frontiers in Genetics

TECHNOLOGY AND CODE
published: 18 September 2019
doi: 10.3389/fgene.2019.00842

iRO-PsekGCC: Identify DNA Replication Origins Based on Pseudo k-Tuple GC Composition

Bin Liu[1,2*†], *Shengyu Chen*[3†], *Ke Yan*[4] *and Fan Weng*[4]

[1] *School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China,* [2] *Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing, China,* [3] *School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, IN, United States,* [4] *School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China*

Summary: Identification of replication origins is playing a key role in understanding the mechanism of DNA replication. This task is of great significance in DNA sequence analysis. Because of its importance, some computational approaches have been introduced. Among these predictors, the iRO-3wPseKNC predictor is the first discriminative method that is able to correctly identify the entire replication origins. For further improving its predictive performance, we proposed the Pseudo k-tuple GC Composition (PsekGCC) approach to capture the "GC asymmetry bias" of yeast species by considering both the GC skew and the sequence order effects of *k*-tuple GC Composition (*k*-GCC) in this study. Based on PseKGCC, we proposed a new predictor called iRO-PsekGCC to identify the DNA replication origins. Rigorous jackknife test on two yeast species benchmark datasets (*Saccharomyces cerevisiae*, *Pichia pastoris*) indicated that iRO-PsekGCC outperformed iRO-3wPseKNC. It can be anticipated that iRO-PsekGCC will be a useful tool for DNA replication origin identification.

Availability and implementation: The web-server for the iRO-PsekGCC predictor was established, and it can be accessed at http://bliulab.net/iRO-PsekGCC/.

Keywords: replication origin identification, pseudo *k*-tuple GC composition, random forest, web-server, DNA sequence analysis

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Xun Lan,
Tsinghua University, China
Qiwen Dong,
East China Normal University, China

Correspondence:
Bin Liu
bliu@bliulab.net

[†] *These authors have contributed equally to this work and share first authorship*

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *05 July 2019*
Accepted: *13 August 2019*
Published: *18 September 2019*

Citation:
Liu B, Chen S, Yan K and Weng F (2019) iRO-PsekGCC: Identify DNA Replication Origins Based on Pseudo k-Tuple GC Composition. Front. Genet. 10:842. doi: 10.3389/fgene.2019.00842

INTRODUCTION

In the process of the cell cycle, DNA replication is one of the most important steps (Shirahige et al., 1998). Since the DNA replication is initiated from a specific region, which is called replication origin, identifying the DNA replication origin is especially important for studying drug developments, cell life activities, genetic engineering, etc. (Méchali, 2010). Experimental methods detect the replication origins by using Chromatin immunoprecipitation (Chip) with high cost (Lubelsky et al., 2012). Therefore, researchers are seeking computational methods to efficiently predict the replication origins only based on the sequence information. Compared with non-replication origins, replication origins show uneven distribution of G (guanine) and C (cytosine) (Lobry, 1996), and the concept of "GC Skew" (Grigoriev, 1998) was proposed. Later, some computational methods incorporated these characteristics into the predictors based on the replication origins (Zhang and Zhang, 1991; Zhang and Zhang, 1994; Grigoriev, 1998; Roten et al., 2002; Thomas et al., 2007; Gao and Zhang, 2008; Luo et al., 2014; Bu et al., 2018). In order to further improve the predictive performance, the discriminative methods were proposed by using both the information of the positive and negative samples (Chen et al., 2012; Li et al., 2015; Zhang et al., 2016), and all of these methods mentioned above achieved the-state-of-the-art performance. A recent method iRO-3wPseKNC incorporated the "GC asymmetry bias" (Lobry, 1996; Grigoriev, 1998; Lubelsky et al., 2012; Li et al., 2014) into the prediction by representing the entire replication origins based on three-window-based PseKNC (3wPseKNC) (Liu et al., 2018b). Feature extraction methods are the keys for the performance improvement. In this regard, many features have been proposed, which can be easily generated by some software tools.

These existing computational methods have significantly enhanced the development of this hot area, but they all suffer from certain disadvantages or limitations, for example, as discussed above the GC Skew is an important feature of replication origins, but all the existing discriminative methods failed to directly use GC Skew to construct the predictors. Furthermore, the existing feature extraction methods cannot reflect the uneven distribution of G and C. To solve these problems, we followed the framework of iRO-3wPseKNC (Liu et al., 2018b), and proposed an improved predictor called iRO-PsekGCC for replication origin identification. iRO-PsekGCC cannot only capture the CG asymmetry bias by using *k*-tuple GC composition (or *k*-GCC), but can also incorporate the GC Skew into the concept of PseKNC (Chen et al., 2014a; Chen et al., 2014b).

MANUSCRIPT FORMATTING

Benchmark Datasets

In order to evaluate the performance of the proposed method, two recently established benchmark datasets of the *Saccharomyces cerevisiae* and *Pichia pastoris* (Liu et al., 2018b) were employed in this study, because they showed clear CG asymmetry distributions, which can be represented as:

$$\mathbb{S}_\tau = \mathbb{S}_\tau^+ \cup \mathbb{S}_\tau^-, \ \tau = \begin{cases} 1 \text{ for } \textit{Saccharomyces cerevisiae} \\ 2 \text{ for } \textit{Pichia pastoris} \end{cases} \tag{1}$$

where the symbol ∪ represents the union, and $\mathbb{S}_-^+$ represents the positive dataset containing 340 replication origins, and $\mathbb{S}_1^-$ represents the negative dataset containing 342 non-replication origins; 305 replication origins are in positive dataset $\mathbb{S}_2^+$, and 302 non-replication origins are in the negative dataset $\mathbb{S}_2^-$. For both of the two benchmark datasets, the redundant samples have been removed by using CD-HIT software tool (Li and Godzik, 2006) with the most stringent cut-off threshold (80%).

