ABSTRACT  Cyclins are a family of proteins that regulate the cell cycle by activating cyclin-dependent kinases or a group of enzymes required in the cell cycle. Constructing a model to classify Cyclins is of importance to understand their function. It is urgent to construct a machine learning based model to identify Cyclins because of low similarity between the sequence of Cyclins. In this study, a method based on support vector machine (SVM) is developed to recognize Cyclins only using amino acid sequence information. 18 feature descriptors with a total of 13151-dimension features were extracted, and the feature dimension were reduced to 8 through feature selection technique. The reserved features show some of feature descriptors such as Autocorrelation, AAC and CTDC are important in the identification of Cyclins. Jackknife cross-validation results indicate our model would classify Cyclins with an accuracy of 91.9%, which is superior to a recent study using the same data set. Our work provides an important tool for discriminating Cyclins.

INDEX TERMS  Cyclin, support vector machine, identification, feature selection, sequence information.

I. INTRODUCTION  Cyclins are a family of proteins that regulate the cell cycle by activating cyclin-dependent kinases or a group of enzymes required in the cell cycle [1]. At different stages of the cell cycle, concentrations of most cyclins alter on some level. Alterations in concentration depend on two mechanisms. One is the floating of cyclin gene expression. The other is the ubiquitin-mediated cyclin degradation pathway [2].

Cyclins can form a complex with cyclin-dependent kinases. Once the complex is phosphorylated, the cdk-active site is activated. The activated complexes in turn play a role in the cell cycle [2]. For example, the complex of Cdk1 (Cdc2a) and cyclin CycBlzm2 mediate mitosis and cytokinesis in root tip cells of corn [3]. Combining with cdk, cyclin can also generate maturation promoting factor (MPF). MPF is the trigger of phosphorylation of some proteins, which contributes to specific events of cell cycle, such as the assembly of microtubules and chromosomal reorganization [4]. Cyclin in these complexes does not have an enzyme-active site, but has a substrate binding site, and can also locate cdks to a specific subcellular structure [2].

In the post-genomic era, biological data, especially sequence data, has increasingly exploded [5]–[8]. But the traditional experimental method not merely costs a long time and a lot of money, but also has a relatively lower rate of success [9]–[13]. Therefore, a quick way is in urgent demand to identify sequences. Although existing calculation methods such as BLAST and FASTA can compare the new sequence with known proteins in common databases, these methods cannot distinguish whether it is a cyclin. Machine learning based classification is becoming mainstream in this area [14]–[25]. In the previous method, StAR [26] and other classifiers using PseAAC could effectively identify cyclins with an accuracy rate of 83.53%. Kalita et al. [16] set up a cyclin prediction server – CyclinPred based on SVM model which offered several prediction strategies. The same data set was used in our work with an improved accuracy of 94.3% through different methods. And the accuracy of Jackknife cross-validation...
was 91.9%. We extracted a total of 18 different feature descriptors, reducing the feature dimensions to 8 after two screenings. This method can be served as a reference in future related research.

II. MATERIALS AND METHODS

A. THE BENCHMARK DATASET

As mentioned above, the data set used in this study is the same as that of Mohabatkar’s work [26], which includes 215 cyclins and 204 acyclic proteins. The training data set has a greater impact on the trained model. Too similar samples may cause overfitting of the proposed model [27]. Through BLASTCLUST, the redundant sequences were removed with a similarity of 90%. 167 cyclins and 167 acyclic proteins may cause overfitting of the proposed model [27]. Through BLASTCLUST, the redundant sequences were removed with a similarity of 90%. 167 cyclins and 167 acyclic proteins remained, which were used to train and test our proposed model. A benchmark dataset with low sequence identity such as 25% could produce reliable results. However, this study did not use such a stringent criterion because number of samples would be too few to have statistical significance.

B. FEATURE EXTRACTION STRATEGIES

Features are the key elements for machine learning-based protein classification issues [28]–[41]. Selecting reasonable features can effectively improve the accuracy of the classifier and the speed of model generation [42]–[45]. The following section will introduce the features utilized in this work.

1) SEQUENCE-ORDER-COUPLING NUMBER (SOCNUMBER)

Sequence-order-coupling number (SOCNumber) reflects the effects of sequence order. The \( d \)-th ranked SOCNumber is defined as \( \tau_d \) at Eq. (1).

\[
\tau_d = \sum_{i=1}^{N-d} (d_{i,i+d})
\]

here \( d_{i,i+d} \) is the distance between the two residues at position \( i \) and \( i + d \). The distance of any two residues is defined as Schneider-Wrede physicochemical distance matrix [46].

2) QUASI-SEQUENCE-ORDER DESCRIPTOR (QSORDER)

Quasi-sequence-order descriptor (QSOrder) also reflects the effects of sequence order [47]. For each amino acid type, a quasi-sequence-order is defined as \( x_r \) at Eq. (2).

\[
x_r = \frac{f_r}{\sum_{r=1}^{20} f_r + \omega \sum_{d=1}^{\text{maxlag}} \tau_d}, \quad r = 1, 2, \ldots, 20
\]

where \( f_r \) is the normalized occurrence for amino acid type \( i \) and \( w \) is a weighting factor (\( w = 0.1 \)).

