SEMal: Accurate Protein Malonylation Site Predictor Using Structural and Evolutionary Information

Shubhashis Roy Dipta[1], Ghazaleh Taherzadeh[2], MD. Wakil Ahmad[1], MD. Easin Arafat[3], Swakkhar Shatabda[1, *], Abdollah Dehzangi[4, 5, *]

[1] Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh

[2] Institute for Bioscience and Biotechnology Research, University of Maryland, College Park, MD, 20742, USA

[3] Institute of Information Technology, Jahangirnagar University, Savar, Dhaka, Bangladesh.

[4] Department of Computer Science, Rutgers University, Camden, NJ, 08102, USA

[5] Center for Computational and Integrative Biology, Rutgers University, Camden, NJ, 08102, USA

[*] Corresponding authors: Swakkhar Shatabda (e-mail: swakkhar@cse.uiu.ac.bd) (S.S.) & Abdollah Dehzangi (e-mail: i.dehzangi@rutgers.edu) (A.D.)

Telephone: +1 (856) 225-6699 (A.D.)

Abstract

Post Transactional Modification (PTM) is a vital process which plays an important role in a wide range of biological interactions. One of the most recently identified PTMs is Malonylation. It has been shown that Malonylation has an important impact on different biological pathways including glucose and fatty acid metabolism. Malonylation can be detected experimentally using mass spectrometry. However, this process is both costly and time-consuming which has inspired research to find more efficient and fast computational methods to solve this problem. This paper proposes a novel approach, called SEMal, to identify Malonylation sites in protein sequences. It uses both structural and evolutionary-based features to solve this problem. It also uses Rotation Forest (RoF) as its classification technique to predict Malonylation sites. To the best of our knowledge, our extracted features as well as our employed classifier have never been used for this problem. Compared to the previously proposed methods, SEMal outperforms them in all metrics such as sensitivity (0.94 and 0.89), accuracy (0.94 and 0.91), and Matthews correlation coefficient (0.88 and 0.82), for *Homo Sapiens* and *Mus Musculus* species, respectively. SEMal is publicly available as an online predictor at: http://brl.uiu.ac.bd/SEMal/.

Keywords: Post-translational modifications, Malonylation, Rotation Forest, Evolutionary Features, Structural Features, Predicted Local Structure

1. Introduction

Post Translational Modifications (PTMs) are defined as the enzymatic changes in the protein after its translation process in the ribosome [1-4]. There is a long list of different PTMs that includes

ubiquitination [5], methylation [6], acetylation [7], and many more [8]. These PTMs have active roles in a wide range of cellular activities. Among 20 amino acids that are building blocks of proteins, Lysine (single-letter amino acid code K) is the most susceptible to PTMs [9]. Some of these modifications lead to a wide range of diseases including arthritis, high blood pressure, hypertension, and coronary heart diseases [1-4, 10].

Malonylation is a recently discovered PTM in which a malonyl group attaches to a lysine amino acid residue. It has been found in bacterial and mammalian cells [11]. The detection of Malonylation site is a cumbersome task. The experimental procedures to detect such sites (mainly mass spectrometry) are time-consuming and costly [12]. Such insufficiencies have given birth to a challenging task of proposing fast and accurate computational methods to identify the Malonylation sites. During the past decade, a wide range of machine learning approaches has been proposed to predict different PTM sites [13-18]. Among them, a few studies have specifically proposed to predict lysine Malonylation.

Xu et al. [13] developed the first method to predict Malonylation sites using machine learning. They used data extracted from sequence order information, position-specific amino acid propensity, and physicochemical properties. They used Support Vector Machine (SVM) as the classifier. Later on, Du et al. [14] also used SVM classifier as well as both sequence and protein functional annotation as features to predict Malonylation sites. The promising results reported in [14] proved that integrating different features could potentially enhance Malonylation sites prediction accuracy.

Later on, Wang et al. [15] developed an SVM based classifier called MaloPred to classify Malonylation sites for specific species (i.e., *E. coli*, *M. musculus*, and *H. sapiens*). They also used different types of features integrated into their model. Given the variation observed in their results, they suggested that different species have different biological processes with respect to Malonylation PTM. Thus, different models should be developed for different species.

At the same time, Xiang et al. [16] developed a computation method to predict Malonylation sites using SVM classifier and the pseudo amino acid composition encoding scheme for feature extraction. Their method achieved promising performance on a relatively small benchmark dataset. Later on, Taherzadeh et al. [17] employed another sequence and structure-based SVM classifier, named Sprint-MAL. Their model was only trained using the *Mus Musculus* (mouse) data. However, its performance was consistent when tested on *Homo Sapiens* (human) protein sequences.

Most recently, Zhang et al. [18] developed an ensemble-based model where they have used different classifiers including Random Forest, SVM, Light Gradient Boosting Machine (LightGBM), K-nearest

neighbor (KNN), Linear Regression, and an ensemble of these classifiers. They showed that ensemble classifiers can be used as a powerful predictor to solve this problem.

Chen et al. [19] developed a new deep learning-based classifier where they used a novel encoding method and physicochemical based information to extract features. To build their classifier, they used a combination of Long Short-Term Memory (LSTM) and RF classifier to predict Malonylation sites. They also used the word2vec concept of natural language processing to get an embedding from the amino acid sequence. To the best of our knowledge, their model obtained the best result for Malonylation site prediction so far. Besides those for Malonylation, there was a large number of studies tackling other PTM sites prediction tasks, which has given us insights about the employed classifiers, feature extraction methods, and imbalanced data handling schemes [20-30]. There is also a wide range of PTM site predictors using machine learning models available at ExPASy as well [31-35].

Despite all the efforts that have been made so far, the prediction of lysine Malonylation sites remain limited. As a result, there is a critical demand to develop more powerful predictors that are also more efficient and accurate. In this paper, we propose a new predictor called SEMal, which incorporates both structural and evolutionary information to extract features. Here we focus on the Malonylation site prediction for *Homo Sapien*s and *Mus Musculus* species [36]. To build SEMal, the position-specific scoring matrix (PSSM) is computed and then incorporated into the profile bigram. After that, structural features are extracted from the predicted local structure for each protein. Finally, with profile bigram and structural information, Rotation Forest classifier was employed for the classification task.

To the best of our knowledge, structural and evolutionary information as well as Rotation Forest classifier have never been used for predicting Malonylation sites. In other words, the main contribution of this study is to employ features and classifiers that have been shown effective for similar problems to tackle Malonylation site prediction. When compared to other predictors [15, 18, 19], SEMal shows more exceptional results than any other predictors. It demonstrates the high accuracy of 94.0% with a sensitivity of 0.94, a specificity of 0.93, and Matthews correlation coefficient (MCC) of 0.88 on the human samples. It also predicts the Malonylation sites for mouse with 91.0% accuracy, 0.89 sensitivity, 0.93 specificity, and 0.82 MCC. SEMal is implemented as an online predictor and is publicly available at: http://brl.uiu.ac.bd/SEMal/.

