UbiSites-SRF: Ubiquitination Sites Prediction Using Statistical Moment with Random Forest Approach

Shazia Murad (Shazykhan57578@gmail.com)
Abdul Wali Khan University

Arwa Mashat
Faculty of Computing and Information Technology, P.O. Box 411, King Abdulaziz University, Rabigh 21911, Jeddah, Saudi Arabia

Alia Mahfooz
Faculty of Computing and Information Technology, P.O. Box 411, King Abdulaziz University, Rabigh 21911, Jeddah, Saudi Arabia

Sher Afzal Khan
Abdul Wali Khan University

Omar Barukab
Faculty of Computing and Information Technology, P.O. Box 411, King Abdulaziz University, Rabigh 21911, Jeddah, Saudi Arabia

Research Article

Keywords: Ubiquitination, Random forest, Statistical movement, Post-translational Modification (PTM), Hahn moments, position relative incidence matrix, 10-fold cross-validation, Jackknife testing

Posted Date: July 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-669582/v1

UbiSites-SRF: Ubiquitination Sites Prediction Using Statistical Moment with Random Forest Approach

[1]Omar BaruKab, [2,3]Shazia Murad, [1]Arwa Mashat, [1]Alia Mahfooz and [3]Sher Afzal Khan,

[1]Faculty of Computing and Information Technology, P.O. Box 411, King Abdulaziz University, Rabigh 21911, Jeddah, Saudi Arabia, obarukab@kau.edu.pk, aasmashat@kau.edu.sa, amalabdali@kau.edu.sa.

[2] Department of Physics, Women University Mardan, Shazykhan57578@gmail.com ,

[3] Department of Computer Sciences, Abdul Wali Khan University, Mardan, Pakistan,

Sher.Afzal@awkum.edu.pk.

Corresponding Author: Shazia Murad, Shazykhan57578@gmail.com.

Abstract

Ubiquitination is the process that supports the growth and development of eukaryotic and prokaryotic organisms. It is helpful in regulating numerous functions such as the cell division cycle, caspase-mediated cell death, maintenance of protein transcription, signal transduction, and restoration of DNA damage. Because of these properties, its identification is essential to understand its molecular mechanism. Some traditional methods such as mass spectrometry and site-directed mutagenesis are used for this purpose, but they are tedious and time consuming. In order to overcome such limitations, interest in computational models of this type of identification is therefore being developed. In this study, an accurate and efficient classification model for identifying ubiquitination sites was constructed. The proposed model uses statistical moments for feature extraction along with random forest for classification. Three sets of ubiquitination are used to train and test the model. The model is assessed through 10-fold cross-validation and jackknife tests. We achieved a 10-fold accuracy of 100% for dataset-1, 99.88% for dataset-2 and 99.84% for the dataset-3, while with Jackknife test we got 100% for the dataset-1, 99.91% for dataset-2 and 99.99%. for the dataset-3. The results obtained are almost the maximum, which is far better as compared to the pre-existing models available in the literature.

Keywords: Ubiquitination, Random forest, Statistical movement, Post-translational Modification (PTM), Hahn moments, position relative incidence matrix, 10-fold cross-validation, Jackknife testing.

1. Introduction

PTM is a process in which polypeptide chains of proteins arebeing modified to convert the immature protein to a mature and useful protein. This approach is used to regulate the function and physical structure of a protein, which helps in many processes inside a cell, such as different signaling passages or networks in cells, gene expression, activation and deactivation of enzymes, and protein-protein interaction (1). Any kind of irregularity in post-translational modification in

protein can lead to many types of diseases such as Alzheimer's, Fanconi anemia, Parkinson's, and Rosenthal fibers (astrocytes). It is therefore very important to examine and identify all such irregularities to early diagnose these diseases and to improve the process of drug development. There are various post-translational modification sites such as crotonylation (2), pseudo uridine (3), SUMOylation (4) (5)-(6), phosphorylation (7), nitro tyrosine (8), methylation (9)-(10), prenylation (11), methyl adenosine (12)-(13)-(14), ubiquitination (15)-(16)-(17).

Ubiquitination is the process, which assistances in the growing and development of eukaryotic and prokaryotic organisms, and different processes such as protein localization, metabolism function, regulation, and degradation. Similarly, it is helpful in the regulation of numerous functions for example cell division cycle, caspase-mediated cell death, maintenance of protein transcription, signal transduction, and DNA damage restoration (18). It is also documented, that any kind of abnormality in ubiquitination modification can result in several human diseases, for example, immunity neurodegenerative, muscular dystrophy, metabolic syndrome, and cancer (19). In the ubiquitination process, some enzymes take part which is Ubiquitin- Activating Enzyme (E1), Ubiquitin-Conjugation Enzyme (E2), and Ubiquitin- Ligase Enzyme (E3), where the E1 enzyme helps the ubiquitin enzyme to the end of carboxyl residue by making thioester bond. This process utilizes energy; therefore, it requires Adenosine triphosphate (ATPs). The E2 enzyme is to transfer the activated ubiquitin towards the cysteine residue of E2. The E3 enzyme works to transfer the ubiquitin from E2 to the targeted protein.

The ubiquitination activating enzyme (E1) acts on it in the presence of ATPs and in this way, ubiquitin is linked with the carboxyl end by forming a Thioester bond. The Ubiquitin-Conjugation Enzyme (E2) transfers the activated ubiquitin in the direction of the cysteine residue of E2 enzymes. The ubiquitin is linked with the E2 enzyme and we have the target protein which needs to be tagged with this ubiquitin-protein. Therefore, the last E3 enzyme called ubiquitination ligase comes and transfers the ubiquitin protein from E2 towards the lysine residue of the target protein by forming the iso-peptide bond. The ubiquitination of protein can be of three types that are mono ubiquitination, multi ubiquitination, and polyubiquitination. For the degradation of protein, the proteasomes are used, and the large protein is converted to short peptides chains. “Figure 1” (20) shows various functions of ubiquitination along with the mechanism.

