International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 18, 2019, pp 1-16
© 2019 Dept. of Statistics, Univ. of Rajshahi, Bangladesh

Computational Prediction of Protein Ubiquitination Sites using the Composition of Amino Acid Pairs

Md. Selim Reza[1*], Samme Amena Tasmia[1], Mst. Ayesha Siddika[2], Adiba Sultana[1], Md. Mehedi Hasan[3], Md. Hadiul Kabir[1] and Md. Nurul Haque Mollah[1]

[1]Bioinformatics Lab., Department of Statistics, Rajshahi University,
Rajshahi-6205, Bangladesh

[2]Microbiology Lab., Department of Veterinary and Animal Sciences,
Rajshahi University, Rajshahi-6205, Bangladesh

[3]School of Life Sciences, State and Key Laboratory of Agrobiotechnolgy,
The Chinese University of Hong Kong, N.T., Hong Kong

[*]Correspondence should be addressed to Md. Selim Reza
(selim.ru4778@gmail.com)

[Received May10, 2019; Revised July 17, 2019; Accepted November 13, 2019]

Abstract

As one of the most reversible protein post-translation modification is ubiqutination, which can involve in lots of biological processes and closely implicated with various diseases. Therefore, identification of ubiquitination site is an important task for understanding the mechanisms of ubiqutinaion. However, the identification of ubiquitationsites in experimental approaches is time consuming and costly. As an alternative, computational identification is more useful and reliable. In this study, we proposed a computational method using some encoding schemes and feature selection methods for analyzing ubiquitination sites. Herein, we compared six machine learning approaches. Finally, random forest classifier with feature selection (i.e. Wilcoxon signed rank test) based on 1:1 ratio for independent test set performed better than other combinations. Our proposed method achieves significantly better performances on both of cross-validation and independent data test. Thus the proposed method anticipates a helpful computational resource for predicting ubiquitination sites.

Keywords: Post-translation modification, Ubiquitination sites prediction, Sequence encoding, Feature selection, random forest classifier.

AMS Classification: 92B10.

1. Introduction

Ubiquitination (also called Ubiquitylation) is an important mechanism and has been found in widespread reversible post-translation modification (PTMs). Ubiquitination occurs when Ubiquitin (Ub) will be linked to specific lysine (k) residues of target proteins. One of the most important and universal PTMs, protein ubiquitination is a rapid and reversible biochemical process in which an isopeptide bond forms covalently the C-terminal double- glycine carboxy group of a ubiquitin protein and the Ɛ-amino group of lysine residues of a substrate protein (Pickart et al., 2001). Three enzymes are engaged in the process of ubiquitination, including Ub-activating (E1), Ub-conjugating (E2) and Ub-ligating (E3) enzymes and the types of ubiquitination are diverse (e.g. the targeting proteins can be jointed with a single Ub or poly-Ub chains) (Haglund et al., 2005; Radivojac et al., 2010; Tung et al., 2008; Hershko et al., 1998). The first discovered function of ubiquitination is to target proteins for subsequence degradation by the ATP-dependent ubiquitin-proteasome system. Subsequently, many regulatory functions of ubiquitination were discovered by the regulation of DNA repair and transcription, control of signal transduction and implication of endocytosis and sorting (Hermann et al., 2007; Welchman et al., 2005). Ubiquitination has been indicated to be involved in regulating a diversity of basic biological cellular process, such as signal transduction,cell division/mitosis, apoptosis, and endocytosis (Sun and Chen et al., 2004; Reinstein and Ciechanover et al., 2006; Hoeller et al., 2006; Hicke, 2001), including the degrationprotein (Hicke L et al., 2001;Pickart, 2001). About 80% of the cellular proteins are degraded by the Ub-proteasome system (Hermann et al. 2007). An aberrance of the ubiquitin-proteasome system (UPS) is associated in manifold pathological diseases, such as Irritantdiseases, neuro-degenerative and cancer (Hoeller et al. 2006; Reinstein and Ciechanover et al., 2006). Researchers have employed several experimental methods to rectify ubiquitylated proteins such as the use of affinity-tagged Ub, Ub antibodies and Ub-binding proteins and high throughput mass-spectrometry (MS) technique (Tomlinson E et al., 2007; Peng J et al., 2003), some user-friendly database such as UbiProt (http://ubiprot.org.ru/) (Chernorudskiy AL et al., 2007), SCUD (http://scud.kaist.ac.kr) (Lee WC et al., 2008) and SysPTM (http://www.sysbio.ac.cn/SysPTM) (Li H et al., 2009). Catic and co-worker, methodically analyzed 135 ubiquitination sites in 95 yeast proteins (Catic et al., 2004), they found that ubiquitination sites promoted to be exposed at the molecular surface and reside in loop regions (Catic et al., 2004).Radivojac also analyzed the structural context of ubiquitination PTMs sites and confirmed that these sites were preferentially located in intrinsically disordered regions (Radivojac et al., 2010).Identification of ubiquitinated proteins PTMs sites is one of the greatest challenges in profiting a full understanding of the regulatory roles of ubiquitination law and the molecular mechanism of the ubiquitin system. It is time consuming and labor-intensive to use conventional experimental approaches to identify the large-scale ubiquitination proteins PTMs sites, such as sitemutagenesis (Lin et al., 2005) and antibodies of Ub (Gentry et al., 2005). Ubiquitination sites prediction is usually presented by a sequence fragment of 2*n* +1 residues with the residue lysine (K) in the central position (i.e. the window size is equal to 2*n*+1). A congenial feature construction or encoding scheme of the sequence fragment is further required for the processing of a prediction algorithm, finally a predictor can be constitute by some statistical based algorithms.Until now, several ubiquitination protein PTMS sites prediction method have been developed elegantly, such as developed an ubiquitination site predictor (UbiPred) (Tung et al., 2008) using a Support Vector Machine (SVM), developed a UbPred by random forest (RF) based predictor (Radivojac et al., 2010), very recently, developed a nearest neighbor algorithm based ubiquittination site predictor (Cai et al., 2011).However, the overall achievement of the above-specified existing predictors is not yet in the satisfactory level and there is further need to improve the prediction performance. In this paper, we develop a new predictor based on combining multiple features (including 1:1 ratio, random forest classifier, Wilcoxon signed rank test). We will observe that our proposed method improves the performance over the existing predictors of ubiquitination site.

