Hindawi
Computational and Mathematical Methods in Medicine
Volume 2021, Article ID 9969751, 9 pages
https://doi.org/10.1155/2021/9969751

Research Article

Genomic Island Prediction via Chi-Square Test and Random Forest Algorithm

Mbulayi Onesime, Zhenyu Yang, and Qi Dai

College of Life Sciences and Medicine, Zhejiang Sci-Tech University, Hangzhou 310018, China

Correspondence should be addressed to Qi Dai; daiailiu04@yahoo.com

Received 29 March 2021; Accepted 14 May 2021; Published 25 May 2021

Academic Editor: Tao Huang

Genomic islands are related to microbial adaptation and carry different genomic characteristics from the host. Therefore, many methods have been proposed to detect genomic islands from the rest of the genome by evaluating its sequence composition. Many sequence features have been proposed, but many of them have not been applied to the identification of genomic islands. In this paper, we present a scheme to predict genomic islands using the chi-square test and random forest algorithm. We extract seven kinds of sequence features and select the important features with the chi-square test. All the selected features are then input into the random forest to predict the genome islands. Three experiments and comparison show that the proposed method achieves the best performance. This understanding can be useful to design more powerful method for the genomic island prediction.

1. Introduction

Horizontal gene transfer (HGT) is one of the main factors affecting bacterial adaptability. Hacker et al. found some viral gene clusters in *E. coli* genomes and did not exist in their close species, and they denoted them as pathogenic islands (PAIs) [1]. Since then, at least a dozen PAIs have been detected, such as "secretion island," "antimicrobial island," and "metabolic island" [2]. They are first expressed as genomic islands (GIs) and further encode them based on the functions related to the complex changes of niche [3]. For example, GIs are responsible for the type III secretion system, iron absorption function, toxin, and adhesion secretion, which enhance the survival ability of pathogens in the host body, leading to diseases [4, 5]. Some researchers reported that pathogenicity can be regulated by selective loss or recovery of specific GIs [6, 7], and PAI can be spontaneously removed from chromosomes at a detectable rate, resulting in different pathogenic phenotypes [8, 9]. Therefore, the detection of different GIs has become an important content of microbial evolution and function research.

With the help of large-scale comparative genomics, researchers found that GIs have different sequence composition, direct flanking duplication, mobility, and tRNA genes. In turn, exploring and utilizing these features can lead to better detection of GIs [3, 10–12]. GIs are scattered among close relatives, which carry some species patterns different from the host. Researchers can identify distant relatives by comparing the differences of 16S rRNA or other homologous sequences [13]. Some alignment-based methods have been developed to detect GIs, such as the basic local alignment method [14] and whole genome alignment method [15]. These tools rely on the observation that, compared with the conserved regions, the genomic regions that are not aligned across multiple genomes or only aligned with one genome are more likely to be hypothetical GIs. For some complex cases, several methods of constructing and applying multilayer or large-scale genome comparison are reported. For example, MobilomeFINDER first finds shared tRNA genes in several related genomes and then uses Mauve to search for GIs in the upstream and downstream regions of homologous tRNA genes [16]. Since the identified GIs with this

method are related to tRNA disruption, the GIs without the tRNA gene as insertion site will be omitted. In order to solve this problem, MOSAIC has developed a method to identify strain-specific regions that do not necessarily insert tRNA [17]. Unfortunately, inversion and translocation are often mistaken for strain-specific regions. IslandPick is one of the most widely used tools for GI detection [18]. Given a genome, IslandPick first automatically selects the appropriate comparative genomes without any deviation and then uses Mauve to construct the whole genome alignment. To avoid duplication, IslandPick uses BLAST as a secondary filter to recheck the areas aligned by mauve. IslandPick has been integrated into the islandviewer website, where the dataset of precomputed GIs can be downloaded [19–21].

In addition to comparative genomics, component-based methods are also very sensitive to GI detection. Considering that GIs usually show significantly different sequence composition from the host, an effective detection algorithm can distinguish the abnormal region from the rest of the genome according to the composition deviation. In practice, component-based methods are desirable because they can rapidly detect GIs from analyzed sequences without the need for additional genomes. CG content and oligonucleotides with lengths 2-9 are widely used to describe the sequence composition in GI detection [10, 22–25]. For example, PAI-Finder calculates G + C content abnormality and codon usage deviation to detect GIs and further evaluates the candidate PAI only when PAI-like region partially or completely crosses GIs [26]. PAI Finder has been integrated into the PAI database, where comprehensive information of all annotated PAIs and predicted PAI in prokaryotic genome can be downloaded [27, 28]. The HMM model has also been introduced to detect abnormal areas containing component deviations [22, 29–31]. For example, SIGI-HMM constructs an HMM model to remove codons using biased ribosomal regions [29, 30], and IslandPath-DIMoB [31] uses HMM to identify migration genes by searching the PFAM37 migration gene map [32] of each prediction gene [11]. Alien_Hunter introduced a scoring system based on the *k*-mers and refined the boundary of prediction GIs using the HMM model [22].

Although the performance of the above algorithms is good, there are still some problems: (1) the comparative genomics relies heavily on the genomes used in the comparison, and so it can be used in the annotation process or when closely related genomes are available. Even if more genomes are available, researchers have to spend more time on selecting genomes from the species of interest. (2) Although these methods based on HMM show better performance in GI detection, they involve relatively more parameters and a lot of training calculation; so, it takes a long time to detect GIs. (3) In recent years, different sequence features have been proposed, but these features are rarely applied to genome island prediction. How to fuse and select some effective features is also a way to improve the efficiency of genomic island detection.

With the above problems in mind, we present a scheme to predict the genomic islands using the chi-square test and random forest algorithm. We first extract seven kinds of widely used sequence features and compare their performance in GI detection. The chi-square test is then used to select the important features. At last, all the selected features are input into the random forest to detect the genome islands. Through a comprehensive comparison and discussion, some novel valuable guidelines for use of the sequence features, feature selection, and prediction methods are obtained.