Pseudo *k*-Tuple GC Composition (PsekGCC)

One of the key steps for constructing machine-learning predictors for analyzing biological sequences is feature extraction. Following the framework of three-window-based PseKNC (3wPseKNC) (Liu et al., 2018b), we proposed a feature extraction method called "Pseudo k-tuple GC composition (PseKGCC)" to directly incorporate the CG asymmetry bias (Lobry, 1996; Grigoriev, 1998; Lubelsky et al., 2012; Li et al., 2014) and GC skew (Grigoriev, 1998) into the predictor. In the following sections, we will introduce how to represent DNA samples by using PseKGCC.

A DNA sequence D can be formulated as follow:

$$\mathbf{D} = \mathrm{N}_1\mathrm{N}_2\mathrm{N}_3\cdots\mathrm{N}_i\cdots\mathrm{N}_L \quad (i = 1,2,3\cdots,L) \tag{2}$$

where *L* denotes the length of **D**, and

$$N_i \in \{\mathrm{A(adenine),C(cytosine),G(guanine),} \\ \mathrm{T(thymine)}\}, \quad (i = 1,2,3,\cdots,L) \tag{3}$$

which represents the *i*-th nucleobase in the sequence, and fi ∈ denotes the "member of"" in the set theory. Following the study (Liu et al., 2018b), **D** is divided into three windows by two parameters ε and δ, including front window, middle window, and rear window respectively. ε and 1 – δ denote the percentage of total nucleobases of **D** in the front window and rear window, respectively. The front window, middle window and rear window can be represented as **D**[1,η], D[η + 1,ξ], and **D**[ξ + 1, *L*], respectively, where η and ξ are defined as (Liu et al., 2018b),

$$\begin{cases} \eta = \mathrm{Int}^{\mathrm{C}}[L \times \varepsilon] \\ \xi = \mathrm{Int}^{\mathrm{C}}[L \times \delta] \end{cases}, \quad (0 < \varepsilon < \delta < 1.0) \tag{4}$$

where the symbol $\mathrm{Int}^{\mathrm{C}}$ represents the ceiling operator, which means to return the smallest integer value greater than or equal to the float number .

According to (Liu et al., 2018b), if D is formulated by the *k*-tuple nucleotide (or *k*-mer) (Liu et al., 2019b; Liu, 2017) based on the three windows strategy, it can be represented as follow:

$$\mathbf{D} = \Big[f_1^{(1)} \cdots f_v^{(1)} \cdots f_{4^k}^{(1)} f_{4^k+1}^{(2)} \cdots f_{4^k+v}^{(2)} \cdots f_{4^k+v}^{(2)} \cdots \\ f_{2\times4^k}^{(2)} f_{2\times4^k+1}^{(3)} \cdots f_{2\times4^k+v}^{(3)} \cdots f_{3\times4^k}^{(3)} \Big]^{\mathbf{T}} \tag{5}$$

where in vector operations, symbol 'T' denotes the transformation symbol, and in the sample D, the normalized frequency values of the corresponding *k*-tuple nucleotides appearing in the front window, middle window and rear window are represented as $f^{(1)}$, $f^{(2)}$, $f^{(3)}$, respectively. The feature vector's dimension is 3×4^k.

This strategy was proposed to capture the patterns of "GC asymmetry bias" in yeast species genomes, and it is able to improve the predictive performance for identifying replication origins among multiple yeast species genomes. However, this approach has the following disadvantages: 1) the three windows strategy can only capture the local GC asymmetry bias of replication origins, but it cannot incorporate the GC asymmetry bias in a global fashion; 2) for large *k* values of *k*-tuple nucleotide, the dimension of the resulting feature vectors is high, which will cause high dimension disaster.

In order to overcome these disadvantages, we proposed a new composition of DNA sequence called "*k*-tuple GC composition (or *k*-GCC)" to capture the GC preference in the replication origins and their global interactions. *k*-GCC treats A (adenine) and T (thymine) as one nucleotide type represented as *. Therefore, the alphabet of *k*-GCC is

$$\mathrm{N}_i \in \{\mathrm{G(guanine)}, \mathrm{C(cytosine)}, *\}, \quad (i = 1, 2, 3, \cdots, L) \tag{6}$$

Therefore, by replacing the *k*-tuple by k-GCC, a DNA sequence D can be represented as:

$$\mathbf{D} = \Big[f_1^{(1)} \cdots f_v^{(1)} \cdots f_{3^k}^{(1)} f_{3^k+1}^{(2)} \cdots f_{3^k+v}^{(2)} \cdots f_{2\times 3^k}^{(2)} f_{2\times 3^k+1}^{(3)} \cdots \\ f_{2\times 3^k+1}^{(3)} \cdots f_{2\times 3^k+v}^{(3)} \cdots f_{3\times 3^k}^{(3)} \Big]^{\mathbf{T}} \tag{7}$$

Compared with Equation 5, the *k*-GCC can efficiently reduce the dimension of the feature vector from 3 × 4^k to 3 × 3^k by focusing on the GC composition.

The proposed Pse-KGCC incorporates both the *k*-GCC and GC skew into the framework of PseKNC (Chen et al., 2014a), which can be represented as:

$$\mathbf{D} = \begin{matrix} [\phi_1 \cdots \phi_{3^k} \cdots \phi_{3^k+\lambda} \ \phi_{3^k+\lambda+1} \cdots \phi_{(3^k+\lambda)+3^k} \cdots \phi_{2\times(3^k+\lambda)} \ \phi_{2\times(3^k+\lambda)+1} \\ \cdots \phi_{2\times(3^k+\lambda)+3^k} \cdots \phi_{3\times(3^k+\lambda)}]^{\mathbf{T}} \end{matrix} \tag{8}$$

