BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 18 2006, pages 2196–2203
doi:10.1093/bioinformatics/btl369

Genome analysis

Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands

Georgios S. Vernikos* and Julian Parkhill

The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Received on May 3, 2006; revised on June 22, 2006; accepted on July 3, 2006
Advance Access publication July 12, 2006
Associate Editor: John Quackenbush

ABSTRACT

Motivation: There is a growing literature on the detection of Horizontal Gene Transfer (HGT) events by means of parametric, non-comparative methods. Such approaches rely only on sequence information and utilize different low and high order indices to capture compositional deviation from the genome backbone; the superiority of the latter over the former has been shown elsewhere. However even high order *k*-mers may be poor estimators of HGT, when insufficient information is available, e.g. in short sliding windows. Most of the current HGT prediction methods require pre-existing annotation, which may restrict their application on newly sequenced genomes.

Results: We introduce a novel computational method, Interpolated Variable Order Motifs (IVOMs), which exploits compositional biases using variable order motif distributions and captures more reliably the local composition of a sequence compared with fixed-order methods. For optimal localization of the boundaries of each predicted region, a second order, two-state hidden Markov model (HMM) is implemented in a change-point detection framework. We applied the IVOM approach to the genome of *Salmonella enterica* serovar Typhi CT18, a well-studied prokaryote in terms of HGT events, and we show that the IVOMs outperform state-of-the-art low and high order motif methods predicting not only the already characterized *Salmonella* Pathogenicity Islands (SPI-1 to SPI-10) but also three novel SPIs (SPI-15, SPI-16, SPI-17) and other HGT events.

Availability: The software is available under a GPL license as a standalone application at http://www.sanger.ac.uk/Software/analysis/alien_hunter

Contact: gsv@sanger.ac.uk

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Genomic regions of 'alien' origin are present in various forms in the prokaryotic genome. These include large inserts of DNA that contain a number of functionally related genes putatively acquired by horizontal transfer, often referred to as genomic islands (GIs). The location of these islands frequently correlates with distinct sequence elements such as stable RNA genes, direct/inverted repeats (DR/IRs) and mobility genes. Other genomic elements with some of the signatures of GIs include bacteriophages, plasmids, extracellular polysaccharide biosynthesis loci (Hacker and Kaper, 2000; Zhang *et al*., 1997) and other gene clusters under specific constraints; these may or may not be recently horizontally acquired. Pathogenicity islands (PAIs) constitute a specific type of GIs that provide virulence properties to bacterial strains. The concept of PAI was established in the late 1980s by Jörg Hacker and colleagues studying the virulence properties of uropathogenic strains of *Escherichia coli* (UPEC) 536 and J96 (Hacker *et al*., 1990; Knapp *et al*., 1986). Examples of other types of GIs involve the symbiosis island in *Mesorhizobium loti* (Sullivan and Ronson, 1998), the metabolic island in *Salmonella senftenberg* and the antibiotic resistance island in *Staphylococcus aureus*. Using models of amelioration to estimate the time of HGT events it has been previously shown (Lawrence and Ochman, 1997) that the *E.coli* chromosome contains >600 kb of horizontally transferred, protein-coding DNA.

It is often assumed that at the time of integration GIs reflect the sequence composition of the donor genome (although other reasons for the observed bias may apply); based on this principle several indices have been exploited to capture deviation at various levels from the host genome composition. It should be noted that those indices will perform badly if the composition of the donor and the recipient genome sequence is similar. Furthermore if the age of the HGT event is reasonably old then owing to the amelioration process (Lawrence and Ochman, 1997) the composition of GIs will be more similar to that of the host, rendering their prediction by means of parametric methods non-trivial. Often a combination of more than one index can be used for a more efficient identification of 'alien' regions. For example both Lawrence and Ochman (1997) and Karlin *et al*. (1998) utilized codon bias and the Codon Adaptation Index (CAI) (Sharp and Li, 1987) to identify atypical regions. In a similar multi-index approach, Karlin (2001) applied the G + C content, dinucleotide frequency difference (δ^* difference), codon bias and amino acid bias to detect alien gene clusters. Most of these indices cause overlapping peaks predicting the same atypical regions; however, there are cases in which one or more indices might perform poorly in the detection of compositionally deviating regions (see Figure 1c therein).

Yoon *et al*. (2005) combined sequence similarities and composition abnormalities to predict PAIs rather than GIs in general. Regions containing both atypical composition and PAI homologous regions are reported as candidate PAIs. Garcia-Vallve *et al*. (2003) developed a database, 'HGT-DB', of predicted horizontally transferred genes, using G + C content, codon and amino acid usage,

*To whom correspondence should be addressed.

and gene position analysis. Mantri and Williams (2004) developed an algorithm, 'Islander', exploiting the principle that islands tend to be preferentially integrated within stable RNA genes. 'Islander' produces a list of tRNA and tmRNA genes and uses each as a query for a BLAST search. IslandPath (Hsiao *et al*., 2003) is another web-based suite for the prediction of GIs utilizing G + C content, δ^* difference, RNA and mobility gene information; annotation features are retrieved from public resources. Tsirigos and Rigoutsos (2005) and Sandberg *et al*. (2001) utilized higher order templates to overcome the weak discrimination power of lower order ones. Both papers provide data in favour of the higher order templates, with the optimal template size found to be 8–9 nt. In the following section we describe a novel method for the prediction of putative horizontally transferred regions by means of variable order compositional distributions. This approach does not require pre-existing annotation and can, therefore, be applied directly to newly sequenced genomes. Moreover we discuss the implementation of region-specific two-state, second order HMMs to optimize the localization of the boundaries of the predicted regions. Finally we describe the pipeline followed to obtain a test dataset of manually curated putative horizontally transferred regions by applying the reciprocal FASTA (Pearson, 1990) approach.