2. Materials & Methods

SEMal uses evolutionary information extracted from the PSSM as well as structural information extracted from SPIDER 2.0 [37-38] as input features. The PSSM was translated to profile bigram and later concatenated with the structural features [28-29]. The resulting matrix was then used to classify

Malonylation sites from Non-Malonylation sites using Rotation Forest algorithm, which is an ensemble-based machine learning classifier [39]. This section describes the dataset, the method used for balancing dataset, extracted features, and employed classifier.

2.1 Benchmark Dataset

The dataset used in this study was extracted from the Protein Lysine Modification Database (PLMD) which contains various PTMs that are accurately determined using experimental techniques. This Malonylation dataset consists of 1,841 *Homo Sapiens* (human) and 1,466 *Mus Musculus* (mouse) proteins. To reduce the bias, first we removed proteins with over 40% sequence similarity using CD-HIT [40-42]. The remaining dataset included 1,621 human proteins with 4,181 Malonylation sites and 72,145 Non-Malonylation sites, and 1,237 mouse proteins with 3,454 and 43,943 Malonylation and Non-Malonylation sites, respectively.

2.2 Balancing Dataset

Given the number of Malonylation and Non-Malonylation sites mentioned above, the approximate ratio of the positive sites to negative sites is 1:18 for human and 1:12 for mouse. This imbalance is justified from a biological point of view. However, it can be a source of bias in the model which affects the prediction. Hence, it is important to balance the data to avoid bias towards negative samples.

There is a wide range of methods proposed to balance the data in the literature. For instance, SMOTe [43] has been used for oversampling the minor class. It uses the Euclidean distance to project more points for the minor class. ADASYN method is also used to up-sample the data [44]. Although the up-sampling of data can make more data available to us, it also makes some artificial instances of data. In this study, we investigated a wide range of balancing methods, among which, down-sampling using K-Nearest Neighbor (KNN) demonstrated the best results [20-22].

To do this, we initially calculated the Euclidean distance for each and every pair of lysine residues and those negative lysine residues were removed from the dataset which has at least one positive point in their k-nearest neighbors. Subsequently, after some trial and error to balance the data to a 1:1 ratio, the optimal k value was 100 for human, and 60 for mouse. This resulted in 4,181 positive sites and 4,220 negative sites for the human and 3,454 positive and 3,481 negative sites for the mouse datasets. These data were then used to train two different models using Rotation Forest algorithm. Both of the trained models were compared against previous state-of-the-art methods [15, 18, 19].

2.3 Independent Train and Test Sets

To assess the generality of our model and avoid overfitting, we separated 10% of our data and used as the independent test dataset while the remaining 90% was used as the training set. The independent test set was not used for parameter tuning and remained untouched and unbalanced.

2.4 Structural and Evolutionary Features

The following sections describe the structural and evolutionary information we extracted and used with the classifier as a feature vector.

2.4.1 Structural Features

The structural features used in this paper are the predicted secondary structure (three states namely, coil, strand, and helix), the Accessible Surface Area (ASA), and the backbone torsion angles (four angles namely, θ, τ, ψ, and φ). It was shown in previous studies that these structural features can provide important discriminatory information for the classification purpose [20-23, 25, 27]. In this study, we used predicted values for these parameters using SPIDER2 [37-38]. SPIDER2 is a machine learning toolbox to predict protein local structure using deep learning architecture. It produces a matrix consisting of predicted values for all the amino acids along the protein sequence. For simplicity, we will call the matrix SSpre. SPIDER 2.0 achieves one of the best results in predicting the secondary structure [45, 46], the ASA [47], and the backbone torsion angles [47] from protein sequences. The details of the structural features used in this study are explained below.

Accessible surface area **(ASA)**: ASA is a measure of an amino acid's accessible area to a solvent. It exposes essential information about the protein structure and how it can interact with other macromolecules. It also specifies which amino acids are in the surface area and hence has more possibility to undergo PTMs. The resulting ASA value is obtained by running SPIDER2 on every protein sequence.

Secondary structure*:* The secondary structure defined the local 3D structure of the proteins. It consists of three local components namely, coil, strand, and helix. Predicted secondary structure gives the possibility for each amino acid to build one of these three local structures. Such information can determine which amino acids are more structured and which are more susceptible to interact with other macromolecules. SPIDER2 produces a L x 3 matrix as the predicted secondary structure, where L denotes the protein length and the columns denotes coil, strand, and helix (pc, pe, and ph, respectively).

Local backbone angles: Local backbone angels are also representatives of the local structure of proteins. Unlike secondary structure which gives us the idea about a local configuration with respect to building coil, strand, or helix shapes, local backbone angles give us continuous information about the local structure of proteins. In other words, the backbone torsion angle provides continuous information about the interaction of local amino acids. SPIDER2 produces probability values for four local backbone angles namely, θ, τ, ψ, and φ which are explained in detail in their original studies [37-38].

2.4.2 Evolutionary Features

Evolutionary features provide information about the substitution probability of the amino acids along the protein sequence through evolution. Such features can be extracted using Position-Specific Scoring Matrix (PSSM) which is produced as the output of PSI-BLAST toolbox [48]. For each protein, PSSM provides the probability of substitution of its amino acids with all the 20 amino acids depending on their position. PSSM is a L × 20 matrix where L is the length of the protein sequence, and columns represent the 20 amino acids. Here we produced PSSM by running PSI-BLAST on the non-redundant protein data bank for three iterations using a cut-off value of 0.001. It has been shown that using 0.001 as the cut off value for PSI-BLAST is sufficient to identify homologous sequences to produce the PSSM matrix [27, 29, 37].

2.5 Formulation of the Features

In this section, we will look into the methods used for the formulation of the features for each lysine residue. We used 15 upstream and 15 downstream amino acids for both human and mouse to extract evolutionary information (PSSM). On the other hand, we used 3 upstream and 3 downstream for the structural information. Such window sizes are selected as they obtained the best results compared to other window sizes. Here, for both sides of the protein tail (c-terminus and n-terminus), where there are less than 15 (evolutionary) or 3 (structural) neighboring amino acids on one side, we use the mirror effect to fill out the missing amino acids. This process is also shown in Figure 1 where lysine is presented as *K*.

For evolutionary information, a segment *P* comprises 15 upstream amino acids, 15 downstream amino acids, and the lysine residue in the middle. This can be formulated as:

$$P = \{A_{-15}, A_{-14}, \ldots., \mathbf{K}, \ldots., A_{14}, A_{15}\} \quad (1)$$

where A_n is the upstream and downstream amino acids, n is positive for downstream amino acids, n is negative for upstream amino acids, and K represents the lysine residue. Therefore, each segment is made of 31 amino acids (15 upstream, 15 downstream, and one lysine residue).

