Computational and Structural Biotechnology Journal 18 (2020) 852–860

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

RF-MaloSite and DL-Malosite: Methods based on random forest and deep learning to identify malonylation sites

Hussam AL-barakati [a,1], Niraj Thapa [a,1], Saigo Hiroto [b], Kaushik Roy [a], Robert H. Newman [c], Dukka KC [d,*]

[a] Department of computational Science and Engineering, North Carolina A&T State University, Greensboro, NC, USA
bFaculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka, JapanFaculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan
cDepartment of Biology, North Carolina A&T State University, Greensboro, NC, USA
[d] Department of Electrical Engineering and Computer Science, Wichita State University, Wichita, KS, USA

ARTICLE INFO

Article history:
Received 16 October 2019
Received in revised form 27 January 2020
Accepted 19 February 2020
Available online 4 March 2020

Keywords:
Malonylation
Post-translational Modification Sites
Random forest
Deep learning
Convolutional neural network

ABSTRACT

Malonylation, which has recently emerged as an important lysine modification, regulates diverse biological activities and has been implicated in several pervasive disorders, including cardiovascular disease and cancer. However, conventional global proteomics analysis using tandem mass spectrometry can be time-consuming, expensive and technically challenging. Therefore, to complement and extend existing experimental methods for malonylation site identification, we developed two novel computational methods for malonylation site prediction based on random forest and deep learning machine learning algorithms, RF-MaloSite and DL-MaloSite, respectively. DL-MaloSite requires the primary amino acid sequence as an input and RF-MaloSite utilizes a diverse set of biochemical, physiochemical and sequence-based features. While systematic assessment of performance metrics suggests that both 'RF-MaloSite' and 'DL-MaloSite' perform well in all metrics tested, our methods perform particularly well in the areas of accuracy, sensitivity and overall method performance (assessed by the Matthew's Correlation Coefficient). For instance, RF-MaloSite exhibited MCC scores of 0.42 and 0.40 using 10-fold cross-validation and an independent test set, respectively. Meanwhile, DL-MaloSite was characterized by MCC scores of 0.51 and 0.49 based on 10-fold cross-validation and an independent set, respectively. Importantly, both methods exhibited efficiency scores that were on par or better than those achieved by existing malonylation site prediction methods. The identification of these sites may also provide important insights into the mechanisms of crosstalk between malonylation and other lysine modifications, such as acetylation, glutarylation and succinylation. To facilitate their use, both methods have been made freely available to the research community at https://github.com/dukkakc/DL-MaloSite-and-RF-MaloSite.

© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Post-translational modifications (PTMs) play a central role in the regulation of nearly all cellular processes. Among the twenty canonical amino acids, lysine residues undergo the most diverse range of modifications [1,2]. For instance, positively-charged lysine residues can be acetylated, ubiquitinated, SUMOylated, glycated, butyrylated, succinylated, and methylated, with different types of modifications leading to different functional outputs in the cell. Recently, malonylation was discovered as yet another type of lysine PTM [3]. Malonylation, which occurs in both eukaryotic and prokaryotic cells, involves a covalent linkage between the ε-amino group of lysine and a malonyl moiety. Unlike acetylation, which involves the addition of relatively small two-carbon chain, malonylation extends the lysine side chain by an additional 4 carbons, making it one of the bulkier acylation modifications. Moreover, because the malonyl moiety contains a negatively-charged carboxylic acid group at the γ-position, malonylation effectively switches the charge of the modified lysine residue from +1 in the unmodified state to −1 following malonylation [3–6]. As a consequence, malonylation can have a profound influence on protein function. Indeed, several studies have found that malonylation influenced signaling processes in a variety of organisms, including humans [4,7], mice [7,8], and the bacterium, *Saccharopolyspora*

* Corresponding author.
E-mail address: dukka.kc@wichita.edu (D. KC).
[1] Equal first author.

https://doi.org/10.1016/j.csbj.2020.02.012
2001-0370/© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

erythreae [9]. Likewise, malonylation has also been implicated in the regulation of inflammation and core metabolic processes [10–12] as well as numerous pathophysiological disorders, such as cancer, cardiac ischemia and muscle weakness [6,7].

Recently, several proteomics strategies have been developed to identify malonylation sites in cells, including mass spectrometry-based methods, chemical probes, affinity enrichment and label-free quantitative methods [3,4,13–15]. However, these approaches are often time-consuming, expensive, and require a high level of technical expertise. As a consequence, several groups have recently developed computational methods to predict malonylation sites *in silico*. These methods, which complement existing experimental methods, not only have the potential to identify novel malonylation sites on proteins that might not be detected by experimental methods (e.g., low abundance proteins or proteins that are only modified under specific cellular conditions), but they also have the potential to offer new insights into the molecular characteristics that lead to malonylation. The first malonylation site predictor, Mal-Ly, was developed using a support vector machine (SVM) algorithm with feature selection of minimum redundancy/maximum relevance (mRMR) based on experimentally identified malonylation sites from *M. musculus* [16]. Subsequent methods, including MaloPred [17], SPRINT-Mal [18], iLMS [19] and the method of Xianget [20], also employed SVMs with distinct feature sets to integrate malonylation site data from various organisms, including humans, mice, *S. erythraea* and *E. coli* [17–21]. Likewise, several methods, such as LEMP [22] and the method developed by Zhang et al. [23], utilized ensemble methods to identify malonylation sites from diverse species. Notably, during the development of LEMP, which is the most recent and best performing malonylation site prediction method developed to date, Chen et al. integrated a deep learning (DL) network classifier based on long short-term memory (LSTM) with word embedding ($LSTM_{WE}$) and RF with enhanced amino acid content (EAAC) features [22]. Nonetheless, there is still room for improvement, as evidenced by MCCs of just 0.246 and 0.244 using 10-fold cross-validation and an independent dataset, respectively. Indeed, though great strides have been made in the performance of malonylation site predictors over the past several years, most successfull methods still suffer with respect to sensitivity (SN) and overall performance (as assessed by MCC). Therefore, we sought to develop malonylation site prediction tools based on DL and RF/Xgboost (Fig. 1). While a DL-based strategy is

