Sc-ncDNAPred: A Sequence-Based Predictor for Identifying Non-coding DNA in \textit{Saccharomyces cerevisiae}

Wenying He\textsuperscript{1}, Ying Ju\textsuperscript{2}, Xiangxiang Zeng\textsuperscript{2}, Xiangrong Liu\textsuperscript{2*} and Quan Zou\textsuperscript{1,*}

\textsuperscript{1} School of Computer Science and Technology, Tianjin University, Tianjin, China, \textsuperscript{2} School of Information Science and Technology, Xiamen University, Xiamen, China, \textsuperscript{3} Shandong Provincial Key Laboratory of Biophysics, Institute of Biophysics, Dezhou University, Dezhou, China

With the rapid development of high-speed sequencing technologies and the implementation of many whole genome sequencing projects, research in the genomics is advancing from genome sequencing to genome synthesis. Synthetic biology technologies such as DNA-based molecular assemblies, genome editing technology, directional evolution technology and DNA storage technology, and other cutting-edge technologies emerge in succession. Especially the rapid growth and development of DNA assembly technology may greatly push forward the success of artificial life. Meanwhile, DNA assembly technology needs a large number of target sequences of known information as data support. Non-coding DNA (ncDNA) sequences occupy most of the organism genomes, thus accurate recognizing of them is necessary. Although experimental methods have been proposed to detect ncDNA sequences, they are expensive for performing genome wide detections. Thus, it is necessary to develop machine-learning methods for predicting non-coding DNA sequences. In this study, we collected the ncDNA benchmark dataset of \textit{Saccharomyces cerevisiae} and reported a support vector machine-based predictor, called Sc-ncDNAPred, for predicting ncDNA sequences. The optimal feature extraction strategy was selected from a group included mononucleotide, dimer, trimer, tetramer, pentamer, and hexamer, using support vector machine learning method. Sc-ncDNAPred achieved an overall accuracy of 0.98. For the convenience of users, an online web-server has been built at: http://server.malab.cn/Sc-ncDNAPred/index.jsp.

Keywords: non-coding DNA, DNA sequence, feature representation, genome synthesis, support vector machine

INTRODUCTION

After the implementation of many whole genome sequencing projects, more and more researches showed that non-coding DNA (ncDNA) is a major component of the biological genome. Numerous studies (Vogel, 1964; Thomas, 1971; Eddy, 2012; Puente et al., 2015; Liu et al., 2017a; Yao et al., 2018) have shown that the complexity of organisms is related to the length of non-coding regions, which are specially transcribed in physiological and disease states. Although the function of most ncDNAs is still unknown (Khurana et al., 2016), some studies (Horn et al., 2013; Huang et al., 2013; Vinagre et al., 2013; Puente et al., 2015; Hu et al., 2017, 2018; Rheinbay et al., 2017; Liao et al., 2018; Zhang W. et al., 2018) have shown that most cancer-related gene mutations are located in
ncDNA regions. How ncDNAs specifically affect tumor formation is also an urgent problem to be solved. In addition, ncDNAs in the genome play an important role in gene expression, regulatory, and inheritance (Khurana et al., 2016).

Especially, with the rapid growth and development of synthetic biology, research in the genomics is advancing from genome sequencing to genome synthesis (Erlich and Zielinski, 2017; Jain et al., 2018; Liu B. et al., 2018). In recent years, various DNA assembly technologies (Ni et al., 2017; Wu et al., 2017; Xie et al., 2017; Zhang et al., 2017b) have been developed according to the principles of atypical enzyme cut connection (Engler et al., 2009; Sleight et al., 2010), single strand annealing and splicing (Gibson et al., 2009; Li and Elledge, 2012) and PCR (Warrens et al., 1997), which provide more rapid technical support for synthetic biology. In the following years, people are committed to improving the efficiency of large scale DNA assembly technologies. With the rapid development of the computer network and the popularity of the Internet, the number of digital information, such as network data, audio data, and video data, is increasing rapidly. It is urgent to establish a new system which has more efficiency than the existing storage system. DNA storage technology (Baum, 1995; Davis, 1996; Carr and Church, 2009) can meet the requirements above. In a new study (Shipman et al., 2017), the researchers introduced a method that encode images and video images into the genome of the Escherichia coli and read the corresponding images and videos from the genome of living bacterial cells. All the above studies require a large amount of DNA data.