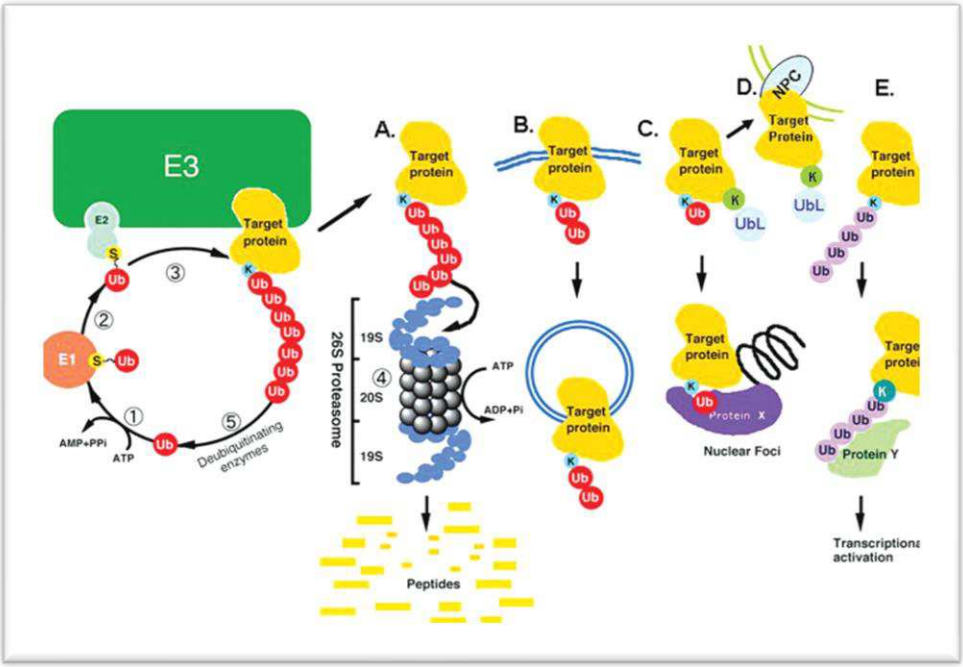

Figure 1. Ubiquitination mechanism and its various functions

2. Literature Review

Considering the numerous benefits of ubiquitination, it is very important to identify these sites to understand their features and make them valuable for the research area. This identification is not only help to understand its molecular mechanism but also provides worthwhile facts for additional studies of its drug development due to its critical regulatory role (21). Some traditional methods are used to identify these sites which are site-directed mutagenesis (22) and mass spectrometry (23). However, these methods are tedious, time-consuming and expensive, because the ubiquitination process is reversible, dynamic, and rapid (24). To design a fast and efficient process, the computational method is combined with biological methods and used to identify the ubiquitination sites (25)-(26). Tung and Ho (27) established the UbiPred a method to predict ubiquitination sites, here Support Vector Machine has utilized a classifier and the physicochemical properties were used selected them by using IPM, this model was helpful because it improves the progress of identification of ubiquitination sites and gives an improvement in accuracy from 72.19% to 84.44%. Another model was developed named UbPred (28) by the combination of amino acid components physicochemical properties, the accuracy achieved by UbPred was 72%. Cai et al. (29) used the KNN algorithm for prediction ubiquitination sites, they achieved greater MMC as compared to UbiPred and UbPred. Chen et al. (30) using the Support Vector Machine developed a model CKSAAP and the accuracy and MMC of this model have been measured as 73.40% and 0.4694, respectively. Using this model ubiquitination sites in Yeast are predicted. hCKSAAP_UbSite (31) is a prediction method used to identify ubiquitination giving an accuracy of 0.770. In 2013, Chen et al. (32) defined different

comparison methods for the identification of the ubiquitination sites of prokaryotes and eukaryotes.

Some researchers did ubiquitination prediction using different feature selection and extraction methods(33). Nguyen et al. (34) did research on protein sequence extraction and applied some methods like amino acid composition (AAC), amino acid pair composition (AAPC), and SVM and for, evaluation 10-fold cross-validation method. The Evolutionary screening algorithm (ESA) is developed by Wang et al. (26) to extract physicochemical properties of protein sequences are extracted. ESA-UbiSites was developed by using the SVM and the given result was 92%. Using the SVM algorithm, Lee et al. (35) developed the UbiSites model, for ubiquitination prediction(36). The LASSO was used as a feature extractor and the PseAAC was integrated with it and used three datasets for training and testing. The acquired accuracy for Dataset-1, Dataset-2, and Dataset-3 are 99%, 88.87%, 84.81%, respectively.

Although several research experiments have been carried out to precisely identify ubiquitination sites. To do this, they used various statistical approaches to feature extraction and machine learning models for training and identification. However, the predictors that are available are time consuming and so far, less efficient (27)- (29)-(30)-(32). With this in mind, we proposed a UbiSites SRF model that contains statistical moments for feature extractions and the random forest as classifiers for model training. In addition, the model is evaluated by 10 fold cross-validation (37), self-consistency (38), and the jackknife validation (39). The proposed model achieved the accuracy for Dataset-1 is 100%, Dataset-2 is 99.91%, and for Dataset-3 is 99.91 which are almost the maximum which is far better as compared to the already existing models, discussed in the literature.

The proposed system will identify ubiquitination sites in less time and with high accuracy. This will help the researchers to know the ubiquitination mechanism more accurately and in the drug development area.

3. Methods and materials

The methodology is implemented by applying the following 5 methods as illustrated in “Figure. 2”.

