Information Dynamics and Applications
https://www.acadlore.com/journals/IDA

Classification of Cyclin Proteins Using Amino Acid Composition and an SVM Approach: An In-Depth Analysis

Muhammad Hassaan*

Department of Computer Science and Information Technology, Virtual University of Pakistan, 54000 Lahore, Pakistan

* Correspondence: Muhammad Hassaan (m.hassaan@vu.edu.pk)

Received: 08-10-2023 **Revised:** 09-18-2023 **Accepted:** 09-25-2023

Citation: M. Hassaan, "Classification of cyclin proteins using amino acid composition and an SVM approach: An in-depth analysis," *Inf. Dyn. Appl.*, vol. 2, no. 3, pp. 153–161, 2023. https://doi.org/10.56578/ida020305.

Abstract: Cyclins, commonly referred to as co-enzymes, are a pivotal family of proteins that modulate cellular growth by activating cell-cycle mediators, proving essential for the cell cycle. Due to the marked dissimilarity in their sequences, effective differentiation among cyclins remains a challenging endeavour. In this study, an innovative methodology was proposed, wherein the amino acid composition was utilized to inform an SVM-based classification approach. SVMs, being supervised machine learning algorithms, are typically employed for classification and regression tasks. From the data analyzed, eighteen (18) feature labels were extracted, culminating in an extensive set of thirteen thousand one hundred and fifty-one (13,151) discernible features. Employing the jackknife cross-validation technique revealed that this SVM-informed approach facilitated the identification of cyclins with an accuracy rate of 91.9%, a notable improvement from prior studies. Such advancements underscore the potential for more accurate and efficient differentiation of cyclins in future endeavours.

Keywords: Cyclin; Cell-cycle; Amino acid; SVM; Classification

1 Introduction

Cyclins, at times referred to as co-enzymes, represent a crucial family of proteins that are implicated in the orchestration of cellular growth. These proteins function by activating cell-cycle mediators, a subset of serine proteases fundamental to the cell-cycle [1]. The presence and concentration of various cyclins fluctuate throughout the cell cycle, as depicted in Figure 1.

Two primary mechanisms are recognized for governing changes in cyclin concentrations. Firstly, variations in cyclin gene expression are often attributed to these shifts [2]. Secondly, the ubiquitin-mediated degradation pathway universally mediates alterations in cyclin concentrations [2]. With the collaboration of cyclin-dependent kinases (CDKS), complexes are formed by cyclins. Following phosphorylation—a process involving the addition of a phosphate group—CDK's active site becomes operational [2]. Subsequently, these activated complexes are known to play pivotal roles in cell cycle progression [2]. When cyclin binds with CDK, the maturation-promoting factor (MPF) is produced. This factor is responsible for the phosphorylation of various proteins, thereby facilitating distinct cell-cycle processes, notably microtubule and chromosomal reorganization [3]. It has been observed that cyclins don't possess an enzymatic active site; however, they present a surface-binding site and have the capability to localise CDKS within specific sub-cellular compartments [2].

The post-genomic era has witnessed a remarkable proliferation of biological data, particularly sequence data [4–7]. Traditional methodologies for processing and understanding this information not only tend to be time-intensive and expensive but also yield relatively low success rates. Thus, swift techniques for sequence identification have become increasingly sought after the references [8–12]. Prevalent computational methods, such as BLAST and FASTA, can facilitate unique nucleotide to peptide sequence database searches, yet they exhibit limitations in cyclin differentiation due to sequence dissimilarities. Hence, machine learning-based classification in this domain has garnered increased attention [13–24]. In prior approaches, like the StAR [25], using pseudo amino acid composition achieved an impressive accuracy rate of 83.53% for co-enzyme detection. After meticulous analysis, eighteen distinct feature labels were extracted, reducing feature dimensions. Through refined methodologies, an impressive accuracy rate of 94.3% was achieved, with a Jackknife cross-validation efficiency of 91.9%.