3) MORAN AUTOCORRELATION DESCRIPTOR

Moran autocorrelation descriptor reflects the interaction between proteins and proteins [48]. Moran autocorrelation is defined as \( I(d) \) at Eq. (3).

\[
I(d) = \frac{1}{N-d} \sum_{i=1}^{N-d} (P_i - \bar{P'}) (P_{i+d} - \bar{P'})
\]

\[
d = 1, 2, \ldots, 30
\]

\[
\bar{P'} = \frac{1}{N} \sum_{i=1}^{N} P_i
\]

In the equation, \( P \) means the amino acid value in a property entry of AAIindex. BHAR880101 [49], CIDH920105 [50] and CHAM810101 [51] are property entries, the meanings of which are average flexibility indices, normalized average hydrophobicity scales and steric parameter respectively. \( \bar{P'} \) means the mean property of an amino acid sequence.

4) COMPOSITION\(\backslash\)TRANSITION\(\backslash\)DISTRIBUTION(C\(\backslash\)T\(\backslash\)D)

Composition\(\backslash\)transition\(\backslash\)distribution (C\(\backslash\)T\(\backslash\)D) describes the global composition of an amino acid sequence, the frequencies of two different contiguous amino acids and the distribution pattern of an amino acid sequence. To calculate CTD of a sequence, the first task is to encode sequences [52]. The amino acids are divided into three classes according to their attributes [53]. And each amino acid is encoded by one of the indices (1, 2, 3) according to which class it belongs to. Hydrophobicity and charge are the other two different classifications, which are shown in Table 1. CTD including composition \( c_r \), transition \( t_i \) and distribution is defined as \( d_{i,m} \) and shown as in Eq. (4).

\[
c_r = \frac{N_r}{N}, \quad r = 1, 2, 3
\]

\[
t_i = \frac{N_{i,j} + N_{j,i}}{N-1}, \quad i = 1, 2, 3j \neq i
\]

\[
d_{i,m} = \frac{N_{i,m}}{N}, \quad i = 1, 2, 3m = 1, 0.25N, \ldots, N
\]

where \( N_r \) is the number of a class. \( N_{i,j} \) is the number of contiguous \( i \)-class and \( j \)-class. \( N_{i,m} \) is the number of amino acids within \( m \)-th of \( i \)-th class along the sequence.

5) AMINO ACID COMPOSITION (AAC)

Amino acid composition (AAC) has been widely applied in protein classification [54]–[57] and here it is defined as the frequency of one in 20 amino acids.

---

**TABLE 1. The classification of hydrophobicity and charge.**

| Hydrophobicity | Class 1 | Class 2 | Class 3 |
|---------------|---------|---------|---------|
| Polar         | R, K, E, D, Q, N | G, A, S, T, P, H, Y | Neutral |
| Neutral       | A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V | C, L, V, I, M, F, W | Negative |

---

**J.-N. Sun et al.: Prediction of Cyclin Protein Using Two-Step Feature Selection Technique**

VOLUME 8, 2020
6) COMPOSITION OF K-SPACED AMINO ACID PAIRS (CKSAAP)
The composition of k-spaced amino acid pairs (CKSAAP) reflects short-range correlation of residues within an amino acid sequence. The CKSAAP is defined as k-spaced residue pairs \(C_{XY}\) in the fragment, which is illustrated as Eq. (5).

\[
C_{XY} = \frac{N_{XY}}{N-k}
\]

where \(X, Y\) mean a type of amino acid, \(N_{XY}\) means the occurrence of dipeptide \(XY\). Generally, \(k\) is always taken 0, 1, 2, 3, 4, 5.

7) DIPEPTIDE COMPETITION (DPC) AND ITS DEVIATION
Dipeptide competition (DPC) encapsulates the global information of an amino acid sequence \([58], [59]\). DPC is defined as \(D_{c(i)}\) at Eq. (6).

Dipeptide deviation from expected mean (DDE) is used to compute the deviation of dipeptide frequencies \([60]\), which is determined by three parameters in Eq. (6).

\[
\begin{align*}
T_{M(i)} &= \frac{C_{i1}}{61} \times \frac{C_{i2}}{61} \\
D_{c(i)} &= \frac{n_i}{N-1} \\
T_{V(i)} &= \frac{T_{M(i)} (1 - T_{M(i)})}{N-1} \\
DDE_i &= \frac{T_{M(i)} - D_{c(i)}}{\sqrt{T_{V(i)}}}
\end{align*}
\]

where \(T_{M(i)}\) means theoretical mean of the codon numbers of amino acids. \(C_{i1}, C_{i2}\) mean, for a given dipeptide \(i\), the number of codons for the first amino acid and the second. \(D_{c(i)}\) represents the dipeptide composition. And \(n_i\) means occurrence of \(i\)-th dipeptide. \(T_{V(i)}\) represents the theoretical variance of the dipeptide \(i\). \(N\) is the length of the amino acid sequence. DDE is computed as \(DDE_i\).

