RESEARCH ARTICLE

A Novel Method to Predict Genomic Islands Based on Mean Shift Clustering Algorithm

Daniel M. de Brito[1], Vinicius Maracaja-Coutinho[2,3,4,5], Savio T. de Farias[3], Leonardo V. Batista[1], Thaís G. do Rêgo[1]*

1 Departamento de Informática, Centro de Informática, Universidade Federal da Paraíba, João Pessoa, Brazil, **2** Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, Santiago, Chile, **3** Departamento de Biologia Molecular, Centro de Ciências Exatas e da Natureza, Universidade Federal da Paraíba, João Pessoa, Brazil, **4** Instituto Vandique, João Pessoa, Brazil, **5** Beagle Bioinformatics, Santiago, Chile

* gaudenciothais@gmail.com

Citation: de Brito DM, Maracaja-Coutinho V, de Farias ST, Batista LV, do Rêgo TG (2016) A Novel Method to Predict Genomic Islands Based on Mean Shift Clustering Algorithm. PLoS ONE 11(1): e0146352. doi:10.1371/journal.pone.0146352

Editor: Dawn L Arnold, University of the West of England, UNITED KINGDOM

Received: August 28, 2015

Accepted: December 16, 2015

Published: January 5, 2016

Data Availability Statement: The software developed in this manuscript in freely available at http://msgip.integrativebioinformatics.me. The software source code was also deposited into GitHub (https://github.com/msgip/msgip).

Funding: VMC developed part of this research as a visiting researcher at Departamento de Biologia Molecular, Centro de Ciencias Exatas e da Natureza, Universidade Federal da Paraiba, through the program "Becas Iberoamerica - Jovenes Profesores e Investigadores", funded by Santander Universidades. The author VMC receives partial salaries from Beagle Bioinformatics. This funder

Abstract

Genomic Islands (GIs) are regions of bacterial genomes that are acquired from other organisms by the phenomenon of horizontal transfer. These regions are often responsible for many important acquired adaptations of the bacteria, with great impact on their evolution and behavior. Nevertheless, these adaptations are usually associated with pathogenicity, antibiotic resistance, degradation and metabolism. Identification of such regions is of medical and industrial interest. For this reason, different approaches for genomic islands prediction have been proposed. However, none of them are capable of predicting precisely the complete repertory of GIs in a genome. The difficulties arise due to the changes in performance of different algorithms in the face of the variety of nucleotide distribution in different species. In this paper, we present a novel method to predict GIs that is built upon mean shift clustering algorithm. It does not require any information regarding the number of clusters, and the bandwidth parameter is automatically calculated based on a heuristic approach. The method was implemented in a new user-friendly tool named MSGIP—*Mean Shift Genomic Island Predictor*. Genomes of bacteria with GIs discussed in other papers were used to evaluate the proposed method. The application of this tool revealed the same GIs predicted by other methods and also different novel unpredicted islands. A detailed investigation of the different features related to typical GI elements inserted in these new regions confirmed its effectiveness. Stand-alone and user-friendly versions for this new methodology are available at http://msgip.integrativebioinformatics.me.

Introduction

In the history of biological systems, many factors have played an important role in the evolutionary success of life. The possibility of generating diversity and spreading it across nature is one of the most important characteristics of life. In the origin of biological system, the capacity to exchange genetic novelties between the first forms of life was very important for survival and

provided support in the form of salaries for author VMC, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of this author are articulated in the "author contributions" section.

Competing Interests: The author VMC receives partial salaries from Beagle Bioinformatics. This funder provided support in the form of salaries for author VMC, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of this author are articulated in the "author contributions" section. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

evolution of organisms or quasispecies [1]. During the ancient period, the horizontal gene transfer was the primary mode of spreading genetic diversity and this contributed strongly to the formation of the first cellular lineages [1]. However, together with the emergence of the first cells, it gave rise to cellular mechanisms to avoid the massive horizontal gene transfer (HGT) and preserve the genome integrity. From this period, vertical gene transfer has become the main process for transmission of genetic material into organisms. Nevertheless, the horizontal gene transfer maintains an important role in the evolution of modern organisms, promoting a rapid spread of important genetic novelties to specific conditions, contributing to the fast adaptation of many organisms. The identification of horizontal gene transfer events can aid in the comprehension of the evolutionary processes and their consequences in the ecological relationship between organisms, as well as in human and animal health.

Regions acquired by HGT are known as genomic islands (GI) and are associated with sequences that are acquired after the emergence of the vertical gene transfer as the main mechanism of genetic heritage [2]. These regions, depending on the time of transmission and the evolutionary distance between the donor and acceptor, can present significant differences in their nucleotides composition and the organization of their biological information, which makes the identification of horizontal gene transfer events in genomics analysis possible.

It is estimated that the bacteria *Escherichia coli* has acquired at least 17% of its genes through HGT [3]. This phenomenon is also observed in higher eukaryotes, with many examples of HGT from prokaryotes-to-eukaryotes and even between eukaryotic organisms [4–7]. These transferred regions are exchanged using transposons, phages and plasmids as classical vectors [8, 9]. In general, the acquired genes are responsible for the birth of important new functions and adaptations associated with pathogenicity (pathogenicity islands, PAIs), antibiotics resistance, symbiosis, degradation and metabolism. These new functions highlight the importance of HGTs in medicine, environment and industry [10]. The identification of these acquired genes is essential for the development of new vaccines, medicines and to understand environmental changes.