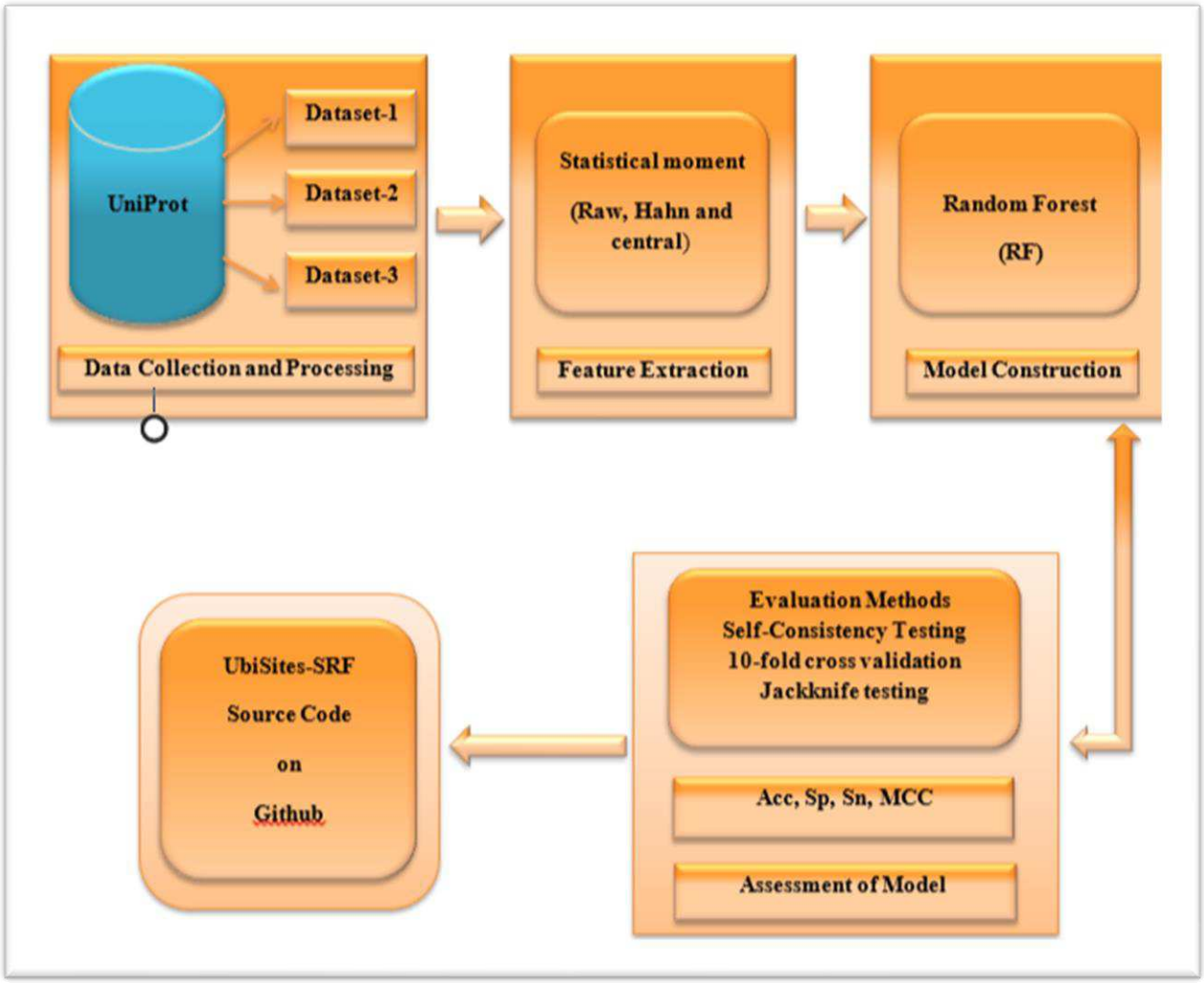

Figure 2 Flowchart of methodology

(1) Collection or construction of a valid benchmark dataset for training and testing (2) transformation of the biological sequences in mathematical form by using the statistical moment as feature extraction method (3) training and testing are performed by Random Forest classifier (4) performance evaluation of the model by 10-fold cross-validation and Jackknife testing. (5) GitHub access.

3.1. Benchmark dataset: To fairly evaluate the performance of the ubiquitination site predictive model and compare it with other existing models available in the literature, an objective and benchmark dataset must be selected. Objective and benchmark datasets play a critical role in evaluating the predictive model of ubiquitination sites, for the correct and valid decision. (18). Here for the prediction of the Ubiquitination Sites dataset is accessed from the UniProt was created by Cai and Jiang BMC Bioinformatics (2016) (40). It contains 3 records named Dataset-1, Dataset-2, and Dataset-3.

- In the Dataset-1, 300 sequences are downloaded from UniProt, which includes 150 positive sequences and the remaining 150 non-ubiquitination (31).
- The Dataset-2 contains 6838 sequences which are accessed from UniProt, where half of them are ubiquitination sites and the remaining half are the non-ubiquitination sites. The datasets are also used in (27).
- Similarly, Dataset-3 containing 12236 protein sequences collected from UniProt, in which half of them are ubiquitination sites and half are non-ubiquitination sites. The datasets are used by (41) for prediction.

3.2. Feature vector construction

Today, biological data and sequences are growing enormously. The most difficult task for us is to express the biological data and sequences into a vector or discrete form without dropping sequence pattern information and its characteristics. Vector formulation is a key task because all computational models including Random Forest (RF) [21] algorithm, K Nearest Neighbors Algorithm (KNN) [22], Support Vector Machine (SVM) (42) use the vector dataset for prediction. However, to avoid the problem of losing important information in this transformation from the biological form to numerical vector, many researchers have developed and applied the popular feature extraction method like Amino Acid Composition (AAC), the Pseudo Amino Acid Composition (PseAAC) (43)-(44) and Statistical moments (45)-(46) In this study, the proposed model has been developed by using statistical moment which consists of the raw moment, Hahn, central moments which is discussed in detail below.

3.2.1. Statistical Moment Calculation

The main purpose of the calculation of statistical moment is to identify the size of the dataset, the composition, and the position of protein sequences. This method is used to extract different features of the protein. Statisticians and mathematicians have suggested different types of the moment that are used to perform the statistical moment calculation by using different functions. (47)-(48). In this study, the raw, central (49) and Hahn moments are calculated (50). Further, a two-dimensional matrix of order n*n is developed based on the protein sequence P as expressed:

$$P = \{\acute{\alpha}_1, \acute{\alpha}_2 \ldots \acute{\alpha}_j\} \quad (1)$$

Where $\acute{\alpha}i$ is the i^{th} amino acid residue component in a primary sequence containing j residues, also let,

$$n = \lceil \sqrt{j} \rceil \quad (2)$$

$$\mathfrak{p} = \begin{bmatrix} \mathfrak{b}_{1_1}, & \mathfrak{b}_{1_2}, & \ldots & \mathfrak{b}_{1_n} \\ \mathfrak{b}_{2_1}, & \mathfrak{b}_{2_2}, & \ldots & \mathfrak{b}_{2_n} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathfrak{b}_{n_1}, & \mathfrak{b}_{n_2}, & \cdots & \mathfrak{b}_{n_n} \end{bmatrix} \quad (3)$$

The above matrix which is denoted by $\mathfrak{p}$ is a 2-dimensional matrix and it shows the primary structure of protein P. Another function called mapping function ω is used to convert the matrix P into $\mathfrak{p}$.

$$\omega(\acute{\alpha}_{\mathbf{m}}) = \mathcal{B}_{\mathbf{ij}} \quad (4)$$

The contents of the 2D matrix $\mathcal{P}$ are used for calculation of raw moment of degree 3 as given:

$$\mathbf{RM_{ij}} = \sum_{\mathbf{p=1}}^{\mathbf{n}} \sum_{\mathbf{q=1}}^{\mathbf{n}} \mathbf{p^i q^j} \mathcal{B}_{\mathbf{pq}} \quad (5)$$

Here i and j represent the order of moments. Moments up to order 3 areas RM_{00},RM_{01},RM_{10}, RM_{11},RM_{20}, RM_{21},RM_{30} and RM_{31}.

