Bioinformatics, 35(12), 2019, 2075–2083
doi: 10.1093/bioinformatics/bty943
Advance Access Publication Date: 14 November 2018
Original Paper

Sequence analysis

Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique

Fu-Ying Dao[1], Hao Lv[1], Fang Wang[1], Chao-Qin Feng[1], Hui Ding[1,*], Wei Chen[1,2,*] and Hao Lin[1,*]

[1]Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China and [2]Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 30, 2018; revised on November 6, 2018; editorial decision on November 9, 2018; accepted on November 13, 2018

Abstract

Motivation: DNA replication is a key step to maintain the continuity of genetic information between parental generation and offspring. The initiation site of DNA replication, also called origin of replication (ORI), plays an extremely important role in the basic biochemical process. Thus, rapidly and effectively identifying the location of ORI in genome will provide key clues for genome analysis. Although biochemical experiments could provide detailed information for ORI, it requires high experimental cost and long experimental period. As good complements to experimental techniques, computational methods could overcome these disadvantages.

Results: Thus, in this study, we developed a predictor called iORI-PseKNC2.0 to identify ORIs in the *Saccharomyces cerevisiae* genome based on sequence information. The PseKNC including 90 physicochemical properties was proposed to formulate ORI and non-ORI samples. In order to improve the accuracy, a two-step feature selection was proposed to exclude redundant and noise information. As a result, the overall success rate of 88.53% was achieved in the 5-fold cross-validation test by using support vector machine.

Availability and implementation: Based on the proposed model, a user-friendly webserver was established and can be freely accessed at http://lin-group.cn/server/iORI-PseKNC2.0. The webserver will provide more convenience to most of wet-experimental scholars.

Contact: hding@uestc.edu.cn or chenweiimu@gmail.com or hlin@uestc.edu.cn

1 Introduction

Genome duplication is the most fundamental and orchestrated step. Although the replication mechanisms are differences among bacteria, archaea and eukaryotes, they have the same core components (Song *et al.*, 2015; Zakrzewska-Czerwińska *et al.*, 2007). One of the most important components is origin of replication (ORI), which is responsible for the initiation of DNA replication. Accurate identification of ORIs can be very helpful for in-depth understanding genome duplication. In addition, scholars are also able to learn about the statistical properties and organizational features of the DNA sequence of the organism, and the development of new drugs for treatment of various diseases caused by errors in genome replication (McFadden and Roos, 1999; Raghu Ram *et al.*, 2007; Soldati, 1999).

In general, most bacterial genomes have a single circular DNA molecule, and typically only one ORI region of replication per circular chromosome (Marczynski and Shapiro, 1993). However, fungal

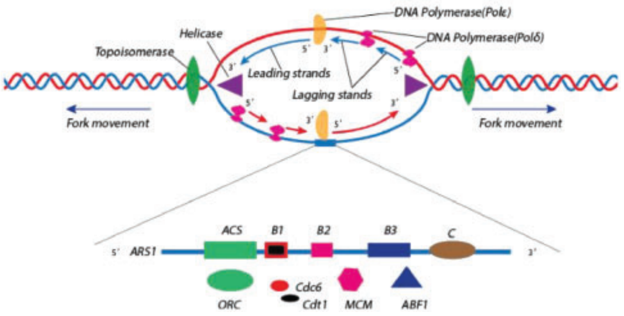

Fig. 1. The unicellular *S.cerevisiae* has a linear DNA molecule with multiple ORIs. And the ARS elements are formed by domain A (ACS), domain B (B1, B2, B3) and domain C

genomes usually own multiple ORI sites on each linear chromosome (Mechali, 2010), so that they can complete replication in a reasonable period of time (Schub *et al.*, 2001). The replication origin regions in unicellular fungus budding yeast *Saccharomyces cerevisiae* (*S.cerevisiae*) have autonomously replicating sequences (ARS), which are formed by three domains namely domain A, B and C (Foureau *et al.*, 2013) (Fig. 1). The three important domains have their own special functions. The A domain contains an essential ARS consensus sequence (ACS) [5-(T/A)TTTAT(A/G)TTT(T/A)-3], which is essential for binding of the origin recognition complex (Lee and Bell, 1997; Rao and Stillman, 1995; Rowley *et al.*, 1995). The B domain includes numerous short sequence motifs that contribute to origin activity (Dhar *et al.*, 2012), additionally, it tends to be helically unstable. The motifs of the C domain are mainly in charge of the interaction between DNA and regulatory protein (Dhar *et al.*, 2012). However, since they are not conservative, these motif sequences are insufficient to identify ORIs (Nieduszynski *et al.*, 2006). Although experimental methods such as chromatin immunoprecipitation could identify the ORIs accurately (Lubelsky *et al.*, 2012), it is still cost-ineffective.

Recent years, with the development of bioinformatics and the accumulation of biological experimental data, it is possible to predict ORIs of microorganisms by computational approaches. For most of bacterial genomes, their ORIs can be correctly identified by computational methods (Gao and Zhang, 2008; Luo *et al.*, 2014; Shah and Krishnamachari, 2012). However, fungal ORIs are difficult to be determined due to multiple ORIs. In the past several years, a series of methods have been proposed to try to solve this challenge. For example, Chen *et al.* (2012) developed a computational model for identifying ORIs in *S.cerevisiae* and found that both DNA bendability and cleavage intensity in core replication regions were significantly lower than those in the linker regions. Li *et al.* (2014a) calculated the GC profile and GC skew to analyze the compositional bias in the *S.cerevisiae* genome. Later on, they developed a predictor called iORI-pseudo *k*-tuple nucleotide composition (PseKNC) for identifying ORIs in *S.cerevisiae* genome (Li *et al.*, 2015b). However, the predictive accuracy is still far from satisfactory. By incorporating the dinucleotide position-specific propensity information into the general pseudo nucleotide composition, a new predictor called iROS-gPseKNC was proposed for identifying ORIs in the *S.cerevisiae* genome (Xiao *et al.*, 2016). However, the efficiency of the corresponding webserver is very slow. Moreover, the position information is not suitable for ORI prediction because the length of experimentally confirmed ORIs is different. Recently, Singh *et al.* (2018) used two feature extraction methods combined with multi-view ensemble learning to predict ORIs in *S.cerevisiae* genome based on three different classification algorithms, and they found the properties of ACS flanking regions are highly distinguishable from ACS sequences (core ARS region).

Although the previous works have indeed achieved encouraging results, unfortunately, all these reported methods have some limitations in terms of accuracy and efficiency. More importantly, most of these methods did not use feature selection technique to select optimal features. And few physicochemical properties were considered in their models. In addition, few webservers were established.