2. Materials and Methods

Datasets

In this study to build comparative study, 203 ubiquitylated substrates which were already accumulated by (Radivojac et al., 2010), downloaded from http://www.ubpred.org/sgd_predictions.txt.gz. These 203 proteins controlled 272 experimentally authenticated ubiquitination sites, that considered as positive sample (i.e. ubiquitination sites) and rest of the K residues can be considered as negative sample (i.e. non-ubiquitination sites), which were not response as ubiquitination sites in these proteins. As previously declared in introduction part, each sample is represented by a sequence fragment with window size 2n+1, optimally 27 window size is selected that our initial computational experiments.

The CKSAAP Encoding Scheme

In this study, a protein ubiquitination or non-ubiquitination site is defined by a sequence fragment of 27 amino acids. The CKSAAP encoding elaborate that the composition of k-spaced amino acids pairs in the fragment. For example, k=0 with non-existing amino acid O, there are (21×21)= 441 types of amino acid pairs (i.e., AA, AC, . . ., YY, OO) then a feature vector of that size is used to represent the composition of these pairs, which can be described as

$$\left(\frac{N_{AA}}{N_{Total}}, \frac{N_{AC}}{N_{Total}}, \frac{N_{AD}}{N_{Total}}, \ldots, \frac{N_{YY}}{N_{Total}}, \frac{N_{OO}}{N_{Total}}\right)_{441}$$

The value of each feature denotes the composition of the corresponding residue pair in the fragment. For an instant, if an AA pair occurs m times in this fragment, the corresponding value in the vector (i.e. N_{AA}) is m. when the value of k increased, the prediction accuracy and the sensitivity would increase, but the computational complexity and the required time for training the models would also increase. So that we consider in this paper k=0, 1, 2, 3, 4, and 5, and the total dimension of the 5-spaced feature vector is 2646.

The Binary Encoding Scheme

The binary encoding is also carried out here to compare with the CKSAAP encoding. As mentioned in the above, there are 21 types' amino acids with non-existing amino acid O in our setting such as ACDEFGHIKLMNPQRSTVWYO. Therefore, each amino acid is represented by a 21-dimensional binary vector, that is, A corresponds to A (100000000000000000000), C corresponds to C (010000000000000000000), . . ., O corresponds to O (000000000000000000001). For each sequence fragment, the central amino acid is always lysine (k), which is not necessary to be defined. Therefore, the total dimension of the binary encoding scheme is 21×26 = 546.

Feature selection

In this study, we used non-parametric test filter method. A non-parametric test likes as, Wilcoxon signed rank test, and at first, we find p-value than ordered this value and extracted optimally features from the dataset.

Let be the sample size, for pairs $i = 1, \dots, N$, let $x_{1,i}$ and $x_{2,i}$ denote the measurements.

H_0 : difference between the pairs follows a symmetric distribution around zero

H_1 : difference between the pairs does not follow a symmetric distribution around zero.

