Mathematical Biosciences and Engineering

MBE, 18(5): 6410–6429.
DOI: 10.3934/mbe.2021318
Received: 17 May 2021
Accepted: 06 July 2021
Published: 23 July 2021

http://www.aimspress.com/journal/MBE

Research article

TYLER, a fast method that accurately predicts cyclin-dependent proteins by using computation-based motifs and sequence-derived features

Jian Zhang[1,*], Xingchen Liang[1], Feng Zhou[1], Bo Li[2] and Yanling Li[1]

[1] School of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, China
[2] College of Electronic Science and Engineering, Jilin University, Changchun 130012, China

* **Correspondence:** Email: jianzhang@xynu.edu.cn.

Abstract: Cyclins and related cyclin-dependent kinases play vital roles in regulating the progression in the cell cycle. Understanding the intrinsic mechanisms of cyclins promises knowledge about cell uncontrolled proliferation and prevention of cancer cells. Therefore, accurate recognition of cyclins is important for the investigation of tumor cells and biomedical engineering. This study proposes a novel sequence-based predictor named TYLER (predicT cYcLin-dEpendent pRoteins) for addressing the long challenge problem of predicting cyclin-dependent proteins (CDPs). We use information theory to compute selectively enriched CDP-related motifs and build the motif-based model. For those proteins without sharing enriched motifs, we compute sequence-derived features and construct machine learning-based models. We optimize the weights of two different models to build a more accurate predictor. We estimate these two types of models by using 5-fold cross-validations on the TRAINING dataset. We prove that the combination of two models and optimization of the corresponding weights promises decent and robust results on both TRAINING and independent TEST dataset. The empirical test demonstrates that TYLER is robust predictor and statistically significantly better than current methods. The runtime assessment reveals TYLER is a high-throughput effective method. We use TYLER to make predictions on the human proteome, and use the results to hypothesize CDPs. The latest experimental verified CDPs and GO analysis proves that some of our novel predictions shall be potential CDPs. TYLER is implemented as a public user-friendly web server at http://www.inforstation.com/webservers/TYLER/. We share all data and source code that used in this research at https://github.com/biocomputinglab/TYLER.git.

Keywords: cyclin-dependent proteins; sequence motifs; machine learning; GO analysis

1. Introduction

Cyclins control the progression in the cell cycle. They usually interact with cyclin-dependent kinases (Cdks) to express function and affect cell period activity [1]. According to the differences in the function and period of activity, cyclins can be categorized into G1, S, G2, and M cyclins [2]. The growth and division of cells are controlled by the above-mentioned different cyclins. Cyclins, together with associated Cdks, allow passage from Gap 1 phase to DNA synthesis phase, Gap 2 phase, and mitosis phase [3–5]. Each of these phases is regulated by corresponding CDPs. Besides meiosis, cyclins and Cdk also active target proteins to regulate the cell cycle. Figure 1 indicates the complete phosphorylation of a target protein by the mediator cyclin-Cdk complex. In the beginning, the cyclins and Cdks are dissociated. An individual Cdk is inactive and cannot work properly [6]. The binding of the corresponding cyclin activates Cdk. It acts like a switch and allows the complex to modify target proteins [7]. The majority of the Cdks bind specific one or several cyclins, which corresponds to the functional specialization during evolution [8]. Then, the target protein attaches to the complex with both been phosphorylated, which results in changing the activity of the target protein [9]. In the end, the target protein no longer needs to attach cyclin-Cdk complex and the cyclin will be destroyed [10]. Cyclins and Cdks widely exist in various types of species. However, the number of Cdks that are used in the cell cycle is different: yeast has only one Cdk [10], while mammals usually have multiple Cdks [11].

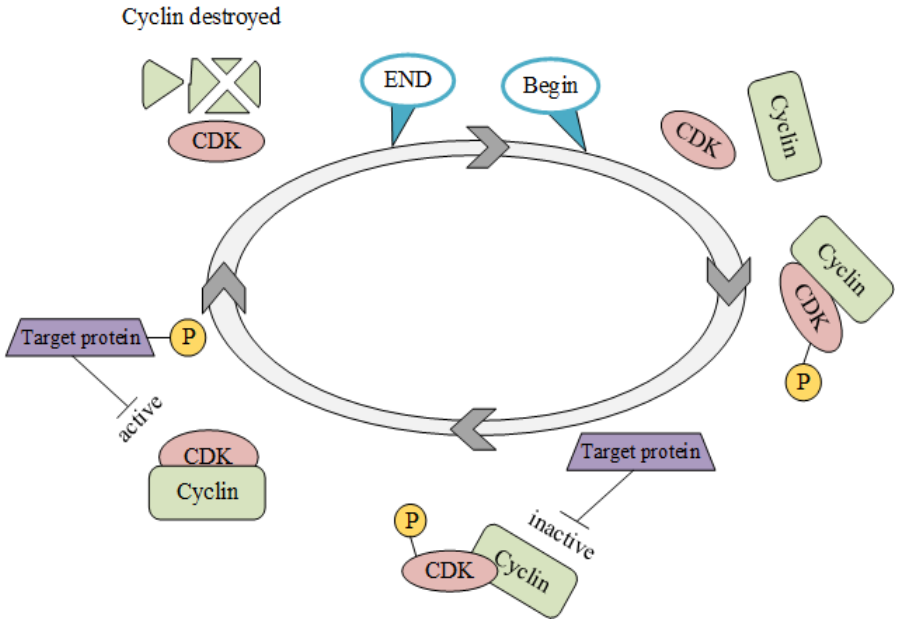

Figure 1. Control of cell cycles by cyclin and specifically related Cdk.

Cyclins play an indispensable role in the cell cycle [12]. If cyclins express their function in a deranged way, the inappropriate cell proliferation leads to tumors or cancers [13,14]. The majority of human tumors is proved to be associated with the alteration of the pRb pathway, which is usually caused by the overexpression of cyclin D1 [15]. E-type cyclins are thought to be pivotal for excessively activating Cdk2 and leading to the excessive proliferation of normal cells [16]. Understanding the intrinsic mechanisms of uncontrolled cell proliferation promises the knowledge of manipulation or

prevention of cancer cells [17–19]. For instance, Cyclin D1 and E2 are proved to be capable of detecting cancers as the diagnostic biomarkers [20,21].

Recent years have witnessed several computational methods for predicting CDPs. CyclinPred was the first predictor that predicted CDPs by using sequence-based features and a support vector machine [22]. Hassan Mohabatkar adopted the concept of Chou's pseudo amino acid composition [23]. The features used in their research included amino acid composition, weight factor, correlation and physicochemical properties. Saxena et al developed a hybrid module by using computed amino acid composition, dipeptide composition, and pseudo amino acid composition [24]. CDKIPred is designed to identify Cdk inhibitors. It used various compositional features and evolutionary information in the form of position-specific scoring matrix profiles [25].

The size of peptides/proteins in the cyclin family varies from ~30 to ~70 kDa [26]. Although most cyclins are sequence heterogeneity, they still share some common peptide patterns [27]. Besides that, the interaction with cyclins and corresponding Cdks needs their substrates to harbor a specific recognition motif [28,29]. Therefore, the conserved regions or motifs help to recognize cyclins. For instance, Gelais et al proved that the cyclin-binding R-L motif plays important role in modulating human SAMHD1 activity [30]. The glycine-rich (conserved '*G-G-G*') motif help to form activation loops [31]. The '*L-C-E*' motif is a cyclin-encoded pattern that assists in substrate recruitment [32]. The '*PEST*' motif is related to protein degradation [33], and the '*L-LL*' motif is involved in D-type cyclin transcriptional function [34]. Figure 2 illustrates three groups of CDPs that share common motifs. For instance, Q9YVL3, P21127, and Q07002 have three enriched motifs, i.e. '*LHRD-K*', '*DFGLA*', and '*V-TLWYR-P*'. P21117 and O76039 share '*V-T-WYR-P*' and '*DMWS*' motifs.