For structural information, a segment Q comprises 3 upstream amino acids, 3 downstream amino acids, and the lysine residue in the middle. Therefore, the segment (Q) is made of 7 amino acids (3 upstream, 3 downstream and the lysine residue in the middle) and can be formulated as:

$$Q = \{A_{-3}, A_{-2}, A_{-1}, \mathbf{K}, A_1, A_2, A_3\} \quad (2)$$

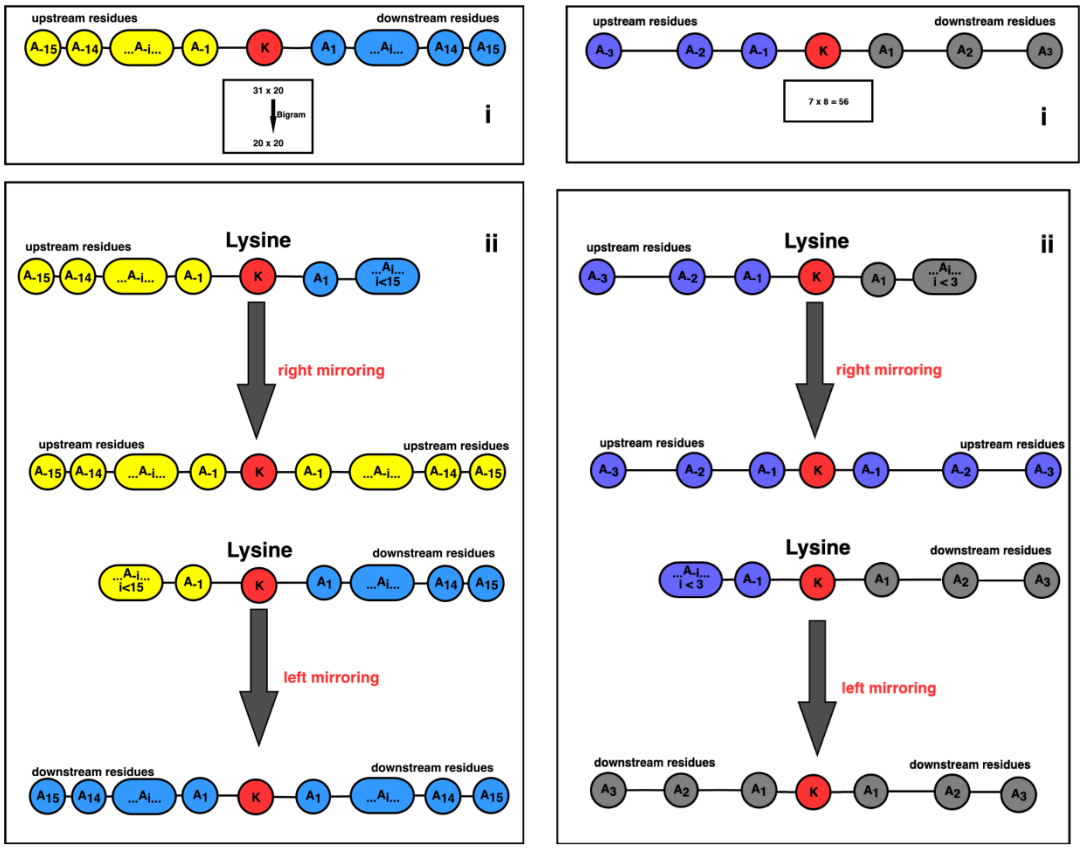

Figure 1: Formulation of the Structural and Evolutionary features

After the acquisition of PSSM and SSpre matrices, we converted PSSM to a frequency vector of bigrams (PSSM + bigram) and SSpre as it is. The dimension of the resulting matrices were 7 x 8 from SSpre (56 features in total) and 20 x 20 from PSSM + bigram (400 features in total). The concept of Bigram used to extract features from PSSM was first introduced in [28-29] and shown to achieve remarkable success. This technique is explained in more detail in the following. For each lysine residue, we then converted both matrices into a 456-dimensional vector (56 + 400). This 456-dimensional vector captures the structural and evolutionary information of the segment P representing each lysine residue.

2.5.1 Profile Bigrams

The technique of profile bigrams has been shown to be an effective method to extract important discriminatory information from PSSM [28-29]. Let's assume matrix *M* represents PSSM for each segment *P* and matrix *E* represents SSpre for each segment *Q*. Every element in matrix *M* is represented by m_{ij}. The dimension of matrix *M* are 31 x 20 and matrix *E* are 7 x 8.

Profile bigram of matrix *M* is formulated as:

$$B_{p,\,q} \sum_{k=1}^{30} m_{k,p} m_{k+1,q} \quad (3)$$

Matrix *B* comprises elements $B_{p,\,q}$ where $p = 1 \ldots 20$ and $q = 1 \ldots 20$. It represents PSSM + Bigram with 20 x 20 dimension. Matrix *B* is then transformed into a 400-dimensional feature vector. This can be formulated as:

$$F = [B_{1,1}, B_{1,2}, \ldots, B_{1,20}, B_{2,1}, \ldots, B_{20,1}, B_{20,2}, \ldots, B_{20,20}] \quad (4)$$

The SSpre matrix, E (7 x 8) is then flattened to a 56-dimensional feature vector. This can be formulated as:

$$G = [E_{1,1}, E_{1,2}, \ldots, E_{1,8}, E_{2,1}, \ldots, E_{7,1}, E_{7,2}, \ldots, E_{7,8}] \quad (5)$$

Then, both of these vectors (*F*, *G*) are merged together to get the 456-dimensional feature vector.

2.6 Rotation Forest

Rotation Forest (RoF) is a tree-based ensemble method that randomly splits the data into *k* subsets, and then applies *Principal Component Analysis (PCA)* to each segment for feature transformation [39]. In the end, the results of classification using decision trees on those transformed features are aggregated using an ensemble to produce the final result. RoF is a supervised learning model in the area of machine learning. The idea behind RoF is to introduce a mechanism to encourage the individual accuracy and diversity of each base classifier at the same time [39]. The decision tree is mainly used as the base classifier for RoF. One of the reasons for choosing the decision tree is that it is sensitive to the rotation of the feature axes and can produce better prediction performance [39]. It has demonstrated promising results for similar studies found in the literature [49-54].

The main parameter for RoF is the number of estimators or the number of decision trees to be used in RoF. Here we investigated the performance of RoF using different values for the number of estimators. Among them, using 100 for this parameter demonstrated the best results. For the max_features parameter, we used the default value 'auto' as it was empirically studied in [55]. Another parameter was n_jobs which sets the number of jobs to run in parallel. Here we used '-1' for n_jobs value which means using all the available processors. For this work, we used the rotation_forest package in python3.

2.7 Evaluation Metrics

Getting the right metrics for performance assessment is one of the essential and critical components of an experiment. We used six metrics to evaluate the performance of our model. The metrics are sensitivity, specificity, precision, accuracy, F1-score, and Matthews correlation coefficient (MCC).