As a complex type of genetic information, DNA sequences have specific characteristics not only in the coding sequence (cDNA) but also in the ncDNA sequences. Currently, the identification of cDNAs and ncDNAs relies mainly on experimental methods. However, traditional experimental methods are time-consuming and laborious, and the amount of genomic data is large and the sequence types are complex. In this context, there is an urgent need to establish accurate and efficient prediction methods to mine the information and knowledge of ncDNAs and cDNAs. Computational methods, which achieve a complementary effect, indeed effectively improved the recognition accuracy (Zhou et al., 2016).

In this study, a SVM-based computational method was first established to recognize the ncDNA sequences in Saccharomyces cerevisiae (S. cerevisiae). Totally several types of features, such as mononucleotide composition (MNC), dimer nucleotide composition (DNC), trimer nucleotide composition (TNC), tetramer nucleotide composition (TrNC), pentamer nucleotide composition (PNC), and hexamer nucleotide composition (HNC) were extracted. The optimal feature extraction strategy was selected using SVM machine learning method. The workflow of constructing the Sc-ncDNAPred model is shown in Figure 1.

METHODS

Benchmark Dataset

In this study, the benchmark dataset was derived from the Ensembl genome database project (Hubbard et al., 2002), which is one of several well-known genome browsers for the retrieval of genomic information. Experimentally validated cDNA sequences of S. cerevisiae were extracted from their database, which contains 6713 samples. Intercepting the ncDNAs of the S. cerevisiae based on the initial marker information of the coding region provided by the original genomic data. By doing so, we obtained 6410 ncDNA samples. To get rid of redundancy, the CD-HIT (Li and Godzik, 2006) was adopted to remove those sequences that had ≥ 75% sequence identity. Finally, we obtained 6030 and 6251 samples in ncDNAs and cDNAs, respectively. Thus, the benchmark dataset can be formulated as

\[ S = S^+ \cup S^- \] (1)

where \( S^+ \) contained 6030 ncDNA samples, \( S^- \) contained 6251 cDNA samples and the symbol \( \cup \) means the ‘union’ in the set theory.

The length distribution of ncDNA samples was shown in Figure 2. According to the graph, the length distribution of ncDNA is mainly between 100 and 800.

Feature Vector Construction

A sample can be simplified by a convenience form as:

\[ P = R_1R_2R_3R_4 \ldots R_{L-1}R_L \] (2)

where \( R_i \) \((i = 1, 2, 3 \ldots L)\) represents the nucleotide at \( i\)-th position in one sequence.

K-mer Composition

K-mer nucleotide composition has been applied in many fields of bioinformatics (Liu et al., 2015b,c; Kim et al., 2017; Matias Rodrigues et al., 2017; Orenstein et al., 2017; Liu, 2018; Liu X. et al., 2018; Rangavittal et al., 2018). MNC equate to \( k = 1 \), DNC equate to \( k = 2 \), TNC equate to \( k = 3 \), TrNC equate to \( k = 4 \), PNC equate to \( k = 5 \), HNC equate to \( k = 6 \). The occurrence frequency of \( k - \text{mer}(i)\) can be represented as:

\[ f_k^i = f(k - \text{mer}(i)) = \frac{n_k^i}{L - k + 1} \]

\[(i = 1, 2, ..., 4^k; k = 1, 2, 3, 4, 5, 6) \tag{3}\]

where \( n_k^i \) denote the number of the \( i \)-th \( k \)-mer, \( L \) is the length of the sample sequence. Thus, each DNA sample can be defined feature vectors in different dimension of size \( 4^k \). The generalized form of whole feature vectors \( X \) can be given by:

\[ X = [f_1^1, f_2^1, \ldots f_{4^k}^1, \ldots f_1^4, f_2^4, \ldots f_{4^k}^4]_T \] (4)