where

$$\phi_u = \begin{cases} \dfrac{f_u^{(1)}}{\sum_{i=1}^{3^k} f_i^{(1)} + w\sum_{j=1}^{\lambda} \theta_j^{(1)}} & 1 \le u \le 3^k \\ \dfrac{w\theta_{u-3^k}^{(1)}}{\sum_{i=1}^{3^k} f_i^{(1)} + w\sum_{j=1}^{\lambda} \theta_j^{(1)}} & 3^k + 1 \le u \le 3^k + \lambda \\ \dfrac{f_u^{(2)}}{\sum_{i=3^k+1}^{2\times 3^k} f_i^{(2)} + w\sum_{j=1}^{\lambda} \theta_j^{(2)}} & 3^k + \lambda + 1 \le u \le 2\times 3^k + \lambda \\ \dfrac{w\theta_{u-(3^k+\lambda)-3^k}^{(2)}}{\sum_{i=3^k+1}^{2\times 3^k} f_i^{(2)} + w\sum_{j=1}^{\lambda} \theta_j^{(2)}} & 2\times 3^k + \lambda + 1 \le u \le 2\times 3^k + 2\lambda \\ \dfrac{f_u^{(3)}}{\sum_{i=2\times 3^k+1}^{3\times 3^k} f_i^{(3)} + w\sum_{j=1}^{\lambda} \theta_j^{(3)}} & 2\times 3^k + 2\lambda + 1 \le u \le 3\times 3^k + 2\lambda \\ \dfrac{w\theta_{u-2\times(3^k+\lambda)-3^k}^{(3)}}{\sum_{i=2\times 3^k+1}^{3\times 3^k} f_i^{(3)} + w\sum_{j=1}^{\lambda} \theta_j^{(3)}} & 3\times 3^k + 2\lambda + 1 \le u \le 3\times 3^k + 3\lambda \end{cases} \tag{9}$$

where λ denotes the highest tier correlation of the *k*-GCC nucleotides in each local window of **D**, whose the value is an integer. *w* is a float number that represents the weight factor, and the value of *w* is between 0 and 1. In the front window, the middle window and the rear window, the correlation factor of the *j*-th tier is represented as $\theta_j^{(1)}$, $\theta_j^{(2)}$, and $\theta_j^{(3)}$, respectively. The GC skew value of the *k*-GCC nucleotides separated by *j* nucleotides is used to represent the correlation factor of the *j*-th tier in each local window. (**Figure 1**). $\theta_j^{(1)}$, $\theta_j^{(2)}$, and $\theta_j^{(3)}$ can be calculated by

$$128 \begin{cases} \theta_j^{(1)} = \dfrac{1}{\mathrm{Int}^{\mathrm{C}}[\frac{\eta-k}{j}]+1} \sum_{i=0}^{\mathrm{Int}^{\mathrm{C}}[\frac{\eta-k}{j}]} \Theta(\mathrm{N}_{i\times j+1}\mathrm{N}_{i\times j+2}\cdots\mathrm{N}_{i\times j+k}) & \\ \theta_j^{(2)} = \dfrac{1}{\mathrm{Int}^{\mathrm{C}}[\frac{\xi-\eta-k}{j}]+1} \sum_{i=0}^{\mathrm{Int}^{\mathrm{C}}[\frac{\xi-\eta-k}{j}]} \Theta(\mathrm{N}_{\eta+i\times j+1}\mathrm{N}_{\eta+i\times j+2}\cdots\mathrm{N}_{\eta+i\times j+k}) & \begin{matrix} j = 1, 2, \cdots, \lambda; \\ \lambda \le min(\eta, \xi-\eta, L-\xi) \end{matrix} \\ \theta_j^{(3)} = \dfrac{1}{\mathrm{Int}^{\mathrm{C}}[\frac{L-\xi-k}{j}]+1} \sum_{i=0}^{\mathrm{Int}^{\mathrm{C}}\left[\frac{L-\xi-k}{j}\right]} \Theta\left(\mathrm{N}_{\xi+i\times j+1}\mathrm{N}_{\xi+i\times j+2}\cdots\mathrm{N}_{\xi+i\times j+k}\right) & \end{cases} \tag{10}$$

where $\mathrm{Int}^{\mathrm{C}}[\dfrac{\eta-k}{j}]+1$ denotes the number of the *k*-GCC in the corresponding local window, and Θ($\mathrm{N}_{i \times j+1}\mathrm{N}_{i \times j+2} \cdots \mathrm{N}_i \times j + k$) is the GC Skew (Lobry, 1996; Grigoriev, 1998; Li et al., 2014) of the *i*-th *k*-GCC in the local window, which can be calculated by

$$\Theta\left(\mathrm{N}_{i\times j+1}\mathrm{N}_{i\times j+2}\cdots\mathrm{N}_{i\times j+k}\right) = \frac{f_G\left(\mathrm{N}_{i\times j+1}\mathrm{N}_{i\times j+2}\cdots\mathrm{N}_{i\times j+k}\right) - f_C\left(\mathrm{N}_{i\times j+1}\mathrm{N}_{i\times j+2}\cdots\mathrm{N}_{i\times j+k}\right)}{f_G\left(\mathrm{N}_{i\times j+1}\mathrm{N}_{i\times j+2}\cdots\mathrm{N}_{i\times j+k}\right) + f_C\left(\mathrm{N}_{i\times j+1}\mathrm{N}_{i\times j+2}\cdots\mathrm{N}_{i\times j+k}\right)} \tag{11}$$

where f_{G}($\mathrm{N}_{i \times j + 1}$ $\mathrm{N}_{i \times j + 2}$ ⋯ N_i × j + k) denotes the frequency of G in the subsequence $\mathrm{N}_{i \times j + 1}$ $\mathrm{N}_{i \times j + 2}$ ⋯ $\mathrm{N}_{i \times j + k}$$f_{\mathrm{C}}$($\mathrm{N}_{i \times j + 1}$ $\mathrm{N}_{i \times j + k}$ ⋯ $\mathrm{N}_{i \times j + k}$) denotes the frequency of C in the subsequence $\mathrm{N}_{i \times j + 1}$ $\mathrm{N}_{i \times j + 2}$ ⋯ $\mathrm{N}_{i \times j + k}$, reflecting the CG asymmetry bias directly. Please note that for the terminal subsequence, if its length is less than *k*, then the GC skew will be calculated by all the available nucleotide residues.

