Identification of cyclin protein using gradient boost decision tree algorithm

Hasan Zulfiqar a, Shi-Shi Yuan a, Qin-Lai Huang a, Zi-Jie Sun a, Fu-Ying Dao a, Xiao-Long Yu b,a, Hao Lin a,a

a School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China
b School of Materials Science and Engineering, Hainan University, Haikou 570228, China

ABSTRACT

Cyclin proteins are capable to regulate the cell cycle by forming a complex with cyclin-dependent kinases to activate cell cycle. Correct recognition of cyclin proteins could provide key clues for studying their functions. However, their sequences share low similarity, which results in poor prediction for sequence similarity-based methods. Thus, it is urgent to construct a machine learning model to identify cyclin proteins. This study aimed to develop a computational model to discriminate cyclin proteins from non-cyclin proteins. In our model, protein sequences were encoded by seven kinds of features that are amino acid composition, composition of k-spaced amino acid pairs, tri peptide composition, pseudo amino acid composition, position, geary correlation, normalized moreau-broto autocorrelation and composition/transition/distribution. Afterward, these features were optimized by using analysis of variance (ANOVA) and minimum redundancy maximum relevance (mRMR) with incremental feature selection (IFS) technique. A gradient boost decision tree (GBDT) classifier was trained on the optimal features. Five-fold cross-validated results showed that our model would identify cyclins with an accuracy of 93.06% and AUC value of 0.971, which are higher than the two recent studies on the same data.

1. Introduction

Cyclin belongs to a group of proteins which are capable to control the cell cycle by triggering Cdk [1]. Cyclin concentration changes on different levels at several stages of the cell cycle. These changes occurred due to the ubiquitin-mediated cyclin degradation [2]. Cyclin combines with cyclin dependent kinases, like cdk1 proteins and p34, to trigger the cyclin dependent kinase active sites. This cdk1, p34 and cyclin combination forms a MPF (maturation-promoting factor) which activates other proteins [3]. However, phosphorylation is needed for the complete activation of cyclin dependent kinase active sites [3]. Therefore, these phosphorylated proteins are liable for the specific movements during the division of cell cycle e.g., chromatin remodeling and the formation of microtubules [3,4].

After the Human Genome Project (HGP), biological sequence data has progressively shattered [5]. The traditional investigational techniques have not only low efficient and expensive but also are time consuming. Therefore, it is urgent to identify sequences efficiently in a short period of time. However, existing tools such as FASTA [6] and BLAST [7] only compare the sequence with the known protein databases [8,9], these tools cannot discriminate whether it is a cyclin or non-cyclin. Now, machine learning classifications are popular in this area [10–13]. In prior methods, StAR [14] and other classifiers using Pseudo-amino acid composition (PseAAC) could identify cyclins with an accuracy of 83.53%. Sun et al. [15] established a cyclin prediction model based on support vector machine (SVM) which could produce an accuracy of 91.90%. Although both cyclin prediction model can produce good outcomes, there is still room for further improvement by extracting more feature information.

To address the aforementioned issues, an ensemble model was established to predict cyclin in multiple eukaryotic genomes. Fig. 1 shows the workflow of the proposed model. First, seven types of feature descriptors, Amino acid composition [16], Tri-peptide composition [17], Composition of K-spaced amino acid composition [18,19], Geary autocorrelation [20], Normalized moreau-broto autocorrelation [21], C/T/D [22] and PseAAC [23,24] were used as features to input into a GBDT classifier [25]. After this, ANOVA [26] and the mRMR [27] with IFS [28] technique was utilized to
get optimal feature vectors. The outcomes were evaluated by using five-fold cross validation.

2. Materials and methods

A reliable and accurate dataset is necessary to establish a prediction model [29–35]. Therefore, the dataset was obtained from Mohabatkar et al. [14]. They collected 215 cyclins and 204 non-cyclin proteins to train and test the methods for cyclins prediction. To reduce the overfitting derived from high similarity of sequences, we applied a cluster database at high identity with tolerance 90% [36] and discarded the sequences that exhibited more than 90% sequence identity. As a result, we attained the 167 cyclin and 167 non-cyclin proteins. Then we divided into 70/30 ratio in order to training and testing the model.

