Chemometrics and Intelligent Laboratory Systems 141 (2015) 100–106

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

iORI-PseKNC: A predictor for identifying origin of replication with pseudo *k*-tuple nucleotide composition

Wen-Chao Li [a], En-Ze Deng [a], Hui Ding [a,*], Wei Chen [b,*], Hao Lin [a,*]

[a] Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
[b] Department of Physics, School of Sciences, Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China

ARTICLE INFO

Article history:
Received 25 June 2014
Received in revised form 21 September 2014
Accepted 23 December 2014
Available online 3 January 2015

Keywords:
Saccharomyces cerevisiae
Origin of replication
Pseudo *k*-tuple nucleotide composition
DNA local structural property

ABSTRACT

The initiation of replication origin is an extremely important process of DNA replication. The distribution of replication origin regions (ORIs) is the major determinant of the timing of genome replication. Thus, correctly identifying ORIs is crucial to understand DNA replication mechanism. With the avalanche of genome sequences generated in the post-genomic age, it is highly desired to develop computational methods for rapidly, effectively and automatically identifying the ORIs in genome. In this paper, we developed a predictor called iORI-PseKNC for identifying ORIs in *Saccharomyces cerevisiae* genome. In the predictor, based on the concept of the global and long-range sequence-order effects of DNA sequence, the feature called "pseudo *k*-tuple nucleotide composition" (PseKNC) was used to encode the DNA sequences by incorporating six local structural properties of 16 dinucleotides. The overall success rate of 83.72% was achieved from the jackknife cross-validation test on an objective benchmark dataset. Comparisons demonstrate that the new predictor is superior to other methods. As a user-friendly web-server, iORI-PseKNC is freely accessible at http://lin.uestc.edu.cn/server/iORI-PseKNC. We hope that iORI-PseKNC will become a useful tool or at least as a complement to existing methods for identifying ORIs.

1. Introduction

In cell division, DNA replication is a highly orchestrated process of producing an identical replica from the original DNA molecule [1]. This process commonly initiates at specific regions called origin of replication regions (ORIs). Most of bacterial genomes have only a single ORI [2]. However, in eukaryotic genome, due to the large size of genomes and the limitation of nucleotide incorporation during DNA synthesis, it is necessary for completing replication in a reasonable period of time using multiple ORIs [3]. In *Saccharomyces cerevisiae*, an autonomously replicating sequence (ARS) element contains the ORI [4], which consists of three domains A, B and C. The A domain contains an essential ARS consensus sequence (T/A)TTTAT(A/G)TTT(T/A), while the B domain tends to be helically unstable and additionally contains a number of short sequence motifs that contribute to origin activity [5,6]. The C domain plays an important role in the interaction between DNA and regulatory protein [5].

During the process of replication, the DNA double helix strands in ORIs are dissociated and unwinded by helicases for allowing access to DNA polymerase [7]. Subsequently, the semiconservative replication strategy is used to synthesize daughter strands based on the parental template strands [8]. The replication is activated only once at each cell cycle to avoid amplification and maintain genome integrity [9]. DNA replication is associated with gene transcription and expression [10]. For example, an analysis from the distribution of ORIs showed that replication initiation events were absent from transcription start sites but were highly enriched in adjacent, downstream sequences [11]. Therefore, it is crucial to understand the regulatory mechanism of cell division and establish the network of a cell cycle so as to reveal the mechanism involved in DNA replication. Accurate identification of ORIs is an essential prerequisite for further studying and understanding the DNA replication mechanisms.

Chromatin immunoprecipitation (Chip) is the most popular technique to determine ORIs [12]. Although the technique can precisely identify the ORIs, with the avalanche of genome sequences generated in the post-genomic era, it is expensive and time-consuming for experimental approaches to perform genome-wide identification of ORIs. In this regard, computational methods can be applied to the entire genome without these disadvantages. Based on the consensus sequence [13], some theoretical works have been proposed in order to accurately identify ORIs. Marie-Claude et al. have predicted the ORIs by analyzing asymmetry indices of sequence [14]. The signal of nucleosome occupancy was used as a likely candidate to determine ORI distribution [15,16]. Although it is of great interest and value, the ACS-based method is not sufficient to predict ORIs [17] because there are 12,000 ACS sites in *S. cerevisiae* genomes, and only 400 associate with ORIs [18]. Recently, two DNA structural

* Corresponding authors.
E-mail addresses: hding@uestc.edu.cn (H. Ding), greatchen@heuu.edu.cn (W. Chen), hlin@uestc.edu.cn (H. Lin).

http://dx.doi.org/10.1016/j.chemolab.2014.12.011
0169-7439/

properties, namely DNA bendability [19] and hydroxyl radical cleavage intensity [20,21], were proposed to predict ORIs in *S. cerevisiae* genome [22]. Although these methods have achieved encouraging results, they are still limited in their accuracy and resolution. Moreover, no web-server was provided to most of these methods, and hence their usage is quite limited, especially for the majority of experimental scientists.

It has been reported that the local DNA structural properties [23] and their impacts to the global sequence effects are important feature signals for DNA functional elements and have been used to identify the nucleosome occupancy [24], recombination spots [25] and exon/intron splice site [26]. DNA conformation may be changed by the ionic bonding effects in the methylation form of specific bases [27]. Besides, the cell differentiation is caused by the dynamical position of nucleosomes due to the chemical reactions, where cell lines have different ORIs [28].