https://doi.org/10.56578/ida020305

Emerging research has highlighted the potential role of cyclin D1 in DNA repair enhancement, potentially safeguarding transformed cells from excessive genomic instability and potentially aiding in shielding breast cancer cells from DNA-damaging treatments. Intriguingly, cyclin D1 has been recently identified as a promoter of whole-genome chromosomal instability [26]. Traditionally, the mitotic protein mechanism has been acknowledged as a vital component of cell membrane transition and regulation. This mechanism, being integral to a plethora of standard and malignant intracellular signaling processes, collaborates with CDKS to regulate the quantity of various cyclin subunits and CDK inhibitors [27]. Peptides autophagy offers a mechanism for chronological execution and harmonization of transcription transitions by working in tandem with cyclin-dependent kinases (Cdks) to modulate the number of different cyclin sub units and Cdk antagonists [28]. D-type co-enzymes, produced by three primary enzymes, have emerged as significant receptors of exogenous signal transduction, transmitting pro-inflammatory signals to the intrinsic circadian tissue generator [29]. Through interactions with CDK4 and CDK6, and by retaining CDK inhibitors p21 and p27, D-type co-enzymes are posited as primary promoters of G1 phase progression in relation to non-inflammatory stimuli. Among these, cyclin D1, one of the most extensively studied D-type cyclins, has been consistently linked to malignancy alterations. One of most investigated D-type cyclin, Cyclin D1, is typically associated with alterations in malignancy, and its abundance having more positive to cell transformation and malignancy [30]. Elevated cyclin D1 levels in tumour cells, largely resulting from aberrant protein ubiquity and stability, have been identified as markers of cancer phenotype and disease progression [31, 32].

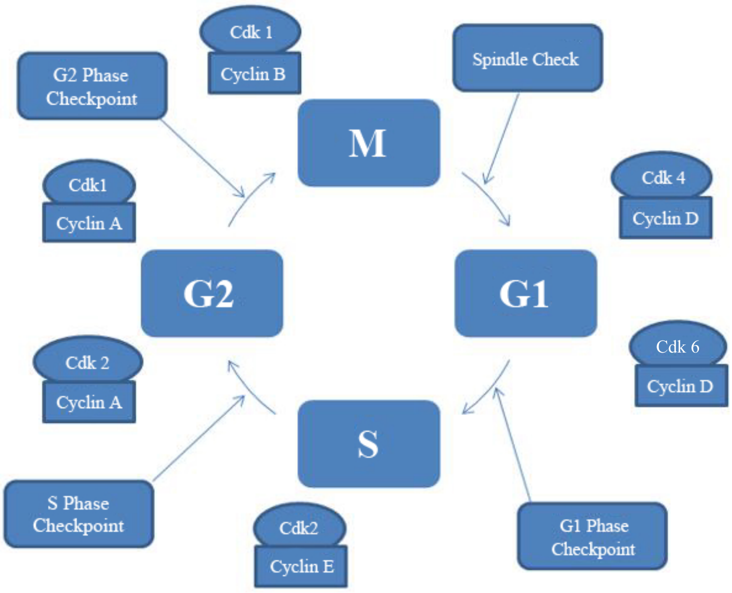

Figure 1. Cyclins in cell-cycle

2 Methodology

As illustrated in Figure 2, an approach was established whereby data were first collected, followed by the application of various feature selection strategies for feature extraction. Subsequently, a range of classifiers were employed, culminating in the validation of the adopted technique.

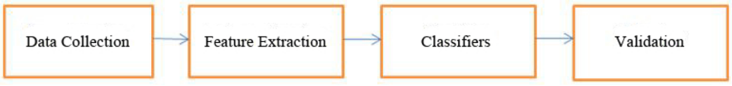

Figure 2. Overview of methodology

2.1 Benchmark Dataset

The datasets employed in this investigation were sourced from UNIPROT, a renowned protein database. Two distinct datasets were utilized: one representing cyclin proteins and the other for acyclic proteins. The search term “cyclin proteins” was used to identify cyclin proteins within the UNIPROT database, while “acyclic proteins” served to identify the latter dataset. Initial collection revealed 297 cyclin proteins and 313 acyclic proteins. To reduce redundancy and potential bias, repetitive sequences exhibiting over 70% similarity were eliminated using the

CD-HIT Suite. Post homology reduction, the datasets comprised 146 cyclin proteins and 13 monocyclic peptides. These peptides were then integrated into the proposed model for both training and testing purposes. Although employing a benchmark dataset with a lower sequence identity threshold, such as 25%, might potentially improve accuracy, it was determined that this would substantially reduce the overall sample size, jeopardizing statistical validity. Consequently, such a stringent threshold was not adopted in this study.