Feature Ranking

Each sample sequence was represented by a large set of features, which leads to the redundant information (Wei and Billings, 2007; Senawi et al., 2017). In order to distinguish the contribution of different features to the prediction model. To analyze these feature vectors, F-score method (Chen W. et al., 2016; Jia and He, 2016; Tang et al., 2016, 2018; He and Jia, 2017) was adopted to
rank the feature, in this study. The $F$-score value of the $i$-th feature is defined as:

$$F-score(i) = \frac{1}{n^+ - 1} \sum_{k=1}^{n^+} (x_{ki}^{(+)} - \bar{x}_i^{ (+)})^2 + \frac{1}{n^- - 1} \sum_{k=1}^{n^-} (x_{ki}^{(-)} - \bar{x}_i^{ (-)})^2$$

(5)

where $\bar{x}_i$, $\bar{x}_i^{ (+)}$ and $\bar{x}_i^{ (-)}$ are the average values of the $i$-th feature in whole, ncDNA and cDNA datasets, respectively. $n^+$ represents the number of ncDNA training samples, $n^-$ represents the number of cDNA training samples, $x_{ki}^{(+)}$ represents the $i$-th feature of the $k$-th ncDNA sample and $x_{ki}^{(-)}$ represents the $i$-th feature of the $k$-th cDNA sample. Obviously, the feature with a greater score value indicates that it has a better discrimination ability.
Support Vector Machine

Support vector machine (SVM) (Hearst et al., 1998) is a widely used two-class classification algorithm based on statistical learning theory. It has been proven to be powerful in many fields of pattern recognition and data classification (Byun and Lee, 2002; Nasrabadi, 2007; Zhang N. et al., 2018). More and more applications also proved that SVM also has strong data processing capabilities in the fields of bioinformatics (Xiong et al., 2011; Jia et al., 2013, 2017; Cao et al., 2014; Liu et al., 2014, 2017b; Wei et al., 2015; Chen X. X. et al., 2016; Jia and He, 2016; Yang et al., 2016; Zou et al., 2016; Xiao et al., 2017; Qiao et al., 2018; Su et al., 2018). A set of ncDNA samples and cDNA samples were represented by the feature vectors. The SVM classifies the data by mapping the input feature vectors to a high-dimensional feature space using a kernel function. In this study, the public LIBSVM package (Chang and Lin, 2011) was implemented to train models for discriminating between ncDNA sequences and cDNA sequences. Here, the radial basis function (RBF) \[ K(S_i, S_j) = \exp(-\gamma \|S_i - S_j\|^2) \] was set as the kernel function. The penalty parameter \( C \) and kernel parameter \( \gamma \) were preliminarily optimized through a grid search strategy.

**Performance Evaluation**

K-fold cross-validation (Chou and Zhang, 1995; Kohavi, 1995; Zhang et al., 2012a,b; Liu et al., 2015a; Chen X. et al., 2016; Li et al., 2016; Luo et al., 2016; Chen et al., 2017b, 2018a,b; Pan et al., 2017a; Xu et al., 2017; He et al., 2018) is one of the widely used approach to examine the ability of prediction model, and other approaches: independent dataset test and jackknife test (Chou and Shen, 2008) are also used in many applications. To reduce the computational cost, 10-fold cross validation was used to examine each model for its effectiveness in identifying ncDNA sequences. The training dataset were randomly divided into 10 subsets of approximately the same size. In each iteration, one subset was chosen as the test set and the remaining 9 subsets were used to train the model. For a complete cycle of a 10-fold cross-validation, the process was repeated 10 times until each subset was chosen as a test set. This 10-fold cross-validation procedure was repeated five times, then the results were averaged.