8) TRIPLETT COMPOSITION (TPC)
Tripeptide composition (TPC) reflects potentially significant starting points for the design of small molecule biological modulators. TPC is defined as \(f_i\) at Eq. (7). \(N_i\) means the number of \(i\)-th tripeptide.

\[
f_i = \frac{N_i}{N-2}
\]

9) GROUPED AMINO ACID COMPOSITION (GAAC) AND ITS EXTENSION
Grouped amino acid composition (GAAC) reflects the physicochemical properties \([61]\). 20 amino acids are grouped into 5 groups by their physicochemical properties (Table 2.). GAAC is computed as the composition of each group.

The composition of k-spaced amino acid group pairs (CKSAAGP) is the combination between CKSAAP and GAAC.

The grouped dipeptide composition (GDPC) is defined as the combination between DPC and GAAC.

### Table 2. The groups of GAAC.

| Physicochemical properties | Amino acid |
|----------------------------|-----------|
| Alphatic                   | G, A, V, I, L, I |
| Aromatic                   | F, Y, W |
| Postivecharge              | K, R, H |
| Negativecharge             | D, E |
| Uncharge                   | S, T, C, P, N, Q |

Similarly, the grouped tripeptide composition (GTPC) is the combination between TPC and GAAC.

10) NORMALIZED MOREAU-BROTO AUTOCORRELATION DESCRIPTOR
Normalized Moreau-Broto autocorrelation (NMBroto) is widely used in studies \([62]\), which resembles Moran. NMBroto is defined as \(I(d)\) at Eq. (8) \([63]\).

\[
I (d) = \frac{\sum_{i=1}^{N-d} (P_i \times P_{i+d})}{N-d} \quad d = 1, 2, \ldots, 30
\]

where \(P_i\) is the value of \(i\)-th amino acid in a property entry of AAIndex.

11) GEARY AUTOCORRELATION DESCRIPTOR
Geary is another autocorrelation descriptor, which resembles Moran \([64]\). Geary is defined as \(I(d)\) at Eq. (9).

\[
I (d) = \frac{N - 1}{2 \times (N - d)} \times \frac{\sum_{i=1}^{N-d} (P_i - P_{i+d})^2}{\sum_{i=1}^{N} (P_i - d)^2}
\]

where \(P_i\) is the value of \(i\)-th amino acid in a property entry of AAIndex.

12) CONJOINT TRIAD
CTriad is proposed to describe the information of protein-protein interactions. It regards any three sequential amino acids as a unit and computes their properties \([65]\). To calculate CTriad, amino acids need to be categorized to 7 classes. So there are \(7 \times 7 \times 7\) different units. Compute the occurrence of each unit as \(f_i (i = 1, 2, \ldots, 343)\). CTriad is defined as \(I(i)\) at Eq. (10).

\[
I (i) = \frac{f_i - \min \{f_1, \ldots, f_{343}\}}{\max \{f_1, \ldots, f_{343}\}}
\]

13) GENERAL PSEUDO-AMINO ACID COMPOSITION (PAAC)
General pseudo-amino acid composition (PAAC) reflects not only the occurrence frequency of each amino acid but also the sequence order effect of an amino acid sequence \([66]\). PAAC consist of \(Xc(i)\) and \(Xc_{\lambda(i)}\). PAAC is defined as Eq. (11).

\[
\left\{ \begin{array}{l}
\theta_i = \frac{\sum_{i=1}^{N-d} (P_i - P_{i+d})^2}{N_p} \\
Xc (i) = \frac{N_i}{1 + \omega \times \sum_{i=1}^{30} \theta_i} \\
Xc_{\lambda(i)} = \frac{1 + \omega \times \sum_{i=1}^{30} \theta_i}{Xc (i)}
\end{array} \right. \quad i = 1, 2, \ldots, 30
\]
100-dimension features, it would require 2 computing resources would be exhausted. Suppose that there used for data. But dealing with high-dimension features, model generation.

After the features were extracted, we removed the redundant features. In this work, iLearn was used to extract all these features.

Different feature descriptors represent different aspects of the physicochemical properties of an amino acid sequence. In order to fully integrate these different properties, this study extracted enough features (Figure 1.) for subsequent screening.

iLearn [68] is a python programmed tool that can help researchers perform feature extraction and subsequent screening. iLearn is known for its full functionality and ease of use. In this work, iLearn was used to extract all these features.

C. FEATURES SELECTION

After the features were extracted, we removed the redundant and noise features, which may have a great impact on the model generation.

In theory, every combination of the features should be used for data. But dealing with high-dimension features, computing resources would be exhausted. Suppose that there were 100-dimension features, it would require $2^{100}$ optimizations and models. Therefore, it is necessary to adopt some strategy for selecting optimal combinations of features. In previous studies, researchers have proposed a variety of excellent feature screening methods, such as single-factor analysis [69], Maximum-Relevance-Maximum-Distance (MRMD) [70], and Minimum Redundancy Maximum Relevance (mRMR) [71]. In this work, we used the following two screening methods.