2 METHODS

2.1 Interpolated variable order motifs

Usage of low order compositional indices may not provide sufficient discrimination of regions with atypical composition (bias in motifs of higher order e.g. 6mers). The total number of all different possible motifs increases exponentially with the size k of the motifs. For k-mers of size k there are 4^k different possible k-mers (parameters). Consequently utilizing high order motifs is more likely to capture deviation from the genome background compositional distribution, as long as there is enough data to produce reliable probability estimates. However for high order motifs, e.g. 8mers in a sliding window of 5 kb, ~60 000 out of 65 536 different possible 8mers will have an observed frequency of zero. Even for 8mers of non-zero frequency the information may not be enough to provide reliable estimates of the local sequence composition of a region, e.g. most 8mers will be present only once in a 5 kb window. An IVOM approach overcomes this problem, implementing variable order k-mers, 'preferring' information derived from high order motifs, but when this information is insufficient, relying more on lower order motifs. Let B be the DNA alphabet, defined as: $B = \{a, t, g, c\}$. In an IVOM approach all k-mers with $1 \leq k \leq 8$ are exploited. Each k-mer can be seen as a linear combination of its component lower order motifs including itself. In a first step, for each k-mer m in the sequence S, its observed frequency $P_m(S)$ is calculated as follows:

$$P_m(S) = \frac{A_m(S)}{N - k + 1}, \tag{1}$$

where $A_m(S)$ is the number of occurrences of m in the sequence S and N is the size of S. Generally a high order motif occurs less frequently (small number of occurrences) in a sequence compared with motifs of lower order, given that the total number of all different possible motifs is higher in the first case. In order to use in combination the different order k-mers, both the difference in the number of occurrences and in the total number of different possible k-mers have to be taken into account. For this reason for each k-mer m in the sequence S, a weight $W_m(S)$ is calculated as follows:

$$W_m(S) = \frac{A_m(S) \cdot |B|^k}{\sum_{j=1}^{8} A_j(S) \cdot |B|^j}, \tag{2}$$

where $|B|^k$ denotes the total number of all different possible motifs of size k. In this framework a high and a low order motif have equal chances of producing bias given that both number of counts and dimensionality have been taken into account. For example if the number of occurrences of a 3mer and a 5mer is 128 and 8, respectively, an IVOM approach treats the two k-mers as equally reliable estimates of the local sequence composition of a region. Having computed the weights for each k-mer, in a second step the IVOM frequency for each k-mer m in the sequence S is calculated as follows:

$$\mathrm{IVOM}(S, m) = \begin{cases} W_m(S) \cdot P_m(S) + [1 - W_m(S)] \cdot \mathrm{IVOM}(S, m_{2,|m|}) \text{ if } |m| \geq 2 \\ W_m(S) \cdot P_m(S) \text{ if } |m| = 1, \end{cases} \tag{3}$$

where $m_{2,|m|}$ denotes the interpolated substring starting at position 2 and ending at position $|m|$ in k-mer m. Using the above equation, it is possible for the observed frequencies of all the interpolated motifs to be combined linearly in such a way that if high order motifs are reliable (sufficient counts) estimates of the local sequence composition, then the corresponding $W_m(S)$ weight will be high enough for the contribution of the lower motifs to be ignored and vice versa. A similar equation is implemented by Salzberg *et al*. (1998) in GLIMMER, a widely used gene prediction method. In GLIMMER, however, the above equation is used in a Markov model-based context, i.e. interpolated Markov models (IMMs). Moreover GLIMMER uses two different criteria (number of occurrences and predictive value) in order to calculate the weight for each k-mer. We chose Equation (2) instead of the aforementioned two criteria to calculate the weights, in order to avoid incorporation of arbitrary threshold values.

2.2 Relative entropy for compositional deviation

In order to predict putatively horizontally transferred regions in microbial genomes, we assume that each genome exhibits a reasonably constant[1] background sequence composition that is the result of the same mutational pressure applied throughout its sequence. Consequently regions of 'atypical' composition within a genome are likely to have been horizontally acquired from a donor genome of different composition. In order to detect compositionally deviating regions, we apply a sliding window approach over raw genomic sequence. In this framework the analysis of atypical regions can be implemented both on annotated and newly sequenced genomes without any level of annotation. In order to converge over the optimal sliding window size l, we experimented on different l values, implementing a ROC analysis and we found that the greatest area under the curve (AUC) for $k \leq 8$ is achieved when the sliding window size and step is set to 5 and 2.5 kb, respectively. It should be noted that increasing the order of the utilized k-mers causes the optimal window size to increase too (Wu *et al*., 2005). The same authors concluded that for symmetric Kullback–Leibler discrepancy as a similarity measure and $2550 \leq l \leq 4950$ the optimal word size k is 8. The step of the sliding window is set to 2.5 kb. However, increasing the step size too much will cause uncertainty about the real boundaries of the predicted 'atypical' regions. We will discuss in the next section how we can overcome this issue. Both for the sliding window w and the genome G we build a compositional vector, defined as

$$\overrightarrow{\mathbf{IVOM}(\boldsymbol{S}, \boldsymbol{m})} = \{\mathrm{IVOM}(S, m) \mid m \in B^8\}. \tag{4}$$

This vector extends over all $(|B|^8)$ the different possible 8mers m in the sequence S. In order to compare the two vectors (of w and G) a distance similarity measure has to be applied. In this study, we implement the relative entropy (Kullback–Leibler distance), defined as follows:

$$d_G(w) = \sum_{m \in B^8} \mathrm{IVOM}(w, m) \log_2 \frac{\mathrm{IVOM}(w, m)}{\mathrm{IVOM}(G, m)}. \tag{5}$$

Implementing Equation (5), a sequence region of 'atypical' composition will have high relative entropy while native-typical regions will have relative entropy close to zero (compositional distribution closer to the genome).