To evaluate the prediction performance of the models, five classic metrics were computed (Chou, 2001; Qiu et al., 2015, 2016; Liu et al., 2017; Pan et al., 2017b; Zhang et al., 2017a; Tang et al., 2018; Yang et al., 2018), including sensitivity (Sn), specificity (Sp), accuracy (Acc), Matthew correlation coefficient (MCC), and the receiver operating characteristic (ROC). These measurements were defined as:

\[
Sn = 1 - \frac{N^-}{N^+}, \\
Sp = 1 - \frac{N^-}{N^-}, \\
Acc = 1 - \frac{N^- + N^-}{N^+ + N^-}, \\
MCC = \frac{1 - (N^- N^- + N^+ N^-)}{\sqrt{(1 + N^- N^-)(1 + N^+ N^-)}}
\]

**TABLE 1** | The 10-fold cross-validation results by different feature methods on the benchmark dataset.

| Methods | Sn (%) | Sp (%) | ACC (%) | MCC   |
|---------|--------|--------|---------|-------|
| MNC     | 80.56  | 87.02  | 83.85   | 0.678 |
| DNC     | 92.64  | 92.62  | 92.64   | 0.853 |
| TNC     | 96.62  | 97.22  | 96.93   | 0.939 |
| TrNC    | 98.01  | 98.51  | 98.26   | 0.965 |
| PNC     | 95.25  | 95.84  | 95.56   | 0.911 |
| HNC     | 90.71  | 91.49  | 91.03   | 0.830 |
| All Features | 95.99  | 96.08  | 96.03   | 0.921 |

The experiments have been executed 5 times and the results were the mean values.

**FIGURE 3** | The ROC curves to assess the predictive performance based on different feature extraction methods.

**FIGURE 4** | Heap map to illustrate the F-score values of 256 different tetramers to identify ncDNA and cDNA.
TABLE 2 | Rules of composition of heat map.

| AAAA | AAAC | AACC | ACAC | ACCA | ACCC | CAAA | CAAC | CACA | CACC | CCAA | CCAC | CCCA | CCCC |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| AAAG | AAAT | AACG | ACAG | ACTA | ACCT | CAAG | CAAT | ACAG | CACT | CCAG | CCA | CCGG | CCCT |
| AAGA | AAGC | AATG | AATT | AGTG | AGCT | AATG | ATGC | ATTC | ATTG | ATGT | ATCC | ATTG | ATTT |
| GAAA | GAAC | GACA | GACC | GCAC | GCCA | GCAC | GCCA | GCCA | GCCC | TAAA | TAAC | TACA | TACC |
| GAAG | GATA | GAGC | GAGT | GAGT | GAGT | GAGT | GAGT | GAGT | GAGT | GAGT | GAGT | GAGT | GAGT |
| GCAA | GGAC | GGCA | GGCG | GTGA | GTAC | GTCA | GTCC | GTCC | GTCC | GTCC | GTCC | GTCC | GTCC |
| GCGA | GCGT | GCGT | GCGT | GCGT | GCGT | GCGT | GCGT | GCGT | GCGT | GCGT | GCGT | GCGT | GCGT |
| GGGG | GGCT | GGGT | GGGT | GGGT | GGGT | GGGT | GGGT | GGGT | GGGT | GGGT | GGGT | GGGT | GGGT |

In these expressions, \(N^+\) and \(N^-\) are the total number of ncDNA and cDNA samples, respectively, while \(N^+_c\) and \(N^-_c\) are respectively the number of ncDNA samples incorrectly predicted as cDNA samples, and the number of cDNA samples incorrectly predicted as ncDNA samples.

