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

Bin Liu1,2*, Shengyu Chen3†, Ke Yan4 and Fan Weng4

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

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) 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:

\[ D = N_1 N_2 N_3 \cdots N_i \cdots N_L \quad (i = 1, 2, 3, \ldots, L) \]  

where \( L \) denotes the length of D, and

\[ N_i \in \{A(adenine), C(cytosine), G(guanine), T(thymine)\}, \quad (i = 1, 2, 3, \ldots, L) \]

which represents the \( i \)-th nucleobase in the sequence, and \( f_i \) denotes the “member of” in the set theory. Following the study (Liu et al., 2018b), D is divided into three windows by two parameters \( \varepsilon \) and \( \delta \), including front window, middle window, and rear window respectively. \( \varepsilon \) and \( 1 - \delta \) 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 \( D[1, \varepsilon \cdot L] \), \( D[\varepsilon \cdot L + 1, L] \), and \( D[L + 1 \varepsilon \cdot L + \delta \cdot L] \), respectively, where \( \eta \) and \( \xi \) are defined as (Liu et al., 2018b),

\[
\left\{ \begin{array}{l}
\eta = \text{Int}^C[L \times \varepsilon] \\
\xi = \text{Int}^C[L \times \delta] \\
\end{array} \right. \quad (0 < \varepsilon < \delta < 1.0)
\]

where the symbol \( \text{Int}^C \) 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:

\[
D = \begin{bmatrix}
\frac{f_1^{(1)}}{f_1^{(2)}} \cdots \frac{f_1^{(2)}}{f_1^{(2)}} \cdots \frac{f_1^{(2)}}{f_1^{(2)}} \cdots \frac{f_1^{(2)}}{f_1^{(2)}} \\
\frac{f_2^{(2)}}{f_2^{(2)}} \cdots \frac{f_2^{(2)}}{f_2^{(2)}} \cdots \frac{f_2^{(2)}}{f_2^{(2)}} \\
\frac{f_3^{(2)}}{f_3^{(2)}} \cdots \frac{f_3^{(2)}}{f_3^{(2)}} \cdots \frac{f_3^{(2)}}{f_3^{(2)}} \\
\frac{f_4^{(2)}}{f_4^{(2)}} \cdots \frac{f_4^{(2)}}{f_4^{(2)}} \cdots \frac{f_4^{(2)}}{f_4^{(2)}} \\
\frac{f_5^{(2)}}{f_5^{(2)}} \cdots \frac{f_5^{(2)}}{f_5^{(2)}} \cdots \frac{f_5^{(2)}}{f_5^{(2)}} \\
\frac{f_6^{(2)}}{f_6^{(2)}} \cdots \frac{f_6^{(2)}}{f_6^{(2)}} \cdots \frac{f_6^{(2)}}{f_6^{(2)}} \\
\frac{f_7^{(2)}}{f_7^{(2)}} \cdots \frac{f_7^{(2)}}{f_7^{(2)}} \cdots \frac{f_7^{(2)}}{f_7^{(2)}} \\
\frac{f_8^{(2)}}{f_8^{(2)}} \cdots \frac{f_8^{(2)}}{f_8^{(2)}} \cdots \frac{f_8^{(2)}}{f_8^{(2)}} \\
\frac{f_9^{(2)}}{f_9^{(2)}} \cdots \frac{f_9^{(2)}}{f_9^{(2)}} \cdots \frac{f_9^{(2)}}{f_9^{(2)}} \\
\frac{f_{10}^{(2)}}{f_{10}^{(2)}} \cdots \frac{f_{10}^{(2)}}{f_{10}^{(2)}} \cdots \frac{f_{10}^{(2)}}{f_{10}^{(2)}} \\
\end{bmatrix}^T
\]

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 \( f_1^{(1)} \), \( f_2^{(2)} \), \( f_3^{(3)} \), respectively. The feature vector’s dimension is \( 3 \times 4^k \).

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

\[ N_i \in \{G\text{(guanine)}, C\text{(cytosine)}, \} \}, \quad (i = 1, 2, 3, \ldots, L) \]  

(6)

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

\[ D = \left[ f^{(1)}_1, f^{(1)}_2, \ldots, f^{(1)}_{3^k}, \ldots, f^{(2)}_1, f^{(2)}_2, \ldots, f^{(2)}_{3^k}, \ldots, f^{(3)}_1, f^{(3)}_2, \ldots, f^{(3)}_{3^k}, \ldots \right]^T \]  

(7)

Compared with Equation 5, the k-GCC can efficiently reduce the dimension of the feature vector from \( 3 \times 4^k \) to \( 3 \times \lambda \) 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:

\[ D = \left[ \phi^{(1)}_1, \phi^{(1)}_2, \ldots, \phi^{(1)}_{3^k}, \phi^{(2)}_1, \phi^{(2)}_2, \ldots, \phi^{(2)}_{3^k}, \phi^{(3)}_1, \phi^{(3)}_2, \ldots, \phi^{(3)}_{3^k}, \right]^T \]  