2. Materials and Methods

2.1. Datasets. Four standard data sets are used in this study. The first data set, PICK108, consists of 108 complete bacterial genome sequences and their annotations. The number of positive and negative GIs in this dataset is 3868 and 679, respectively [33]. The second set of data is referenced as CF15 which consists of 15 complete bacterial genome sequences and their annotations. The number of positive and negative GIs in this data set is 6070 and 5833, respectively [34]. The third data set, denoted as RGP104, consists of 104 complete bacterial genomes and their annotations. The number of positive and negative GIs is 1846 and 3267, respectively, in this dataset [35].

2.2. Sequence Features. Seven kinds of widely used sequence features are extracted for genome island detection. They are composition of *k*-spaced nucleic acid pairs (CKSNAP), dinucleotide composition (DNC), nucleic acid composition (NAC), pseudodinucleotide composition (PseDNC), electron-ion-interaction pseudopotentials of trinucleotide (PSEIIP), reverse compliment *k*-mer (RCKmer), and trinucleotide composition (TNC). The above features are obtained by iLearn that is a comprehensive python-based toolkit that integrates entity extraction, computation, entity analysis, and construction of predictor variables [36].

2.2.1. Reverse Compliment k-Mer (RCKmer). Reverse compliment *k*-mer is a variant of *k*-mer, which ignores the complementary sequences of adjacent nucleotide sequences. For example, there are 16 types of 2-mer: "AA," "CC," "GG," "TT,""AC," "CA," "GA," "TA," "AG," "CG," "GC," "GT," "AT," "CT," "TC," and "TG.". Because "TT" is the reverse completion *k*-mer of "AA," it can be left out. Therefore, there are only 10 kinds of 2-mer in this method: "AA," "CC," "AC," "CA," "GA," "AG," "CG," "GC," "AT," and "TA." The frequency of each *k*-mer is calculated in turn [37].

2.2.2. Composition of k-Spaced Nucleic Acid Pairs (CKSNAP). CKSNAP feature represents the composition of nucleotide pairs that are separated by k (k=0, 1, 2, 5) nucleotides, and it reflects the short-range interactions of nucleic acids within the sequence [38]. Using $k = 0$ as an example, 16 0-spaced nucleotide pairs (i.e., "AA," "AC," "AG," "AT," "CA," "CC," "CT," "CG," "GA," "GC," "GG," "GT," "TA," "TC," "TG," and "TT") are generated. Then, a feature vector is defined as

$$\left(\frac{N_{\mathrm{AA}}}{N_{\mathrm{Total}}}, \frac{N_{\mathrm{AC}}}{N_{\mathrm{Total}}}, \frac{N_{\mathrm{AG}}}{N_{\mathrm{Total}}}, \frac{N_{\mathrm{AT}}}{N_{\mathrm{Total}}}, \cdots, \frac{N_{\mathrm{TT}}}{N_{\mathrm{Total}}}\right)_{K=0}. \quad (1)$$

In this study, all nucleotide pairs for k (0, 1, …, 5) were considered, and they are encoded to a 96-dimensional digital vector as follows:

$$\left(\frac{N_{AA}}{N_{Total}}, \frac{N_{AC}}{N_{Total}}, \frac{N_{AG}}{N_{Total}}, \cdots, \frac{N_{TT}}{N_{Total}}\right)_{K=0}, \cdots, \left(\frac{N_{AA}}{N_{Total}}, \frac{N_{AC}}{N_{Total}}, \frac{N_{AG}}{N_{Total}}, , \cdots, \frac{N_{TT}}{N_{Total}}\right)_{K=5}. \quad (2)$$

2.2.3. Dinucleotide Composition (DNC). DNC expresses the composition of consecutive pairs of nucleotides [36, 39]. The coding of the DNC characteristics uses 16 descriptors defined as follows:

$$D(i, j) = \frac{N_{(ij)}}{N-1}, i, j \in \{A, C, G, T\}, \quad (3)$$

where N_{ij} donates the number of dinucleotides represented by nucleotide types i and j.

2.2.4. Trinucleotide Composition (TNC). TNC refers to the composition of three consecutive nucleotides in biological sequences [40]. The coding of TNC 64 descriptors described as follows: ("AAA," "AAC," "AAG," "AAT," …, "TTT"), which can be defined as

$$D(i, j, k) = \frac{N_{(ijk)}}{N-2}, i, j, k \in \{A, C, G, T\}, \quad (4)$$

where N_{ijk} donates the number of trinucleotide pairs represented by nucleotide types i, j, and k.

2.2.5. Pseudodinucleotide Composition (PseDNC). PseDNC converts the local sequence arrangement and global sequence information into the feature vector [39]. The PseDNC is expressed as follows:

$$P = (p_1, p_2, \cdots, p_{16}, p_{16+1}, \cdots, p_{16+\lambda})^T,$$
$$p_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j}, & (1 \le k \le 16) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j}, & (17 \le k \le 16+\lambda) \end{cases}, \quad (5)$$

where f_k $(k = 1, 2 \cdots 16)$ reflects the normalized frequency of occurrence of dinucleotides, λ represents the highest counted rank of the correlation along the biological sequences, w (0 to 1) is the weight factor, and θ_j $(j = 1, 2 \cdots \lambda)$ is the j-tier correlation factor, which is defined as

$$\theta_1 = \frac{1}{L-2}\sum_{i=1}^{L-2}\Theta(R_iR_{i+1}, R_{i+1}R_{i+2}),$$
$$\theta_\lambda = \frac{1}{L-1-\lambda}\sum_{i=1}^{L-1-\lambda}\Theta(R_iR_{i+1}, R_{i+\lambda}R_{i+\lambda+1}), \quad (6)$$

where the correlation function is defined as

$$\Theta\left(R_iR_{i+1}, R_jR_{j+1}\right) = \frac{1}{u}\sum_{u=1}^{u}\left(C_u(R_iR_{i+1}) - C_u\left(R_jR_{j+1}\right)\right)^2, \quad (7)$$

where μ denotes the number of physicochemical indexes, $C_u(R_iR_{i+1})$ is the numerical value of the u^{th} physicochemical index of the dinucleotide R_iR_{i+1}, and $C_u(R_jR_{j+1})$ denotes the corresponding value of the dinucleotide R_jR_{j+1} at position j.