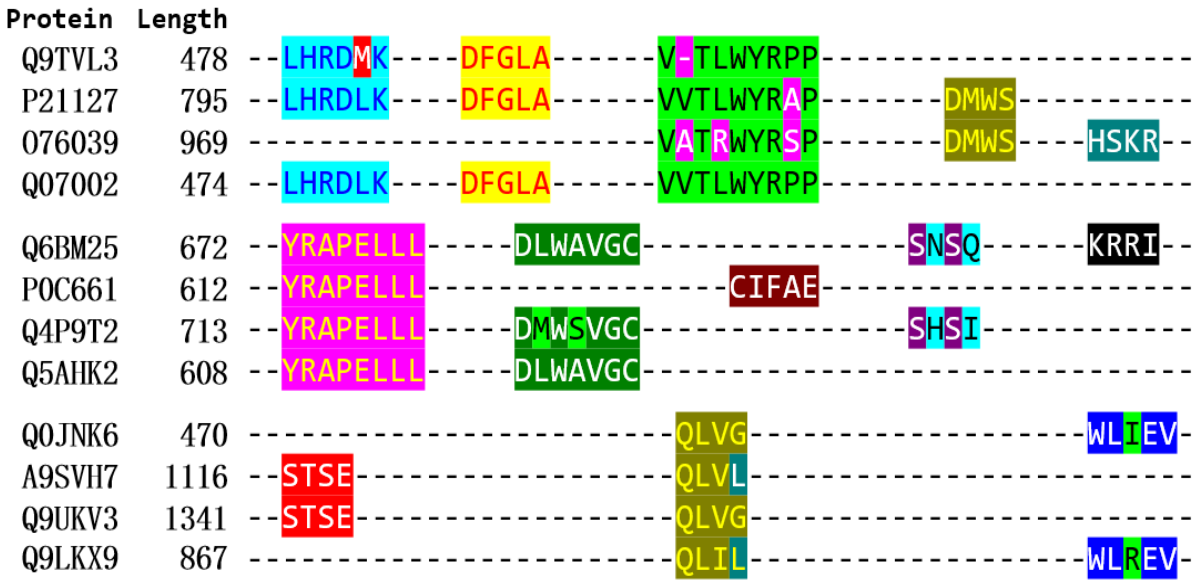

Figure 2. Three groups of CDPs that share common motifs.

These methods contributed to the knowledge of the cyclins and Cdks. However, our survey demonstrates several drawbacks of the current method. *First*, these methods used different types of features to decode CDPs without proving the effectiveness of the considered features. *Second*, cyclin-related motifs have been proved to widely exist and associated with the specific Cdk, helping to recognize related CDPs, which needs further investigation. *Third*, our survey reveals that the current methods are slow and not

suitable for large-scale predictions, which limits the application of these methods.

We introduce a novel method, TYLER (predicT cYcLin dEpendent pRoteins), that adopts motif- and machine learning-based models to accurately predict CDPs. The motif-based model computes the information gain rate to determine whether the query protein share enriched motifs that are related to CDPs. The machine learning model builds a general predictor. We optimize the weights between the motif- and the machine learning-based models. Blind test on the TEST dataset proves the efficiency of the proposed method. We also use TYLER to make predictions on the human proteome to explore potential CDPs. We implement the proposed method as a public user-friendly web server at http://www.information.com/webservers/TYLER/. We share all data and source code that used in this research at https://github.com/biocomputinglab/TYLER.git.

2. Materials and methods

2.1. Benchmark datasets

The CDPs and non-CDPs used in this study are sourced from Uniprot database [35] in October of 2020. We collect 2572 proteins with molecular function annotated as '*cyclin*', '*cyclin-dependent*' or biological process that includes '*cell cycle*' from Swissprot. To avoid data bias from the homology proteins, we use blastclust [36] to perform clustering for these proteins with a cutoff of 30%. We randomly pick one presentative from each cluster and obtain 1043 proteins. Next, we collect 1043 non-CDPs, which share less than 30% similarities with the above-mentioned CDPs. We randomly pick 700 CDPs and 700 non-CDPs to construct the TRAINING dataset, the remaining 343 CDPs and 343 non-CDPs are used as the TEST dataset. The data used in this study is publicly available at the TYLER web server.

2.2. Selective of enriched CDP-related sequence motifs

Protein motifs are short conserved regions of sequence which are shared among different proteins in the same family [37]. Motifs are usually associated with a distinct spatial structure, determinating a particular chemical or biological function [38]. Motifs promise insights for those newly found proteins with unknown functions. This study introduces information theory to compute motifs that are enriched in CDPs versus non-CDPs. First, we compute the original information entropy of CDPs and non-CDPs. Then, we iteratively generate a l-length peptide pattern. This research we define the length of the peptide pattern is between four and ten. The number of gaps in each considered pattern is less than four and the gap length is less than one. For the target pattern, we compute its appearance frequencies in CDPs and non-CDPs, respectively. We set a minimal frequency to avoid introducing too much computation. Next, we update the information entropy for each eligible pattern and compute the corresponding information gain. We use the information gain ratio to quantify the difference of the information gain between CDPs and non-CDPs. Particularly, to avoid data bias, we randomly select ten different negative samples (non-CDPs). Algorithm 1 details the strategy of the selection of enriched CDP-related sequence motifs. The final selected motifs are obtained by averaging the ranking of the IGR obtained in ten repeats.

Algorithm 1. The computation of the peptide/protein sequence motif.

Input: $D_C = \{S^+ | S_i^+ \in \{CDPs\}\},\ i \in \{1,2,3,\dots,m\};\ D_{nC}^{\tau} = \{S^- | S_i^- \in \{non-CDPs\}\},\ i \in \{1,2,3,\dots,m\},\ \tau \in \{1,2,3,\dots,10\}$.
Initalization: $L_{min} = 4$ (minimal length of motif), $L_{max} = 10$ (maximal length of motif), $T = 5$ (minimal frequency).
$H(D_C) = -\sum_i P(S_i^+) \cdot logP(S_i^+), H(D_{nC}^{\tau}) = -\sum_i P(S_i^-) \cdot logP(S_i^-)$.
Iteration:
Generate a sequence pattern p_t with the length of l, $L_{min} \le l \le L_{max}$, number of gaps $\le$ 4, gap length $\le$ 1.
Compute the occurrence frequency of p_t in D_C and D_{nC}^{τ}, i.e., $f(p_t, D_C)$ and $f(p_t, D_{nC}^{\tau})$.
if $f(p_t, D_C) > T > f(p_t, D_{nC}^{\tau})$
Update information entropy: $H'(D, p_t) = f(p_t, D) \cdot H(p_i|p_t) + (1 - f(p_t, D)) \cdot H(p_i|\overline{p_t})$
Compute the information gain: $IG(D, p_t) = H(D) - H'(D, p_t)$
Compute the information gain ratio: $IGR(p_t) = IG(D_C, p_t)/IG(D_{nC}, p_t)$
End if
End Iteration
Output: $motifs = \{p_1, p_2, \dots p_q\}$

2.3. Feature construction

Studies [39–43] have shown that proteins that share same functions tend to have common conserved protein regions. These evolutionary conserved regions can be quantified by using the position-specific scoring matrix (PSSM), which has been widely used in bioinformatics research [44–46]. This study we run MMseqs [47] against the swissprot database to obtain multi sequence alignments and compute PSSMs. Secondary structure is proved to be associated with the 3D structure and function of proteins [48]. Specifically, the spatial composition of the majority Cdks is a two-lobed structure [49]. It has two different sizes of lobes: the small amino-terminal lobe locates at the top and is composed of β-sheets, the large carboxy-terminal lobe locates at the bottom and is consists of α-helices. The active sites are located deeply in-between [49]. Some cyclins also show preference on certain secondary structure segments. For instance, the helix-loop-helix segment is essential for the expression of the secretin gene and related protein such as MyoD [50]. Cyclin D1 is proved to repress the expression of muscle-specific genes [51]. Amino acids are the basic building blocks of proteins. Different amino acids show various propensity in the microscopic environment. Considering this, this study introduces eleven physicochemical properties. These properties are used individually for encoding CDPs as well as incorporating with PSSMs to quantify the evolutionary substitutions among various groups of properties. We detail the computation of three types of features in the Table S1.

2.4. Evaluation criteria

The proposed method is evaluated by using six common used metrics, including sensitivity (SN), specificity (SP), precision (PRE), accuracy (ACC), F1-measure (F1), and Mathew's Correlation Coefficient (MCC). SN indicates the correctly predicted CDPs among all putative CDPs; SP means the correctly predicted non-CDPs among all putative non-CDPs; PRE quantifies the correctly predicted proteins among all putative samples; ACC stands for the correctly predicted proteins from all samples; F1 measures the harmonic mean of both precision and sensitivity; MCC balances the assessment in both the positive and negative samples.

$$SN = \frac{TP}{TP + FN} \tag{1}$$

$$SP = \frac{TN}{TN + FP} \tag{2}$$

$$PRE = \frac{TP}{TP + FP} \tag{3}$$

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{4}$$

$$F1 = 2 \times \frac{SN \times PRE}{SN + PRE} \tag{5}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}} \tag{6}$$

Besides above binary prediction criteria, this research also illustrates two curves to demonstrate the overall predictive quantity, namely receiver operating characteristic curve (ROC curve) and precision-recall curve (PR curve). ROC curve plots the TPR (true positive rate) against the FPR (false positive rate) at various thresholds. PR curve measures the PRE against TPR at various thresholds. Moreover, we also calculate the area under ROC curve (AUC), area under PR curve (AUPRC).