2.2 Feature Extraction Strategies

The efficacy of machine learning-based protein classification largely hinges on the robustness of the features selected. By judiciously selecting optimal features, both the classifier's performance and the model-building process can be significantly enhanced. In this section, the features employed in this study will be elucidated.

A) Frequency Vector (FV)

Insightful details about the benchmark datasets in relation to each protein constituent within the population are furnished by frequency analysis. For each peptide position, its occurrence is computed and then encapsulated within a vector, known as the *FV*. The *FV* is adept at preserving data concerning the magnitude and composition of protein samples. The *FV* is derived as:

$$FV = [r_1, r_2, \cdots, r_{20}] \tag{1}$$

In this equation, each r_i signifies the frequency of a unique amino acid position in the sequence, arranged in alphabetical order.

B) Computation of Position Relative Incidence Matrix (PRIM)

The primary sequence is instrumental in deciphering concealed attributes of a protein. The underlying mathematical paradigm of this model rests on the segmentation of acids within proteins and a subset of proteins from the benchmark dataset. A 20×20 matrix, labelled as H_{PRIM}, is utilized to capture the relative positioning data of peptides or amino acids. This matrix is constructed from the relative segmentation of protein sample acids. Its derivation is:

$$H_{PRIM} = \begin{bmatrix} H_{1\to 1} & H_{1\to 2^2}\cdots & H_{1\to j'}\cdots & H_{1\to 20} \\ H_{2\to 1} & H_{2\to 2^2}\cdots & H_{2\to j'}\cdots & H_{2\to 20} \\ \vdots & \vdots & & \vdots \\ H_{i\to 1} & H_{i\to 2}\cdots & H_{i\to j'} & H_{i\to 20} \\ \vdots & \vdots & & \vdots \\ H_{N\to 1} & H_{N\to 2'} & H_{N\to j}\cdots & H_{N\to 20} \end{bmatrix} \tag{2}$$

In the H_{PRIM} matrix, every $H_{i-} > j$ possesses an aggregate value. This value is discerned as the relative positioning of the j-th residue concerning the initial appearance of the i-th residue. This process yields 400 coefficients. To streamline these coefficients, a set of computations is conducted, eventually resulting in 30 coefficients.

C) Computation of Reverse RPRIM

To unveil nuanced features in proteins with certain ambiguities, H_{RPRIM} is determined, drawing on information from the reverse-sequenced protein sample. The formulation of H_{RPRIM} is:

$$H_{RPRIM} = \begin{bmatrix} H_{1\to 1} & H_{1\to 2}\cdots & H_{1\to j'}\cdots & H_{1\to 1} \\ H_{2\to 1} & H_{2\to 2^2}\cdots & H_{2\to j'}\cdots & H_{2\to 20} \\ \vdots & \vdots & & \vdots \\ H_{i\to 1} & H_{i\to 2'} & H_{i\to j'}\cdots & H_{i\to 20} \\ \vdots & \vdots & \vdots & \vdots \\ H_{N\to 1} & H_{N\to 2^2} & H_{N\to j'} & H_{N\to 20} \end{bmatrix} \tag{3}$$

H_{RPRIM} furnishes 400 coefficients akin to H_{PRIM}, undergoing an identical coefficient reduction process, and eventually generating 30 empirical coefficients.

D) Computation of Accumulative Absolute Position Incidence Vector (AAPIV)

A frequency matrix was devised to retain the locational data of peptides or amino acids and to unveil the nuanced characteristics of protein sequences associated with configurational data. However, it omits details on the relative positioning of peptides. To bridge this gap, AAPIV is employed, calculated for a set of 20 indigenous peptides:

$$\text{AAPIV } = [u_1, u_2, u_3, \ldots, u_{20}] \tag{4}$$

Herein, each u_i represents an AAPIV component, its formulation being:

$$u_i = \sum_{k=1}^{n} p_k \tag{5}$$

E) Computation of Reverse Accumulative Absolute Position Incidence Vector (RAAPIV)