1. For $i = 1, \dots, N$, calculate $|x_{2,i} - x_{1,i}|$ and $sign(x_{2,i} - x_{1,i})$
2. Exclude pairs with$|x_{2,i} - x_{1,i}| = 0$. Let N_r be the reduced sample size
3. Order the remaining, $N_r$pairs from smallest absolute difference to largest absolute difference, $|x_{2,i} - x_{1,i}|$.
4. Rank the pairs, starting with the smallest as 1. Ties receive a rank equal to the average of the ranks they span. Let $R_i$denote the rank.
5. Calculate the test statistics W
$$W = \sum_{i=1}^{N_r} [sign(x_{2,i} - x_{1,i}). R_i]$$
6. Under H_0, W follows a specific distribution with no simple expression. This distribution has an expected value 0 and a variance of $\frac{N_r(N_r+1)(2N_r+1)}{6}$
7. As N_r increases, the sampling distribution of W converges to a normal distribution. Thus, $N_r \geq 10$, a Z-score can be calculated as $Z = \frac{W}{\sigma_W}$, where $\sigma_W = \sqrt{\frac{N_r(N_r+1)(2N_r+1)}{6}}$ and finally P-value can be calculated.

Linear Discriminant Analysis (LDA)

LDA is a generalization of Fisher's linear discriminant. For optimal classification, we need to know the class posteriors probability p$(\pi i/D)$ and Bayes theorem gives us following relation $P(\frac{\pi i}{D})=\frac{P\left(\frac{D}{\pi i}\right)Pi}{\sum_{i=1}^{k}P\left(\frac{D}{\pi i}\right)Pi}$. We see that in terms of ability to classify, having the p $(D/\pi i)$ is almost equivalent to having the quality p $(\pi i/D)$. Suppose that we model each class density as multivariate Gaussian, Pi(x)=$\frac{1}{(2\pi)^{\frac{p}{2}}|\sum|^{\frac{1}{2}}}\ exp[-\frac{1}{2}(X-\mu i)^{T}\sum^{-1}(X-\mu i)$; Where,$\mu i$ is a mean vector of x and $\sum$ is a covariance matrix.Linear discriminant analysis (LDA) arises in the special cases when we assume that the class have a common covariance matrix$\sum i=\sum\ \forall\ i$. In comparing two classes the ratio of two densities is,

$$\frac{p1(x)}{p2(x)}=\frac{exp[-\frac{1}{2}(X-\mu 1)^{T}\sum^{-1}(X-\mu 1)}{exp[-\frac{1}{2}(X-\mu 2)^{T}\sum^{-1}(X-\mu 2)}.$$

Naive Bayes (NB)

Naïve Bayes is a predictive algorithm based on the statistical learning theory of Bayesian theorem. Let the input *X*= ($x_1x_2\ldots x_p$) for the Naïve Bayes classifier produce a binary class C∈{1, -1}, where 1 denotes the residues was predicted as ubiquitination sites and -1 denotes the residues non- ubiquitination sites. The NB was trained using a set of labeled training dataset (X, C). The NB classifier can be defined as,

$$\frac{P(C=1|X=x1,x2,\ldots,xp)}{P(C=-1|X=x1,x2,\ldots,xp)}=\frac{P(C=1)\Pi_{i=1}^{p}P(xi|C=1)}{P(C=-1)\Pi_{i=1}^{p}P(xi|C=-1)}$$

$$\text{If}\ \frac{P(C=1|X=x1,x2,\ldots,xp)}{P(C=-1|X=x1,x2,\ldots,xp)}\geq\theta$$

Then the residue of the input X was classified as 1 (ubiquitination sites) otherwise -1 (non- ubiquitination sites) and θ is the classification threshold.

Support Vector Machine (SVM)

The original SVM algorithm was invented by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. In 1992, Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik suggested a way to create nonlinear classifiers by applying the kernel trick to maximum-margin hyperplanes (Burden TS, 1976). Our training data consists of N pairs (x_1,y_1), (x_2,y_2),…, (x_N, y_N), with $x_i \in {}_{R}{}^{p}$ and $x_i \in$ {-1,1}, where 1 denotes that the residue was predicted a ubiquitination sites and -1 denotes the residues non-ubiquitination sites. Define a hyperplane by**,** {*x*: *f(x)*=

$x^{T}\ \beta + \beta_{0}$}; Where β is a unit vector: || β|| = 1. SVM solves the following optimization problem:

$$\min_{\beta,\beta 0}\frac{1}{2}||\beta||^{2} + \gamma \sum_{i=1}^{N} \xi i$$

$$\text{Subject to }_{,}\ \xi_{i} \geq 0,\ y_{i}(x_{i}{}^{T}\ \beta + \beta_{0}) \geq (1 - \xi_{i})\ \forall_{I}$$

AdaBoost (ABoost)

AdaBoost, short for "Adaptive Boosting", is a machine learning meta-algorithm conveyed by Yoav Freund and Robert Schapire who won the Gödel Prize in 2003 for their work. AdaBoost is a method for combining many weak classifiers to make a strong classifier. Input: Set of weak classifiers {φ_{μ}(x): μ = 1,…,M}. Labelled data X = {(x^{i},y^{i}): i = 1,...,N} with y^{i}∈ {±1}.Output: Strong classifier:

$$S(x) = \text{sign } (\textstyle\sum_{\mu=1}^{M} \theta\mu\varphi\mu(x))$$

Where the {θ_{μ}} are weights to be learned. We generally want most of the θ_{μ} = 0, which means that the corresponding weak classifier φ_{μ}(.) is not selected.