Sensitivity (Eq. 1) measures the ability of the predictor to classify Malonylation sites, correctly. The range of the value is from 0 to 100 percent, where 100% indicates an excellent estimator, and 0% shows an incompetent predictor.

Specificity (Eq. 2) measures the ability of the predictor to classify the non-Malonylation sites. The range of the value is also 0 to 100 percent, where 100% indicates that the predictor can successfully predict the Non-Malonylation sites, and 0% means it fails to predict any Non-Malonylation site.

Precision (Eq. 3) indicates the number of correctly classified Malonylation and Non-Malonylation sites among all sites. This metric gives us the measurement of the predictor's ability to label a site as Malonylation if the site is Non-Malonylation. This metric is retrieved by dividing the true positives by the total number of data.

Accuracy (Eq. 4) shows us the predictor's ability to measure the correctly classified lysine residues. It ranges from 0 to 100 percent, where 0% means the least accurate predictor, and 100% means the most accurate predictor.

The F1-score (Eq. 5) gives a balanced score between sensitivity and specificity. It also ranges from 0 (worst balance) to 1 (perfect balance between sensitivity and specificity).

MCC (Eq. 6) gives us the classification quality of a predictor. It ranges from -1 to +1, where -1 indicates an entirely negative correlation, and +1 means a highly positive relationship.

The six metrics can be calculated as:

$$SN = \frac{TP}{TP+FN} * 100 \quad (6)$$

$$SP = \frac{TN}{TN+FP} * 100 \quad (7)$$

$$Precision = \frac{TP}{TN+FP} \quad (8)$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} * 100 \quad (9)$$

$$F1 = \frac{2*SN*PR}{PR+SN} \quad (10)$$

$$MCC = \frac{(TP)(TN)-(FP)(FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (11)$$

where TP (True Positive) presents the number of correctly classified Malonylation sites, TN (True Negative) presents the number of correctly classified Non-Malonylation sites, FP (False Positive) presents the number of misclassified Non-Malonylation sites as Malonylation sites, and FN (False Negative) presents the number of misclassified Malonylation sites as Non-Malonylation sites.

Any method that performs the highest in all of these metrics would be the ideal predictor. However, a good one should at least show a higher sensitivity when compared with other approaches because a predictor with lower sensitivity cannot effectively detect Malonylation sites.

2.8 Cross-validation Strategy

To accurately assess our model, we utilized cross-validation strategy on our training. The cross-validation is done in the following manner:

1. Randomly partition the data into n roughly equal parts.
2. Among them, n – 1 sets are used for the training data and the remaining one as the validation data.
3. Repeated step 2 for n times for all n folds.

In this model, we used n = 10 for n-fold cross-validation data whose results are presented in the next sections. The general architecture of our proposed method is shown in Figure 2.

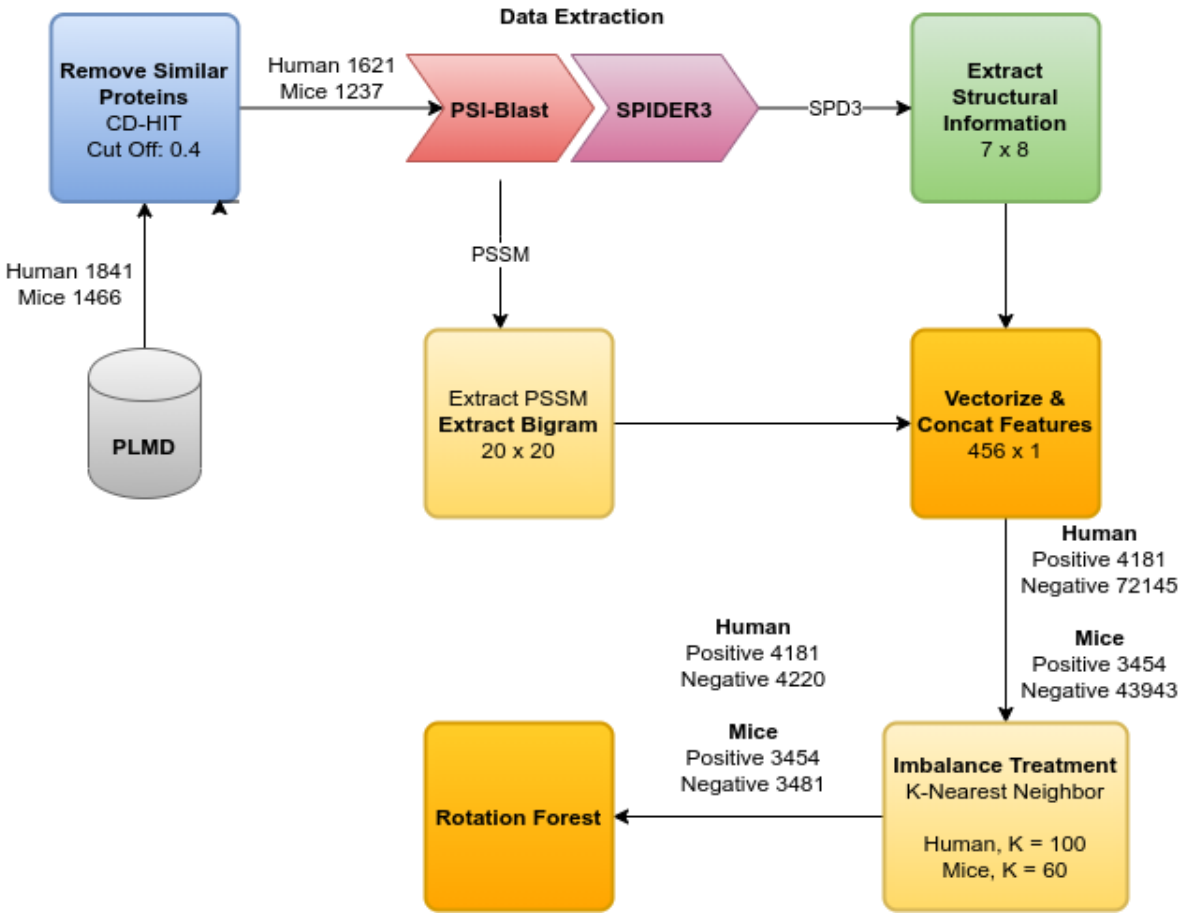

Figure 2: The general architecture of our proposed model (SEMal)

3. Results and Discussion

In this section, we present our results, compare them with previous studies found in the literature, and then discuss them.

3.1 Comparison of SEMal with Current State-of-the-Art Predictors

Here we compare our model to three previously proposed models namely, kmal-sp [18], MaloPred [15] and LEMP [19], All of which have produced promising results for the Malonylation site prediction problem. We first compare SEMal with the models experimented and presented by Zhang et al., [18].