Current methodologies for genomic islands detection

Genomic islands are genomic regions exchanged between different organisms. This transfer may result in a characteristic different G+C content and codon usage in that region in comparison to the rest of the genome [11]. The genes of a particular species are normally similar in their base composition and patterns of codon usage, in a way such that the sequence fragments acquired horizontally from other organisms can be distinguished computationally [12]. Current methods for the GI prediction based on sequence composition depends strongly on this fact. However, it leads to the non-identification of genomic islands in which the genomes of both donor and receptor organisms have similar or identical G+C content [3].

The identification of these genomic signatures can be represented using a metric. Two different commonly used measures are based on the G+C content and oligonucleotides composition (k-mers) [2]. In both cases, genome fragments (windows) are usually considered to measure the compositional bias and to compare it with the expected value for the entire genome, which corresponds to the whole sequence of composition average. Regions that differ from the rest of the genome are considered as GIs. Despite the windows being widely used, it is difficult to adjust their sizes, since small values lead to a large statistical fluctuation and higher values lead to a low resolution, making it impossible to detect small variations in the GC content [13]. Among the main methods that use this approach, we can highlight: SIGI-HMM [14], IslandPath [15], PAI-IDA [16], Centroid [17] and Alien_Hunter [18]. In 2004, Zhang and Zhang [13] proposed a method for the calculation of the G+C content of a genomic region without the use of windows, but a region of interest should be defined for the method be applied in GIs identification.

Other methods make use of multiple full genomic sequences of closely related phylogenetic species in order to investigate the presence of GIs. The full genome is aligned between two or more closely related organisms, and regions that are only present in the genome of interest are considered. This approach may be limited by the small number of sequences available for the group of organism under study. Also, it is important to note that the use of phylogenetically distant genomes can potentially result in the prediction of false positive GIs [2], while the use of closely related organisms can fail in the detection of recently transferred GIs that are acquired before the genomes start to differ from each other, leading to false negative islands. Among the main methods that use this approach we can highlight: MibolomeFinder [19] and IslandPick [20]. Finally, some tools have the option to use multiple methods for the prediction of genomic islands, allowing its visualization and comparison. Examples are EDIG [21], GIST [22] and IslandViewer [23].

Even with the availability of a broad range of methods, none of them are capable of predicting precisely the complete repertory of GIs in a genome. The difficulties arise due to the changes in performance of different algorithms in the face of the variety of nucleotide distribution in different species. It makes necessary the development of programs to combine various methods in order to increase the predictive accuracy in predicting GIs. This way, it is equally opportune to elaborate novel methods to employ different approaches for GI prediction in order to take full advantage of the benefits gained from identifying these regions. In this paper, we propose an alternative method for the prediction of GIs in bacteria using mean shift clustering algorithm. The mean shift does not require the definition of the number of clusters for the operation, but needs the bandwidth parameter to be set, which influences the number of clusters formed and affects the speed of algorithm convergence. The definition of bandwidth parameter is critical, strongly influencing the result of the algorithm. Thus, a heuristic was developed for automatic calculation. The method was implemented in a new tool named MSGIP—*Mean Shift Genomic Island Predictor*, freely available at http://msgip.integrativebioinformatics.me.

Materials and Methods

Clustering using mean shift

Mean shift is a non-parametric clustering algorithm that neither requires the prior definition of the number of clusters nor restricts the shape of the clusters [24]. The density estimation forms the basis of mean shift. For each data point, also referred to as feature vector, the algorithm executes a gradient ascend on local estimated density until its convergence. The stationary points represent density modes. The mathematical formulation is described as follows [25]:

Given a set of n data points $\mathbf{x}_i$ in d-dimensional space, where $i = 1, \ldots, n$, the multivariate density estimation with kernel $K(\mathbf{x})$, computed in the point $\mathbf{x}$ can be defined as follows:

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \tag{1}$$

where h is called the bandwidth and defines the kernel size.

The kernel function is defined as $K(\mathbf{x}) = c_{k,\, d}\, k(\|\mathbf{x}\|^2)$, where $c_{k,\, d}$ is the normalization constant that integrates $K(\mathbf{x})$ to 1, and $k(x)$ is called kernel profile. Taking the gradient of the density estimator defined in Eq 1, we have the equation below after some algebraic manipulation, assuming the function, $g(x) = -k'(x)$:

$$\nabla\hat{f}(\mathbf{x}) = \underbrace{\frac{2c_{k,d}}{nh^{d+2}}\left[\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)\right]}_{\text{first term}} \underbrace{\left[\frac{\sum_{i=1}^{n}\mathbf{x}_i\, g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x}\right]}_{\text{second term}} \tag{2}$$

The first term of Eq 2 is proportional to Eq 1, and the second term is the mean shift vector $\mathbf{m}_h(\mathbf{x})$, described below:

$$\mathbf{m}_h(\mathbf{x}) = \frac{\sum_{i=1}^{n}\mathbf{x}_i\, g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x} \tag{3}$$

The mean shift vector always points in the direction of the maximum increase of density. The algorithm procedure performed for a given data point $\mathbf{x}_i$ is described in the following points:

1. Calculates the mean shift vector $\mathbf{m}_h(\mathbf{x}_i^t)$;
2. Move the density estimation window $\mathbf{x}_i^{t+i} = \mathbf{x}_i^t + \mathbf{m}_h(\mathbf{x}_i^t)$;
3. Repeat the above steps until convergence, i.e. when, $\mathbf{x}_i^{t+1} - \mathbf{x}_i^t < \epsilon$.

where the superscript t is the current procedure iteration, and ϵ, the threshold. The procedure application for a data point $\mathbf{x}_i$ is illustrated in Fig 1

The number of clusters is automatically generated through the mode calculation of each point. Further, points that have the same mode are assigned to the same cluster.