Random Forest

Being widely used in bioinformatics (Zhao et al., 2014; Su et al., 2019), Random Forest (RF) (Ho, 1995; Barandiaran, 1998) is a machine learning classifier. Its training process can prevent overfitting (Hastie et al., 2008). The Random Forest model was implemented by calling the command line RandomForestClassifier ("max_features='sqrt', min_samples_leaf=1, min_samples_split=2, criterion = 'gini', $\mathcal{F}$ = optimize-d value ") with the help of the Scikit-learn package (Pedregosa et al., 2011), where the values of $\mathcal{F}$ represents the number of the trees in the forest, and it was set as 600 for both the two benchmark datasets (cf. Equation 1).

Ensemble Learning

Previous studies (Zou et al., 2015; Liu et al., 2016a; Chen et al., 2016b; Chen et al., 2017a; Chen et al., 2017b; Liu et al., 2018a) have demonstrated that fusing a series of individual predictors

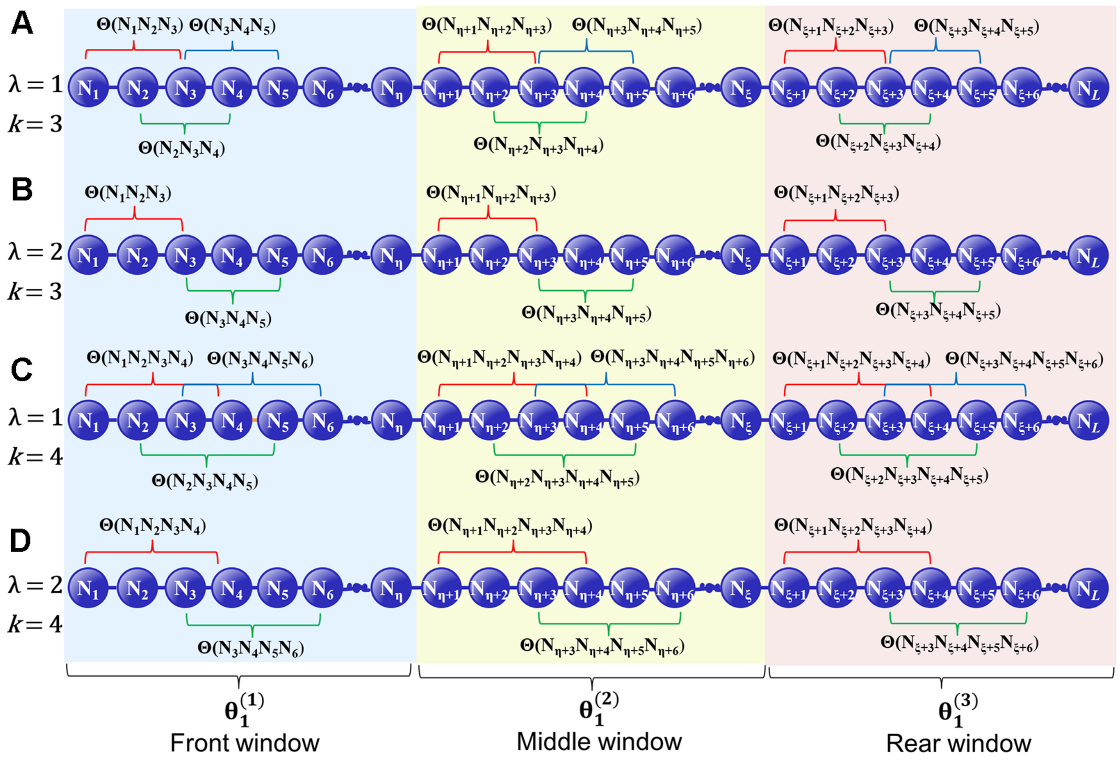

A
λ = 1
k = 3
$\Theta(N_1N_2N_3)$ $\Theta(N_3N_4N_5)$
$\Theta(N_2N_3N_4)$
$\Theta(N_{\eta+1}N_{\eta+2}N_{\eta+3})$ $\Theta(N_{\eta+3}N_{\eta+4}N_{\eta+5})$
$\Theta(N_{\eta+2}N_{\eta+3}N_{\eta+4})$
$\Theta(N_{\xi+1}N_{\xi+2}N_{\xi+3})$ $\Theta(N_{\xi+3}N_{\xi+4}N_{\xi+5})$
$\Theta(N_{\xi+2}N_{\xi+3}N_{\xi+4})$
B
λ = 2
k = 3
$\Theta(N_1N_2N_3)$
$\Theta(N_3N_4N_5)$
$\Theta(N_{\eta+1}N_{\eta+2}N_{\eta+3})$
$\Theta(N_{\eta+3}N_{\eta+4}N_{\eta+5})$
$\Theta(N_{\xi+1}N_{\xi+2}N_{\xi+3})$
$\Theta(N_{\xi+3}N_{\xi+4}N_{\xi+5})$
C
λ = 1
k = 4
$\Theta(N_1N_2N_3N_4)$ $\Theta(N_3N_4N_5N_6)$
$\Theta(N_2N_3N_4N_5)$
$\Theta(N_{\eta+1}N_{\eta+2}N_{\eta+3}N_{\eta+4})$ $\Theta(N_{\eta+3}N_{\eta+4}N_{\eta+5}N_{\eta+6})$
$\Theta(N_{\eta+2}N_{\eta+3}N_{\eta+4}N_{\eta+5})$
$\Theta(N_{\xi+1}N_{\xi+2}N_{\xi+3}N_{\xi+4})$ $\Theta(N_{\xi+3}N_{\xi+4}N_{\xi+5}N_{\xi+6})$
$\Theta(N_{\xi+2}N_{\xi+3}N_{\xi+4}N_{\xi+5})$
D
λ = 2
k = 4
$\Theta(N_1N_2N_3N_4)$
$\Theta(N_3N_4N_5N_6)$
$\Theta(N_{\eta+1}N_{\eta+2}N_{\eta+3}N_{\eta+4})$
$\Theta(N_{\eta+3}N_{\eta+4}N_{\eta+5}N_{\eta+6})$
$\Theta(N_{\xi+1}N_{\xi+2}N_{\xi+3}N_{\xi+4})$
$\Theta(N_{\xi+3}N_{\xi+4}N_{\xi+5}N_{\xi+6})$
N_1 N_2 N_3 N_4 N_5 N_6 N_η $N_{\eta+1}$ $N_{\eta+2}$ $N_{\eta+3}$ $N_{\eta+4}$ $N_{\eta+5}$ $N_{\eta+6}$ N_ξ $N_{\xi+1}$ $N_{\xi+2}$ $N_{\xi+3}$ $N_{\xi+4}$ $N_{\xi+5}$ $N_{\xi+6}$ N_L
$\theta_1^{(1)}$
Front window
$\theta_1^{(2)}$
Middle window
$\theta_1^{(3)}$
Rear window

