iROS-gPseKNC: Predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition

Xuan Xiao[1,2,5], Han-Xiao Ye[1], Zi Liu[3], Jian-Hua Jia[1], Kuo-Chen Chou[4,5]

[1]Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, 333403, China

[2]Information School, ZheJiang Textile and Fashion College, NingBo, 315211, China

[3]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China

[4]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, 21589, Saudi Arabia

[5]Gordon Life Science Institute, Boston, Massachusetts, 02478, USA

Correspondence to: Xuan Xiao, **email:** xxiao@gordonlifescience.org, jdzxiaoxuan@163.com

Keywords: *origin of replication, position-specific dinucleotide propensity, general pseudo nucleotide composition, random forest, iROS-gPseKNC*

Received: *March 02, 2016* ***Accepted:*** *April 09, 2016* ***Published:*** *April 27, 2016*

ABSTRACT

DNA replication, occurring in all living organisms and being the basis for biological inheritance, is the process of producing two identical replicas from one original DNA molecule. To in-depth understand such an important biological process and use it for developing new strategy against genetics diseases, the knowledge of duplication origin sites in DNA is indispensible. With the explosive growth of DNA sequences emerging in the postgenomic age, it is highly desired to develop high throughput tools to identify these regions purely based on the sequence information alone. In this paper, by incorporating the dinucleotide position-specific propensity information into the general pseudo nucleotide composition and using the random forest classifier, a new predictor called iROS-gPseKNC was proposed. Rigorously cross–validations have indicated that the proposed predictor is significantly better than the best existing method in sensitivity, specificity, overall accuracy, and stability. Furthermore, a user-friendly web-server for iROS-gPseKNC has been established at http://www.jci-bioinfo.cn/iROS-gPseKNC, by which users can easily get their desired results without the need to bother the complicated mathematics, which were presented just for the integrity of the methodology itself.

INTRODUCTION

During the cell-replicating process, the genome duplication is an indispensable step. Although the processes of DNA replications are different for bacteria, archaea, and eukaryotes, they all share the same core components as elaborated in [1–2]. For in-depth understanding the genome duplication, it is important to find the "origin of replication region" (Ori), or "replication origin" (RO) (Figure 1).

For small DNAs, such as those in bacterial plasmids and small viruses, a single origin would be sufficient to ensure a complete and opportune replication for each cell cycle in the entire genome. It is quite different, however, for eukaryotic genomes that contain substantially more origins [2–3]. Actually, it is quite natural to establish the replication forks at multiple locations [3] in order for timely duplicating their larger linear chromosomes. Therefore, to in-depth understand the process of cell reproduction, it is fundamentally important to acquire the RO information [1].

There are many experimental methods that can be used to determine the RO sites, such as chromatin immunoprecipitation (Chip), ChIp sequencing, and surface plasmon resonance (SPR). But it would take much longer time and spend more money to purely use experimental methods alone to acquire this kind of information. Therefore, it would be wise to develop computational methods to do the job, or at least as a complementary tool to the traditional experimental approach.

Actually, many scientists have endeavored to do so, as reported in a series of publications [2–12]. Unfortunately, all these reported methods have some limitations, such as in limited accuracy and practical application value. Particularly, most of these methods are without a web-server, and can hardly be used by most experimental scientists. In view of this, further work in such an important and urgent area is definitely needed.

According to Chou's five guidelines [13] and many recent publications [14–20], to develop a sequence-based statistical predictor useful not only for theoretical scientists but also broad experimental scientists, we should observe the following five guidelines and make their concrete processes crystal clear: (1) how to prepare benchmark dataset; (2) how to formulate the biological sequence samples; (3) how to operate the prediction engine; (4) how to validate the predictor's results; (5) how to provide a publically accessible web-server for the predictor. In the rest of this paper, we are to address these five aspects one-by-one. To fit in the style of the Oncotarget journal, however, their order may be subject to some sort of change.

RESULTS AND DISCUSSION

A new predictor with its web-server and user guide

A new and much more accurate sequence-based method, called iROS-gPseKNC, was developed for predicting replication origin sites in DNA. Moreover, to attract most experimental scientists and maximize their convenience [11, 21], the server of iROS-gPseKNC has been established along with its instructions, as given below.

(1) Click the web-server at http://www.jci-bioinfo.cn/iROS-gPseKNC, the top page of the iROS-gPseKNC will be prompted on your computer screen (Figure 2).

(2) Enter your query DNA sequences into the central input box (Figure 2) by using either typing or copying/pasting operation. The entered query sequences should be in the FASTA format. If you are not familiar with it, please click the Example button nearby.

(3) You can see the prediction results by clicking the Submit button. For example, if your query DNA sequences are none but those listed in the Example window, the following results will be shown on the screen: (1) DNA region 1 is the replication origin site; (2) DNA region 2 is non-replication origin site. All these outcomes were confirmed by experiments.

(4) If you have a lot of query sequences and need much longer computational time, you are also allowed to use the batch prediction. To do this, just use the Browse button to select the desired file (in FASTA format of course) and follow the online instruction.

(5) The benchmark dataset used in this study is available by clicking the button of Supporting Information on the top of Figure 2.

(6) To see the papers relevant to the development of this server, just click on the button of Citation.