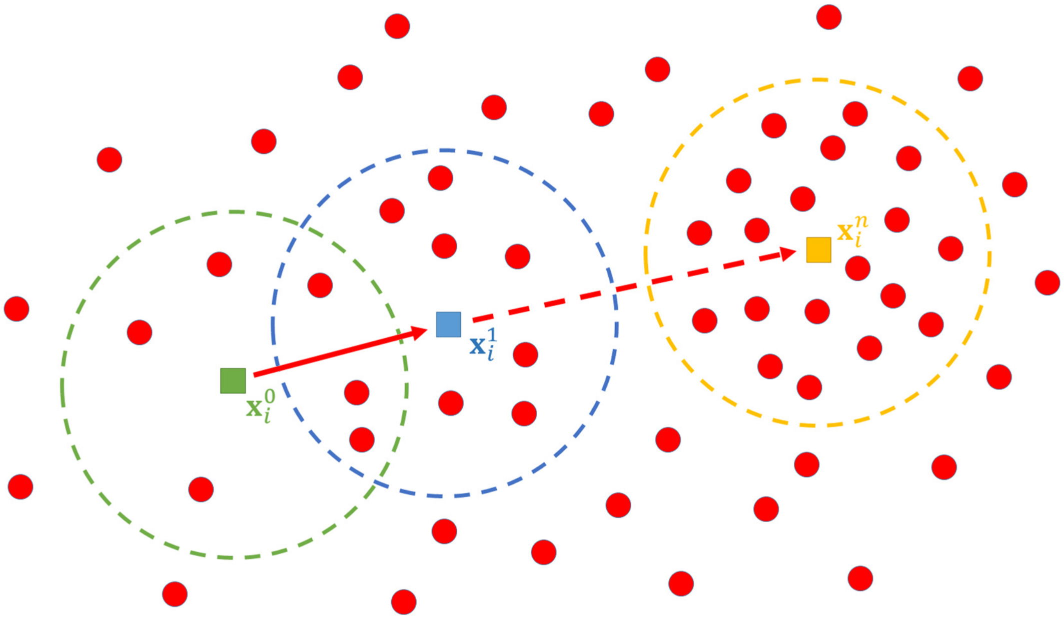

Fig 1. Mean shift algorithm procedure for a data point $\mathbf{x}_i$. The bold filled circles with arrows represent the iteration, while the pointed circles represent the window used in density estimation until the convergence is achieved at the *n*th iteration.

doi:10.1371/journal.pone.0146352.g001

Automatically setting the optimal bandwidth size

A common problem found in most partitional clustering algorithms is the necessity to specify a priori the number of clusters to be formed. However, mean shift does not require this parameter; it only requires the bandwidth length to be defined beforehand. The definition of this parameter is not a trivial task. Here, we choose to develop a heuristic to automatically calculate this parameter through the insertion of artificial genes. First, the user defines a fixed number of artificial fragments that will be extracted from the genome of different bacteria and inserted into the genome of interest. The fragments' sizes might be similar to the length defined by the user for the window size parameter. Then, the mean shift algorithm is executed until all windows from the original genome are grouped into different clusters. The bandwidth value is decreased and the algorithm is executed again. The process is repeated until the algorithm sets all the artificially inserted windows into different clusters. It is believed at this moment that the windows that host genes potentially acquired horizontally will also be kept in separate groups from the rest of the genomes, allowing them to be identified.

The selection process of artificial windows is described as follows. For each one of the nucleotides (A, T, C, G), the mean and standard deviation of its sum are calculated in all windows of the investigated genome. The user might specify the length of window *j*. In a set of genomes, artificial windows are selected in accordance with the following criteria:

$$\sum_{i=0}^{j} Y_i < \bar{X} - c\sigma_x \text{ or } \sum_{i=0}^{j} Y_i > \bar{X} + c\sigma_x \tag{4}$$

where *c* is a positive constant bigger or equal to 1, that might be provided as input to this method.

The higher the selected coefficient for standard deviation, the more distant is the artificially inserted window in comparison to the original genome. One should also ensure that there is no similarity between the artificially inserted windows. In this work, we consider artificial sequences that differ by more than 7.5%, in the sum of their four nucleotides, from the other selected artificial sequences.

Identification of genomic islands

It is expected that the windows of genes acquired horizontally differ in their nucleotides composition in relation to the rest of the host genome sequence. After the mean shift clustering execution using the ideal bandwidth value, these windows will be grouped separately in relation to the rest of the genome.