FIGURE 1 | A schematic diagram to illustrate how to calculate the GC Skew in the front, middle, and rear windows along a DNA sequence. **(A)** The coupling between all the contiguous *k*-GCC (*k* = 3); **(B)** The coupling between the second most contiguous *k*-GCC (*k* = 3); **(C)** The coupling between all the contiguous *k*-GCC (*k* = 4); **(D)** The coupling between the second most contiguous *k*-GCC (*k* = 4).

by a voting strategy can improve the predictive performance. In this regard, in this study an ensemble predictors was constructed by fusing 10 top performing individual predictors constructed by different parameter combinations of PseKGCC (see **Supplementary Information S1**), which can be represented as (Liu et al., 2016a):

$$\mathbb{RF}^{\mathrm{E}} = \mathrm{RF}(1) \,\forall\, \mathrm{RF}(2) \,\forall\, \cdots \forall\, \mathrm{RF}(i) \,\forall \cdots \forall\, \mathrm{RF}(10) = \forall_{i=1}^{10}\mathrm{RF}(i) \tag{12}$$

where $\mathbb{RF}^{\mathrm{E}}$ represents the ensemble classifier, ∀ represents the fusing operator, and RF(i) represents the basic Random Forest predictor.

The ensemble predictor is constructed based on the fusion score ß of the probabilities predicted by the 10 basic predictors, which can be calculated by

$$ß = \sum_{i=1}^{10} q_i P_i \tag{13}$$

where q_i is the weight of the i-th basic RF predictor, which was optimized by the genetic algorithm (Mitchell, 1998), and their values were listed in **Supplementary Information S1**. If the value of ß is higher than 0.5, it is a replication origin, otherwise, it is a non-replication origin. The flowchart of the iRO-PseKGCC is illuminated in **Figure 2**.

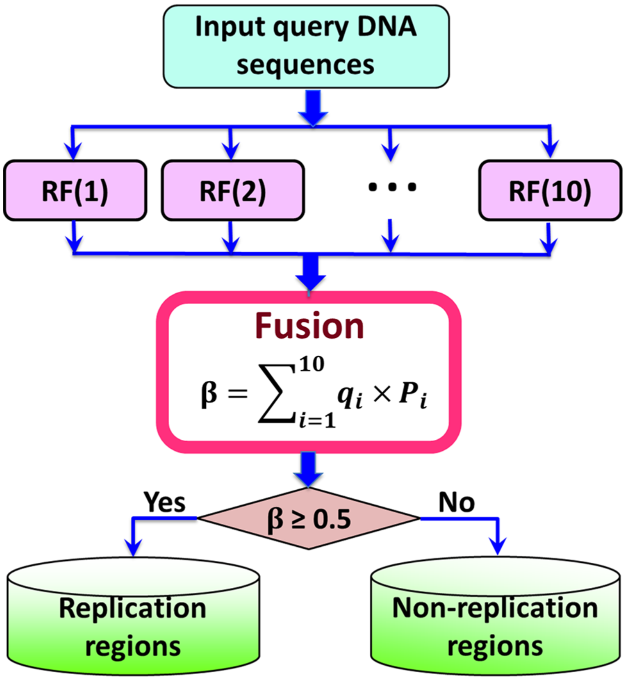


FIGURE 2 | A flowchart illustration to show how the iRO-PseKGCC predictor works.

Cross Validation

Three widely used cross-validation strategies include: i) independent test, ii) K-fold cross validation, and iii) jackknife test. Among these methods, only the jackknife test can achieve the unique results for the same benchmark dataset. Therefore, in this study, the jackknife test was employed to give the final predictive results. However, considering its high computational cost, during the parameter optimization process, the 5-fold cross-validation was used to reduce the computational cost (see *Optimize Parameters* section).

Evaluation Method of Performance

To evaluate the quality of the classifier for prediction of the replication origins, the four metrics are used (Feng et al., 2013; Chen et al., 2016c; Chen et al., 2019): i) the sensitivity, Sn, ii) the specificity, Sp, iii) the overall accuracy of the predictive results, Acc, iv) the Mathew's correlation coefficient, MCC, and v) Arear under ROC Curve, AUC (Chen et al., 2016a), defined as:

$$\begin{cases} \mathrm{Sn} = 1 - \frac{N_-^+}{N^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \frac{N_+^-}{N^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{ACC} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\frac{N_-^+ + N_+^-}{N^+ + N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- + N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ + N_+^-}{N^-}\right)}} & -1 \le \mathrm{MCC} \le 1 \\ AUC & \text{Arear under ROC Curve} \end{cases} \tag{14}$$

where N^+ denotes the number of all the positive samples (replication origins), N^- denotes the number of all the negative samples (non-replication origins), N_-^+ denotes the number of the positive samples (replication origins) incorrectly predicted as the negative samples (non-replication origins), N_+^- denotes the number of the negative samples (non-replication origins) incorrectly predicted as the positive samples (replication origins). More information of these performance measures can refer to Liu et al. (2016b).

RESULTS AND DISCUSSION

Optimize Parameters

There are five parameters in PseKGCC according to Equations 4–9. These parameters were optimized by the following equations:

$$\begin{cases} 0.15 \le \varepsilon \le 0.5, & \text{with step}\triangle\varepsilon = 0.05 \\ 0.5 < \delta \le 0.85, & \text{with step}\triangle\delta = 0.05 \\ 3 \le k \le 7, & \text{with step}\triangle k = 1 \\ 1 \le \lambda \le 10, & \text{with step}\triangle\lambda = 3 \\ 0.1 \le w \le 1, & \text{with step}\triangle w = 0.1 \end{cases} \tag{15}$$

The fivefold cross-validation was employed to search the optimal parameters by gridding method so as to reduce the time consumption, and the predictive results of the top 10 performing predictors, and their optimized parameters were listed in **Supplementary Information S1**.