In view of this, the present study was initiated in an attempt to develop a new method for predicting *S. cerevisiae* ORIs based on the physicochemical properties of DNA. At first, a valid benchmark database was constructed to train and test the proposed method. Subsequently, the DNA sequences were encoded with the pseudo *k*-tuple nucleotide composition (PseKNC), which can reflect the intrinsic correlation between local/global features and the ORIs. Thirdly, a powerful algorithm SVM was used to operate the prediction by using rigorous jackknife cross-validation test to evaluate the performance of the proposed method. Finally, based on the proposed method, a user-friendly web-server, called iORI-PseKNC, was established for basic academic study and application of ORIs.

2. Materials and methods

2.1. Datasets

A total of 740 *S. cerevisiae* ORIs were collected from OriDB [29] (http://www.oridb.org/). The following steps were used to construct a reliable benchmark dataset. Firstly, the ORIs with ambiguous annotation such as "likely" and "dubious" were excluded because they lack confidence. Then, we obtained 410 experimental-confirmed ORIs with the length of 300 bp. Subsequently, the 410 non-ORI samples with 300 bp long were extracted from −600 bp to −300 bp upstream of the 410 ORIs. It is well known that high similarity data can lead to erroneous estimation of the performance of the methods. To get rid of redundancy and avoid bias, the CD-HIT software [30] was used to remove those samples that have more than 75% pairwise sequence identity to any other. After strictly following the above procedures, we finally obtained 405 ORIs and 406 non-ORIs which can be freely downloaded from our website (http://lin.uestc.edu.cn/server/iOriPseKNC/data.html).

2.2. Pseudo k-tuple nucleotide composition

In pattern recognition, one of the key points is to generate a set of informative parameters. It has become a challenge in DNA functional region prediction to formulate DNA sequences with an effective mathematical expression for truly reflecting the intrinsic properties of DNA functional fragments. In the past two decades, various sequence parameters such as *k*-tuple nucleotide composition [31,32], Z-curve [33] and position weight matrix (PWM) [34] have been successfully employed to predict gene coding region and promoter. However, these methods ignored the local DNA structural properties and their impacts to the global sequence effects. Recently, a novel feature vector, called 'pseudo *k*-tuple nucleotide composition' (PseKNC), was developed to represent DNA sequence samples by incorporating the global and long-range sequence-order effects of DNA sequence and has been applied for predicting the recombination spots [25], nucleosome [24], and exon/intron splice site [35]. Thus, in this work, the PseKNC was used in the ORI prediction. The basic principle of PseKNC is introduced briefly as below.

Consider an ORI (or non-ORI) DNA sequence D with L nucleic acid residues; i.e.,

$$D = R_1 R_2 R_3 \cdots R_{L-2} R_{L-1} R_L \tag{1}$$

where R_1 denotes the nucleic acid residue at the sequence position 1, R_2 denotes the nucleic acid residue at position 2, and so forth. Here, L is 300. Then the ORI (or non-ORI) sample can be denoted as a $4^k + \lambda$ dimension vector which is formulated as:

$$D = \left[d_1 d_2 \ldots d_{4^k} d_{4^k+1} \ldots d_{4^k+\lambda-1} d_{4^k+\lambda}\right] \tag{2}$$

where

$$d_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{4^k} f_i + \omega \sum_{j=1}^{\lambda} \theta_j} & \left(1 \le u \le 4^k\right) \\ \dfrac{\omega \theta_{u-4^k}}{\sum_{i=1}^{4^k} f_i + \omega \sum_{j=1}^{\lambda} \theta_j} & \left(4^k + 1 \le u \le 4^k + \lambda\right) \end{cases} \tag{3}$$

where f_u $(u = 1, 2, \ldots, 4^k)$ denotes the normalized occurrence frequency of the u-th k-tuple nucleotide composition, λ is the number of the total counted ranks (or tiers) of the correlations along a DNA sequence, and ω is the weight factor. The concrete values for λ and ω as well as k will be further discussed later. The θ_j is the j-th tire structural correlation factor that reflects the local structure correlation between all the j-th most contiguous dinucleotide and can be given by

$$\theta_j = \frac{1}{L-j-1} \sum_{i=1}^{L-j-1} \Theta\left(R_i R_{i+1}, R_{i+j} R_{i+j+1}\right) \quad (j = 1, 2, \cdots, \lambda; \lambda < L) \tag{4}$$

where the $\Theta(R_i R_{i+1}, R_{i+j} R_{i+j+1})$ is the correlation function and can be defined by

$$\Theta\left(R_i R_{i+1}, R_{i+j} R_{i+j+1}\right) = \frac{1}{\mu} \sum_{v=1}^{\mu} \left[P_v(R_i R_{i+1}) - P_v\left(R_{i+j} R_{i+j+1}\right)\right]^2 \tag{5}$$

where μ is the number of local DNA structural properties considered that is equal to 6 in the current study as will be explained below; $P_v(R_i R_{i+1})$ is the numerical value of the v-th ($v = 1, 2, \ldots, \mu$) DNA local structural 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 $i + j$.

The spatial arrangements of any two successive base pairs could be characterized by six types of local structural parameters, of which three are local translational parameters (shift, slide and rise) and the other three are local angular parameters (twist, tilt and roll) [24,36]. In recent years, more and more researches have demonstrated that the six DNA structural properties play important roles in many biological processes [37,38]. There are sixteen kinds of dinucleotides, so the total number of local structural parameters is $6 \times 16 = 96$. The parameter values can be found from Ref. [24].