2.1. Feature descriptors

Selecting the feature-encodings that are instructive and autonomous is an important step in creating machine learning models [37–40]. Expressing the protein sequences with a mathematical formulation is key and difficult in functional element identification [37–40]. Therefore, seven types of feature-encoding approaches were presented to describe the protein sequences.

2.1.1. Amino acid composition descriptor (AAC)

AAC calculates the frequency of single type of amino acids in a protein sequence [16,45–50]. The frequencies $f(p)$ of 20 residues can be calculated as

$$f(p) = \frac{N(p)}{N} \cdot p \in \{ACDEFGHIKLMNPQRSTVWY\}$$ (1)

where $N(p)$ is the number of the $p$-th residue in a protein sequence with the length of $N$ residues.

2.1.2. Composition of k-spaced amino acid pairs descriptor (CKSAAP)

The encoding technique composition of k-spaced nucleic acid pairs embodies the incidence of nucleotide pairs disconnected by any K nucleotides ($K = 0, 1, 2, 3, 4, 5$). The CKSAAP [18,45,51] is defined as k-spaced residue pairs $Q_{xy}$, which is illustrated as

$$Q_{xy} = \frac{N_{xy}}{N(N-k)}$$ (2)

where $N_{xy}$ is the number of residue pairs and $k$ denotes the number of nucleotides. In this study, $k = 3$ and the dimension of the composition of k-spaced amino acid pairs feature was 1600.

2.1.3. Pseudo amino acid composition descriptor (PseAAC)

PseAAC describes the occurrence of the amino acid frequency and the correlation of between two residues’ physicochemical properties [23]. It consists of $A_c_i$ and $A_{c_0}$.

$$A_{c_i} = \frac{N_i}{1 + \alpha \sum_{i=1}^{20} \theta_i} \left( \frac{Q_i - Q_{i-d}}{N_q} \right)_{i=1, 2, 3 \cdots , 20}$$ (3)

$$A_{c_0} = \frac{\omega \times \theta_i}{1 + \omega \times \sum_{i=1}^{20} \theta_i}$$ (here, $\omega = 0.05$) (4)

where $N_q$ is the number of properties and $N_i$ is the $i$-th amino acid occurrence. $Q_i$ is the $i$th amino acid property value and $\theta_i$ is the sequence order factor.

2.1.4. Tri-peptide composition descriptor (TPC)

TPC are three amino acid molecules joined together and reflects hypothetically substantial starting points for the design of small biotic modulators [17]. Tripeptide composition is defined as

$$f_{lmn} = \frac{N_{lmn}}{N} (-I, m, n \in (A, B, C \cdots , Z))$$ (5)

where $N_{lmn}$ represents the number of tripeptide amino acid type $l$, $m$, $n$.

2.1.5. Composition|transition|distribution descriptor (C/T/D)

C/T/D defines the global composition of an amino acid sequence and the frequencies of two different adjoining amino acids and the distribution pattern of an amino acid sequence. Sequence scrambling is the first job to compute the composition, transition and distribution [52]. On the basis of their attributes, amino acids are alienated into three classes (class 1, class 2 and class 3) [24], also named reduced or simplified amino acids [53,54]. Classifications of charge and hydrophobicity are shown in Table 1. C/T/D with composition $A_c$, transition $T_b$ and distribution is defined as $D_{b,z}$.

| Table 1 | Attribute classification. |
|--------|--------------------------|
| C 1    | C 2                      | C 3 | Attributes              |
| R, K   | A, N, C, Q, H, I, S, L   | D, E| Charge                  |
| F, P, S, M | W, Y, V, T                |     |
| Polar  | Not +tive nor -tive      |     |
| D, E, K, N, Q, R | A, G, H, P, S, T, Y     | | Hydrophobic             |
| Hydrophobic | C, F, I, M, V, W     | | Hydrophobicity           |
\[ C_a = \frac{N_c}{N} (a = 1, 2, 3 \cdots) \]  
\[ T_b = \frac{N_{bc} + N_{cb}}{N-1} (b = 1, 2, 3 \cdots c \neq b) \]  
\[ D_{b2} = \frac{N_{b2}}{N} (b = 1, 2, 3 \cdots z = 1, 0.15N \cdots N) \]  

where \( N \) is the class number, \( N_{bc} \) is the adjoining number of class \( b \) and \( c \), \( N_{b2} \) is the number of those AA which are in \( z \)-th of \( b \)-th class.