(8)

where

\[ \phi_a = \begin{bmatrix} \sum_{i=1}^{3^k} f^{(1)}_i w^{\lambda \mathbf{j}^{(1)}} & \sum_{i=1}^{3^k} f^{(2)}_i w^{\lambda \mathbf{j}^{(2)}} & \sum_{i=1}^{3^k} f^{(3)}_i w^{\lambda \mathbf{j}^{(3)}} \end{bmatrix} \]  

(9)

where \( \lambda \) denotes the highest tier correlation of the k-GCC nucleotides in each local window of \( D \), whose 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 \( \theta^{(1)}_j, \theta^{(2)}_j \), and \( \theta^{(3)}_j \), 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). \( \theta^{(1)}_j, \theta^{(2)}_j \), and \( \theta^{(3)}_j \) can be calculated by

\[ \theta^{(1)}_j = \frac{1}{\ln \left( \left\lceil \frac{N_i}{j+1} \right\rceil + 1 \right)} \sum_{k=1}^{N_i} \frac{f(C_{ik})}{\left\lceil \frac{N_i}{j+1} \right\rceil + 1} \]  

(10)

where \( \ln \left( \left\lceil \frac{N_i}{j+1} \right\rceil + 1 \right) \) denotes the number of the k-GCC in the corresponding local window, and \( \Theta(N_{i,j+1} \ldots N_{i,j+k}) \) 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

\[ \Theta(N_{i,j+1} \ldots N_{i,j+k}) = \frac{f(C_{i,j+1} \ldots C_{i,j+k})}{f(C_{i,j+1} \ldots C_{i,j+k}) + f(C_{i,j+1} \ldots C_{i,j+k})} \]  

(11)

where \( f(C_{i,j+1} \ldots C_{i,j+k}) \) denotes the frequency of C in the subsequence \( N_{i,j+1} \ldots N_{i,j+k} \) 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} = \text{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
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):

$$E_{RF} = \forall RF(1) \lor RF(2) \lor \ldots \lor RF(10) = \lor_{i=1}^{10} RF(i)$$

(12)

where $E_{RF}$ represents the ensemble classifier, $\forall$ represents the fusing operator, and $RF(i)$ represents the basic Random Forest predictor.

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

$$\beta = \sum_{i=1}^{10} q_iP_i$$

(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 $\beta$ 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.
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:

\[
\begin{align*}
\text{Sn} &= 1 - \frac{N^-}{N^+}, \\
\text{Sp} &= 1 - \frac{N^-}{N^-}, \\
\text{Acc} &= 1 - \frac{N^+ + N^-}{N^+ + N^-}, \\
\text{MCC} &= \frac{1 - \left(\frac{N^+ + N^-}{N^+ + N^-}\right)}{\sqrt{\left(1 + \frac{N^+ + N^-}{N^+ + N^-}\right)\left(1 + \frac{N^+ + N^-}{N^+ + N^-}\right)}}, \\
\text{AUC} &= \text{Arear under ROC Curve}
\end{align*}
\]

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:

\[
\begin{align*}
0.15 \leq \varepsilon \leq 0.5, \quad &\text{with step} \Delta \varepsilon = 0.05 \\
0.5 < \delta \leq 0.85, \quad &\text{with step} \Delta \delta = 0.05 \\
3 \leq k \leq 7, \quad &\text{with step} \Delta k = 1 \\
1 \leq \lambda \leq 10, \quad &\text{with step} \Delta \lambda = 3 \\
0.1 \leq w \leq 1, \quad &\text{with step} \Delta w = 0.1
\end{align*}
\]

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
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 \( \lambda \) 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 **iRO-PseKGCC**: identify DNA replication origins based on Pseudo \( k \)-tuple GC composition.

![Web Server interface](http://bliulab.net/iRO-PsekGCC/)

**Entry the query sequences in FASTA format (Example):**

- Choose File
- No file chosen

**Please select predicting species:** *Saccharomyces cerevisiae* ▼

**Input your e-mail (optional):** 

Submit | Reset

![FIGURE 3](http://bliulab.net/iRO-PsekGCC/)

**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/](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 site 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

**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 via high-order statistics, in *Bioinformatics and Biomedical Engineering (BIBM), 2018 IEEE International Conference on*. New York: Springer series in statistics New York.

Chen, J., Guo, M., Li, S., and Liu, B. (2017b). Protdec-ltr2.0: an improved method bioinformatics/btx429. 10.1093/bib/btx429

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., Zhang, X., Booker, 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* (Springer: New York), 569–593. doi: 10.1007/978-1-4612-4380-9_41

Feng, P.-M., Chen, W., Lin, H., and Chou, K.-C. (2013). iHSP-PseRAAC: 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, F., 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/gkm2368

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., and 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

**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. KJQSCX20170327161949608), Guangdong Natural Science Funds for Distinguished Young Scholars (2016A030360008), 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
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/bbx085

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/gkz1740

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., Wang, 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 archael 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 Genomics (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., Shirai, 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 intuitive 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.

Copyright © 2019 Liu, Chen, Yan and Weng. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.