Before substituting the six types of parameters of dinucleotides into Eq. (5), all the original values must be subjected to a standard conversion, as described by the following equation:

$$P_v(\xi) = \frac{P_v^0(\xi) - \sum_{\xi=1}^{16} \left[P_v^0(\xi)/16\right]}{\sqrt{\dfrac{\sum_{\xi=1}^{16} \left\{P_v^0(\xi) - \sum_{\xi=1}^{16} \left[P_v^0(\xi)/16\right]\right\}^2}{16}}} \tag{6}$$

where the $P_v(\xi)$ and $P_v^0(\xi)$ denote the standard value and original value of ξ-th dinucleotide of v-th local structural parameter, respectively. The converted values obtained by Eq. (6) will have a zero mean value over

the 16 different dinucleotides, and will remain unchanged if going through the same conversion procedure again.

2.3. Support vector machine (SVM)

Support vector machine (SVM) is a popular supervised machine learning method which has been widely used in bioinformatics [22,24,39] and chemometrics [40–42]. In this study, we also used SVM to discriminate ORIs from non-ORIs. The basic principle of SVM is to transform the input vector into a high-dimension Hilbert space and seek a separating hyperplane with the maximal margin in this space. Detailed descriptions about SVM can be referred to Ref. [43].

The software toolbox (LibSVM) for implementing SVM can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm. The radial basis kernel function (RBF) was used in the current work due to its effectiveness and speed in non-linear classification process. A grid search method was used to optimize the regularization parameter C and kernel parameter γ through 10-fold cross-validation for time-saving. The search spaces for C and γ are $[2^{15}, 2^{-5}]$ and $[2^{-5},2^{-15}]$ with the steps of 2^{-1} and 2, respectively.

2.4. Criteria for performance evaluation

In statistical prediction, three cross-validation methods, namely independent dataset test, sub-sampling (e.g., 2, 5 or 10-fold cross-validation) test, and jackknife test are often used to evaluate the performance of the predicted methods in practical application [44]. Among the three test methods, the jackknife test can always yield a unique result for a given benchmark dataset. Therefore, the jackknife test has been increasingly and widely adopted by investigators to test the power of various prediction methods [45,46]. In this study, we also used the jackknife test to examine the anticipated success rates of the predictor. In the jackknife test, each sequence in the training dataset is in turn singled out as an independent test sample and all the rule-parameters are calculated without including the one being identified.

The performance of the predictor was evaluated by the following four metrics: sensitivity (Sn), specificity (Sp), overall accuracy (Acc) and Mathew's correlation coefficient (Mcc), which are expressed as follows:

$$Sn = \frac{TP}{TP + FN} \tag{8}$$

$$Sp = \frac{TN}{TN + FP} \tag{9}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Mcc = \frac{TP \times TN - FP \times FN}{(TP + FN) \times (TN + FN) \times (TP + FP) \times (TN + FP)} \tag{11}$$

where TP denotes the number of the correctly predicted ORIs, FN denotes the number of the ORIs predicted as non-ORIs, FP denotes the number of the non-ORIs predicted as ORIs, and TN denotes the number of correctly predicted non-ORIs.

To describe the performance of models across the entire range of SVM decision values, the receiver operating characteristic (ROC) curves were also provided. The quality of the proposed method can be objectively evaluated by measuring the area under the receiver operating characteristic curve (auROC).

3. Result and discussion

3.1. Profile analyze for local structural property

The specific conformation of DNA sequence can be recognized by regulatory proteins [47–49]. In order to explore the specific features possessed by ORI sequences, six structural parameters (twist, tilt, roll, shift, slide and rise) of both ORI and non-ORI sequences was calculated to characterize the local geometry with the step of one base-pair in the *S. cerevisiae* genome. Using graphic approaches to study ORIs can provide an intuitive picture and useful insights for revealing complicated relations in the DNA replication origin system. Thus, we carried out a graphic profile comparison between ORIs and non-ORIs. By using a sliding window approach with a window size of 50 bp and a step size of 1 bp, the average structural property profiles for both ORI and non-ORI sequences were plotted in Fig. 1, which consists of 6 panels corresponding to the 6 local structural properties of DNA sequences.

Replication initiation is the first stages of DNA replication in eukaryotic genomes. During initiation, the DNA helix is unwound by helicase to form replication forks at the ORI. As shown in Fig. 1, except for few sites in roll, shift and twist panels, the differences between the ORI and non-ORI sequences are quite remarkable in all six panels ($P < 10^{-11}$, U-test). These observations suggest that the distinctive flexibility and stiffness curvature in ORI sequences may be one of the key factors that can promote regulatory proteins and helicases binding to ORI regions for activating the replication, and that may be also contributable for discriminating ORI from non-ORI sequences.