In view of this, we devoted to enhance the prediction capability in recognizing yeast ORIs from the aforementioned disadvantages. At first, 90 physicochemical properties were incorporated into PseKNC to characterize the DNA sequences of ORI and non-ORI samples. Meanwhile, *F*-score (Lin *et al.*, 2014) and mRMR (Mundra and Rajapakse, 2010) were utilized to optimize features. The support vector machine (SVM) was proposed to perform classification. Based on the proposed model, we established a webserver named iORI-PseKNC2.0 which will provide great assistance to fungal ORIs research.

2 Materials and methods

2.1 Benchmark dataset

Establish a reliable and stringent benchmark dataset is the first important step for building a reliable predictor. In this study, the original data was collected from OriDB (http://www.oridb.org/) (Nieduszynski *et al.*, 2007), which contained experimentally confirmed ORIs in *S.cerevisiae*. A total of 740 ORIs in *S.cerevisiae* was obtained from OriDB (Nieduszynski *et al.*, 2007). For guaranteeing to constructed dependable dataset, those ORIs with lacking confidence which are ambiguous annotation such as 'likely' and 'dubious' were removed, and only keep the 'confirmed' 410 ORIs verified by experiment. And then, the CD-HIT software was used to get rid of redundant samples by setting the threshold as 0.75. Finally, 405 ORIs with 300 bp long were extracted as positive samples, at the same time, 406 non-ORIs were extracted as negative samples. The dataset can be freely downloaded from the website (http://lin-group.cn/server/iOri-PseKNC2.0/download.html). Therefore, the benchmark dataset *S* can be formulated as:

$$S = S^{+} \cup S^{-} \tag{1}$$

where S^{+} represents the ORI dataset (or called positive subset), S^{-} represents the non-ORI dataset (or called negative subset). The symbol U denotes the union in the set theory.

In the dataset *S*, each sample sequence has 300 bp and can be represented by the following formula:

$$P = P_1 P_2 \cdots P_{300} \tag{2}$$

where *P* represents the DNA sample, and $P_1,\ P_2, \cdots$ denote the first nucleotide, the second nucleotide and so on.

2.2 Feature vector construction

The PseKNC is commonly used to characterize nucleotide sequences in DNA/RNA sequence classification (Chen *et al.*, 2014c, 2015; Lin *et al.*, 2017; Su *et al.*, 2018; Tang *et al.*, 2018a; Yang *et al.*, 2018a, b). The sequence information of DNA can be converted into vectors from two aspects, one is the frequency of oligonucleotide components, and the other is the correlation of physicochemical properties between two dinucleotides. This strategy incorporates both the short-range and long-range information of DNA sequence. Therefore, most information in DNA sequences can be extracted. Thus, it has been adopted to formula DNA/RNA samples for DNA elements prediction, such as the promoter identification (Lin *et al.*, 2014), nucleosome position prediction (Guo *et al.*, 2014) and so on. These studies all utilized Type-I PseKNC to formulate their samples. In this work, we used the other kind of PseKNC namely Type-II PseKNC to represent DNA samples. Accordingly, each ORI (or non-ORI) DNA sequence sample can be denoted as a $4^k + n\lambda$ dimensional vector which is formulated as:

$$P = [x_1, x_2, \cdots, x_{4^k}, x_{4^k+1}, \cdots, x_{4^k+n\lambda-1}, x_{4^k+n\lambda}]^T \tag{3}$$

in which

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{4^k} f_i + \omega \sum_{j=1}^{n\lambda} \tau_j}, & (1 \le u \le 4^k) \\ \dfrac{\omega \tau_{u-4^k}}{\sum_{i=1}^{4^k} f_i + \omega \sum_{j=1}^{n\lambda} \tau_j}, & (4^k + 1 \le u \le 4^k + n\lambda) \end{cases} \tag{4}$$

where the value n represents the number of the physicochemical properties and the parameter λ represents the counted rank of the correlation along a DNA sequence. The weight factor ω is used to adjust the ratio between nucleotide composition and correlation effect. f_u $(u = 1,\ 2, \cdots, 4^k)$ is the normalized occurrence frequency of the u-th k-tuple nucleotide in the DNA segment, and τ_j is the j-th tire correlation factor that reflects the sequence-order correlation between all the j-th most contiguous dinucleotides along a DNA sequence. The f_u and τ_j can be formulated by:

$$f_u = \frac{n_u}{\sum_{i=1}^{L-k+1} n_i} \tag{5}$$

$$\tau_j = \frac{1}{\mathrm{L} - j - 1} \sum_{i}^{L-j-1} \Theta(R_i R_{i+1}; R_{i+j} R_{i+1+j}) \ (j = 1,\ 2, \cdots, \lambda\ <\ 300) \tag{6}$$

In the above two equations, the value L represents the length of sample. n_u denotes the occurrences number of u-th k-tuple nucleotide in DNA sample P. $\Theta(R_i R_{i+1}; R_{i+j} R_{i+1+j})$ in Equation (6) is the correlation function of the physicochemical properties between two dinucleotides and can be defined by:

$$\Theta(R_i R_{i+1}; R_{i+j} R_{i+1+j}) = \frac{1}{\mu} \sum_{v=1}^{\mu} [P_v(R_i R_{i+1}) - P_v(R_{i+j} R_{i+j+1})]^2 \tag{7}$$

where μ is the number of local DNA structural properties considered in the current study as described in the following section; $P_v(R_i R_{i+1})$ is the numerical value of the v-th ($v = 1, 2, \cdots, \mu$) DNA local physicochemical property for the dinucleotide $R_i R_{i+1}$ at position i and $P_v(R_{i+j} R_{i+j+1})$ is the corresponding value for the dinucleotide $R_{i+j} R_{i+j+1}$ at position $i + j$.

2.3 DNA local structural property parameters

Many reports have showed that DNA local structural properties play crucial roles in a series of biological processes (Goni *et al.*, 2007, 2008; Miele *et al.*, 2008). Li *et al.* (2015a, b) predicted ORIs in *S. cerevisiae* genome by using six basic physicochemical properties, including Slide, Shift, Rise, Roll, Tilt and Twist. It was shown that the physicochemical properties did play an important role in ORIs recognition. We also examined the above six properties in Type-II PseKNC method with feature selection methods, but the accuracy was still less than satisfactory. Thus, we guessed that the six kinds of physicochemical properties cannot afford enough information for describing the dinucleotides' correlation. Accordingly, a total of 90 kinds of physicochemical properties (Chen *et al.*, 2014c) were proposed for gaining more information.