1) INCREMENTAL FEATURE SELECTION

We sorted features in descending order according to ANOVA results and used incremental feature selection (IFS). In this way, the dimension of features drops rapidly so as that calculation speed is fast.

Analysis of variance (ANOVA) [72] was applied to rank features, not only due to its excellent characteristics, but also its short calculation time. According to the samples of two groups, Cyclin and nonCyclin, the f-score for each feature can be calculated based on the following formula.

$$f = \frac{\sum_{i=1}^{n} \frac{n_i \tilde{x}_i - \bar{x}}{1}}{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_{ij} - \bar{x}_j) / (\sum_{i=1}^{n} n_i - 2)}$$ (13)

In Eq. (13), $n_i$ represents the number of samples in the $i$-th group, $\tilde{x}_i$ represents the average value of the samples in the $i$-th group, $\bar{x}_j$ reflects the average value of all samples, and $x_{ij}$ is the value of $j$-th sample in the $i$-th group. The features can be ranked in descending order according to their f-score.

Feature set starts with the first feature with the maximum f-score. And the parameters of SVM are optimized in cross-validation. The accuracy of the feature set with these parameters was calculated. Then, the second feature in the ranked feature set was added to the above first one to produce a new feature set, the accuracy of which was also calculated by SVM. Given 100-dimension features as well, only 100 models are generated in this way, which greatly reduces the calculation time required for feature selection. After calculating the accuracy of each feature set, the feature set with the maximum accuracy, which is called SM, was obtained.

2) GREEDY ALGORITHM

Greedy algorithm [73] was used to further optimize the feature set.

The main process of greedy algorithm is as follows. As for 100-dimension features, the model is firstly trained using only 1-dimensional features. After 100 training sessions, the feature with the highest accuracy was selected. Then, another dimensional features are added to the previous features, and the model with 2-dimensional features is trained. 99 comparisons were performed to find the feature set with the highest accuracy. Repeat this step until the accuracy of new feature set is lower than that in the previously added features.

D. SUPPORT VECTOR MACHINE

In machine learning area, support vector machine (SVM) [74] is a supervised learning model, which is widely applied in
bioinformatics researches, especially in data classification and regression analysis [20], [75]–[83]. In this paper, we used LIBSVM [84] for data classification. Radial basis function was selected as kernel function. The grid search are used to optimize the kernel parameter $\gamma$ and the regularization parameter $C$.

E. EVALUATION OF MODEL

The generated model needs to be tested to verify its performance. Among many verification methods, cross-validation is often utilized because of its objectivity [85]–[91]. Three methods commonly used to validate models are: independent verification, n-fold cross-validation, and Jackknife cross-validation. Jackknife cross-validation is considered the best here because it is suitable for small-sample issues and can produce a unique result [92]. In Jackknife cross-validation, the sensitivity ($Sn$), specificity ($Sp$), and overall accuracy ($Acc$) were calculated as follows.

$$
\begin{align*}
Sn &= 1 - \frac{FN}{TP + FN} \\
Sp &= 1 - \frac{FP}{TN + FP} \\
Acc &= 1 - \frac{TP + FN}{TP + TN + FP + FN}
\end{align*}
$$

In Eq. (14), TP, TN, FP, FN represent the number of positive samples which are predicted as positive, the negative samples which are predicted as negative, the negative samples which are predicted as positive and the positive samples which are predicted as negative. We also used the receiver operation characteristic curve (ROC) and the area under ROC (AUC) to display the mode’s performance.

III. RESULTS AND DISCUSSION

In this work, a total of 18 different types of features were extracted, which constituted 13151-dimensional features. Based on the $f$-score of each feature calculated by ANOVA, we sorted the features in descending order.

After the first-step screening, the accuracy of trained models constructed for each feature set is illustrated in Figure 2.

When feature set has 130 features, the 5-fold cross-validation accuracy reaches the highest (95.2). Although the accuracy is superior, the feature dimension is still high.

Thus, we further used the greedy algorithm to further reduce feature dimension. After the second-step screening, the dimension of optimal feature set dramatically dropped to eight. Using these features to train the model, we obtained an accuracy rate of 91.9% in jackknife cross-validation with the AUC of 0.9159. The accuracy of new feature subset is slightly lower than that in the first-step feature selection. However, we should note that the feature dimension decreases from 130 to 8, which could guarantee the reliable of the model and reduce the risk of overfitting. Therefore, we prefer using the 8-dimension features to construct the final model. In Mohabatka’s work, the feature set has not been screened and they constructed their model using the whole 21 discrete factors [26].

In a recent study [26], Mohabatkar etc. used the same training data set as in this paper, and the accuracy of training results is 83.53%. Comparison in Table 3 shows that our proposed model is better than the published one.