3.2. Prediction accuracy

As it can be seen from Eqs. (2)–(3), we must adjust the three parameters, namely k, λ and ω, to achieve the best prediction accuracy. The parameter k of k-tuple reflects the local or short-range sequence order effect. The parameter λ represents the global or long-range sequence order effect. The parameter ω in Eq. (3) is the weight factor of long-range effect usually within the range from 0 to 1. Generally speaking, the greater the k is, the more local sequence-order information the model contains. Moreover, the greater the λ is, the more global sequence-order information the model contains. However, if k or λ is too large, it would reduce the cluster-tolerant capacity so as to lower down the cross-validation accuracy due to overfitting or "high dimension disaster" problem [24]. Therefore, our searching for the optimal values for the three parameters were carried out according to the following

$$\begin{cases} 2 \le k \le 5 & (\text{with step } \Delta = 1) \\ 1 \le \lambda \le 50 & (\text{with step } \Delta = 1) \\ 0.1 \le \omega \le 1.0 & (\text{with step } \Delta = 0.1) \end{cases}. \tag{12}$$

According to Eq. (12), total of $4 \times 50 \times 10 = 2000$ individual combinations (or points in the 3D parameter space) should be investigated for finding the optimal parameter combination. To reduce the computational time, the 10-fold cross-validation approach was used to assess the accuracies of 2000 combinations in the process of parameter optimization. Once the optimal values of the three parameters are determined, the rigorous jackknife test will be performed to finally evaluate the anticipated success rate of the model. As a result, the maximum Acc of 83.72% was obtained with the Sn of 84.69%, Sp of 82.76% and Mcc of 0.6746 (Table 1) when the parameters k, λ and ω are equal to 3, 50 and 0.6, respectively. That means the optimal feature set contains 114 features, of which 64 are 3-tuple oligonucleotides ($4^3 = 64$) reflect the short-range information, and 50 are the correlation of structural properties reflecting the long-range information. From the results, we concluded that the global or long-range correlation of structural properties play an important role in ORI recognition.

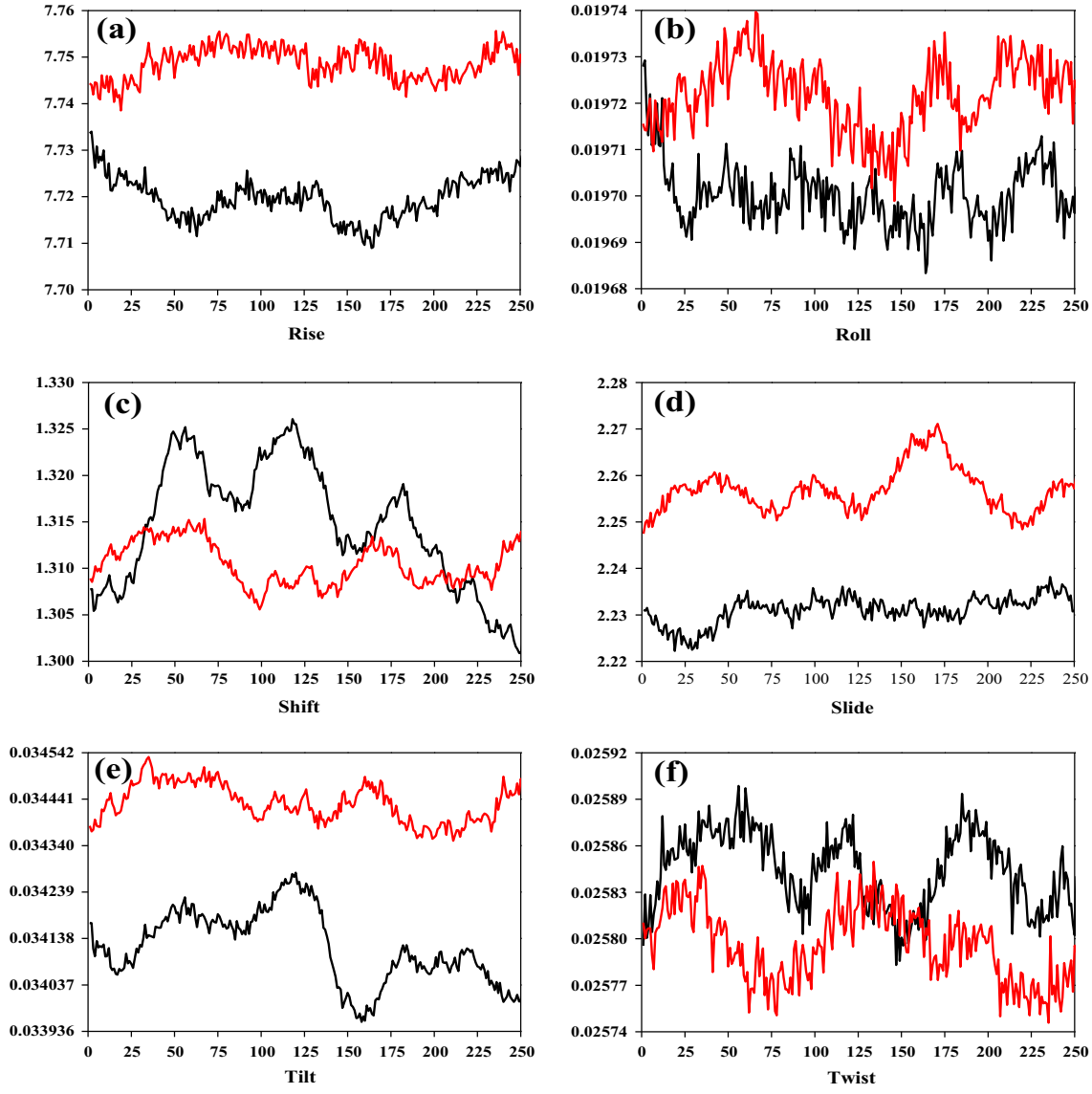

Fig. 1. Graphic profiles to show the difference between ORI (red) and non-ORI (black) sequences. It contains 6 sub-graphs corresponding to six local structural parameters: (a) rise, (b) roll, (c) shift, (d) slide, (e) tilt and (f) twist. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. Comparison with published method