RESULTS AND DISCUSSION
Prediction Results of Models
We used six types of effective feature extraction methods, such as MNC, DNA, TNC, TrNC, PNC, and HNC, as input of SVM to establish six models. The ability of each feature extraction method to discriminate between ncDNA and cDNA samples was compared by the 10-fold cross-validation (Table 1). As we can see from Table 1, the model for a combination SVM and TrNC yielded the best prediction performance, with the accuracy of 98.26%, the sensitivity of 98.01%, the specificity of 98.51%, and the MCC of 0.965, respectively. Then, the following second best prediction performance was yielded by TNC with the accuracy of 96.93%, the sensitivity of 96.62%, the specificity of 97.22%, and the MCC of 0.939, respectively. Besides, in the case of PNC, the corresponding model still obtained a good prediction results, which are 95.56% of accuracy, 95.25% of sensitivity, 95.84% of specificity and 0.911 of MCC, respectively.

To further investigate the overall prediction performance of each model, we showed the ROC curves and AUC values of different models for the 10-fold cross-validation in Figure 3. With the increase of \(k\)-mer value, the performance first increased and then decreased. Comparison demonstrated that the TrNC could produce the best results. Thus, the feature TrNC was adopted as the final model for Sc-ncDNAPred.

To further optimize the model, we performed multiple rounds of experiments on TrNC to select the appropriate subset of all 256 features (see Additional file 1: Table S1 for full details); however, the results showed no significant improvement in the corresponding performance. The possible reason is that the selected feature cannot burden enough information for the discrimination.

Compositional Analysis
To understand the 256 different tetramers bias in ncDNAs and cDNAs, a heap map was provided in Figure 4. Each square in the heat map corresponds to the \(F\)-score value of one tetramer (see Table 2 for full details). Deep red in the heat map corresponds to a strong recognition ability.

Heap map analysis revealed that tetramers include TATA, TTGG, CAAG, CCAA, ATAT, TAAA, TGGG, TTAT, ATGG, ATAA, AATA, and CTGG are with the \(F\)-score values ranking...
top twelve in all tetramers. In addition, we also analyzed the other k-mer components based on the F-score method, respectively. Among them, the two key nucleotides G and T from MNC, the top five key dimer nucleotide composition (TA, CG, GA, TT, and CA) from DNC, (TGG, ATA, CCA, TAT, and TTT) from TNC, (TTTTT, ATATA, TAAAA, TATAT, and TTTTTA) from PNC, and (TTTTTT, ATTTTT, TTTTTC, and CTTTTT) from HNC. These key features are presented in a radar diagram (Figure 5). The study of these key features can deepen the understanding of the overall structure of the genome, which not only promotes the annotation of the genome, but also promotes the study of biological evolution.

Comparison With Other Classifiers
To the best of our knowledge, this is the first time that machine learning method has been used to identify ncDNA in S. cerevisiae. In order to further testify the superiority of proposed model Sc-ncDNAPred, the predictive results of it were compared with that of other powerful and widely used classifiers, i.e., k-Nearest Neighbor (KNN), Naïve Bayes, Random Forest, and J48 Tree as implemented in WEKA (Frank et al., 2004). The 10-fold cross validation results of these four classifier for identifying ncDNA in the same benchmark dataset were shown in Additional file 1: Table S2. The results showed that the four metrics as defined in Eq. 6 of the proposed model Sc-ncDNAPred are all higher than those of k-Nearest Neighbor (KNN), Naïve Bayes, Random Forest, and J48 Tree.

Web-Server
Based on the benchmark dataset defined in Eq.1, a predictor called Sc-ncDNAPred was established, where “Sc” stands for S. cerevisiae and “Pred” stands for “Prediction.” For conveniences of users’ community, a step-by-step guide about how to use the web-server is provided as follows:

Step 1. Open the web-server at: http://server.malab.cn/Sc_ncDNAPred/index.jsp, you will see the home page of Sc-ncDNAPred, as shown in Figure 6. Click the “About” button to see a brief introduction of the server.
Step 2. Paste the query DNA sequences into the input box. The input sequence should be in FASTA format. For the example of DNA sequences in FASTA format, click the “example” button top above the input box.
Step 3. Click on the “Submit” button to start the prediction. If the prediction result of a sequence is positive, its output is “ncDNA.” Otherwise, its output is “cDNA.”
Step 4. Click on the “DataSet” button to download the benchmark dataset.
Step 5. Click on the “Contact” button to contact us.