We train the model on the TRAINING dataset using 5-fold cross-validation. In detail, the TRAINING dataset is equally divided into 5 subsets. In each iteration, we use 4 out of 5 subsets to train the model and make predictions on the remaining one subset. The iteration repeats five times until each of the 5 subsets has been predicted. We estimate the results by computing the average of these five repeats. Besides that, we use the undersampling schema on the TEST dataset. We randomly select 50% samples from the TEST dataset ten times. The average performance of the ten times is treated as its final result to avoid potential data bias. Moreover, we also evaluate the significance of the differences in predictively quantify among various methods. For two lists of prediction results, we first use the Anderson-Darling test [52] to determine whether the data is normal distribution or not. We run the t-test for the data with < 0.05 significance under the Anderson-Darling test, otherwise we adopt the Wilcoxon rank sum test [53]. We set the cut-offs at 0.05 to judge the statistical significance.

2.5. Architecture of the TYLER predictor

Figure 3 illustrates the architecture of the proposed method. TYLER is designed with motif-based and machine learning-based models using the newly compiled benchmark dataset. For the motif-based model, we first compute the IG after introducing a new candidate pattern that has a frequency higher than the preset minimal. We compute IGR to quantify the relative value of the motifs and build the model by accumulating the IGRs. Although the motif-based model achieves decent results on some CDPs, it only covers parts of proteins. Considering this, we also introduce machine learning models. Specifically, we construct the feature space using sequence-derived features, including PSSM-based, physicochemical properties, and secondary structure-based features. Next, four different popular machine learning algorithms are adopted to construct models

and perform evaluations. We choose the logistic regression algorithm since it produces good results and is capable of building a high-throughput model. We optimize the weights between motif-based and machine learning-based models for aiming of constructing a more robust model. The proposed method is tested on the independent TEST dataset and is implemented as a web server.

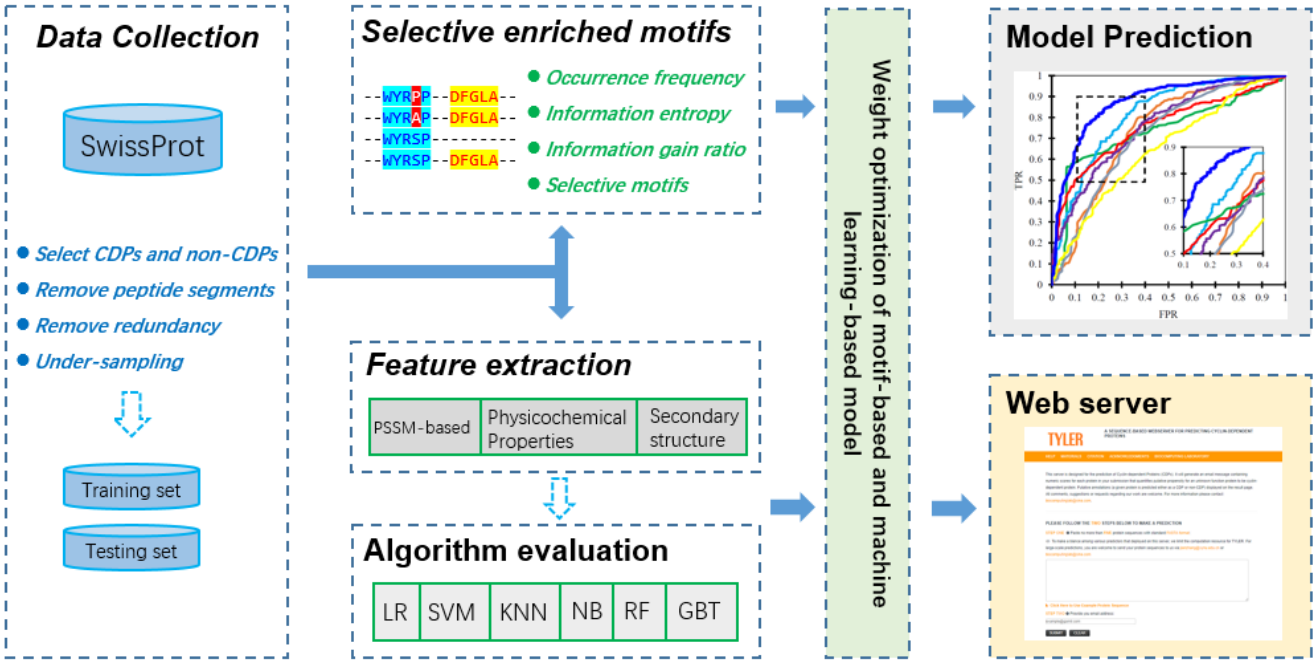

Figure 3. Overall framework of the proposed TYLER.

3. Results and discussion

3.1. Selective computational motifs and quantity assessment

This study uses information gain ratio to quantify the relative enrichments of peptide patterns (motifs). Shown in Table S2 is the complete list of the considered motifs. The motif with a higher IGR value indicates that it is more favored in the CDPs compared with that in the non-CDPs. For an unknown protein, we explore whether it shares the target CDP-related motifs. We quantify the enrichment by accumulating the IGR values. That is, the protein with more target motifs tends to have higher probabilities to be native CDPs. For comparison, we also generate a random predictor. Figure 4 evaluates the quantity of the selectively enriched motifs on the TRAINING and TEST datasets respectively. The motif-based method achieves the AUC of 0.742 and 0.736, and the AUPRC of 0.736 and 0.729 on the TRAINING and TEST datasets, respectively.

We observe a sharp rise in the starting (left bottom) region of the ROC curves of the predictions on both TRANING and TEST datasets. When setting the threshold where FPR = 10% (specificity = 90%), the motif-based model achieves the SN of 33 and 58% on the TRANING and TEST dataset, respectively. We further find that these motifs cover 391 (or 27.9%), and 194 (or 28.2%) CDPs on the TRAINING and TEST dataset. Moreover, about 12% proteins in the TEST dataset have more than four selectively motifs, while the fraction is only 5% for that in the TRAINING dataset. This explains why predictions on the TEST datasets show better performance on the low part of the ROC curve. This

experiment demonstrates that the selective computational motifs contribute to the recognition of CDPs. However, not all native CDPs share these selective motifs since we set the threshold of minimal appearance frequency to filter selection. This is the reason that the tails of blue and green curves are close to black curves in both sub plots.

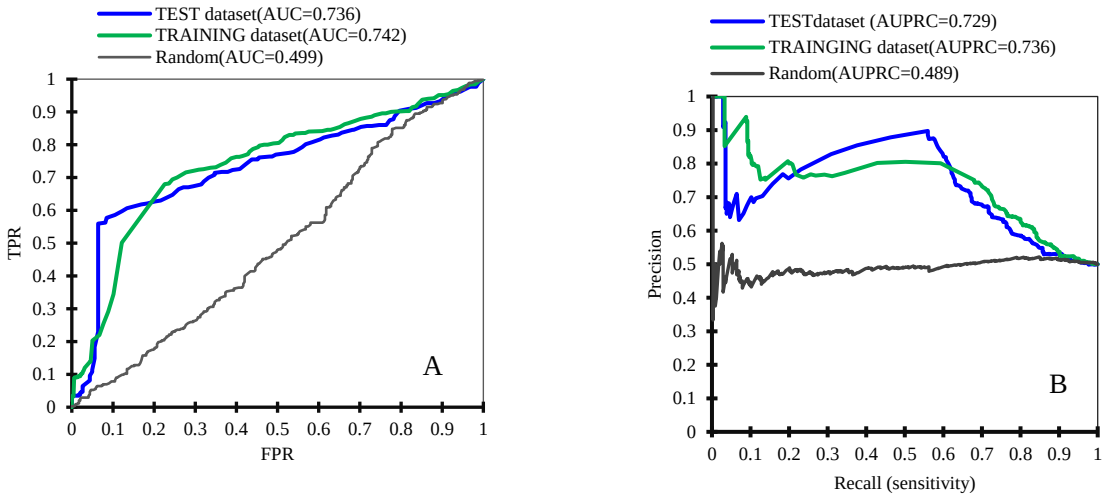

Figure 4. Quantity assessment of the selectively enriched motifs of CDPs.

3.2. Empirical analysis of the effectiveness of the considered features

Although the motif-based model produces decent results when compared with the random predictor, it has limitations when being used to identify proteins that have no selectively enriched cyclin-related motifs. This results in a general machine learning-based model, which provides further predictions for target unknown proteins. Here, we choose to utilize logistic regression to build the predictive model since it is effective and unlikely to trap overfitting [54,55]. Particularly, logistic regression is suitable for building high-throughput predictors for large-scale computations[54]. We use three types of features, including evolutionary profile, secondary structure, and physicochemical properties to encode the CDPs.