It is worth noting that all the original values must be subjected to a standard conversion before we substitute the values of the 90 parameters of dinucleotides into Equation (7). For the consistency of parameters, a standard conversion should be made as follows.

$$P_v(R_i R_{i+1}) = \frac{P_v(R_i R_{i+1}) - < P_v >}{SD(P_v)} \tag{8}$$

where the symbol <> means the average value of dinucleotides, and *SD* denotes the corresponding standard deviation. The standardized 90 kinds of physicochemical properties can be found in http://lin-group.cn/server/iOri-PseKNC2.0/download.html.

2.4 SVM

SVM was originally developed by VapnikVladimir (1997), which is a machine-learning method based on the statistical learning theory. It has been successfully used in the realm of bioinformatics (Chen *et al.*, 2016, 2017; Lai *et al.*, 2017; Manavalan and Lee, 2017; Manavalan *et al.*, 2018; Song *et al.*, 2018b; Stephenson *et al.*, 2018; Yang *et al.*, 2016). The basic idea of SVM is to transform the input vector into a high-dimensional Hilbert space and seek a separating hyperplane in this space. It targets on minimizing the structural risk and uses kernel function to tackle non-linearly separable problem (Keerthi and Lin, 2003).

In this study, we downloaded the LIBSVM package from http://www.csie.ntu.edu.tw/~cjlin/libsvm/ to implement SVM (Chang and Lin, 2011). The software provides executable files and the corresponding source code. The user can modify and extend the software according to their own needs. The radial basis kernel function was used in this study due to its effectiveness and speed in non-linear classification process. The best values for the regularization parameter C and kernel parameter γ could be obtained by applying a grid search strategy with n-fold cross-validation test. The search spaces for C and γ were defined by:

$$\begin{cases} 2^{-5} \le C \le 2^{15} & \textit{with step of } 2 \\ 2^{-15} \le \gamma \le 2^{-5} & \textit{with step of } 2^{-1} \end{cases} \tag{9}$$

2.5 Feature selection technique

Feature selection can not only gain relevant modeling variables, but also improve the understandability, scalability and accuracy of the proposed models (Yuan *et al.*, 2013). In this study, the number of physicochemical properties is 90. Thus, the dimension of feature vector will dramatically increase with the increase of k and λ. To avoid the high-dimensional disaster and improve prediction accuracy (Ding *et al.*, 2012), we performed the two-step feature selection to optimize the features. In fact, the two-step feature selection strategy has been applied in related bioinformatics and sequence analysis studies (Feng *et al.*, 2019; Jia and He, 2016; Li *et al.*, 2014b, 2015a, 2016; Song *et al.*, 2018b; Wang *et al.*, 2012; Wei *et al.*, 2018). By doing so, we not only reduced the over-fitting risk and information redundancy, but also improved the computational efficiency.

2.5.1 *F*-score technique

F-score is a simple and valid feature selection method, and is usually used to measure the degree of difference between two real number sets (Lin *et al.*, 2014). For a given training sample x_d ($d = 1, 2, \cdots, m$), there are n^+ positive samples and n^- negative samples, then the *F*-score of the *i*-th feature is described as:

$$F_i = \frac{\left(\overline{x}_i^{(+)} - \overline{x}_i\right)^2 + \left(\overline{x}_i^{(-)} - \overline{x}_i\right)^2}{\frac{1}{n^+-1}\sum_{k=1}^{n^+}\left(\overline{x}_{d,i}^{(+)} - \overline{x}_i^{(+)}\right)^2 + \frac{1}{n^--1}\sum_{d=1}^{n^-}\left(\overline{x}_{d,i}^{(-)} - \overline{x}_i^{(-)}\right)^2} \tag{10}$$

where $\overline{x}_i^+$, $\overline{x}_i^-$, and $\overline{x}_i$ are the average frequency of the *i*-th feature in the positive samples, the negative samples and the whole samples respectively; $\overline{x}_{d,i}^{(+)}$ and $\overline{x}_{d,i}^{(-)}$ represent the value of the *i*-th feature of the *d*-th sequence in the positive samples and the negative samples, respectively. In Equation (10), numerator describes the ability to distinguish between positive and negative samples, denominator represents the ability to distinguish within two samples. The larger the *F* value means the better the predictive capability the feature has. *F*-score could achieve an effective feature selection with strict mathematical definition.

2.5.2 mRMR technique

The core idea of mRMR (Mundra and Rajapakse, 2010; Peng *et al.*, 2005) is to maximize the correlation between features and categorical labels while minimize the correlation between features and features. If there are two random variables *x* and *y*, and their probability density function were defined as *p*(*x*), *p*(*y*) and *p*(*x*, *y*). The mutual information between them can be defined as.

$$I(x;y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \tag{11}$$

According to mutual information, finding a feature subset *S* with *m* optimal features $\{x_i\}$ is the purpose of feature screening that has the largest dependency on the target class *c*.

The maximum relevance has the following form:

$$\max D(S,c),\ \ D = \frac{1}{|s|}\sum_{x_i\in S} I(x_i;c)\ \ (i = 1,\cdots,m) \tag{12}$$

The minimum redundancy is defined as:

$$\min R(S,c),\ \ R = \frac{1}{|s|^2}\sum_{x_i,x_j\in S} I(x_i;x_j). \tag{13}$$

The final selection criteria are formulated as:

$$\max\varnothing(D,R), \varnothing = D - R. \tag{14}$$

The final feature subset can keep the correlation between features and the categories maximum, and at the same time, ensure the redundancy among the features minimum. As a result, the 'purest' feature subset is obtained. Thus, the mRMR algorithm was performed easily and efficiently as well as could achieve robust model.

2.6 Evaluation metrics

In order to objectively evaluate the performance of proposed models, the following indexes: sensitivity (*Sn*), specificity (*Sp*), overall accuracy (*Acc*) and Matthews correlation coefficient (*MCC*) (Li *et al.*, 2018; Lin *et al.*, 2014; Song *et al.*, 2018a; Tang *et al.*, 2017; Zhu *et al.*, 2018) were used in the current work and were defined as:

$$\begin{cases} Sn = 1 - \dfrac{N_-^+}{N^+}, \ \ 0 \le Sn \le 1 \\[2ex] Sp = 1 - \dfrac{N_+^-}{N^-}, \ \ 0 \le Sp \le 1 \\[2ex] Acc = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-}, \ \ 0 \le Acc \le 1 \\[2ex] MCC = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}}, \ \ 0 \le MCC \le 1 \end{cases} \tag{15}$$

where N^+ denotes the total number of ORIs sequences in investigated benchmark dataset, N_-^+ denotes the number of ORIs sequences misclassified as non-ORIs sequences; similarly, N^- denotes the total number of non-ORIs sequences on investigated benchmark dataset, N_+^- denotes the number of non-ORIs misclassified as ORIs.