CONCLUSIONS
DNA assembly technology needs a large number of target sequences of known information as data support. Non-coding DNA (ncDNA) sequences occupy most of the organism genomes, thus accurate recognizing of them is necessary. In this study, an efficient computational model was proposed to identify ncDNAs in S. cerevisiae. The tetramer nucleotide composition (TrNC) was adopted to extract features. The F-score method was used to analyze these feature vectors and find the key features. The high accuracy indicated that Sc-ncDNAPred was a powerful tool for predicting ncDNA. Finally, a free web-server was developed based on the proposed model. We hope...
that the predictor will provide convenience to most of scholars. Currently, annotations for the genomic sequences of most species are lacking or unavailable. To analyze the ncDNA data of these organisms, we can obtain data and methodological support in a cross-species manner from annotated species. For example, we could try to use the model built from *S. cerevisiae* dataset to analyze other species of bacteria that have not been explored in depth. In addition, we will also apply this computational model for the prediction of potential disease related non-coding DNA. In the future, we will apply this computational model for the prediction of potential disease related non-coding RNA (Chen and Huang, 2017; Chen et al., 2017a, 2018c,d; You et al., 2017).

**AUTHOR CONTRIBUTIONS**

WH, QZ, and XL wrote the paper. XZ and YJ participated in preparation of the manuscript. QZ, WH, XL, XZ, and YJ participated in the research design. WH and QZ developed the web server. WH, YJ, XL, XZ, and QZ read and approved the final manuscript.

**REFERENCES**

Baum, E. B. (1995). Building an associative memory vastly larger than the brain. Science 268, 583–585.

Byun, H., and Lee, S. W. (2002). Applications of support vector machines for pattern recognition: a survey. In: *Pattern Recognition With Support Vector Machines*. Springer p (Niagara Falls, ON), 213–236.

Cao, R., Wang, Z., Wang, Y. H., and Cheng, J. (2014). SMOQ: a tool for predicting Chao, R., Wang, Z., Wang, Y. H., and Cheng, J. (2014). SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. *BMC Bioinform.* 15:120. doi: 10.1186/1471-2105-15-120

Carr, P. A., and Church, G. M. (2009). Genome engineering. *Nat. Biotechnol.* 27, 1151–1162. doi: 10.1038/nbt.1590

Chang, C. C., and Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intel. Syst. Technol.* 2:27. doi: 10.1145/1961189.1961199

Chen, X., Tang, H., Li, W. C., Wu, H., Chen, W., Ding, H., et al. (2016). Identification of bacterial cell wall lysases via pseudo amino acid composition. *Biomol. Res. Int.* 2016:1654623. doi: 10.1155/2016/1654623

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct. Funct. Bioinform.* 43, 246–55. doi: 10.1002/prot.1033

Chou, K. C., and Shen, H. B. (2008). Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* 3, 153–162. doi: 10.1038/nprot.2007.494

Engler, C., Gruetzner, R., Kandzia, R., and Marillonnet, S. (2009). Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PloS ONE* 4:e5553. doi: 10.1371/journal.pone.0005553

Engler, C., Gruetzner, R., Kandzia, R., and Marillonnet, S. (2008). Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PloS ONE* 4:e5553. doi: 10.1371/journal.pone.0005553

Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics* 20, 2479–2481. doi: 10.1093/bioinformatics/bth261

Gibson, D. G., Young, L., Chuang, R. Y., Venter, J. C., Hutchison, C. A. III., and Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6, 343–345. doi: 10.1038/nmeth.1318

He, W., and Jia, C. (2017). EnhancerPred2. 0: predicting enhancers and their strength based on position-specific trinucleotide propensity and electron–ion interaction potential feature selection. *Mol. Biol. Syst.* 13, 767–74. doi: 10.1093/mbe/sbw054

He, W., Jia, C., Duan, Y., and Zou, Q. (2018). 70Propred: a predictor for discovering sigma70 promoters based on combining multiple features. *BMC Syst. Biol.* 12:44. doi: 10.1186/s12918-018-0570-1

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intell. Syst. Appl.* 13, 18–28.