Table 1 lists the prediction results of different individual and combined features on the TRAINING dataset. Three types of individual features, namely EP, SS, and PC, produce the average AUC of 0.68~0.73, and MCC of 0.18~0.25 on the TRAINING dataset. Compared with individual features, the combination of two types of features obtains higher performance than individual features. Particularly, the combination of three types of features achieves the highest predictive results with the average AUC of 0.81, MCC of 0.40, SN of 0.45, PRE of 0.82 and F1 of 0.58, which is about 4.7~73.9% higher than the second-best EP + PC. The corresponding p-values prove the improvement are statistical significant (p-value < 0.026). Therefore, this study uses the combination of three types of features to construct the predictive model.

Table 1. Comparison of the predictive performance of different individual and combined features on the TRAINING dataset. The threshold is set where FPR equals 10% (SP = 0.90). The results are computed over the 5-fold cross-validation on the TRAINING dataset. We report the corresponding averages and standard deviations over the 5 subsets. The best performance for each measure is given in bold font. We also evaluate the statistical significance of these results, the differences with p-value < 0.05 are shown with italics font.

Features	SN	SP	PRE	ACC	F1	MCC	AUC
EP[1]	0.30± 0.02	0.90±0.00	0.75±0.01	0.60±0.01	0.43±0.02	0.25±0.02	0.73±0.01
	2.44E-06	1.00	*3.63E-06*	*2.46E-06*	*2.29E-06*	*2.49E-06*	*7.19E-09*
SS[2]	0.30±0.02	0.90±0.00	0.75±0.01	0.60±0.01	0.43±0.02	0.25±0.02	0.73±0.01
	2.91E-08	1.00	*2.27E-08*	*2.94E-08*	*1.62E-08*	*2.26E-08*	*4.57E-11*
PC[3]	0.23±0.01	0.90±0.00	0.70±0.00	0.57±0.01	0.35±0.01	0.18±0.01	0.68±0.00
	2.07E-08	1.00	*7.71E-10*	*2.13E-08*	*6.52E-09*	*1.09E-08*	*5.37E-11*
EP+SS	0.34±0.03	0.90±0.00	0.77±0.01	0.62±0.01	0.47±0.03	0.29±0.02	0.74±0.01
	3.30E-06	1.00	*3.25E-06*	*3.40E-06*	*2.78E-06*	*3.18E-06*	*1.07E-08*
EP+PC	0.43±0.02	0.90±0.00	0.81±0.01	0.66±0.01	0.56±0.02	0.37±0.02	0.80±0.00
	0.11	1.00	0.11	0.11	0.11	0.11	*8.28E-04*
SS+PC	0.28±0.01	0.90±0.00	0.74±0.01	0.59±0.01	0.41±0.01	0.23±0.01	0.70±0.00
	2.00E-07	1.00	*5.18E-08*	*2.06E-07*	*1.07E-07*	*1.49E-07*	*1.03E-10*
EP+SS+PC	**0.45±0.03**	**0.90±0.00**	**0.82±0.01**	**0.68±0.01**	**0.58±0.02**	**0.40±0.01**	**0.81±0.00**

[1]indicates evolutionary profile based features, [2]stands for secondary structure based features, and [3]means physicochemical properties based features.

3.3. Optimization of the weights between motif- and machine learning-based model

TYLER uses two different models to specifically recognizing CDPs. The motif-based model uses selected enriched cyclin-related motifs to quantify the propensities of an unknown protein being a CDP, while the machine learning-based model adopts a pre-constructed model to make predictions. Here, we set different groups of weights to balance the predictive power between these two models.

As shown in Table 2, when only use the motif-based model (w1 = 100%, and w2 = 0%), TYLER achieves the average SN of 0.33 and AUC of 0.74 on the TRAINING dataset. The introduction of a machine-learning model helps increase the recognition of those CDPs without selectively enriched motifs. Thus, we observe a rise of SN with the increase of w2. When the weight of the machine-learning model continues to increase and is higher than 50%, the average MCC and AUC values reveal a trend of gradual decline. When w1 and w2 are set as 0.4 and 0.6 respectively, TYLER achieves the highest results with the average SN of 0.65, PRE of 0.87, ACC of 0.78, F1 of 0.74, MCC of 0.57, and AUC of 0.87.

3.4. Comparison between TYLER and other methods

Besides the logistic regression that is used in this study, many other popular algorithms have been applied in bioinformatics. These algorithms include K nearest neighbors (KNN), support vector machine (SVM), naïve Bayes (NB), random forest (RF), and gradient boosted trees (GBT). KNN predicts an unknown sample by measuring the distance between the query example and the

current examples from the data [56]. SVM aims to explore a hyperplane that has the maximum margin to separate data. It uses kernel functions on non-linearity and high-dimension data [57]. NB is a probabilistic algorithm that assumes conditional independence between a feature and any other features [58]. It is suitable for large highly sophisticated data. RF is an ensemble learning method consisting of a number of decision trees [59]. GBT is another ensemble approach. However, GBT build trees one by one. The newly introduced tree helps to correct errors made by previous ones [60]. For a strict comparison, we first optimize the parameters for each algorithm on the TRAINING dataset by using 5-fold cross-validation. We use the algorithm and corresponding optimized parameters to train the model on the TRAINING dataset. Then, we perform predictions on the TEST dataset.

Table 2. Comparison of the predictive performance of different sets of weights on the motif- and machine learning-based layers on the TRAINING dataset using 5-fold cross-validation. Same as Table 1, we report the averages and standard deviations for each measure. The best performance is given in bold font, and the differences with statistical significance (p-value < 0.05) are shown with italics font. w1 indicates the weight of the motif-based model, while w2 means that of the machine-learning model.

Weights		SN	SP	PRE	ACC	F1	MCC	AUC
w1=1.0	w2=0.0	0.33±0.02	0.90±0.00	0.74±0.01	0.67±0.01	0.47±0.02	0.28±0.03	0.74±0.01
		2.10E-09	1.00	*4.48E-08*	*2.12E-09*	*5.06E-09*	*3.15E-09*	*2.69E-09*
w1=0.9	w2=0.1	0.59±0.03	0.90±0.00	0.85±0.01	0.74±0.01	0.69±0.02	0.51±0.02	0.84±0.01
		3.65E-03	1.00	*4.41E-03*	*3.62E-03*	*3.96E-03*	*3.62E-03*	*9.65E-06*
w1=0.8	w2=0.2	0.58±0.02	0.90±0.00	0.85±0.00	0.74±0.01	0.69±0.01	0.51±0.01	0.85±0.01
		3.74E-05	1.00	*3.44E-05*	*3.60E-05*	*3.52E-05*	*3.71E-05*	*2.43E-04*
w1=0.7	w2=0.3	0.58±0.01	0.90±0.00	0.85±0.00	0.74±0.01	0.69±0.01	0.51±0.01	0.85±0.00
		1.09E-04	*1.00*	*1.41E-04*	*1.08E-04*	*1.18E-04*	*1.08E-04*	*1.32E-04*
w1=0.6	w2=0.4	0.59±0.02	0.90±0.00	0.85±0.01	0.75±0.01	0.70±0.02	0.52±0.02	0.86±0.00
		1.15e-03	1.00	*1.60E-03*	*1.14E-03*	*1.32E-03*	*1.14E-03*	*8.42E-03*
w1=0.5	w2=0.5	0.58±0.02	0.90±0.00	0.85±0.00	0.74±0.01	0.69±0.02	0.50±0.02	0.86±0.01
		1.28E-04	1.00	*1.92E-04*	*1.28E-04*	*1.49E-04*	*1.28E-04*	*4.58E-03*
w1=0.4	w2=0.6	**0.65±0.02**	0.90±0.00	**0.87±0.00**	**0.78±0.01**	**0.74±0.01**	**0.57±0.02**	**0.87±0.01**
w1=0.3	w2=0.7	0.60±0.02	0.90±0.00	0.86±0.00	0.75±0.01	0.71±0.01	0.53±0.02	0.87±0.01
		0.02	1.00	*0.02*	*0.02*	*0.02*	*0.02*	0.85
w1=0.2	w2=0.8	0.61±0.03	0.90±0.00	0.86±0.01	0.76±0.01	0.72±0.02	0.54±0.03	0.87±0.01
		0.20	1.00	0.17	0.20	0.19	0.20	0.62
w1=0.1	w2=0.9	0.61±0.01	0.90±0.00	0.86±0.00	0.76±0.00	0.71±0.01	0.53±0.01	0.87±0.00
		2.90E-03	1.00	*2.75E-03*	*2.78E-03*	*2.88E-03*	*2.88E-03*	0.34
w1=0.0	w2=1.0	0.45±0.03	0.90±0.00	0.82±0.01	0.68±0.01	0.58±0.02	0.40±0.01	0.81±0.00
		4.13E-08	1.00	*1.45E-07*	*4.12E-08*	*6.11E-08*	*4.55E-08*	*1.12E-08*

Table 3. Comparison of the predictive performance of different methods on the TEST dataset. The best performance for each measure (average ± standard deviation) is shown in bold font. We compute the differences between each method and the best one. The differences with statistical significance (p-value < 0.05) are shown with italics font.