[1]However there are several exceptions to this very general rule e.g. the ribosomal protein coding and rRNA genes.

2.3 Change-point detection

As mentioned in the previous section the choice of the step for the sliding window approach is crucial, given that the window slides over raw genomic sequence (unknown gene boundaries), decreasing the window step will increase the computation required, while increasing it will reduce the accuracy of the localization of the predicted 'atypical' regions. For these reasons we implement a second order, two-state hidden Markov model (HMM) in a change-point detection framework. HMMs can be described by two processes (Durbin *et al*., 1998). The hidden state process $\pi = (\pi_1, \ldots, \pi_L)$ also known as the *path* and the observed process $x = (x_1, \ldots, x_L)$ which corresponds to the observed symbols, in our case the bases of a DNA sequence. In an n-th order HMM each base x_i depends on the previous bases $(x_{i-n}, \ldots, x_{i-1})$ as well as on the i-th state π_i in the path. In the current study, we use two states: the 'native'state that corresponds to regions of typical composition and the 'alien' state that models each compositionally deviating, 'atypical' region. Under this framework, a change-point corresponds to switching from one state to the other; in our case we want to infer the boundaries of the predicted regions, where a state transition occurs. This change-point will represent the new optimized boundary of each prediction, offering higher predictive accuracy in terms of boundary localization. In order to detect the point where the transition from the native to the alien state occurs and vice versa, we pursue the following approach:

Each predicted 'atypical' region is extended further upstream in order to incorporate sequence of typical composition. This hybrid sequence of one typical and one atypical subsequence is used to train the HMM on-the-fly (the same approach is also applied on the downstream boundary). We implement an Expectation Maximization technique, the Baum–Welch (BW) (Baum, 1972) algorithm to train the parameters (transition and emission probabilities) of the model, in an iterative fashion until some convergence criteria are met (Supplementary Table I). Given that we do not know beforehand for how long the system remains in the native state before it makes the transition to the alien state we start with multiple starting points (prior expectations) over the transition probability:

$$\alpha_{NA} = P(\pi_i = A \mid \pi_{i-1} = N), \tag{6}$$

where α_{NA} denotes the transition probability from the native N to the alien A state (different starting parameter values strongly affect the local maxima which the BW will converge over). In a change-point detection framework with a single change-point, once the α_{NA} transition occurs, the model persists at the alien state until the end. For this reason we choose to train only the α_{NA} transition probability while the transition probability from the alien to the typical state is set to be zero ($\alpha_{AN} = 0$, untrainable). For the emission probabilities we start with two trainable, uniform second order compositional distributions.

In a second step for each starting point, upon BW training, we implement the Viterbi algorithm with the updated-trained parameters. The Viterbi algorithm is a dynamic programming algorithm widely used in inferring the most probable state path π^* given the observations. Keeping track of the probability of the most probable path predicted by the Viterbi algorithm, the iteration (over different starting points) with the highest probable path among the most probable state paths, will be the one which best describes the data (the true transition point); the procedure is summarized as a pseudo code in Supplementary Table I.

2.4 Reciprocal FASTA—HGT test dataset

In order to evaluate the performance of the described method, we created a test dataset of putative horizontally transferred genes. Previous approaches involved simulation of HGT events by inserting genes from various donor genomes into the genome under study. As mentioned earlier, such approaches simulate only very recent HGT events, thus they do not take into account the amelioration (Lawrence and Ochman, 1997) of horizontally transferred genes, a time-dependent process. For this reason we chose to build a test dataset of putative HGT events, based on real data. We selected the genome of *S.typhi* CT18, a well-studied prokaryote in terms of HGT events. *S.typhimurium* LT2 was selected as a sister lineage to *S.typhi* while the genome of *E.coli* K12 was chosen as an outgroup of *S.typhi* and *S.typhimurium*. The main idea is that genes that are present in all the three genomes form a set of core genes, while the rest of the genes represent either species or strain specific genes, thus, are considered putative candidates for HGT. The choice of two sister lineages and one outgroup increases the chances of capturing older HGT events, which otherwise might be indistinguishable; e.g. SPI-1 and SPI-2 are species-specific, but not strain-specific. Moreover a comparative analysis between two sister taxa and one outgroup enables a more reliable discrimination between gene loss and gene gain. *E.coli* seems to form a good outgroup organism, given that the estimated divergence of *E.coli* and *S.enterica* from the common ancestor occurred ~100 million years ago (Doolittle *et al*., 1996; Ochman and Wilson, 1987). We took the following approach in order to extract all the putative horizontally transferred genes in *S.typhi*:

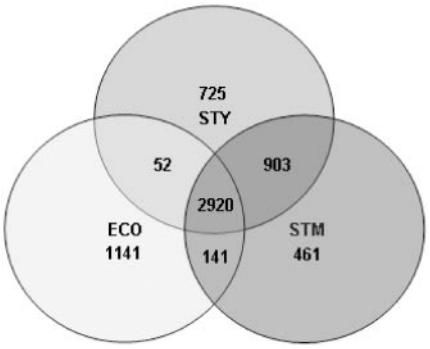

Fig. 1. Venn diagram illustrating the unique and the orthologous genes present in the genome of *E.coli* (ECO), *S.typhi* (STY) and *S.typhimurium* (STM).

