sefOri: selecting the best engineered sequence features to predict DNA replication origins

Chenwei Lou[1,#], Jian Zhao[1,#], Ruoyao Shi[2], Qian Wang[1], Wenyang Zhou[1], Yubo Wang[1], Guoqing Wang[3,*], Lan Huang[1], Xin Feng[1], Fengfeng Zhou[1,*].

1 BioKnow Health Informatics Lab, College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin, China, 130012.

2 BioKnow Health Informatics Lab, College of Life Sciences, Jilin University, Changchun, Jilin, China, 130012.

3 Department of Pathogenobiology, The Key Laboratory of Zoonosis, Chinese Ministry of Education, College of Basic Medicine, Jilin University, Changchun, China, 130021.

[#] These authors contributed equally to this study.

* Correspondence may be addressed to Fengfeng Zhou: FengfengZhou@gmail.com or ffzhou@jlu.edu.cn . Lab web site: http://www.healthinformaticslab.org/ . Additional correspondence may also be addressed to Guoqing Wang: qing@jlu.edu.cn

Feature engineering from DNA sequences

The pseudo k-tuple nucleotide composition (PseKNC) (Chen, et al., 2014) was widely used to describe the sequence level characteristics of DNA or RNA elements and have been successfully utilized for predicting various genetic elements, e.g., genomic replication origins (Liu, et al., 2018), enhancers (Liu, et al., 2018), and N(6)-methyladenosine sites (Chen, et al., 2015), etc. PseKNC calculated the frequencies of the oligonucleotide components and the inter-nucleotide physicochemical properties (Chen, et al., 2014). So PseKNC described both short-range and long-range information of a given DNA sequence and its successful applications in predicting various genetic elements suggested that PseKNC may have captured the inherent properties of a genetic element.

This study utilized the parallel correlation PseKNC features, which was the type I PseKNC (Chen, et al., 2014). The type I PseKNC calculated (4t+r) features from a given DNA sequence, as defined in the following formula:

$$\begin{cases} \theta_1 = \frac{1}{L-t}\sum_{i=1}^{L-t} C_{i,i+1} \\ \theta_2 = \frac{1}{L-t-1}\sum_{i=1}^{L-t-1} C_{i,i+2} \\ \theta_3 = \frac{1}{L-t-2}\sum_{i=1}^{L-t-2} C_{i,i+3} \\ \quad \dots\dots\dots\dots\dots \\ \theta_r = \frac{1}{L-t-r+1}\sum_{i=1}^{L-t-r+1} C_{i,i+r} \end{cases} \quad r < L - t \qquad (1)$$

The variable L is the length of the given DNA sequence. The two parameters of the type I PseKNC were the tuple size t and the correlation rank r. The xth tier correlation factor θx was defined as the sequence order correlation between all the xth most contiguous t-tuple nucleotides. The correlation rank r describes the correlation rank (or tier), so r is smaller than (L-t). The detailed illustration may be found in (Chen, et al., 2014).

The correlation function is defined as:

$$C_{i,i+j} = \frac{1}{\Delta}\sum_{\xi=1}^{\Delta}[H_\xi(R_i R_{i+1} \dots R_{i+t-1}) - H_\xi(R_{i+j} R_{i+j+1} \dots R_{i+j+t-1})]^2 \, (2)$$

where i∈{1, 2, …, L-t+1}, j∈{1, 2, …, r}, and r<L-t. Ri is the nucleotide in the position i, and Ri∈{A, C, G, T}. Hξ(RiRi+1…Ri+t-1) is the ξth physiochemical property for the t-tuple RiRi+1…Ri+t-1 in the given DNA sequence. And Hξ(Ri+jRi+j+1…Ri+j+t-1) is that of the next jth t-tuple. The total number of the correlation functions is defined as Δ. After the standard

normalization procedure (Chen, et al., 2014), a feature vector of length (4t+r) may be calculated as:

$$D=[d_1, d_2, \dots, d_{4^t}, d_{4^{t+1}}, \dots, d_{4^t+r}]^{\mathrm{T}}, (r<L\text{-}t) \quad (3)$$