It has been reported that the ORI regions have lower DNA bendability and cleavage intensity than non-ORI regions [22]. DNA bendability reflects the non-parallel tendency, major or minor grove bias, and electrostatic potential energy of consecutive base pairs in a DNA sequence [50–52]. Cleavage intensity reflects the likelihood of DNA cleavage by hydroxyl radicals and provides a map of local variation in the shape of DNA backbone [37,38]. Based on the two physicochemical parameters, the *S. cerevisiae* ORIs were predicted by using SVM algorithm [22]. Thus, it is necessary to investigate whether the PseKNC-based method has a better performance than the existing method (called BC-based method) on discriminating ORI from non-ORI sequences. Due to differences in

Table 1
The predicted result using different parameters.

Methods	*Sn*	*Sp*	*Acc*	*Mcc*	auROC
Bendability + cleavage intensity	0.8123	0.8030	0.8076	0.6153	0.8563
PseKNC	0.8469	0.8276	0.8372	0.6746	0.8848

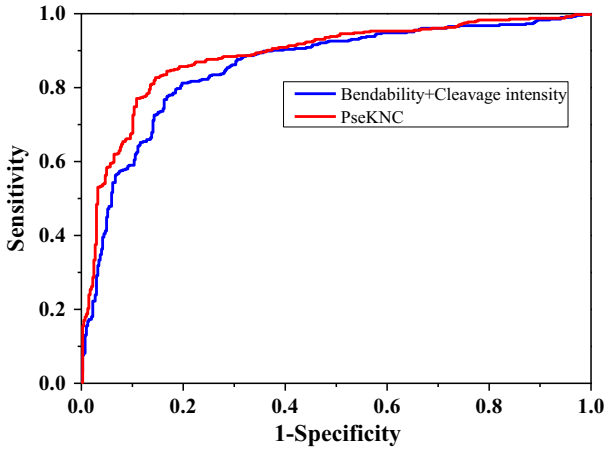

Fig. 2. Two ROC curves for PseKNC and bendability + cleavage intensity method.

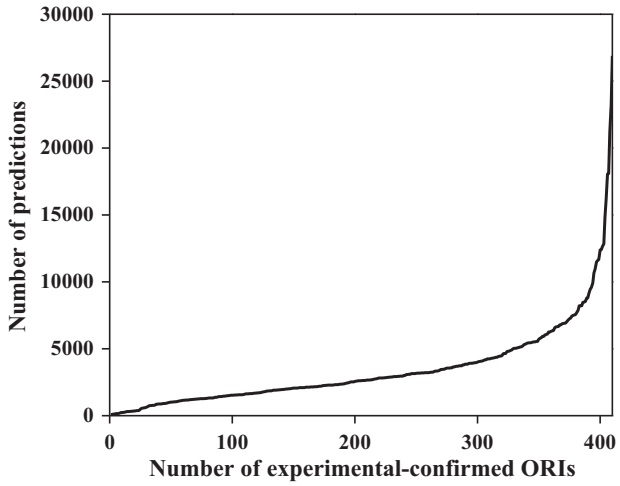

Fig. 3. The genome-scanned results using the PseKNC-based method. The abscissa denotes the number of ORIs; the ordinate denotes the number of predictions.

benchmark dataset and experimental protocol between this study and Ref. [22], we cannot provide a direct comparison with published results. Consequently, we repeated the BC-based method [22] on our benchmark dataset for investigating the performance of ORI prediction. The results in Table 1 show that the BC-based method can produce the *Acc* of 80.76% in jackknife cross-validation. It is obvious that the PseKNC-based method outperformed the BC-based method in identifying ORIs. This comparison demonstrates again that local nucleotide correlation and long-range structural properties are very important in DNA replication initiation. For evaluating the performance of the two models across the entire range of SVM decision values, two ROC curves were plotted in Fig. 2. The auROC of the PseKNC-based method is 0.8848, which is larger than that (0.8563) of the BC-based method.

3.4. Prediction of ORIs in S. cerevisiae genome

To further investigate the performance of our proposed method, according to the strategy of Ref. [53], we must use the PseKNC-based method to predict the ORIs in *S. cerevisiae* genome for comparing the computational results with experimental-confirmed ORIs. Thus, the PseKNC-based method was used to scan the 16 *S. cerevisiae* chromosomes using the window of 300 bp with the step of 1 bp. Then we obtained 1.21×10^7 subsequence, and got the prediction probability of each subsequence. For providing a detailed analysis, we firstly ranked the prediction probabilities of the 410 experimentally confirmed ORIs in the benchmark dataset in a descending order. Then, each prediction probability was selected as a cutoff denoted as cutoff[i] ($i = 1, 2, \ldots, 410$). For an arbitrary subsequence to be predicted with a probability higher than the cutoff[i], it will be regarded as a possible ORI. If this subsequence locates in the region from 200 bp upstream to 200 bp downstream of an experimentally confirmed ORI, the predicted ORI would be regarded as a true positive (TP). If the distance of two predictions is less than 300 bp, the two predictions are considered as one prediction. This process was repeated until 410 cutoffs were used. The result of the number of TP versus the number of predictions was shown in Fig. 3 and Table S1. We noticed that if the cutoff of the predicted probability is set to 0.5, 385 experimental-confirmed ORIs (Sn = 93.9%) can be correctly identified; and 8208 additional predictions were obtained. These predictions are maybe the potential ORIs.