Each CDS (*a*) from the genome (*A*) was searched, with FASTA, against the CDSs of the other genome (*B*). If the top hit covered at least 80% of the length of both sequences with at least 30% identity, a reciprocal FASTA search of the top hit sequence (*b*) was launched against the CDSs of the first genome. If the reciprocal top hit is the same as the original query CDS then (*a*) and (*b*) are considered orthologous genes of (*A*) and (*B*). Genes that are unique in, or are orthologs between *S.typhi* and *S.typhimurium* but do not have an ortholog in *E.coli* form our initial dataset of putative HGT events. In a second step, in order to validate the results, we performed a BLASTN and TBLASTX comparison between the three genomes to check for a syntenic relationship among the putative orthologs and visualized the results using ACT (Carver *et al*., 2005). It should be recognized that this procedure will also identify genes that have been uniquely deleted in *E.coli* as putative HGT events (see below).

2.5 Comparative analysis—distribution of novel SPIs

In order to analyze the distribution of the three predicted novel SPIs and other HGT events in the *Salmonella* lineage we performed a comparative analysis between *E.coli* and eight representatives of the *Salmonella* lineage (Supplementary Table II). Genome comparisons were generated using BLASTN and the results were inspected using ACT.

3 RESULTS

3.1 Manually curated HGT dataset

Implementing the reciprocal FASTA approach described above, we were able to identify four different groups of genes in *S.typhi*: The first group involves 725 genes that are unique in *S.typhi*. The second and third group includes orthologous genes between *S.typhi* and *E.coli* (52) and *S.typhi* and *S.typhimurium* (903). In the last group are 2920 core genes that are shared between all the three genomes (Fig. 1). Excluding the 2920 predicted core genes and the

Table 1. Characteristics of the three novel predicted SPIs (SPI-15, SPI-16, SPI-17) in the genome of *S.typhi*

SPI	Location	Insertion site	Repeats	Integrase	Score[a]	Size (bp)	Potential virulence determinants
SPI-15	3053654..3060017	tRNA Gly	22 nt (DR)[b]	Phage integrase	18.893	6364	Unknown
SPI-16	605515..609992	tRNA Arg	43 nt (DR)	Phage integrase	20.949	4478	Serotype conversion by O-antigen glucosylation
SPI-17	2460793..2465914	tRNA Arg	—	—	23.953	5122	Serotype conversion by O-antigen glucosylation

[a]The score reflects the compositional deviation of those regions from the genome background compositional distribution. The higher the score the more atypical the corresponding sequence. In *S.typhi*, the score threshold is 6.157.
[b]Degenerate repeats with two internal mismatches.

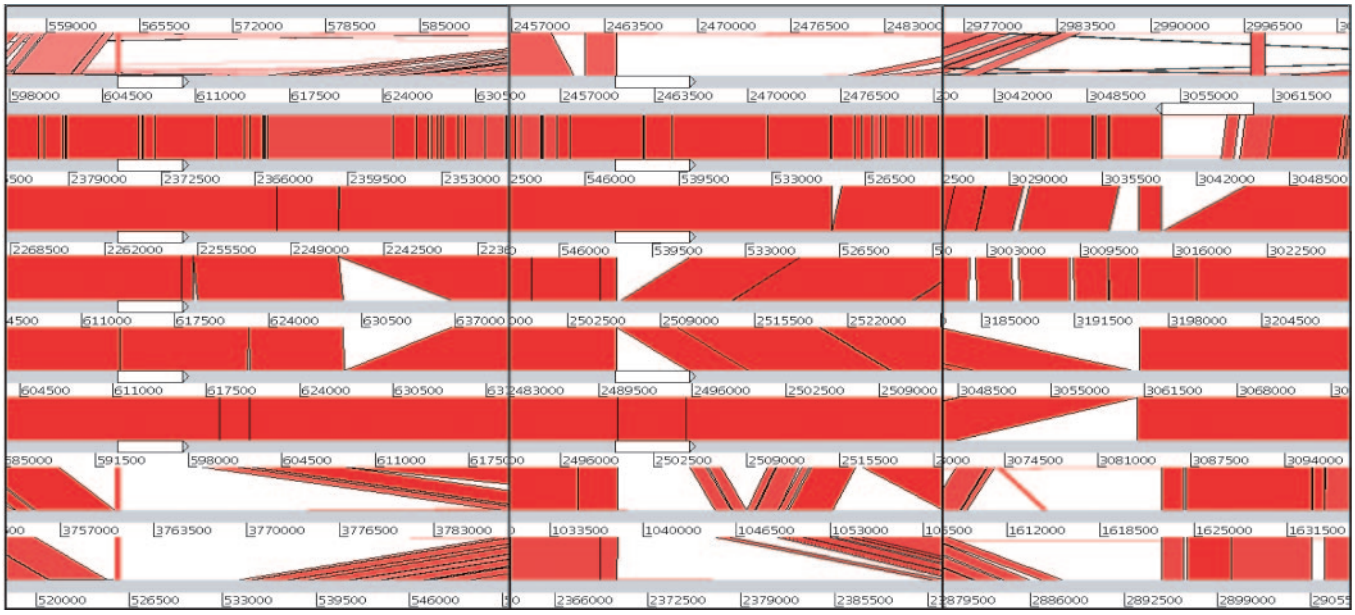

Fig. 2. ACT screenshot: BLASTN comparison between *E.coli* and eight *Salmonella* genomes (from top to bottom): *E.coli* K12, *S.typhi* CT18, *S.typhi* TY2, *S.paratyphi* A, *S.typhimurium* LT2, *S.gallinarum* 287/91, *S.enteritidis* PT4, *S.arizonae* RSK2980, *S.bongori* 12419. Regions within the nine genomes with sequence similarity are joined by red colored bands that represent the matching regions. The three novel SPIs are illustrated as white colored features (from left to right: SPI-16, SPI-17, SPI-15). The above screenshot is a mosaic picture of three individual screenshots at different locations along the genomes that have been concatenated for ease of visualization.