In addition to the above indicators, the receiver operating characteristic curve (ROC) (Metz, 1989), a visual indicator with the abscissa of 1-*Sp* and the ordinate of *Sn*, was used to measure the performance of the predictor across the entire range of SVM decision values. The quality of the predictors can be objectively evaluated by measuring the area under the ROC (auROC).

3 Results and discussion

3.1 Cross-validation

Cross-validation, a kind of statistical analysis method, is always used to objectively estimate the effectiveness and performance of a predictor in practical application. Three kinds of cross-validation methods including independent dataset test, *n*-fold cross-validation test and jackknife cross-validation test (Chou and Zhang, 1995) are commonly used in the performance evaluation of machine learning. Although jackknife cross-validation can achieve unique outcomes, in order to saving computational time, we used the 5-fold cross-validation test to confirm the optimal feature subset. And the jackknife cross-validation was used to examine the performance of the final model and further to compare with published models.

3.2 Parameter optimization

According to the section of feature vector construction, three parameters *k*, *λ* and *ω* must be determined in advance to obtain an optimal classification model. The parameter *k* is the tier of the oligonucleotides, which describes the local pattern information (short-range information) of a DNA sequence; *λ* is the correlation tier of physicochemical properties between two dinucleotides, which describes the global pattern sequence-order effect (long-range information) along a DNA sequence; *ω* is the weight factor to adjust the ratio between short-range effect and long-range effect. It is obvious that the greater the *k* is, the more local sequence-order information the model contains. Similarly, the greater the *λ* is, the more global sequence-order information it contains. Generally, the search space of the three parameters was described as

$$\begin{cases} 1 \le k \le k_0 \ \ with\ step\ \ \Delta\ = 1 \\ 1 \le \lambda \le \lambda_0 \ \ with\ step\ \ \Delta = 1 \\ 0.1 \le \omega \le \omega_0 \ \ with\ step\ \ \Delta = 0.1 \end{cases}. \tag{16}$$

Based on Equation (16), the performances of the $k_0 \times \lambda_0 \times \omega_0$ combinations should be investigated by grid search. The dimension will raise dramatically with the increases of parameters k_0, λ_0 and ω_0, which will reduce the calculation efficiency. Moreover, for each combination, feature selection should be made to obtain the best accuracy. Thus, it is extremely time-consuming to investigate performances of these feature subsets. For saving the computing time, the parameter *k* was changed from 2 to 6 with the step of 1 for finding the best short-range correlation factor, and the value of *λ* was adjusted from 1 to 50 with the step of 1 for finding the optimal long-range correlation factor. As the weight factor *ω* does not influence the outcome dramatically in our pre-experiments, the value of *ω* was set to 1. Thus, the searching space was formulated as

$$\begin{cases} 2 \le k \le 6 \ \ with\ step\ \ \Delta\ = 1 \\ 1 \le \lambda \le 50 \ \ with\ step\ \ \Delta = 1 \\ \omega = 1 \end{cases}. \tag{17}$$

Accordingly, there are $5 \times 50 \times 1 = 250$ combinations. For each combination, feature selection should be made to obtain the best accuracy. However, it is still extremely slow to investigate performances of these feature subsets. Thus, for further improving calculating efficiency, we initially fixed the parameter λ to 50 in order to include the global information as more as possible, and the parameter k was changed from 2 to 6 with the step of 1 for determining the best parameter k. Results were listed in following

$$\lambda = 50 \begin{cases} Acc = 85.08\%, & k = 2 \\ Acc = 85.11\%, & k = 3 \\ Acc = 85.20\%, & k = 4 \\ Acc = 85.25\%, & k = 5 \\ Acc = 85.70\%, & k = 6 \end{cases}. \quad (18)$$

One may notice from Equation (18) that the accuracies reached to maximum (85.70%) when the optimal value of k was selected as 6. Subsequently, we will determine λ. For each λ, there are total of $(4^6 + 90 \times \lambda)$ features which is still a high dimension vector. Thus, it is necessary to pick out the useful information from original features for building a high-quality model. It is obvious that the best feature subset could be obtained by investigating the performance of all combinations of features. However, the combination of features is so large that any computer cannot burden such calculation. Let's use a 400-dimension feature vector as an example. The number of all possible combinations will be $>2.58 \times 10^{120}$.

Therefore, we must use a valid technique to reduce the computational complex for saving computing time and resource. *F*-score and mRMR are two wonderful feature ranking techniques based on statistical and informational theories. However, the running speed of *F*-score is faster than that of mRMR. Thus, we firstly utilized the *F*-score combined with incremental feature selection (IFS) (Li *et al.*, 2015a, 2016) to obtain the optimal feature subset which could produce the best prediction performance of model. At first, *F*-score was used to calculate the score of each feature. Subsequently, the features were ranked from large to small according to their scores. And $4^6 + 90 \times \lambda$ feature subsets were constructed on the basis of the IFS rule that the i-th feature subset contains the first i feature in the ranking feature set. Third, the SVM was employed to perform classification by using 5-fold cross-validation. Finally, we could find the best feature subset which could produce the maximum accuracy. Here, we varied λ from 1 to 50. Figure 2 showed the change of the 50 highest accuracies with correlation factor λ. We found that, when $\lambda = 1$, maximum accuracy of 88.41% can be obtained by 1246 features. Figure 3 showed the IFS process (blue curve). The detailed predictive performance of these 1246 features was shown in Table 1.

In the feature selection process of *F*-score with IFS, the correlation between any two features was not considered. The mRMR could keep the correlation between features and the categories maximum, and at the same time, ensure the redundancy among the features minimum. Thus, the mRMR combined with IFS was utilized to further optimize the 1246 features obtained by *F*-score. The IFS curves for mRMR were drawn in Figure 3 (red curve). One may observe that the maximum accuracy was increased to 88.53% in 5-fold cross-validation. Although this improvement is not dramatically, the feature dimension was decreased from 1246 to 1048. Thus, we could conclude that the two-step feature selection can do improve the prediction accuracy and reduce the feature vector dimension. The details of prediction results were listed in Table 1. For further measuring and comparing the performances of the predictor in different feature optimization statuses, the ROC curves were drawn in Figure 4. Obviously, the final prediction model should be constructed based on the 1048 best features.