2.1.6. Geary descriptor (GD)

Geary descriptor is a kind of correlation descriptor and have a maximum similarity with Moran descriptor [55]. It is well-defined as \( Q(r) \):
\[
Q(r) = \frac{N-1}{2 \times (N-r)} \times \sum_{i=1}^{N-r} \frac{(P_i - \bar{P})^2}{\sum_{i=1}^{N} (P_i - \bar{P})^2} (i = 1, 2, \ldots, 20)
\]

where \( P_i \) is the property value of the \( i \)-th amino acids in AA index.

2.1.7. Normalized moreau-broto autocorrelation descriptor (NMBroto)

NMBroto is also a type of autocorrelation [21] and have a likeness with Moran as shown in below equation.
\[
Q(r) = \frac{\sum_{i=1}^{N-r} (P_i \times P_{i+r})}{N-r} (i = 1, 2, \ldots, 20)
\]

where \( P_i \) is the property value of the \( i \)-th amino acids in AA index.

2.2. Feature selection

The noise in feature vector might result in the unsatisfactory performance of a model [56–63]. Therefore, the selection of features is an obligatory phase to remove the less important features and increase the productivity of a model [37,64–69]. Many feature selection and ranking techniques are available, such as ANOVA, F-score [70], mRMR [27], Chi-square [71], LGBM [72,73]. A high feature dimension also both can create overfitting and information redundancy and produce poor accuracy of the cross-validation prediction. Therefore, ANOVA is good option to tackle these issues because it consumes less time and gave efficient results. The combination of some of the top-executing features does not mean that the top predictive results can be attained. These features are probably to have a high degree of correlation, which leads to additional redundant information in the feature vectors. Therefore, mRMR is a good option to tackle these issues due to less time consuming and efficient results. These techniques are also used in many high dimensional protein features selection. In this study, the ANOVA and mRMR [27] with IFS [56] was applied to obtain the optimal feature subset. The comparison with other state of the art feature selection techniques is given in Fig. 25 in Supplementary file 1.

2.2.1. ANOVA

ANOVA is used for significance test of mean difference between two or more samples. \( F \)-value is the ratio of variance between groups and variance within groups [74]. If the \( F \)-value will be larger, then the ability of distinguishing positive and negative samples will be better. Therefore, all features can be sorted according to this \( F \)-value.
\[
Q_m^2(\xi) = \frac{\sum_{i=1}^r l_i (\overline{x}_i - x)^2}{df_m}
\]
\[
Q_n^2(\xi) = \frac{\sum_{i=1}^r \sum_{j=1}^r (x_{ij} - x)^2}{df_n}
\]

where \( P(\gamma_i) \) is the risk minimization parameter of the new decision tree which is shown in below equation.
\[
\delta_k = \text{argmin} \sum_{i=1}^n L(y_i, f_{k-1}(x_i) + T(x; \theta_k)) : \text{Listhelossfunction}
\]

Gradient boost decision tree algorithm calculates the final assessment in a forwarding mode.
\[
f_0(x) = f_{k-1}(x) + T(x; \theta_k)
\]
\[
\text{Finally, Loss function } f_{k-1}, \text{ of negative gradient is used for residual calculation.}
\]
At the end, we trained the model by all \( R_{ki} \) to calculate the risk minimization parameter \( h_k \). This type of decision trees logically models the relations amongst predictor variables, e.g., mapping the parameters input space \( X \) in to \( J \) split sections \( R_1 \ldots R_J \), and the output is \( Z_j \) for region \( R_j \).

\[
T(x; \theta) = \sum_{j=1}^{J} z_j I(x \in R_j)
\]

The pseudo code of gradient boost decision tree is given below in Algo 1.