Furthermore, we listed the 8 optimal features in Table 4. These feature descriptors have also been described in MATERIALS AND METHODS. Here we explain these features briefly. As known, Schneider-Wrede is a matrix of amino acid distance, and lag25 means that the SOCNumber of this reserved feature is 25th. BHAR880101.lag22 is Moran autocorrelation using BHAR880101, where d is 22. And CIDH920105.lag5 means the same as BHAR880101.lag22. CHAM810101.lag7 is NMBroto autocorrelation using CHAM810101, where d is 7. Hydrophobicity_PRAM900101.C1 means the composition of Class 1 classified by Hydrophobicity_PRAM900101, and Charge.C2 means the composition of Class 2 classified by Charge in $C \setminus T \setminus D$.  

### TABLE 3. Evaluation and comparison of models.

|                   | This paper | Mohabatkar etc’s model |
|-------------------|------------|------------------------|
| Acc               | 91.90%     | 83.53%                 |
| Sp                | 92.80%     |                        |
| Sn                | 91.00%     | 87.44%                 |
| AUC               | 0.9159     | 0.8944                 |
| Feature Dimensions| 8          | 21                     |

### TABLE 4. The 8-dimensional optimal features.

| Feature descriptors | Description of the feature |
|---------------------|-----------------------------|
| SOCNumber           | Schneider-Wrede.lag25       |
| AAC                 | Frequency of amino acid G   |
| Moran               | BHAR880101.lag22            |
| CTDC                | hydropobicity_PRAM900101.C1 |
| NMBroto             | CHAM810101.lag7             |
| Moran               | CIDH920105.lag5             |
| QSOder              | Schneider-Wrede.Xr.L        |
| CTDC                | Charge.C2                   |

FIGURE 2. The feature selection by using ANOVA.
Schneider-Wrede.Xr.L is QSOrder of amino acid L using Schneider-Wrede matrix. It is by these 8-dimension features that we could train a model with good results. And consequently the biochemical and physical-chemical significance represented by these features has certain function for cyclin. The second and seventh features reveal amino acid G and L is relevant to the nature of cyclin. The fourth and eighth features mean that neutral charge and neutral hydrophobicity are also relevant to cyclin. Coincidentally amino acid G and L are neutral hydrophobicity and neutral charge respectively. Meanwhile the autocorrelation of amino acids is also important to cyclin according to other features.

IV. CONCLUSION
A method based on SVM was developed to classify Cyclins using amino acid sequences. Through feature extraction among 18 types of feature descriptors and feature selection method including ANVOA and greedy algorithm, 8-dimension features were reserved. SVM model constructed through the 8-dimension features was better than recent study. Jackknife cross validation demonstrated this method is helpful for classifying Cyclins.