To expose salient features of patterns in terms of relative positioning data, RAAPIV is derived from the inverse sequence of protein samples, following a methodology analogous to AAPIV:

$$\text{RAAPIV } = [u_1, u_2, u_3, \ldots, u_{20}] \tag{6}$$

2.3 Feature Selection

Upon feature extraction, duplicated and chaotic features were excluded, given their potential to significantly influence model construction. While theoretically every feature permutation could be utilized for data, processing becomes cumbersome with increasing feature dimensions. For instance, with a feature dimensionality of 100, 2^{100} potential optimizations and models would need to be considered. Thus, an effective method for isolating optimal trait combinations becomes paramount. Notable feature screening methodologies have been previously discussed in the literature, encompassing single-factor analysis [33], Maximum-Relevance-Maximum-Distance (MRMD) [34], and Minimum Redundancy Maximum Relevance (mRMR) [35]. In this study, two screening methodologies were employed.

A) Incremental Feature Selection (IFS)

Incremental feature selection was utilized, with features being ordered in a descending manner based on Analysis of Variance (ANOVA) values. This approach led to a rapid reduction in feature dimensionality, facilitating efficient computations. ANOVA [36] was chosen for feature ranking not only due to its distinguished attributes but also its computational efficiency. The f-score for each feature can be deduced as:

$$f = \frac{\sum_{i=1}^{2} n_i \left(\bar{x}_i - \bar{x}\right)/1}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)/\left(\sum_{i=1}^{2} n_i - 2\right)} \tag{7}$$

In this equation, n_i is indicative of the observations in the i-th class, x_i represents the average value of said observations, x_{ij} is the value of the j-th observation in the i-th group, and x_i is the average rate of all measurements. Features can then be ranked based on their respective f-scores. The feature boasting the highest f-score occupies the premier position in the selection hierarchy. The coefficients of SVM are adjusted via cross-validation. Using these values, the precision of the feature set was assessed. Thereafter, by incorporating the subsequent highest-ranked feature, a new feature set was formed and its precision was also evaluated using SVM. With a 100-dimensional feature set, this procedure yields merely 100 models, marking a substantial reduction in computational time needed for feature selection. Upon computing the accuracy for each feature set, the optimal feature set, denoted as the statistical model, was discerned.

B) Greedy Algorithm

To further refine the feature set, the greedy algorithm [37] was applied. The foundational steps of the greedy algorithm entail training the model initially with a singular dimensional feature, progressing eventually to features of 100 dimensions. Post 100 training iterations, the feature associated with the highest precision was identified. Subsequently, by adding more spatial features to the preceding ones, the classifier was trained with two-dimensional data. After 99 such comparisons, the optimal feature set was identified. This procedure was reiterated until the precision of an additionally introduced feature was found to be inferior to its predecessors.

C) SVM

The SVM [38] is a supervised learning model that has been extensively adopted in bioinformatics investigations, especially in data regression and classification challenges [39–47]. For the classification tasks in this study, LIBSVM [48] was used. A linear function was selected as the kernel function. Both the kernel γ variable and the regularization variable C were fine-tuned via grid search.

2.4 Model Evaluation

For the effective functioning of a model, rigorous testing is paramount. Owing to its inherent independence, cross-validation has been identified as one of the pre-eminent methods for this purpose [49–55]. Within the realm of model validation, three predominant approaches emerge: Jackknife cross-validation, n-fold cross-validation, and independent validation. Of these, Jackknife cross-validation has been highlighted as especially potent for compact datasets, though it occasionally yields atypical results [56].

Eq. (8) delineates the calculations for sensitivity (*Sn*), specificity (*Sp*), and overall accuracy (*Acc*) within the context of Jackknife cross-validation.

$$\begin{cases} Sn = 1 - \frac{FN}{TP+FN} \\ Sp = 1 - \frac{FP}{TN+FP} \\ Acc = 1 - \frac{TP+TN}{TP+TN+FP+FN} \end{cases} \tag{8}$$

In this equation, the terms True Positive (*TP*), True Negative (*TN*), False Positive (*FP*), and False Negative (*FN*) represent the number of positive instances correctly identified as positive, negative instances correctly identified as negative, negative instances incorrectly identified as positive, and positive instances incorrectly identified as negative, respectively.

The receiver operating characteristic curve (ROC) alongside the area under the curve (AUC) were utilized to elucidate the performance metrics of the model.