K-Nearest Neighbor (KNN)

The most basic instance-based algorithm is the k-Nearest Neighbor (KNN) algorithm. It assumes all instance correspond to points in the n-dimensional space R^{n}, the distance between instances is usually taken as the Euclidian distance, i.e., if an instance x_{i} is x_{i}= [$x_{i}^{1}, \dots, x_{i}^{n}$], Where x_{i}^{T} denote the value of the r-th feature of instance x_{i} ,then the distance between two instance x_{i} and x_{j} is

$$d(x_{i}, x_{j})=\sqrt{\textstyle\sum_{r=1}^{n}(x_{i}^{r} - x_{j}^{r})^{2}}$$

Other distance matrices can be used as well the KNN algorithm, (Mitchell, 1977) assigns a quarry sample to the class that has a maximum number 0f representatives among the k training samples closets to it. Ties are usually broken at random. If k=1 then the KNN algorithm assigns the query to the class of the nearest training sample.

Random Forests

In this study, we have used the random forest algorithm, which is shown to have the capability of handling many input variables and avoiding model over fitting.

(Breiman L et al., 2001). Random forests (RF) use a combination of independent decision trees to improve classifier performance. Specifically, each decision tree in a forest is constructed using a bootstrap sample from the training data,and the class with the most votes will be output as the predicted class of the random forest. Each tree is constructed using the following procedure:

a) Suppose the number of training cases is N, take N samples at random with replacement from the original data.
b) If there are M input variables, choose a number m variable which should be much less than M variables. At each node, m variables are selected randomly from the original M variables and the most optimized split on these m variables is employed to split the node. The value of m does not change during the growth of the forest.
c) Each tree is fully grown and not pruned.

Performance assessment

In this study, there are four measurements: Sensitivity (Sn), Specificity (Sp), Accuracy (Ac) and Matthew’s correlation coefficient (MCC) (Baldi et al. 2000). They are defined as follows:

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$Ac = \frac{TP + TN}{TP + FN + TN + FP}$$

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

Where TP, TN, FP and FNare the true positive, true negative, false positive and false negative respectively. In addition, the prediction validity is often examined by Receiver Operating Characteristic (ROC) curve (Centor et al. 1991; Gribskov et al. 1996), which plots the true positive rate (i.e. Sn) against the false positive rate (i.e. 1-Sp) for all possible thresholds. Besides, the overall performance of CKSAAP_UbSite can also be quantified by the corresponding area under the ROC curve (AUC).

3. Result and Discussion

Performance assessment on the training dataset

In nature, the Ubiquitination sites and non- Ubiquitination sites datasets are highly unbalanced. It has been established that statistical learning algorithms become computationally intractable and the accuracy is strongly affected due to the nature of the unbalanced datasets. To address this issue, many PTM site prediction studies employ a relatively balanced ratio between the positive and negative samples during the training of the classifiers (e.g. the ratio of positives versus negatives is controlled at 1 : 1 or 1 : 2), including the Ubiquitination sites prediction as well. The Ubiquitination sites prediction made up by trained and tested on a balanced dataset through a 5- fold cross-validation. The balanced dataset was prepared by 1:1 ratio of positive and negative sample (i.e. 272 ubiquitination sites and 272 non-ubiquitination sites selected from Radivojac_dataset). The detailed performance measurements for different method such as random forest (RF), support vector machine (SVM), naïve bayes (NB), linear discriminant analysis (LDA), K-nearest neighbor (KNN), adaboostbased on 1:1 ratio of positive and negative sample in Table-1. We showed that RF classifier was best performance than other method like as (Sn = 99.9%, Sp = 99.9%, Acc = 99.9%, Mcc = 99.1%) in Table-1 and the area under roc curve (AUC) was 99.9%. Furthermore, the ROC curve of different classifier was plotted in Figure 1.We extracted feature by non-parametric method as Wilcoxon signed rank test from 1:1 ratio of positive and negative sample and the detailed performance measurements was summarized in Table-2. In this case we showed that RF classifier was best performance than other method like as (Sn = 99.9%, Sp = 99.9%, Acc = 99.9%, Mcc = 99.9%) in Table-2 and the area under ROC curve (AUC) was 99.9% and the ROC curve of different method was plotted in Figure 2. Now the balanced dataset was prepared by 1:2 ratio of positive and negative sample (i.e 272 ubiquitination sites and 544 non-ubiquitination sites selected from Radivojac_dataset). The detailed performance measurement was summarized in Table-3 and the ROC curve of different classifier was plotted in Figure 3. In this case we showed that RF classifier was best performance than other method like as (Sn = 92.0%, Sp = 99.1%, Acc = 99.2%, Mcc = 99.6%) in Table-3 and the area under ROC curve (AUC) was 98.5% and the ROC curve of different classifier was plotted in Figure 3. After feature selection the detailed performance measurement was summarized in Table-4 and the ROC curve of different classifier was plotted in Figure 4 and RF classifier was best performance than other method like as (Sn = 90.0%, Sp = 99.1%, Acc = 98.5%, Mcc = 90.2%) in Table-4 and the area under ROC curve (AUC) was 98.0% and the ROC curve of different classifier was plotted in Figure 4.