Result analysis and comparison

The success scores achieved by iROS-gPseKNC on the benchmark dataset (Supporting Information S1) by the jackknife tests are given in Table 1. Shown in that table are also the corresponding scores obtained by the existing methods. It can be seen from Table 1 that iROS-gPseKNC

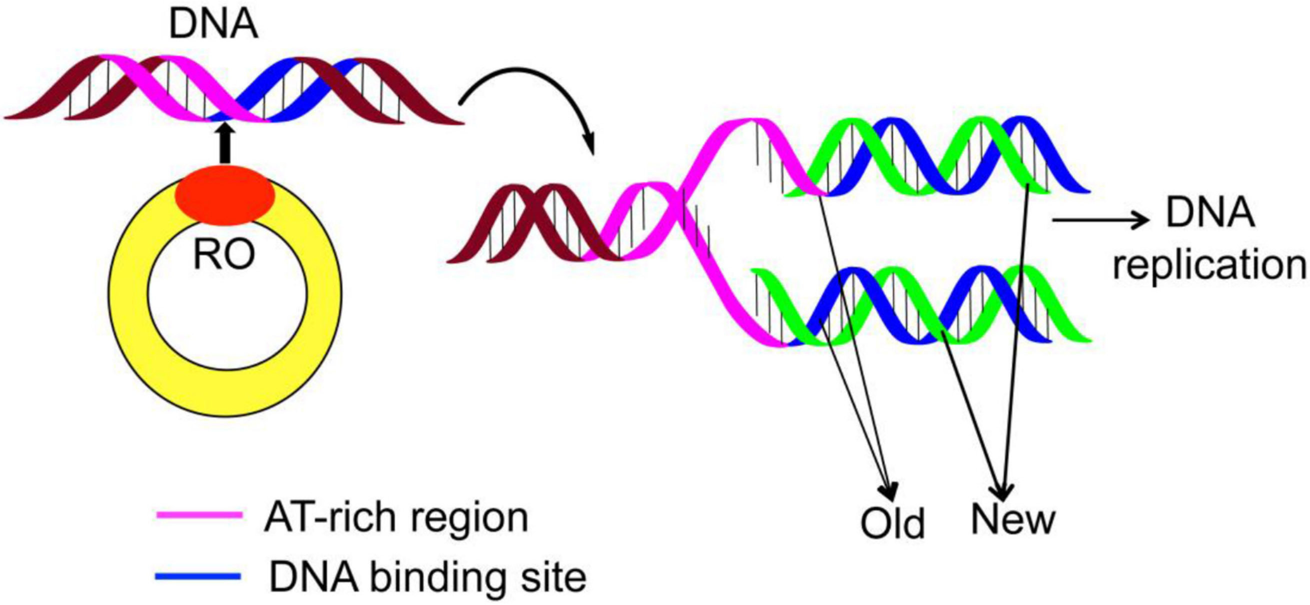


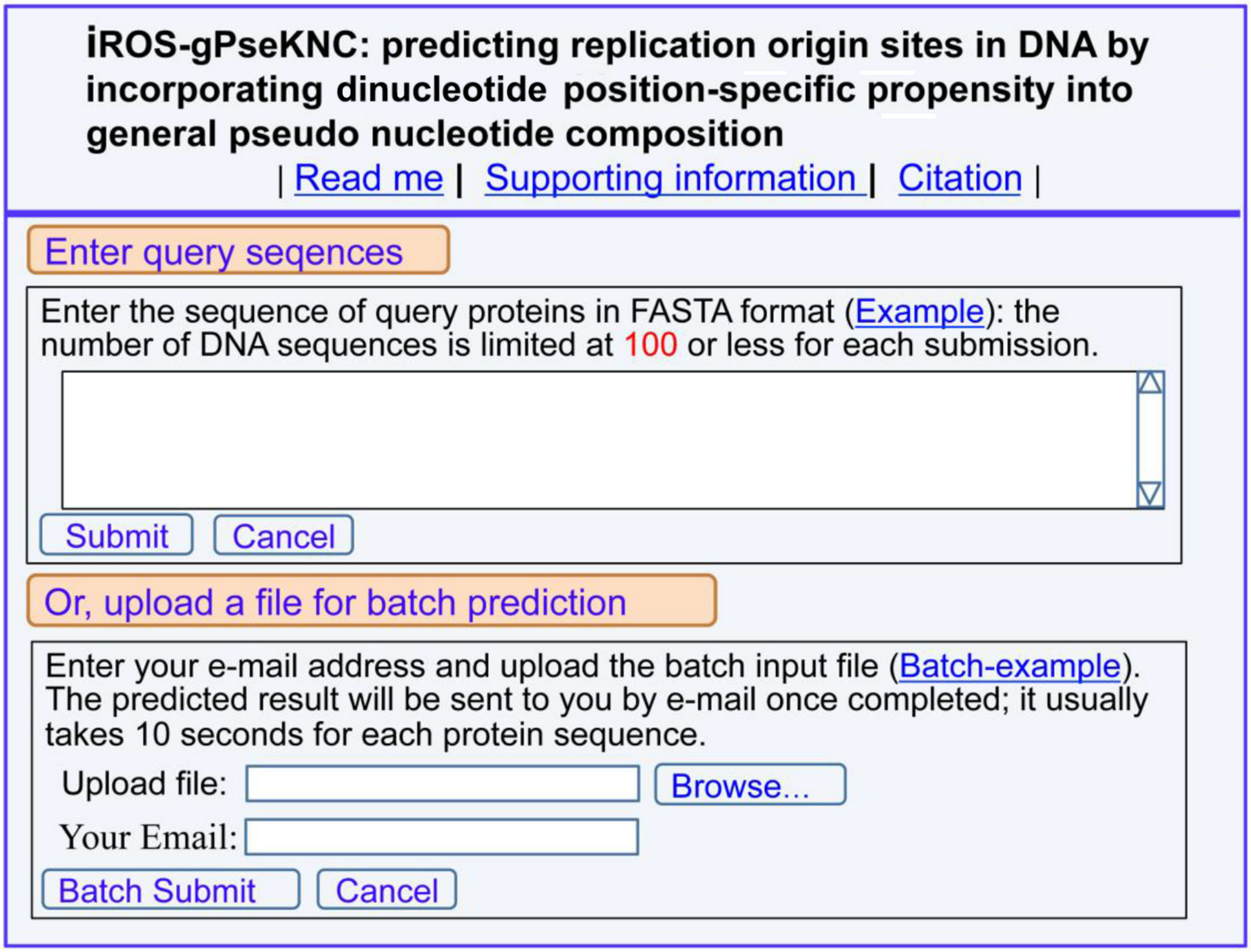
Figure 1: A schematic drawing to show the DNA replication origin (RO).