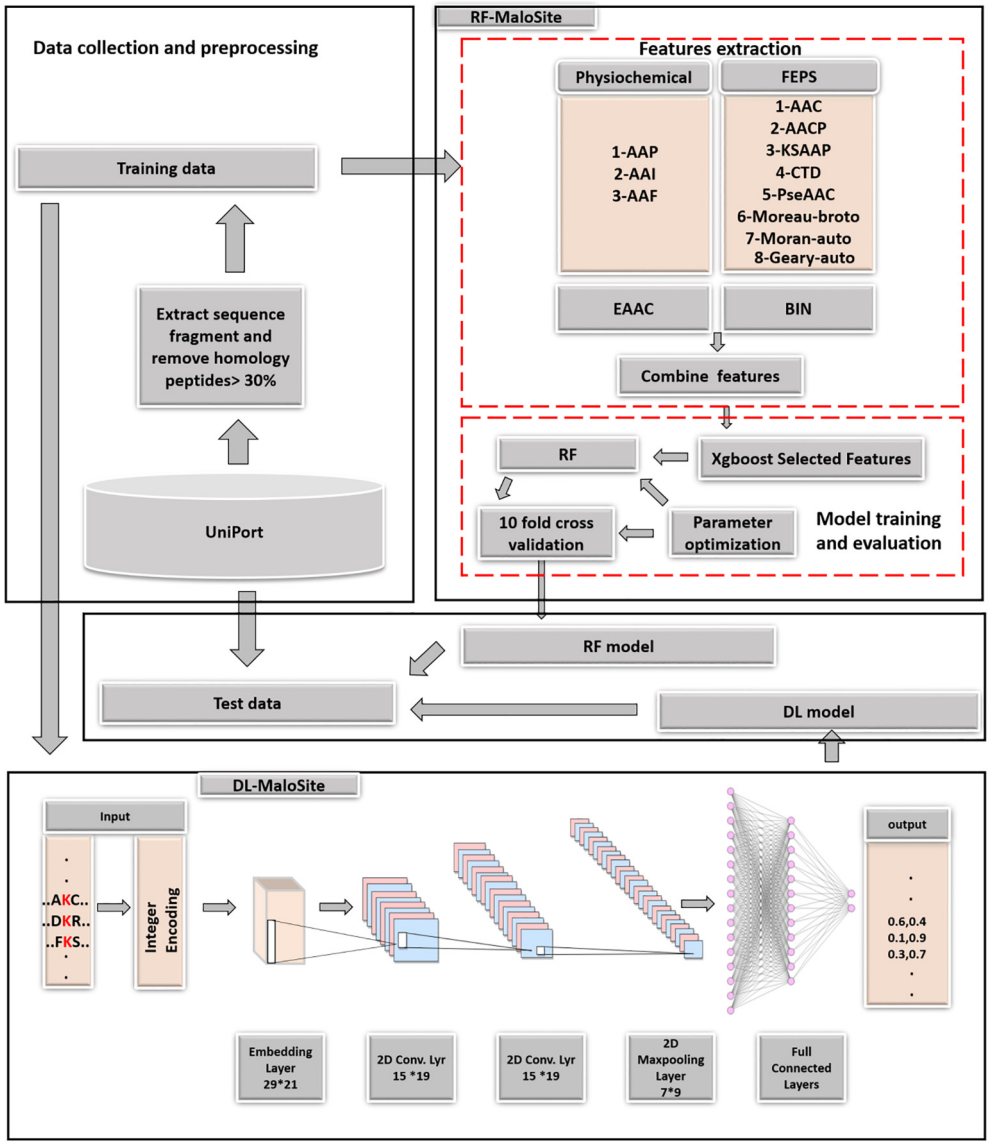

Fig 1. Schematic diagrams illustrating the architectures of RF-MaloSite (A) and DL-MaloSite (B).

highly efficient and not dependent on manually extracted features, the use of RF/Xgboost permits identification of informative features along with different numbers of features. Our methods, which we termed DL-MaloSite and RF-MaloSite, respectively, each improved prediction of malonylation sites considerably and provide complementary information about malonylation site selection.

2. Materials and methods

2.1. Dataset and pre-processing

The original dataset used in the development of our methods was retrieved from previous studies and contained experimentally identified malonylation sites from mice and humans [8,13]. The training and independent sets were assembled in a manner similar to that described by Chen et al. [22]. We also employed the same procedure outlined in Chen et al. [24] and Olsen et al. [25] to keep only non-redundant data. To this end, we used CD-Hit to remove homologous sequences that shared greater than 30% identity [26]. Any sequences that contained "placeholder" resides, such as X, O, U, Z, or J were discarded. Briefly, we first defined experimentally-verified malonylation as positive sites and lysine residues not known to be malonylated as negative sites. We then took a certain number of residues upstream and downstream of the positive/negative site to define a window. To determine the optimal window size, different window sizes were tested and the window with the highest MCC was selected. Finally, we randomly split our dataset so that one-fifth of the dataset was set aside for the independent set and the remaining four-fifths were used for training. Summary statistics for the training and independent sets are given in Table 1.

Feature encoding. To develop a computational tool to distinguish malonlaytion sites, we first needed to convert each residue in the protein sequences into a numerical value. Recently, several features have been proposed to convert protein primary amino acid sequences into mathematical expressions, including amino acid composition (AAC), pseudo amino acid composition (PSAAC), gene ontology (GO), position specific scoring matrix (PSSM), and various other physicochemical properties [24–31]. Likewise, other studies have generated features based on structural information, such as secondary structure, super secondary structure, accessible surface area and local backbone angles [32–36]. Each of these features can provide insights into distinctive types of posttranslational modification. There have been some ensemble approaches for bioinformatics problems [37,38]. In this study, we extracted some features from the Features Extraction from Protein Sequence (FEPS) web application [39]. In total, we extracted 4246 features from the FEPS server. We also combined these features with three types of physiochemical properties, such as amino acid properties (AAP), binary encoding (BE) and enhanced encoding features (EAAC). The final number of features was 6590. The feature classes, many of which are similar to those used to develop our glutarylation site predictor, RF-GlutarylSite, are summarized in Table 2 [40]. A detailed description of each feature class can be found in [40]. The other features are EAAC, amino acid index (AAI) derived from this study [22], and AAP that has been implemented in prediction of S-sulfenylation sites [41]. Here, we briefly describe EAAC.

Table 1
The number of positive and negative sites in the training and independent sets before (left) and after (right) balancing:

Original data	Positive sites (before/after)	Negative sites (before/after)
Training	3978/3978	68,595/3978
Testing	988/988	16,097/988

Table 2
Composition of the complete feature set used during initial model development. Detailed descriptions of each feature class except EAAC, AAP, and AAI are provided in [40].

NO	Feature Class	Abbreviation	Feature Length
1	Binary feature	BIN	560
2	Amino Acid Composition	AAC	20
3	Composition of amino acid pairs	AACP	400
4	K-space amino acid pairs	KSAAP	2400
5	Composition, Transition and Distribution	CTD	147 (C:21; T:21, D:105)
6	Conjoint triad	CT	512
7	Pseudo-amino acid composition	PseAAC	47
8	Autocorrelation	A_u	720
9	Amino acid factor	AAF	140
10	Amino acid properties	AAP	392*
11	Amino acid index	AAI	812*
12	Enhanced amino acid composition	EAAC	440*
13	Combined all features	Total	6590

2.2. Enhanced amino acid composition

EAAC, which was applied by Chen et al. during the development of LEMP [22], is an enhanced version of amino acid composition that counts the frequency of sliding window across the peptide or sequences. Instead of counting the frequency of individual resides within a given fragment, EEAC starts upstream of the site and ends downstream. In our work, we calculated the occurrence of a sliding window that consisted of eight residues, yielding a final dimensional vector of 440.