Performance assessment on the test dataset

In this section, at firstthe test dataset result for the model of trained on 1:1 ratio of positive and negative sample was summarized in Table-5 and the ROC curve of was plotted in Figure 5. In this case we showed that RF classifier was best performance than other method like as (Sn = 79.0%, Sp = 97.8%, Acc = 97.6%, Mcc = 80.0%) in Table-5and the area under ROC curve (AUC) was 97.8% and the ROC curve of different method was plotted in Figure 5. The detailed performance measurement for the model of trained on 1:2 ratio of positive and negative sample was summarized in Table-6 and the ROC curve was plotted in Figure 6. In this case we showed that RF classifier was best performance than other method like as (Sn = 80.0%, Sp = 97.5%, Acc = 97.2%, Mcc = 86.0%) in Table-6 and the area under ROC curve (AUC) was 96.9% and the ROC curve of different method was plotted in Figure 6.

Table 1: Comparison of different method with 1:1 ratio of positive and negative sample.

Classifier	Sensitivity	specificity	accuracy	AUC	MCC	Error.rat
LDA	0.992	0.999	0.978	0.995	0.116	0.009
NB	0.148	0.987	0.589	0.567	0.253	0.410
SVM	0.996	0.940	0.942	0.970	0.477	0.058
ABoost	0.965	0.972	0.972	0.968	0.784	0.028
KNN	0.995	0.985	0.989	0.988	0.977	0.001
RF	**0.999**	**0.999**	**0.999**	**0.999**	**0.991**	**0.001**

Table 2: Comparison of different method with 1:1 ratio of positive and negative sample for feature selection

Classifier	Sensitivity	specificity	Accuracy	AUC	MCC	Error.rat
LDA	0.533	0.924	0.923	0.729	0.116	0.076
NB	0.970	0.996	0.994	0.983	0.961	0051
SVM	0.999	0.925	0.925	0.962	0.181	0.074
ABoost	0.778	0.944	0.939	0.861	0.464	0.060
KNN	0.857	0.924	0.924	0.890	0.135	0.075

RF	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.001**

Table 3: Comparison of different method with 1:2 ratio of positive and negative sample

Classifier	Sensitivity	specificity	Accuracy	AUC	MCC	Error.rat
LDA	0.823	0.979	0.969	0.901	0.179	0.030
NB	0.123	0.983	0.539	0.553	0.208	0.460
SVM	0.999	0.938	0.939	0.969	0.384	0.060
ABoost	0.731	0.947	0.941	0.839	0.433	0.058
KNN	0.840	0.933	0.932	0.887	0.234	0.067
RF	**0.920**	**0.991**	**0.992**	**0.985**	**0.996**	**0.001**

Table 4: Comparison of different classifier with 1:2 ratio of positive and negative sample for feature selection

Classifier	Sensitivity	specificity	Accuracy	AUC	MCC	Error.rat
LDA	0.285	0.9289	0.927	0.607	0.0449	0.072
NB	0.670	0.995	0.962	0.833	0.779	0.037
SVM	1.000	0.931	0.931	0.965	0.194	0.068
ABoost	0.614	0.934	0.931	0.774	0.227	0.069
KNN	0.820	0.933	0.932	0.876	0.242	0.067
RF	**0.900**	**0.991**	**0.985**	**0.980**	**0.902**	**0.001**