2.2.6. Nucleic Acid Composition (NAC). NAC assesses the frequency of each nucleic acid along the sequence. The frequencies of all 4 natural nucleic acids (i.e., "ACGT") can be calculated:

$$f(t) = \frac{N(t)}{N}\ t \in \{A, C, G, T\}, \quad (8)$$

where $N(t)$ represents the number of nucleic acid type t, while N is the length of a nucleotide sequence [36].

2.2.7. Electron-Ion-Interaction Pseudopotentials of Trinucleotide (PSeEIIP). EIIPA, EIIPT, EIIPG, and EIIPC represent the EIIP measurements of nucleotides A, T, G, and C, respectively. The average EIIP of the trinucleotides in each sample is exploited for the construction of the feature vector, which is described as follows:

$$Q = [\mathrm{EIIP}_{AAA} \times \mathrm{f}_{AAA}, \mathrm{EIIP}_{AAc} \times \mathrm{f}_{AAc}, \mathrm{EIIP}_{AAG} \times \mathrm{f}_{AAG}, \mathrm{EIIP}_{AAT} \times \mathrm{f}_{AAT}], \quad (9)$$

where f_{xyz} represents the normalized frequency of the i^{th} trinucleotide, $\mathrm{EIIIP}_{xyz} = \mathrm{EIIP}_x + \mathrm{EIIP}_y + \mathrm{EIIP}_z$ represents the EIIP value of a trinucleotide and $x, y, z \in \{A, C, G, T\}$ [36].

2.3. Chi-Square Test. All kinds of sequence features will be fused together in order to improve the prediction efficiency, but the redundancy of different features cannot be ignored. Therefore, one of the primary tasks involved in genomic island prediction is to select the best features from the given dataset to achieve the best prediction. This work uses the chi-square test to select the best features for genomic island prediction.

The chi-square $(\mathcal{X}^2)$ test measures the deviation from the expected distribution [40, 41]. Statistically, $\mathcal{X}^2$ tests the independence of two variables, where two variables A and B are defined as independent if $P(AB) = P(A)P(B)$ or $P(A \mid B) = P(A)$ $(P(B \mid A) = P(B))$. In feature selection, the two variables are the term occurrence and the class occurrence. The terms in relation to the quantity are classified as follows:

$$X^2(D, i, j) = \sum_{w_i \in \{0,1\}}\sum_{w_j \in \{0,1\}}\frac{\left(N_{w_iw_j} - F_{w_iw_j}\right)^2}{F_{w_iw_j}}, \quad (10)$$

where N is the observed frequency in D and F. w_i and w_j are defined as

$$I(U, C) = \sum_{w_i \in \{1.0\}}\sum_{w_j \in \{1.0\}} P\left(U = w_i, C = w_j\right)\log_2\frac{P\left(U = \mathrm{w}_i, C = w_j\right)}{P(U = w_i)P\left(C = w_j\right)}, \quad (11)$$

where U is a random variable that takes values $w_i = 1$ (the

presence of the feature i) and $w_i = 0$ (absence of the feature i), and C is a random variable that takes values $e_j = 1$ (the presence of the feature in class j) and $e_j = 0$ (absence of the feature in class j). We write U_i and U_j if it is not clear from context which features i and class j we are referring to and got the following equation:

$$I(U,C) = \frac{F_{11}}{F}\log_2\frac{FF_{11}}{F_1F_1} + \frac{F_{01}}{F}\log_2\frac{FF_{01}}{F_0F_1} + \frac{F_{10}}{F}\log_2\frac{FF_{10}}{F_1F_0} + \frac{F_{00}}{F}\log_2\frac{FF_{00}}{F_0F_0}, \tag{12}$$

where the N are counts of features that have the values of w_i and w_j that are indicated by the two subscripts. For example, F_{10} is the number of features that contain i ($w_i = 1$) and are not in j($w_j = 0$). $F_1 = F_{10} + F_{11}$ is the number of features that contain i ($w_i = 1$), and we count features independent of class membership $w_i \in \{0, 1\}$. $F = F_{00} + F_{01} + F_{10} + F_{11}$ is the total number of documents [42].

$\mathcal{X}^2$ is a measure of how much expected counts E and observed counts N deviate from each other. A high value of $\mathcal{X}^2$ indicates that the hypothesis of independence, which implies that expected and observed counts are similar, is incorrect. An arithmetically simpler way of computing $\mathcal{X}^2$ is the following:

$$X^2(D,i,j) = \frac{(F_{11} + F_{10} + F_{01} + F_{00}) * (F_{11} + F_{00} - F_{10}F_{01})^2}{(F_{11} + F_{01}) * (F_{11} + N_{10}) * (F_{10} + F_{00}) * (F_{01} + F_{00})}. \tag{13}$$

2.4. Prediction Algorithm. Random forest (RF) is among the best classification algorithms and widely applied to manage many biological problems. It works by building small groups of weak classifiers, to finally combine them and form a strong classifier. This is a configuration learning method that can build models that create multiple decision trees during training and will remove modal classes from classes predicted by a single tree. It is a fusion of tree predictors, where each tree depends on the value of an independent sampled random vector and the same distribution of all trees in the forest [43].