Methods	SN	SP	PRE	ACC	F1	MCC	AUC
Motif	0.58±0.02	0.90±0.00	0.85±0.01	0.74±0.01	0.69±0.02	0.51±0.02	0.72±0.01
	8.14E-03	1.00	0.097	*0.033*	*0.031*	*0.035*	*1.19E-16*
SVM	0.24±0.09	0.90±0.00	0.69±0.09	0.57±0.04	0.36±0.11	0.19±0.10	0.71±0.08
	3.45E-10	1.00	*2.23E-05*	*4.19E-10*	*1.22E-08*	*6.37E-09*	*1.29E-05*
KNN	0.44±0.03	0.90±0.00	0.82±0.01	0.67±0.01	0.57±0.03	0.38±0.03	0.73±0.02
	8.57E-10	1.00	*1.92E-09*	*2.30E-09*	*1.59E-09*	*2.48E-09*	*3.95E-15*
NB	0.39±0.04	0.90±0.00	0.79±0.02	0.64±0.02	0.52±0.04	0.33±0.03	0.75±0.01
	9.75E-11	1.00	*1.62E-10*	*1.47E-10*	*1.90E-10*	*1.51E-10*	*2.44E-16*
RF	0.19±0.05	0.90±0.00	0.66±0.06	0.55±0.03	0.30±0.07	0.13±0.06	0.70±0.02
	8.98E-14	1.00	*3.31E-09*	*1.12E-13*	*1.41E-12*	*1.04E-12*	*1.33E-16*
GBT	0.29±0.04	0.90±0.00	0.74±0.02	0.60±0.02	0.42±0.04	0.24±0.04	0.71±0.02
	1.37E-13	1.00	*8.39E-12*	*2.81E-13*	*2.91E-13*	*4.40E-13*	*4.72E-14*
LR	0.45±0.02	0.90±0.00	0.82±0.01	0.67±0.01	0.58±0.02	0.39±0.02	0.83±0.01
	7.00E-10	1.00	*7.91E-10*	*2.07E-09*	*1.10E-09*	*2.18E-09*	*5.49E-10*
TYLER	**0.62±0.04**	**0.90±0.00**	**0.86±0.01**	**0.76±0.03**	**0.72±0.03**	**0.54±0.04**	**0.87±0.01**

Shown in Figure S2 are the ROC and PRC curves of different methods on the TRAINING dataset. When only using motif-based models, we obtain the AUC and AUPRC of 0.742 and 0.736, respectively. Six types of machine learning algorithms produce the AUC between 0.688 and 0.817, and the AUPRC between 0.669 and 0.789. RF and GBT are both ensemble learning algorithms and yield similar AUC values of 0.80. Among machine learning algorithms, LR secures the highest AUC of 0.817, and RF achieves the highest AUPRC of 0.800. Next, we use the optimized model to perform the independent test on the TEST dataset. As listed in Table 3, the motif-based model yields an SN of 0.58, which is even higher than the four machine learning models. Our previous observation demonstrates that about 28.2% of proteins on the TEST dataset have selectively enriched cyclin-related motifs. However, the AUC of the motif-based model is only about 0.72, which is similar or slightly lower than other machine learning algorithms. Six machine learning models produce the SN values between 0.19 and 0.45, and the AUC scores between 0.70 and 0.83. LR, RF and GBT obtain decent results with the AUC > 0.8 and AUPRC > 0.78. We observe that RF and GBT models produce a much worse performance on the TEST dataset compared to that on the TRAINING dataset (AUC = 0.72 vs. 0.81 for RF, and AUC = 0.70 vs. 0.80 for GBT). Considering the small number of samples in the TRAINING dataset, we speculate that RF and GBT have strapped into the overfitting. While LR secures similar good performance both on the TRAINING and TEST dataset. Therefore, we use LR to construct the machine learning-based model.

We empirically find that when combining motif-based and LR-based models, TYLER gives out the highest results both on the TRAINING and the TEST dataset. Specifically, TYLER secures the AUC of 0.879 and the AUPRC of 0.807 on the TRAINING dataset using 5-fold cross-validation (Figure S2). It also achieves the highest performance on the TEST dataset with the average SN = 0.62, PRE = 0.86, ACC = 0.76, MCC = 0.54, and AUC = 0.87. Figure 5 shows the ROC and PRC curves of various methods. TYLER achieves the highest TPR over the entire range of FPR (Figure 5A), and the highest PRE over the

entire range of SN (Figure 5B), both with wide margins ahead of the second-best method.

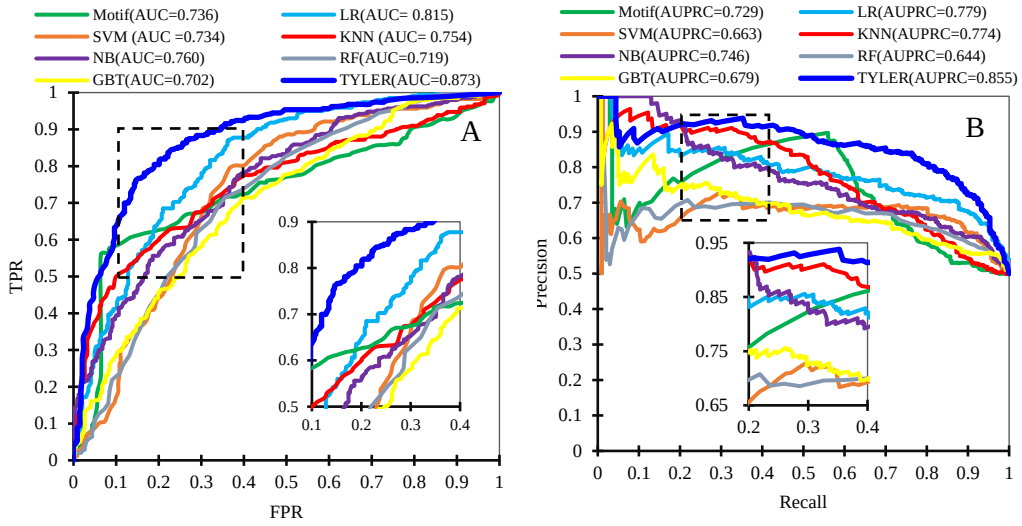

Figure 5. ROC curves and PRC curves for different methods on the TEST dataset.

3.5. Comparison between TYLER and state-of-the-art methods

We focus on comparative assessment on the methods that identify CDPs. After a comprehensive survey, we choose CyclinPred and cdkipred, which are the only two current public available predictors, to perform the independent test. We note that CyclinPred and cdkipred only produce binary predictions. In other words, they predict whether an unknown protein is a CDP or not without giving the probabilities. Therefore, we mainly use binary assessment to evaluate the considered methods. Particularly, for a consistent comparison, we adjust the thresholds for TYLER to obtain similar SP values as CyclinPred and cdkipred, respectively.

Table 4 reveals that TYLER outperforms CyclinPred and cdkipred on the TEST dataset. Generally, CyclinPred shows better performance than cdkipred. It is because that CyclinPred was designed for predicting cyclin protein sequences, while cdkipred mainly predicted cyclin/cyclin-dependent kinase inhibitors. TYLER secures the average SN of 0.490 and 0.457 with specificity = 0.938 and 0.947, respectively. By contrast, CyclinPred and cdkipred recognize less than 28% CDPs. We note that TYLER achieves the better F1 scores (0.629 vs. 0.416 with p-value = 1.68E-04, and 0.601 vs. 0.228 with p-value = 4.63E-06) and MCC values (0.481 vs. 0.286 with p-value = 7.57E-05, and 0.466 vs. 0.137 with p-value = 1.77E-06) compared to CyclinPred and cdkipred. This empirical test demonstrates that TYLER is statistically significantly better than current predictors for the identification of CDPs.

Table 4. Comparison of the predictive performance between TYLER and state-of-the-art methods on the TEST dataset.

Methods	SN	SP	PRE	ACC	F1	MCC
CyclinPred	0.280±0.040	0.936±0.022	0.815±0.057	0.608±0.023	0.416±0.048	0.286±0.054
TYLER	0.490±0.068	0.938±0.032	0.891±0.024	0.714±0.030	0.629±0.052	0.481±0.026
cdkipred	0.136±0.022	0.944±0.022	0.713±0.079	0.540±0.014	0.228±0.032	0.137±0.048
TYLER	0.457±0.088	0.947±0.027	0.902±0.031	0.702±0.030	0.601±0.070	0.466±0.035

3.6. Analysis of predictive performance on datasets with different distributions of CDPs vs. non-CDPs

In living organisms, the number of CDPs is much smaller than that of the non-CDPs. The distribution of CDPs vs. non-CDPs might somehow influence the predictive performance of the proposed method. To test the robustness of TYLER on unknown numbers and distributions of CDPs, we further design several testing datasets with various ratios of CDPs vs. non-CDPs. First, we randomly pick 100 CDPs from our source CDPs (after removing the proteins in the TRAINING dataset). Second, we randomly pick 200, 400, 600, 800, and 1000 proteins from the Swissprot (after removing the proteins in the source CDPs). The under-sampling procedure repeats five times to avoid potential data bias.