**Algo 1: Gradient Boosting Decision Tree Algorithm**

**Input:** Training Data: \( (x_i, y_i)_{i=1}^n \)

Where, \( x_i \) is a data point and \( y_i \) is the label for \( x_i \)

Loss function: \( L(y, f(x)) \)

1. Initialize the model \( f_1(x) = \arg\min_{f(x)} \sum_{i=1}^{n} L(y_i, f(x)) \)

2. for \( k = 1, 2, 3 \ldots K \) Do

3. for \( l = 1, 2, 3 \ldots n \) Do

4. By calculating the Pseudo residual error:

\[
R_{ki} = \frac{\partial L(y_i, f(x))}{\partial f(x)}
\]

5. End

6. End

7. By constructing a new Decision Tree \( T_k(x; \theta_k) \), based on \( R_{ki}, \theta_k = \{R_{kj} = [1, 2, 3 \ldots J]\} \)

8. for \( j = 1, 2, 3 \ldots J \) Do

9. \( z_{kj} = \arg\min_{\theta_0} \sum_{x \in R_{kj}} L(y, f_{k-1}(x) + z) \)

10. End

11. Updating the model \( f_k(x) = f_{k-1}(x) + \sum_{j=1}^{J} z_{kj} I(x \in R_{kj}) \)

12. \( f(x) = \sum_{k=1}^{K} \sum_{j=1}^{J} z_{kj} I(x \in R_{kj}) \)

**Output:** The decision tree function \( f(x) \)

Scikit-learn package (v = 0.22.1) [91] was used to execute the random forest classifiers. Firstly, we used randomized search cross-validation and then grid search cross-validation to tune hyperparameter. The best tuned parameters of the proposed model are given in Table 2.

| Table 2 | Best parameters of the proposed model. |
|---------|----------------------------------------|
| Best Parameters | \( 'Max-depth' \) | 20 |
| | \( 'Max-features' \) | 05 |
| | \( 'Min-samples-leaf' \) | 03 |
| | \( 'Learning-rate' \) | 0.05 |
| | \( 'Min-samples-split' \) | 02 |
| | \( 'N-estimators' \) | 80 |
| | \( 'Mean square error' \) | 0.1287 |

Scikit-learn package (v = 0.22.1) [91] was used to execute the random forest classifiers. Firstly, we used randomized search cross-validation and then grid search cross-validation to tune hyperparameter. The best tuned parameters of the proposed model are given in Table 2.

**Fig. 2.** Plot showing the Incremental Feature Selection (IFS) procedure for identifying Cyclins in 5-fold cross-validation. (A) Firstly, 5711 features were selected from a total of 10,200 features by ANOVA. (B) 304 optimal features were further obtained from the 5711 features by using mRMR. The Acc increases from 88.92% to 93.06%. (C) Feature descriptor contribution in GBDT-based fusion model to predict cyclins. (D) Comparison between single-encodings and fusion features on different machine learning classifiers.
3. Results and discussion

3.1. Performance evaluation

First, the training data were converted into feature vectors using feature descriptors (amino acid composition, composition of k-spaced amino acid pairs, tri peptide composition, pseudo amino acid composition, geary correlation, normalized moreau-broto autocorrelation and composition/transition/distribution), and the feature vectors of each encoding model were evaluated by gradient boost decision tree algorithm using a five-fold CV test. Firstly, the ANOVA and mRMR with IFS were used to pick the best feature subset for the sake of better prediction accuracy. Fig. 2(A) and (B) shows the incremental feature selection curve of optimal features and comparison of single encodings and fusion on different machine learning classifiers on the basis of AUC. Table 3 shows the efficiency of the optimized single-encoding models and the feature fusion model on different machine learning methods. The performance of single-encoding models and the fusion model on different machine learning classifiers before feature selection is recorded in Table 1S in Supplementary file 1. We also visualized the single-encoding features and fusion features using t-SNE (t-distributed Stochastic Neighbor Embedding) method before and after feature selection. The t-SNE visualization of single-encoding and fusion before feature selection is available in Fig. 15 in Supplementary file 1 and the t-SNE visualization of the optimized single-encodings and the fusion is shown in Fig. 3. The AUCs of single-encoding models are 0.827, 0.526, 0.825, 0.506, 0.896, 0.854, and 0.890, respectively for AAC, CKSAAP, PseAAC, TPC, C/T, D, GD, and NMBroto. The AUC of composition/transition/distribution was around 0.6% ~ 39% higher as compared with those of the other encodings. On the contrary, the Acc, Sp, Sn, MCC, and AUC of the feature fusion model were 93.06%, 94.00%, 92.00%, 0.862% and 0.971, respectively. The Acc, Sp, Sn, MCC, and AUC on independent data were 89.36%, 90.10%, 89.45% and 0.823%. ROC with the AUC of 0.954 is given in Fig. 35 in Supplementary file 1. In order to check the better performance and reliability of our model, we further randomly extracted 50 non-cyclin sequences from the public databases and checked the performance by running our model. We found quite reasonable results. The Accuracy, specificity, sensitivity and matthews correlation coefficient were 90.05%, 91.11%, 89.45%, and 0.829.