A random forest is a collection of tree predictor $h(X;\omega_i)$, $i = 1, \cdots, I$, where X represents the observed input (covariate) vector of length p with associated random vector X and ω_i. They are independent and identically distributed (iid) random vectors. As mentioned, we focus on the regression setting for which we have a numerical outcome Y, but we make some points of contact with classification (categorical outcome) problems [44]. The observed (training) data is assumed to be independently drawn from the joint distribution of (X, Y) and comprises $n(p+1)$-tuples $X(x_1, y_1), \cdots, (x_n, y_n)$.

For regression, the random forest prediction is the weighted average over the collection

$$h(y) = \left(\frac{1}{k}\right)\sum_{i=1}^{I} h(X;\omega_i). \tag{14}$$

As $i \to \infty$, the law of large numbers ensures

$$E_{X,Y}\left(Y - \bar{h}(X)^2 \to E_{X,Y}\left(Y - E_\omega\bar{h}(X,\omega)\right)\right)^2. \tag{15}$$

The quantity on the right is the prediction (or generalization) error for the random forest, denoted as PE_f^*. The convergence implies that random forests do not overfit. Now, define the average prediction error for an individual tree $h(X, \omega)$

$$PE_t^* = E_\omega E_{X,Y}(Y - h(X,\omega))^2. \tag{16}$$

Assume that for all the tree is unbiased, i.e., $EY = E_X h(X, \omega)$. Then,

$$PE_f^* \le \bar{\mu}PE_t^*, \tag{17}$$

where $\bar{\mu}$ is the weighted correlation between residuals $Y - h(X, \omega)$ and $h(X;\omega)$ for independent ω, ω^k. The above inequality pinpoints what is required for accurate random forest regression: low correlation between residuals of differing tree members of the forest and low prediction error for the individual trees [44]. Further, the random forest will decrease the individual tree error (PE_t^*), by the factor $\bar{\mu}$.

2.5. Performance Evaluation. This work introduces crossvalidation to evaluate the proposed method and calculates accuracy, recall, F-measure, precision specificity, sensitivity, and precision as standard performance indicators. They are defined as follows:

$$\begin{aligned}
\text{Acc} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100,\\
\text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100,\\
\text{Prec} &= \frac{\text{TP}}{\text{TP} + \text{FP}},\\
\text{Sn} &= \frac{\text{TP}}{\text{TP} + \text{FN}},\\
\text{Sp} &= \frac{\text{TN}}{\text{TN} + \text{FP}},\\
F1 &= \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}},\\
\text{MCC} &= \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},
\end{aligned} \tag{18}$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN is the number of false negatives.

3. Results and Discussion

3.1. Performance of the Proposed Prediction Method. To build the prediction model, seven kinds of sequence features are extracted, fused, and filtered by the chi-square test and then

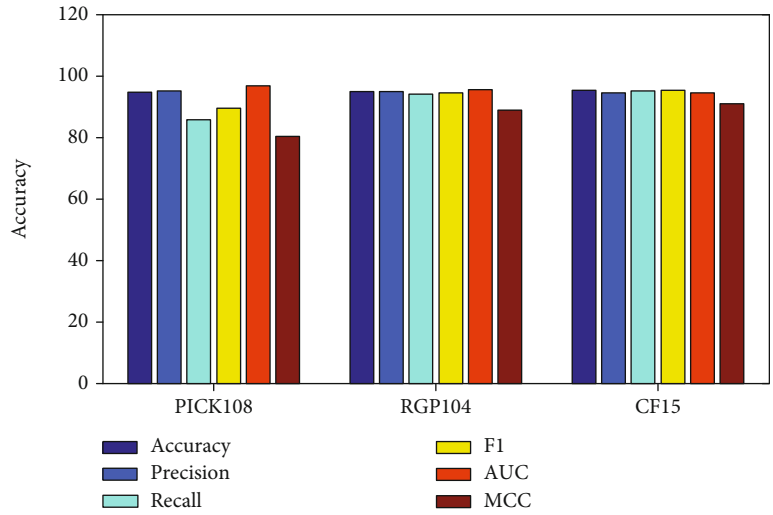

Figure 1: Comparison of the accuracy, precision, recall, F1, AUC, and MCC of the PICK108, CF15, and RGP104 datasets.

input into the random decision tree for genomic island prediction. Accuracy, F1, MCC, precision, recall, and AUC are calculated based on 10 times crossvalidation, which are summarized in Figure 1.

Figure 1 shows that the proposed method achieves good performance among four datasets. As for PICK108, its accuracy, precision, recall, F1, AUC, and MCC are 94.6%, 95.1%, 85.7%, 89.5%, 96.8%, and 80.3%, respectively. For dataset CF15, the overall precision is 94.9%, and precision, recall, F1, AUC, and MCC are 94.8%, 94.0%, 94.4%, 95.6%, and 88.8%, respectively. As for RGP104, its accuracy, precision, recall, F1, AUC, and MCC are 95.4%, 94.4%, 95.2%, 95.4%, 94.5%, and 90.9%, respectively.

We further compare the proposed method with the current methods. For the convenience of comparison, we compare our results with that of the published results with the existing methods. Therefore, different datasets choose different evaluation methods, which are summarized in Tables 1–3.

As for PICK108, the proposed method is compared with the Centroid [45], INDeGenIUS [46], MTGIpick [33], SigHunt [47], and Zisland Explore [48]. Table 1 indicates that the proposed method achieves the highest accuracy, precision, and recall with the values of 94.6%, 95.1%, and 85.7%, respectively. Compared with the second best method, the accuracy, precision, and recall of the proposed method are 8.4%, 22.3%, and 38.5% higher than that of MTGIpick, respectively.

In the RGP104 dataset, PanRGP [35], IslandViewer [19, 20], IslandPath-Dimob [31], IslandCafe, and SIGI-HMM [29, 30] are compared with the proposed method. Table 2 shows that the proposed method outperforms the others in term of MCC, F1, accuracy, and recall. Specifically, the MCC, F1, ACC, and recall of the proposed method are 11%, 12.4%, 3.2%, and 15.2%, respectively, higher than that of the PanRGP model [35], but its accuracy is 0.1% lower than that of the PanRGP model.