3.3 Compared with published methods

It is necessary to compare our proposed method with other published methods. Our first work investigated the distribution of DNA bendability and cleavage intensity around ORIs and found that the properties of ORIs regions are lower than those of non-ORI regions (Chen *et al.*, 2012). Based on the observation, a model was constructed by using SVM classifier. The auROC of 0.8563 was obtained in jackknife cross-validation. Later on, our group (Li *et al.*, 2015b) used the Type-I PseKNC to formulate samples for discriminating ORIs from non-ORIs by considering the concepts of the local and global sequence-order effects of DNA sequence. Better results were obtained. On the basis of the model, we built a webserver named iORI-PseKNC (http://lin-group.cn/server/iOri-PseKNC).

However, it is unfair to compare the above results with published results because the protocol of cross-validation is different. Based on the same dataset, we must compare the proposed method in this paper with the two previous models using the same assessment criteria and jackknife cross-validation. Thus, we further examined the performance of the 1048 optimal features using jackknife

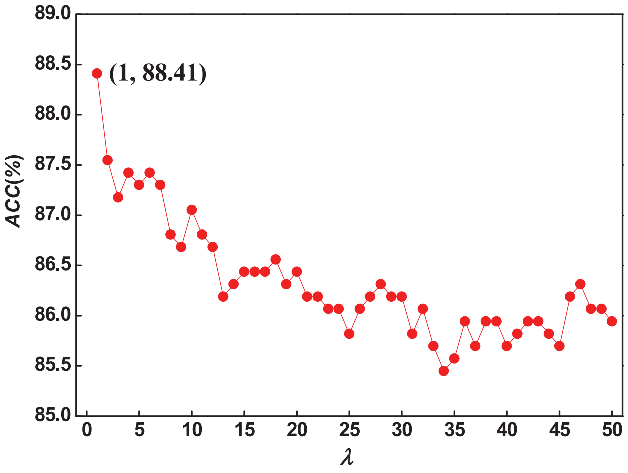

Fig. 2. A plot showing the 50 optimal accuracies with their respective best feature set based on *F*-score when $k=6$, $1 \le \lambda \le 50$. The peak is 88.41% when $k=6$, $\lambda=1$ with top 1246 features in 5-fold cross-validation

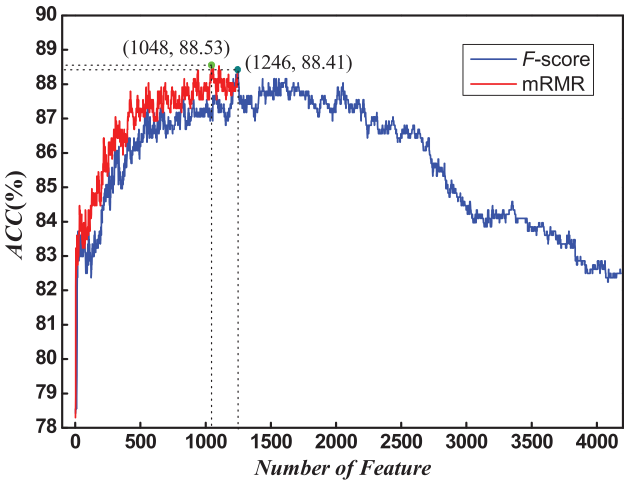

Fig. 3. A plot showing the IFS procedure for identifying ORIs. When the top 1246 features optimized by *F*-score were used to perform prediction, the overall success rate reaches an IFS peak of 88.41% in 5-fold cross-validation

Table 1. The predictive performance three feature subsets based on 5-fold cross-validation

Features	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*	auROC
$k=6,\ \lambda=50$ 8596-D	84.69	82.27	85.70	0.6698	0.8704
F-score 1246-D	90.12	86.70	88.41	0.7687	0.9022
mRMR 1048-D	90.37	86.70	88.53	0.7712	0.9045

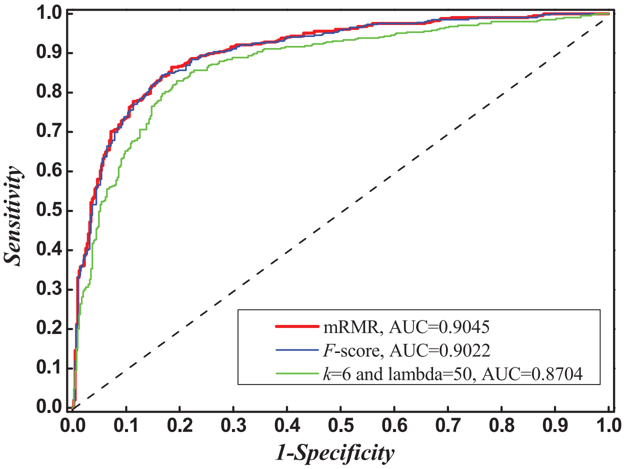

Fig. 4. ROC curves for three feature subsets with the dimensions of 1048, 1246 and 8596

Table 2. Comparative results for identifying ORIs based on different methods by jackknife cross-validation

Methods	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*	auROC
Bendability + cleavage intensity (Chen *et al.*, 2012)	81.23	80.30	80.76	0.6153	0.8563
Type-I PseKNC (Li *et al.*, 2015b)	84.69	82.76	83.72	0.6746	0.8848
Type-II PseKNC	89.63	85.96	87.79	0.7564	0.9110

cross-validation. All compared results were recorded in Table 2. It is obvious that the model proposed in this paper is superior to other published models for identifying ORIs.

Recently, Liu *et al.* (2018) developed a predictor called 'iRO-3wPseKNC' to classify four yeast species by rigorous cross validations. To provide a fair comparison, we applied the method proposed in this paper on the *S.cerevisiae* dataset constructed by Liu *et al.* (2018). The compared details can be found in Table 3. Obviously, our method is superior to iRO-3wPseKNC.

Using the same dataset, Xiao *et al.* (2016) incorporated the dinucleotide position-specific propensity information into the general pseudo nucleotide composition to formulate ORIs and non-ORIs samples. A very high accuracy was obtained. However, the feature about position information was not suitable for ORI identification because of the following reason. Currently, biochemical experimentally-based method cannot determine the ORI precisely. It can only limit some small regions which contain ORIs. Thus, in order to include all possible ORIs, the length of positive samples was set to 300 bp in benchmark dataset. An example was as shown in Figure 5. In the figure, blue boxes denote the ORIs regions. Pink boxes denote flanking regions of ORIs, which are also regarded as a part of positive samples. The seq *1*, seq *2*,. . ., seq *5* can express as positive samples in training set. It is clear that the position distribution of the ORIs in positive samples is not fixed in alignment. Therefore, from the biological viewpoint, it is unreasonable to use position information to represent sequence characteristics for ORI prediction.