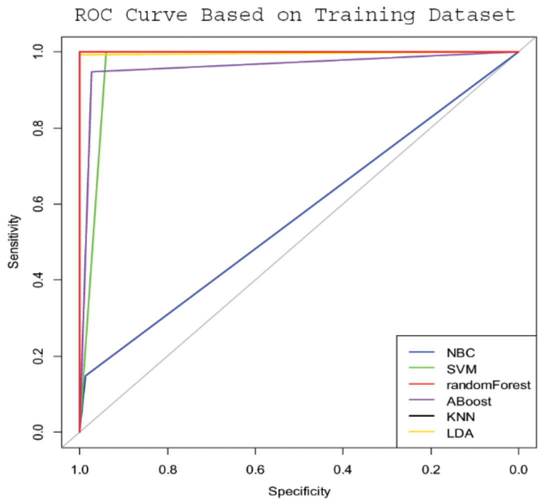
Table 5: Based on test dataset with 1:1 ratio of positive and negative sample

Classifier	Sensitivity	specificity	Accuracy	AUC	MCC	Error.rat
LDA	0.088	0.940	0.938	0.514	0.006	0.061
Knn	0.084	0.940	0.940	0.512	0.0028	0.077
ABoost	0.078	0.941	0.926	0.511	0.011	0.073
RF	**0.790**	**0.978**	**0.976**	**0.978**	**0.800**	**0.010**

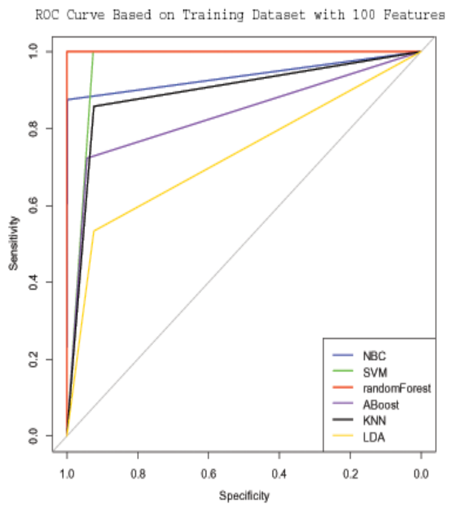
Table 6: Based on test dataset with 1:2 ratio of positive and negative sample

Classifier	Sensitivity	specificity	Accuracy	AUC	MCC	Error.rat
LDA	0.085	0.941	0.939	0.513	0.005	0.061
NB	0.606	0.993	0.959	0.799	0.719	0.040
Knn	0.092	0.941	0.940	0.516	0.005	0.072
ABoost	0.099	0.941	0.932	0.520	0.018	0.067
RF	**0.800**	**0.975**	**0.972**	**0.969**	**0.860**	**0.011**

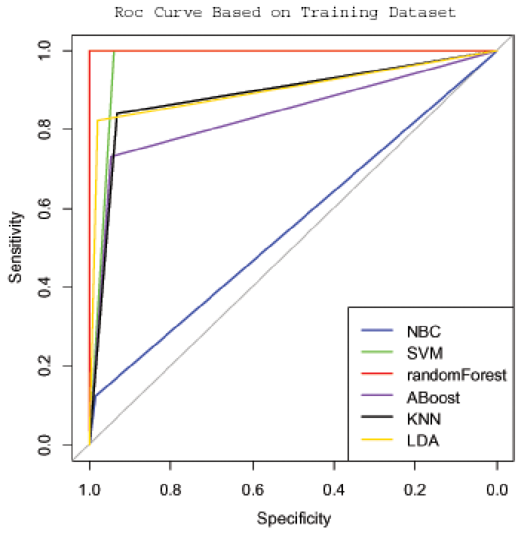
(1)


ROC Curve Based on Training Dataset
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
NBC
SVM
randomForest
ABoost
KNN
LDA

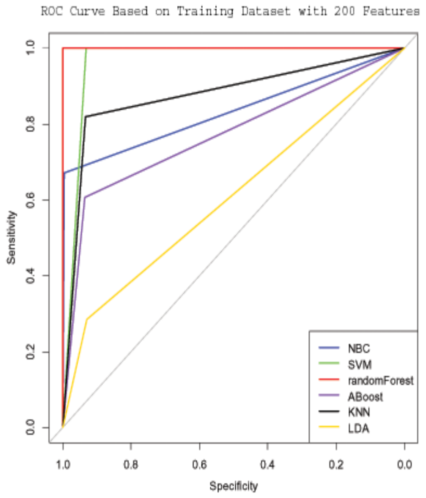
(2)


ROC Curve Based on Training Dataset with 100 Features
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
NBC
SVM
randomForest
ABoost
KNN
LDA