52 *S.typhi* and *E.coli* unique orthologs, the remaining gene set (1628 genes) forms the initial dataset of putatively horizontally transferred genes. In a second step, the above dataset was manually curated for gene position consistency using ACT, and the initial number was reduced to 1560 manually curated putative horizontally transferred genes[2], which form the basis of the analysis described in the following sections.

[2]It should be noted that this analysis yields a significantly high number of putative HGT events in the genome of *S.typhi* CT18. The reliable estimation of true HGT strongly depends on the evolutionary sample at hand; going well back in the evolutionary history of an organism offers more reliable detection of sequences that have been transferred horizontally from other sources. For example, some of the *Salmonella* lineage-specific genes might not necessarily represent HGT events (gene loss in *E. coli*). However this analysis provides a more reliable estimation of putative HGT events (taking into account the amelioration process), given that it is based on real data rather on simulated events.

3.2 Three novel *Salmonella* Pathogenicity Islands

Running the IVOM approach on the genome of *S. typhi*, all the previously annotated SPIs and bacteriophages were successfully predicted. Moreover this analysis revealed three novel putative SPIs, SPI-15, SPI-16 and SPI-17 (Table 1). SPI-11, 12 and SPI-13, 14 have been previously described (Chiu *et al*., 2005; Shah *et al*., 2005). SPI-15 represents an insertion of ∼6.5 kb, inserted in the 3′ end of a Gly tRNA; the insertion has duplicated a 22 nt tRNA fragment, which forms the downstream boundary of SPI-15. Adjacent to the tRNA, there is an integrase gene of putative phage origin and further downstream four hypothetical protein-coding genes. Among the eight *Salmonella* genomes, SPI-15 is only present in *S.typhi* CT18 (Fig. 2). In *S.typhi* TY2, there is a similar insertion of different gene content, at the same position, which also forms two DRs, 22 bp long.

The second SPI, SPI-16 is a 4.5 kb long island, inserted in an Arg tRNA. Two DRs of 43 bp form the boundaries of SPI-16 while

a phage integrase (pseudogene) is located near the tRNA gene. Encoded within this island are two bactoprenol-linked glucose translocases (*gtrA* and *gtrB*) that along with the integrase pseudogene show high percentage identity (93, 97 and 78%, respectively) to homologous genes in the genome of bacteriophage P22 (Figure I in Supplementary Material). *gtrA* and *gtrB* have been previously described to be involved in serotype conversion through O-antigen glycosylaion mediated by bacteriophages (Guan *et al*., 1999; Mavris *et al*., 1997).

Also present in SPI-16 is the STY0605 gene that encodes a putative membrane protein with nine predicted transmembrane segments (TMs). Although there is no sequence similarity to the *gtrC* gene in P22 bacteriophage, both genes encode proteins with TMs in equivalent positions (data not shown). It seems possible that those two genes have similarity on the structural level rather on the sequence level which might indicate similar function. Moreover the DR at the 5′ end of SPI-16 has significant sequence similarity (74% in 23 nt) with the 23 bp P22 bacteriophage attP attachment site (see alignment in Supplementary Figure I). These data support the phage origin of SPI-16 and indicate that this island seems to have been originated from a phage that shares similarities with P22 bacteriophage family. SPI-16 is absent from *E.coli*, *S.bongori* and *S.arizonae* while it is present in the rest of the *Salmonella* lineage (Fig. 2). Interestingly in *S.bongori* at the same tRNA location, there is a different insertion (8155 bp) with a phage integrase, suggesting that this tRNA location might represent a hotspot for integration of different SPIs in the *Salmonella* lineage.

The third novel island, SPI-17 is 5.1 kb long, inserted in an Arg tRNA. An integrase and DRs/IRs seem to be absent from this island, which is present in all the *Salmonella* genomes used in this study, apart from *S.bongori*, *S.arizonae* and *S.typhimurium*. This observation may indicate a possible recent deletion event that took place in the genome of *S.typhimurium*. SPI-17 seems to belong to the same phage family as SPI-16 given that the two serotype converting genes (*gtrA* and *gtrB*) are also present in the former island and both show high similarity with homologous genes in P22 bacteriophage; moreover in SPI-17 there is a pseudogene (STY2621a) with similarity with the P22 phage bifunctional tail protein (TSPE_BPP22), suggesting an island of phage origin with two well-defined boundaries (*gtrA* and the phage tail protein coding gene).

3.3 Change-point detection in boundary optimization

Other putative horizontally transferred regions (confirmed by comparative analysis—data not shown) were also predicted by this method, but given the lack of GI-related signatures, e.g. tRNA, integrase genes, were not classified as SPIs. As mentioned earlier, given that the current method is sliding window-based, the step of the window significantly affects the accuracy of the localization of the predicted boundaries. The implementation of a HMM model in a change-point detection framework seems to provide an effective way of dealing with this (Supplementary Table III). Indeed the average absolute error δx for the predicted boundaries with the implementation of the HMMs is much lower (3830 bp) than that without the boundary optimization (4936 bp). Interestingly the HMM-based approach gives an average δx quite close to the W8 method (3543 bp). W8 is a gene-based method, thus it is expected to provide quite accurate predicted boundaries of HGT events. Overall this indicates that the implementation of HMMs in a change-point detection framework significantly improves the localization of the predicted boundaries. An example is illustrated in Figure 3. This region is absent from the genome of *E.coli* and *S.typhimurium* and the BLASTN comparison indicates a well defined putative horizontally transferred region, 5223 bp long, consisting of four genes (STY3343, STY3344, STY3345, STY3347: putative membrane and putative hypothetical genes of no significant database hits).