3. Feature selection

Though using a larger number of relevant features can provide clearer information about the target, if the feature set consists of many irrelevant or correlated features, classification performance can actually be diminished with an expanded feature set. Moreover, large numbers of features cause major issues for computational problems and can increase computational time. Therefore, it is often advantageous to reduce the number of irrelevant and redundant features when employing machine-learning methods [42,43]. To this end, we employed the Gradient Boosted Trees method, Xgboost [44], to detect non-linear associations from the features set. Xgboost has been applied to extract the highest valuable features in many studies [40,45,46]. We applied Xgboost [44] in Python with the Scikit-learn (v 0.19.0) package [47] to capture significant attributes from our training set.

Essentially, we calculated Gini impurity for each feature to find the one that had the best split. This means that we reduced the impurity for each node and reached to the terminal node. High impurity values would suggest more uncertain in identifying class labels. It required more split until reaching to the leaf node.

Next, we computed information gain for each feature and found the one that had maximum values, which was used as important features for the first tree. The same procedure was repeated for other trees. After that, we determined the average importance of each feature from aggregated trees and recorded those as important features for our method.

Gini impurity can be described by:

$$G = \sum_{m=1}^{c_n} t_m(1 - t_m) \quad (1)$$

where C_n is the number of classes and t_m is the probability value of m. Operating values of each node in the gradient boosted trees, we computed Gini Importance by Gini Importance formula as follows:

$$A = G_{parent} - G_{child1} - G_{child2} \quad (2)$$

Any attribute (A) with a significance value lower than 0.002 as threshold was considered as an unimportant attribute and was therefore removed from the feature set. We selected only 104 features as informative features for our approach whereas the rest of the attributes were rejected from the training set due to small importance impact.

Random forest method. RF is a supervised machine-learning technique that uses a bootstrapping algorithm [48–50]. It has been broadly applied to many bioinformatics problems, such as those described in [51–53]. The construction of a random forest is achieved by aggregation of decision trees. Each tree is a random subset of data points (d) from samples (S) with a random subset of attributes (A). It is assumed that data point (d) was randomly chosen with replacements from samples (S) with a random subset of attributes (A). Following that, the best select node was designated from A. Next, each decision tree was expanded as likely without trimming. Finally, using this information, each tree will assign positive or negative class, but the last categorization is made based on maximum voting from the entire tree. In this work, RF was employed using Python (v 3.6.0) with the Scikit-learn (v 0.19.0) [47], and pandas (v 0.20.3) [54], packages were executed to build our approach.

Deep learning based approach. One of the principle advantages of DL over traditional machine learning approaches is the ability to learn relevant features itself. The input for our DL approach, which we named 'DL-MaloSite', is the FASTA format sequence windows. A window size of 29 gave us the optimal results, so it was used in this study. First, the alphabetic sequence for each peptide was converted to respective integers from 0 to 20 for 20 amino acids. These integers were encoded with the window size 29, which were then fed into the DL network.

Model architecture. The overall DL architecture is shown in Fig. 1. The integer encoded input is passed to the embedding layer [55]. Embedding dimension of 21 gave us the optimal results. At the beginning, weights are initialized randomly but the model learns gradually for the improved vector representation. Each vectorization is an orthogonal representation in another dimension, so commonly co-occurring items are together in the vector space, thus performing better than one-hot encoding used in [56,57]. The output dimension is 29×21. A lambda layer was then used to add pseudo dimension before passing to 2-dimensional convolutional layer.

By using 2D convolutional layer, a filter size of 15×3 was possible, which in turn allowed us to include the central PTM site in every stride. The use of this filter size, along with the disabling of padding, allowed the model to be optimized for training time without compromising performance. To avoid overfitting, a relatively high cutoff of 0.6 was used. In this model, two convolutional layers, one maxpooling layer, a fully connected layer with two dense layers and an output layer were used. A rectified linear unit (ReLU) was also used as an activation function for all layers and Adam [58] was used as the optimizer for our architecture. More information about the model and its important parameters are provided in supplementary materials.

Transfer learning. The performance of different DL architectures greatly relies -on the quantity of dataset, which adversely affects the outcome. To circumvent the problems posed by a limited dataset, we employed a transfer learning approach similar to that used during the development of DeepPhos [57]. Specifically, a model developed for a similar task was reused and retrained on a new dataset. In our study, for the second dataset, we used a model trained on a mouse dataset as a base for training on a human dataset to improve the performance.

We have implemented a similar approach during the development of our DL model, DeepSuccinylSite [59] for succinylation site prediction, with better performance metrics than previous machine learning models without the use of manually extracted features.

Model assessment and performance metrics. To assess the implementation of our approach and the current malonylation site predictors' methods, we evaluated our methods using two assessment techniques. The first technique was established by ten-fold cross-validation. Specifically, we divided our data into ten chunks, of which one was used for validation while the other nine were used for training. We then evaluated the performance for each model. This procedure was iterated ten times for 10 models by changing training data points and test data. Lastly, we computed the average results from each iteration. The second technique was using an independent test set to evaluate the overall quality of our approaches.

To evaluate the quality of RF-MaloSite and DL-MaloSite, four statistical metrics were adopted. These metrics were accuracy (ACC), sensitivity (SN), specificity (SP), and Matthew's correlation coefficient (MCC). They have been employed to assess the performance of many approaches in different works [39,60,61].

These measurements can be computed as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100 \tag{3}$$

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100 \tag{4}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \times 100 \tag{5}$$

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

From these metrics, TP denotes true positives (i.e., correctly assigned malonylation sites), TN denotes true negatives (i.e., correctly assigned non-malonylation sites), FP represents false positives (i.e., non-malonylation sites that were incorrectly assigned as malonylation sites) and FN specifies false negatives (i.e., *bona fide* malonylation sites that were incorrectly assigned as non-malonylation sites). Since MCC reports proportions of TP, TN, FP and FN, the assessment value of MCC is generally reflected as alternate for complete quality of approach [62–65].

Receiver Operating Characteristics (ROC) curves were adopted as an additional quality metric. The ROC curve is a graphical plot that shows the performance of sensitivity against 1 – specificity with every likely cutoff [66] while the area under curve (AUC) denotes the level or quantity of separability.