Comparison With Other Methods

To the best knowledge of ours, iRO-3wPseKNC (Liu et al., 2018b) is the only existing predictor that is able to predict the entire replication origins. All the other predictors can only predict the fragments of replication origins. Therefore, the performance of the proposed iRO-PseKGCC was compared with iRO-3wPseKNC on the two benchmark datasets, and the results were listed in **Table 1**, from which we can see that iRO-PseKGCC obviously outperformed iRO-3wPseKNC in terms of the five performance measures (cf. Equation 14), indicating that the proposed PseKGCC feature is able to capture the GC asymmetry bias, and incorporate the GC skew into the predictor. Therefore, iRO-PseKGCC is an efficient approach for improving the predictive performance.

Feature Analysis

Random forest is a combination classifier model composed of decision tree classifiers. During the process of constructing each tree by the "Bootstrap" method (Efron, 1992), samples not extracted for training the corresponding tree can be used to make "Out Of Bag" (OOB) error estimate (Breiman, 1996) to evaluate the generalization performance of a predictor. Based on the OOB error, the Mean Decrease Accuracy (MDA) (Jiang et al., 2007) can

TABLE 1 | The results of the iRO-PseKGCC Predictor and comparison with iRO-PseKGCC on the two benchmark datasets (cf. Equation 1) obtained by using jackknife test.

Species	Method	Acc(%)	MCC	Sn(%)	Sp(%)	AUC
Saccharomyces cerevisiae $\mathbb{S}_1$	iRO-PseKGCC[a]	76.46	0.5298	73.90	78.13	0.8129
	iRO-3wPseKNC[b]	72.95	0.4594	70.67	75.22	0.8084
Pichia pastoris $\mathbb{S}_2$	iRO-PseKGCC[a]	74.22	0.4844	74.51	73.93	0.8002
	iRO-3wPseKNC[c]	71.10	0.4222	69.93	72.28	0.7962

[a] *The parameters are listed in* ***Supplementary Information S1****.*

[b] *The predictor reported in (Liu et al., 2018b) with parameter* ε *= 0.25,* δ *= 0.85, k = 5,* λ*= 6, w = 0.3, and* $\mathcal{F}$ *= 700.*

[c] *The predictor reported in (Liu et al., 2018b) with parameter* ε *= 0.15,* δ *= 0.55, k = 4,* λ *= 9, w = 0.3, and* $\mathcal{F}$ *= 800.*

be used to estimate the importance of the features. The details of the process are (Jiang et al., 2007): 1) When training a Random Forest model, using the OOB samples to test the accuracy of each tree in the model; 2) Randomly disturb the value of the feature variable *v* in the OOB samples, and retest the accuracy of each tree; 3) Calculate the mean value of the decreasing accuracy between the two tests in all decision trees in the Random Forest model. The MDA value can reflect the importance of the corresponding feature.

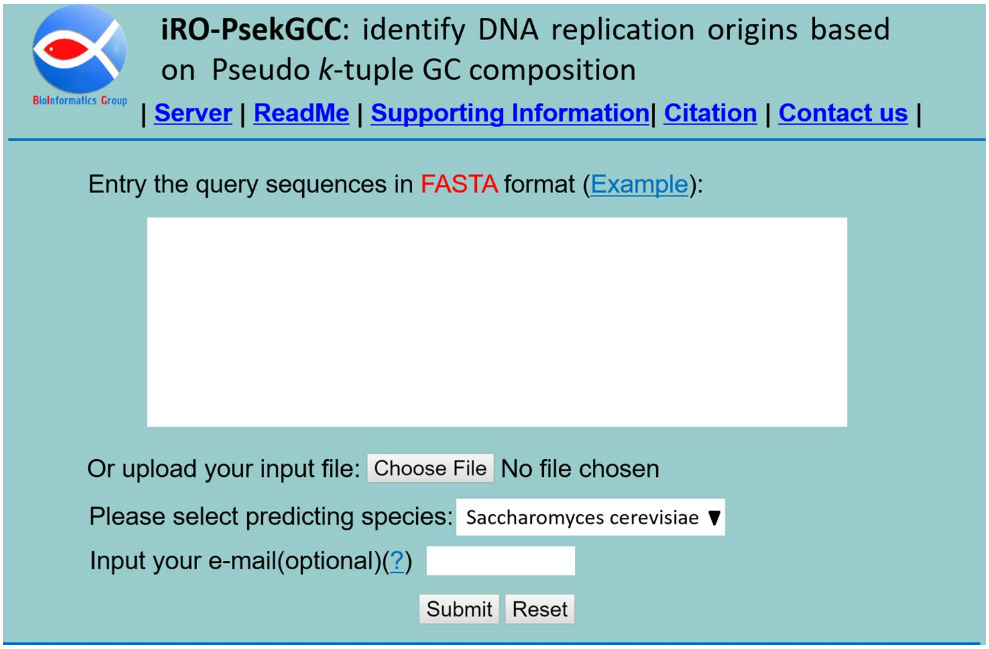
As shown in previous studies (Liu and Zhu, 2019; Liu et al. 2019a), feature analysis is critical for exploring the characteristics of the predictors. To explore the reason why the proposed predictor iRO-PseKGCC works so well, we analyzed the features of the two top performing iRO-PseKGCC predictors (see **Supplementary Information S1**) on the two benchmark datasets (cf. Equation 1) by MDA approach, and the results are listed in the **Table 2**, from which we can see that: 1) for both the two RF-based predictors, their most important features are the "***" and "*****," indicating the importance of the *k*-GCC; 2) The global sequence order effects measured by different λ values and GC skew values contribute to the performance improvement; 3) Features in certain local window show more discriminative powers than those in other windows, for examples, for *Pichia pastoris*, all the top 10 most important features are in the middle window, which is consistent with the previous observations that the nucleobase composition distribution is uneven along the replication origins (Lobry, 1996; Grigoriev, 1998; Frank and Lobry, 1999; Tillier and Collins, 2000; Liu et al., 2018b).

