers of length six to eight (we implemented heptamers). Despite the algorithmic sophistication of these methods, the same problems remained: compositionally ambiguous genes were not sorted robustly into native and alien gene classes and the use of positional information again resulted in a significant decrease in error rates (Tables 2 and 3). Therefore, we posit that a high degree of computational sophistication alone does not eliminate compositionally ambiguous genes, and additional information must be used to identify the potentially alien genes among them.

Application to clustering methods

While native genes are similar to each other, alien genes are most often described as being 'not native', rather than possessing properties of their own. But, owing to their arrival on GIs from a non-random selection of donor genomes, alien genes may be identified by their similarity to each other as much as by their dissimilarity to native genes. Clustering methods use this similarity among sets of alien genes to identify them and have been implemented using several different approaches (13,14). While these methods also offer improvement over single-threshold methods, the use of multiple clusters alone does not eliminate the problem of compositionally ambiguous genes; such genes would still not be assigned to any single cluster robustly.

We previously implemented a two-tier approach to use genomic information to improve the performance of

clustering methods (14). We examined gene-context information to reassign genes between clusters based on the cluster assignment of their flanking genes. To begin, similar genes were grouped by the JS clustering method using conservative significance thresholds, leading to a large number of robust clusters. Positional information was used to merge clusters with genes that were physically associated within a genome. Gene-context information was again invoked to refine the final set of gene clusters, moving genes between clusters if flanking genes were robust members of a different cluster. When tested on chimeric artificial genomes, this two step procedure minimized the classification errors well in comparison to the standard JS method which assigns genes into different clusters by invoking JS distance alone (14).

However, the efficiency of this approach greatly depends on the selection of thresholds. In our earlier study, the optimal performance was achieved within the threshold range 0.2–0.3. For example, the optimal threshold for *E. coli* K12 was found to be 0.2 while that for *E. coli* O157 was around 0.168 (Supplementary Table S10). Further, slight variations from the optimal-threshold range may cause the unwanted demerger of mostly smaller native clusters from the largest (native) cluster, or unwanted merger of almost all smaller clusters to the largest cluster apparently induced by the recursive inclusion of incorrect ('mixed' or 'alien') clusters into the largest one (Supplementary Table S10). One can address this issue through heuristics, for example, by examining the relative change in cluster size in the process of merger; alternatively, to eliminate this subjectivity, a separate clustering approach could be pursued as described below. Here, we propose to invoke gene-context information to group similar genes from the initial steps of the clustering procedure, in contrast to using this information in a post-processing step. We first grouped only contiguous genes that were similar to each other, and then recursively grouped the proximal gene clusters with similar compositional bias in the hypothesis testing framework (Figure 2B, 'JSD/codon bias proximity'). Significant improvement in performance was observed when compared to our original approach whereby the most similar genes located at any genomic position were grouped first (Figure 2B, 'JSD/codon bias'). Further improvements were gained when we reintroduced this approach in a multiple-threshold framework (Figure 2B, 'JSD/codon bias augmented'). When compared to the use of positional information in refining cluster composition, this approach was far less sensitive to variation in threshold parameters (Supplementary Tables S10 and S11). Further this approach also raised the accuracy bar significantly (compare 96% *E. coli* O157 island gene detection when 35% of total 5360 genes were predicted alien with 75% detection from the previous approach that predicted 31% as alien, Supplementary Tables S10 and S11).

Testing clustering methods on genuine genomes

We again utilized the set of genuine GIs to evaluate the efficacy of both position-aware JSD clustering approaches relative to the original, position-blind approach. As seen for single-threshold approaches, the use of positional information required far fewer alien gene predictions to achieve comparable sensitivity in detecting both GI-borne and phylogenetically unique *S. enterica* Typhi genes (Tables 2 and 3). For equivalent number of predictions, the sensitivity increased by 19% for *E. coli* O157 EDL933 and up to 38% for *E. coli* O157 Sakai, the greatest improvement observed (Table 2). A remarkably large improvement in the accuracy of this method clearly demonstrates the effectiveness of gene-context information in grouping compositionally similar genes. Efficient grouping of genes is critical to the success of this class of methods which have been shown to outperform single-threshold methods for alien gene detection consistently (14).

We also assessed the performance of augmented versions of parametric methods on the HGT-DB database (28), which is more comprehensive in its inclusion of suspected alien genes. However, this database was compiled using parametric methods including G+C bias and codon usage bias, and so is not ideal for assessing the methodologies being presented here. We observed elevation of accuracy for each method, though it was more remarkable for Karlin's dinucleotide and codon bias methods (Supplementary Tables S6 and S7). These results clearly demonstrate the utility and the promise of our proposed approach in all techniques of alien gene detection.