Each feature dm was defined as:

$$d_m = \begin{cases} \dfrac{f_m^{t-\text{tuple}}}{\sum_{i=1}^{4^t} f_i^{t-\text{tuple}} + a\sum_{j=1}^{r}\theta_j}, (1 \le m \le 4^t) \\ \dfrac{a\theta_{m-4^t}}{\sum_{i=1}^{4^t} f_i^{t-\text{tuple}} + a\sum_{j=1}^{r}\mathrm{\theta_j}}, (4^t + 1 \le m \le 4^t + r\ ) \end{cases} \quad (4)$$

where $f_i^{t-\text{tuple}}$ is the normalized occurrence frequency of the ith nucleotide t-tuple in the given DNA sequence, and the weight a may tune how much the pseudo nucleotide component contributes to the overall features.

In addition to the above sample formulation process, this study also calculated the probabilities of converting from one nucleotide to another. For example, the probability of the A to T conversion was calculated as:

$$P_{\text{A to T}} = \frac{count(\text{AT})}{count(\text{A})} \quad (5)$$

where count(AT) is the number of the subsequence “AT”, and count(A) is the number of the subsequence “A”. So we got 16 more features in this way.

Our final sample formulation consisted of six groups of features from PseKNC(t=3) with different values of r as well as the probabilities of nucleotide conversion. The total number of features was 550 for a given DNA sequence, as shown in Supplementary Table S1.

Supplementary Table S1

Seven groups of features calculated from a given DNA sequence.

Name	PseKNC, m=0	PseKNC, m=10	PseKNC, m=20	PseKNC,m=30
FeatureNum	64	74	84	94
Name	PseKNC, m=40	PseKNC, m=50	NU conversion	Total
FeatureNum	64	74	84	94

The group of features “PseKNC,m=i” was the PseKNC features calculated for t=3 and m=i. Row “FeatureNum” gave the number of features calculated for each group of features .

DNA physicochemical properties

DNA physicochemical properties were essential functional elements in various biomolecular processes, e.g., protein-DNA interactions (Lyubchenko, et al., 2009; Rachofsky, et al., 2001), transcriptional regulation (Ponomarenko, et al., 1999), and nucleosome occupancy (Chen, et al., 2012), etc. Moreover, the relatively constrained DNA physicochemical properties have been proved to correlate with functional noncoding regions like enhancers. This study utilized two types of DNA physicochemical properties to predict the DNA replication origins, i.e., MW-daltons (MW) and Nucleosome (NU). These features were calculated using the C++ programming language based on the .NET framework and Python version 3.6.

Selecting features to improve the predictions

Feature selection algorithms may not only improve the model prediction accuracy (Chatterjee, et al., 2018; Deshpande, et al., 2019), but also find the biologically essential genes for a better understanding of the investigated biological process (Guo, et al., 2014). This study calculated 550 features from each given DNA sequence. In order to avoid the overfitting problem, the feature selection step was utilized in order to eliminate the redundant or weakly correlated features. Four feature selection algorithms were evaluated for their feature screening capabilities on predicting DNA replication origins, i.e., chi-squared test (Chi2) (Jin, et al., 2006), McTwo (Ge, et al., 2016), random-forest based recursive feature elimination (RF-RFE) (Granitto, et al., 2006) and support vector machine based recursive feature elimination (SVM-RFE) (Duan, et al., 2005).

Two more popular feature selection algorithms were evaluated. T-test based feature ranking algorithm was widely used to evaluate the phenotype-association of each feature, and usually the incremental feature selection (IFS) was utilized to find the best number of top-ranked features (Gharbali, et al., 2018; Ye, et al., 2017). The Lasso algorithm evaluated the features by minimizing the L1-penalty and assigned a weight to each feature (Deshpande, et al., 2019; Kumar, et al., 2017).