The midpoint from where data are distributed in every direction in terms of its weight average. It just acts as the center of gravity and it is very easy to calculate centroid moment after the calculation of raw moment. It is given as a point $\bar{k}, \bar{l}$ as defined:

$$\mathbf{k} = \mathbf{RM_{10}}/\boldsymbol{RM_{00}} \text{ and } \bar{\mathbf{l}} = \mathbf{RM_{01}}/\mathbf{RM_{00}} \quad (6)$$

The centroid is used to find the value of a central moment. Mathematically it can be computed by the following relation.

$$C_{ij} = \sum_{\mathbf{p=1}}^{\mathbf{n}} \sum_{\mathbf{q=1}}^{\mathbf{n}} (\mathbf{p} - \bar{\mathbf{k}})^{\mathbf{i}} (\mathbf{q} - \bar{\mathbf{l}})^{\mathbf{j}} \mathcal{B}_{\mathbf{pq}} \quad (7)$$

Furthermore, the Hahn polynomial order of n is given as

$$\mathbf{H_i^{y,z}(r, M)} = \mathbf{(M+V-1)_i (M-1)_i} \times \sum_{\mathbf{j=0}}^{\mathbf{i}} (-\mathbf{1})^{\mathbf{j}} \frac{\mathbf{(-i)_j (-r)_j (2M+y+z-i-1)_j}}{\mathbf{(M+z-1)_j (M-1)_j}} \frac{\mathbf{1}}{\mathbf{j!}} \quad (8)$$

The pochhammer symbol given in the equation above can be presented as

$$(\mathcal{b})_{\mathbf{j}} = \mathcal{b}.(\mathcal{b}+\mathbf{1}) \cdots (\mathcal{b}+\mathbf{j}-\mathbf{1}) \quad (9)$$

And is simplified using the Gamma operator.

$$(\mathcal{b})_{\mathbf{j}} = \frac{\boldsymbol{\Gamma}(\mathcal{b}+\mathbf{j})}{\boldsymbol{\Gamma}(\mathcal{b})} \quad (10)$$

To scale the raw values of Hahn moments are by using weighting function and square norm is given as $\mathbf{H_i^{y,z}(r, M)}=$

$$(\mathbf{H_i^{y,z}(r, M)} \sqrt{\frac{\mathbf{p(r)}}{\mathbf{d_i^2}}}, \mathbf{n = 0, 1, \cdots, M-1} \quad (11)$$

While

$$\mathbf{p(r)} = \frac{\boldsymbol{\Gamma}\mathbf{(y+r+z)(y+r+1)(y+z+r+1)_M}}{\mathbf{(y+z+2r+1)n!(M-r-1)!}} \quad (12)$$

The equation given below is used to calculate the Hahn moment for 2-dimensional discrete data.

$$\mathbf{h_{ij}} = \sum_{\mathbf{q=0}}^{\mathbf{N-1}} \sum_{\mathbf{p=0}}^{\mathbf{N-1}} \mathcal{B}_{\mathbf{ij}} \mathbf{H_i^{y,z}(q, M) h_j^{\widetilde{u,v}}(p, M)}, \quad \mathbf{m, n = 0, 1, \cdots M-1} \quad (13)$$

The raw, central, and Hahn moments are computed for each primary sequence up to order 3 and obtained a 2-dimensional matrix.

3.2.2. Determination of Position Relative Incidence Matrix (PRIM)

To quantize the relative position of the specific protein sequence is very much important (51). Therefore, the Position Relative Incidence Matrix (PRIM) is constructed and is a 20x20 matrix. Using this matrix, the features that are related to the relative position of amino acid residue are calculated.

$$\epsilon_{\mathbf{PRIM}} = \begin{bmatrix} \varepsilon_{1\to1}, & \varepsilon_{1\to2}, \cdots & \varepsilon_{1\to j}, \cdots & \varepsilon_{1\to20} \\ \varepsilon_{2\to1}, & \varepsilon_{2\to2}, \cdots & \varepsilon_{2\to j}, \cdots & \varepsilon_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_{i\to1}, & \varepsilon_{i\to2}, \cdots & \varepsilon_{i\to j}, \cdots & \varepsilon_{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_{n\to1}, & \varepsilon_{n\to2}, \cdots & \varepsilon_{n\to j}, \cdots & \varepsilon_{n\to20} \end{bmatrix} \quad (14)$$

In the above matrix, the summation of relative positions of j[th] residues concerning the first occurrence of i[th] residue is represented in each element$\varepsilon_{i\to j}$. The matrix produced 400 coefficients. For further reduction in the number of coefficients, statistical movements are computed and reduced to 30 coefficients.

3.2.3. Determination of Reverse Position Relative Incidence Matrix (RPRIM)

Some primary sequence of protein has incomprehensible unknown features. Our work is to detect these hidden attributes of protein with homologous ambiguities in protein sequence and for this purpose, we implemented Reverse Position Relative Incidence Matrix (RPRIM) which produces the same matrix (20x20 dimensions) of 400 coefficients as PRIM but for the reversed sequence (51). RPRIM can be mathematically defined as

$$''\Omega_{RPRIM} = \begin{bmatrix} \varrho_{1\to1}, & \varrho_{1\to2}, \cdots & \varrho_{1\to j}, \cdots & \varrho_{1\to20} \\ \varrho_{2\to1}, & \varrho_{2\to2}, \cdots & \varrho_{2\to j}, \cdots & \varrho_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \varrho_{i\to1}, & \varrho_{i\to2}, \cdots & \varrho_{i\to j}, \cdots & \varrho_{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \varrho_{n\to1}, & \varrho_{n\to2}, \cdots & \varrho_{n\to j}, \cdots & \varrho_{n\to20} \end{bmatrix} \quad (15)$$

To decrease the dimensionality of RPRIM, statistical moments are calculated for RPRIM which yields a set of 24 elements.

3.2.4. Generating Frequency Matrix

The total amount of amino acids in the sequences is represented by frequency and the mathematical principle which is used to precisely determine the frequency matrix distribution is shown as

$$\boldsymbol{\sigma = \vartheta_1, \vartheta_2, \vartheta_3, \ldots\ldots\ldots\ldots \vartheta_{20}} \quad (16)$$

Here ϑ_n is the frequency of occurrence of n[th] residue. The main purpose behind computing the frequency matrix is to extract compositional information of the sequence.