Undoubtedly, the experimental-based methods (i.e. Ref. [53]) can accurately identify ORIs. However, it is expensive and time-consuming to perform genome-wide identification of ORIs. Our computational method can scan complete genome rapidly and find potential ORIs without many false positives. We hope that our results will provide clues to the identification of ORIs by web-experiments.

3.5. Web-server guide

Establishing a user-friendly web-server will improve the efficiency and avoid repeating a complicated mathematics and program for studying ORIs. The predictor established via aforementioned procedures is called iORI-PseKNC, where "i" stands for "identify", "ORI" for "origin of replication", "Pse" for "pseudo", "K" for "*k*-tuple", "N" for "nucleotide", and "C" for "composition". For the convenience of the vast majority of experimental scientists, we provided a guide to help experimental scientists to use the web-server to get the desired results.

Firstly, open the web server at http://lin.uestc.edu.cn/server/iORI-PseKNC and you will see the top page of iORI-PseKNC on your computer screen, as shown in Fig. 4. Click on the Read Me button to see a brief introduction about the predictor and the caveat when using it. Click on the Data button to download the benchmark datasets used to train and test the iORI-PseKNC predictor. Click on the Citation button to

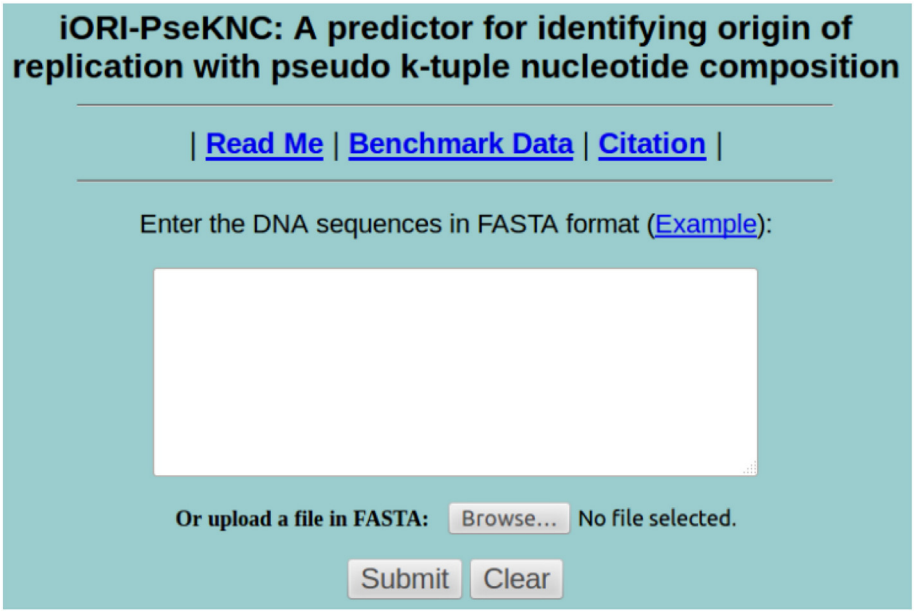

Fig. 4. A semi-screenshot for the top page of the iORI-PseKNC web-server at http://lin.uestc.edu.cn/server/iORI-PseKNC.

find the relevant papers that document the detailed development and algorithm of iORI-PseKNC. Secondly, either type or copy/paste the query DNA sequences into the input box at the center of Fig. 4. The input sequence should be in the FASTA format. Example sequences in the FASTA format can be seen by clicking on the Example button right above the input box. Thirdly, click on the Submit button to see the predicted result. It should be noted that each of the input query sequences should exclude all illegal characters: such as 'N', 'W', 'Y'. The length of input query sequences should not be less than 300 bp.

4. Conclusion

Correct identification of ORIs is the first step of understanding the replication mechanisms. The current study developed a PseKNC-based method which can incorporate the local and global sequence-order information for identifying the ORIs. The physiochemical properties were proposed to formulate the DNA sequences. Statistical analysis shows that ORI sequences are dramatically different from the non-ORI sequences in the sequence structure which may be the key feature recognized by regulatory proteins. Based on this method, a predictor called iORI-PseKNC was constructed for the convenience of the vast majority of experimental scientists. The predictor can correctly identify 84.69% ORIs in the jackknife test. Hence, we anticipated that the iORI-PseKNC will become an important tool in relevant fields.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.chemolab.2014.12.011.

Conflict of interests

The author has no conflict of interests concerning this work.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Nature Scientific Foundation of China (No. 61202256, 61301260 and 61100092), the Nature Scientific Foundation of Hebei Province (No. C2013209105), and the Fundamental Research Funds for the Central Universities (No. ZYGX2013J102).