Classification of DNA replication origins

Classification algorithms were used to separate the positive samples from the negative ones in each of the four datasets {D(Sc), D(Sp), D(Kl), D(Pp)}. This study utilized five classification algorithms to evaluate their capabilities of predicting DNA replication origins, i.e., support vector machine (SVM) (Weston, et al., 2001), random forest (RF) (Jang, et al., 2018; Li, et al., 2018), multinomial naïve Bayes (MNB) classifier (Pan, et al., 2018), gradient boosting decision tree (GBDT) (Liang, et al., 2018; Wang, et al., 2019), and back propagation neural network (BPNN) (Rumelhart, et al., 1986).

Another two popular classifiers were used for further validation. Xgboost was a gradient-boosting-based classification algorithm and generated quite a few successful applications (Deng, et al., 2019; Qiang, et al., 2018; Turki and Wei, 2018). It has been widely used for the diagnosis of cancers based on transcriptomic datasets (Turki and Wei, 2018). Xgboost outperformed the existing algorithms on both the RNA modification residues and protein-RNA binding (Deng, et al., 2019; Qiang, et al., 2018), etc. Extreme learning machine (ELM) was proposed by Prof. Guangbing Huang to solve the challenge of the slow training process of feed-forward neural networks (Huang, et al., 2006). ELM significantly speeds up the learning speed of the generalized feed-forward network and allows the single hidden layer in this network to be un-tuned (Huang, et al., 2011). ELM has been widely utilized to predict various bioinformatics questions, e.g., protein complex (Li, et al., 2019) and biomedical imaging data (Zhang, et al., 2019), etc.

Evaluation Method of Performance

This study used six performance metrics to evaluate how a classification algorithm performed on the investigated binary classification problem. Five metrics were sensitivity (Sn), specificity (Sp), overall accuracy (Acc), balanced accuracy (bAcc), and Matthews correlation coefficient (MCC) (Feng, et al., 2018; Xu, et al., 2018). These five performance metrics were defined as:

$$\begin{cases} Sn = TP/P,\ \ and\ 0 \le Sn \le 1 \\ Sp = TN/N,\ \ and\ 0 \le Sp \le 1 \\ Acc = (TP+TN)/(P+N),\ \ and\ 0 \le Acc \le 1 \\ bAcc = (Sn+Sp)/2,\ \ and\ 0 \le bAcc \le 1 \\ MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(FP+TP)\times(FN+TP)\times(FP+TN)\times(FN+TN)}}, and -1 \le MCC \le 1 \end{cases} \qquad (6)$$

The numbers of positive samples with correct and incorrect predictions were True Positive (TP) and False Negative (FN), respectively. While True Negative (TN) and False Positive (FP) were the numbers of negative samples with and without correct predictions, respectively. So the number of positive samples was P=TP+FN, and the number of negative samples was N=TN+FP. The area under the ROC curve (AUC) may be calculated by the integral calculus (Nguyen and Rebello, 2011).

A classifier was evaluated for optimizing its parameters by the 5-fold cross-validation strategy, and its classification performance was calculated by the leave-one-out validation strategy for the comparison with the existing studies. The k-fold cross validation strategy randomly split both the positive and negative datasets into k equally-sized bins and iteratively tested each pair of one positive and one negative bin with the model trained over the other samples (Wang, et al., 2018; Wang, et al., 2016; Zhao, et al., 2018). The final performance metrics were averaged over all the samples. This strategy ensured that each sample was tested for once. So the model parameters were tuned to optimize the performance metrics of the 5-fold cross validation strategy. The existing studies provided their performance metrics by the leave-one-out (LOO) validation, and a comparative analysis was carried out using LOO in this study.

Comparison with other PseKNC features

A further investigation was carried out to evaluate whether the other PseKNC feature groups may improve our models, as shown in Supplementary Figure S1. There are 12 feature groups for each of the types 1 and 2 of PseKNC. This study utilized the feature groups MW and NU of type 1, defined as Type1-2. These two feature groups under type 2 were defined as Type2-2. The sets of all the 12 feature groups of types 1 and 2 were defined as Type1-12 and Type2-12.