2.2.6. Generating Accumulative Absolute Position Incidence Vector (AAPIV)

To extract compositional information from ubiquitination sequences, the frequency matrix is calculated, but here the problem arises that it did not provide information on the relative positions of the residues, hence, the cumulative absolute impact vector of the position (AAPIV) of the length 20 elements is calculated. In AAPIV, the sum of all ordinal values for

each native amino acid, arising in the major sequence is a presence at their localities. AAPIV can be denoted as

$$\underline{K} = \boldsymbol{\varphi_1, \varphi_2, \varphi_3, \ldots \ldots \ldots \varphi_{20}} \quad (17)$$

From this, an arbitrary j^{th} element of AAPIV is computed as

$$\boldsymbol{\varphi_j} = \sum_{\mathbf{k=1}}^{\mathbf{n}} p_{\mathbf{k}} \quad (18)$$

2.2.7. Generating Reverse Accumulative Absolute Position Incidence Vector (RAAPIV)

One more step is to extract deep and incomprehensible information about the relative positions of residues in the sequence, to solve this problem, we used an accumulative reverse absolute position impact vector (RAAPIV) that is built by moving back the key sequence and production of the AAPIV from this reverse sequence. RAAPIV is symbolized as

$$\forall = \{\boldsymbol{\partial_1, \partial_2, \partial_3, \ldots \ldots \ldots \partial_{20}}\} \quad (19)$$

3.3 Prediction Model

Random Forest is one of the machine learning methods used for classification and regression under the category of supervised learning techniques. The theory used by random forest (52) is ensemble learning, in which several decision tree classifiers are combined for problem-solving to evaluate the performance of the model. Where each tree classifier collecting a random data sampled vector independently from the input dataset and performs independent training and prediction. After the training of the classifiers, the overall decision is taking place based on majority voting, which predicts the final output. As maximum the number of trees, the highest will be the accuracy, and the problem of overfitting will also be eliminated. The random forest works based on the following steps which are also shown in "Figure. 3":

- First-class random data points as a dataset.
- Make a model by integrating different decision trees according to that datasets.
- Repeats these steps again and again and make many decision trees.
- Now decide predictions by every tree.
- The final decision is based on the majority decision.

In the proposed study we have used the RF model for ubiquitination site predictions. Three datasets are passed through the feature extraction process for calculating the raw moment, Hahn, central moments are calculated. Other related terms for example PRIM, RPRIM, FV, AAPIV, and RAAPIV are also calculated for the datasets. All the information is saved in a vector called feature vector which includes the information about the composition and relative positions of the related datasets.

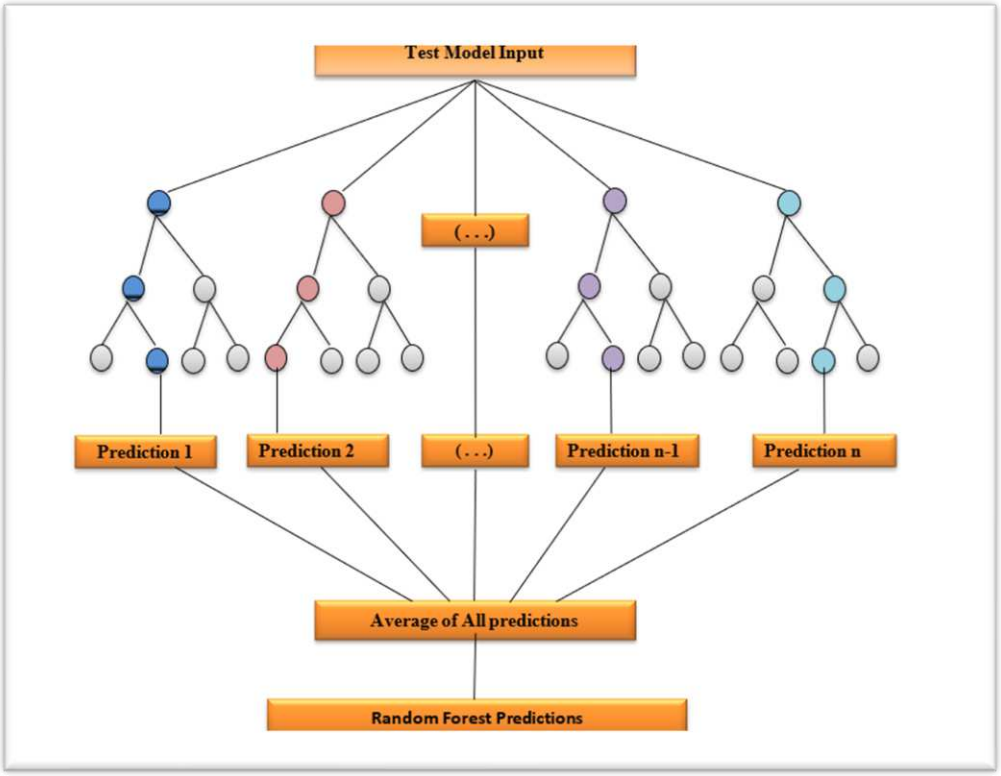

Figure 3. Architecture of Random Forest

4. Results and Discussions

4.1 Estimated accuracy

In this, we review the performance of the UbiSites-SRF, which is the most important part of our research because to decide about the correctness and reliability of our model. For this purpose, several matrices will be used to evaluate the performance of the model under various cross-validation tests.

4.2 Matrices Formation

Various evaluation methods are used to measure the machine learning models. Some measurement matrices that are us ed for the evaluation purpose are accuracy, sensitivity, specificity, and MCC (53)-(54)-(55). All these matrices are calculated to check the stability of the model. The mathematical representation of these matrixes is given below:

$$\boldsymbol{Sn} = \mathbf{1} - \frac{L_{-}^{+}}{L^{+}} \qquad 0 \leq \mathrm{Sn} \leq 1 \tag{8}$$

$$\boldsymbol{Sp} = \mathbf{1} - \frac{L_{+}^{-}}{L_{-}} \qquad 0 \leq \mathrm{Sn} \leq 1 \tag{9}$$

$$\boldsymbol{Acc} = \mathbf{1} - \frac{L_{-}^{+} + L_{+}^{-}}{L^{+} + L^{-}} \qquad 0 \leq \mathrm{Acc} \leq 1 \tag{10}$$

$$MCC = \frac{\left(1 - \frac{L_{-}^{+} + L_{+}^{-}}{L^{+} + L^{-}}\right)}{\sqrt{\left(1 + \frac{L_{+}^{-} + L_{-}^{+}}{L^{+}}\right)\left(1 - \frac{L_{-}^{+} + L_{+}^{-}}{L^{-}}\right)}} \quad -1 \leq Sn \leq 1 \tag{11}$$