achieved remarkably higher scores than its counterparts in all the four metrics, clearly indicating that, compared with its counterparts, the proposed predictor has the highest sensitivity, specificity, overall accuracy, and stability.

Why could the proposed method yield so high success rates? It is not easy to give a simple and intuitive answer for this problem. Fortunately, many biological systems and the complicated relations therein could be revealed via the intuitive graphical approaches (see, e.g. [22–31]).

In this study, using the intuitive graphic method, we obtained various statistical distributions for different dinucleotide occurrence frequencies along the 300 bp region as shown in Figure 3, where panel (A) is for dinucleotide AA, and panel (B) for dinucleotide TT. Of course, we could draw a total of 16 such panels, but two are more than enough to make the point clear. It can be seen from Figure 3A that the AA profile for the positive samples (blue) is remarkably different from that for the negative samples (red). The same is true for the two TT profiles as shown in Figure 3B. Consequently, it is self-evident why the proposed method, which was established by including the dinucleotide position-specific propensity with the general PseKNC (see Material and Methods section), is so successful.

To provide an intuitive comparison of the proposed predictor with its counterpart, the graph of ROC (receiver operating characteristic) [32, 33] was adopted as shown in Figure 4, where the ROC curves for the iROS-gPseKNC and iORI-PseKNC [12] are in blue and red, respectively. The greater the AUC (area under the ROC curve) value is, the better the corresponding predictor will be [32, 33]. It can be easily seen from Figure 4 that the area under the blur curve is substantially greater than that under the red one, clearly indicating that the proposed predictor is no doubt superior to iORI-PseKNC [12], the best existing predictor for identifying the origins of replication in DNA sequences. Accordingly, we anticipate that iROS-gPseKNC will become a very useful computational tool for predicting DNA RO sites.



iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition

| Read me | Supporting information | Citation |
Enter query seqences

Enter the sequence of query proteins in FASTA format (Example): the number of DNA sequences is limited at 100 or less for each submission.

Submit
Cancel
Or, upload a file for batch prediction

Enter your e-mail address and upload the batch input file (Batch-example). The predicted result will be sent to you by e-mail once completed; it usually takes 10 seconds for each protein sequence.

Upload file:

Your Email:

Figure 2: A semi-screenshot for the top page of the web-server iROS-gPseKNC at http://www.jci-bioinfo.cn/iROS-gPseKNC.

Table 1: A comparison of the proposed predictor with the existing methods via the jackknife tests on a same benchmark dataset of Supporting Information S1

Predictor	Sn (%)[d]	Sp (%)[d]	Acc (%)[d]	MCC[d]
BC-based[a]	81.23	80.30	80.76	61.53
iORI-PseKNC[b]	84.69	82.76	83.72	67.46
iROS-gPseKNC[c]	**96.42**	**99.74**	**98.03**	**96.11**

[a] The prediction method developed by Chen [4].
[b] The prediction method developed by Li et al. [12]} that was deemed the most powerful one among the existing methods for the same purpose.
[c] The prediction method proposed in this paper.
[d] See Eq.7 for the definition of the metrics.