Web Server and User Guide

Web-servers are important for the researchers to implement the corresponding computational predictors. In this regard, for the user's convenience, we established a web-server named

TABLE 2 | The top 10 most important features of the top two performing RF-based predictors on the two benchmark datasets (cf. Equation 1).

Rank	Saccharomyces cerevisiae			Pichia pastoris		
	Feature	Window	MDA (%)	Feature	Window Index	MDA (%)
1	***	Rear window	20.49	*****	Middle window	15.89
2	***	Middle window	19.62	****G	Middle window	5.69
3	*GG	Rear window	9.04	G****	Middle window	5.38
4	GG*	Rear window	8.35	*C***	Middle window	5.23
5	*GG	Middle window	8.26	*G***	Middle window	5.14
6	λ = 1	Rear window	7.67	*CGCG	Middle window	3.99
7	GG*	Middle window	7.45	****C	Middle window	3.94
8	CC*	Middle window	7.31	***G*	Middle window	3.77
9	G*G	Rear window	6.64	*C*GG	Middle window	3.47
10	λ = 2	Rear window	6.12	C**G*	Middle window	3.40



iRO-PsekGCC: identify DNA replication origins based on Pseudo *k*-tuple GC composition

| Server | ReadMe | Supporting Information| Citation | Contact us |

Entry the query sequences in FASTA format (Example):

Or upload your input file: Choose File No file chosen

Please select predicting species: Saccharomyces cerevisiae ▼

Input your e-mail(optional)(?)

FIGURE 3 | A semi-screen shot to show the homepage of the web-server iRO-PseKGCC, which can be accessed at http://bliulab.net/iRO-PsekGCC/.

"iRO-PseKGCC." For users' convenience, a detailed user guide explaining how to use the web-server is given.

Step 1. Click on the web sites address http://bliulab.net/iRO-PsekGCC/ to open the web-server, then the main pages on the website as shown in **Figure 3** will appear in front of you. To see a brief introduction about the server, please click on the "Read Me" button.

Step 2. Choose the one specie from *Saccharomyces cerevisiae* or *Pichia pastoris*.

Step 3. The input sequences should be in the FASTA format. The sequence data can be uploaded *via* the "Browse" button or copy and paste or type into the input box directly.

Step 4. To see the predicted results, please click on the "Submit" button. For example, if the four query DNA sequences in the Example window are used as the queried data, the predictive results are the 1st and 2nd query sequences are replication origins, and the 3rd and 4th are non-replication origins.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. These data can be found here: https://academic.oup.com/bioinformatics/article-abstract/34/18/3086/4978052?redirectedFrom=fulltext

AUTHOR CONTRIBUTIONS

BL provided the main idea of the manuscript and wrote the manuscript. SC did the experiments and revised the manuscript. KY revised the manuscript and did the typesetting. FW did the experiments.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 61672184, 61732012, 61822306), Fok Ying-Tung Education Foundation for Young Teachers in the Higher Education Institutions of China (161063), Shenzhen Overseas High Level Talents Innovation Foundation (Grant No. KQJSCX20170327161949608), Guangdong Natural Science Funds for Distinguished Young Scholars (2016A030306008), and Scientific Research Foundation in Shenzhen (JCYJ20180306172207178).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00842/full#supplementary-material

REFERENCES

Barandiaran, I. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 20 (8), 832–844. doi: 10.1109/34.709601

Breiman, L. (1996). "Out-of-bag estimation". Citeseer).

Bu, H. D., Hao, J. Q., Guan, J. H., and Zhou, S. G. (2018). Predicting enhancers from multiple cell lines and tissues across different developmental stages based on svm method. *Curr. Bioinform.* 13 (6), 655–660. doi: 10.2174/1574893613666180726163429

Chen, H., Peng, S., Dai, L., Zou, Q., Yi, B., Yang, X., et al. (2017a). Oral microbial community assembly under the influence of periodontitis. *PloS One* 12 (8), e0182259. doi: 10.1371/journal.pone.0182259

Chen, J., Guo, M., Li, S., and Liu, B. (2017b). Protdec-ltr2. 0: an improved method for protein remote homology detection by combining pseudo protein and supervised learning to rank. *Bioinformatics* 33 (21), 3473–3476. doi: 10.1093/bioinformatics/btx429

Chen, J., Guo, M., Wang, X., and Liu, B. (2016a). A comprehensive review and comparison of different computational methods for protein remote homology detection. *Brief. Bioinform.* 19 (2), 231–244. doi: 10.1093/bib/bbw108

Chen, J., Long, R., Wang, X.-l., Liu, B., and Chou, K.-C. (2016b). dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation. *Sci. Rep.* 6, 32333. doi: 10.1038/srep32333

Chen, J., Wang, X., and Liu, B. (2016c). IMiRNA-SSF: improving the identification of MicroRNA precursors by combining negative sets with different distributions. *Sci. Rep.* 6, 19062. doi: 10.1038/srep19062

Chen, W., Feng, P., and Lin, H. (2012). Prediction of replication origins by calculating DNA structural properties. *Febs Letters* 586 (6), 934–938. doi: 10.1016/j.febslet.2012.02.034

Chen, W., Lei, T.-Y., Jin, D.-C., Lin, H., and Chou, K.-C. (2014a). PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.* 456, 53–60. doi: 10.1016/j.ab.2014.04.001

Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: Identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics.* doi: 10.1093/bioinformatics/btz015

Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K.-C. (2014b). PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics* 31 (1), 119–120. doi: 10.1093/bioinformatics/btu602

Efron, B. (1992). "Bootstrap methods: another look at the jackknife," in *Breakthroughs in statistics* (New York: Springer), 569–593. doi: 10.1007/978-1-4612-4380-9_41

Feng, P.-M., Chen, W., Lin, H., and Chou, K.-C. (2013). iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.* 442 (1), 118–125. doi: 10.1016/j.ab.2013.05.024

Frank, A., and Lobry, J. (1999). Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* 238 (1), 65–77. doi: 10.1016/S0378-1119(99)00297-8

Gao, F., and Zhang, C.-T. (2008). Ori-Finder: a web-based system for finding oriC s in unannotated bacterial genomes. *BMC Bioinform.* 9 (1), 79. doi: 10.1186/1471-2105-9-79

Grigoriev, A. (1998). Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.* 26 (10), 2286–2290. doi: 10.1093/nar/26.10.2286

Hastie, T., Tibshirani, R., and Friedman, J., (2008). *The elements of statistical learning (2nd ed.)*. New York: Springer series in statistics New York.