L^{-} is the total accurately predicted negative samples and the L_{+}^{-} is the non-ubiquitination sites that are predicted incorrectly as ubiquitination. Similarly, L^{+} is set of positive ubiquitination dataset predicted positively and L_{-}^{+} show the positive samples of ubiquitination that are predicted as negative samples. All the matrices are explained through the formulas in the above equations (Equation 22, 23, 24, 25), at this point if the value of $L_{-}^{+} = 0$ all the positive samples are predicted correctly and thus sensitivity value will be Sn = 1. If $L_{-}^{+} = L^{+}$, it means all the positive samples are predicted as negative samples. Here the value of sensitivity will be zero. Moving forward, if we have $L^{+} = 0$, all the negative samples in the dataset are predicted accurately the specificity Sp = 1; on the other side if we have $L_{+}^{-} = L^{-}$, here the negative samples are predicted inaccurately as positive samples and the value of Sp = 0. If we have $L_{-}^{+} = L_{+}^{-}$, all the positive and negative samples which are present in the datasets are predicted accurately and here the MCC = 1 and Acc = 1; if we have $L_{-}^{+} = L^{+}$ and $L_{+}^{-} = L^{-}$, it means the model has detected all the positive samples a negative samples and negative samples as positive samples. Here the MCC = 1 and Acc = 0. All these measurement matrixes are used while evaluating the model by using different testing and evaluation methods.

4.3. Testing methods

In statistical prediction, many test methods are used to measure the power of the predictor. Some of the testing methods we used in this proposed model are (1) self-consistency tests, (2)10-fold cross-validation, and (3) Jackknife tests.

4.3.1 Testing via K-fold cross-validation

K-fold Cross-validation is a technique, used for the estimation of the prediction model. This method splits the dataset into k number of uniform folds, further, the model is trained by k-i folds and tests for the remaining i^{th} fold to estimate accuracy, sensitivity, specificity, and MCC, where i = 1, 2, …, k (56)-(46). Finally, all the outcomes of each fold are averaged to calculate the overall performance. In this study, the value of k is kept at 10 and the obtained results are shown in "Table 1" along the ROC curve in "Figure. 4", "Figure 5", and "Figure 6".

Predictor	**Accuracy Matrix**											
UbiSites-SRF	**Dataset-1**				**Dataset-2**				**Dataset-3**			
K-Fold Iterations	**Acc%**	**Sp**	**Sn**	**MCC**	**Acc%**	**Sp**	**Sn**	**MCC**	**Acc%**	**Sp**	**Sn**	**MCC**
1	100	1.0	1.0	1.0	100	1.0	1.0	1.0	99.92	1.0	1.0	1.0
2	100	1.0	1.0	1.0	99.71	0.99	1.0	0.99	99.84	1.0	1.0	1.0
3	100	1.0	1.0	1.0	99.85	1.0	1.0	1.0	100	1.0	1.0	1.0
4	100	1.0	1.0	1.0	99.85	1.0	1.0	1.0	99.92	1.0	1.0	1.0
5	100	1.0	1.0	1.0	99.85	1.0	1.0	1.0	100	1.0	1.0	1.0
6	100	1.0	1.0	1.0	99.71	1.0	1.0	0.99	100	1.0	1.0	1.0
7	100	1.0	1.0	1.0	99.85	1.0	1.0	1.0	100	1.0	1.0	1.0
8	100	1.0	1.0	1.0	100	1.0	1.0	1.0	99.92	1.0	1.0	1.0
9	100	1.0	1.0	1.0	100	1.0	1.0	1.0	100	1.0	1.0	1.0
10	100	1.0	1.0	1.0	100	1.0	1.0	1.0	99.84	1.0	1.0	1.0

	Final 10CV Score = 100.0	Final 10CV Score99.88	Final 10CV Score99.84

Table 1 10-fold Cross-Validation results for UbiSites-SRF

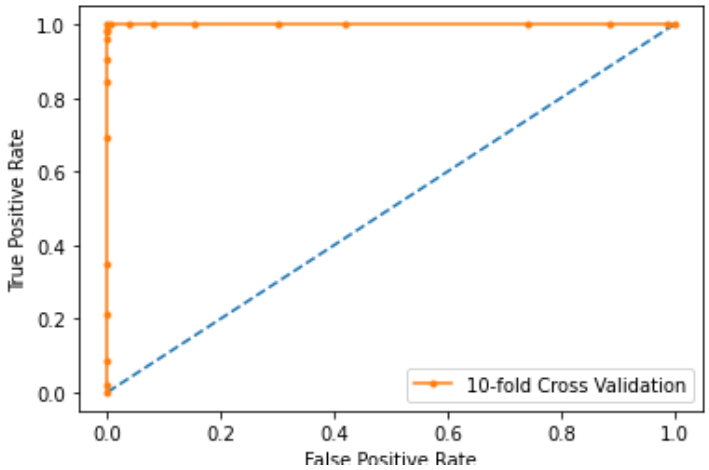

Figure 4. 10-Fold ROC Curve for Datase-1

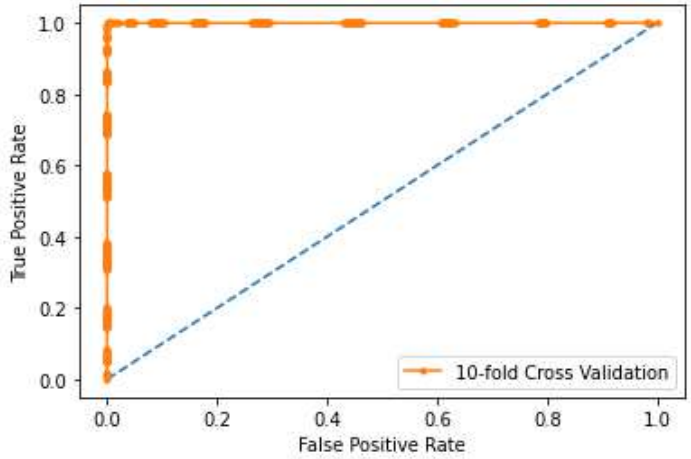

Figure 5 10-Fold ROC Curve for Dataset-2

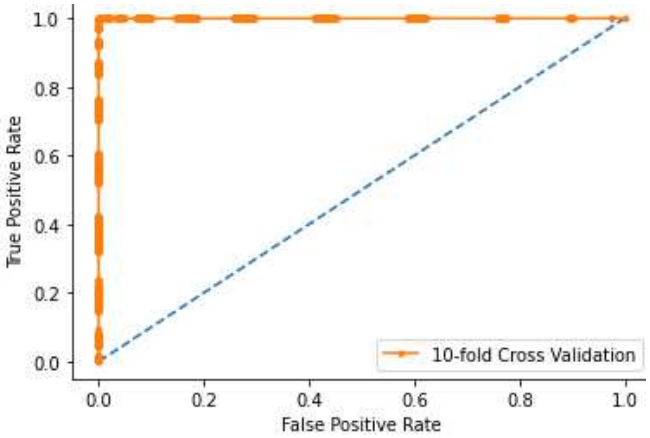

Figure 6. 10-Fold ROC Curve for Dataset-3

4.3.2. Self-consistency testing

This method is applied to know the accuracy of UbiSites-SRF, applied on all three datasets to train and test the model individually. Self-consistency testing is usually performed for the true positive (TP) values that are already known(48). The results of this testing are illustrated in “Table 2”, which shows the actual and predicted classification implemented by the proposed computational model. The ROC curve for the self-consistency testing is show in “Figure 7”, “Figure 8”, and “Figure 9” for Dataset-1, Dataset-2 and Dataset3, respectively.