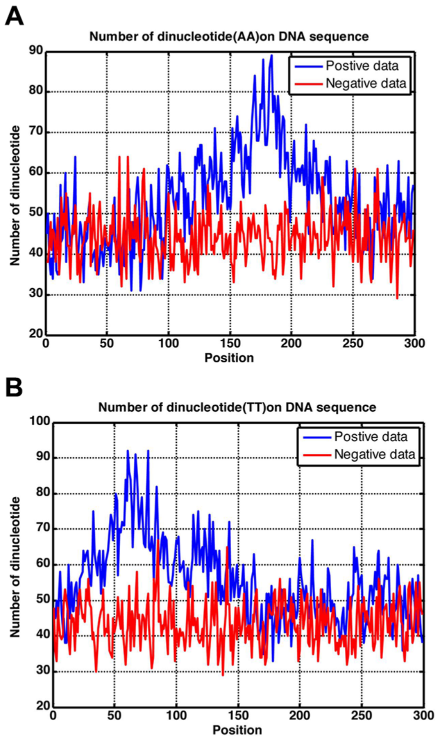


Figure 3: Graph to show the statistical distribution of the dinucleotide occurrence frequency for (**A**) AA and (**B**) TT along the 300 bp region. See the text for further explanation.

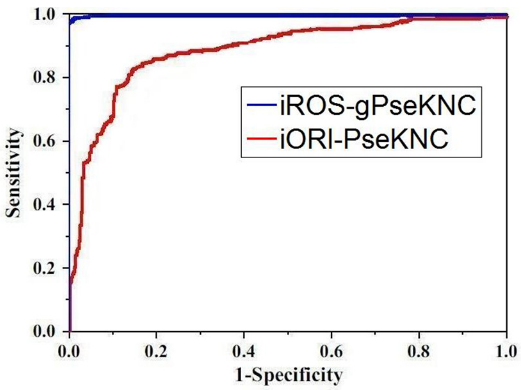


Figure 4: Graph to show the ROC curve [32, 33]. The one with red is for iORI-PseKNC predictor [12]}; while the one with blue is for the proposed predictor iROS-gPseKNC. The area under the blue curve is remarkably larger than that under the red curve. See the text for further explanation.

MATERIALS AND METHODS

Benchmark dataset

In this study, we used the same dataset recently constructed by Li et al. [12] that was specialized for studying the replication origin sites. The reasons are as follows. (1) The dataset was constructed rigorously based on experiment-confirmed reports only, and hence is more reliable. (2) None of samples included had pairwise sequence identity to any other, and hence the dataset is more stringent in excluding homology bias than the other relevant ones. (3) Most important, it will facilitate the comparison of our new prediction method with the existing ones since a fair comparison should be based on a same benchmark dataset and same cross-validation approach.

In literature, the benchmark dataset usually consists of a training dataset and a testing dataset: the former is constructed for the purpose of training a proposed model, while the latter for the purpose of testing it. As pointed out by a comprehensive review [34], however, there is no need to separate a benchmark dataset into a training dataset and a testing dataset for validating a prediction method if it is tested by the jackknife or subsampling (K-fold) cross-validation because the outcome thus obtained is actually from a combination of many different independent dataset tests. Thus, the benchmark dataset taken from Li et al. [12] for the current study can be formulated as

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

where the positive subset $\mathbb{S}^{+}$ contains 405 replication origin samples, the negative subset $\mathbb{S}^{-}$ contains 406 non-replication origin samples, and the symbol $\cup$ denotes the union in the set theory. The 405 + 406 = 811 DNA samples

are each consist of 300 bp [12], as can be generally formulated by

$$\mathrm{D} = \mathrm{N}_1\mathrm{N}_2\mathrm{N}_3\cdots\mathrm{N}_i\cdots\mathrm{N}_{300} \tag{2}$$

For readers' convenience, their sequences are given in Supporting Information S1.

Feature vector construction

Biology is a natural science with historic dimension. All biological species have developed beginning from a very limited number of ancestral species. It is true for the biological sequences as well. Their evolution involves changes of single amino or nucleic acid residues, insertions and deletions of several residues, gene doubling, and gene fusion. With these changes accumulated for a long period of time, many apparent similarities between the initial and resultant biological sequences have been gradually disappearing, but the corresponding sequences may still share some essential common features. That is why the 3D (three-dimensional) structure of a protein derived from the template [35] of a remote homologous protein [36] is often quite successful although their sequence similarity may not be high [37, 38]. Also, it has been reported that the bacterial replication origins share similar nucleotide sequence motifs. Therefore, the key is how to "unearth" this kind of motifs deeply "buried" in extremely complicated DNA sequences.

Actually, with the avalanche of biological sequences generated in the post-genomic age, one of the most challenging problems in computational biology is how to formulate a biological sequence with a discrete model or vector, yet still considerably keep its sequence pattern or order information. This is because almost all the existing machine-learning algorithms were developed to handle vector but not sequence samples, as elaborated in [21]. But a vector defined in a discrete model may completely lose this kind of sequence-pattern information. To overcome this problem, the "pseudo amino acid composition" [39] or Chou's PseAAC [40, 41] was developed to deal with protein/peptide sequences. Encouraged by its successes in computational proteomics, the idea of PseAAC was recently extended to dealing with DNA/RNA sequences in many important problems of genome analysis [12, 16, 18, 42–47] by introducing the pseudo nucleotide composition or PseKNC [9, 10, 14, 48, 49].

According to a recent review paper [11], the general form of PseKNC for a DNA sequence can be formulated as

$$\mathbf{D} = [\phi_1 \;\; \phi_2 \cdots \phi_u \cdots \phi_z]^{\mathbf{T}} \tag{3}$$

where T is the transpose operator, while Z an integer to reflect the vector's dimension. The value of Z as well as the components ϕ_u $(u = 1, 2, ..., z)$ in Eq.3 will depend on how to extract the desired information from the DNA sequence.