In the CF15 experiment, IslandCafe [34], IslandViewer [19, 20], IslandPath-Dimob [31], Zisland Explorer [48] and

Table 1: Comparison of the proposed method with other reported results on the PICK108 dataset.

Method	Accuracy	Precision	Recall
Centroid	82.4	61.4	27.6
INDeGenIUS	82.4	67.9	19.9
MTGIpick	86.2	72.8	47.2
SigHunt	80.5	51.0	24.0
Zisland Explorer	83.8	75.9	25.5
This paper	94.6	95.1	85.7

Table 2: Comparison of the proposed method with other reported results on the RGP104 dataset.

Method	MCC	F1	ACC	Precision	Recall
PanRGP	77.8	80.9	92.4	94.9	76.4
IslandViewer	76.2	82.0	91.1	90.8	78.8
IslandPath	52.3	57.0	78.1	89.1	47.7
IslandCafe	37.7	44.4	76.1	76.9	35.5
SIGI-HMM	33.8	45.5	75.6	65.5	37.6
This paper	88.8	94.4	95.6	94.8	94.0

Table 3: Comparison of the proposed method with other reported results on the CF15 dataset.

Method	Recall	Precision	F1	MCC
IslandCafe	71.0	61.0	66.0	62.0
IslandViewer	72.0	59.0	65.0	59.0
IslandPath-Dimob	53.0	67.0	59.0	55.0
Zisland Explorer	45.0	56.0	50.0	46.0
SIGI-HMM	24.0	57.0	33.0	32.0
This paper	95.4	95.4	95.4	90.9

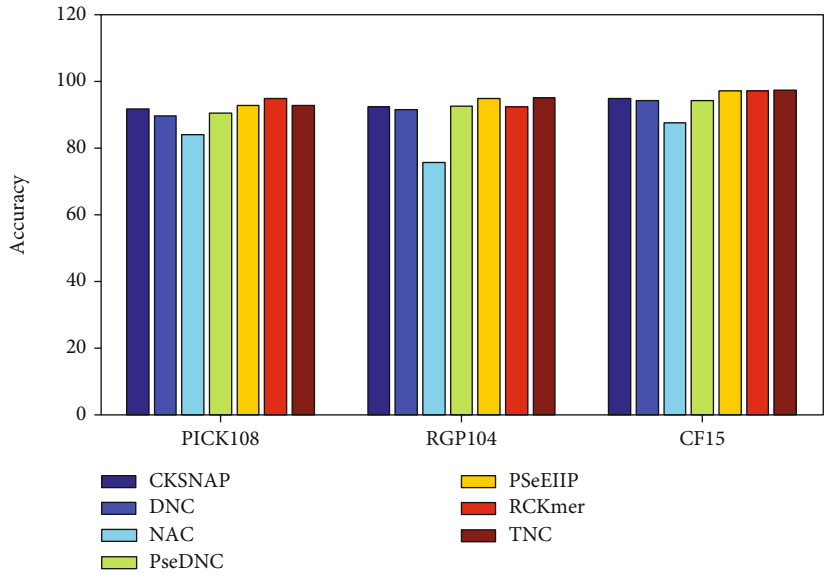

Figure 2: Comparison of the overall prediction accuracies of seven kinds of the sequence features.