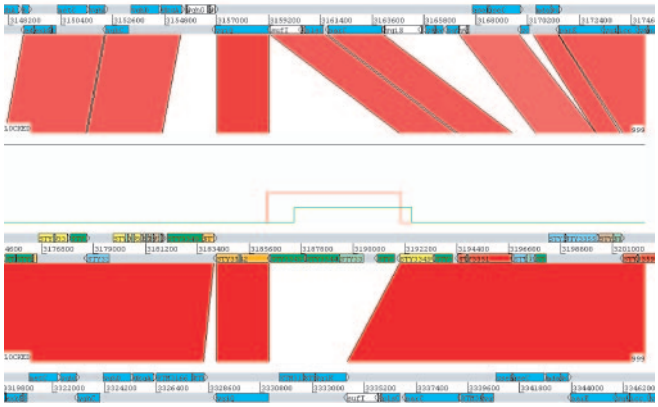

Fig. 3. ACT screenshot: BLASTN comparison between (from top to bottom): *E.coli* K12, *S.typhi* CT18, *S.typhimurium* LT2. An example of a predicted putative horizontally transferred region in the genome of *S.typhi* is indicated with two peaks in the IVOM score plot (above *S.typhi*). This region seems to be absent in the other two genomes compared. The red and the green colored score plots represent IVOM predictions with optimized (HMM) and unoptimized boundaries respectively.

As illustrated in the score plot in Figure 3, the unoptimized boundaries (green colored plot) were predicted in the middle of STY3343 and STY3349 genes. Applying the HMM approach, the true transition points were successfully identified (red plot), predicting the exact downstream and upstream boundaries of this region, diminishing the uncertainty of the localization of the predicted regions caused by the sliding window approach. The reason why we chose not to apply a purely HMM-based approach was the fact that a significant number of GIs (e.g. SPI-2) show a very mosaic structure, a result of several individual acquisitions, perhaps of different origin. Given that a HMM implementation requires the properties of the regions modeled to remain constant throughout their whole length, such an approach is not readily applicable to the prediction of GIs in microbial genomes.

3.4 Prediction accuracy—comparison with other methods

In order to test the performance of the IVOM method, a dataset of 1560 manually curated putative horizontally transferred genes in the genome of *S.typhi* was used. In this study we compared the IVOM method with four other published methods for the prediction of putative HGT events (Table 2): Islander, IslandPath, HGT-DB, and the W8 method of Tsirigos *et al*. Further the above methods and the method for the prediction of PAIs introduced by Yoon *et al*. (2005) were tested in terms of percentage coverage of the 10 previously described SPIs (SPI-1–SPI-10) and the five annotated bacteriophages (Table 3). Overall, the IVOM method shows higher predictive accuracy (AC = 0.764) compared with the other four methods (Table 2). Interestingly, the second most accurate method

Table 2. Performance comparison of IVOM with other prediction methods

Method	TP	FP	TN	FN	Number of predictions	SN	SP	AC	CC
IVOM	1013	539	2501	547	1552	0.649	0.653	0.764	0.473
W8	968	538	2502	592	1506	0.620	0.643	0.754	0.447
IslandPath_GC (δ^*)	611 (301)	467 (492)	2573 (2548)	949 (1259)	1078 (793)	0.392 (0.193)	0.567 (0.380)	0.692 (0.619)	0.266 (0.039)
Islander	275	89	2951	1285	364	0.176	0.755	0.701	0.258
HGT-DB	435	116	2924	1125	551	0.279	0.789	0.730	0.351

The comparison was based on the manually curated dataset of 1560 putative horizontally transferred genes, described in the text. TP: true positives, FP: false positives, TN: true negatives, FN: false negatives, SN: sensitivity, SP: specificity, AC: accuracy, CC: Matthews correlation coefficient.
The performance of IslandPath was evaluated based on two compositional indices: G + C content and dinucleotide bias (δ^*difference).

Table 3. Performance comparison of IVOM with other prediction methods based on a dataset of 10 previously described SPIs (SPI-1–SPI-10) and 5 annotated bacteriophages (SopE and P4 bacteriophages were ignored because they overlap with SPI-7 and SPI-10 respectively)

Annotated HGT	Number of CDS	IVOM		Islander		IslandPath_GC (δ^*)		HGT-DB		W8		Yoon *et al*. (2005)	
		%	counts	%	counts	%	counts	%	counts	%	counts	%	counts
SPI-6	60	81.7	49	0.0	0	51.7 (41.7)	31 (25)	40.0	24	70.0	42	0.0	0
Prophage10	63	81.0	51	100.0	63	23.8 (39.7)	15 (25)	25.4	16	96.8	61	0.0	0
SPI-5	8	100.0	8	100.0	8	75.0 (100.0)	6 (8)	100.0	8	100.0	8	100.0	8
Bacteriophage	53	100.0	53	0.0	0	39.6 (5.7)	21 (3)	34.0	18	86.8	46	0.0	0
SPI-2	44	77.3	34	0.0	0	61.4 (18.2)	27 (8)	68.2	30	77.3	34	100.0	44
Bacteriophage	71	88.7	63	0.0	0	33.8 (8.5)	24 (6)	35.2	25	94.4	67	0.0	0
SPI-9	4	25.0	1	0.0	0	25.0 (50.0)	1 (2)	0.0	0	25.0	1	0.0	0
Bacteriophage 27	19	89.5	17	0.0	0	36.8 (0.0)	7 (0)	5.3	1	73.7	14	26.3	5
SPI-1	44	95.5	42	0.0	0	54.5 (25.0)	24 (11)	77.3	34	88.6	39	40.9	18
SPI-8	16	100.0	16	0.0	0	68.8 (0.0)	11 (0)	68.8	11	68.8	11	0.0	0
Bacteriophage	46	89.1	41	0.0	0	37.0 (6.5)	17 (3)	23.9	11	60.9	28	0.0	0
SPI-3	14	85.7	12	0.0	0	28.6 (0.0)	4 (0)	14.3	2	42.9	6	100.0	14
SPI-4	7	100.0	7	0.0	0	85.7 (0.0)	6 (0)	100.0	7	100.0	7	100.0	7
SPI-7	149	100.0	149	31.5	47	31.5 (32.2)	47 (48)	28.2	42	81.2	121	10.1	15
SPI-10	29	100.0	29	44.8	13	44.8 (62.1)	13 (18)	6.9	2	72.4	21	0.0	0
ALL	627	91.2	572	20.9	131	40.5 (25.0)	254 (157)	36.8	231	80.7	506	17.7	111