To ensure that the model is working properly, it must be tested. Cross-validation is one of the most popular verification methods because of its independence [49–55]. Jack knifing pass, n-fold pass, and independent confirmation are three common approaches for validating models. The best method is Jackknife cross-validation, which is excellent for tiny problems and can yield a weird solution [56]. The sensitivity (*Sn*), specificity (*Sp*), and overall accuracy (*Acc*) of Jackknife cross-validation were calculated as follows.

3 Results

In the course of this investigation, 18 distinct categories of attributes were procured, yielding a cumulative 13151-dimensional feature set. These features were systematically ranked based on their respective f-scores, determined through a one-way ANOVA.

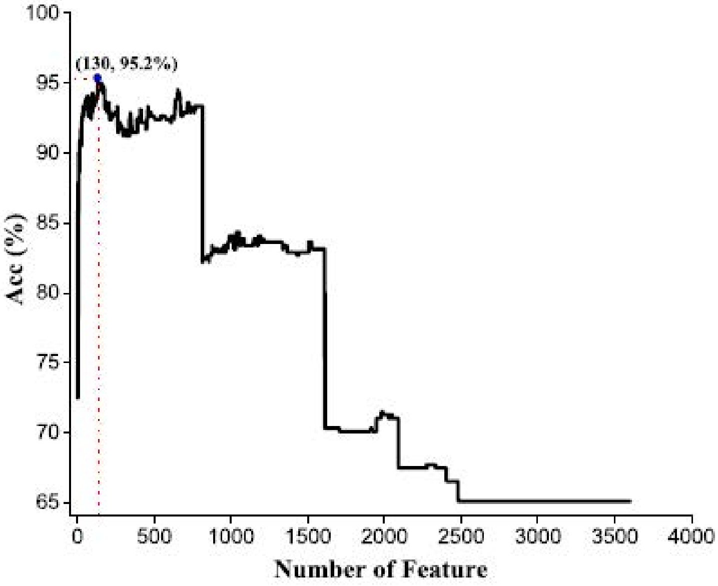

Figure 3. ANOVA-based feature selection

Table 1. Model evaluation and comparative analysis

Metric/Study	Current Study	Mohabatkar [25]
Accuracy (Acc)	91.90%	83.53%
Specificity (Sp)	92.80%	-
Sensitivity (Sn)	91.00%	87.44%
Area Under Curve (AUC)	0.9159	0.8944
Dimensions of Feature	8	21

Figure 3 elucidates the efficacy of models trained for each feature set after the initial evaluation phase. The pinnacle of precision, observed during a 5-fold cross-validation, materialises when the feature set is comprised of 130 features, registering at 95.2% (as depicted in Figure 3). Despite the laudable precision, the dimensionality remains rather substantial. Consequently, a greedy approach was adopted to further condense the feature set's dimensions. After the secondary screening, the feature set dimension was discernibly truncated to eight. With this refined set, an accuracy of 91.9% was achieved in the jackknife validation, paired with an area under the curve of 0.9159. Although the precision exhibited by this compacted subset is marginally eclipsed by that of the original selection from the primary phase, the dimensionality reduction from 130 to 8 attributes could bolster the model's robustness and diminish overfitting susceptibility. Thus, the final model was constructed employing these

8-dimensional attributes. In contrast, in the study conducted by Mohabatkar [25], the feature set underwent no filtration, leading to a model built upon all 21 discrete attributes.

Table 1 posits that the methodology delineated here exhibits superior performance compared to previously published models.

4 Discussion

The identification of Cyclins through amino acid composition has been a topic of significant interest given its relevance to understanding cell cycle regulation and its potential applications in numerous biological fields. This study's methodological approach of combining SVM with advanced feature extraction techniques offers promising advancements in the quest for accurate Cyclin identification.

One of the main contributions of this study is the effective reduction of features. From an extensive pool of eighteen categories and over thirteen thousand dimensional features, the optimal feature space was distilled down to just eight. This not only simplifies the model, making it computationally more efficient, but also potentially increases the generalizability by reducing the risk of overfitting. The use of ANOVA and the greedy algorithm in this feature extraction process is noteworthy. Both techniques have been utilized in various applications across disciplines, but their effectiveness in this particular context underscores their potential for broader applications in bioinformatics.