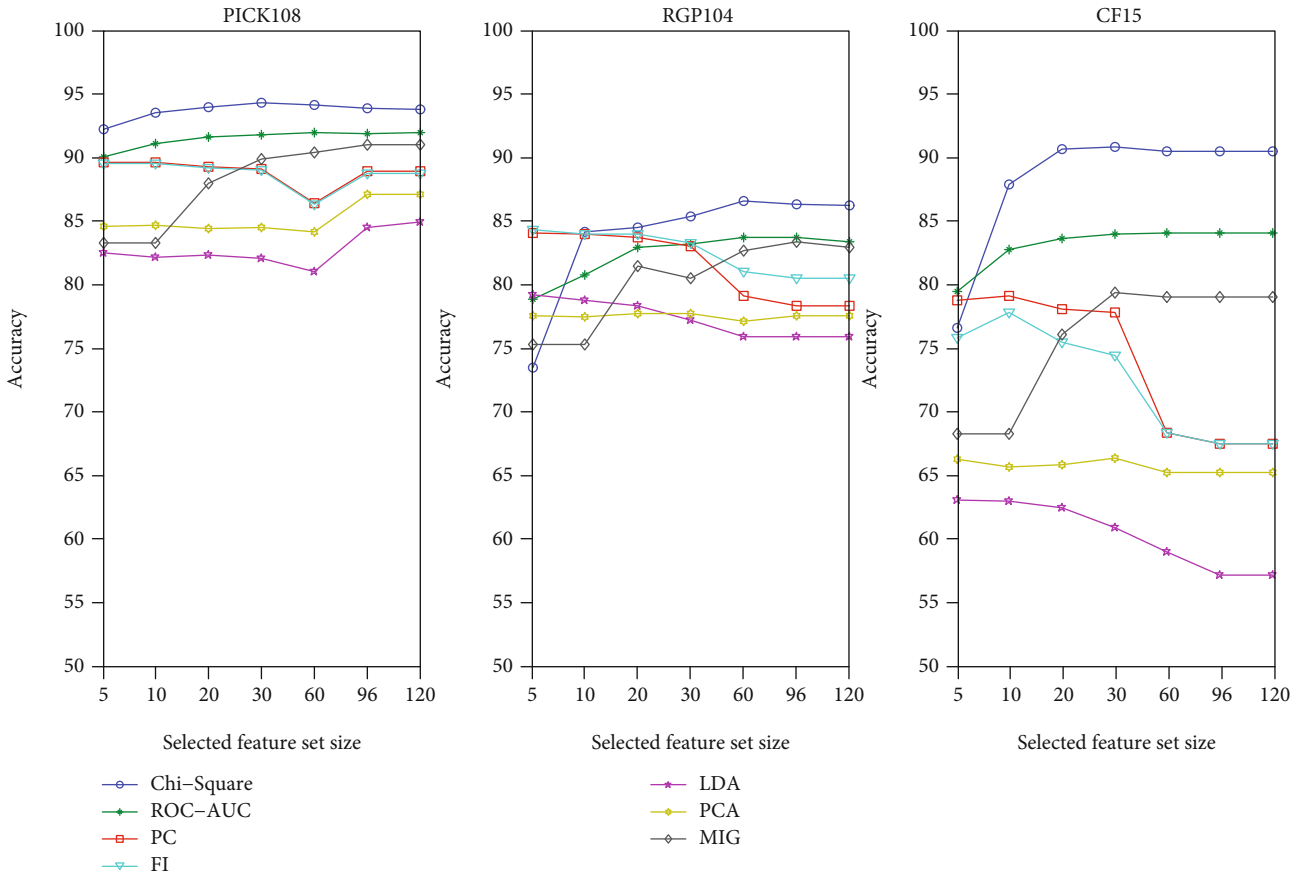

Figure 3: The comparison of the overall accuracies of all experiments with the selected feature sets for three datasets.

SIGI-HMM [29, 30] are compared with the proposed method. Table 3 indicates that the proposed method achieves the highest recall, precision, F1, and MCC with the values of 95.4%, 95.4%, 95.4%, and 90.9%, respectively, which are 23.4%, 28.4%, 29.4%, and 28.9% higher than that of the next competitive method [34].

The above results show that the proposed method outperforms the available genomic island prediction methods,

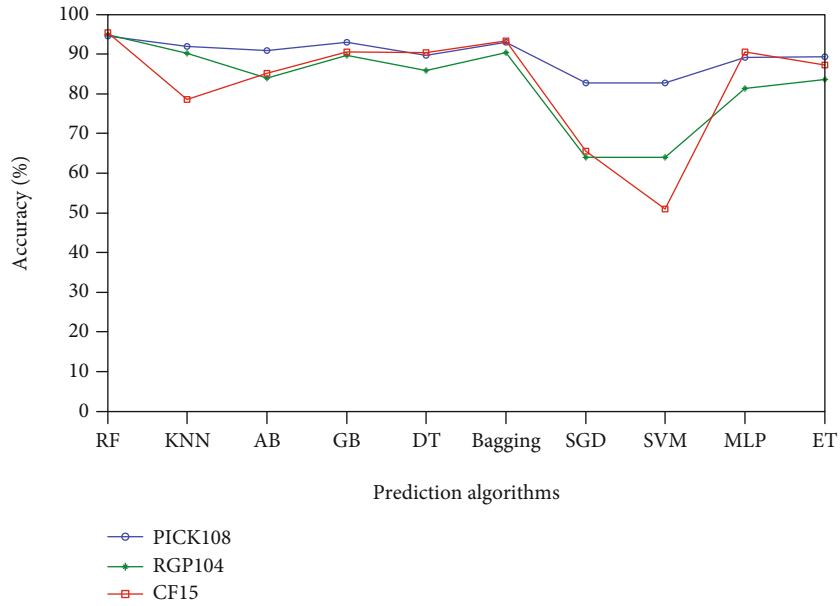

Figure 4: The comparison of the overall accuracies of different prediction algorithms with the selected feature sets for three datasets.

indicating that the combination of different features, feature selection based on the chi-square test, and prediction algorithm is very effective to advance the prediction. This understanding can be used to develop more powerful genomic island prediction methods.

3.2. Influence of the Different Features. To predict genomic islands, we use seven kinds of protein features: reverse compliment k-mer (RCKmer), composition of k-spaced nucleic acid pairs (CKSNAP), dinucleotide composition (DNC), trinucleotide composition (TNC), pseudodinucleotide composition (PseDNC), nucleic acid composition (NAC), and electron-ion-interaction pseudopotentials of trinucleotide (PSeEIIP). To evaluate the contribution of each kind of the sequence features, we present the comparison of the accuracies of seven kinds of the sequence features in Figure 2.

Figure 2 indicates that each feature makes its own positive contributions to the predictions; although, different features have certain preferences for different data sets. On the whole, PSeEIIP, RCKmer, and TNC achieve the best performance among all kinds of the sequence features. It is easy to note that PSeEIIP and RCKmer not only reflect the content of components but also focus the local sequence arrangement and global sequence information and calculate the energy of delocalized electrons in nucleotides as the electron-ion interaction. Compared with the ANC and DNC, PSeEIIP and RCKmer are more closely related to the genomic islands, and this is why they achieve the better performance in the genomic island prediction.

3.3. Influence of the Different Feature Selections. A feature of the proposed method is the feature selection based on the chi-square test. For a better understanding of the feature selection, we select the feature set with size from 5 to 120. All experiments are performed with each selected feature set using the 10 times crossvalidation test, and overall accuracy is chosen to represent the score in this prediction. Figure 3 is the overall accuracies of all experiments with the selected feature sets for three datasets.

As would be expected, the overall accuracy first increases and then decreases as the selected feature size continues to increase. When the selected feature set size is less than 30, all data sets have reached the best prediction. As the increase of the number of selected features, the overall accuracy decreases. The chi-square is further compared with feature importance (FI), Pearson correlation (PC), ROC-AUC, mutual information gain (MIG), linear discriminant analysis (LDA), and principal component analysis (PCA), and it is easy to note that the chi-square test achieves the best performance among seven feature selection method.

3.4. Influence of the Different Prediction Algorithms. Random forest (RF) was employed as a classifier in this work. To compare different classifiers' performance, support vector machine (SVM), k-nearest neighbor (KNN), gradient boosting (GB), adaBoost (AB), decision tree (DT), bagging, extra trees (ET), stochastic gradient descent (SGD), and layer perceptron (MLP) were also adopted for protein structural class prediction. All experiments are performed with each selected feature set using the 10 times crossvalidation test, and overall accuracy is chosen to represent the score in this prediction. Figure 4 summarizes the overall accuracies of all experiments with the different prediction algorithms for three datasets.