For each annotated HGT, the number of predicted CDSs as well as the % CDS coverage of each method has been calculated.
The genomic locations of annotated bacteriophages (from top to bottom) are 1008747..1051266, 1538899..1572919, 1887450..1933558, 2759733..2782364, 3515397..3549055.

is W8, which utilizes higher order motifs (i.e. 8mers). These data suggest that the utilization of interpolated variable order motifs, improves both the sensitivity SN (IVOM: 0.649, W8: 0.62) and the specificity SP (IVOM: 0.653, W8: 0.643) compared with fixed-order methods; similarly this analysis confirms the superiority of higher order motif methods, discussed in the introduction. The sensitivity of IVOM is much higher compared to the other four methods which in turn reflects an increased ability to predict novel, putative horizontally transferred regions as well as already known examples. In terms of specificity the IVOM method is third from the top, following the Islander and the HGT-DB. Perhaps this can be attributed to the increased number of predictions provided by the IVOM method (1552) compared with the Islander (364) and HGT-DB (551) as well as to the fact that the IVOM method runs on raw genomic sequence without gene position information. Compared to the W8 method, although the IVOM provides higher number of predictions, both its sensitivity and specificity are higher.

In the second performance analysis, based on the percentage coverage of previously described HGT events, the IVOM predictions overlap with 91.2% of the CDSs present in SPIs and bacteriophages giving the highest number of complete GIs in *S.typhi*, followed by the W8 method with 80.7% coverage. These data suggest that the IVOM method is capable of detecting not only novel GIs but also can identify the majority of the already known regions of 'alien' origin. Overall the IVOM method predicts six complete structures (SPI-5, the bacteriophage at 1538899..1572919, SPI-8, SPI-4, SPI-7 and SPI-10), while in the case of SPI-2 predicts 34 out of 44 genes; it has been shown previously (Hensel *et al*., 1999) that SPI-2 is a mosaic island of at least two individual acquisitions. The mosaic nature of this SPI is also apparent in the G + C content (44.08 and 52.85%, respectively). This observation might explain the fragmented prediction for this SPI by all the methods except for the method of Yoon *et al*. (2005). The latter combines a method for capturing sequence deviation and similarity matches to already known PAIs to predict PAIs instead of GIs in general. Such methods will be powerful approaches in the detection of complete PAIs structures of similar gene content with previously annotated ones. Overall the

W8 method only outperforms the IVOM approach twice: in the first case it predicts 96.8% (IVOM: 81%) of the complete structure of prophage10 and in the second case 94.4% (IVOM: 88.7%) of the bacteriophage located at position 1887450..1933558. The Islander provides the lowest number of predictions (364) perhaps owing to the fact that it is restricted to predict only complete GI structures. In the case of known *S.typhi* islands, Islander predicts three SPIs (SPI-5, SPI-7, SPI-10) and one bacteriophage (prophage 10). The rest of the already known SPIs were not predicted although some of them (e.g. SPI-8) have both tRNA and integrase genes.

4 DISCUSSION

In this article, we have introduced and described a novel computational method for the prediction of putative horizontally transferred regions. This method, IVOM, exploits compositional biases at various levels (e.g. codon, dinucleotide and aminoacid bias, structural constraints) by implementing variable order motif distributions. Under this framework, the local sequence composition can be captured more reliably, compared with fixed-order methods. The IVOM approach relies more on higher order motifs to make more accurate predictions, but when the underlying information is insufficient for high order motifs, it takes into account information obtained from lower order motifs. Moreover, an IVOM approach can be applied even on newly sequenced genomes, given that it does not require any level of pre-existing annotation or gene position information. We discussed also the implementation of a HMM-based approach in a change-point detection framework for the optimization of the boundaries of the predicted regions and we showed that the uncertainty of the localization of the predictions caused by a sliding window method can be sufficiently handled by such an approach enabling more accurate localization of putative HGT events. Applying the IVOM method on the genome of *S.typhi*, all the previously annotated SPIs and bacteriophages were successfully predicted; moreover, the analysis of *S.typhi* revealed the presence of three novel SPIs, SPI-15–SPI-17, that have not been previously described. SPI-16 and SPI-17 represent islands of putative phage origin that may be implicated in serotype conversion by O-antigen glycosylation.

The performance benchmark of IVOM against four published methods indicates that IVOM is more sensitive in detecting compositionally deviating, putative HGT regions. On the other hand IVOM shows fairly poor specificity compared with HGT-DB and Islander. This observation seems to indicate that the last two methods are more reliable in terms of SP compared with the IVOM method. One obvious reason behind the lower SP of IVOM is the increased number of predictions (1552). HGT-DB and Islander show the highest SP owing to the low number of predictions (551 and 364 respectively); in other words they sacrifice SN for SP, predicting only a very small fraction of the already annotated HGT regions (Table 3). However if both SP and number of predictions are taken into account, the IVOM provides the highest number of predictions and at the same time its SP is even higher than W8s, although the latter provides lower number of predictions (1506). Overall this indicates that IVOM can be more sensitive and accurate compared to other methods that provide equally high number of predictions. It should be noted that this performance benchmark is based on a reciprocal FASTA approach that might penalize older HGT regions that were inserted prior to the divergence of *E.coli* and *Salmonella* lineages and were predicted by the IVOM method. Such cases are considered false positives based on this analysis, although they might represent true HGT events, and significantly affect the assigned SP of IVOM.