As listed in Table 5, with the ratio of negative samples to positive samples increases, we observe a gradual decline in PRE (from average 0.85 to 0.38), F1 (from average 0.68 to 0.46), and MCC (from average 0.50 to 0.39). However, we notice an interesting phenomenon, which is that the SN doesn't show an obviously decreasing trend accordingly. TYLER secures good SN values in different ratios of CDPs vs. non-CDPs. We attribute this good generalization to the introduction of the motif-based model. Our empirical analysis demonstrates that the motif-based model still shows good capability in recognizing CDPs given high SP scores. More importantly, the AUC fluctuates in a small range (from the lowest 0.84 to the highest 0.86). Overall, we prove that TYLER reveals good generalization and can be used in practical applications.

Table 5. Analysis of predictive performance on datasets with different distribution of CDPs vs. non-CDPs.

Ratio (CDPs vs. non-CDPs)	SN	SP	PRE	ACC	F1	MCC	AUC
1: 1	0.58±0.10	0.90±0.00	0.85±0.02	0.74±0.05	0.68±0.07	0.50±0.08	0.86±0.03
1: 2	0.55±0.10	0.90±0.00	0.73±0.04	0.78±0.03	0.63±0.08	0.49±0.09	0.84±0.03
1: 4	0.47±0.08	0.90±0.00	0.54±0.04	0.81±0.02	0.50±0.06	0.39±0.07	0.84±0.02
1: 6	0.51±0.11	0.90±0.00	0.45±0.05	0.84±0.02	0.48±0.08	0.39±0.09	0.84±0.02
1: 8	0.57±0.05	0.90±0.00	0.41±0.02	0.86±0.01	0.48±0.03	0.41±0.04	0.86±0.01
1: 10	0.56±0.05	0.90±0.00	0.38±0.02	0.87±0.01	0.46±0.03	0.39±0.04	0.85±0.01

3.7. Comparative evaluation of runtime

Runtime is an important criterion for assessing a high-throughput method, particularly when considering proteome-level predictions. According to our investigations, the most time-consuming computation is related to evolutionary profiles, which are obtained by PSI-BLAST [36] for most current methods. In this study we use much faster MMseqs [47] to compute the same information. Here, we consider the complete computation time of TYLER (including motif-based and machine learning-based models) and use the runtime of PSI-BLAST alone to estimate the computation time for other methods. All computations are performed using the same hardware and operating system (PC with i5 CPU, 8GB RAM, and Ubuntu 18.04 64-bit OS) allowing a direct compare of the results.

Figure 6 illustrates the comparison of runtime between TYLER and the lower bound of the other methods. We randomly picked 50 proteins for each chain length interval from less than 100 residues to 900 to 1000 residues. TYLER offers between 12.9 and 46.4 times faster runtime than other methods. Generally, both runtime measurements scale linearly with the protein length, with TYLER having the

smaller/better slope of the linear fit. We quantify the correlation between the linear fit and the scatters, and obtain the PCC of 0.59 and 0.68 for PSI-BLAST-based methods and our TYLER, respectively. For an average of 300-residues-length protein, TYLER costs about 2 seconds, while PSI-BLAST uses more than 70 seconds. Considering the runtime of PSI-BLAST is just the lower bounds of current methods, these results indisputably show that TYLER offers substantially faster predictions.

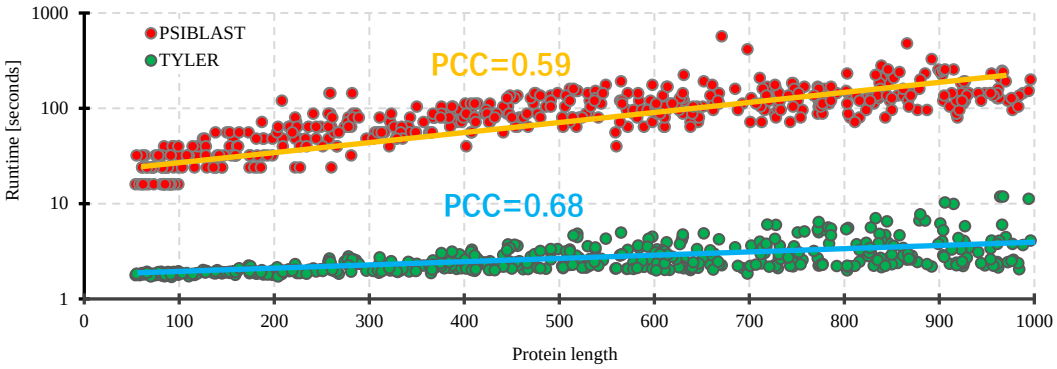

Figure 6. Scatter plot of the runtime of TYLER and PSI-BLAST (lower bound of the other methods). We run PSI-BLAST to compute PSSMs by using the SwissProt database with 3 iterations. TYLER runs MMseqs by using the same database and computes secondary structure by adopting fast PSI-PRED. Each point illustrates the runtime and protein length. The yellow and blue lines represent linear fit into the data for PSI-BLAST and TYLER, respectively. We also report the corresponding values of the Pearson correlation coefficient (PCC).

3.8. Prediction and validation of putative CDPs on the human proteome

The experiment in Section 3.5 proves the effectiveness of the proposed method. Here, we use TYLER to make predictions on human proteome. We define a putative CDP with two criteria, the query protein shares at least two cyclin-related motifs and the putative probability is higher than the threshold where the false positive rate equals 1% on the TRANING dataset. Finally, TYLER predicts 867 cyclins, which occupy about 4.3% of the human proteome. Next, we compare our predictions with current known CDPs.

First, we search SwissProt to find 479 proteins that are annotated as CDPs in the human proteome. Second, we search for 60 CDPs from the Reactome [61], which stores signaling and metabolic molecules and their relations organized into biological pathways and processes. Third, we search the Pfam database [62] and collect 77 CDPs. After combining these CDPs and removing overlap, we establish a final set of 507 verified CDPs.

Among these current known CDPs in the human proteome, TYLER successfully recognizes 109 or 21.5% (109/507 = 21.5%) of them. We compare our predictions with a randomized baseline and find an overlap of random 867 proteins and the 507 verified CDPs. Sampling is repeated 1000 times to avoid data bias and establish confidence intervals. The corresponding average and standard deviation for the overlap of the baseline are 4.3% ± 0.9%. In contrast, the above-mentioned 21.5% overlap produced by TYLER is about 4.9 times larger (p-value < 0.001) than the average of the baseline.

We use PANTHER [63] to perform three types of GO analysis, including cellular components, molecular functions, and biological processes.