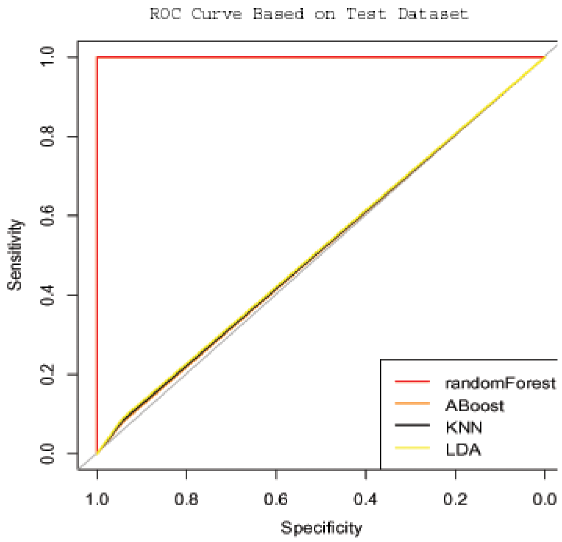
(3)


Roc Curve Based on Training Dataset
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
NBC
SVM
randomForest
ABoost
KNN
LDA

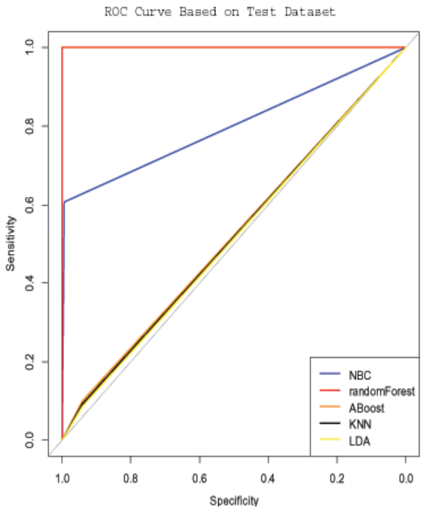
(4)


ROC Curve Based on Training Dataset with 200 Features
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
NBC
SVM
randomForest
ABoost
KNN
LDA

(5)



(6)



For the training dataset, the amino acid propensities of surrounding ubiquitylated sites compared to the non- ubiquitylated sites were displayed by Two Sample Logos software (**Figure 7**). Briefly, in the two sample logo, only over- or under-represented residues at each position are plotted above and under the X-axis, respectively. The height of the letter was in proportion to the percentage of positive (if over-represented) or negative samples (if under-represented) protecting the corresponding residue. The Y-axis reports the cumulative percentage of these over or under represented residues. We can see that some amino acids are over/under represented in the specific positions (**Figure 7**), which indicates that the positional amino acid encoding was an efficient method to identify the ubiquitylated sites.

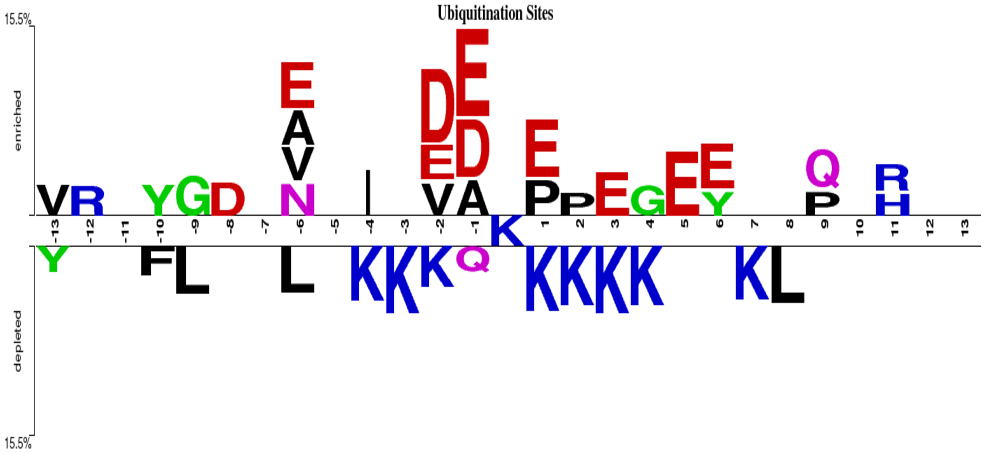


Figure 7: The amino acid propensities of surrounding Ubiquitination sites compared to non-Ubiquitination sites, as displayed with the Two Sample Logos software. It also shows that the position between the compositional amino acids of the ubiquitilyted and non- ubiquitilyted peptides had a wide difference, especially those located in the positions from ~ -13 to -1 and +1 to +13.

4. Conclusions

For class prediction in the independent dataset, we observed that RF predictor performed better than other predictors a feature selection method like as Wilcoxon signed rank test were carried out to identify the significant rules from the RF model, which helps to better understanding of the important rules that underlie the ubiquitylated proteins. It would help to decrease in the overall cost and time period for disease diagnosis and drug or vaccine discovery in protein ubiquitination sites. In future, we would like to pay more attention to develop an individual organism specific predictor for improving the performance of ubiquitination sites prediction.