Table 3. Comparison between proposed method in this paper and iRO-3wPseKNC on *S.cerevisiae* dataset

Methods	*Sn*(%)	*Sp*(%)	*Acc*(%)	*MCC*	auROC
iRO-3wPseKNC	70.7	75.2	73.0	0.459	0.808
Our method	76.3	80.2	78.2	0.565	0.831

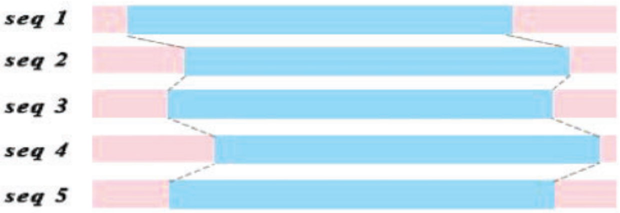

Fig. 5. A schematic diagram of experimental-confirmed ORI sequences in gene. The seq *1*, seq *2*,. . ., seq *5* can express positive samples in training set with the length of 300 bp and blue boxes denote the ORIs regions, pink boxes denote flanking regions of ORIs

We believe that pseudo-components are the most reasonable method for representing a sequence, and it keeps considerable sequence-order information, particularly the global or long-range sequence-order information via the physicochemical properties of its constituent dinucleotides, however, incorporating position-specific propensity information into pseudo-components has no enough biological significance to formulate ORI samples. Using features without biological significance to perform prediction maybe result in model overtraining and results overestimation.

3.4 Prediction of ORIs in *S.cerevisiae* genome-wide

In order to further evaluate the generalization ability of the proposed model, we performed the model to identify the ORIs in *S.cerevisiae* genome by scanning its 16 chromosomes with the window of 300 bp and the step of 1 bp. As a results, total of 1.21×10^7 subsequences were examined by evaluating their prediction probabilities with our proposed model. The prediction probabilities of 410 experimental-confirmed ORIs were also achieved and then ranked in a descending order. Therefore, there are 410 cutoff[i] ($i = 1, 2, \cdots, 410$) for prediction probabilities of 410 experimental-confirmed ORIs. Here, the following two principles were utilized to evaluate subsequences: (i) the predicted subsequence would be regarded as a true positive if it locates in the region from 200 bp upstream to 200 bp downstream of an experimentally-confirmed ORI; (ii) two predictions are regarded as one prediction if their distance is <300 bp. According to above rules, we counted the number of subsequences whose predicted probabilities are greater than the corresponding cutoff[i]. The details were shown in Figure 6. When we set cutoff of the predicted probability as 0.5, total of 373 experimental-confirmed ORIs ($Sn = 90.9\%$) can be correctly identified. At same time, we also obtained 5135 potential ORIs in genome-wide. Obviously, comparing to the experimental-based methods, our

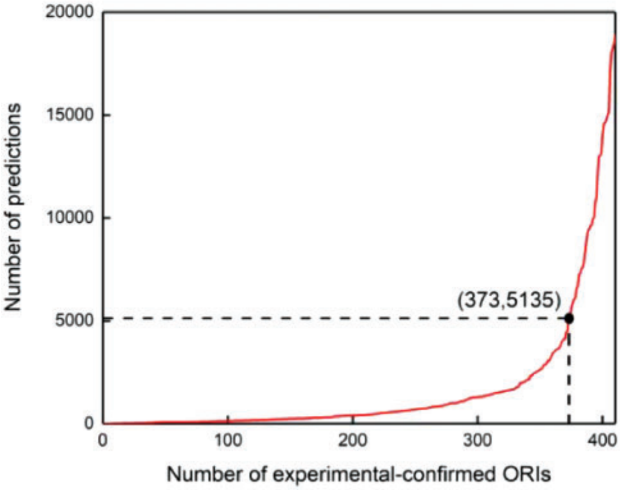

Fig. 6. The genome-scanned results using the classifier of iORI-PseKNC2.0. The abscissa is the number of experimental-confirmed ORIs; the ordinate denotes the number of predictions from genome-wide

computational method can scan genome-wide quickly and identify potential ORIs with high generalization ability.

3.5 Webserver

Webserver and database are popular in the internet age because they could provide more convenience to the majority of wet-experiments scholars, for whom it is difficult to understand the mathematics and calculation process (Cao *et al.*, 2016, 2017a, b; Cheng *et al.*, 2018; Cui *et al.*, 2018; He *et al.*, 2018; Li *et al.*, 2017; Liang *et al.*, 2017; Liu *et al.*, 2015; Tang *et al.*, 2018b; Yi *et al.*, 2017; Zhang *et al.*, 2017; Zhu *et al.*, 2015; Zou *et al.*, 2016). Thus, based on our proposed model, the predictor iORI-PseKNC was improved to version 2.0. A step-by-step guide on how to use the webserver is given as follows:

Step 1: open the URL (http://lin-group.cn/server/iORI-PseKNC2.0/) and the screen contained brief introduction about the predictor will appear on your computer (Fig. 7). You can click on the 'User guide' to obtain the relevant information.

Step 2: click on the 'Web server' button, either type or copy/paste the query DNA sequences into the input box at the center of Figure 6, or upload a file in FASTA format. It is worth noting that the input sequence should be in the FASTA format and the length of each query sequence should be more than 300 bp but not too long due to the high dimension features. Click on the 'example' button below the input box to see the sample sequence in the FASTA format.

Step 3: click on the 'submit' button to obtain the predicted result.

4 Conclusions

Pseudo nucleotide composition is a very popular formulation for describing nucleotide sequence samples, and has been successfully applied in the recognition of nucleosome positioning (Guo *et al.*, 2014), promoter (Lin *et al.*, 2017), splice site (Chen *et al.*, 2014b) and translation initial site (Chen *et al.*, 2014a). The Type-I PseKNC has also been used in ORI identification (Li *et al.*, 2015b; Zhang *et al.*, 2016). However, in the Type-I PseKNC, the different physiochemical properties were added in one formulation so that some

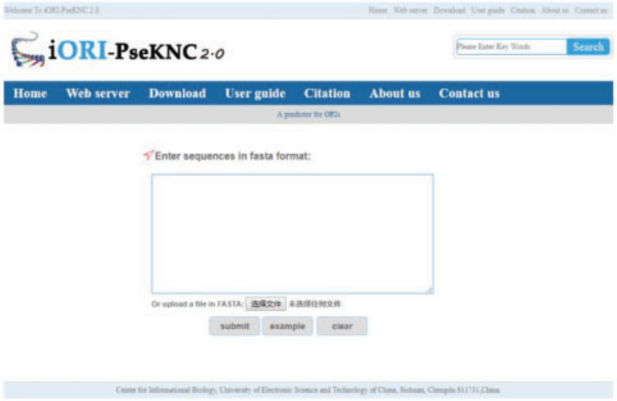
Welcome To iORI-PseKNC 2.0
Home Web server Download User guide Citation About us Contact us
iORI-PseKNC 2.0
Please Enter Key Words
Search
Home
Web server
Download
User guide
Citation
About us
Contact us
A predictor for ORIs
Enter sequences in fasta format:
Or upload a file in FASTA:
选择文件
未选择任何文件
submit
example
clear
Center for Informational Biology, University of Electronic Science and Technology of China, Sichuan, Chengdu 611731,China