According to Hacker and Kaper [3], the sizes of genomic islands typically vary between 10kb and 200kb. However, here we consider clusters containing up to 200kb, varying according to the size of the selected windows. We choose this approach due to the variation in the sizes of genomic islands, so that a genomic island can correspond to the length of multiple consecutive windows. Roos and Van Passel [26] pointed that a unique bacterium can be responsible for multiple horizontal transferences in a unique receptor organism. Because of that, we considered GIs as those clusters in which the sum of windows does not surpass 200kb, regardless of the order in which they appear. Other values were also tested and nearby numbers did not influence the results, while small or large values were influenced negatively.

Datasets

In this work, we have used the genome sequence from bacteria with known genomic islands. The full sequences for each species or strain were downloaded from the Bacteria database from NCBI (ftp.ncbi.nih.gov/genomes/Bacteria) and are listed in Table 1.

Table 1. Genomes used for testing mean shift and other related algorithms.

Genome	Length(Mb)	Accession Number	References
Corynebacterium glutamicum ATCC 13032	3.309	NC_003450.3	[27]
Vibrio vulnificus CMCP6 chromosome I	3.281	NC_004459.3	[28]
Rhodopseudomonas palustris CGA009	5.459	NC_005296.1	[29]
Streptococcus mutans UA159	2.032	NC_004350.2	[30]
Vibrio cholerae O1 biovar El Tor str. N16961 chromosome II	1.072	NC_002506.1	[31]
Vibrio vulnificus YJ016 chromosome I	3.354	NC_005139.1	[32]
Mycoplasma genitalium G37	0.580	NC_000908.2	[33]
Rickettsia prowazekii str. Breinl	1.109	NC_020993.1	[34]

doi:10.1371/journal.pone.0146352.t001

Results and Discussion

In this section, the mean shift method is applied in bacterial genomes, where genomic islands were previously described in the literature. The G+C cumulative profile [13] is used to visualize the obtained results. The curve generated by this approach can represent different interesting characteristics of a genome sequence. Zhang and Zhang [13] showed that the G+C content inside genomic islands is reasonably homogeneous in comparison to the rest of the host genome, in a way that these regions are represented in the curve as almost straight lines. A jump or a fall in the curve indicates an abrupt decrease or increase in the G+C content, pointing to a potential source of horizontal transfers in that region.

We applied variations on the set of parameters for this method to find the best option. Finally, we have obtained consistent results with all tested genomes when the following parameters were used: (i) window length: 50kb; (ii) quantity of artificially inserted genomic sequences fixed in 5; and (iii) a standard deviation of 1.

In Table 2, we have listed all genomic islands detected by the mean shift and other methods for all genomes and species listed in Table 1. The z' curve for every studied genome can be observed in Fig 2. The regions containing a genomic island are highlighted by black lines in the graphic.

Corynebacterium glutamicum ATCC 13032

Corynebacterium glutamicum ATCC 13032 is a gram-positive bacterium that is important for industrial production of amino acids [39], with a genome size estimated in 3.309 Mb. Here, the mean shift method identified one genomic island (Table 2). The region of 200kb corresponding to the position 1.800–2.000 Mb is represented in Table 2. The z' curve generated for this genome is displayed in Fig 2A. This genomic island was predicted in the same region as described by [13], where 82.9% of the identified genes encode hypothetical proteins (with unknown function), while for the remainder of the genome, only 26.6% encodes these proteins. This difference indicates that the genes present in these regions were acquired from other organisms, and can therefore be described as GIs.

Vibrio vulnificus CMCP6 chromosome I

Vibrio vulnificus CMCP6 is a pathogenic estuarine bacterium with a genome estimated in 3.281 Mb, which affects a human being's underlying hepatic diseases and other immune-compromised conditions [40]. As listed in Table 2, the mean shift method identified three different genomic islands in this genome. All of them correspond to previously described genomic islands by [13]. The GI VVC01 corresponds to the position 0.350–0.400 Mb (see Table 2 and

Table 2. Genomic islands detected by mean shift and its comparison with other methods previously used for each species.

Genome	Identifier	Detected GI (Mb) (This work)	Corresponding GI (Mb) (Other Methods)	Characteristics
Corynebacterium glutamicum **ATCC 13032**	CGGI01	1.800–2.000	1.776–1.987 [13]	Hypothetical proteins (with unknown function)
Vibrio vulnificus **CMCP6 chromosome** I	VVCGI01	0.350–0.400	0.355–0.395 [13]	Hypothetical proteins and invasion-associated proteins
	VVCGI02	2.450–2.600	2.438–2.605 [13]	Invasion-associated proteins
	VVCGI03	3.250–3.281	3.248–3.281 [13]	Transporter protein, transposase, phage and hypothetical proteins
Rhodopseudomonas palustris **CGA009**	RPGI01	—	2.481–2.564 [35]	IV secretion genes for conjugal transfer of DNA, arsenate reductase pump modifier and an arsenical pump membrane protein
	RPGI02	3.750–3.800	3.729–3.807 [35]	Hypothetical proteins
	RPGI03	4.400–4.450	—	Hypothetical proteins and flagellar proteins
	RPGI04	4.550–4.650	4.578–4.678 [35]	Multidrug efflux and transporter related genes
Streptococcus mutans **UA159**	SMGI01	1.250–1.300	1.250–1.300 [36, 37]	TnSMU2 (nonribosomal peptide synthetases (NRPS), polyketide synthases (PKS), accessory proteins, transporters, and transcription regulators)
Vibrio cholerae **chromosome II**	VCGI01	0.300–0.450	0.302–0.436 [38]	Chloramphenicol acetyltransferase, killer protein, antidote protein, haemagglutinin, others copies of acetyltransferase and Hypothetical protein
Vibrio vulnificus **YJ016 chromosome I**	VVYGI01	—	0.159–0.167 [38]	Lactoglutathione lyase
	VVYGI02	1.800–1.950	1.757–1.936 [38]	Hypothetical proteins and transposases
	VVYGI03	2.200–2.250	—	Hypothetical proteins, region started and finished