REFERENCES
[1] U. Galderisi, F. P. Jori, and A. Giordano, “Cell cycle regulation and neural differentiation,” Oncogene, vol. 22, pp. 5208–5219, Aug. 2003.
[2] D. O. Morgan, The Cell Cycle, Principles of Control, vol. 273, London, U.K.: New Science Press, 2007.
[3] P. C. John, M. Mews, and R. Moore, “Cyclin/Cdk complexes: Their involvement in cell cycle progression and mitotic division,” Protoplasma, vol. 216, nos. 1–4, pp. 119–142, 2001.
[4] I. Ferby, M. Blazquez, A. Palmer, R. Eritja, and A. R. Nebreda, “A novel p34cdc2-binding and activating protein that is necessary and sufficient to trigger G2/M progression in Xenopus oocytes,” Genes Develop., vol. 13, pp. 2177–2189, Aug. 1999.
[5] C. Liang, Q. Changlu, Z. He, F. Tongze, and X. Xue, “gutMDisorder: A comprehensive database for dysbiosis of the gut microbiota in disorders and interventions,” Nucleic Acids Res., vol. 48, no. 1, pp. D55–D560, 2019.
[6] L. Cheng, H. Yang, H. Zhao, X. Pei, H. Shi, J. Sun, Y. Zhang, Z. Wang, and M. Zhou, “MetSigDis: A manually curated resource for the metabolic signatures of diseases,” Briefings Bioinf., vol. 20, no. 1, pp. 203–209, Jan. 2019.
[7] Y. Xiao, J. Zhang, and L. Deng, “Prediction of IncRNA–protein interactions using HeteSim scores based on heterogeneous networks,” SCIL Rep., vol. 7, no. 1, pp. 3664, Dec. 2017.
[8] H. Liu, W. Zhang, B. Zou, J. Wang, Y. Deng, and L. Deng, “DrugCombDB: A comprehensive database of drug combinations toward the discovery of combinatorial therapy,” Nucleic Acids Res., vol. 48, pp. D871–D881, Oct. 2019.
[9] L. Zheng, S. Huang, N. Ma, H. Zhang, J. Zhang, Y. Chang, L. Yang, and Y. Zuo, “RAACBook: A Web server of reduced amino acid alphabet for sequence-dependent inference by using Chou’s five-step rule,” Database, vol. 2019, p. baxz131, Jan. 2019.
[10] D. Liu, G. Li, and Y. Zuo, “Function determinants of TET proteins: The arrangements of sequence motifs with specific codes,” Briefings Bioinf., vol. 20, pp. 1826–1835, Sep. 2019.
[11] L. Cheng, H. Zhuang, G. P. S. Raghava, and D. Gupta, “CyclinPred: A SVM-based method for predicting cyclin protein sequences,” PLos ONE, vol. 3, no. 7, p. e2605, Jul. 2008.
[12] Q. Zou, “Latest machine learning techniques for biomedicine and bioinformatics,” Current Bioinf., vol. 14, no. 3, pp. 176–177, Mar. 2019.
[13] B. Liu, X. Gao, and H. Zhang, “BioSeq-analysis2.0: An updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches,” Nucleic Acids Res., vol. 47, no. 20, p. e127, Nov. 2019.
[14] M. P. Mazanetz, O. Ichihara, R. J. Law, and M. Whittaker, “Prediction of cyclin-dependent kinase 2 inhibitor potency using the fragment molecular orbital method,” J. Cheminform., vol. 3, no. 1, p. 2, Jan. 2011.
[15] E. J. Chang, R. Begum, B. T. Chait, and T. Gaasterland, “Prediction of cyclin-dependent kinase phosphorylation substrates,” PLos ONE, vol. 2, no. 8, p. e656, Aug. 2007.
[16] M. K. Kalita, U. K. Nandal, A. Pattnaik, A. Sivalingam, G. Ramasamy, M. Kumar, G. P. S. Raghava, and D. Gupta, “CyclinPred: A SVM-based method for predicting cyclin protein sequences,” PLos ONE, vol. 3, no. 7, p. e2605, Jul. 2008.
[17] Q. Zou, “Latest machine learning techniques for biomedicine and bioinformatics,” Current Bioinf., vol. 14, no. 3, pp. 176–177, Mar. 2019.
[18] L. Cheng, Y. Jiang, J. Hu, J. Sun, J. Peng, M. Zhou, and Y. Hu, “InfAcrOnt: Calculating cross-ontology term similarities using information flow by a random walk,” BMC Genomics, vol. 19, no. S1, p. 919, Jan. 2018.
[19] X. Chen, W. Shi, and L. Deng, “Prediction of disease comorbidity using HeteSim scores based on multiple heterogeneous networks,” Current Gene Therapy, vol. 19, no. 4, pp. 232–241, Nov. 2019.
[20] Q. Jiang, G. Wang, S. Jin, Y. Li, and Y. Wang, “Predicting human microRNA-disease associations based on support vector machine,” Int. J. Data Mining Bioinform., vol. 8, no. 3, pp. 282–293, 2013.
[21] L. Cheng, Y. Hu, J. Sun, M. Zhou, and Q. Jiang, “DineRNA: A comprehensive Web-based bioinformatics toolkit for exploring disease associations and ncRNA function,” Bioinformatics, vol. 34, no. 11, pp. 1953–1956, Jun. 2018.
[22] L. Yu, F. Xu, and L. Gao, “Predict new therapeutic drugs for hepatocellular carcinoma based on gene mutation and expression,” Frontiers Biomed. Biotechnol., vol. 8, p. 8, Jan. 2020.
[23] Y. Pan, Z. Wang, W. Zhan, and L. Deng, “Computational identification of binding energy hot spots in protein–RNA complexes using an ensemble approach,” Bioinformatics, vol. 34, no. 9, pp. 1473–1480, May 2018.
[24] B. Liu, “BioSeq-analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches,” Briefings Bioinf., vol. 20, no. 4, pp. 1280–1294, Jul. 2019.
[25] K. Yan, X. Fang, Y. Xu, and B. Liu, “Protein fold recognition based on multi-view modeling,” Bioinformatics, vol. 35, no. 17, pp. 2982–2990, Sep. 2019.
[26] M. Mohabatkar, “Prediction of cyclin proteins using Chou’s pseudo amino acid composition,” Protein Pept. Lett., vol. 17, pp. 1207–1214, Oct. 2010.
[27] F. Meng, C. Wang, and L. Kurgan, “DTECT webserver: Fast predictor of propensity for protein production, purification, and crystallization,” BMC Bioinf., vol. 18, p. 580, 2018.
[28] K. Pailil and U. Chouhan, “Relevance of machine learning techniques and various protein features in protein fold classification: A review,” Current Bioinf., vol. 14, no. 8, pp. 688–697, Dec. 2019.
[29] W. Chen, P. Feng, T. Liu, and D. Jin, “Recent advances in machine learning methods for predicting heat shock proteins,” Current Drug Metabolism, vol. 20, no. 3, pp. 224–228, May 2019.
[30] S. Wang, Q. Zhang, C. Yu, Y. Cao, Y. Zuo, and L. Yang, “Immune cell infiltration-based signature for prognosis and immunogenicomic analysis in breast cancer,” Briefings Bioinf., Mar. 2020, doi: 10.1093/bib/bbaa026.
[31] L. Cheng, “Computational and biological methods for gene therapy,” Current Gene Therapy, vol. 19, no. 14, p. 210, 2019.
[32] L. Cheng and Y. Hu, “Human Disease System Biology,” Current Gene Therapy, vol. 18, no. 5, pp. 255–256, Nov. 2018.
[33] X. Zhao, Q. Jiao, H. Li, Y. Wu, H. Wang, S. Huang, and G. Wang, “ECFS-DEA: An ensemble classifier-based feature selection for differential expression analysis on expression profiles,” BMC Bioinf., vol. 21, no. 1, p. 43, Feb. 2020.
[34] G. Wang, Y. Yang, M. Teng, D. Zhang, L. Li, and Y. Liu, “Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon γ-stimulated HeLa cells,” PLos ONE, vol. 5, no. 10, Art. no. e11794, Oct. 2010.
[35] G. Wang, Y. Yang, W. Feng, X. Wang, J. Y. Yang, Y. Zhao, Y. Wang, and Y. Liu “Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells,” BMC Genomics, vol. 9, no. 2, p. 522, Sep. 2008.
[36] S. Basith, B. Manavalan, T. H. Shin, and G. Lee, “Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening,” Med. Res. Rev., Jan. 2020, doi: 10.1002/med.21658.
Z. Lv, S. Jin, H. Ding, and Q. Zou, “A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features,” *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 215, Sep. 2019.