From Figure 4, it is easy to note that the random forest (RF) achieves the best performance among the ten classifiers. Specifically, the average overall prediction accuracy is 95% for PICK108, RGP104, and CF15 datasets compared with 91% of the gradient boosting (GB) and 92% of the bagging. These results indicate that the random forest is a more powerful classifier for the genomic island prediction.

4. Conclusion

Genome islands are related to the rapid adaptation of prokaryotes, which have important medical, economic, or environmental significance. Some methods usually evaluate all features and focus on whether the local features of a certain area are significantly different from the host. Although these methods have achieved good experimental results, various feature extraction methods have been proposed, but they are rarely used to predict genomic islands. With these problems in mind, we present a scheme to predict the genomic islands using the chi-square test and random forest algorithm. We extract seven kinds of widely used sequence features and select the important features with the chi-square test. At last, all the selected features are input into the random forest to predict the genome islands. Three experiment results show that the proposed method has better performance than previous methods.

The first contribution can be seen from the influence of the different features, and we find that PSeEIIP, RCKmer, and TNC are more closely related to the genomic islands and achieve the best performance among all kinds of the sequence features. The second contribution can be indicated from the influence of the different feature selections, and the chi-square test achieves the best performance among seven feature selection method. The final contribution can be seen from the influence of the different prediction algorithms, and we notice that the random forest (RF) achieved the best performance among the ten classifiers; its accuracy is 3% higher than that of the next one. This understanding can be then used to develop more powerful methods for genomic island prediction.

Data Availability

All the data used to support the findings of this study are available on https://github.com/Onesime243/Chi_square_Genomic_Islands_predicton_data-and-result.git.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61772028) and research Grants from Zhejiang Provincial Natural Science Foundation of China (LY20F020016).

References

[1] J. Hacker, L. Bender, M. Ott et al., "Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extra intestinal Escherichia coli isolates," *Microbial Pathogenesis*, vol. 8, no. 3, pp. 213–225, 1990.

[2] J. Hacker and J. B. Kaper, "Pathogenicity islands and the evolution of microbes," *Annual Reviews in Microbiology*, vol. 54, no. 1, pp. 641–679, 2000.

[3] Y. Nakamura, T. Itoh, H. Matsuda, and T. Gojobori, "Biased biological functions of horizontally transferred genes in prokaryotic genomes," *Nature Genetics*, vol. 36, no. 7, pp. 760–766, 2004.

[4] O. Gal-Mor and B. B. Finlay, "Pathogenicity islands: a molecular toolbox for bacterial virulence," *Cellular Microbiology*, vol. 8, no. 11, pp. 1707–1719, 2006.

[5] U. Dobrindt, B. Hochhut, U. Hentschel, and J. Hacker, "Genomic islands in pathogenic and environmental microorganisms," *Nature Reviews Microbiology*, vol. 2, no. 5, pp. 414–424, 2004.

[6] J. G. Lawrence, "Common themes in the genome strategies of pathogens," *Current Opinion in Genetics & Development*, vol. 15, no. 6, pp. 584–588, 2005.

[7] J. M. Manson and M. S. Gilmore, "Pathogenicity island integrase cross-talk: a potential new tool for virulence modulation," *Molecular Microbiology*, vol. 61, no. 3, pp. 555–559, 2006.

[8] B. Middendorf, B. Hochhut, K. Leipold, U. Dobrindt, G. Blum-Oehler, and J. Hacker, "Instability of pathogenicity islands in uropathogenic Escherichia coli 536," *Journal of Bacteriology*, vol. 186, no. 10, pp. 3086–3096, 2004.

[9] B. B. Finlay and S. Falkow, "Common themes in microbial pathogenicity revisited," *Microbiology and Molecular Biology Reviews*, vol. 61, no. 2, pp. 136–169, 1997.

[10] S. Karlin, "Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes," *Trends in Microbiology*, vol. 9, no. 7, pp. 335–343, 2001.

[11] W. W. Hsiao, K. Ung, D. Aeschliman, J. Bryan, B. B. Finlay, and F. S. Brinkman, "Evidence of a large novel gene pool associated with prokaryotic genomic islands," *PLoS Genetics*, vol. 1, no. 5, article e62, 2005.

[12] G. S. Vernikos and J. Parkhill, "Resolving the structural features of genomic islands: a machine learning approach," *Genome Research*, vol. 18, no. 2, pp. 331–342, 2008.

[13] M. A. Ragan, "Detection of lateral gene transfer among microbial genomes," *Current Opinion in Genetics & Development*, vol. 11, no. 6, pp. 620–626, 2001.

[14] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[15] A. C. Darling, B. Mau, F. R. Blattner, and N. T. Perna, "Mauve: multiple alignment of conserved genomic sequence with rearrangements," *Genome Research*, vol. 14, no. 7, pp. 1394–1403, 2004.

[16] H.-Y. Ou, L.-L. Chen, J. Lonnen et al., "A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria," *Nucleic Acids Research*, vol. 34, no. 1, pp. e3–e3, 2006.

[17] H. Chiapello, I. Bourgait, F. Sourivong et al., "Systematic determination of the mosaic structure of bacterial genomes: species backbone versus strain-specific loops," *BMC Bioinformatics*, vol. 6, no. 1, p. 171, 2005.

[18] M. G. Langille, W. W. Hsiao, and F. S. Brinkman, "Evaluation of genomic island predictors using a comparative genomics approach," *BMC Bioinformatics*, vol. 9, no. 1, p. 329, 2008.

[19] M. G. Langille and F. S. Brinkman, "IslandViewer: an integrated interface for computational identification and visualization of genomic islands," *Bioinformatics*, vol. 25, no. 5, pp. 664-665, 2009.