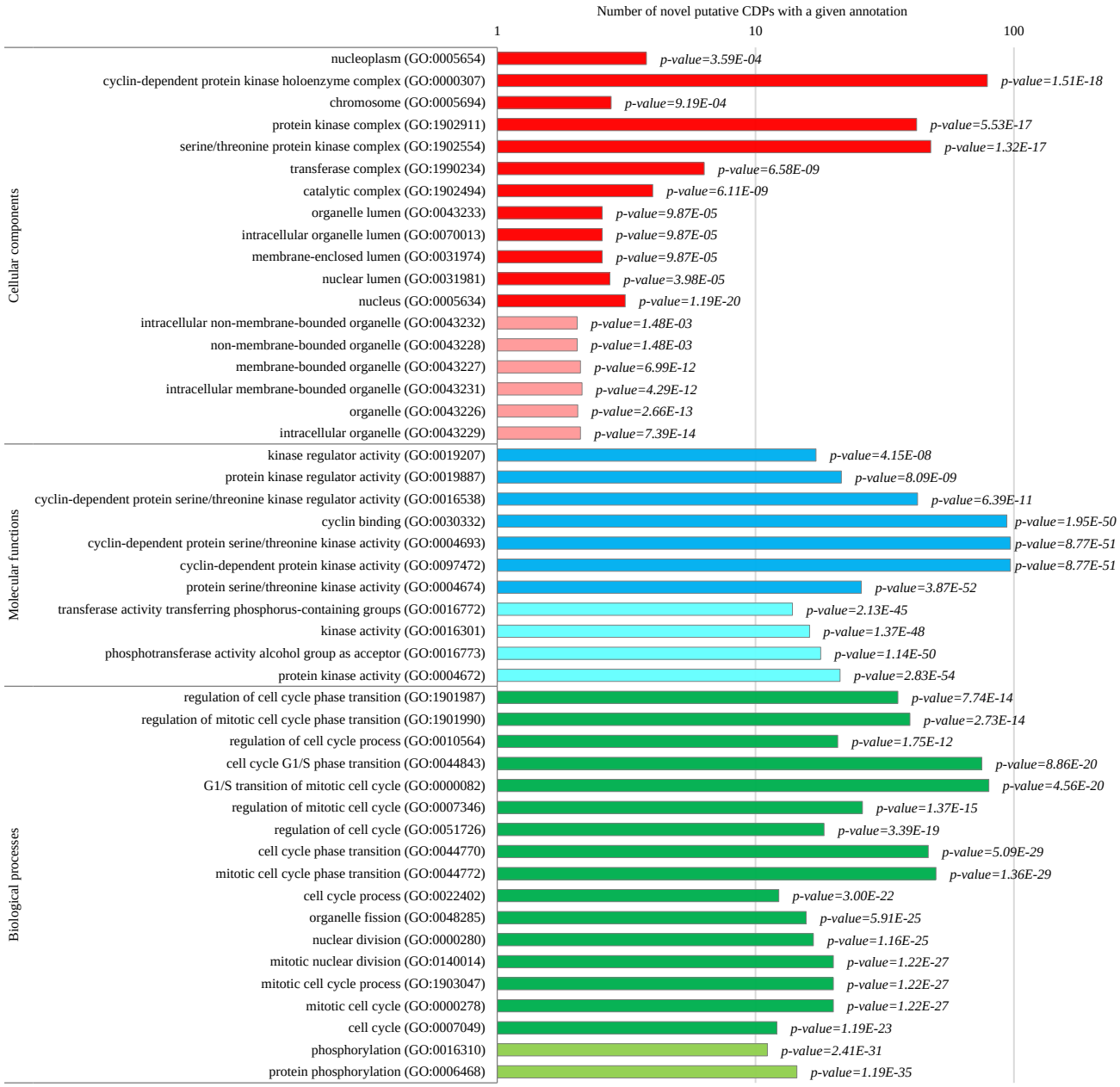

Figure 7. Geno ontology analysis of the novel putative CDPs. Cellular components, molecular functions, and biological processes are shown in the top, middle, and bottom of the figure, respectively. The dark red, dark blue, and dark green represent the significantly enriched GO items are shared with the verified CDPs. The light red, light blue, and light green indicates the corresponding GO items exist only in putative CDPs.

Figure 7 illustrates the subcellular locations and functions that are significantly enriched in the putative CDPs. The cellular components show that the putative CDPs are significantly involved in the cyclin-dependent protein kinase holoenzyme complex, protein kinase complex, and serine/threonine protein kinase complex, which are overlap with that of current known CDPs. Molecular functions

demonstrated that the novel putative CDPs participate in the activity related to various types of kinase regulators, cyclin-dependent protein kinase, and cyclin binding, which is consistent with that for the verified CDPs. When considering the biological processes, we notice that 16 out of 18 enriched cellular components of the novel putative CDPs are involved in the cell cycle. This is again evidence that some of the novel predictions shall be potential CDPs. Details of the computation and analysis are given in the Supplement. In comparison, we also use the same analysis for a random set that contains 867 randomly picked proteins. There are no significantly enriched cellular components, molecular functions, and biological processes. We share the complete set of the putative CDPs on the TYLER website.

3.9. TYLER webserver

We implement the proposed novel method TYLER as a free public available user-friendly web server at http://www.inforstation.com/webservers/TYLER/. Since TYLER is based on protein primary sequence, it requires only requires FASTA-formatted protein sequences as input. A single query can submit up to ten protein sequences. TYLER produces the predictions of motif-based and machine learning-based models, respectively. Besides that, the server will list all enriched motifs for the query proteins. After computation, the server sends emails to the users as well as shows results in the browser window. Moreover, we also provide large-scale proteome-level calculation services.

4. Conclusions

This study aims to propose a novel high-throughput computation-based predictor for the recognition of CDPs. Our investigation finds that CDPs share some common sequence patterns, which can be quantified by using information theory. Then, the selective CDP motifs are used to compute the propensity of an unknown protein being potential CDP. The motif-based model shows decent predictions on both TRAINING and TEST dataset. However, the motif-based predictor cannot cover all CDPs. Therefore, we also introduce a machine-learning model for further predicting CDPs by using sequence-derived features, including evolutionary conservation profile, secondary structure, and physicochemical properties. We prove that the combination of different types of features contributes a better predictor than individual features. Besides that, we also optimize the weights between the motif-based and machine learning-based model to achieve decent performance. The empirical test on the TEST dataset proves that the prosed method, named TYLER, is statistically significantly better than current methods. We also prove that TYLER is featured by good generalization and can be widely used in practical applications. TYLER is a high-throughput predictor, which offers at least 12 times faster runtime than current methods. We use TYLER to make predictions on the human proteome, and use the results to hypothesize CDPs and compare them with current known CDPs. The GO analysis suggests that at least some of the novel predictions constitute promising leads to recognize previously unknown CDPs. We share all data used in this work and implement the method as a publicly available web server.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 61802329), by the Innovation Team Support Plan of University Science and Technology of Henan

Province (grant 19IRTSTHN014), by the Key Science Research Project of Colleges and Universities in Henan Province (No.21A520039), and by Nanhu Scholars Program for Young Scholars of the Xinyang Normal University.

Conflict of interest

The authors declare that they have no competing interests.

References

1. S. Lim, P. Kaldis, Cdks, cyclins and CKIs: roles beyond cell cycle regulation, *Development*, **140** (2013), 3079–3093.
2. M. C. Horne, G. L. Goolsby, K. L. Donaldson, D. Tran, M. Neubauer, A. F. Wahl, Cyclin G1 and cyclin G2 comprise a new family of cyclins with contrasting tissue-specific and cell cycle-regulated expression, *J. Biol. Chem.*, **271** (1996), 6050–6061.
3. D. Fisher, Control of DNA replication by cyclin-dependent kinases in development, *Results Probl. Cell Differ.*, **53** (2011), 201–217.
4. N. Bouftas, K. Wassmann, Cycling through mammalian meiosis: B-type cyclins in oocytes, *Cell Cycle*, **18** (2019), 1537–1548.
5. G. Zheng, H. Yu, Cyclin a turns on bora to light the path to mitosis, *Dev. Cell*, **45** (2018), 542–543.
6. A. W. Murray, Recycling the cell cycle: cyclins revisited, *Cell*, **116** (2004), 221–234.
7. D. W. Stacey, Cyclin D1 serves as a cell cycle regulatory switch in actively proliferating cells, *Curr. Opin. Cell Biol.*, **15** (2003), 158–163.
8. D. J. Wood, J. A. Endicott, Structural insights into the functional diversity of the CDK-cyclin family, *Open Biol.*, **8** (2018), 180112.
9. M. Örd, K. Möll, A. Agerova, R. Kivi, I. Faustova, R. Venta, et al., Multisite phosphorylation code of CDK, *Nat. Struct. Mol. Biol.*, **26** (2019), 649–658.
10. P. Gutiérrez-Escribano, P. Nurse, A single cyclin-CDK complex is sufficient for both mitotic and meiotic progression in fission yeast, *Nat. Commun.*, **6** (2015), 1–13.
11. C. Gérard, A. Goldbeter, From quiescence to proliferation: Cdk oscillations drive the mammalian cell cycle, *Front. Physiol.*, **3** (2012), 413.
12. G. Bertoni, Cell cycle regulation by chlamydomonas cyclin-dependent protein kinases, *Plant Cell*, **30** (2018), 271.
13. M. Stamatakos, V. Palla, I. Karaiskos, K. Xiromeritis, I. Alexiou, I. Pateras, et al., Cell cyclins: triggering elements of cancer or not?, *World J. Surg. Oncol.*, **8** (2010), 111.
14. L. C. Leal-Esteban, L. Fajas, Cell cycle regulators in cancer cell metabolism, *Biochem. Biophys. Acta, Mol. Basis Dis.*, **1866** (2020), 165715.
15. E. A. Musgrove, C. E. Caldon, J. Barraclough, A. Stone, R. L. Sutherland, Cyclin D as a therapeutic target in cancer, *Nat. Rev. Cancer*, **11** (2011), 558–572.
16. Y. Geng, W. Michowski, J. M. Chick, Y. E. Wang, M. E. Jecrois, K. E. Sweeney, et al., Kinase-independent function of E-type cyclins in liver cancer, *Proc. Nat. Acad. Sci. U. S. A.*, **115** (2018), 1015–1020.
17. R. V. Dross, P. J. Browning, J. C. Pelling, Do truncated cyclins contribute to aberrant cyclin expression in cancer?, *Cell Cycle*, **5** (2006): 472–477.