Predictor	Accuracy Matrix											
UbiSites-SRF	Dataset-1				Dataset-2				Dataset-3			
	Acc%	Sp	Sn	MCC	Acc%	Sp	Sn	MCC	Acc%	Sp	Sn	MCC
	100	1.0	1.0	1.0	99.58	0.99	1.0	0.99	99.71	1.0	1.0	1.0

Table 2 Self-consistency testing results for UbiSites-SRF

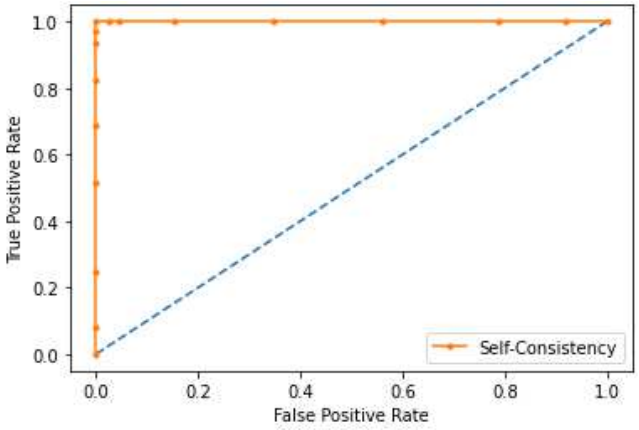

Figure 7. Self Consistency ROC Curve for Dataset-1

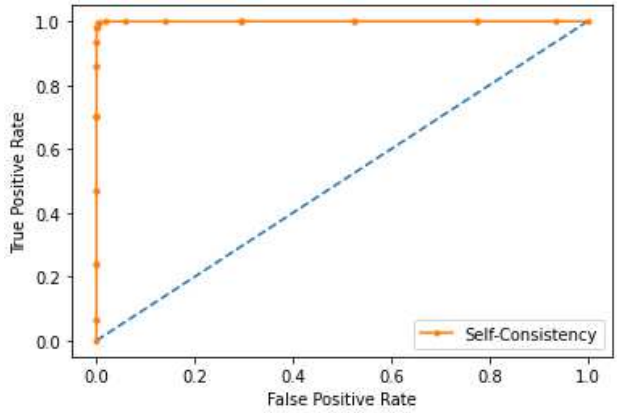

Figure 8. Self Consistency ROC Curve for Dataset-2

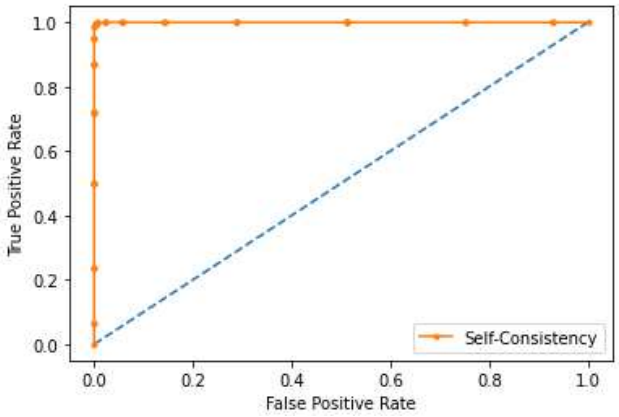

Figure 9. Self Consistency ROC Curve for Dataset-3

4.3.3. Jack-Knife testing

It is one of the cross-validation methods to estimate the performance of the machine learning model, it is a common resampling method that produces bias estimates and a standard error of an estimate by computing the estimate from subsamples of the existing sample (57). The method has some resemblances to the bootstrap method but can provide diverse bootstrap results in actual uses. [36]. In this proposed study, this technique is also used to calculate the UbiSites-SRF predictor and the accuracy for Dataset-1 is 100%, Dataset-2 is 99.91%, and for Dataset-3 is 99.91.

5. Comparison to Existing Models

The proposed model is compared with all existing models to assess its performance. UbiSites-SRF is compared to UbiSitePred (36) which is developed by using three datasets of ubiquitination and non-ubiquitination downloaded from UniProt. We used the same datasets for the training and testing of our model. The UbiSitePred achieved a dominant result as compared to the prediction model developed by Cai et al. (40) which used several computational methods such as Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Bayesian multivariate classifier (EBMC) along with Feature Selection NB (FSNB), Least Absolute Shrinkage and Select Operator (LASSO), Model Averaged NB (MANB), for the identification of physic-chemical properties of ubiquitination. “Table 3” is illustrated based on the accuracy of these predictors. UbiSites-SRF performs the prediction of ubiquitination sites based on various positions and composition variant features by using the statistical moment as a feature extraction method and Random Forest as a classifier. In this table, the accurate result for UbiSites-SRF is based on 10-fold cross-validation.

Dataset	**UbiSites-SRF**	UbiSitePred	EBMC	NB	FSNB	MANB	SVM	LR	LASSO
Set 1	**100**	99.98	67.14	52.89	56.13	55.45	65.97	72.44	69.33
Set 2	**99.88**	88.87	64.67	53.30	55.82	55.02	60.39	61.40	60.41
Set 3	**99.44**	84.81	66.67	51.41	56.33	51.92	61.02	64.76	61.29

Table 3 Comparative analysis of UbiSites-SRF with existing models

6. Conclusion

This proposed research has been conducted to identify ubiquitination sites that are helpful in various cell-related tasks such as transcriptional regulation, proteasomal degradation, and restoration of DNA damage. For this purpose, the statistical moment is used for the feature extraction to transform the biological data into the equivalent numerical vector and the Random forest classifier is used for further its classification and prediction. Finally, to measure the model performance various tests are used which are self-consistency, 10-fold cross-validation, and jackknife. As result, we achieved 100% accuracy for Dataset-1 identification using 10-fold cross-validation, self-consistency, and jackknife testing. For Dataset-2 the obtained accuracy is 99.88% through 10-fold cross-validation, 99.58% through self-consistency, and 99.91% through Jackknife testing. Finally, for Dataset-3 the model acquired the results of 99.84% through 10-fold cross-validation, 99.71% through self-consistency, and 99.91 for Jackknife testing. From the above, the UbiSites-SRF has the highest identification capability to identify ubiquitination as compared to all existing models available in the literature on the same datasets. The use of other popular predictors based on deep / machine learning methods along with statistical moments and their impact on identifying of other complex PTM sites is left as future work.