### Table 3

Performance of optimized single-encodings and fusion models on different machine learning classifiers.

| Descriptor | Acc | Sp | Sn | MCC | AUC |
|------------|-----|----|----|-----|-----|
| AAC        | 76.34 | 74.70 | 79.60 | 0.528 | 0.827 |
| CKSAAP     | 82.33 | 81.40 | 83.80 | 0.647 | 0.896 |
| Geary      | 78.74 | 77.30 | 81.40 | 0.576 | 0.854 |
| NMBroto    | 76.04 | 74.90 | 78.40 | 0.522 | 0.825 |
| PAAC       | 50.89 | 80.00 | 24.00 | 0.074 | 0.506 |
| Fusion All | 92.35 | 92.55 | 92.11 | 0.971 | 0.956 |

### 2.4. Evaluation metrics

Sensitivity (Sn), specificity (Sp), accuracy (Acc), and matthews correlation coefficient (MCC) [92–106] were used in this study to check the overall efficiency of the model defined as Equation (23).

\[
\begin{align*}
Sn &= \frac{TP}{TP + FN} \\
Sp &= \frac{TN}{TN + FP} \\
Acc &= \frac{TP + TN}{TP + TN + FP + FN} \\
MCC &= \frac{(TP \times TN - FP \times FN)^{\frac{1}{2}}}{(TP + FN)(TN + FP)(TP + TN)(FP + TN))^{\frac{1}{4}}}
\end{align*}
\]

where TP represents the overall cyclins sequences in benchmark data and FP signifies the cyclins sequences false-classified as non-cyclins. Likewise, TN represents the overall non-cyclins sequences in the data and FN signifies the non-cyclin sequences, which were false-classified as cyclins. Consequently, the receiver operating characteristic (ROC) curve was used to illustrate the efficiency of the model graphically. The ROC curvature could assess the projecting ability of the proposed model on the whole assortment of resultant values. The area under the curve (AUC) was premeditated to check the efficiency of the model. A good classifier gave $AUC = 1$, and the arbitrary performance gave $AUC = 0.5$. 

H. Zulfiqar, Shi-Shi Yuan, Qin-Lai Huang et al. Computational and Structural Biotechnology Journal 19 (2021) 4123–4131
3.2. Performance evaluation of different ML algorithms

Single-encoding AAC, CKSAAP, PseAAC, TPC, C/T/D, GD, NMBroto and feature fusion models were inputted into different machine learning classifiers such as Ada boost, SVM, and Naive bayes algorithm. Their performances were compared with that of gradient boost decision tree classifier-based models. A five-fold cross-validation test was used to evaluate these model performances. Results were shown in Table 3. We may notice that the accuracies of feature fusion models were always higher than those of single-encoding models, indicating that the multiple information was effective to achieve better results. Fig. 2 (C) showed the feature descriptor contribution in GBDT-based fusion model. The optimized fusion model consists of 304 features of seven descriptors. AAC descriptor contributed 3.28 % in final fusion model because their 10 features were participated in the fusion model. CKSAAP descriptor contributed 16.11 % in final model because their 49 features were participated in the fusion model. CTD descriptor contributed 13.15 % in final model because their 40 features were in the final fusion model. Geary descriptor contributed 32.89 % because their 100 features were participated in the fusion model. NMBroto descriptor contributed 26.31 % in the final optimized model due to their 80 best features. PAAC contributed 4.93 % in the model with their 15 features and TPC contributed 3.28 % in the final optimized model with their best 10 features. Fig. 2 (D) exhibited that the GBDT-based fusion model performed best among all methods. Particularly, the AUC of GBDT classifier was almost 3.5% – 17.7% higher than that of the other models, indicating that the GBDT-based model was the best for cyclin identification.
3.3. Comparison with existing models

In recent studies, Mohabatkar et al. [14] and Sun et al. [15] used the similar dataset for training their models by using jackknife cross-validation. The accuracies of their models were 83.53% and 91.90%, respectively. We also used the same dataset and applied GBDT algorithm. Results on jackknife cross-validation and five-fold cross-validation showed that our model is better than the two existing models. The comparison of two existing models with our model has been shown in Table 4 and Fig. 4.