The same feature selection procedure was applied to the four datasets, i.e., Type1-2, Type1-12, Type2-2, and Type2-12. Unfortunately, no features were selected from the dataset Type2-12. So we selected 50 features top-ranked by SVM-RFE from Type2-12, which is the same feature number compared to that of Type1-2. A subset of top-ranked 200 features was also selected from Type2-12, which is in proportion to the 50 features in Type1-2.

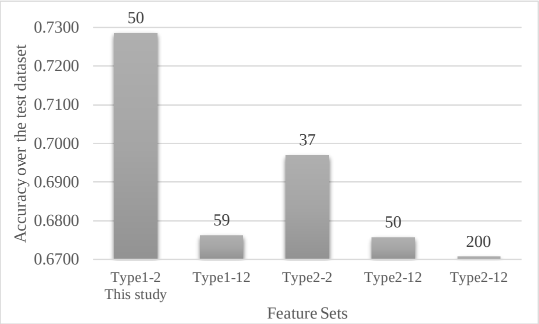
The experimental data demonstrated that our model performed the best compared with all the other PseKNC feature sets, as shown in Supplementary Figure S1. Although BPNN was the best classifier in this study, its learning speed was very slow compared with the other classifiers

(Gu, et al., 2016). So only the feature subsets selected from the original features were evaluated for their BPNN-based classification performances. If all the 12 feature groups were utilized, BPNN achieved at least 0.05 in the accuracy improvement. The prediction accuracy of the 37 features selected from Type2-2 was 0.6969, which was improved by 0.0316 compared with the 50 features selected from Type1-2.

So the two feature groups MW-daltons (MW) and Nucleosome (NU) from type 1 generated the best prediction accuracy of the DNA replication origins.

Supplementary Figure S1

Performances of different PseKNC features on the species Schizosaccharo mycres pombe.



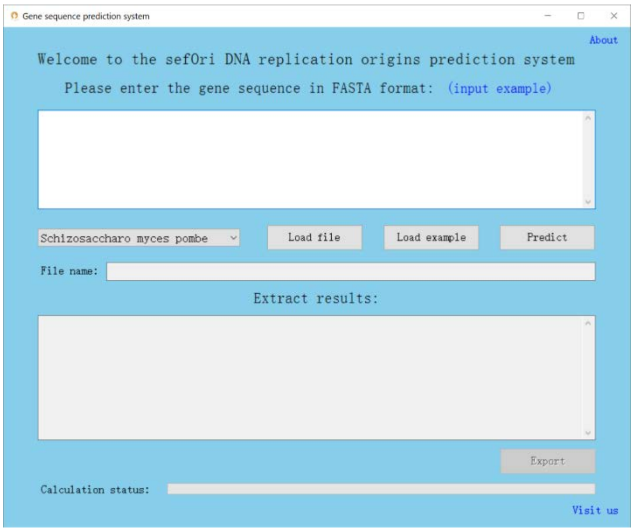
Type1-2 and Tpe2-2 are the MW and NU feature groups of type 1 and 2, respectively. Type1-12 and Type2-12 are the collections of all the 12 feature groups of type 1 and 2, respectively. A feature subset was selected from each feature set, and the feature number was given on the top of each column. All the prediction accuracies were calculated by the five-fold cross-validation strategy of the classifier BPNN.

An easy-to-use GUI-based prediction software

The prediction models for the four yeast genomes were developed as an easy-to-use software, sefOri version 1.0, as shown in Supplementary Figure S2. The user may choose to use one of the four prediction models for the four yeast genomes in the dropbox. The DNA sequences in the FASTA format may be input in the top textbox by copying-pasting from a text editor or clicking the button “Load file” to load the sequences from a FASTA file. Two example sequences may be loaded in the top textbox by clicking the button “Load example”. Click the button “Predict” will start the prediction procedure, and the final results will be given in a popup window. The prediction procedure will be illustrated by the progress bar in the bottom. The bottom textbox gives the features calculated from the query DNA sequences, and these features may be output as a CSV file by clicking the button “Export”. The user may carry out further model improvements by using these features.