V. Saravanan and N. Gautham, “Harnesinng computational biology for exact linear B-cell epitope prediction: A novel amino acid composition-based feature descriptor,” *OMICS*, vol. 19, pp. 648–658, Oct. 2015.

Z. Chen, P. Zhao, F. Li, A. Leier, T. T. Marquez-Lago, Y. Wang, G. I. Webb, A. I. Smith, R. J. Daly, K.-C. Jou, and J. Song, “Feature: A Python package and web server for feature extraction and selection from protein and peptide sequences,” *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, Jul. 2018.

Z.-P. Feng and C.-T. Zhang, “Prediction of membrane protein types based on the hydrophobic index of amino acids,” *J. Protein Chem.*, vol. 19, pp. 269–275, May 2000.

Z. R. Li, H. H. Lin, L. Y. Han, L. Jiang, X. Chen, and Y. Z. Chen, “PROFEAT: A Web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence,” *Nucleic Acids Res.*, vol. 34, pp. W32–W37, Jul. 2006.

R. R. Sokal and B. A. Thomson, “Population structure inferred by local spatial autocorrelation: An example from an American tribal population,” *Amer. J. Phys. Anthropol.*, vol. 129, pp. 121–131, Jan. 2006.

J. Shen, J. Zhang, X. Luo, W. Zhi, K. Yu, K. Chen, Y. Li, and H. Jiang, “Predicting protein–protein interactions based only on sequences information,” *Proc. Nat. Acad. Sci. USA*, vol. 104, pp. 4337–4341, Mar. 2007.

K. C. Chou, “Prediction of protein cellular attributes using pseudo-amino acid composition,” *Proteins*, vol. 43, pp. 246–255, May 2001.

K. C. Chou, “Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes,” *Bioinformatics*, vol. 21, pp. 9–10, Jan. 2005.

Z. Chen, P. Zhao, F. Li, T. T. Marquez-Lago, A. Leier, J. Revote, Y. Zhi, D. R. Powell, T. Akutsu, G. I. Webb, K.-C. Jou, A. I. Smith, R. J. Daly, J. Li, and J. Song, “ILearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data,” *Briefings Bioinf.*, vol. 21, no. 3, pp. 1047–1057, Apr. 2019.

H. Ding and D. Li, “Identification of mitochondrial proteins of malaria parasite using analysis of variance,” *Amino Acids*, vol. 47, pp. 329–333, Feb. 2015.

Q. Zhou, J. Zeng, L. Cao, and R. Ji, “A novel features ranking metric with application to scalable visual and bioinformatics data classification,” *Neurocomputing*, vol. 173, pp. 346–354, Jan. 2016.

H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1228–1238, Aug. 2005.

P. M. Feng, H. Ding, W. Chen, and H. Lin, “Naive Bayes classifier with feature selection to identify phage virus proteins,” *Comput. Math. Methods Med.*, vol. 2013, May 2013, Art. no. 530069.

H. Piriaviash, D. Ramanan, and C. C. Fowlkes, “Global-optimal greedy algorithms for tracking a variable number of objects,” in *Proc. CVPR*, Providence, RI, USA, 2011, pp. 1201–1208.

C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, pp. 273–297, Sep. 1995.

C. Meng, S. Jin, L. Wang, F. Guo, and Q. Zou, “AOPs-SVM: A sequence-based classifier of antioxidant proteins using a support vector machine,” *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 224, Sep. 2019.