4. Results and discussion

To facilitate the computational prediction of malonylation sites in proteins, we sought to develop a series of malonylation site predictors based on machine learning strategies. To this end, we first established datasets for training and testing of the algorithms. The dataset used in this work was initially extracted from Chen's study [22]. However, this dataset had a highly imbalanced number of positive and negative samples, with a ratio of approximately 16 negative sites for every positive site (i.e., a 1:16 positive-to-negative ratio). Previous studies suggest that imbalanced datasets can cause serious complications with machine learning algorithms, potentially leading to reduced accuracy of prediction [67,68]. To address this potential complication, several computational methods have been described to balance datasets, including over-sampling and under-sampling [52,69,70]. Here, we employed an under-sampling strategy [71], which was a technique that randomly selected the same number of instances from non-malonylation sites that equally balanced with the size of instances

of malonylation sites. This led to a 1:1 positive-to-negative ratio for the training sets used to establish our methods.

Once the training and testing sets were established, we next used the training set to evaluate several statistical algorithms, including SVM, RF, DL, neural network (NN) and k-nearest neighbor (KNN) methods. Initially, we integrated all 6590 features uniformly into these algorithms and compared the performance of each algorithm based on 10-fold cross-validation (Tables S1 and S2). These analyses suggested that inclusion of the entire feature set did not improve the methods' ability to distinguish between those sites that were malonylated and those that were not (and may have even hindered performance). This is likely due to the inclusion of a large number of redundant and non-associated variables. Because decision trees, such as RF and Xgboost, provide information about the relative contribution of each feature to overall method performance, we used Xgboost to identify those features that contributed most substantially to malonylation site prediction (Figs. S1 and S2; Tables S3 and S4). Based on these analyses, we selected 104 features with importance greater than 0.002. As can be seen in Fig. 2, a high percentage of the selected features came from pseudo-amino acid composition (PseAAC; 20%), enhanced amino acid composition (EAAC; 14%) and composition, transition and distribution (CTD; 14%). Indeed, all 10 of the most important features came from one of these three feature classes, with 5 of the 10 coming from CTD, 3 of the 10 coming from PseAAC and 2 of the 10 coming from EAAC (Fig. 3; Table 3). Together, these data suggest that the physiochemical properties of the amino acids surrounding the modified lysine (and particularly the presence of positively charged residues) plays a major role in malonylation site selection. On the other hand, sequence-based features, such as composition of amino acid pairs (AACP) was the least important feature, exhibiting the lowest percentage among all of the important features.

Once determined, the 104 important features were integrated into the RF-, SVM-, NN- and KNN-based frameworks (since DL does not utilize feature information, it was not impacted by feature selection). This strategy helped to improve the performance of each algorithm substantially (Table 4). Not only did feature selection help to simplify and focus the models, but it also has the added benefit of reducing computational time. Importantly, it also suggested that DL- and RF-based methods performed the best with respect to ACC, SN, SP and MCC. Similar results were observed when the independent set was used for evaluation (Table 5) Therefore, we focused on these classifiers during subsequent model development.

In addition to a balanced dataset and feature selection, the window size used during method development is also an important parameter in prediction quality. Indeed, when using the primary amino acid sequence as input, evaluation of the window size helps to identify the impact that contiguous residues have on a given method's ability to distinguish between positive and negative sites. Therefore, to determine the window size(s) that yielded the best performance, we examined window sizes ranging from 25 to 33 residues. During these analyses, we used MCC based on 10-fold cross-validation as a surrogate of overall method performance. While examination of the optimal window size for the DL method entailed a simple sliding window strategy with no feature

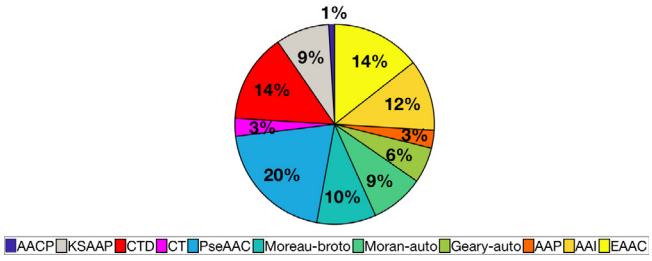

Fig 2. The distributions of each kind of attribute for optimal features with the corresponding percentage score. The total number of selected attributes was 104.

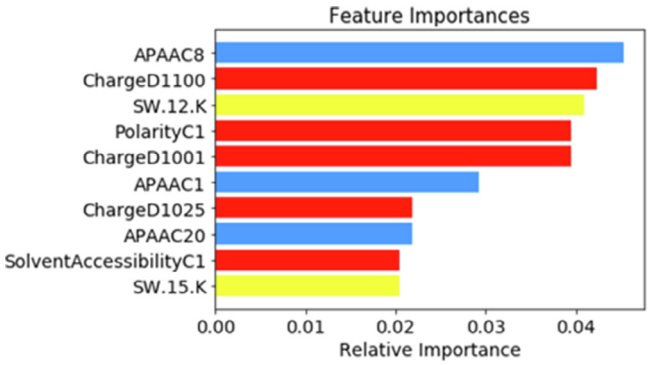

Fig 3. Top ten most important features with corresponding weights.

Table 3
Top ten features ranked by information gain score. PseAAC: Pseudo-amino acid composition; EAAC: Enhanced amino acid composition; CTD: Composition, distribution and transition.

Rank	Features Type	Description
1	APAAC8	PseAAC feature at a particular site
2	ChargeD1100	Distribution feature denoting positive charge of class 1 at last position
3	SW.12.K	EAAC feature denoting enrichment of lysine at portion 12
4	PolarityC1	Composition feature of class 1 for polarity property
5	ChargeD1001	Distribution feature which denotes positive charge of class 1 at first position
6	APAAC1	PseAAC feature at particular site
7	ChargeD1025	Distribution feature which denote positive charge of class 1 at second location
8	APAAC20	PseAAC feature at particular site
9	SolventAccessibilityC1	Composition feature of class 1 for Solvent Accessibility
10	SW.15.K	EAAC feature denoting enrichment of lysine at portion 15

Table 4
Comparison between various machine learning algorithms based on 10-fold cross-validation using 104 features. ACC: Accuracy; SN: Sensitivity; SP: Specificity; MCC: Matthew's Correlation Coefficient.

Classifier	ACC (%)	SN (%)	SP (%)	MCC
DL	**75**	**82**	**69**	**0.51**
RF	**71**	**79**	**63**	**0.42**
SVM	64	71	58	0.30
KNN	61	67	56	0.23
NN	69	73	65	0.37

Table 5
Comparison between our methods and various machine learning algorithms based on an independent test set using 104 features. ACC: Accuracy; SN: Sensitivity; SP: Specificity; MCC: Matthew's Correlation Coefficient.