References

[1] Breiman, L. (2001). Random forests. Machine Learning, 45:5-32.

[2] Burden, T. S. (1976). Succinylation of proteins associated with the ribosomal attachment site on microsomalmembranes. Hoppe-Seyler's Zeitschrift fur physiologische Chemie. 357(10): 1353–7.PMID: 992562.

[3] Cai Y., Huang T., Hu L., Shi X., Xie L., et al. (2011). Prediction of lysine ubiquitination with mRMR feature selection and analysis. Amino Acids.

[4] Catic, A., Collins, C., Church, G. M. and Ploegh, H. L. (2004). Preferred in vivo ubiquitination sites. Bioinformatics 20: 3302–3307.

[5] Centor, R. M. (1991). Signal detectability: the use of ROC curves and their analyses. Med Decis Making 11: 102–106.

[6] Chernorudskiy, A. L., Garcia, A., Eremin, E. V., Shorina, A. S., Kondratieva, E. V., et al. (2007). UbiProt: a database of ubiquitylated proteins. BMC Bioinformatics 8: 126.

[7] Gentry, M. S., Worby, C. A. and Dixon, J. E. (2005). Insights into lafora disease: Malin is an e3 ubiquitin ligase that ubiquitinates and promotes the degradation of laforin. Proc Natl AcadSci USA 102(24): 8501–8506

[8] Gribskov, M. and Robinson, N. L. (1996). Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. ComputChem 20: 25–33.

[9] Haglund, K. and Dikic, I.. (2005) Ubiquitylation and cell signaling. EMBO J 24: 3353–3359.

[10] Herrmann, J, Lerman, L. O. and Lerman, A. (2007). Ubiquitin and ubiquitin-like proteins in protein regulation. Circ Res, 100(9):1276-1291.

[11] Hershko, A. and Ciechanover, A. (1998). The ubiquitin system.Annu Rev Biochem 67: 425–479.

[12] Hicke, L. (2001). Protein regulation by monoubiquitin. Nat Rev Mol Cell Biol 2: 195–201.

[13] Hicke, L. (2001). Protein regulation by monoubiquitin. Nat Rev Mol Cell Biol 2(3):195–201.

[14] Hoeller, D., Hecker, C. M. and Dikic, I. (2006). Ubiquitin and ubiquitin-like proteins in cancer pathogenesis. Nat Rev Cancer 6(10): 776–788.

[15] Lee, W. C., Lee, M., Jung, J. W., Kim, K. P. and Kim, D. (2008). SCUD: Saccharomyces cerevisiae ubiquitination database. BMC Genomics 9: 440.

[16] Li, H., Xing, X., Ding, G., Li, Q., Wang, C., et al. (2009). SysPTM: a systematic resource for proteomic research on post-translational modifications. Mol Cell Proteomics 8: 1839–1849.

[17] Lin, D. H., Sterling, H., Wang, Z., Babilonia, E., Yang, B., Dong, K., Hebert, S. C., Giebisch, G. and Wang, W. H. (2005). Romk1 channel activity is regulated by monoubiquitination. Proc Natl AcadSci USA 102(12):4306–4311.

[18] Peng, J., Schwartz, D., Elias, J. E., Thoreen, C. C., Cheng, D., et al. (2003). A proteomics approach to understanding protein ubiquitination. Nat Biotechnol 21: 921–926.

[19] Pickart, C. M. (2001). Mechanisms underlying ubiquitination.Annu Rev Biochem 70:503–533.

[20] Pickart, C. M. (2001). Ubiquitin enters the new millennium. Mol Cell 8: 499–504.

[21] Radivojac, P., Vacic, V., Haynes, C., Cocklin, R. R., Mohan, A., et al. (2010). Identification, analysis, and prediction of protein ubiquitination sites. Proteins 78: 365–380.

[22] Reinstein, E. and Ciechanover, A. (2006). Narrative review: protein degradation and human diseases: the ubiquitin connection. Ann Intern Med 145(9):676–684.

[23] Sun, L. and Chen, Z. J. (2004). The novel functions of ubiquitination in signaling. CurrOpin Cell Biol 16(2):119–126.

[24] Tomlinson, E., Palaniyappan, N., Tooth, D. and Layfield, R. (2007). Methods for the purification of ubiquitinated proteins. Proteomics 7: 1016–1022.

[25] Tung, C. W. and Ho, S. Y. (2008). Computational identification of ubiquitylation sites from protein sequences. BMC Bioinformatics 9: 310.

[26] Welchman, R.L., Gordon, C. and Mayer, R. J. (2005). Ubiquitin and ubiquitin-like proteins as multifunctional signals. Nat Rev Mol Cell Biol, 6(8):599-609.