The prediction of the three novel SPIs in *S.typhi* CT18, raises the following question: What is the minimum size of PAIs or GIs that still maintain their ability to mobilize (integrate-excise)? Usually GIs are expected to be large ($\geq$10kb), distinct chromosomal regions (Schmidt and Hensel, 2004). The three novel SPIs described in this analysis seem to represent exceptions to this rule, with a size of 4–6 kb. For example SPI-17 is a minute PAI, and is absent from the genome of *S.typhimurium* LT2, possibly indicating a recent deletion or recombination event. The size of these regions may be the reason why they have not been previously reported.

SPI-15 encodes four hypothetical protein-coding genes with unknown function. Moreover while SPI-15 is only present in *S.typhi* CT18 and TY2, it can also be found in *Shigella flexneri* serovar 2a, strains 301 and 2457T (data not shown). Given that SPI-15 or similar structures are present in *S.flexneri* and *S.typhi* but not in *E.coli* (K-12, EDL933, O157:H7 and CFT073, data not shown) or other *Salmonella*, it would be interesting to further investigate the functionality of SPI-15 with respect to the biology of *S.typhi* and *S.flexneri*, given that both organisms are human-restricted enteric pathogens.

The annotation of horizontally transferred regions (e.g. GIs, phages) is a key task in annotation pipelines, especially in the case of pathogens since it reveals pathogenic aspects and characteristics of newly sequenced genomes. Prediction methods that reliably detect regions of 'alien' origin, requiring a minimum level of annotation, can form a powerful tool for the understanding and analysis of the biology for the genome at hand.

ACKNOWLEDGEMENTS

The authors would like to thank WUSTL for making *S.arizonae* RSK2980 data available, Nicholas Thomson for his valuable comments on the manuscript, Thomas Down for technical support regarding the BioJava source code and comments on the manuscript, David Carter for his valuable suggestions on the implementation of the HMM theory and Tim Carver for his helpful suggestions and technical support. G.S.V. is funded by the Wellcome Trust through a Sanger Institute Ph.D. studentship. Funding to pay the Open Access publication charges was provided by the Wellcome Trust.

Conflict of Interest: none declared.

REFERENCES

Baum,L.E. (1972) An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities*, **627**, 1–8.

Carver,T.J. *et al.* (2005) ACT: the Artemis Comparison Tool. *Bioinformatics*, **21**, 3422–3423.

Chiu,C.H. *et al.* (2005) The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. *Nucleic. Acids. Res.*, **33**, 1690–1698.

Doolittle,R.F. *et al.* (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science*, **271**, 470–477.

Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Garcia-Vallve,S. *et al*. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic. Acids. Res*., **31**, 187–189.

Guan,S. *et al*. (1999) Functional analysis of the *O* antigen glucosylation gene cluster of *Shigella flexneri* bacteriophage SfX. *Microbiology*, **145**, 1263–1273.

Hacker,J. *et al*. (1990) Deletions of chromosomal regions coding for fimbriae and hemolysins occur *in vitro* and *in vivo* in various extraintestinal *Escherichia coli* isolates. *Microb. Pathog*., **8**, 213–225.

Hacker,J. and Kaper,J.B. (2000) Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol*., **54**, 641–679.

Hensel,M. *et al*. (1999) Molecular and functional analysis indicates a mosaic structure of *Salmonella* pathogenicity island 2. *Mol. Microbiol*., **31**, 489–498.

Hsiao,W. *et al*. (2003) IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*, **19**, 418–420.

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

Karlin,S. *et al*. (1998) Codon usages in different gene classes of the *Escherichia coli* genome. *Mol. Microbiol*., **29**, 1341–1355.

Knapp,S. *et al*. (1986) Large, unstable inserts in the chromosome affect virulence properties of uropathogenic *Escherichia coli* O6 strain 536. *J. Bacteriol*., **168**, 22–30.

Lawrence,J. and Ochman,H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol*., **44**, 383–397.

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.

Mavris,M. *et al*. (1997) Mechanism of bacteriophage SfII-mediated serotype conversion in *Shigella flexneri*,. *Mol. Microbiol*., **26**, 939–950.

Ochman,H. and Wilson,A.C. (1987) Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol*., **26**, 74–86.

Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*., **183**, 63–98.

Salzberg,S.L. *et al*. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*., **26**, 544–548.

Sandberg,R. *et al*. (2001) Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res*., **11**, 1404–1409.

Schmidt,H. and Hensel,M. (2004) Pathogenicity islands in bacterial pathogenesis. *Clin. Microbiol. Rev*., **17**, 14–56.

Shah,D.H. *et al*. (2005) Identification of *Salmonella gallinarum* virulence genes in a chicken infection model using PCR-based signature-tagged mutagenesis. *Microbiology*, **151**, 3957–3968.

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

Sullivan,J.T. and Ronson,C.W. (1998) Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. *Proc. Natl Acad. Sci. USA*, **95**, 5145–5149.

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

Wu,T.J. *et al*. (2005) Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. *Bioinformatics*, **21**, 4125–4132.

Yoon,S.H. *et al*. (2005) A computational approach for identifying pathogenicity islands in prokaryotic genomes. *BMC Bioinformatics*, **6**, 184.

Zhang,L. *et al*. (1997) Molecular and chemical characterization of the lipopolysaccharide O-antigen and its role in the virulence of *Yersinia enterocolitica* serotype O:8. *Mol. Microbiol*., **23**, 63–76.