Recently, by incorporating the dipeptide position-specific propensity into the general PseAAC [13], Xu et al. developed two predictors for identifying posttranslational modification (PTM) sites for proteins: one for cysteine S-nitrosylation sites [50], and the other for hydroxyproline and hydroxylysine sites [51]. Stimulating by their approach, here we are to develop a new method for predicting the replication origin sites by incorporating the dinucleotide position-specific propensity into the general PseKNC [11] or Eq.3.

There are $4^2 = 16$ dinucleotides: AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, and TT. Thus, for a DNA sample with 300 bp (Eq.2) as given in Supporting Information S1, its profile (or detailed information) of the dinucleotide position-specific propensity can be summarized by the following 16 × 299 matrix:

$$\mathbb{D} = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,299} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,299} \\ \vdots & \vdots & \cdots & \vdots \\ P_{16,1} & P_{16,2} & \cdots & P_{16,299} \end{bmatrix} \tag{4}$$

where

$$P_{i,j} = Q^{+}\left(2\text{mer}_i|j\right) - Q^{-}\left(2\text{mer}_i|j\right)\ \ (i = 1,\ 2, \cdots, 16) \\ (j = 1,\ 2, \cdots, 299) \tag{5}$$

In the above equation, 2mer_1 = AA, 2mer_2 = AC, 2mer_3 = AG, 2mer_4 = AT, 2mer_{15} = TG, 2mer_{16} = TT, and $Q^+ (2\text{mer}_i\,|j)$ is the occurrence frequency of the *i*-th dinucleotide (2mer_i) at the *j*-th subsite on the sequence of Eq.2 that can be easily derived from the positive dataset $\mathbb{S}^+$, while $Q^- (2\text{mer}_i\,|j)$ is the corresponding occurrence frequency, but from the negative dataset $\mathbb{S}^-$.

Thus, the DNA sample of Eq.2 can be uniquely defined via the general form of PseKNC (cf. Eq.3) with its dimension $Z = 299$ and its *u*-th component given by

$$\phi_u = \begin{cases} P_{1,u} & \text{when } \mathrm{N}_u\mathrm{N}_{u+1} = \mathrm{AA} \\ P_{2,u} & \text{when } \mathrm{N}_u\mathrm{N}_{u+1} = \mathrm{AC} \\ P_{3,u} & \text{when } \mathrm{N}_u\mathrm{N}_{u+1} = \mathrm{AG}\ (1 \le u \le 299) \\ \vdots & \qquad\vdots \\ P_{16,u} & \text{when } \mathrm{N}_u\mathrm{N}_{u+1} = \mathrm{TT} \end{cases} \tag{6}$$

Random forest classifier

The random forests (RF) algorithm is a powerful algorithm and has been used in many areas of computational biology (see, e.g. [52–56]). The essence of BF is to randomly generate many trees by the recursive partitioning approach, followed by aggregating the results. Its detailed procedures and formulation have been very clearly described in [57], and hence there is no need to repeat here.

After training by the relevant benchmark dataset, the RF classifier can quickly indicate which attribute an input query sample belongs to. For the current study, the input are DNA sequences, while the output are which of them belong to the replication origins and which of them do not.

The predictor obtained via the aforementioned procedures is called iROS-gPseKNC, where "i" stands for "identify", "ROS" for "replication origin site", and "gPseKNC" for "general PseKNC" approach.

As pointed out in the beginning of this paper, in developing a new predictor it is very important to clearly report how to evaluate its anticipated success rates [13]. To realize this, let us consider the following two things: one is what metrics we should use to quantitatively measure the predictor's quality; the other is what kind of test approach we should adopt to calculate the metrics rates.

A set of four metrics for measuring prediction quality

In statistical prediction, four metrics were often used to measure the quality of a predictor; they are: (1) overall accuracy or Acc; (2) Mathew's correlation coefficient or MCC; (3) sensitivity or Sn; and (4) specificity or Sp [58]. But their conventional formulations are not quite intuitive, and most experimental scientists feel difficult to understand them, particularly for the MCC metrics. Fortunately, if using the formulation introduced by Chou [59] in studying the signal peptides, the set of four metrics can be equivalently defined as follows [60, 61]:

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = \wedge = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+})\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-})\right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases} \tag{7}$$

where N^+ stands for the total number of replication origin samples investigated, whereas N_-^+ for the number of replication origin samples incorrectly predicted to be of non-replication origin; N^- for the total number of non-replication origin samples investigated, whereas N_+^- for the number of non-replication origin samples incorrectly predicted to be of replication origin. With such formulation as given in Eq.7, the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient and their rate scopes would become more intuitive and easier-to-understand, particularly for the Mathew's correlation coefficient, as concurred by many investigators in their recent publications [20, 55, 56, 60, 62–72]}[16, 20].

It is instructive to point out, however, the set of metrics in Eq.7 is valid only for the single-label systems. For the multi-label systems as emerging increasingly frequent in system biology [73–75] and system medicine [76], a completely different set of metrics is needed as elucidated in [77].