Fig. 7. A semi-screenshot for the webserver page of the iORI-PseKNC2.0 webserver at http://lin-group.cn/server/iORI-PseKNC2.0

important information was annihilated. Thus, to consider different correlation information independently, in this paper, the Type-II PseKNC was used to describe ORI samples. Due to high-dimensional features could bring out dimensionality disaster which results in low calculation efficiency, noise which reduces the prediction accuracy of the proposed method and over-fitting problem which overestimates the performance of proposed model and reduces the robust of models, we introduced a two-step feature selection strategy by combining *F*-score and mRMR techniques to reduce the feature dimension. The strategy not only overcomes the low running speed of mRMR, but also picks out the optimal features and obtains the maximum accuracy. As a result, the final model could produce the overall accuracy of 88.53% in 5-fold cross-validation. Based on the proposed model, a new predictor called iORI-PseKNC2.0 was established. We anticipated that the predictor will become a useful tool in relevant fields. In addition, the algorithm in this study takes full account of the sequence component information and physicochemical property information, and screening features by two feature selection methods, that shows, the final feature set is best to characterize ORI sequence. We advise researchers can focus on the other ORI prediction in bacteria, yeast and human genomes by using our method.

Funding

This work was supported by the National Nature Scientific Foundation of China [61772119, 31771471]; the Natural Science Foundation for Distinguished Young Scholar of Hebei Province [grant number C2017209244]; the Fundamental Research Funds for the Central Universities of China [grant numbers ZYGX2015Z006, ZYGX2016J118, ZYGX2016J125, ZYGX2016J223]; and the Science Strength Promotion Program of UESTC.

Conflict of Interest: none declared.

References

Cao,R. *et al.* (2017a) ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules*, **22**, E1732.

Cao,R.Z. *et al.* (2017b) QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics*, **33**, 586–588.

Cao,R.Z. *et al.* (2016) DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC Bioinformatics*, **17**, 495.

Chang,C.C. and Lin,C.J. (2011) LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 1–27.

Chen,W. *et al.* (2012) Prediction of replication origins by calculating DNA structural properties. *FEBS Lett.*, **586**, 934–938.

Chen,W. *et al.* (2014a) iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal. Biochem.*, **462**, 76–83.

Chen,W. *et al.* (2014b) iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *BioMed Res. Int.*, **2014**, 623149.

Chen,W. *et al.* (2014c) PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.*, **456**, 53–60.

Chen,W. *et al.* (2017) iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics*, **33**, 3518–3523.

Chen,W. *et al.* (2015) PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics*, **31**, 119–120.

Chen,X.X. *et al.* (2016) Identification of bacterial cell wall lyases via pseudo amino acid composition. *BioMed Res. Int.*, **2016**, 1654623.

Cheng,J.H. *et al.* (2018) Prediction of bacteriophage proteins located in the host cell using hybrid features. *Chemometr. Intell. Lab. Syst.*, **180**, 64–69.

Chou,K.C. and Zhang,C.T. (1995) Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, **30**, 275–349.

Cui,T. *et al.* (2018) MNDR v2.0: an updated resource of ncRNA-disease associations in mammals. *Nucleic Acids Res.*, **46**, D371–D374.

Dhar,M.K. *et al.* (2012) Structure, replication efficiency and fragility of yeast ARS elements. *Res. Microbiol.*, **163**, 243–253.

Ding,C. *et al.* (2012) Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions. *J. Proteomics*, **77**, 321–328.

Feng,C.Q. *et al.* (2019) iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics*, **35**, 1469–1477.

Foureau,E. *et al.* (2013) Characterization of an autonomously replicating sequence in Candida guilliermondii. *Microbiol. Res.*, **168**, 580–588.

Gao,F. and Zhang,C.T. (2008) Ori-Finder: a web-based system for finding oriCs in unannotated bacterial genomes. *BMC Bioinformatics*, **9**, 79.

Goni,J.R. *et al.* (2008) DNAlive: a tool for the physical analysis of DNA at the genomic scale. *Bioinformatics*, **24**, 1731–1732.

Goni,J.R. *et al.* (2007) Determining promoter location based on DNA structure first-principles calculations. *Genome Biol.*, **8**, R263.

Guo,S.H. *et al.* (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics*, **30**, 1522–1529.

He,W.Y. *et al.* (2018) 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. *BMC Syst. Biol.*, **12**, 44.

Zakrzewska-Czerwińska,J. *et al.* (2007) Regulation of the initiation of chromosomal replication in bacteria. *FEMS Microbiol. Rev.*, **31**, 378–387.

Jia,C. and He,W. (2016) EnhancerPred: a predictor for discovering enhancers based on the combination and selection of multiple features. *Sci. Rep.*, **6**, 38741.

Keerthi,S.S. and Lin,C.J. (2003) Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Comput.*, **15**, 1667–1689.

Lai,H.Y. *et al.* (2017) Sequence-based predictive modeling to identify cancerlectins. *Oncotarget*, **8**, 28169–28175.

Lee,D.G. and Bell,S.P. (1997) Architecture of the yeast origin recognition complex bound to origins of DNA replication. *Mol. Cell. Biol.*, **17**, 7159–7168.

Li,F. *et al.* (2018) Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. *Bioinformatics*, **34**, 4223–4231.

Li,F. *et al.* (2016) GlycoMine(struct): a new bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features. *Sci. Rep.*, **6**, 34595.

Li,F. *et al.* (2015a) GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. *Bioinformatics*, **31**, 1411–1419.

Li,N. *et al.* (2017) PSBinder: a web service for predicting polystyrene surface-binding peptides. *BioMed Res. Int.*, **2017**, 1.

Li,W.C. *et al.* (2015b) iORI-PseKNC: a predictor for identifying origin of replication with pseudo k -tuple nucleotide composition. *Chemometr. Intell. Lab. Syst.*, **141**, 100–106.

Li,W.C. *et al.* (2014a) Sequence analysis of origins of replication in the *Saccharomyces cerevisiae* genomes. *Front. Microbiol.*, **5**, 574.