Comparatively, when set against findings from Mohabatkar's work [25], the model proposed herein showcased enhanced accuracy and performance while operating on a reduced feature set. The reduction in feature dimensions from 21 to 8, coupled with improved performance metrics, indicates the efficacy of the methodological innovations introduced.

However, as with all studies, there are considerations to be made. Although the eight-dimensional features led to improved accuracy in the SVM model, one must question the biological relevance of each feature. Future studies might delve deeper into the understanding of why these specific features were critical in the identification process. Moreover, the external validity of the model should be tested across different datasets to ascertain its broad applicability.

Jackknife cross-validation, a significant aspect of the validation process in this research, further corroborates the model's reliability. However, exploring other validation techniques in conjunction could provide a more comprehensive view of the model's robustness.

In conclusion, the findings of this study provide a foundational step for future research in this domain. With continued advancements and refinements, the path towards precise and efficient identification of Cyclins seems clearer than ever.

5 Conclusions

In the ever-evolving realm of bioinformatics, the identification of Cyclins based on amino acid composition has been a notable challenge. This study took a leap forward by leveraging the strengths of the SVM in tandem with innovative feature extraction techniques. Out of a comprehensive pool of eighteen distinct categories of features, the method distilled the dimensions down to just eight, without compromising the accuracy. This reduction in feature dimensions, achieved through methods such as ANOVA and the greedy algorithm, not only enhanced the model's accuracy but also provided an essential safeguard against the ever-looming threat of overfitting.

In direct comparison to previous models, particularly the one outlined by Mohabatkar [25], the approach delineated in this research was demonstrably superior. Beyond mere numerical supremacy, the findings have broader implications. They underscore the invaluable role of meticulous feature extraction in the bioinformatics domain, potentially paving the way for more streamlined and efficient models in the future.

Moreover, the research fills an important gap in the existing literature, serving as an exemplar of how judicious computational methods can be harmoniously combined with biological data to yield robust results. As the world stands on the cusp of more advanced computational and biological integrations, the methodologies and results from this study provide a solid foundation for future endeavours aiming to solve complex biological puzzles.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] U. Galderisi, F. Jori, and A. Giordano, "Cell cycle regulation and neural differentiation," *Oncogene*, vol. 22, no. 33, pp. 5208–5219, 2003. https://doi.org/10.1038/sj.onc.1206558

[2] D. O. Morgan, *The Cell Cycle, Principles of Control.* London, U.K.: New Science Press, 2007. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2248297/

[3] 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, 1999. https://doi.org/10.1101/gad.13.16.2177

[4] C. Liang, Q. Changlu, Z. He, F. Tongze, and Z. Xue, "gutMDisorder: A comprehensive database for dysbiosis of the gut microbiota in disorders and interventions," *Nucleic Acids Res.*, vol. 48, no. 1, pp. D554–D560, 2019. https://doi.org/10.1093/nar/gkz843

[5] 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," *Brief. Bioinf.*, vol. 20, no. 1, pp. 203–209, 2019. https://doi.org/10.1093/bib/bbx103

[6] Y. Xiao, J. Zhang, and L. Deng, "Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks," *Sci. Rep.*, vol. 7, no. 1, p. 3664, 2017. https://doi.org/10.1038/s41598-017-03986-1

[7] 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, 2019. https://doi.org/10.1093/nar/gkz1007

[8] L. Zheng, S. Huang, N. Mu, 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. baz131, 2019. https://doi.org/10.1093/database/baz131

[9] D. Liu, G. Li, and Y. Zuo, "Function determinants of TET proteins: The arrangements of sequence motifs with specific codes," *Brief. Bioinf.*, vol. 20, pp. 1826–1835, 2019. https://doi.org/10.1093/bib/bby053

[10] L. Cheng, H. Zhuang, S. Yang, H. Jiang, S. Wang, and J. Zhang, "Exposing the causal effect of C-reactive protein on the risk of type 2 diabetes mellitus: A mendelian randomization study," *Front. Genet.*, vol. 9, p. 657, 2018. https://doi.org/10.3389/fgene.2018.00657