The comparison is given in Table 1 for Human and Table 2 for Mouse using 10-fold cross-validation. In this table, we present different models that have been tried by Zhang et al., and the result from our model is presented at the bottom of the table (bolded). The results shown in Tables 1 and 2 are collected from [18], directly. We can see that our model outperforms all the previous models and shows good performance over the others.

Method	**PRE**	**SN**	**SP**	**F1**	**ACC**	**MCC**
1. RF	0.83	84.3%	83.4%	0.84	83.8%	0.68
2. SVM	0.83	83.9%	83.7%	0.84	83.8%	0.68
3. LightGBM	0.85	86.3%	85.4%	0.86	85.8%	0.72
4. KNN	0.83	74.9%	85.0%	0.79	80.0%	0.60
5. LR	0.84	82.3%	84.4%	0.83	83.3%	0.67
{1, 2}	0.85	84.6%	85.7%	0.85	85.2%	0.70
{1, 2, 3}	0.85	84.6%	85.7%	0.85	85.2%	0.70
{1, 2, 3, 4}	0.86	85.6%	86.7%	0.86	85.7%	0.71
{1, 2, 3, 4, 5}	0.86	84.6%	86.7%	0.86	85.7%	0.71
{3, 4, 5}	0.86	84.9%	87.0%	0.86	86.0%	0.72
SEMal	**0.95**	**94.3%**	**94.7%**	**0.94**	**94.5%**	**0.89**

Table 1: Performance of the **SEMal** compared to Zhang et al., individual and ensemble predictors for human.

Here in Tables 1 and 2, the experimented models used by Zhang et al. [18] are RF {1}, SVM {2}, Light Gradient Boosting Machine (LightGBM) {3}, KNN {4}, Linear Regression (LR) {5}, and an ensemble of some of these algorithms. Also, {1, 2, 3, 4, 5} means ensemble of all these five classifiers, together. In addition, PRE stands for Precision, SN for Sensitivity, SP for Specificity, ACC for Accuracy, and MCC for Matthews correlation coefficient. As shown in Tables 1 and 2, SEMal shows a significant improvement over the experimented models presented in Zhang et al., [18]. These results clearly indicate considerable improvement in predicting Malonylation sites.

Method	**PRE**	**SN**	**SP**	**F1**	**ACC**	**MCC**
1. RF	0.81	84.3%	80.4%	0.83	82.3%	0.65
2. SVM	0.82	82.9%	81.7%	0.82	82.3%	0.65
3. LightGBM	0.81	82.6%	80.7%	0.82	81.7%	0.63
4. KNN	0.81	72.9%	83.1%	0.77	78.0%	0.56
5. LR	0.81	82.9%	80.4%	0.82	81.7%	0.63
{1, 2}	0.82	82.6%	82.1%	0.82	82.3%	0.65
{1, 2, 3}	0.81	82.3%	80.4%	0.82	81.3%	0.63
{1, 2, 3, 4}	0.83	83.9%	82.4%	0.83	83.2%	0.66
{1, 2, 3, 4, 5}	0.83	83.6%	82.7%	0.83	83.2%	0.66
{1, 2, 4}	0.84	82.9%	83.7%	0.83	83.3%	0.67
SEMal	**0.91**	**88.3%**	**90.8%**	**0.89**	**89.6%**	**0.79**

Table 2: Performance of the **SEMal** compared to Zhang et al.'s individual and ensemble predictors for mouse.

We then compared our predictor to MaloPred [15] and kmal-sp [18] for both Mouse and Human species. These studies reported promising results for the Malonylation site prediction problem with respect to these two species. These comparisons for 10-fold cross-validation are presented in Table 3 for both human and mouse. The results shown in Table 3 for Kmal-sp and MaloPred are collected from their papers [15, 18], directly. As shown in Table 3, for human, we achieved around 9%, 11%, 9%, 11%, 10%, and 23% improvement for PRE, SN, SP, F1-Score, ACC, and MCC, respectively. Similarly, for mouse, we achieved over 8%, 7%, 8%, 8%, 8%, and 20% improvement for PRE, SN, SP, F1-Score, ACC, and MCC, respectively.

Species	Method	PRE	SN	SP	F1	ACC	MCC
Human	MaloPred	0.82	82.9%	82.4%	0.82	82.7%	0.65
	kmal-sp	0.87	84.9%	87.0%	0.85	86.0%	0.72
	SEMal	**0.95**	**94.3%**	**94.7%**	**0.94**	**94.5%**	**0.89**
Mouse	MaloPred	0.80	80.6%	79.7%	0.80	80.2%	0.60
	kmal-sp	0.84	82.9%	87.0%	0.83	83.3%	0.66
	SEMal	**0.91**	**88.3%**	**90.8%**	**0.89**	**89.6%**	**0.79**

Table 3: Performance of the **SEMal**, MaloPred [15] and kmal-sp [18]

We also evaluated our model for the independent test set. For this comparison, we have used the state-of-the-art model, named LEMP [19]. LEMP is the most recent proposed tool for the Malonylation site prediction problem. Although LEMP is a peptides-based predictor, its webserver (which is available at: http://www.bioinfogo.org/lemp/index.php) accepts protein sequences as input. Hence, we passed our independent test set to their server and recorded the results. The comparisons for both human and mouse are given in Table 4. As shown in this table, the results for the independent test set are consistent with those reported using 10-fold cross-validation. Such consistency demonstrates the generality and robustness of our proposed model. Moreover, as shown in Table 4, SEMal outperforms LEMP in all the evaluation metrics. As shown in this table, SEMal achieves 20% and 17% better accuracy for human and mouse, respectively. Such significant enhancement is consistent with all the evaluation metrics.

Species	Method	PRE	SN	SP	F1	ACC	MCC
Human	LEMP	0.15	82.4%	73.0%	0.25	73.5%	0.28
	SEMal	**0.93**	**94.3%**	**93.4%**	**0.94**	**93.8%**	**0.88**
Mouse	LEMP	0.18	77.5%	73.1%	0.29	73.4%	0.28
	SEMal	**0.93**	**89.0%**	**92.8%**	**0.91**	**90.9%**	**0.82**

Table 4: Performance of the SEMal compared to LEMP [19] on the independent test set for Human and Mouse samples.

Our results in general illustrate the effectiveness of using SEMal in discriminating Malonylation sites from Non-Malonylation sites. These results are mainly achieved by incorporating both structural and evolutionary information together as well as using RoF. As mentioned in the Introduction Section, Kmal-sp is built using evolutionary information [18], MaloPred is built using sequence, physicochemical, and evolutionary base information [15], and LEMP is built using sequence and physicochemical base information to extract features [19]. However, to the best of our knowledge, structural information has never been used for the Malonylation site classification. As it was shown in previous studies, the combination of structural and evolutionary information potentially provides significant discriminatory information for other PTM sites prediction problems [20-23, 25, 27].