4. Conclusions

Cyclin proteins are capable to regulate the cell cycle and forms a complex with cyclin-dependent kinases. This complex activates cell cycle but the full activation requires phosphorylation. Cyclin protein have low similarity between their sequences. To date, numerous predictors have been established to classify cyclins in diverse species [14,15,107]. In this study, an advanced ensemble model was developed to identify cyclins. In the proposed model, protein sequences were encoded by using AAC, CKSAAP, PseAAC, TPC, C/T/D, GD, and NMBroto. Then, these encoding-features were optimized by using ANOVA and mRMR with IFS technique. On the basis of top feature subset, the finest scoring model was achieved by the gradient boost decision tree classifier using five-fold CV test. The estimated outcomes on training data showed that the proposed model provided outstanding generalization capability. The data and codes are also available in the Supplementary file 2. Further studies will aim to create a user-friendly web server for the projected model. Also, additional feature selection methods and algorithms will be implemented to further improve the efficiency to classify cyclins [108–117].

Declaration of Competing Interest

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (61772119), Sichuan Provincial Science Fund for Distinguished Young Scholars (2020DJQ0012).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.07.013.

References

[1] Gallerisi U, Jori FP, Giordano A. Cell cycle regulation and neural differentiation. Oncogene 2003;22(33):5208–19.
[2] Morgan DO, The cell cycle: principles of control. 2007: New science press.
[3] Ferby I, Blazquez M, Palmer A, Eritja R, Nebreda AR. A novel p34cdc2-binding and activating protein that is necessary and sufficient to trigger G2/M progression in Xenopus oocytes. Genes Dev 1999;13(16):2177–89.
[4] Robinson DR, Gull K. Basal body movements as a mechanism for mitochondrial genome segregation in the trypomastigote cell cycle. Nature 1991;352(6337):731–3.
[5] Lee TF, The Human Genome Project: Cracking the genetic code of life. 2013: Springer.
[6] Pearson WR, Finding protein and nucleotide similarities with FASTA. Current protocols in bioinformatics, 2016. 53(1): p. 3.9. 1-3.9. 25.
[7] Madden T, The BLAST sequence analysis tool, In The NCBI Handbook [Internet]. 2nd edition, 2013, National Center for Biotechnology Information (US).
[8] Xu Baofang, Liu Dongyang, Wang Zerong, Tian Ruixia, Zulfiqar Hasan, Wang Xiao-Feng, et al. Computational and Structural Biotechnology Journal 2021;19:4123–4131.
[9] Zulfiqar Hasan, Sun Jia-Nan, Yang Hua-Yi, Yao Jing, Ding Hui, Han Shu-Guang, Wu Cheng-Yan, et al. Screening of Prospectve Plant Compounds as H1R and CL1R inhibitors and its antiallergic efficacy through molecular docking approach, Comput Math Methods Med 2021;2021:1–9.
[10] Zhang Ziding, et al. Prediction of ubiquitination sites by using reduced amino acid composition descriptor. Evolutionary Bioinformatics, 2019, 15: p. 117693419867088.
[11] Wu Jianping, Aluko Rotimi E. Quantitative structure-activity relationship studies of bitter di- and tri-peptides including relationship with antinociceptive l-enzyme converting inhibitory activity. J Peptide Sci 2007;13(1):63–9.
[12] Chen Zhen, Chen Yong-Zi, Wang Xiao-Feng, Wang Chuan, Yuan Shi-Shi, Huang Qin-Lai et al. Prediction of cyclin protein using two-step feature selection technique. IEEE Access 2020;8:109535–42.
[13] Sun Jia-Nan, Yang Hua-Yi, Yao Jing, Ding Hui, Han Shu-Guang, et al. Prediction of cyclin protein and related amino acid sequence features. Genomics, Proteomics 2020;18:1084–91.
[14] Yang Yu-He, Ma Chi, Wang Jia-Shu, Yuan Shi-Shi, Ding Hui, Han Shu-Guang, et al. Prediction of N7-methylguanosine sites in human RNA based on optimal sequence features. Genomics, Proteomics 2021;19(6):4342–7.
[15] Mohabatkar H. Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept Lett 2010;17(10):2017–4.