[11] L. Cheng, P. Wang, R. Tian, S. Wang, Q. Guo, M. Luo, W. Zhou, G. Liu, H. Jiang, and Q. Jiang, "LncRNA2Target v2.0: A comprehensive database for target genes of lncRNAs in human and mouse," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D140–D144, 2019. https://doi.org/10.1093/nar/gky1051

[12] 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, 2019. https://doi.org/10.1093/nar/gkz740

[13] 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, 2011. https://doi.org/10.1186/1758-2946-3-2

[14] 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, 2007. https://doi.org/10.1371/journal.pone.0000656

[15] 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, 2008. https://doi.org/10.1371/journal.pone.0002605

[16] Q. Zou, "Latest machine learning techniques for biomedicine and bioinformatics," *Curr. Bioinf.*, vol. 14, no. 3, pp. 176–177, 2019. https://doi.org/10.2174/157489361403190220112855

[17] L. Cheng, Y. Jiang, H. Ju, 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, 2018. https://doi.org/10.1186/s12864-017-4338-6

[18] X. Chen, W. Shi, and L. Deng, "Prediction of disease comorbidity using hetesim scores based on multiple heterogeneous networks," *Curr. Gene Ther.*, vol. 19, no. 4, pp. 232–241, 2019. https://doi.org/10.2174/1566523219666190917155959

[19] 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. https://doi.org/10.1504/IJDMB.2013.056078

[20] L. Cheng, Y. Hu, J. Sun, M. Zhou, and Q. Jiang, "DincRNA: A comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function," *Bioinformatics*, vol. 34, no. 11, pp. 1953–1956, 2018. https://doi.org/10.1093/bioinformatics/bty002

[21] L. Yu, F. Xu, and L. Gao, "Predict new therapeutic drugs for hepatocellular carcinoma based on gene mutation and expression," *Front. Bioeng. Biotechnol.*, vol. 8, p. 8, 2020. https://doi.org/10.3389/fbioe.2020.00008

[22] 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, 2018. https://doi.org/10.1093/bioinformatics/btx822

[23] B. Liu, "BioSeq-analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches," *Brief. Bioinf.*, vol. 20, no. 4, pp. 1280–1294, 2019. https://doi.org/10.1093/bib/bbx165

[24] 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, 2019. https://doi.org/10.1093/bioinformatics/btz040

[25] H. Mohabatkar, "Prediction of cyclin proteins using chou's pseudo aminoacid composition," *Protein Pept. Lett.*, vol. 17, pp. 1207–1214, 2010.

[26] M. C. Casimiro, M. Crosariol, E. Loro, Z. Li, and R. G. Pestell, "Cyclins and cell cycle control in cancer and disease," *Genes Cancer*, vol. 3, no. 11-12, pp. 649–657, 2012. https://doi.org/10.1177/1947601913479022

[27] A. Hershko and A. Ciechanover, "The ubiquitin system," *Annu. Rev. Biochem.*, vol. 67, pp. 425–479, 1998. https://doi.org/10.1146/annurev.biochem.67.1.425

[28] K. I. Nakayama and K. Nakayama, "Ubiquitin ligases: Cell-cycle control and cancer," *Nat. Rev. Cancer*, vol. 6, pp. 369–381, 2006. https://doi.org/10.1038/nrc1881

[29] C. J. Sherr, "D-type cyclins," *Trends Biochem. Sci.*, vol. 20, pp. 187–190, 1995. https://doi.org/10.1016/s0968-0004(00)89005-2

[30] M. C. Casimiro, M. Velasco-Velázquez, C. Aguirre-Alvarado, and R. G. Pestell, "Overview of cyclins D1 function in cancer and the CDK inhibitor landscape: Past and present," *Expert Opin. Investig. Drugs*, vol. 23, no. 3, pp. 295–304, 2014. https://doi.org/10.1517/13543784.2014.867017

[31] E. A. Musgrove, C. E. Caldon, J. Barraclough, A. Stone, and R. L. Sutherland, "Cyclin D as a therapeutic target in cancer," *Nat. Rev. Cancer*, vol. 11, no. 8, pp. 558–572, 2011. https://doi.org/10.1038/nrc3090

[32] J. P. Alao, "The regulation of cyclin D1 degradation: Roles in cancer development and the potential for therapeutic invention," *Mol. Cancer*, vol. 6, p. 24, 2007. https://doi.org/10.1186/1476-4598-6-24