Supplementary Figure S2

Graphical User Interface (GUI) of the software sefOri version 1.0. The top textbox is to load the DNA sequences in FASTA format. The bottom textbox gives the features calculated from the query DNA sequences in the top textbox.



Welcome to the sefOri DNA replication origins prediction system

Please enter the gene sequence in FASTA format: (input example)

Biological implications of the chosen features

The two feature groups Molecular Weight (MW) and Nucleosome Occupancy (NO) demonstrated the best prediction performances. The feature group Molecular Weight (MW) described the molecular weight in daltons of the raw DNA sequences. Different short DNA motifs represented by the PseKNC algorithm usually have varied molecular weight (Iguchi-Ariga, et al., 1993). So MW was considered as a representative feature group for predicting the DNA replication origins of the yeast genomes.

The other feature group NO described the status of the nucleosome occupancy and was also known to be associated with DNA replication origin sites (Yin, et al., 2009). For example, both yeast and human genome tend to deploy the DNA replication origins in the nucleosome depletion regions (Yin, et al., 2009). The DNA replication origins were also frequently observed in the nucleosome position enriched regions (Eaton, et al., 2010). So there exists a molecular machinery that can recognize the DNA replication origins.

The type 1 feature groups MW and NO demonstrated a very good prediction performance. The experimental data in the above sections supported that these two feature groups were significantly associated with the class labels “DNA replication origin”. We took the list of 50 features chosen from the dataset of Schizosaccharomycres pombe as an example and discussed the biological implications of these 50 chosen features. The type 1 PseKNC features consist of 4k composite features and λ correlation features, where k=3 and λ has multiple values.

The 50 chosen features consist of 38 composite features and 12 correlated features for Schizosaccharo mycres pombe, as shown in Supplementary Table S1. We evaluated the statistical significance of each chosen feature with the phenotype DNA replication origin using t-test.

Four tri-mers (AAA, AAT, TTA, and TTT) were chosen in at least two choices of λ values and were statistically differentially represented (Pvalue<1e-5) between the DNA replication origins and the controls, as shown in Supplementary Table S1. Multiple consensus sequences were observed in the DNA replication origin regions and the 11-bp sequence [5/(A/T)TTTA(T/C)(A/G)TTT(A/T)-3] supports all the known cases in S. cerevisiae (Linskens and Huberman, 1988). Another shorter sequence motif A/TTA/T was found to be conserved in six Saccharomyces species (Chang, et al., 2008). Both studies suggested that the tri-mer TTA may be an essential element in the DNA replication origins (Leonard and Mechali, 2013). The DNA replication origins were also known to be AT-rich and these four tri-mer features (AAA, AAT, TTA, and TTT) suggested that a prediction model also relies on these AT-rich tri-mers for achieving accurate predictions (Lee, et al., 2001; Segurado, et al., 2003).

Supplementary Table S2

Features chosen for Schizosaccharomyces pombe.

Schizosaccharomyces pombe					
λ=0	λ=10	λ=20	λ=30	λ=40	λ=50
CAT	AAG	GCT	CAT	GCC	GGC
	CAT	AGC	CCC	GCT	AGC
	TGA	AAG	CCT	TGC	AGT
	TTT	AAA	TGA	AAG	CGC
		AAT	TCA	AAA	CAG
		TTA	TCC	AAT	TGA
			92	CGC	TCG
				CAT	TTA
				CTA	64
				TGT	65
				TAT	69
				TCC	81
				TTT	90
				95	91
					92
					95
					98
					99

References

Chang, F., *et al.* Analysis of chromosome III replicators reveals an unusual structure for the ARS318 silencer origin and a conserved WTW sequence within the origin recognition complex binding site. *Mol Cell Biol* 2008;28(16):5071-5081.