[16] Su Wei, Liu Meng-Lu, Yang Yu-He, Wang Jia-Shu, Li Shi-Hao, Lv Hao, et al. Prediction of cyclin and cyclin-dependent kinase protein sequences by using AAC, CKSAAP, PseAAC, TPC, C/T/D, GD, and NMBroto. IEEE Access 2018;6:22930–3.
[17] Chen Wei, Peng Pangmin, Nie Fulei, IATP: A sequence based method for identifying anti-tubercular peptides. Med Chem 2020;16(5):620–5.
[18] Sokal RR, Thomson BA. Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. Am J Phys Anthropol 2006;129(1):121–31.
[19] Horne David S. Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicity. Biopolymers 1991;33(3):451–77.
[20] Cai CZ, Han LY, Ji ZL, Chen YZ. Enzyme family classification by support vector machine protein interactions: Feature extraction and machine learning approaches. Comput Struct Biotechnol J 2020;18:66–76.
[21] Chou Kuo-Chen. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins Struct Funct Bioinf 2001;43(3):246–55.
[22] Su Wei, Liu Meng-Lu, Yang Yu-He, Wang Jia-Shu, Li Shi-Hao, Lv Hao, et al. PPD: a manually curated database for experimentally verified prokaryotic promoters, J Mol Biol 2019;431(11):166860. https://doi.org/10.1016/j.jmb.2021.166860.
[23] Ning L et al., MNDR v3.0: mammal ncRNA-disease repository with increased coverage and annotation. Nucleic Acids Res, 2021. 49(D1): p. D160-d164.
[24] Zuo Yongchun, Li Yuan, Chen Yingli, Li Guangpeng, Yang Zeng, et al. PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. Bioinformatics 2017;33(1):122–4.
[25] Su G et al. Lightgbm: a high efficiency gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30:3146–54.
[26] Tang Hua, Zhao Ya-Wei, Zou Ping, Zhang Chun-Mei, Chen Rong, Huang Po, et al. HBPre: a tool to identify growth hormone-binding proteins. Int J Biol Sci 2018;14(8):557–64.
[27] De Jay N et al., mRMR-e: An package for paralleled mRMR ensemble feature selection. Bioinformatics, 2013. 29(18): p. 2365-2368.
[28] Wang Ziru, Xing Jian, Huang Jian, Ding Hui, Liu Hao. A Brief Survey of Machine Learning Methods in Protein Sub-Golgi Localization. Curr Bioinform 2019;14(3):234–40.
[29] Su Wei, Liu Meng-Lu, Yang Yu-He, Wang Jia-Shu, Li Shi-Hao, Lv Hao, et al. PPD: a manually curated database for experimentally verified prokaryotic promoters. J Mol Biol 2019;431(11):166860. https://doi.org/10.1016/j.jmb.2021.166860.
[30] Ning L et al., MNDR v3.0: mammal ncRNA-disease repository with increased coverage and annotation. Nucleic Acids Res, 2021. 49(D1): p. D160-d164.
[31] Liang ZY et al. PseDR: a database for experimentally verified sigma-54 promoters. Bioinformatics 2017;33(3):467–9.
[32] Hong Z et al. Identifying enhancer–promoter interactions with neural network based on pre-trained RNA vectors and attention mechanism. Bioinformatics 2020;36(4):1037–43.
[33] Zeng X et al., deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics, 2019. 35(24): p. 5191-5198.
[34] Yu Liang, Wang Meng, Yang Yang, Xu Fengdan, Zhang XU, Xie Fei, et al. Predicting therapeutic potential for hepatocellular carcinoma based on tissue-specific pathways. PLoS Comput Biol 2021;17(2):e1008696.
[35] Zhao Xudong, Jiao Qing, Li Hangyu, Wu Yiming, Wang Hanxu, Huang Shan, et al. UCSAAE-DF: an amino acid classifier-based feature selection for differential expression analysis on expression profiles. BMC Bioinf 2020;21(1):43.
[36] Fu L et al., CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 2012. 28(23): p. 3150-3152.
[37] Zhang Nan, Wang Kairui, Zhan Weihua, Deng Lei. Targeting virus-host protein interactions: Feature extraction and machine learning approaches. Curr Drug Metab 2019;20(3):177–84.
[38] Zeng S et al., Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. Briefings in bioinformatics, 2020. 21(4): p. 1425-1436.
Yu Liang, Zhou Dandan, Gao Lin, Zha Yunhong. Prediction of drug response in multilayer networks based on fusion of multiomics data. Methods (San Diego, Calif.) 2021;192:85–92.