doi:10.1371/journal.pone.0146352.t002

Fig 2B) and host genes associated with hypothetical proteins and invasion-associated proteins; and the GI VVC02 corresponds to the region 2.438–3.605 Mb and also the host genes related to invasion-associated proteins. Finally, a last island was identified at the end of the genome (position 3.250–3.281 Mb) and host site-specific recombinase, phage, integrase, transposase, multidrug transporter and hypothetical proteins.

Rhodopseudomonas palustris CGA009 CMCP6 chromosome I

Rhodopseudomonas palustris CGA009 is a gram-negative bacterium, recognized by its versatile metabolism to produce energy by using light and inorganic and organic compounds [29]. For this bacterium, the mean shift method identified three different genomic islands (see Table 2). Two of them were previously identified by [35], while the remaining were identified for the first time. The z' curve for this genome is displayed in Fig 2C. The island RPGI01, corresponds to several hypothetical proteins, typical to a genomic island as described by [2]. However, this genomic island was not identified by our method.

The region RPGI02 hosts genes associated with hypothetical proteins and consist of a previously described genomic island [35]. RPGI03 was also predicted by [35], and hosts a high presence of hypothetical proteins, as well as transport proteins associated with the flagellar system. The last predicted genomic island hosts proteins associated with two multidrug efflux and transport related genes, suggesting that horizontal gene transfer may play a role in the antibiotic resistance of *R. palustris*.

Streptococcus mutans UA159

Streptococcus mutans UA159 is a gram-positive bacterium, and it is known to be the leading cause of dental caries [30]. The mean shift method identified only one genomic island in this