Y. Wang, F. Shi, L. Cao, N. Dey, Q. Wu, A. S. Ashour, R. S. Sherratt, V. Rajinkanth, and L. Wu, “Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images,” *Current Bioinf.*, vol. 14, no. 4, pp. 282–294, Apr. 2019.

N. Zhang, Y. Sa, Y. Guo, W. Lin, P. Wang, and Y. Feng, “Discriminating ramos and jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine,” *Current Bioinf.*, vol. 13, no. 1, pp. 50–56, Feb. 2018.

L. Chao, L. Wei, and Q. Zou, “SecProMTB: A SVM-based classifier for secretory proteins of mycobacterium tuberculosis with imbalanced data set,” *Proteomics*, vol. 19, no. 1, 2019, Art. no. e1900007.

J. Zhao, D. Li, Y. Wang, L. Li, and Q. Zou, “Cancer diagnosis through IsomiR expression with machine learning method,” *Current Bioinf.*, vol. 13, no. 1, pp. 57–63, Feb. 2018.

K. Liu and W. Chen, “IMRM: A platform for simultaneously identifying multiple kinds of RNA modifications,” *Bioinformatics*, vol. 36, no. 11, pp. 3336–3342, Mar. 2020.

Y. Zhao, F. Wang, and L. Juan, “MicroRNA promoter identification in Arabidopsis using multiple histone markers,” *Biomed. Res. Int.*, vol. 2015, Sep. 2015, Art. no. 861402.
[82] L. Deng, J. Wang, and J. Zhang, “Predicting gene ontology function of human MicroRNAs by integrating multiple networks,” Frontiers Genet., vol. 10, p. 3, Jan. 2019.

[83] B. Liu, C.-C. Li, and K. Yan, “DeepSVM-fold: Protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks,” Briefings Bioinf., Oct. 2019, doi: 10.1093/bib/bbz098.

[84] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 1–27, Apr. 2011.

[85] L. Cheng, H. Zhao, P. Wang, W. Zhou, M. Luo, T. Li, J. Han, S. Liu, and Q. Jiang, “Computational methods for identifying similar diseases,” Mol. Therapy Nucleic Acids, vol. 18, pp. 590–604, Dec. 2019.

[86] M. M. Hasan, B. Manavalan, M. S. Khatun, and H. Kurata, “i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome,” Int. J. Biol. Macromol., Dec. 2019, doi: 10.1016/j.ijbiomac.2019.12.009.

[87] M. M. Hasan, B. Manavalan, W. Shoombuatong, M. S. Khatun, and H. Kurata, “i6mA-Fuse: Improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation,” Plant Mol. Biol., vol. 103, nos. 1–2, pp. 225–234, Mar. 2020.

[88] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, “AtbPpred: A robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees,” Comput. Struct. Biotechnol. J., vol. 17, pp. 972–981, Jan. 2019.

[89] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, “Meta-4mCpred: A sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation,” Mol. Therapy Nucleic Acids, vol. 16, pp. 733–744, Jun. 2019.

[90] J. Yuan, Y. Zhang, H. Liu, Z. Tian, X. Li, Y. Zheng, Q. Gao, L. Song, X. Xiao, J. Sun, Z. Wang, and B. Li, “Clinical observation of patients with Leber’s hereditary optic neuropathy before gene therapy,” Current Gene Therapy, vol. 18, no. 6, pp. 386–392, 2018.

[91] Y. Pan, D. Liu, and L. Deng, “Accurate prediction of functional effects for variants by combining gradient tree boosting with optimal neighborhood properties,” PLoS ONE, vol. 12, no. 6, Jun. 2017, Art. no. e0179314.

[92] M.-L. Liu, W. Su, Z.-X. Guan, D. Zhang, W. Chen, L. Liu, and H. Ding, “An overview on predicting protein subchloroplast localization by using machine learning methods,” Current Protein Peptide Sci., vol. 21, Jan. 2020.

JIA-NAN SUN is currently pursuing the bachelor’s degree with the Center for Informational Biology, University of Electronic Science and Technology of China. His research interest includes bioinformatics.

HUA-YI YANG is currently pursuing the bachelor’s degree with the Center for Informational Biology, University of Electronic Science and Technology of China. Her research interest includes bioinformatics.

JING YAO is currently pursuing the bachelor’s degree with the Center for Informational Biology, University of Electronic Science and Technology of China. Her research interest includes bioinformatics.

HUA TANG is currently a Professor with the Department of Pathophysiology, Southwest Medical University. Her research interests include bioinformatics and pathophysiology.

HUI DING received the Ph.D. degree from Inner Mongolia University, in 2009. She is currently an Associate Professor with the School of Life Science and Technology, University of Electronic Science and Technology of China. Her research interest includes systems biology.

SHU-GUANG HAN is currently a Lecturer with the Center for Informational Biology, University of Electronic Science and Technology of China. His research interest includes bioinformatics.

CHENG-YAN WU is currently an Lecturer with the Baotou Teacher’s College, Inner Mongolia University of Science and Technology. Her research interest includes bioinformatics.