We also assessed the efficacy of combining augmented methods. As has been seen for standard methods (29), sets of alien genes predicted by more than one method include fewer false positives (Table 6). Proper strategy for combining predictions can help in achieving high sensitivity at the cost of negligible additional false positives. This is apparent from the performance by the unison of JS method predictions with the predictions shared among at least three of the other four methods; this approach clearly outperforms the rather naïve approach of combining predictions from all five methods, identifying more island genes at lesser total predictions (Table 6). Notably the use of positional information and the use of multiple methods for detecting alien genes are complementary in their ability to reduce errors in alien gene detection.

CONCLUSIONS

Identifying horizontally-acquired genes has remained a challenging task despite significant progress made in recent years, partly because of the large spectrum of variability reflected in the compositional properties of both native and acquired genes. Parametric methods strive to balance type I and II error of misclassification by selecting an appropriate threshold, yet this approach is inherently ineffective in classifying a large fraction of compositionally ambiguous genes; we assert that this problem cannot be addressed by invoking parametric methods alone. Here we show that by incorporating gene context and operon structure information within the model framework of parametric techniques, the performance of parametric methods can be improved substantially. This necessitated usage of multiple thresholds as opposed to one threshold

Table 6. Performance in detecting island borne genes by the combined methods

Predicted by at least[a]	Genome analyzed											
	Escherichia coli O157 Sakai[b]			*Escherichia coli* O157 EDL933[c]			*Escherichia coli* CFT073[d]			*Salmonella enterica* Typhi CT18[e]		
	Predicted	Detected	Percent	Predicted	Detected	Percent	Predicted	Detected	Percent	Predicted	Detected	Percent
1 of 5 methods	3150	326	99.6	3168	705	99.2	3228	821	95.5	2327	802	88.8
2 of 5 methods	2298	321	98.1	2241	684	96.3	2242	765	89.0	1604	738	81.7
3 of 5 methods	1731	292	89.2	1690	610	85.9	1712	692	80.5	1231	675	74.7
4 of 5 methods	1259	243	74.3	1178	494	69.5	1168	578	67.2	904	568	62.9
5 of 5methods	692	154	47.0	598	303	42.6	576	392	45.6	532	372	41.1
1 of 4 methods or JSD	3150	326	99.6	3168	705	99.2	3228	821	95.5	2327	802	88.8
2 of 4 methods or JSD	2502	324	99.0	2491	691	97.3	2436	787	91.6	1729	761	84.2
3 of 4 methods or JSD	2205	322	98.4	2206	685	96.4	2086	767	89.2	1454	720	79.7
4 of 4 methods or JSD	2049	317	96.9	2027	673	94.7	1805	757	88.1	1280	686	75.9

[a]Augmented methods using multiple thresholds. '5 methods' denotes augmented versions of Karlin's codon bias, Karlin's dinucleotide, HTI/hexamer, Wn/heptamer and JSD/codon bias. '4 methods' denotes augmented versions of Karlin's codon bias, Karlin's dinucleotide, HTI/hexamer and Wn/heptamer.
[b]Predicted: total number of putative alien genes predicted. Detected: number of the 327 genes from the Islander database that were among the total number of predicted. Percent: fraction of the database-archived alien genes detected.
[c]Seven hundred and ten genes from the tRNAcc database.
[d]Eight hundred and fifty-nine genes from the tRNAcc database.
[e]Nine hundred and three genes as reported by Vernikos and Parkhill (27).

to classify genes based on their composition, genomic context and intergenic spacer length. The improvements we observe demonstrate the importance of using additional biological information within more flexible, multiple-threshold model frameworks for deciphering the evolutionary history of bacterial genes. While the emergence of more accurate, sophisticated methods for alien gene detection is highly desired, we propose that future efforts should be focused on integrating diverse evidence encoded in genomes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Grant GM078092 from the US National Institutes of Health. Funding for open access charge: NIH (GM078092).

Conflict of interest statement. None declared.