The main reason behind this is that structural properties provide significant information on the local structure of the proteins and how the amino acids interact, locally. Moreover, to the best of our knowledge, RoF has never been used for the Malonylation site prediction problem. The main advantage of RoF over RF or other similar classifiers is that it can give similar or better performance with a smaller number of trees [39, 49, 50]. Hence, using a smaller number of samples, we can get better results. Though LEMP [19] uses a more sophisticated machine learning classifier, the main advantage of RoF is that it can do implicit feature selection [39], that does not require scaling, missing value treatment, or outlier removal. It also performs better than deep learning architecture that is used in [19] when the number of samples is limited as it is the case in this study.

Figures 3 and 4 illustrate the Receiver Operating Characteristic (ROC) curves of our results for human and mouse with our model (Left) and with the LEMP model (Right), respectively. In the ROC curve the x-axis is the TP Rate (TPR) and the y-axis is the FP Rate (FPR). TPR is the same as sensitivity and FPR means 1 – specificity. The optimal point is where TPR is maximum and FPR is minimum.

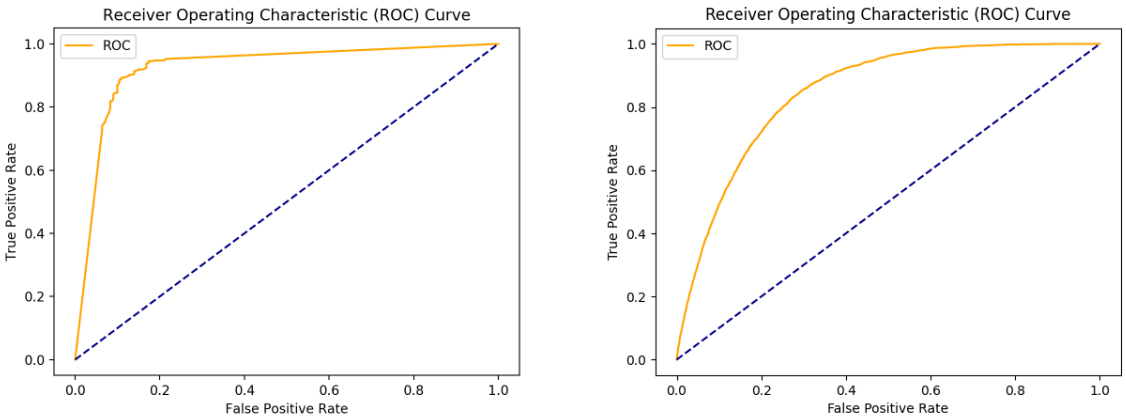

Figure 3: ROC Curve of Human of SEMal (Left) and LEMP (Right)

As shown in Figure 3 (human), the Area Under the Curve (AUC) score for SEMal is 0.92, whereas the AUC score for LEMP is 0.85. Similarly, according to Figure 4 (mouse), the AUC score for SEMal is 0.90 while the AUC score for LEMP is 0.82. These results highlight the better performance of SEMal over LEMP with respect to AUC values as well.

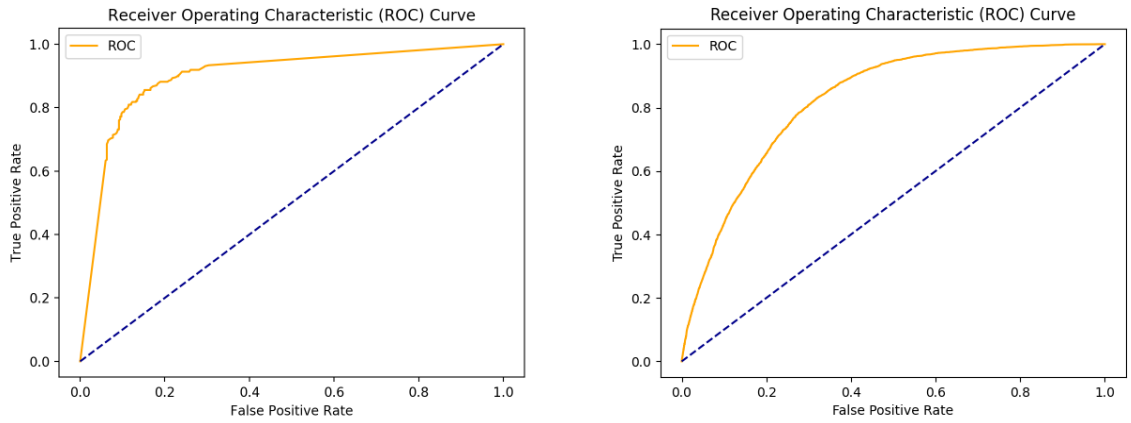

Figure 4: ROC Curve of Mice of SEMal (Left) and LEMP (Right)

3.2 Investigating the Impact of our Extracted Features on the Achieved Results

To investigate the impact of our extracted features on the achieved results, we utilize four widely used machine learning algorithms namely, AdaBoost, RF, RoF (used to build SEMal), and SVM on different combinations of our extracted feature groups. Here, we used three sets of features from PSSM, SPD3, and a combination of PSSM and SPD3. Results are presented in Table 5.

As shown in Table 5, results achieved using features extracted from the PSSM are better than those extracted from SPD3. It shows the importance of evolutionary-based features in enhancing Malonylation site prediction accuracy. However, the best results are achieved when the combination of features extracted from PSSM and SPD3 are used simultaneously. This indicates that the achieved results are dependent on both the evolutionary and structural features proposed in this study.

Also, as shown in Table 5, we achieve the best results for both human and mouse datasets using RoF on the combination of all features together. In other words, by using RoF on the combination of both structural and evolutionary-based features we achieve constantly better results than other classifiers on both human and mouse datasets. Such results demonstrate the importance of both structural and evolutionary features extracted in this study as well as the use of RoF as our classifier on the prediction enhancement achieved in this study.