Chatterjee, S., *et al.* Clinical application of modified bag-of-features coupled with hybrid neural-based classifier in dengue fever classification using gene expression data. *Med Biol Eng Comput* 2018;56(4):709-720.

Chen, W., *et al.* iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal Biochem* 2015;490:26-33.

Chen, W., *et al.* PseKNC: A flexible web server for generating pseudo K-tuple nucleotide composition. *Analytical Biochemistry* 2014;456:53-60.

Chen, W., *et al.* iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. 2012;7(10):e47843.

Deng, L., Sui, Y. and Zhang, J. XGBPRH: Prediction of Binding Hot Spots at Protein(-)RNA Interfaces Utilizing Extreme Gradient Boosting. *Genes (Basel)* 2019;10(3).

Deshpande, S., *et al.* PLIT: An alignment-free computational tool for identification of long non-coding RNAs in plant transcriptomic datasets. *Comput Biol Med* 2019;105:169-181.

Duan, K.-B., *et al.* Multiple SVM-RFE for gene selection in cancer classification with expression data. 2005;4(3):228-234.

Eaton, M.L., *et al.* Conserved nucleosome positioning defines replication origins. *Genes Dev* 2010;24(8):748-753.

Feng, X., *et al.* Selecting Multiple Biomarker Subsets with Similarly Effective Binary Classification Performances. *J Vis Exp* 2018(140).

Ge, R., *et al.* McTwo: a two-step feature selection algorithm based on maximal information coefficient. *BMC Bioinformatics* 2016;17:142.

Gharbali, A.A., Najdi, S. and Fonseca, J.M. Investigating the contribution of distance-based features to automatic sleep stage classification. *Comput Biol Med* 2018;96:8-23.

Granitto, P.M., *et al.* Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. 2006;83(2):83-90.

Gu, T., *et al.* Pressure prediction model based on artificial neural network optimized by genetic algorithm and its application in quasi-static calibration of piezoelectric high-pressure sensor. *Rev Sci Instrum* 2016;87(12):125005.

Guo, P., *et al.* Gene expression profile based classification models of psoriasis. *Genomics* 2014;103(1):48-55.

Huang, G.-B., *et al.* Extreme learning machines: a survey. 2011;2(2):107-122.

Huang, G.-B., Zhu, Q.-Y. and Siew, C.-K.J.N. Extreme learning machine: theory and applications. 2006;70(1-3):489-501.

Iguchi-Ariga, S.M., *et al.* Identification of the initiation region of DNA replication in the murine immunoglobulin heavy chain gene and possible function of the octamer motif as a putative DNA replication origin in mammalian cells. 1993;1172(1-2):73-81.

Jang, B.S., *et al.* Prediction of Pseudoprogression versus Progression using Machine Learning Algorithm in Glioblastoma. *Sci Rep* 2018;8(1):12516.

Jin, X., *et al.* Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In, *International Workshop on Data Mining for Biomedical Applications*. Springer; 2006. p. 106-115.

Kumar, S., Mamun, K. and Sharma, A. CSP-TSM: Optimizing the performance of Riemannian tangent space mapping using common spatial pattern for MI-BCI. *Comput Biol Med* 2017;91:231-242.

Lee, J.K., *et al.* The Schizosaccharomyces pombe origin recognition complex interacts with multiple AT-rich regions of the replication origin DNA by means of the AT-hook domains of the spOrc4 protein. *Proc Natl Acad Sci U S A* 2001;98(24):13589-13594.

Leonard, A.C. and Mechali, M. DNA replication origins. *Cold Spring Harb Perspect Biol* 2013;5(10):a010116.

Li, J., *et al.* RNAm5Cfinder: A Web-server for Predicting RNA 5-methylcytosine (m5C) Sites Based on Random Forest. *Sci Rep* 2018;8(1):17299.