Classifier	**ACC (%)**	**SN (%)**	**SP (%)**	**MCC**
DL	**74**	**80**	**68**	**0.49**
RF	**70**	**76**	**63**	**0.40**
SVM	65	69	62	0.29
KNN	62	66	58	0.23
NN	68	76	58	0.35

selection, during the evaluation of the RF algorithm, every proposed window size used the optimal features that had been selected by Xgboost. As shown in Table S5 and Fig. S3, the optimal window size for both the RF- and DL-based algorithms was found to be 29. For instance, the MCC of the RF method peaked at 0.42 using a window size of 29, suggesting that this window size contains more useful information for malonylation site identification than other windows. It is interesting to note that both the RF- and DL-based methods showed the best performance using this window size. This may suggest that 29 residues (i.e., 14 residues on each side of a central lysine residue) represents the preferred peptide length for malonylation to occur (and/or to prevent demalonylation by Sirtuin family members). However, the molecular basis for this observation remains to be determined.

While both the final DL- and RF-based methods, which we termed DL-MaloSite and RF-MaloSite, respectively, performed comparably well with respect to ACC, SN, SP and MCC, in general, DL-MaloSite outperformed RF-MaloSite in malonylation site identification. The same was true based on AUC and precision recall (Figs. S4 and 4).

5. Comparison between our methods and existing methods

To determine how DL-MaloSite and RF-MaloSite compare to existing malonylation site prediction methods, we conducted a series of side-by-side comparisons between our methods and existing malonylation site prediction methods. During these analyses, we restricted our comparisons to datasets used during the development of the existing methods, which in some cases required us to implement transfer learning for DL-MaloSite. For instance, for comparisons between our methods and SPRINT-Mal [18] and iLMS [19], we used the same training and independent datasets used by Taherzadeh et al. [18]. Because these datasets were organism-specific (i.e., from *H. sapiens*, *M. musculus* and *S. erythraea*) and DL-MaloSite was trained on a mouse dataset that contained a comparatively larger number of malonylation sites than the *H. sapiens* or *S. erythraea* datasets, we implemented transfer learning for DL-MaloSite when comparisons were made using datasets from these organisms. To this end, the same model was used as the initial phase to be retrained against the new dataset. The evaluations using 10-fold cross-validation against the mouse dataset suggest that, while SPRINT-Mal outperforms our methods with respect to ACC and SP by an average of 16% and 21%, respectively, DL-MaloSite and RF-MaloSite each achieved higher SN scores than the other methods. Indeed, the SN scores exhibited by DL-MaloSite and RF-MaloSite were ~45% and ~47% higher than those achieved by SPRINT-Mal and iLMS (Table 6). As a consequence, DL-MaloSite and RF-MaloSite exhibited the highest MCC scores of all of the methods tested, with MCC scores that were 54% and 38% higher than those achieved by iLMS, respectively (and 90% and 71% higher than those achieved by SPRINT-Mal, respectively). Similar trends were observed using the independent dataset from mouse, where DL-MaloSite and RF-MaloSite achieved MCC scores that were 50% and 19% higher than those exhibited by iLMS, respectively (and 95% and 55% higher than SPRINT-Mal's MCC score) (Table 7). The gains in MCC by our methods were likely due to improved TP rates, which led to substantially higher SN scores without sacrificing SP and ACC, where SPRINT-Mal achieved the highest scores.

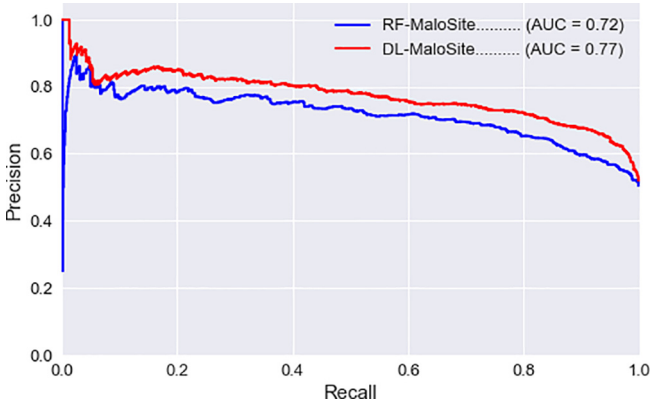

Fig 4. Precision recall (PR) curve based on the independent test set for DL-MaloSite (red) and RF-MaloSite (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The comparisons based on 10-fold cross validation using the human dataset showed that, while the SP scores exhibited by iLMS outpace those of DL-MaloSite and RF-MaloSite by 23% and 29%, respectively, our methods achieved SN scores that were 75% and 48% higher than those scored by iLMS[2] (Table 8). Likewise, DL-MaloSite performed the best with respect to MCC using the human dataset, where it achieved MCC scores that were 2.2-fold higher than iLMS and 47% higher than RF-MaloSite. Similarly, performance evaluation using an independent dataset from humans suggests that, though SPRINT-Mal achieved the highest SP and ACC scores, DL-MaloSite and RF-MaloSite performed well with respect to all metrics (particularly SN and MCC) (Table 9). For instance, the SN scores observed for RF-MaloSite and DL-MaloSite were 88% and 2.2-fold higher than that achieved by SPRINT-Mal. Likewise, DL-MaloSite exhibited the highest MCC score using the independent dataset from humans (0.39), followed by RF-MaloSite (0.24) and then by SPRINT-Mal (0.20).

The differences between method performance were even more pronounced when the independent dataset from *S. erythraea* was used for comparison. Indeed, though the overall trends were similar to those observed using datasets from the other organisms, SPRINT-Mal outperformed RF-MaloSite and DL-MaloSite by 73% and 96%, respectively, with respect to SP and by 32% and 26%, respectively, with regard to ACC (Table 10). Conversely, DL-MaloSite and RF-MaloSite achieved SN scores that were a remarkable 3.9- and 3.4-fold higher than those exhibited by SPRINT-Mal. This culminated in MCC scores for DL-MaloSite and RF-MaloSite that were 3.3- and 2.67-fold higher than the scores observed for SPRINT-Mal.

Finally, we compared our methods to LEMP, the current gold-standard for malonylation site prediction [22]. For these analyses, we used the training and independent test sets described by Chen during the development of LEMP. Since these datasets combined malonylation sites from human, mouse and bacteria, we did not need to employ transfer learning under these circumstances. Based on 10-fold cross-validation, the ACC scores obtained for RF-MaloSite and DL-Malosite were 19% and 15% lower than those achieved by LEMP, respectively (Table 11). Likewise, the SP scores obtained by RF-MaloSite and DL-MaloSite were 31% and 24% lower than those exhibited by LEMP, respectively. In contrast, the SN and MCC scores for both RF-MaloSite and DL-MaloSite were markedly higher than those achieved by LEMP. For instance, the SN scores for RF-MaloSite and DL-MaloSite based on 10-fold cross-validation were 89% and 97% higher, respectively, than the SN score observed for LEMP. Importantly, the MCC scores for RF-MaloSite and DL-MaloSite were 75% and 113% higher, respectively, than that achieved by LEMP.