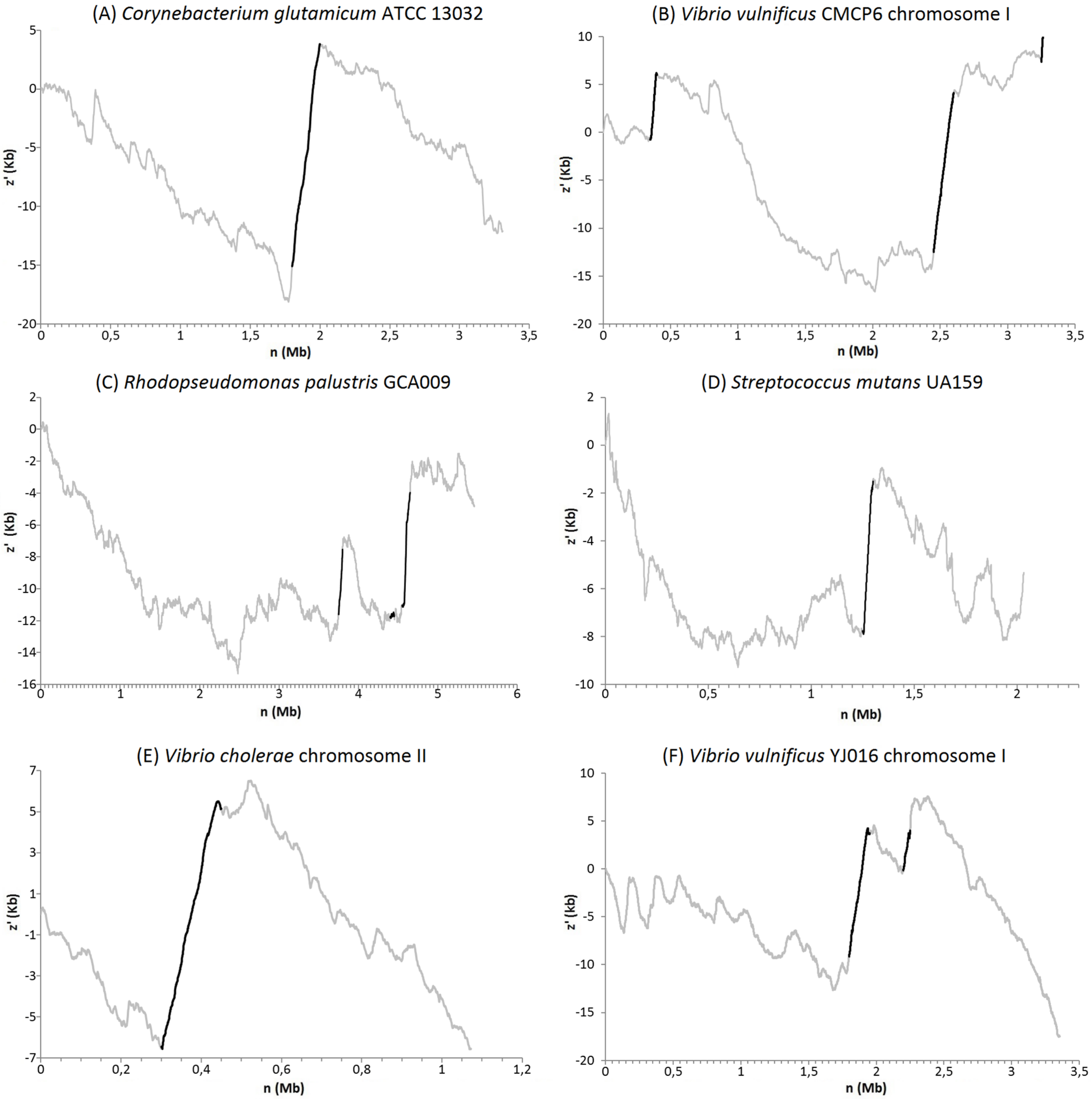

Fig 2. Graphics representing the *z*′ curves for following six different bacteria genomes. (A) *Corynebacterium glutamicum* ATCC 13032, (B) *Vibrio vulnificus* CMCP6 chromosome I, (C) *Rhodopseudomonas palustris* CGA009, (D) *Streptococcus mutans* UA159, (E) *Vibrio cholerae* chromosome II e (F) *Vibrio vulnificus* YJ016 chromosome I. Black lines represents the genomic islands identified by the mean shift method.

doi:10.1371/journal.pone.0146352.g002

bacterium (Table 2 and Fig 2D). The GI consists of the region *TnSmu2*, which is known to have much lower G+C content than the rest of the genome [36, 37]. This region carries genes coding for non-ribosomal peptide synthetases, polyketide synthases, and accessory proteins that are responsible for the biosynthesis of the pigment mutanobactin carried by *S. mutans*.

Mutational analysis further demonstrated that this gene cluster (named the *mub* locus) is involved in oxygen tolerance, H2O2 resistance, and biofilm formation in *S. mutans*[41].

Vibrio cholerae chromosome II

Vibrio cholerae is the etiological agent of the cholera disease [42]. For this bacterium, the mean shift identified only one region as a potential genomic island (Table 2 and Fig 2E). This island's host genes are associated with chloramphenicol acetyltransferase, killer protein, antidote protein, haemagglutinin and other copies of acetyltransferase. Also, it was observed many hypothetical and conserved hypothetical proteins. This island was also previously predicted by [38].

Vibrio vulnificus YJ016 chromosome I

As previously described, *Vibrio vulnificus* is a human pathogenic bacterium. For this different strain, we identified two different genomic islands with the mean shift method (Table 2 Fig 2F). The GI VVY02 was previously described by [38]. The other region identified only by our method (VVY03) hosts several hypothetical proteins and phages, and begins and ends with transposases, which are known to be associated with genomic islands insertions. The mean shift method was not capable of identifying GI VVY01 (position: 0.159–0.167 Mb), as described in [38].

Mycoplasma genitalium G37 e *Rickettsia prowazekii*

It has been reported that some bacterium does not have genomic islands in its genome composition. This is the case with *Mycoplasma genitalium*[13], the smallest known genome of any free-living organism, which was originally isolated from urethral specimens of patients with non-gonococcal urethritis. It lives in a parasitic association with ciliated epithelial cells of primate genital and respiratory tracts [33]. *Rickettsia prowazekii*, the agent of epidemic typhus, and a potential biothreat agent [43], also does not have any genomic islands [44]. We used the mean shift method on the genome sequence of both bacteria to evaluate the method efficiency and the potential prediction of false positive islands. As expected, genomic islands were not detected for both bacteria.

MSGIP—*Mean Shift Genomic Island Predictor tool*

The mean shift method described here was implemented in a standalone and user-friendly tool named MSGIP—*Mean Shift Genomic Island Predictor*. MSGIP was developed in Java and is compatible with any operating system with Java Runtime Environment installed. MSGIP outputs the predicted GIs in a text-like format that can be further saved into a file. The usage requires only a FASTA ".fna" file, containing the completely genomic sequence of the investigated bacteria. The user has to set the three parameters discussed in the text or, alternatively, he can use the default parameters tested by the authors in this study. The dataset of bacteria used to extract the artificial fragments is provided by MSGIP, but the user can add other sequences (in.fna format) to the folder "genomes" containing the sequences file. MSGIP source code and friendly version are freely available and can be downloaded at: http://msgip.integrativebioinformatics.me.

Conclusions

This work presents a new method for genomic island prediction based on the mean shift clustering algorithm. Our results prove that this new method can produce results that are consistent with other methods described in the previous literature. We consider the simplicity and easy implementation of this method as an advantage in comparison to other methods used for genomic island prediction, since the current model uses only the base vector (A, T, C, G) of its own genome sequence as a measure of composition. However, the calculation of bandwidth parameter must be studied in greater detail in order to identify the possible bias generated by artificially selected genes during the determination of bandwidth parameter. We expect this method to be used in discovering new GIs, which are not predicted by other methods and can also be used in combination with other methods.

Acknowledgments

VMC developed a part of this research as a visiting researcher at the *Departamento de Biologia Molecular, Centro de Ciências Exatas e da Natureza, Universidade Federal da Paraíba*, through the program *"Becas Iberoamérica—Jóvenes Profesores e Investigadores"*, funded by Santander Universidades.