Cross validation

With a set of well-defined metrics to measure the quality of a predictor, the next thing is what kind of validation method should be used to score these metrics.

In predictive analytics, the following three cross-validation methods are often used: (1) independent dataset test, (2) subsampling (or K-fold cross-validation) test, and (3) jackknife test [78]. Of these three, however, the jackknife test is deemed the least arbitrary that can always yield a unique outcome for a given benchmark dataset as elucidated in [13]. Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (see, e.g., [79, 80] [81–84]). Therefore, the jackknife test was also adopted in this study to score the metrics of Eq.7. In the jackknife test, each of the samples in the benchmark dataset is singled out one-by-one and tested by the predictor trained by the remaining samples. During the jackknifing process, both the training dataset and testing dataset are literally open, and each sample is in turn moved between the two. The jackknife test can exclude the “memory” effect; it can also avoid the arbitrariness problem occurring in the independent dataset test and subsampling test as pointed out in [13] because the outcome obtained by the jackknife test is always unique for a given benchmark dataset.

CONCLUSIONS

DNA replication is one of the most important life processes at the cellular level. To really understand such vitally important biological process, the knowledge of duplication origin sites is fundamentally important. The iROS-gPseKNC predictor presented in this paper can be used to identify the duplication origin sites based on the DNA sequence information alone. Its accuracy is better than the best existing predictor in this area. By running the iROS-gPseKNC web-server according to its step-by-step guide, users can easily obtain their desired results without the need to go through the detailed mathematics, which were presented in this paper just for its integrity.

Although the new predictor can yield significantly higher success rates than the existing ones, there still are plenty rooms to further improve it from the following two angles. One is with the increase of experimental data available in future, the dataset used to train the current model can be further refined and its coverage scope being much wider, and hence the predictor will be even more powerful. The other one is that many studies [80, 85–94] have indicated a predictor formed by fusing an array of individual classifiers may significantly enhance the prediction power; we will try to develop an ensemble predictor in this regard by fusing an array of individual classifiers with each being based on different modes of PseAAC [13, 39, 95, 96].

SUPPORTING INFORMATION

Supporting Information S1. The original benchmark dataset. It contains 811 DNA segments, of which 405 are ORIs or positive samples, and 406 are non-ORIs or negative samples, where the benchmark dataset was taken from Li et al. [12]. Each segment sample contains 300 nucleotide residues. None of the samples include here is identical to any other. See the main paper for further explanation.

ACKNOWLEDGMENTS AND FUNDING

The authors wish to thank the three anonymous reviewers for their constructive comments, which were very useful for strengthening the presentation of this study. This work was partially supported by the National Nature Science Foundation of China (No.31260273, 31560316,61261027), Natural Science Foundation of Jiangxi Province, China (No 20122BAB211033, 20122BAB201044) the Department of Education of JiangXi Province (GJJ12490), the LuoDi plan of the Department of Education of JiangXi Province(KJLD12083), and the JiangXi Provincial Foundation for Leaders of Disciplines in Science (20113BCB22008) and the Graduate Innovation Fund of Jingdezhen ceramic institute (JYC201522).

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

1. Song C, Zhang S, Huang H. Choosing a suitable method for the identification of replication origins in microbial genomes. Frontiers in microbiology. 2015; 6:1049.
2. Zakrzewska-Czerwinska J, Jakimowicz D, Zawilak-Pawlik A, Messer W. Regulation of the initiation of chromosomal replication in bacteria. FEMS Microbiol Rev. 2007; 31: 378–387.
3. Breier AM, Chatterji S, Cozzarelli NR. Prediction of Saccharomyces cerevisiae replication origins. Genome Biology. 2004; 5:60–60.
4. Chen W, Feng P, Lin H. Prediction of replication origins by calculating DNA structural properties. Febs Letters. 2012; 586:934–938.
5. Brukner I, Sánchez R, Suck D, Pongor S. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. Embo Journal. 1995; 14:1812–1818.
6. Kang JH, Kim SM. DNA cleavage by hydroxyl radicals generated in the Cu,Zn-superoxide dismutase and hydrogen peroxide system. Molecules & Cells. 1998; 7:777–782.
7. Bishop EP, Rohs R, Parker SC, West SM, Liu P, Mann RS, Honig B, Tullius TD. A Map of Minor Groove Shape and Electrostatic Potential from Hydroxyl Radical Cleavage Patterns of DNA. Acs Chemical Biology. 2011; 6:1314–1320.
8. Marsolier-Kergoat MC. Asymmetry Indices for Analysis and Prediction of Replication Origins in Eukaryotic Genomes. Plos One. 2012; 7:e45050–e45050.
9. Chen W, Lei TY, Jin DC, Lin H. PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. Anal Biochem. 2014; 456:53–60.
10. Chen W, Zhang X, Brooker J, Lin H. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics. 2015; 31:119–120.
11. Chen W, Lin H, Chou KC. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol BioSyst. 2015; 11:2620–2634.
12. Li W-C, Deng E-Z, Ding H, Chen W, Lin H. iORI-PseKNC: A predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. Chemometrics and Intelligent Laboratory Systems. 2015; 141:100–106.

13. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). J Theor Biol. 2011; 273:236–247.

14. Liu B, Liu F, Fang L, Wang X. repRNA: a web server for generating various feature vectors of RNA sequences. Molecular Genetics and Genomics. 2016; 291:473–481.

15. Jia J, Liu Z, Xiao X, Liu B. iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets. Molecules. 2016; 21:95.

16. Liu B, Fang L, Long R. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition Bioinformatics. 2016; 32:362–389.

17. Jia J, Liu Z, Xiao X, Liu B. iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset Anal Biochem. 2016; 497: 48–56.

18. Liu Z, Xiao X, Yu DJ, Jia J. pRNAm-PC: Predicting N-methyladenosine sites in RNA sequences via physical-chemical properties Anal Biochem. 2016; 497:60–67.

19. Jia J, Liu Z, Xiao X, Liu B. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach J Theor Biol. 2016; 394:223–230.

20. Chen W, Ding H, Feng P, Lin H, Chou KC. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 2016. doi: 10.18632/oncotarget.7815.

21. Chou KC. Impacts of bioinformatics to medicinal chemistry. Medicinal Chemistry. 2015; 11:218–234.

22. Jiang SP, Liu WM, Fee CH. Graph theory of enzyme kinetics: 1. Steady-state reaction system. Scientia Sinica. 1979; 22:341–358.

23. Forsen S. Graphical rules for enzyme-catalyzed rate laws. Biochem J. 1980; 187:829–835.

24. Zhou GP, Deng MH. An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. Biochem J. 1984; 222:169–176.

25. Chou KC. Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophysical Chemistry. 1990; 35:1–24.

26. Althaus IW, Gonzales AJ, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J Biol Chem. 1993; 268:14875–14880.

27. Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Aristoff PA, Tarpley WG, Reusser F. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry. 1993; 32:6548–6554.

28. Chou KC. Graphic rule for drug metabolism systems. Current Drug Metabolism. 2010; 11:369–378.

29. Wu ZC, Xiao X. 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J Theor Biol. 2010; 267:29–34.

30. Lin WZ, Xiao X. Wenxiang: a web-server for drawing wenxiang diagrams Natural Science. 2011; 3:862–865.

31. Zhou GP. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. J Theor Biol. 2011; 284:142–148.

32. Fawcett JA. An Introduction to ROC Analysis. Pattern Recognition Letters. 2005; 27:861–874.

33. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning: ACM. 2006; 233–240.

34. Chou KC, Shen HB. Review: Recent progresses in protein subcellular location prediction. Anal Biochem. 2007; 370: 1–16.

35. Chou KC. Review: Structural bioinformatics and its impact to biomedical science. Current Medicinal Chemistry. 2004; 11:2105–2134.

36. Liu B, Zhang D, Xu R, Xu J, Wang X. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics. 2014; 30:472–479.

37. Watenpaugh KD, Heinrikson RL. A Model of the complex between cyclin-dependent kinase 5 (Cdk5) and the activation domain of neuronal Cdk5 activator. Biochemical & Biophysical Research Communications. 1999; 259:420–428.

38. Howe WJ. Prediction of the tertiary structure of the beta-secretase zymogen. Biochem Biophys Res Commun. 2002; 292:702–708.

39. Chou KC. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins (Erratum: 2001; 44:60). 2001; 43:246–255.

40. Cao DS, Xu QS, Liang YZ. propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics. 2013; 29:960–962.

41. Du P, Gu S, Jiao Y. PseAAC-General: Fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. International Journal of Molecular Sciences. 2014; 15:3495–3506.

42. Chen W, Feng PM, Deng EZ. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal Biochem. 2014; 462:76–83.

43. Chen W, Feng PM, Lin H. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. Biomed Research International. 2014; 2014:623149.

44. Guo SH, Deng EZ, Xu LQ, Ding H. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics. 2014; 30:1522–1529.

45. Lin H, Deng EZ, Ding H. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 2014; 42:12961–12972.

46. Qiu WR, Xiao X. iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and

pseudo amino acid components. Int J Mol Sci. 2014; 15:1746–1766.

47. Chen W, Feng P, Ding H. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition Analytical Biochemistry (also, Data in Brief, 2015, 5: 376–378). 2015; 490:26–33.

48. Liu B, Liu F, Fang L, Wang X. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics. 2015; 31:1307–1309.

49. Liu B, Liu F, Wang X, Chen J. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences Nucleic Acids Res. 2015; 43:W65–W71.

50. Xu Y, Shao XJ, Wu LY, Deng NY. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ. 2013; 1:e171.

51. Xu Y, Wen X, Shao XJ. iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. Int J Mol Sci. 2014; 15:7594–7610.

52. Lin WZ, Fang JA, Xiao X. iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. PLoS ONE. 2011; 6:e24756.

53. Kandaswamy KK, Moller S, Suganthan PN, Sridharan S, Pugalenthi G. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. J Theor Biol. 2011; 270:56–62.

54. Pugalenthi G, Kandaswamy KK, Kolatkar P. RSARF: Prediction of Residue Solvent Accessibility from Protein Sequence Using Random Forest Method. Protein & Peptide Letters. 2012; 19:50–56.