Li, Y., Niu, M. and Zou, Q. ELM-MHC: An Improved MHC Identification Method with Extreme Learning Machine Algorithm. *J Proteome Res* 2019;18(3):1392-1401.

Liang, S., *et al.* Classification of First-Episode Schizophrenia Using Multimodal Brain Features: A Combined Structural and Diffusion Imaging Study. *Schizophr Bull* 2018.

Linskens, M.H. and Huberman, J.A. Organization of replication of ribosomal DNA in Saccharomyces cerevisiae. *Mol Cell Biol* 1988;8(11):4927-4935.

Liu, B., *et al.* iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. *Bioinformatics* 2018;34(22):3835-3842.

Liu, B., *et al.* iRO-3wPseKNC: identify DNA replication origins by three-window-based PseKNC. *Bioinformatics* 2018;34(18):3086-3093.

Lyubchenko, Y.L., Shlyakhtenko, L.S. and Gall, A.A. Atomic force microscopy imaging and probing of DNA, proteins, and protein DNA complexes: silatrane surface chemistry. *Methods Mol Biol* 2009;543:337-351.

Nguyen, D.-H. and Rebello, N.S.J.P.R.S.T.-P.E.R. Students' understanding and application of the area under the curve concept in physics problems. 2011;7(1):010112.

Pan, Y., *et al.* Identification of Bacteriophage Virion Proteins Using Multinomial Naive Bayes with g-Gap Feature Tree. *Int J Mol Sci* 2018;19(6).

Ponomarenko, J.V., *et al.* Conformational and physicochemical DNA features specific for transcription factor binding sites. 1999;15(7):654-668.

Qiang, X., *et al.* M6AMRFS: Robust Prediction of N6-Methyladenosine Sites With Sequence-Based Features in Multiple Species. *Front Genet* 2018;9:495.

Rachofsky, E.L., Ross, J.B. and Osman, R. Conformation and dynamics of normal and damaged DNA. *Comb Chem High Throughput Screen* 2001;4(8):675-706.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. Learning representations by back-propagating errors. *Nature* 1986;323(6088):533.

Segurado, M., de Luis, A. and Antequera, F. Genome-wide distribution of DNA replication origins at A+T-rich islands in Schizosaccharomyces pombe. *EMBO Rep* 2003;4(11):1048-1053.

Turki, T. and Wei, Z. Boosting support vector machines for cancer discrimination tasks. *Comput Biol Med* 2018;101:236-249.

Wang, C., *et al.* Predicting Future Driving Risk of Crash-Involved Drivers Based on a Systematic Machine Learning Framework. *Int J Environ Res Public Health* 2019;16(3).

Wang, L., *et al.* Combining High Speed ELM Learning with a Deep Convolutional Neural Network Feature Encoding for Predicting Protein-RNA Interactions. *IEEE/ACM Trans Comput Biol Bioinform* 2018.

Wang, Y., *et al.* Machine learning based detection of age-related macular degeneration (AMD) and diabetic macular edema (DME) from optical coherence tomography (OCT) images. *Biomed Opt Express* 2016;7(12):4928-4940.

Weston, J., *et al.* Feature selection for SVMs. In, *Advances in neural information processing systems*. 2001. p. 668-674.

Xu, C., *et al.* An OMIC biomarker detection algorithm TriVote and its application in methylomic biomarker detection. *Epigenomics* 2018;10(4):335-347.

Ye, Y., *et al.* RIFS: a randomly restarted incremental feature selection algorithm. *Sci Rep* 2017;7(1):13013.

Yin, S., *et al.* The impact of nucleosome positioning on the organization of replication origins in eukaryotes. 2009;385(3):363-368.

Zhang, F., *et al.* Voxel-Based Morphometry: Improving the Diagnosis of Alzheimer's disease based on an Extreme Learning Machine method from the ADNI cohort. *Neuroscience* 2019.

Zhao, R., *et al.* TriZ-a rotation-tolerant image feature and its application in endoscope-based disease diagnosis. *Comput Biol Med* 2018;99:182-190.