Ho, T. K. (1995). "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition* (Washington: IEEE), 278–282.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., and Lu, Z. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* 35 (suppl_2), W339–W344. doi: 10.1093/nar/gkm368

Li, W.-C., Deng, E.-Z., Ding, H., Chen, W., and Lin, H. (2015). iORI-PseKNC: a predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. *Chemom. Intell. Lab. Syst.* 141, 100–106. doi: 10.1016/j.chemolab.2014.12.011

Li, W.-C., Zhong, Z.-J., Zhu, P.-P., Deng, E.-Z., Ding, H., Chen, W., et al. (2014). Sequence analysis of origins of replication in the Saccharomyces cerevisiae genomes. *Front. Microbiol.* 5, 574. doi: 10.3389/fmicb.2014.00574

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22 (13), 1658–1659. doi: 10.1093/bioinformatics/btl158

Liu, B. (2017). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief. Bioinform*. doi: 10.1093/bib/bbx165

Liu, B., Gao, X., and Zhang, H. (2019b). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* doi: 10.1093/nar/gkz740

Liu, B., Li, C., and Yan, K. (2019a). DeepSVM-fold: Protein fold recognition by combining Support Vector Machines and pairwise sequence similarity scores generated by deep learning networks. *Brief. Bioinform.* doi: 10.1093/bib/bbz098

Liu, B., Li, K., Huang, D.-S., and Chou, K.-C. (2018a). iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. *Bioinformatics* 34 (22), 3835–3842. doi: 10.1093/bioinformatics/bty458

Liu, B., Long, R., and Chou, K.-C. (2016a). iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. *Bioinformatics* 32 (16), 2411–2418. doi: 10.1093/bioinformatics/btw186

Liu, B., Wang, S., Long, R., and Chou, K.-C. (2016b). iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics* 33 (1), 35–41. doi: 10.1093/bioinformatics/btw539

Liu, B., Weng, F., Huang, D.-S., and Chou, K.-C. (2018b). iRO-3wPseKNC: Identify DNA replication origins by three-window-based PseKNC. *Bioinformatics* 34 (18), 3086–3093. doi: 10.1093/bioinformatics/bty312

Liu, B., and Zhu, Y. (2019). ProtDec-LTR3.0: protein remote homology detection by incorporating profile-based features into Learning to Rank. *IEEE Access* 7, 102499–102507. doi: 10.1109/ACCESS.2019.292963

Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* 13 (5), 660–665. doi: 10.1093/oxfordjournals.molbev.a025626

Lubelsky, Y., MacAlpine, H. K., and MacAlpine, D. M. (2012). Genome-wide localization of replication factors. *Methods* 57 (2), 187–195. doi: 10.1016/j.ymeth.2012.03.022

Luo, H., Zhang, C.-T., and Gao, F. (2014). Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes. *Front. Microbiol.* 5, 482. doi: 10.3389/fmicb.2014.00482

Méchali, M. (2010). Eukaryotic DNA replication origins: many choices for appropriate answers. *Nat. Rev. Mol. Cell Biol.* 11 (10), 728. doi: 10.1038/nrm2976

Mitchell, M. (1998). *An introduction to genetic algorithms*. Boston: MIT press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12 (Oct), 2825–2830. doi: 10.1524/auto.2011.0951

Roten, C.-A. H., Gamba, P., Barblan, J.-L., and Karamata, D. (2002). Comparative Genometrics (CG): a database dedicated to biometric comparisons of whole genomes. *Nucleic Acids Res.* 30 (1), 142–144. doi: 10.1093/nar/30.1.142

Shirahige, K., Hori, Y., Shiraishi, K., Yamashita, M., Takahashi, K., Obuse, C., et al. (1998). Regulation of DNA-replication origins during cell-cycle progression. *Nature* 395 (6702), 618. doi: 10.1038/27007

Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. *Methods (San Diego, Calif.)* 166 (2019), 91–102. doi: 10.1016/j.ymeth.2019.02.009

Thomas, J. M., Horspool, D., Brown, G., Tcherepanov, V., and Upton, C. (2007). GraphDNA: a Java program for graphical display of DNA composition analyses. *BMC Bioinform.* 8 (1), 21. doi: 10.1186/1471-2105-8-21

Tillier, E. R., and Collins, R. A. (2000). The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.* 50 (3), 249–257. doi: 10.1007/s002399910029

Zhang, C.-J., Tang, H., Li, W.-C., Lin, H., Chen, W., and Chou, K.-C. (2016). iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget* 7 (43), 69783–69793. doi: 10.18632/oncotarget.11975

Zhang, C.-T., and Zhang, R. (1991). Analysis of distribution of bases in the coding sequences by a digrammatic technique. *Nucleic Acids Res.* 19 (22), 6313–6317. doi: 10.1093/nar/19.22.6313

Zhang, R., and Zhang, C.-T. (1994). Z curves, an intutive tool for visualizing and analyzing the DNA sequences. *J. Biomol. Struct. Dyn.* 11 (4), 767–782. doi: 10.1080/07391102.1994.10508031

Zhao, X., Zou, Q., Liu, B., and Liu, X. (2014). Exploratory predicting protein folding model with random forest and hybrid features. *Curr. Proteomics* 11 (4), 289–299. doi: 10.2174/157016461104150121115154

Zou, Q., Guo, J., Ju, Y., Wu, M., Zeng, X., and Hong, Z. (2015). Improving tRNAscan-SE annotation results *via* ensemble classifiers. *Mol. Inf.* 34 (11–12), 761–770. doi: 10.1002/minf.201500031

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.