Author Contributions

Conceived and designed the experiments: VMC STF LVB TGR DMB. Performed the experiments: DMB. Analyzed the data: DMB VMC STF LVB TGR. Contributed reagents/materials/analysis tools: DMB VMC. Wrote the paper: DMB VMC STF TGR.

References

1. Woese C. The universal ancestor. Proceedings of the National Academy of Sciences. 1998; 95 (12):6854–6859. doi: 10.1073/pnas.95.12.6854
2. Langille MG, Hsiao WW, Brinkman FS. Detecting genomic islands using bioinformatics approaches. Nature Reviews Microbiology. 2010; 8(5):373–382. doi: 10.1038/nrmicro2350 PMID: 20395967
3. Hacker J, Kaper JB. Pathogenicity islands and the evolution of microbes. Annual Reviews in Microbiology. 2000; 54(1):641–679. doi: 10.1146/annurev.micro.54.1.641
4. Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics. 2008; 9(8):605–618. doi: 10.1038/nrg2386 PMID: 18591983
5. Andersson JO. Gene transfer and diversification of microbial eukaryotes. Annual review of microbiology. 2009; 63:177–193. doi: 10.1146/annurev.micro.091208.073203 PMID: 19575565
6. Marcet-Houben M, Gabaldón T. Acquisition of prokaryotic genes by fungal genomes. Trends in Genetics. 2010; 26(1):5–8. doi: 10.1016/j.tig.2009.11.007 PMID: 19969385
7. Gladieux P, Ropars J, Badouin H, Branca A, Aguileta G, Vienne DM, et al. Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. Molecular ecology. 2014; 23(4):753–773. doi: 10.1111/mec.12631 PMID: 24341913
8. Che D, Hasan MS, Chen B. Identifying pathogenicity islands in bacterial pathogenomics using computational approaches. Pathogens. 2014; 3(1):36–56. doi: 10.3390/pathogens3010036 PMID: 25437607
9. Osborn AM, Böltner D. When phage, plasmids, and transposons collide: genomic islands, and conjugative-and mobilizable-transposons as a mosaic continuum. Plasmid. 2002; 48(3):202–212. doi: 10.1016/S0147-619X(02)00117-8 PMID: 12460536
10. Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS microbiology reviews. 2009; 33(2):376–393. doi: 10.1111/j.1574-6976.2008.00136.x PMID: 19178566
11. Greub G, Collyn F, Guy L, Roten CA. A genomic island present along the bacterial chromosome of the *Parachlamydiaceae* UWE25, an obligate amoebal endosymbiont, encodes a potentially functional F-like conjugative DNA transfer system. BMC microbiology. 2004; 4(1):48. doi: 10.1186/1471-2180-4-48 PMID: 15615594
12. Lawrence JG, Ochman H. Molecular archaeology of the *Escherichia coli* genome. Proceedings of the National Academy of Sciences. 1998; 95(16):9413–9417. doi: 10.1073/pnas.95.16.9413
13. Zhang R, Zhang CT. A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I. Bioinformatics. 2004; 20(5):612–622. doi: 10.1093/bioinformatics/btg453 PMID: 15033867

14. Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, et al. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC bioinformatics. 2006; 7(1):142. doi: 10.1186/1471-2105-7-142 PMID: 16542435

15. Hsiao W, Wan I, Jones SJ, Brinkman FS. IslandPath: aiding detection of genomic islands in prokaryotes. Bioinformatics. 2003; 19(3):418–420. doi: 10.1093/bioinformatics/btg004 PMID: 12584130

16. Tu Q, Ding D. Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. FEMS microbiology letters. 2003; 221(2):269–275. doi: 10.1016/S0378-1097(03)00204-0 PMID: 12725938

17. Rajan I, Aravamuthan S, Mande SS. Identification of compositionally distinct regions in genomes using the centroid method. Bioinformatics. 2007; 23(20):2672–2677. doi: 10.1093/bioinformatics/btm405 PMID: 17724060

18. Vernikos GS, Parkhill J. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. Bioinformatics. 2006; 22(18):2196–2203. doi: 10.1093/bioinformatics/btl369 PMID: 16837528

19. Ou HY, He X, Harrison EM, Kulasekara BR, Thani AB, Kadioglu A, et al. MobilomeFINDER: web-based tools for in silico and experimental discovery of bacterial genomic islands. Nucleic acids research. 2007; 35(suppl 2):W97–W104. doi: 10.1093/nar/gkm380 PMID: 17537813

20. Langille MG, Hsiao WW, Brinkman FS. Evaluation of genomic island predictors using a comparative genomics approach. BMC bioinformatics. 2008; 9(1):329. doi: 10.1186/1471-2105-9-329 PMID: 18680607

21. Che D, Hasan MS, Wang H, Fazekas J, Huang J, Liu Q. EGID: an ensemble algorithm for improved genomic island detection in genomic sequences. Bioinformation. 2011; 7(6):311. doi: 10.6026/007/97320630007311 PMID: 22355228

22. Hasan MS, Liu Q, Wang H, Fazekas J, Chen B, Che D. GIST: Genomic island suite of tools for predicting genomic islands in genomic sequences. Bioinformation. 2012; 8(4):203. doi: 10.6026/97320630008203 PMID: 22419842

23. Dhillon BK, Laird MR, Shay JA, Winsor GL, Lo R, Nizam F, et al. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. Nucleic acids research. 2015;p. gkv401.