Charoenkwan Phasit, Kanthawong Sakawrat, Nantasenamat Chanin, Hasan Md Meheri, Shoombuatong Watshara. iDPPIV-SCM: a sequence-based predictor for identifying and analyzing dipeptidyl peptidease IV (DPP-IV) inhibitory peptides using a scoring card method. J Proteome Res 2020;19(10):4125–36.

Charoenkwan Phasit, Yana Janchai, Nantasenamat Chanin, Hasan Md Meheri, Shoombuatong Watshara. iUmami-SCM: a novel sequence-based predictor for prediction and analysis of umami peptides using a scoring card method with propensity scores of dipeptides. J Chem Inf Model 2020;60(12):6666–78.

Wang G, et al., MedReaders: a database for transcription factors that bind to methylated DNA. Nucleic Acids Res, 2018. 46(D1): p. D146-D151.

Stephenson Natalie, Shane Emily, Chase Jessica, Rowland Jason, Ries David, Justice Nicola, et al. Survey of machine learning techniques in drug discovery. Curr Drug Metab 2019;20(3):185–93.

Cao Renzhi, Freitas Colton, Chan Leong, Sun Xiao, Jiang Haiqing, Chen Zhangxin. Protein function prediction using neural machine translation based on a recurrent neural network. Molecules 2017;22(10).

Kalita Mridul K, Nandal Umesh K, Pattnaik Ansuman, Sivalingam Anandhan, Ramasamy Gowthaman, Kumar Manish, et al. CyclinPred: a SVM-based method for predicting cyclin protein sequences. PLoS ONE 2008;3(7):e2605.

Lv Z, et al., Anticancer peptides prediction with deep representation learning features. Briefings in bioinformatics, 2021.

Ahmad Fareed, Farooq Amjad, Ghani Khan Muhammad Usman, Shabbir Muhammad Zubair, Rabban Masood, Hussain Irshad. Identification of most relevant features for classification of franciella tularensis using machine learning. Curr Bioinform 2021;15(10):1197–212.

Amanat Saba, Ashraf Adeel, Hussain Waqar, Rasool Nouman, Khan Yaser D. Identification of lysine carboxylation sites in proteins by integrating statistical moments and position relative features via general PseAAC. Curr Bioinform 2020;15(5):396–407.

Ayachit Garima, Shaikh Inayatullah, Pandya Himanshu, Das Jayashankar. Salient Features, Data and Algorithms for MicroRNA Screening from Plants: A Review on the Gains and Pitfalls of Machine Learning Techniques. Curr Bioinform 2021;15(10):1051–103.

Kong Liang, Zhang Lichao, He Shiqian. Improving multi-type gram-negative bacterial secreted protein prediction via protein evolutionary information and feature ranking. Curr Bioinform 2020;15(6):538–46.

Li Hong-Dong, Zhang Wenjing, Luo Yuwen, Wang Jianxin. IsoDetect: detection of splice isoforms from third generation long reads based on short feature sequences. Curr Bioinform 2021;15(10):1168–77.

Zhang Ge, Yu Pan, Wang Jianlin, Yao Chaokun. Feature selection algorithm for high-dimensional biomedical data using information gain and improved chemical reaction optimization. Curr Bioinform 2021;15(8):912–26.

Zhang Tianjiao, Wang Rongjie, Jiang Qinghua, Wang Yadong. An information gain-based method for evaluating the classification power of features towards identifying enhancers. Curr Bioinform 2020;15(6):574–80.

Hasan Md Meheri, Manavalan Balachandran, Khatun Mst Shamima, Kurata Hiroyuki. i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. Int J Biol Macromol 2020;157:752–8.

Hasan MM, et al., Meta-i6mA: an interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. Brief Bioinform, 2020.