18. C. Sanchez-Martinez, L. M. Gelbert, M. J. Lallena, A. de Dios, Cyclin dependent kinase (CDK) inhibitors as anticancer drugs, *Bioorg. Med. Chem. Lett.*, **25** (2015), 3420–3435.
19. W. Rozpędek, D. Pytel, A. Nowak-Zduńczyk, D. Lewko, R. Wojtczak, J. A. Diehl, et al., Breaking the DNA damage response via serine/threonine kinase inhibitors to improve cancer treatment. *Curr. Med. Chem.*, **26** (2019), 1425–1445.
20. J. A. Diehl, Cycling to cancer with cyclin D1, *Cancer Biol. Ther.*, **1** (2002), 226–231.
21. J. K. Kim, J. A. Diehl, Nuclear cyclin D1: an oncogenic driver in human cancer, *J. Cell Physiol.*, **220** (2009), 292–296.
22. M. K. Kalita, U. K. Nandal, A. Pattnaik, A. Sivalingam, G. Ramasamy, M. Kumar, CyclinPred: a SVM-based method for predicting cyclin protein sequences, *PLoS One*, **3** (2008), e2605.
23. H. Mohabatkar, Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein Pept. Lett.*, **17** (2010), 1207–1214.
24. Saxena, K. Pant, B. Pant, N. Adlakha, Hybrid based SVM model for prediction of CDKs and cyclins, in *2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)*, **5** (2010), 504–508.
25. J. Ramana, D. Gupta, Machine learning methods for prediction of CDK-inhibitors, *Plos One*, **5** (2010), e13357.
26. P. Loyer, J. H. Trembley, Roles of CDK/Cyclin complexes in transcription and pre-mRNA splicing: Cyclins L and CDK11 at the cross-roads of cell cycle and regulation of gene expression, *Semin. Cell Dev. Biol.*, **107** (2020), 36–45.
27. S. Bandyopadhyay, S. Bhaduri, M. Örd, N. E. Davey, M. Loog, P. M. Pryciak, Comprehensive analysis of G1 cyclin docking motif sequences that control CDK regulatory potency in vivo, *Curr. Biol.*, **30** (2020), 4454–4466.
28. D. Y. Takeda, J. A. Wohlschlegel, A. Dutta, A bipartite substrate recognition motif for cyclin-dependent kinases, *J. Biol. Chem.*, **276** (2001), 1993–1997.
29. J. A. Wohlschlegel, B. T. Dwyer, D. Y. Takeda, A. Dutta, Mutational analysis of the Cy motif from p21 reveals sequence degeneracy and specificity for different cyclin-dependent kinases, *Mol. Cell. Biol.*, **21** (2001), 4868–4874.
30. C. S. Gelais, S. H. Kim, V. V. Maksimova, O. Buzovetsky, K. M. Knecht, C. Shepard, et al., A cyclin-binding motif in human SAMHD1 is required for its HIV-1 restriction, dNTPase activity, tetramer formation, and efficient phosphorylation, *J. Virol.*, **92** (2018), e01787–17.
31. D. J. Wood, J. A. Endicott, Structural insights into the functional diversity of the CDK-cyclin family, *Open Biol.*, **8** (2018), 180112.
32. M. W. Landis, N. E. Brown, G. L. Baker, A. Shifrin, M. Das, Y. Geng, et al., The LxCxE pRb interaction domain of cyclin D1 is dispensable for murine development, *Cancer Res.*, **67** (2007), 7613–7620.
33. I. Quilis, J. C. Igual, A comparative study of the degradation of yeast cyclins Cln1 and Cln2, *FEBS Open Bio.*, **7** (2017), 74–87.
34. M. Wei, Q. He, Z. Yang, Z. Wang, Q. Zhang, B. Liu, et al., Integrity of the LXXLL motif in Stat6 is required for the inhibition of breast cancer cell growth and enhancement of differentiation in the context of progesterone, *BMC Cancer*, **14** (2014), 1–17.
35. A. Bateman, UniProt: a worldwide hub of protein knowledge, *Nucleic Acids Res.*, **47** (2019), D506–d515.

36. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 25 (1997), 3389–3402.
37. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif, *Bioinformatics*, **27** (2011), 1017–1018.
38. M. U. Johansson, V. Zoete, O. Michielin, N. Guex, Defining and searching for structural motifs using DeepView/Swiss-PdbViewer, *BMC Bioinf.*, **13** (2012), 1–11.
39. J. Zhang, H. Chai, G. Yang, Z. Ma, Prediction of bioluminescent proteins by using sequence-derived features and lineage-specific scheme, *BMC Bioinf.*, **18** (2017), 1–13.
40. H. Chai, J. Zhang, Identification of mammalian enzymatic proteins based on sequence-derived features and species-specific scheme, *IEEE Access*, **6** (2018), 8452–8458.
41. J. Zhang, H. Chai, S. Guo, H. Guo, Y. Li, High-throughput identification of mammalian secreted proteins using species-specific scheme and application to human proteome, *Molecular*, **23** (2018), 1448.
42. J. Zhang, Y. Zhang, Z. Ma, In-silico prediction of human secretory proteins in plasma based on discrete firefly optimization and application to Cancer biomarkers identification, *Front. Genet.*, **10** (2019), 542.
43. K. V. Gunbin, V. V. Suslov, I. I. Turnaev, D. A. Afonnikov, N. A. Kolchanov, Molecular evolution of cyclin proteins in animals and fungi, *BMC Evol. Biol.*, **11** (2011), 1–20.
44. J. Zhang, Y. Zhang, Y. Li, S. Guo, G. Yang, Identification of cancer biomarkers in human body fluids by using enhanced physicochemical-incorporated evolutionary conservation scheme, *Curr. Trends Med. Chem.*, **20** (2020), 1888–1897.
45. X. Zhao, J. Zhang, Q. Ning, P. Sun, Z. Ma, M. Yin, Identification of protein pupylation sites using bi-profile Bayes feature extraction and ensemble learning, *Math. Probl. Eng.*, **2013** (2013).
46. H. Chai, J. Zhang, G. Yang, Z. Ma, An evolution-based DNA-binding residue predictor using a dynamic query-driven learning scheme, *Mol. BioSyst.*, **12** (2016), 3643–3650.
47. M. Hauser, M. Steinegger, J. Söding, MMseqs software suite for fast and deep clustering and searching of large protein sequence sets, *Bioinformatics*, **32** (2016), 1323–1330.
48. L. Khalatbari, M. R. Kangavari, S. Hosseini, H. Yin, N. M. Cheung, MCP: a multi-component learning machine to predict protein secondary structure, *Comput. Biol. Med.*, **110** (2019), 144–155.
49. M. Malumbres, Cyclin-dependent kinases, *Genome Biol.*, **15** (2014), 1–10.
50. C. Ratineau, M. W. Petry, H. Mutoh, A. B. Leiter, Cyclin D1 represses the basic helix-loop-helix transcription factor, BETA2/NeuroD, *J. Biol. Chem.*, **277** (2002), 8847–8853.
51. U. K. Bhawal, F. Sato, Y. Arakawa, K. Fujimoto, T. Kawamoto, K. Tanimoto, et al., Basic helix-loop-helix transcription factor DEC1 negatively regulates cyclin D1, *J. Pathol.*, **224** (2011), 420–429.
52. A. N. Pettitt, A two-sample Anderson-Darling rank statistic, *Biometrika*, **63** (1976), 161–168.
53. S. Siegel, Nonparametric statistics, *Am. Stat.*, **11** (1957), 13–19.
54. J. Zhang, L. Kurgan, SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences, *Bioinformatics*, **35** (2019), i343–i353.
55. J. Zhang, Z. Ma, L. Kurgan, Comprehensive review and empirical analysis of hallmarks of DNA-, RNA-and protein-binding residues in protein chains, *Briefings Bioinf.*, **20** (2019), 1250–1268.
56. G. Beliakov, G. Li, Improving the speed and stability of the k-nearest neighbors method, *Pattern Recognit. Lett.*, **33** (2012), 1296–1301.
57. D. A. Pisner, D. M. Schnyer, Support vector machine, in *Machine Learning*, (2020), 101–121.

58. T. Calders, S. Verwer, Three naive bayes approaches for discrimination-free classification, in *Data Mining and Knowledge Discovery*, **21** (2010), 277–292.
59. A. L. Boulesteix, S. Janitza, J. Kruppa, I. R. König, Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics, in *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **2** (2012), 493–507.
60. Y. Zhang, X. Zhang, A. N. Lane, T. W. M. Fan, J. Liu, Inferring gene regulatory networks of metabolic enzymes using gradient boosted trees, in *IEEE journal of biomedical and health informatics*, **24** (2019), 1528–1536.
61. B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, et al., The reactome pathway knowledgebase, *Nucleic Acids Res.*, **48** (2020), D498–D503.
62. J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A.Salazar, E. L.Sonnhammer, et al., Pfam: The protein families database in 2021, *Nucleic Acids Res.*, **49** (2021), D412–D419.
63. H. Mi, D. Ebert, A. Muruganujan, C. Mills, L. P. Albou, T. Mushayamaha, et al., PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API, *Nucleic Acids Res.*, **49** (2021), D394–D403.

AIMS Press