24. Georgescu B, Shimshoni I, Meer P. Mean shift based clustering in high dimensions: A texture classification example. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. IEEE; 2003. p. 456–463.

25. Derpanis KG. Mean shift clustering. Lecture Notes http://www cse yorku ca/ ~ kosta/CompVis_Notes/mean_shift pdf. 2005;.

26. Roos TE, van Passel MW. A quantitative account of genomic island acquisitions in prokaryotes. BMC genomics. 2011; 12(1):427. doi: 10.1186/1471-2164-12-427 PMID: 21864345

27. Ikeda M, Nakagawa S. The Corynebacterium glutamicum genome: features and impacts on biotechnological processes. Applied Microbiology and Biotechnology. 2003; 62(2–3):99–109. doi: 10.1007/s00253-003-1328-1 PMID: 12743753

28. Kim HU, Kim SY, Jeong H, Kim TY, Kim JJ, Choy HE, et al. Integrative genome-scale metabolic analysis of Vibrio vulnificus for drug targeting and discovery. Molecular systems biology. 2011; 7(1):460. doi: 10.1038/msb.2010.115 PMID: 21245845

29. Larimer FW, Chain P, Hauser L, Lamerdin J, Malfatti S, Do L, et al. Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. Nature biotechnology. 2004; 22(1):55–61. doi: 10.1038/nbt923 PMID: 14704707

30. Ajdić D, McShan WM, McLaughlin RE, Savić G, Chang J, Carson MB, et al. Genome sequence of Streptococcus mutans UA159, a cariogenic dental pathogen. Proceedings of the National Academy of Sciences. 2002; 99(22):14434–14439. doi: 10.1073/pnas.172501299

31. Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, et al. DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. Nature. 2000; 406(6795):477–483. doi: 10.1038/35020000 PMID: 10952301

32. Chen CY, Wu KM, Chang YC, Chang CH, Tsai HC, Liao TL, et al. Comparative genome analysis of Vibrio vulnificus, a marine pathogen. Genome research. 2003; 13(12):2577–2587. doi: 10.1101/gr.1295503 PMID: 14656965

33. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, et al. The minimal gene complement of *Mycoplasma genitalium*. Science. 1995; 270(5235):397–404. doi: 10.1126/science.270.5235.397 PMID: 7569993

34. Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Pontén T, Alsmark UCM, Podowski RM, et al. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature. 1998; 396(6707):133–140. doi: 10.1038/24094 PMID: 9823893

35. Zhang CT, Zhang R. Genomic islands in *Rhodopseudomonas palustris*. Nature biotechnology. 2004; 22(9):1078–1079. doi: 10.1038/nbt0904-1078b PMID: 15340470

36. Waterhouse JC, Russell RR. Dispensable genes and foreign DNA in *Streptococcus mutans*. Microbiology. 2006; 152(6):1777–1788. doi: 10.1099/mic.0.28647-0 PMID: 16735740

37. Chattoraj P, Banerjee A, Biswas S, Biswas I. ClpP of *Streptococcus mutans* differentially regulates expression of genomic islands, mutacin production, and antibiotic tolerance. Journal of bacteriology. 2010; 192(5):1312–1323. doi: 10.1128/JB.01350-09 PMID: 20038588

38. Nag S, Chatterjee R, Chaudhuri K, Chaudhuri P. Unsupervised statistical identification of genomic islands using oligonucleotide distributions with application to *Vibrio* genomes. Sadhana. 2006; 31 (2):105–115. doi: 10.1007/BF02719776

39. Mormann S, Lömker A, Rückert C, Gaigalat L, Tauch A, Pühler A, et al. Random mutagenesis in *Corynebacterium glutamicum* ATCC 13032 using an IS6100-based transposon vector identified the last unknown gene in the histidine biosynthesis pathway. BMC genomics. 2006; 7(1):205. doi: 10.1186/1471-2164-7-205 PMID: 16901339

40. Kim YR, Lee SE, Kim CM, Kim SY, Shin EK, Shin DH, et al. Characterization and pathogenic significance of *Vibrio vulnificus* antigens preferentially expressed in septicemic patients. Infection and immunity. 2003; 71(10):5461–5471. doi: 10.1128/IAI.71.10.5461-5471.2003 PMID: 14500463

41. Wu C, Cichewicz R, Li Y, Liu J, Roe B, Ferretti J, et al. Genomic island TnSmu2 of *Streptococcus mutans* harbors a nonribosomal peptide synthetase-polyketide synthase gene cluster responsible for the biosynthesis of pigments involved in oxygen and H2O2 tolerance. Applied and environmental microbiology. 2010; 76(17):5815–5826. doi: 10.1128/AEM.03079-09 PMID: 20639370

42. Kim EJ, Lee CH, Nair GB, Kim DW. Whole-genome sequence comparisons reveal the evolution of *Vibrio cholerae* O1. Trends in microbiology. 2015;.

43. Bishop-Lilly KA, Ge H, Butani A, Osborne B, Verratti K, Mokashi V, et al. Genome sequencing of four strains of *Rickettsia prowazekii*, the causative agent of epidemic typhus, including one flying squirrel isolate. Genome announcements. 2013; 1(3):e00399–13. doi: 10.1128/genomeA.00399-13 PMID: 23814035

44. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature. 2000; 405(6784):299–304. doi: 10.1038/35012500 PMID: 10830951