Similar results were obtained using the LEMP independent dataset, with RF-MaloSite and DL-MaloSite outperforming LEMP

[2] SPRINT-Mal was not evaluated by 10-fold cross-validation using the human dataset.

Table 6
Comparison between SPRINT-Mal [18], iLMS [19], RF-MaloSite and DL-MaloSite based on 10-fold cross-validation using a M. musculus dataset. ACC: Accuracy; SN: Sensitivity; SP: Specificity; MCC: Matthew's Correlation Coefficient.

Features	Organism	ACC (%)	SN (%)	SP (%)	MCC	AUC
SPRINT-Mal [18]	*M. musculus*	**80**	49	**81**	0.21	0.74
iLMS [19]		–	49	80	0.26	0.74
RF-MaloSite		68	**72**	65	**0.36**	**0.75**
DL-MaloSite		70	**71**	68	**0.40**	0.73

Table 7
Comparison between SPRINT-Mal [18], iLMS [19], RF-MaloSite and DL-MaloSite based on an independent test set from M. musculus. ACC: Accuracy; SN: Sensitivity; SP: Specificity; MCC: Matthew's Correlation Coefficient.

Features	Organism	ACC (%)	SN (%)	SP (%)	MCC	AUC
SPRINT-Mal [18]	*M. musculus*	**90**	33	**92**	0.20	**0.74**
iLMS [19]		–	–	–	0.26	0.72
RF-MaloSite		65	**65**	65	**0.31**	0.72
DL-MaloSite		68	**85**	51	**0.39**	0.72

Table 8
Comparison between SPRINT-Mal [18], iLMS [19] RF-MaloSite and DL-MaloSite based on 10-fold cross-validation using a H. sapiens dataset. ACC: Accuracy; SN: Sensitivity; SP: Specificity; MCC: Matthew's Correlation Coefficient.

Features	Organism	ACC (%)	SN (%)	SP (%)	MCC	AUC
SPRINT-Mal [18]	*H. sapiens*	–	–	–	–	–
iLMS [19]		–	48	**80**	0.23	0.74
RF-MaloSite		67	71	62	0.34	0.74
DL-MaloSite (Transfer Learning)		**75**	**84**	65	**0.50**	**0.78**

Table 9
Comparison between SPRINT-Mal [18], iLMS [19] RF-MaloSite and DL-MaloSite based on an independent test set from H. sapiens. ACC: Accuracy; SN: Sensitivity; SP: Specificity; MCC: Matthew's Correlation Coefficient.

Features	Organism	ACC (%)	SN (%)	SP (%)	MCC	AUC
SPRINT-Mal [18]	*H. sapiens*	**91**	35	**89**	0.20	0.70
iLMS [19]		–	–	–	–	–
RF-MaloSite		62	66	59	0.24	0.66
DL-MaloSite (Transfer Learning)		69	**78**	60	**0.39**	**0.75**

Table 10
Comparison between SPRINT-Mal [18], iLMS [19] RF-MaloSite and DL-MaloSite based on an independent test set from S. erythraea. ACC: Accuracy; SN: Sensitivity; SP: Specificity; MCC: Matthew's Correlation Coefficient.

Features	Organism	ACC (%)	SN (%)	SP (%)	MCC	AUC
SPRINT-Mal [18]	*S. erythraea*	**86**	23	**92**	0.12	0.64
iLMS [19]		–	–	–	–	–
RF-MaloSite		65	77	53	**0.32**	**0.67**
DL-MaloSite		68	**90**	47	**0.40**	**0.75**

Table 11
Comparison between DL-MaloSite, RF-MaloSite and LEMP based on 10-fold cross-validation. ACC: Accuracy; SN: Sensitivity; SP: Specificity; MCC: Matthew's Correlation Coefficient.

Features	ACC (%)	SN (%)	SP (%)	MCC	AUC
LEMP [22]	**88**	42	**91**	0.24	**0.82**
RF-MaloSite	71	79	63	0.42	0.77
DL-MaloSite	75	**82**	69	**0.51**	0.81

with respect to MCC by 1.6- and 2.0-fold, respectively (Table 12). Likewise, the SN scores exhibited by RF-MaloSite and DL-MaloSite were 72% and 82% higher than the SN score achieved by LEMP.

Taken together, these the side-by-side comparisons with LEMP suggest that both RF-MaloSite and DL-MaloSite are able to identify malonylation sites with higher sensitivity and higher MCC scores than LEMP. In contrast, though their SP and ACC scores were decent, our methods did not perform as well as LEMP with respect to SP and ACC.

More broadly, we observed that both of our methods performed better than existing methods in predicting malonylation sites, as evidenced by their relatively high SN and MCC scores. In contrast, our methods had some challenges determining non-malonylation sites (i.e., TNs in our datasets). This may be a function of our under-sampling strategy or a consequence of the potentially large

Table 12
Comparison between DL-MaloSite, RF-MaloSite and LEMP based on the independent test. ACC: Accuracy; SN: Sensitivity; SP: Specificity; MCC: Matthew's Correlation Coefficient.

Features	ACC (%)	SN (%)	SP (%)	MCC	AUC
LEMP [22]	**87**	44	**90**	0.24	**0.82**
RF-MaloSite	70	76	63	0.40	0.75
DL-MaloSite	74	**80**	68	**0.49**	0.81

number of as-yet-unidentified malonylation sites. There are ways to increase the performance of our methods, especially with respect to the SP and accuracy metrics. For instance, additional structural features can be introduced to help the classifiers distinguish between malonylation and non-malonylation sites. Likewise, as the number of experimentally-determined malonylation sites grows, we can also increase the number of samples, which can help our method learn more about the problem and enhance prediction.

6. Conclusions

Here, we described the development of two independent machine-learning methods designed to predict malonylation sites in proteins based on their primary amino acid sequences. Our methods, which we termed 'Rf-MaloSite' and 'DL-MaloSite', outperformed existing approaches with respect to SN and MCC. For RF-MaloSite, this highlighted the power of using important features generated via the Xgboost technique and integrated into our RF algorithm. Interestingly, our DL-based model, DL-MaloSite, was able to perform as well or better than existing methods without the need of manual feature extraction. This reduces hassle for manual feature extraction, which can introduce bias as well. For further improvements in our DL model, we can use multi-windows input in a manner similar to DeepPhos [57]. We can also add feature information and position specific scoring matrices (PSSM) in addition to that of the current embedding encoded input to improve performance. However, the main challenge would be the size of data, which will likely be remedied as the number of experimentally-verified malonylation sites continues to grow. Our method can be beneficial for the signaling community and biologists interested in understanding and exploring the impact of malonylation sites on physiological and pathophysiological states. Likewise, these methods will help researchers explore crosstalk between malonylation and other similar types of lysine PTMs, such as acetylation, glutarylation and succinylation.