REFERENCES

1. Doolittle,W.F. (2000) Uprooting the tree of life. *Sci. Amer.*, **282**, 90–95.
2. Gogarten,J.P., Doolittle,W.F. and Lawrence,J.G. (2002) Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.*, **19**, 2226–2238.
3. Ochman,H., Lawrence,J.G. and Groisman,E. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
4. Ragan,M.A. (2001) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Lett.*, **201**, 187–191.
5. Ragan,M.A. (2001) Detection of lateral gene transfer among microbial genomes. *Curr. Opin. Genet. Dev.*, **11**, 620–626.
6. Arvey,A.J., Azad,R.K., Raval,A. and Lawrence,J.G. (2009) Detection of genomic islands via segmental genome heterogeneity. *Nucleic Acids Res.*, **37**, 5255–5266.
7. Lawrence,J.G. and Ochman,H. (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl Acad. Sci., USA*, **95**, 9413–9417.
8. Lawrence,J.G. and Ochman,H. (2002) Reconciling the many faces of lateral gene transfer. *Trends Microbiol.*, **10**, 1–4.
9. Karlin,S. (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.*, **1**, 598–610.
10. Karlin,S., Mrazek,J. and Campbell,A.M. (1998) Codon usages in different gene classes of the *Escherichia coli* genome. *Mol. Microbiol.*, **29**, 1341–1355.
11. Vernikos,G.S. and Parkhill,J. (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics*, **22**, 2196–2203.
12. Hayes,W.S. and Borodovsky,M. (1998) How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res.*, **8**, 1154–1171.
13. Azad,R.K. and Lawrence,J.G. (2005) Use of artificial genomes in assessing methods for atypical gene detection. *PLoS Comp. Biol.*, **1**, e56.
14. Azad,R.K. and Lawrence,J.G. (2007) Detecting laterally transferred genes: use of entropic clustering methods and genome position. *Nucleic Acids Res.*, **35**, 4629–4639.
15. Koski,L.B., Morton,R.A. and Golding,G.B. (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol. Biol. Evol.*, **18**, 404–412.

16. Dobrindt,U., Hochhut,B., Hentschel,U. and Hacker,J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.*, **2**, 414–424.
17. Cortez,D., Forterre,P. and Gribaldo,S. (2009) A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. *Genome Biol.*, **10**, R65.
18. Chiapello,H., Bourgait,I., Sourivong,F., Heuclin,G., Gendrault-Jacquemard,A., Petit,M.A. and El Karoui,M. (2005) Systematic determination of the mosaic structure of bacterial genomes: species backbone versus strain-specific loops. *BMC Bioinformatics*, **6**, 171.
19. Chiapello,H., Gendrault,A., Caron,C., Blum,J., Petit,M.A. and El Karoui,M. (2008) MOSAIC: an online database dedicated to the comparative genomics of bacterial strains at the intra-species level. *BMC Bioinformatics*, **9**, 498.
20. Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Contrl.*, **AC-19**, 716–723.
21. Nakamura,Y., Itoh,T., Matsuda,H. and Gojobori,T. (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat. Genet.*, **36**, 760–766.
22. Tsirigos,A. and Rigoutsos,I. (2005) A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res.*, **33**, 922–933.
23. Mantri,Y. and Williams,K.P. (2004) Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res.*, **32**, D55–D58.
24. Sridhar,J. and Rafi,Z.A. (2007) Identification of novel genomic islands associated with small RNAs. *In Silico Biol.*, **7**, 601–611.
25. Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18(Suppl. 1)**, S329–S336.
26. Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. (2000) Operons in *Escherichia coli:* genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.
27. Vernikos,G.S. and Parkhill,J. (2008) Resolving the structural features of genomic islands: a machine learning approach. *Genome Res.*, **18**, 331–342.
28. Garcia-Vallve,S., Guzman,E., Montero,M.A. and Romeu,A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.*, **31**, 187–189.
29. Langille,M.G. and Brinkman,F.S. (2009) IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics*, **25**, 664–665.
30. Lawrence,J.G. and Ochman,H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.*, **44**, 383–397.
31. Sandberg,R., Winberg,G., Branden,C.I., Kaske,A., Ernberg,I. and Coster,J. (2001) Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res.*, **11**, 1404–1409.
32. Hooper,S.D. and Berg,O.G. (2002) Detection of genes with atypical nucleotide sequence in microbial genomes. *J. Mol. Evol.*, **54**, 365–375.
33. Zhang,R. and Zhang,C.T. (2004) 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**, 612–622.
34. Merkl,R. (2004) SIGI: score-based identification of genomic islands. *BMC Bioinformatics*, **5**, 22.
35. Dufraigne,C., Fertil,B., Lespinats,S., Giron,A. and Deschavanne,P. (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res.*, **33**, e6.
36. Chatterjee,R., Chaudhuri,K. and Chaudhuri,P. (2008) On detection and assessment of statistical significance of Genomic Islands. *BMC Genomics*, **9**, 150.
37. Tsirigos,A. and Rigoutsos,I. (2005) A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. *Nucleic Acids Res.*, **33**, 3699–3707.