Species		Method	PRE	SN	SP	F1	ACC	MCC
Human	PSSM	AdaBoost	0.92	92.6%	90.8%	0.92	91.7%	0.83
		Random Forest	**0.94**	**94.3%**	**94.3%**	0.94	**94.3%**	**0.89**
		Rotation Forest	0.89	89.5%	89.3%	0.89	89.4%	0.79
		SVM	0.90	92.8%	90.3%	0.92	91.6%	0.83
	SPD3	AdaBoost	0.81	82.3%	80.6%	0.82	81.5%	0.63
		Random Forest	0.82	90.5%	80.6%	0.86	85.4%	0.71
		Rotation Forest	0.76	76.1%	76.3%	0.76	76.2%	0.52
		SVM	0.81	87.1%	80.1%	0.84	83.6%	0.67
	PSSM + SPD3	AdaBoost	0.90	91.5%	90.0%	0.91	90.7%	0.81
		Random Forest	**0.94**	**94.3%**	**94.3%**	**0.94**	**94.3%**	**0.89**
		Rotation Forest	**0.93**	**94.3%**	93.4%	**0.94**	93.80%	**0.88**
		SVM	**0.94**	93.6%	94.1%	**0.94**	93.8%	**0.88**
Mouse	PSSM	AdaBoost	0.87	84.7%	87.4%	0.86	86.0%	0.72
		Random Forest	0.91	84.9%	91.4%	0.88	88.2%	0.77
		Rotation Forest	0.88	82.1%	88.8%	0.85	85.5%	0.71
		SVM	0.88	81.2%	89.1%	0.85	85.2%	0.71
	SPD3	AdaBoost	0.74	69.4%	75.9%	0.72	72.7%	0.45
		Random Forest	0.74	77.7%	72.8%	0.76	75.3%	0.51
		Rotation Forest	0.66	64.5%	66.5%	0.65	65.5%	0.31
		SVM	0.72	78.6%	70.2%	0.75	74.4%	0.49
	PSSM + SPD3	AdaBoost	0.87	85.3%	87.1%	0.86	86.2%	0.72
		Random Forest	0.90	84.9%	91.1%	0.88	88.1%	0.76
		Rotation Forest	**0.93**	**89.0%**	**92.8%**	**0.91**	**90.9%**	**0.82**
		SVM	0.88	83.5%	88.3%	0.86	85.9%	0.72

Table 5: Investigating the impact of our extracted feature groups using four different classifiers namely, AdaBoost, Random Forest, Rotation Forest, and SVM.

3.3 Web Server Implementation

SEMal has been made available as an online webserver at: http://brl.uiu.ac.bd/SEMal/. The use of the web server is very intuitive. The user has to upload the PSSM and SPD3 file, which can be generated using PSI-BLAST [48] and SPIDER2 [38] respectively. The trained model is also available on the server. All our codes from feature extraction to model prediction are available at: https://github.com/dipta007/SEMal to facilitate the use of our paper. The limitation of our model is that it requires PSSM and SPD3 as the input. It is mainly done to make sure that we can regularly update and maintain our webserver and to avoid version control with SPIDER 2.0, PSI-BLAST, and protein databank. The other limitation of our model is that it only works for human and mouse species. For our future direction, we aim at employing SEMal to predict Malonylation sites for other species.

To implement our model, we used Python 3.6, Scikit Learn 0.23.2, and TensorFlow 2.1.0. To implement our frontend and server, we used Flask 1.1.0, React 16.12.0, and Bootstrap 4.4.1. We have also used PSI-BLAST 2.10.1 and Spider 2.0 to produce PSSM and SPD3. We commit to regularly update our web server and adjust it based on the new versions of the tools that we used to implement SEMal. We will also make sure that our online tool would be compatible with the future version of the PSI-BLAST and SPIDER 2.0.

4. Conclusion

In this paper, we have presented a novel Malonylation sites predictor called SEMal, which effectively uses a combination of PSSM + bigram and SSpre features. The structural and evolutionary features provide important discriminatory information to enhance the Malonylation site prediction task. Moreover, we used RoF to build our model, which to the best of our knowledge has never been used for this task. Our achieved results demonstrate that SEMal is able to outperform previous studies found in the literature both for *Homo Sapiens* (Human) and *Mus Musculus* (Mouse). Such results demonstrate the effectiveness of SEMal for predicting Malonylation as one of the most recently identified PTMs. For our future direction, we aim at investigating the application of RoF in conjunction with other classifiers to build an effective ensemble of different classifiers to enhance Malonylation site prediction accuracy even further. We also aim at providing more data interpretation ability to our webserver using graphical support.

Data Availability

All the data used in this article are available at: http://plmd.biocuckoo.org/
Our proposed method is available as an online predictor at: http://brl.uiu.ac.bd/SEMal/
Our codes from feature extraction to model prediction at: https://github.com/dipta007/SEMal

References:

1. Gallego M, Virshup DM. Post-translational modifications regulate the ticking of the circadian clock. Nat Rev Mol Cell Biol 2007;8:139–48.
2. Westermann S, Weber K. Post-translational modifications regulate microtubule function. Nat Rev Mol Cell Biol 2003;4:938–47.
3. Harmel R, Fiedler D. Features and regulation of nonenzymatic post-translational modifications. Nat Chem Biol 2018;14:244–52.
4. Johnson LN. The regulation of protein phosphorylation. Biochem Soc Trans 2009;37:627–41.

5. Qiu, Wang-Ren, et al. "iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model." Journal of Biomolecular Structure and Dynamics 33.8 (2015): 1731-1742.
6. Qiu, Wang-Ren, et al. "iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach." BioMed research international 2014 (2014).
7. Hou, Ting, et al. "LAceP: lysine acetylation site prediction using logistic regression classifiers." PloS one 9.2 (2014): e89575.
8. Consortium, U. Uniprot: a worldwide hub of protein knowledge. Nucleic acids research 47, D506–D515 (2018).
9. Z. Xie, J. Dai, L. Dai, M. Tan, Z. Cheng, Y. Wu, J. D. Boeke, and Y. Zhao, "Lysine succinylation and lysine malonylation in histones," Molecular & Cellular Proteomics, vol. 11, no. 5, pp. 100–107, 2012.
10. Harmel R, Fiedler D. Features and regulation of non-enzymatic post-translational modifications. Nat Chem Biol 2018;14:244–52.
11. Oughtred, R.et al.Biogrid: a resource for studying biological interactions in yeast.Cold Spring Harb. Protoc.2016,pdb–top080754 (2016).
12. Xu, Yan, et al. "Prediction of posttranslational modification sites from amino acid sequences with kernel methods." Journal of theoretical biology 344 (2014): 78-87.
13. Xu Y, Ding Y, Ding J, et al. Mal-Lys: prediction of lysine malonylation sites in proteins integrated sequence-based features with mRMR feature selection. Nat Publ Gr 2016;1–7.
14. Du Y, Zhai Z, Li Y, et al. Prediction of protein lysine acylation by integrating primary sequence information with multiple functional features. J Proteome Res 2016;15:4234–44
15. Wang LN, Shi SP, Xu HD, et al. Computational prediction of species-specific malonylation sites via enhanced characteristic strategy. Bioinformatics 2017;33:1457–63
16. Xiang Q, Feng K, Liao B, et al. Prediction of lysine malonylation sites based on pseudo amino acid compositions. Comb Chem. High Throughput Screen 2017;20:1.
17. Taherzadeh G, Yang Y, Xu H, et al. Predicting lysine malonylation sites of proteins using sequence and predicted structural features. J Comput Chem 2018, doi:10. 1002/jcc.25353.
18. Zhang, Yanju, et al. "Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework." Briefings in Bioinformatics 5 (2018).
19. Chen, Zhen, et al. "Integration of a deep learning classifier with a random forest approach for predicting malonylation sites." Genomics, proteomics & bioinformatics 16.6 (2018): 451-459.
20. Dehzangi, A., Lopez, Y., Lal, S. P., Taherzadeh, G., Sattar, A., Tsunoda, T., & Sharma, A. (2018). Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams. PloS one, 13(2), e0191900.