CRediT authorship contribution statement

Hussam AL-barakati: Methodology, Software, Writing - original draft. **Niraj Thapa:** Methodology, Software, Writing - original draft. **Saigo Hiroto:** Conceptualization, Writing - review & editing. **Kaushik Roy:** Conceptualization, Writing - review & editing. **Robert H. Newman:** Conceptualization, Writing - review & editing. **Dukka KC:** Conceptualization, Writing - review & editing, Supervision.

Declaration of Competing Interest

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

Acknowledgments

This work was supported by National Science Foundation (NSF) grant nos. 2021734, 1564606 and 1901793 (to DK). RHN is supported by an HBCU-UP Excellence in Research Award from NSF (1901793) and an SC1 Award from the National Institutes of Health National Institute of General Medical Science (5SC1GM130545). HS was supported by JSPS KAKENHI Grant Numbers JP18H01762 and JP19H04176.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.02.012.

References

[1] Liu Z, Wang Y, Gao T, Pan Z, Cheng H, Yang Q et al. CPLM: a database of protein lysine modifications. Nucl Acids Res 2014, 42(Database issue):D531-D536.
[2] Lanouette S, Mongeon V, Figeys D, Couture JF. The functional diversity of protein lysine methylation. Mol Syst Biol 2014;10:724.
[3] Peng C, Lu Z, Xie Z, Cheng Z, Chen Y, Tan M et al. The first identification of lysine malonylation substrates and its regulatory enzyme. Mol Cell Proteomics 2011, 10(12):M111. 012658.
[4] Bao X, Zhao Q, Yang T, Fung YME, Li XD. A chemical probe for lysine malonylation. Angew Chem Int Ed 2013;52(18):4883–6.
[5] Olsen CA. Expansion of the lysine acylation landscape. Angew Chem Int Ed Engl 2012;51(16):3755–6.
[6] Xie Z, Dai J, Dai L, Tan M, Cheng Z, Wu Y, et al. Lysine succinylation and lysine malonylation in histones. Mol Cell Proteomics 2012;11(5):100–7.
[7] Saggerson D. Malonyl-CoA, a key signaling molecule in mammalian cells. Annu Rev Nutr 2008;28:253–72.
[8] Colak G, Pougovkina O, Dai L, Tan M, te Brinke H, Huang H, et al. Proteomic and biochemical studies of lysine malonylation suggest its malonic aciduria-associated regulatory role in mitochondrial function and fatty acid oxidation. Mol Cell Proteomics 2015;14(11):3056–71.
[9] Xu J-Y, Xu Z, Zhou Y, Ye B-C. Lysine malonylome may affect the central metabolism and erythromycin biosynthesis pathway in Saccharopolyspora erythraea. J Proteome Res 2016;15(5):1685–701.
[10] He W, Newman JC, Wang MZ, Ho L, Verdin E. Mitochondrial sirtuins: regulators of protein acylation and metabolism. Trends Endocrinol Metab 2012;23(9):467–76.
[11] Lin H, Su X, He B. Protein lysine acylation and cysteine succination by intermediates of energy metabolism. ACS Chem Biol 2012;7(6):947–60.
[12] Qian L, Nie L, Chen M, Liu P, Zhu J, Zhai L, et al. Global profiling of protein lysine malonylation in Escherichia coli reveals its role in energy metabolism. J Proteome Res 2016;15(6):2060–71.
[13] Nishida Y, Rardin MJ, Carrico C, He W, Sahu AK, Gut P, et al. SIRT5 regulates both cytosolic and mitochondrial protein malonylation with glycolysis as a major target. Mol Cell 2015;59(2):321–32.
[14] Du Y, Cai T, Li T, Xue P, Zhou B, He X, et al. Lysine malonylation is elevated in type 2 diabetic mouse models and enriched in metabolic associated proteins. Mol Cell Proteomics 2015;14(1):227–36.
[15] Hirschey MD, Zhao Y. Metabolic regulation by lysine malonylation, succinylation, and glutarylation. Mol Cell Proteomics 2015;14(9):2308–15.
[16] Xu Y, Ding YX, Ding J, Wu LY, Xue Y. Mal-Lys: prediction of lysine malonylation sites in proteins integrated sequence-based features with mRMR feature selection. Sci Rep 2016;6:38318.
[17] Wang L-N, Shi S-P, Xu H-D, Wen P-P, Qiu J-D. Computational prediction of species-specific malonylation sites via enhanced characteristic strategy. Bioinformatics 2016;33(10):1457–63.
[18] Taherzadeh G, Yang Y, Xu H, Xue Y, Liew AWC, Zhou Y. Predicting lysine-malonylation sites of proteins using sequence and predicted structural features. J Comput Chem 2018;39(22):1757–63.
[19] Hasan MM, Kurata H. iLMS, Computational Identification of lysine-malonylation sites by combining multiple sequence features. In: 2018 IEEE 18th international conference on bioinformatics and bioengineering (BIBE): 2018. IEEE: 356–359.
[20] Xiang Q, Feng K, Liao B, Liu Y, Huang G. Prediction of lysine malonylation sites based on pseudo amino acid. Comb Chem High Throughput Screening 2017;20(7):622–8.
[21] Du Y, Zhai Z, Li Y, Lu M, Cai T, Zhou B, et al. Prediction of protein lysine acylation by integrating primary sequence information with multiple functional features. J Proteome Res 2016;15(12):4234–44.
[22] Chen Z, He N, Huang Y, Qin WT, Liu X, Li L. Integration of a deep learning classifier with a random forest approach for predicting malonylation sites. Genomics Proteomics Bioinf 2018;16(6):451–9.