**FUNDING**

The work was supported by the National Natural Science Foundation of China (Nos. 61771331, 61472333, 61772441, 61472335, 61425002), Funding from Shandong Provincial Key Laboratory of Biophysics, Project of marine economic innovation and development in Xiamen (No. 16PFW034SF02), National Science Foundation of the Higher Education Institutions of Fujian Province (No. JZ160400), Natural Science Foundation of Fujian Province (No. 2017J01099), President Fund of Xiamen University (No. 20720170054), and Shenzhen Overseas High Level Talents Innovation Foundation (No. KJSCX20170327161949608). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**SUPPLEMENTARY MATERIAL**

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.02174/full#supplementary-material
Shipman, S. L., Nivala, J., Macklis, J. D., and Church, G. M. (2017). CRISPR-cas encoding of a digital movie into the genomes of a population of living bacteria. Nature 547, 345–349. doi: 10.1038/nature23017

Sleigh, S. C., Bartley, B. A., Lieviant, J. A., and Sauro, H. M. (2010). Infection biobrick assembly and re-engineering. Nucleic Acids Res. 38, 2624–2636. doi: 10.1093/nar/gkq417

Su, Z. D., Huang, Y., Zhang, Z. Y., Zhao, Y. W., Wang, D., Chen, W., et al. (2018). iLoc-IncRNA: predict the subcellular location of IncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics. doi: 10.1093/bioinformatics/bty508. [Epub ahead of print].

Tang, H., Chen, W., and Lin, H. (2016). Identification of immunoglobulins using Chou’s pseudo amino acid composition with feature selection technique. Mol. Biol. Sys. 12, 1269–1275. doi: 10.1039/c5mb00883b

Tang, H., Zhao, Y. W., Zou, P., Zhang, C. M., Chen, R., Huang, P., et al. (2018). HBPred: a tool to identify growth hormone-binding proteins. Int. J. Biol. Sci. 14, 957–964. doi: 10.7150/ijbs.24174

Thomas, C. A. Jr. (1971). The genetic organization of chromosomes. Annu. Rev. Genet. 5, 237–256.

Vinagre, J., Almeida, A., Pópulo, H., Batista, R., Lyra, J., Pinto, V., et al. (2013). Enhancing protein fold prediction by using asymmetric amplification: an improved technique for the generation of hybrid proteins of immunological interest. Gene 186, 29–35.

Wei, H. L., and Billings, S. A. (2007). Feature subset selection and ranking for data dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. 29, 162–6. doi: 10.1109/TPAMI.2007.11

Wei, L., Liao, M., Gao, X., and Zou, Q. (2015). Enhanced protein fold prediction method through a novel feature extraction technique. IEEE Trans. Nanobiosci. 14, 649–659. doi: 10.1109/TNBN.2015.2450233

Wu, Y., Li, B. Z., Zhao, M., Mitchell, L. A., Xie, Z. X., Lin, Q. H., et al. (2017). Bug mapping and testing fitness of chemically synthesized chromosome X. Science 355:eaaf6706. doi: 10.1126/science.aaf6706

Xiao, Y., Zhang, L., and Deng, L. (2017). Prediction of IncRNA-protein interactions using HeteSim scores based on heterogeneous networks. Sci. Rep. 7:3664. doi: 10.1038/s41598-017-03986-1

Xie, Z. X., Li, B. Z., Mitchell, L. A., Wu, Y., Qi, X., Jin, Z., et al. (2017). “Perfect” designer chromosome V and behavior of a ring derivative. Science 355:eaaf4704. doi: 10.1126/science.aaf4704

Xiong, Y., Liu, J., and Wei, D. Q. (2011). An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins 79, 509–517. doi: 10.1002/prot.22898