References

[1] T.D. Halazonetis, Conservative DNA replication, Nat. Rev. Mol. Cell Biol. 15 (2014) 300.
[2] G.T. Marczynski, L. Shapiro, Bacterial chromosome origins of replication, Curr. Opin. Genet. Dev. 3 (1993) 775–782.
[3] O. Schub, G. Rohaly, R.W.P. Smith, A. Schneider, S. Dehde, I.L. Dornreiter, H.P. Nasheuer, Multiple phosphorylation sites of DNA polymerase alpha-primase cooperate to regulate the initiation of DNA replication in vitro, J. Biol. Chem. 276 (2001) 38076–38083.
[4] E. Foureau, V. Courdavault, S.M.N. Gallon, S. Besseau, A.J. Simkin, J. Creche, L. Atehortua, N. Giglioli-Guivarc'h, M. Clastre, N. Papon, Characterization of an autonomously replicating sequence in *Candida guilliermondii*, Microbiol. Res. 168 (2013) 580–588.
[5] M.K. Dhar, S. Sehgal, S. Kaul, Structure, replication efficiency and fragility of yeast ARS elements, Res. Microbiol. 163 (2012) 243–253.
[6] A. Crampton, F. Chang, D.L. Pappas, R.L. Frisch, M. Weinreich, An ARS element inhibits DNA replication through a SIR2-dependent mechanism, Mol. Cell 30 (2008) 156–166.
[7] C. Tiengwe, C.A. Marques, R. McCulloch, Nuclear DNA replication initiation in kinetoplastid parasites: new insights into an ancient process, Trends Parasitol. 30 (2014) 27–36.
[8] F. Coin, B. Reina-San-Martin, G. Giglia-Mari, M. Berneburg, DNA in 3R: repair, replication, and recombination, Mol. Biol. Int. 2012 (2012) 658579.
[9] C. Cayrou, P. Coulombe, A. Puy, S. Rialle, N. Kaplan, E. Segal, M. Mechali, New insights into replication origin characteristics in metazoans, Cell Cycle 11 (2012) 658–667.
[10] T. Valovka, M. Schonfeld, P. Raffeiner, K. Breuker, T. Dunzendorfer-Matt, M. Hartl, K. Bister, Transcriptional control of DNA replication licensing by Myc, Sci. Rep. 3 (2013) 9.
[11] M.M. Martin, M. Ryan, R. Kim, A.L. Zakas, H. Fu, C.M. Lin, W.C. Reinhold, S.R. Davis, S. Bilke, H. Liu, J.H. Doroshow, M.A. Reimers, M.S. Valenzuela, Y. Pommier, P.S. Meltzer, M.I. Aladjem, Genome-wide depletion of replication initiation events in highly transcribed regions, Genome Res. 21 (2011) 1822–1832.
[12] Y. Lubelsky, H.K. MacAlpine, D.M. MacAlpine, Genome-wide localization of replication factors, Methods 57 (2012) 187–195.
[13] J.V. Van Houten, C.S. Newlon, Mutational analysis of the consensus sequence of a replication origin from yeast chromosome III, Mol. Cell. Biol. 10 (1990) 3917–3925.
[14] M.C. Marsolier-Kergoat, Asymmetry indices for analysis and prediction of replication origins in eukaryotic genomes, PLoS ONE 7 (2012) e45050.
[15] S. Yin, W. Deng, L. Hu, X. Kong, The impact of nucleosome positioning on the organization of replication origins in eukaryotes, Biochem. Biophys. Res. Commun. 385 (2009) 363–368.
[16] M.L. Eaton, K. Galani, S. Kang, S.P. Bell, D.M. MacAlpine, Conserved nucleosome positioning defines replication origins, Genes Dev. 24 (2010) 748–753.
[17] M. Mechali, Eukaryotic DNA replication origins: many choices for appropriate answers, Nat. Rev. Mol. Cell Biol. 11 (2010) 728–738.
[18] C.A. Nieduszynski, Y. Knox, A.D. Donaldson, Genome-wide identification of replication origins in yeast by comparative genomics, Genes Dev. 20 (2006) 1874–1879.
[19] I. Brukner, R. Sanchez, D. Suck, S. Pongor, Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides, EMBO J. 14 (1995) 1812–1818.
[20] E.P. Bishop, R. Rohs, S.C. Parker, S.M. West, P. Liu, R.S. Mann, B. Honig, T.D. Tullius, A map of minor groove shape and electrostatic potential from hydroxyl radical cleavage patterns of DNA, ACS Chem. Biol. 6 (2011) 1314–1320.
[21] J.H. Kang, S.M. Kim, DNA cleavage by hydroxyl radicals generated in the Cu, Zn-superoxide dismutase and hydrogen peroxide system, Mol. Cell 7 (1997) 777–782.
[22] W. Chen, P. Feng, H. Lin, Prediction of replication origins by calculating DNA structural properties, FEBS Lett. 586 (2012) 934–938.
[23] A. Jayamani, V. Thamilarasan, V. Ganesan, N. Sengottuvelan, Structural, electrochemical, DNA binding and cleavage properties of nickel(II) complex Ni(H(2)biim)(2)(H_2O)(2) (2+) of 2,2′-biimidazole, Bull. Korean Chem. Soc. 34 (2013) 3695–3702.
[24] S.-H. Guo, E.-Z. Deng, L.-Q. Xu, H. Ding, H. Lin, W. Chen, K.-C. Chou, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics (Oxford, England) 30 (2014) 1522–1529.
[25] W. Chen, P.M. Feng, H. Lin, K.C. Chou, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e68.
[26] Y. Zuo, P. Zhang, L. Liu, T. Li, Y. Peng, G. Li, Q. Li, Sequence-specific flexibility organization of splicing flanking sequence and prediction of splice sites in the human genome, Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology, 2014.
[27] I. Jimenez-Useche, C. Yuan, The effect of DNA CpG methylation on the dynamic conformation of a nucleosome, Biophys. J. 103 (2012) 2502–2512.
[28] K. Hizume, M. Yagura, H. Araki, Concerted interaction between origin recognition complex (ORC), nucleosomes and replication origin DNA ensures stable ORC-origin binding, Genes Cells 18 (2013) 764–779.
[29] C.A. Nieduszynski, S. Hiraga, P. Ak, C.J. Benham, A.D. Donaldson, OriDB: a DNA replication origin database, Nucleic Acids Res. 35 (2007) D40–D46.
[30] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.
[31] H. Lin, Q.Z. Li, Eukaryotic and prokaryotic promoter prediction using hybrid approach, Theory Biosci. (Theor. Biowissenschaften), 130 (2011) 91–100.
[32] X. Zhou, Z. Li, Z. Dai, X. Zou, Predicting methylation status of human DNA sequences by pseudo-trinucleotide composition, Talanta 85 (2011) 1143–1147.
[33] F.B. Guo, H.Y. Ou, C.T. Zhang, ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes, Nucleic Acids Res. 31 (2003) 1780–1789.
[34] Q.Z. Li, H. Lin, The recognition and prediction of sigma70 promoters in *Escherichia coli* K-12, J. Theor. Biol. 242 (2006) 135–141.
[35] W. Chen, P.-M. Feng, H. Lin, K.-C. Chou, iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition, BioMed. Res. Int. 2014 (2014) 12.
[36] W. Chen, T.Y. Lei, D.C. Jin, H. Lin, K.C. Chou, PseKNC: a flexible web server for generating pseudo k-tuple nucleotide composition, Anal. Biochem. 456 (2014) 53–60.
[37] Y.C. Zuo, Q.Z. Li, The hidden physical codes for modulating the prokaryotic transcription initiation, Phys. A 389 (2010) 4217–4223.
[38] S. Soltani, H. Askari, N. Ejlali, R. Aghdam, The structural properties of DNA regulate gene expression, Mol. BioSyst. 10 (2014) 273–280.
[39] L.J. Jensen, A. Bateman, The rise and fall of supervised machine learning techniques, Bioinformatics 27 (2011) 3331–3332.
[40] Y.X. Zhang, An improved QSPR method based on support vector machine applying rational sample data selection and genetic algorithm-controlled training parameters optimization, Chemometr. Intell. Lab. Syst. 134 (2014) 34–46.
[41] X. Huang, D.S. Cao, Q.S. Xu, L. Shen, J.H. Huang, Y.Z. Liang, A novel tree kernel support vector machine classifier for modeling the relationship between bioactivity and molecular descriptors, Chemometr. Intell. Lab. Syst. 120 (2013) 71–76.
[42] C. Nantasenamat, K. Srungboonmee, S. Jamsak, N. Tansila, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Quantitative structure–property relationship study of spectral properties of green fluorescent protein with support vector machine, Chemometr. Intell. Lab. Syst. 120 (2013) 42–52.
[43] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[44] H. Lin, W. Chen, H. Ding, AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes, PLoS ONE 8 (2013) 6.
[45] P.M. Feng, W. Chen, H. Lin, K.C. Chou, iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition, Anal. Biochem. 442 (2013) 118–125.
[46] P.M. Feng, H. Lin, W. Chen, Identification of antioxidants from sequence information using Naive Bayes, Comput. Math. Meth. Med. 2013 (2013) ID 567529.