[20] B. K. Dhillon, T. A. Chiu, M. R. Laird, M. G. Langille, and F. S. Brinkman, "IslandViewer update: improved genomic island discovery and visualization," *Nucleic Acids Research*, vol. 41, no. W1, pp. W129–W132, 2013.

[21] A. J. Arvey, R. K. Azad, A. Raval, and J. G. Lawrence, "Detection of genomic islands via segmental genome heterogeneity," *Nucleic Acids Research*, vol. 37, no. 16, pp. 5255–5266, 2009.

[22] G. S. Vernikos and J. Parkhill, "Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands," *Bioinformatics*, vol. 22, no. 18, pp. 2196–2203, 2006.

[23] S. Karlin, J. Mrázek, and A. M. Campbell, "Codon usages in different gene classes of the Escherichia coli genome," *Molecular Microbiology*, vol. 29, no. 6, pp. 1341–1355, 1998.

[24] R. Sandberg, G. Winberg, C.-I. Bränden, A. Kaske, I. Ernberg, and J. Cöster, "Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier," *Genome Research*, vol. 11, no. 8, pp. 1404–1409, 2001.

[25] A. Tsirigos and I. Rigoutsos, "A new computational method for the detection of horizontal gene transfer events," *Nucleic Acids Research*, vol. 33, no. 3, pp. 922–933, 2005.

[26] S. H. Yoon, C.-G. Hur, H.-Y. Kang, Y. H. Kim, T. K. Oh, and J. F. Kim, "A computational approach for identifying pathogenicity islands in prokaryotic genomes," *BMC Bioinformatics*, vol. 6, no. 1, pp. 1–11, 2005.

[27] S. H. Yoon, Y.-K. Park, S. Lee et al., "Towards pathogenomics: a web-based resource for pathogenicity islands," *Nucleic Acids Research*, vol. 35, no. Database, suppl_1, pp. D395–D400, 2007.

[28] S. H. Yoon, Y.-K. Park, and J. F. Kim, "PAIDB v2. 0: exploration and analysis of pathogenicity and resistance islands," *Nucleic Acids Research*, vol. 43, no. D1, pp. D624–D630, 2015.

[29] R. Merkl, "SIGI: score-based identification of genomic islands," *BMC Bioinformatics*, vol. 5, no. 1, pp. 1–14, 2004.

[30] S. Waack, O. Keller, R. Asper et al., "Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models," *BMC Bioinformatics*, vol. 7, no. 1, pp. 1–12, 2006.

[31] W. Hsiao, I. Wan, S. J. Jones, and F. S. Brinkman, "IslandPath: aiding detection of genomic islands in prokaryotes," *Bioinformatics*, vol. 19, no. 3, pp. 418–420, 2003.

[32] R. D. Finn, J. Tate, J. Mistry et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 36, suppl_1, pp. D281–D288, 2007.

[33] Q. Dai, C. Bao, Y. Hai et al., "MTGIpick allows robust identification of genomic islands from a single genome," *Briefings in Bioinformatics*, vol. 19, no. 3, pp. bbw118–bbw373, 2016.

[34] M. Jani and R. K. Azad, "IslandCafe: compositional anomaly and feature enrichment assessment for delineation of genomic islands," *G3 Genes Genomes Genetics*, vol. 9, no. 10, pp. 3273–3285, 2019.

[35] A. Bazin, G. Gautreau, C. Médigue, D. Vallenet, and A. Calteau, "panRGP: a pangenome-based method to predict genomic islands and explore their diversity," *Bioinformatics*, vol. 36, Supplement_2, pp. i651–i658, 2020.

[36] Z. Chen, P. Zhao, F. Li et al., "iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data," *Briefings in Bioinformatics*, vol. 21, no. 3, pp. 1047–1057, 2020.

[37] Q. Li, L. Xu, Q. Li, and L. Zhang, "Identification and classification of enhancers using dimension reduction technique and recurrent neural network," *Computational and Mathematical Methods in Medicine*, vol. 2020, 9 pages, 2020.

[38] Y. Zhou, P. Zeng, Y.-H. Li, Z. Zhang, and Q. Cui, "SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features," *Nucleic Acids Research*, vol. 44, no. 10, pp. e91–e91, 2016.

[39] D. Che, T. Shafer, and P. Tian, "Classification of endangered languages using decision tree based algorithms," in *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pp. 1814–1821, Guilin, China, 2017.

[40] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 80–89, 2004.

[41] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[42] P. P. Shinde and S. Shah, "A review of machine learning and deep learning applications," in *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, pp. 1–6, Pune, India, 2018.

[43] M. R. Segal, "Machine Learning Benchmarks and Random Forest Regression," *UCSF: Center for Bioinformatics and Molecular Biostatistics*, 2004.

[44] X.-B. Wang, L.-Y. Wu, Y.-C. Wang, and N.-Y. Deng, "Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs," *Protein Engineering, Design & Selection*, vol. 22, no. 11, pp. 707–712, 2009.

[45] I. Rajan, S. Aravamuthan, and S. S. Mande, "Identification of compositionally distinct regions in genomes using the centroid method," *Bioinformatics*, vol. 23, no. 20, pp. 2672–2677, 2007.

[46] S. Shrivastava, C. V. Siva Kumar Reddy, and S. S. Mande, "INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms," *Journal of Biosciences*, vol. 35, no. 3, pp. 351–364, 2010.

[47] K. S. Jaron, J. C. Moravec, and N. Martínková, "SigHunt: horizontal gene transfer finder optimized for eukaryotic genomes," *Bioinformatics*, vol. 30, no. 8, pp. 1081–1086, 2014.

[48] W. Wei, F. Gao, M.-Z. Du, H.-L. Hua, J. Wang, and F.-B. Guo, "Zisland Explorer: detect genomic islands by combining homogeneity and heterogeneity properties," *Briefings in Bioinformatics*, vol. 18, no. 3, pp. 357–366, 2017.