[33] H. Ding and D. Li, "Identification of mitochondrial proteins of malaria parasite using analysis of variance," *Amino Acids*, vol. 47, no. 2, pp. 329–333, 2015. https://doi.org/10.1007/s00726-014-1862-4

[34] Q. Zou, 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, 2016. https://doi.org/10.1016/j.neucom.2014.12.123

[35] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 2005, p. 1226–1238. https://ieeexplore.ieee.org/abstract/document/1453511

[36] P. M. Feng, H. Ding, W. Chen, and H. Lin, "Naive bayes classifier with feature selection to identify phage virion proteins," *Comput. Math. Methods Med.*, vol. 2013, p. 1530696, 2013. https://doi.org/10.1155/2013/530696

[37] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *CVPR 2011*, Colorado Springs, CO, USA, 2011, pp. 1201–1208. https://doi.org/10.1109/CVPR.2011.5995604

[38] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995. https://doi.org/10.1007/BF00994018

[39] 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, 2019. https://doi.org/10.3389/fbioe.2019.00224

[40] Y. Wang, F. Shi, L. Cao, N. Dey, Q. Wu, A. S. Ashour, R. S. Sherratt, V. Rajinikanth, 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, 2019. https://doi.org/10.2174/1574893614666190304125221

[41] 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," *Curr. Bioinf.*, vol. 13, no. 1, pp. 50–56, 2018. https://doi.org/10.2174/1574893611666160608102537

[42] C. Meng, L. Wei, and Q. Zou, "SecProMTB: Support Vector Machine-based classifier for secretory proteins using imbalanced data sets applied to mycobacterium tuberculosis," *Proteomics*, vol. 19, no. 1, p. e1900007, 2019. https://doi.org/10.1002/pmic.201900007

[43] Z. Liao, D. Li, X. Wang, L. Li, and Q. Zou, "Cancer diagnosis through IsomiR expression with machine learning method," *Curr. Bioinf.*, vol. 13, no. 1, pp. 57–63, 2018. https://doi.org/10.2174/1574893611666160609081155

[44] K. Liu and W. Chen, "IMRM: A platform for simultaneously identifying multiple kinds of RNA modifications," *Bioinformatics*, vol. 36, no. 11, pp. 3336–3342, 2020. https://doi.org/10.1093/bioinformatics/btaa155

[45] Y. Zhao, F. Wang, and L. Juan, "MicroRNA promoter identification in Arabidopsis using multiple histone markers," *Biomed. Res. Int.*, vol. 2015, pp. 1–10, 2015. https://doi.org/10.1155/2015/861402

[46] L. Deng, J. Wang, and J. Zhang, "Predicting gene ontology function of human MicroRNAs by integrating

multiple networks," *Front. Genet.*, vol. 10, no. 3, 2019. https://doi.org/10.3389/fgene.2019.00003

[47] 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," *Brief. Bioinform.*, vol. 21, no. 5, pp. 1733–1741, 2019. https://doi.org/10.1093/bib/bbz098

[48] 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, 2011. https://doi.org/10.1145/1961189.1961199

[49] 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. Ther. Nucleic Acids*, vol. 18, pp. 590–604, 2019. https://doi.org/10.1016/j.omtn.2019.09.019

[50] 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.*, vol. 157, pp. 752–758, 2019. https://doi.org/10.1016/j.ijbiomac.2019.12.009

[51] 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, pp. 225–234, 2020. https://doi.org/10.1007/s11103-020-00988-y

[52] 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, 2019. https://doi.org/10.1016/j.csbj.2019.06.024

[53] 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. Ther. Nucleic Acids*, vol. 16, pp. 733–744, 2019. https://doi.org/10.1016/j.omtn.2019.04.019

[54] 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," *Curr. Gene Ther.*, vol. 18, no. 6, pp. 386–392, 2018. https://doi.org/10.2174/1566523218666181105125245

[55] 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, p. e0179314, 2017. https://doi.org/10.1371/journal.pone.0179314

[56] 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," *Curr. Protein Pept. Sci.*, vol. 21, no. 16, pp. 1229–1241, 2020. https://doi.org/10.2174/1389203721666200117153412