[23] Zhang Y, Xie R, Wang J, Leier A, Marquez-Lago TT, Akutsu T, et al. Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework. Briefings Bioinf 2018:5.
[24] Fujiwara Y, Asogawa M. Prediction of subcellular localizations using amino acid composition and order. Genome Informatics 2001;12:103–12.
[25] Cheng X, Xiao X, Chou K-C. pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. Genomics 2018;110(1):50–8.
[26] Xiao X, Cheng X, Su S, Mao Q, Chou K-C. pLoc-mGpos: incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins. Nat Sci 2017;9(09):330.
[27] Jung I, Matsuyama A, Yoshida M, Kim D. PostMod: sequence based prediction of kinase-specific phosphorylation sites with indirect relationship. BMC Bioinf 2010;11(1):S10.
[28] Bui V-M, Weng S-L, Lu C-T, Chang T-H, Weng JT-Y, Lee T-Y. SOHSite: incorporating evolutionary information and physicochemical properties to identify protein S-sulfenylation sites. In: BMC genomics: 2016. BioMed Central: 9.
[29] Yu L, Zhang Y, Gutman I, Shi Y, Dehmer M. Protein sequence comparison based on physicochemical properties and the position-feature energy matrix. Sci Rep 2017;7:46237.
[30] Bao W, Yang B, Bao R, Chen Y. LipoFNT: lipoylation sites identification with flexible neural tree. Complexity 2019;2019.
[31] Bao W, Yuan C-A, Zhang Y, Han K, Nandi AK, Honig B, Huang D-S. Mutli-features prediction of protein translational modification sites. IEEE/ACM Trans Comput Biol Bioinf 2017;15(5):1453–60.
[32] Bodén M, Yuan Z, Bailey TL. Prediction of protein continuum secondary structure with probabilistic models based on NMR solved structures. BMC Bioinf 2006;7(1):68.
[33] MacCarthy E, Perry D. Advances in protein super-secondary structure prediction and application to protein structure prediction. In: Protein supersecondary structures. Springer; 2019. p. 15–45.
[34] Deng L, Xu X, Liu H. PredCSO: an ensemble method for the prediction of S-sulfenylation sites in proteins. Mol Omics 2018;14(4):257–65.
[35] Reddy HM, Sharma A, Dehzangi A, Shigemizu D, Chandra AA, Tsunoda T. GlyStruct: glycation prediction using structural properties of amino acid residues. BMC Bioinf 2019;19(13):547.
[36] Chandra A, Sharma A, Dehzangi A, Ranganathan S, Jokhan A, Chou K-C, et al. PhoglyStruct: prediction of phosphoglycerylated lysine residues using structural properties of amino acids. Sci Rep 2018;8(1):17923.
[37] Yang B, Chen Y. Somatic mutation detection using ensemble of flexible neural tree model. Neurocomputing 2016;179:161–8.
[38] Yang B, Chen Y, Jiang M. Reverse engineering of gene regulatory networks using flexible neural tree models. Neurocomputing 2013;99:458–66.
[39] Ismail HD, Jones A, Kim JH, Newman RH. Kc DB: RF-Phos: a novel general phosphorylation site prediction tool based on random forest. BioMed Res Int 2016;2016.
[40] AL-barakati HJ, Saigo H, Newman RH, Kc DB: RF-GlutarySite: a random forest based predictor for glutarylation sites. Mol Omics 2019;15(3):189–204.
[41] AL-barakati HJ, McConnell EW, Hicks LM, Poole LB, Newman RH, Kc DB: SVM-SulfoSite: a support vector machine based predictor for sulfenylation sites. Sci Rep 2018;8(1):11288.
[42] Barbu A, She Y, Ding L, Gramajo G. Feature selection with annealing for computer vision and big data learning. IEEE Trans Pattern Anal Mach Intell 2017;39(2):272–86.
[43] Wang R, Perez-Riverol Y, Hermjakob H, Vizcaíno JA. Open source libraries and frameworks for biological data visualisation: a guide for developers. Proteomics 2015;15(8):1356–74.
[44] Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM; 2016. p. 785–94.
[45] White C, Ismail HD, Saigo H, Kc DB: CNN-BLPred: a convolutional neural network based predictor for β-lactamases (BL) and their classes. BMC Bioinf 2017;18(16):577.
[46] Stahl K, Schneider M, Brock O. EPSILON-CP: using deep learning to combine information from multiple sources for protein contact prediction. BMC Bioinf 2017;18(1):303.
[47] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Machine Learn Res 2011;12 (Oct):2825–30.
[48] Breiman L. Random forests. Machine Learn 2001;45(1):5–32.
[49] Polikar R. Ensemble based systems in decision making. IEEE Circuits Syst Mag 2006;6(3):21–45.
[50] Rokach L. Ensemble-based classifiers. Artif Intell Rev 2010;33(1–2):1–39.
[51] Li C, Wang X-F, Chen Z, Zhang Z, Song J. Computational characterization of parallel dimeric and trimeric coiled-coils using effective amino acid indices. Mol BioSyst 2015;11(2):354–60.
[52] Li Y, Wang M, Wang H, Tan H, Zhang Z, Webb GI, et al. Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features. Sci Rep 2014;4:5765.
[53] Zhou Y, Liu S, Song J, Zhang Z. Structural propensities of human ubiquitination sites: accessibility, centrality and local conformation. PLoS ONE 2013;8(12): e83167.
[54] McKinney W. Data structures for statistical computing in python. In: Proceedings of the 9th python in science conference: 2010. Austin, TX: 51–56.
[55] Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model. 2001. URL http://www iro umontreal ca/~lisa/pointeurs/nips00_lm ps.
[56] Wang D, Zeng S, Xu C, Qiu W, Liang Y, Joshi T, et al. MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. Bioinformatics 2017;33(24):3909–16.
[57] Luo F, Wang M, Liu Y, Zhao X-M, Li A. DeepPhos: prediction of protein phosphorylation sites with deep learning. Bioinformatics 2019.
[58] Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv:14126980 2014.
[59] Wang S, Peng J, Ma J, Xu J. Protein secondary structure prediction using deep convolutional neural fields. Sci Rep 2016;6:18962.
[60] Geng H, Lu T, Lin X, Liu Y, Yan F. Prediction of protein-protein interaction sites based on naive Bayes classifier. Biochem Res Int 2015;2015.
[61] Chen Y-Z, Tang Y-R, Sheng Z-Y, Zhang Z. Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. BMC Bioinf 2008;9(1):101.
[62] Šícho M, de Bruyn Kops C, Stork C, Svozil D, Kirchmair J. FAME 2: simple and effective machine learning model of cytochrome P450 Regioselectivity. J Chem Inf Model 2017;57(8):1832–46.
[63] Chen C-W, Lin J, Chu Y-W. iStable: off-the-shelf predictor integration for predicting protein stability changes. In: BMC bioinformatics: 2013. BioMed Central: S5.
[64] Chen Y-Z, Chen Z, Gong Y-A, Ying G. SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties. PLoS ONE 2012;7(6):e39195.
[65] Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16 (5):412–24.
[66] Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett 2006;27 (8):861–74.
[67] Hasan MM, Zhou Y, Lu X, Li J, Song J, Zhang Z. Computational identification of protein pupylation sites by using profile-based composition of k-spaced amino acid pairs. PLoS ONE 2015;10(6):e0129635.
[68] Chen Z, Chen Y-Z, Wang X-F, Wang C, Yan R-X, Zhang Z. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. PLoS ONE 2011;6(7):e22930.
[69] Chen X, Qiu J-D, Shi S-P, Suo S-B, Liang R-P. Systematic analysis and prediction of pupylation sites in prokaryotic proteins. PLoS ONE 2013;8(9):e74002.
[70] Shi S-P, Qiu J-D, Sun X-Y, Suo S-B, Huang S-Y, Liang R-P. PLMLA: prediction of lysine methylation and lysine acetylation by combining multiple features. Mol BioSyst 2012;8(5):1520–7.
[71] Estabrooks A, Jo T, Japkowicz N. A multiple resampling method for learning from imbalanced data sets. Comput Intell 2004;20(1):18–36.