21. Islam, M. M., Saha, S., Rahman, M. M., Shatabda, S., Farid, D. M., & Dehzangi, A. (2018). iProtGly-SS: Identifying protein glycation sites using sequence and structure based features. Proteins: Structure, Function, and Bioinformatics, 86(7), 777-789.
22. Reddy, H. M., Sharma, A., Dehzangi, A., Shigemizu, D., Chandra, A. A., & Tsunoda, T. (2019). GlyStruct: glycation prediction using structural properties of amino acid residues. BMC bioinformatics, 19(13), 547.
23. Dehzangi, A., Paliwal, K., Lyons, J., Sharma, A., & Sattar, A. (2013, June). Enhancing protein fold prediction accuracy using evolutionary and structural features. In IAPR International Conference on Pattern Recognition in Bioinformatics (pp. 196-207). Springer, Berlin, Heidelberg.
24. Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., & Sattar, A. (2015). Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou′ s general PseAAC. Journal of theoretical biology, 364, 284-294.
25. Chowdhury, S. Y., Shatabda, S., & Dehzangi, A. (2017). iDNAprot-es: Identification of DNA-binding proteins using evolutionary and structural features. Scientific reports, 7(1), 14938.
26. Dehzangi, A., Paliwal, K., Sharma, A., Lyons, J., & Sattar, A. (2013, December). Protein fold recognition using an overlapping segmentation approach and a mixture of feature extraction models. In Australasian Joint Conference on Artificial Intelligence (pp. 32-43). Springer, Cham.
27. Shatabda, S., Saha, S., Sharma, A., & Dehzangi, A. (2017). iPHLoc-ES: identification of bacteriophage protein locations using evolutionary and structural features. Journal of theoretical biology, 435, 229-237.
28. Dehzangi, A., López, Y., Lal, S. P., Taherzadeh, G., Michaelson, J., Sattar, A., ... & Sharma, A. (2017). PSSM-Suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction. Journal of theoretical biology, 425, 97-102.
29. Sharma, A., Lyons, J., Dehzangi, A., & Paliwal, K. K. (2013). A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. Journal of theoretical biology, 320, 41-46.
30. M. W. Ahmad, M. E. Arafat, G. Taherzadeh, A. Sharma, S. R. Dipta, A. Dehzangi, & S. Shatabda, "Mal-Light: Enhancing Lysine Malonylation Sites Prediction Problem Using Evolutionary-based Features," IEEE Access. DOI: 10.1109/ACCESS.2020.2989713
31. Wang, Chenwei, et al. "Gps 5.0: An update on the prediction of kinase-specific phosphorylation sites in proteins." Genomics, Proteomics & Bioinformatics (2020).
32. Steentoft, Catharina, et al. "Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology." The EMBO journal 32.10 (2013): 1478-1488.
33. Ren, Jian, et al. "CSS-Palm 2.0: an updated software for palmitoylation sites prediction." Protein Engineering, Design & Selection 21.11 (2008): 639-644.
34. Julenius, Karin. "NetCGlyc 1.0: prediction of mammalian C-mannosylation sites." Glycobiology 17.8 (2007): 868-876.

35. Juncker, Agnieszka S., et al. "Prediction of lipoprotein signal peptides in Gram-negative bacteria." Protein Science 12.8 (2003): 1652-1662.
36. Xu, Haodong, et al. "PLMD: An updated data resource of protein lysine modifications." Journal of Genetics and Genomics 44.5 (2017): 243-250.
37. Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., ... & Zhou, Y. (2015). Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Scientific reports, 5, 11476.
38. Yang, Y., Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., ... & Zhou, Y. (2017). Spider2: A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. In Prediction of Protein Secondary Structure (pp. 55-63). Humana Press, New York, NY.
39. Rodriguez, Juan José, Ludmila I. Kuncheva, and Carlos J. Alonso. "Rotation forest: A new classifier ensemble method." IEEE transactions on pattern analysis and machine intelligence 28.10 (2006): 1619-1630.
40. Fu, Limin, et al. "CD-HIT: accelerated for clustering the next-generation sequencing data." Bioinformatics 28.23 (2012): 3150-3152.
41. Li, Weizhong, and Adam Godzik. "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics 22.13 (2006): 1658-1659.
42. Huang, Ying, et al. "CD-HIT Suite: a web server for clustering and comparing biological sequences." Bioinformatics 26.5 (2010): 680-682.
43. Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.
44. He, Haibo, et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE, 2008.
45. Faraggi, Eshel, et al. "SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles." Journal of computational chemistry 33.3 (2012): 259-267.
46. Xu, Xiaoqian, et al. "A spine X-ray image retrieval system using partial shape matching." IEEE Transactions on Information Technology in Biomedicine 12.1 (2008): 100-108.
47. Lyons, James, et al. "Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network." Journal of computational chemistry 35.28 (2014): 2040-2046.
48. Altschul, Stephen F., and Eugene V. Koonin. "Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases." Trends in biochemical sciences 23.11 (1998): 444-447.

49. Dehzangi, A., Sohrabi, S., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., & Sattar, A. (2015). Gram-positive and gram-negative subcellular localization using rotation forest and physicochemical-based features. BMC bioinformatics, 16(4), S1.
50. Dehzangi, A., Phon-Amnuaisuk, S., Manafi, M., & Safa, S. (2010, April). Using rotation forest for protein fold prediction problem: An empirical study. In European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (pp. 217-227). Springer, Berlin, Heidelberg.
51. Bustamam, A., Musti, M.I.S., Hartomo, S. et al. Performance of rotation forest ensemble classifier and feature extractor in predicting protein interactions using amino acid sequences. BMC Genomics 20, 950 (2019).
52. Wang, L., You, Z., Yan, X. et al. Using Two-dimensional Principal Component Analysis and Rotation Forest for Prediction of Protein-Protein Interactions. Sci Rep 8, 12874 (2018).
53. Wang, Lei, et al. "Rfdt: A rotation forest-based predictor for predicting drug-target interactions using drug structure and protein sequence information." Current Protein and Peptide Science 19.5 (2018): 445-454.
54. You, Zhu-Hong, Xiao Li, and Keith CC Chan. "An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers." Neurocomputing 228 (2017): 277-282.
55. Geurts, Pierre, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." Machine learning 63.1 (2006): 3-42.

Author Contributions

S.R. Dipta, A. Dehzangi designed and performed the experiments. S.R. Dipta developed the web-server. S.R. Dipta, G. Taherzadeh, A. Dehzangi wrote the manuscript. W. Ahmad, E. Arafat helped with figures and literature review. A. Dehzangi, S. Shatabda mentored and analytically reviewed the paper. All the authors reviewed the article.

Competing interests

The author(s) declare no competing interests.