Xu, Q., Xiong, Y., Dai, H., Kumari, K. M., Xu, Q., Ou, H. Y., et al. (2017). PDC-SGB: prediction of effective drug combinations using a stochastic gradient boosting algorithm. J. Theor. Biol. 417, 1–7. doi: 10.1016/j.jtbi.2017.01.019

Yang, H., Qiu, W. R., Liu, G. Q., Guo, F. B., Chen, W., Chou, K. C., et al. (2018). iRSpot-PseNC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC. Int. J. Biol. Sci. 14, 883–891. doi: 10.7150/ijbs.24616

Yang, H., Tang, H., Chen, X. X., Zhang, C. J., Zhu, P. P., Ding, H., et al. (2016). Identification of secretory proteins in Mycobacterium tuberculosis using pseudo amino acid composition. Biomed. Res. Int. 2016:5413903. doi: 10.1155/2016/5413903

Yao, Y. H., Li, X. H., Geng, L. L., Nan, X. Y., Qi, Z. H., and Liao, B. (2018). Recent progress in long noncoding RNAs prediction. Curr. Bioinformatics 13, 344–351. doi: 10.2174/157489361266170905153933

Yau, Z. H., Huang, Z. A., Zhu, Z., Yan, G. Y., Li, Z. W., and Wen, Z. (2017). PRMDA: a novel and effective path-based computational model for miRNA-disease association prediction. PLoS Comput. Biol. 13:e1005455. doi: 10.1371/journal.pcbi.1005455

Zhang, N., Yu, S., Guo, Y., Wang, L., Wang, P., and Feng, Y. (2018). Discriminating Ramos and Jurkat Cells with image textures from diffraction imaging flow cytometry based on a support vector machine. Curr. Bioinform. 13, 50–6. doi: 10.2174/157489361166608102537

Zhang, W., Bojorguez-Gomez, A., Velez, D. O., Xu, G., Sanchez, K. S., Shen, J. P., et al. (2018). A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. Nat. Genet. 50, 613–620. doi: 10.1038/s41588-018-0091-2

Zhang, W., Liu, J., Zhao, M., and Li, Q. (2012a). Predicting linear B-cell epitopes by using sequence-derived structural and physicochemical features. Int. J. Data Min. Bioinform. 6, 557–569. doi: 10.1504/IJDMB.2012.049298

Zhang, W., Niu, Y., Xiong, Y., Zhao, M., Yu, R., and Liu, J. (2012b). Computational prediction of conformational B-cell epitopes from antigen primary structures by ensemble learning. PloS ONE 7:e3575. doi: 10.1371/journal.pone.0043575

Zhang, W., Niu, Y., Zou, H., Luo, L., Liu, Q., and Wu, W. (2015). Accurate prediction of immunogenic T-cell epitopes from epitope sequences using the genetic algorithm-based ensemble learning. PloS ONE 10:e0128194. doi: 10.1371/journal.pone.0128194

Zhang, W., Zhao, G., Luo, Z., Lin, Y., Wang, L., Guo, Y., et al. (2017b). Engineering the ribosomal DNA in a megabase synthetic chromosome. Science 355:eaaf3981. doi: 10.1126/science.aaf3981

Zhang, W., Zhu, X., Fu, Y., Tsuji, J., and Weng, Z. (2017a). Predicting human splicing branchpoints by combining sequence-derived features and multi-label learning methods. BMC Bioinform. 18(Suppl. 13):464. doi: 10.1186/s12859-017-1875-6

Zhou, L. Q., Li, R., and Hu, L. (2016). Enhanced prediction of small non-coding RNA in bacterial genomes based on improved inter-nucleotide distances of genomes. Curr. Bioinform. 11, 169–72. doi: 10.2174/1574893611666022301114

Zou, Q., Liu, W., Merler, M., and Ji, R. (2016). Advanced learning for large-scale heterogeneous computing. Neurocomputing 217, 1–2. doi: 10.1016/j.neucom.2016.06.009

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 He, Ju, Zeng, Liu and Zou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.