55. Jia J, Liu Z, Xiao X. iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J Theor Biol. 2015; 377:47–56.

56. Jia J, Liu Z, Xiao X, Liu B. Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. Journal of Biomolecular Structure & Dynamics, 2015; doi:10.1080/07391102.07392015.1095116.

57. Breiman L. Random forests. Machine learning. 2001; 45:5–32.

58. Chen J, Liu H, Yang J. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids. 2007; 33:423–428.

59. Chou KC. Prediction of signal peptides using scaled window. Peptides. 2001; 22:1973–1979.

60. Xu Y, Ding J, Wu LY. iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition PLoS ONE. 2013; 8:e55844.

61. Chen W, Feng PM, Lin H. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition Nucleic Acids Res. 2013; 41:e68.

62. Feng PM, Chen W, Lin H. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Anal Biochem. 2013; 442:118–125.

63. Xiao X, Min JL, Wang P. iGPCR-Drug: A web server for predicting interaction between GPCRs and drugs in cellular networking. PLoS ONE. 2013; 8:e72234.

64. Xiao X, Min JL, Wang P. iCDI-PseFpt: Identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints. J Theor Biol. 2013; 337C: 71–79.

65. Min JL, Xiao X. iEzy-Drug: A web server for identifying the interaction between enzymes and drugs in cellular networking. BioMed Research International. 2013; 2013:701317.

66. Ding H, Deng EZ, Yuan LF, Liu L. iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed Research International. 2014; 2014:286419.

67. Fan YN, Xiao X, Min JL. iNR-Drug: Predicting the interaction of drugs with nuclear receptors in cellular networking. Intenational Journal of Molecular Sciences. 2014; 15:4915–4937.

68. Qiu WR, Xiao X, Lin WZ. iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach. Biomed Res Int. 2014; 2014:947416.

69. Xu Y, Wen X, Wen LS, Wu LY. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS ONE. 2014; 9:e105018.

70. Qiu WR, Xiao X, Lin WZ. iUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model Journal of Biomolecular Structure and Dynamics. 2015; 33:1731–1742.

71. Xiao X, Min JL, Lin WZ, Liu Z. iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. Journal of Biomolecular Structure & Dynamics. 2015; 33:2221–2233.

72. Xu R, Zhou J, Liu B, He YA. Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach. Journal of Biomolecular Structure & Dynamics. 2015; 33:1720–1730.

73. Chou KC, Wu ZC, Xiao X. iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Molecular Biosystems. 2012; 8:629–641.

74. Lin WZ, Fang JA, Xiao X. iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins Molecular BioSystems. 2013; 9: 634–644.

75. Wu ZC, Xiao X. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. Molecular BioSystems. 2011; 7:3287–3297.

76. Xiao X, Wang P, Lin WZ, Jia JH. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem. 2013; 436:168–177.

77. Chou KC. Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. Molecular Biosystems. 2013; 9:1092–1100.

78. Chou KC, Zhang CT. Review: Prediction of protein structural classes. Crit Rev Biochem Mol Biol. 1995; 30:275–349.

79. Chou KC, Cai YD. Prediction of membrane protein types by incorporating amphipathic effects. Journal of Chemical Information and Modeling. 2005; 45:407–413.

80. Shen HB. Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers. 2007; 85:233–240.

81. Ali F, Hayat M. Classification of membrane protein types using Voting Feature Interval in combination with Chou's Pseudo Amino Acid Composition. J Theor Biol. 2015; 384:78–83.

82. Khan ZU, Hayat M, Khan MA. Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. J Theor Biol. 2015; 365:197–203.

83. Kumar R, Srivastava A, Kumari B, Kumar M. Prediction of beta-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. J Theor Biol. 2015; 365:96–103.

84. Kabir M, Hayat M. iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples. MGG. 2016; 291:285–296.

85. Chou KC, Shen HB. Predicting protein subcellular location by fusing multiple classifiers. J Cell Biochem. 2006; 99:517–527.

86. Chou KC, Shen HB. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. Journal of Proteome Research. 2006; 5:1888–1897.

87. Chou KC, Shen HB. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Comm. 2007; 357:633–640.

88. Liu DQ, Liu H, Shen HB, Yang J. Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments. Amino Acids. 2007; 32: 493–496.

89. Shen HB. Nuc-PLoc: A new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Engineering, Design & Selection. 2007; 20:561–567.

90. Chou KC, Shen HB. ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochem Biophys Res Comm. 2008; 376:321–325.

91. Shen HB. QuatIdent: A web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. Journal of Proteome Research. 2009; 8:1577–1584.

92. Shen HB, Song JN. Prediction of protein folding rates from primary sequence by fusing multiple sequential features Journal of Biomedical Science and Engineering. 2009; 2:136–143.

93. Chou KC, Shen HB. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. Journal of Proteome Research. 2007; 6:1728–1734.

94. Shen HB. Virus-mPLoc: A Fusion Classifier for Viral Protein Subcellular Location Prediction by Incorporating Multiple Sites. J Biomol Struct Dyn. 2010; 28:175–186.

95. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics. 2005; 21:10–19.

96. Chou KC. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Current Proteomics. 2009; 6:262–274.