Li,Y. *et al.* (2014b) Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features. *Sci. Rep.*, **4**, 5765.

Liang,Z.Y. *et al.* (2017) Pro54DB: a database for experimentally verified sigma-54 promoters. *Bioinformatics*, **33**, 467–469.

Lin,H. *et al.* (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.*, **42**, 12961–12972.

Lin,H. *et al.* (2017) Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, doi: 10.1109/TCBB.2017.266614.

Liu,B. *et al.* (2015) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.*, **43**, W65–W71.

Liu,B. *et al.* (2018) iRO-3wPseKNC: identify DNA replication origins by three-window-based PseKNC. *Bioinformatics*, **34**, 3086–3093.

Lubelsky,Y. *et al.* (2012) Genome-wide localization of replication factors. *Methods*, **57**, 187–195.

Luo,H. *et al.* (2014) Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes. *Front. Microbiol.*, **5**, 482.

Manavalan,B. and Lee,J. (2017) SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics*, **33**, 2496–2503.

Manavalan,B. *et al.* (2018) PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.*, **9**, 476.

Marczynski,G.T. and Shapiro,L. (1993) Bacterial chromosome origins of replication. *Curr. Opin. Genet. Dev.*, **3**, 775–782.

McFadden,G.I. and Roos,D.S. (1999) Apicomplexan plastids as drug targets. *Trends Microbiol.*, **7**, 328–333.

Mechali,M. (2010) Eukaryotic DNA replication origins: many choices for appropriate answers. *Nat. Rev. Mol. Cell Biol.*, **11**, 728–738.

Metz,C.E. (1989) Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest. Radiol.*, **24**, 234–245.

Miele,V. *et al.* (2008) DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.*, **36**, 3746–3756.

Mundra,P.A. and Rajapakse,J.C. (2010) SVM-RFE with MRMR filter for gene selection. *IEEE Trans. Nanobioscience*, **9**, 31–37.

Nieduszynski,C.A. *et al.* (2007) OriDB: a DNA replication origin database. *Nucleic Acids Res.*, **35**, D40–D46.

Nieduszynski,C.A. *et al.* (2006) Genome-wide identification of replication origins in yeast by comparative genomics. *Genes Dev.*, **20**, 1874–1879.

Peng,H. *et al.* (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 1226–1238.

Raghu Ram,E.V. *et al.* (2007) Nuclear gyrB encodes a functional subunit of the Plasmodium falciparum gyrase that is involved in apicoplast DNA replication. *Mol. Biochem. Parasitol.*, **154**, 30–39.

Rao,H. and Stillman,B. (1995) The origin recognition complex interacts with a bipartite DNA binding site within yeast replicators. *Proc. Natl. Acad. Sci. USA*, **92**, 2224–2228.

Rowley,A. *et al.* (1995) Initiation complex assembly at budding yeast replication origins begins with the recognition of a bipartite sequence by limiting amounts of the initiator, ORC. *EMBO J.*, **14**, 2631–2641.

Schub,O. *et al.* (2001) Multiple phosphorylation sites of DNA polymerase alpha-primase cooperate to regulate the initiation of DNA replication in vitro. *J. Biol. Chem.*, **276**, 38076–38083.

Shah,K. and Krishnamachari,A. (2012) Nucleotide correlation based measure for identifying origin of replication in genomic sequences. *Biosystems*, **107**, 52–55.

Singh,V.K. *et al.* (2018) Prediction of replication sites in *Saccharomyces cerevisiae* zgenome using DNA segment properties: multi-view ensemble learning (MEL) approach. *Biosystems*, **163**, 59–69.

Soldati,D. (1999) The apicoplast as a potential therapeutic target in and other apicomplexan parasites. *Parasitol. Today*, **15**, 5–7.

Song,C. *et al.* (2015) Choosing a suitable method for the identification of replication origins in microbial genomes. *Front. Microbiol.*, **6**, 1049.

Song,J. *et al.* (2018a) PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics*, **34**, 684–687.

Song,J. *et al.* (2018b) iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Brief. Bioinform.*, doi: 10.1093/bib/bby028.

Stephenson,N. *et al.* (2018) Survey of machine learning techniques in drug discovery. *Curr. Drug Metab.*, doi: 10.2174/1389200219666180820112457.

Su,Z.D. *et al.* (2018) iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics*, **34**, 4196–4204.

Tang,H. *et al.* (2017) A two-step discriminated method to identify thermophilic proteins. *Int. J. Biomath.*, **10**, 1750050.

Tang,H. *et al.* (2018a) HBPred: a tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.*, **14**, 957–964.

Tang,W. *et al.* (2018b) Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics*, **34**, 398–406.

VapnikVladimir,N. (1997) The nature of statistical learning theory. *IEEE Trans. Neural Netw.*, **8**, 1564.

Wang,M. *et al.* (2012) FunSAV: predicting the functional effect of single amino acid variants using a two-stage random forest model. *PLoS One*, **7**, e43847.

Wei,L. *et al.* (2018) ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics*, **34**, 4007–4016.

Xiao,X. *et al.* (2016) iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. *Oncotarget*, **7**, 34180–34189.

Yang,H. *et al.* (2018a) iRNA-2OM: a sequence-based predictor for identifying 2'-O-methylation sites in Homo sapiens. *J. Comput. Biol.*, **25**, 1266–1277.

Yang,H. *et al.* (2018b) iRSpot-Pse6NC: identifying recombination spots in *Saccharomyces cerevisiae* by incorporating hexamer composition into general PseKNC. *Int. J. Biol. Sci.*, **14**, 883–891.

Yang,H. *et al.* (2016) Identification of secretory proteins in mycobacterium tuberculosis using pseudo amino acid composition. *BioMed Res. Int.*, **2016**, 1.

Yi,Y. *et al.* (2017) RAID v2.0: an updated resource of RNA-associated interactions across organisms. *Nucleic Acids Res.*, **45**, D115–D118.

Yuan,L.F. *et al.* (2013) Prediction of the types of ion channel-targeted conotoxins based on radial basis function network. *Toxicol. In Vitro*, **27**, 852–856.

Zhang,C.J. *et al.* (2016) iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget*, **7**, 69783–69793.

Zhang,T. *et al.* (2017) RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res.*, **45**, D135–D138.

Zhu,P.P. *et al.* (2015) Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition. *Mol. Biosyst.*, **11**, 558–563.

Zhu,X.J. *et al.* (2018) Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl. Based Syst.*, doi: 10.1016/j.knosys.2018.10.007.

Zou,Q. *et al.* (2016) Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.*, **10**, 114.