[47] X.Y. Zhang, C.K. Asiedu, P.C. Supakar, R. Khan, K.C. Ehrlich, M. Ehrlich, Binding sites in mammalian genes and viral gene regulatory regions recognized by methylated DNA-binding protein, Nucleic Acids Res. 18 (1990) 6253–6260.
[48] R.H. Costa, D.R. Grayson, K.G. Xanthopoulos, J.E. Darnell Jr., A liver-specific DNA-binding protein recognizes multiple nucleotide sites in regulatory regions of transthyretin, alpha 1-antitrypsin, albumin, and simian virus 40 genes, Proc. Natl. Acad. Sci. U. S. A. 85 (1988) 3840–3844.
[49] S. Cogoi, M. Paramasivam, B. Spolaore, L.E. Xodo, Structural polymorphism within a regulatory element of the human KRAS promoter: formation of G4-DNA recognized by nuclear proteins, Nucleic Acids Res. 36 (2008) 3765–3780.
[50] O.N. Ozoline, A.A. Deev, E.N. Trifonov, DNA bendability—a novel feature in *E. coli* promoter recognition, J. Biomol. Struct. Dyn. 16 (1999) 825–831.
[51] E.N. Trifonov, Base pair stacking in nucleosome DNA and bendability sequence pattern, J. Theor. Biol. 263 (2010) 337–339.
[52] I. Gabdank, D. Barash, E.N. Trifonov, Nucleosome DNA bendability matrix (*C. elegans*), J. Biomol. Struct. Dyn. 26 (2009) 403–411.
[53] C.A. Muller, M. Hawkins, R. Retkute, S. Malla, R. Wilson, M.J. Blythe, R. Nakato, M. Komata, K. Shirahige, A.P.S. de Moura, C.A. Nieduszynski, The dynamics of genome replication using deep sequencing, Nucleic Acids Res. 42 (2014) 11.