7. Acknowledgment:

This project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. IFPHI-360-830-2020. The authors, therefore, acknowledge with thanks to DSR for technical and financial support.

REFERENCES

1. Mann M, Jensen ON. Proteomic analysis of post-translational modifications [Internet]. Vol. 21, Nature Biotechnology. Nature Publishing Group; 2003 [cited 2021 Feb 2]. p. 255–61. Available from: https://www.nature.com/articles/nbt0303-255
2. Sabari BR, Tang Z, Huang H, Yong-Gonzalez V, Molina H, Kong HE, et al. Intracellular Crotonyl-CoA Stimulates Transcription through p300-Catalyzed Histone Crotonylation. Mol Cell. 2015 Apr 16;58(2):203–15.
3. Tahir M, Tayara H, Chong KT. iPseU-CNN: Identifying RNA Pseudouridine Sites Using Convolutional Neural Networks. Mol Ther - Nucleic Acids [Internet]. 2019;16(June):463–70. Available from: https://doi.org/10.1016/j.omtn.2019.03.010
4. Chang C-C, Tung C-H, Chen C-W, Tu C-H, Chu Y-W. SUMOgo: Prediction of sumoylation sites on lysines by motif screening models and the effects of various post-translational modifications OPEN. Available from: www.nature.com/scientificreports
5. Yavuz AS, Sezerman OU. Predicting sumoylation sites using support vector machines

based on various sequence features, conformational flexibility and disorder [Internet]. Available from: http://www.biomedcentral.com/1471-2164/15/S9/S18
6. Dehzangi A, López Y, Taherzadeh G, Sharma A, Tsunoda T. molecules SumSec: Accurate Prediction of Sumoylation Sites Using Predicted Secondary Structure. 2018; Available from: www.mdpi.com/journal/molecules
7. Zhao X, Zhang W, Xu X, Ma Z, Yin M. Prediction of Protein Phosphorylation Sites by Using the Composition of k-Spaced Amino Acid Pairs. PLoS One. 2012;7(10).
8. Jia C, Zhang M, Fan C, Li F, Song J. Formator: predicting lysine formylation sites based on the most distant undersampling and safe-level synthetic minority oversampling. IEEE/ACM Trans Comput Biol Bioinforma. 2019;PP(c):1–1.
9. Chiang PK, Gordon RK, Tal J, Zeng GC, Doctor BP, Pardhasaradhi K, et al. S-Adenosylmetliionine and methylation. FASEB J. 1996;10(4):471–80.
10. Maros ME, Capper D, Jones DTW, Hovestadt V, von Deimling A, Pfister SM, et al. Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data. Nat Protoc [Internet]. 2020 Feb 1 [cited 2021 Feb 2];15(2):479–512. Available from: https://www.nature.com/articles/s41596-019-0251-6
11. Xu Y, Wang Z, Li C, Chou K-C. iPreny-PseAAC: Identify C-terminal Cysteine Prenylation Sites in Proteins by Incorporating Two Tiers of Sequence Couplings into PseAAC. Med Chem (Los Angeles) [Internet]. 2017 Apr 20 [cited 2021 Jan 5];13(6). Available from: https://pubmed.ncbi.nlm.nih.gov/28425870/
12. Beemon K, Keith J. Localization of N6-methyladenosine in the Rous sarcoma virus genome. J Mol Biol. 1977 Jun 15;113(1):165–79.
13. Ji P, Wang X, Xie N, Li Y. N6-methyladenosine in RNA and DNA: An epitranscriptomic and epigenetic player implicated in determination of stem cell fate. Vol. 2018, Stem Cells International. Hindawi Limited; 2018. p. 3256524.
14. Huang H, Weng H, Sun W, Qin X, Shi H, Wu H, et al. Recognition of RNA N 6 -methyladenosine by IGF2BP proteins enhances mRNA stability and translation. Nat Cell Biol. 2018 Mar 1;20(3):285–95.
15. Finley D, Chau V. Ubiquitination. Annu Rev Cell Biol [Internet]. 1991 Nov 28 [cited 2021 Feb 2];7(1):25–69. Available from: http://www.annualreviews.org/doi/10.1146/annurev.cb.07.110191.000325
16. Herrmann J, Lerman LO, Lerman A. Ubiquitin and ubiquitin-like proteins in protein regulation. Circ Res. 2007;100(9):1276–91.
17. Hershko A, Ciechanover A. The ubiquitin system. Vol. 67, Annual Review of Biochemistry. 1998. p. 425–79.
18. Haglund K, Dikic I. Ubiquitylation and cell signaling [Internet]. Vol. 24, EMBO Journal. EMBO J; 2005 [cited 2021 Jan 5]. p. 3353–9. Available from: https://pubmed.ncbi.nlm.nih.gov/16148945/
19. Pickart CM. Mechanisms underlying ubiquitination. Vol. 70, Annual Review of Biochemistry. 2001. p. 503–33.
20. Ciechanover A, Iwai K. The ubiquitin system: From basic mechanisms to the patient bed. IUBMB Life. 2004;56(4):193–201.
21. Chen Z, Zhou Y, Zhang Z, Song J. Towards more accurate prediction of ubiquitination sites: A comprehensive review of current methods, tools and features. Brief Bioinform. 2014 Jul 10;16(4):640–57.

22. Gentry MS, Worby CA, Dixon JE. Insights into Lafora disease: Malin is an E3 ubiquitin ligase that ubiquitinates and promotes the degradation of laforin. Proc Natl Acad Sci U S A [Internet]. 2005 Jun 14 [cited 2021 Jan 5];102(24):8501–6. Available from: www.pnas.orgcgidoi10.1073pnas.0503285102
23. de Hoffmann E. Mass Spectrometry. In: Kirk-Othmer Encyclopedia of Chemical Technology [Internet]. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2005 [cited 2021 Feb 2]. Available from: http://doi.wiley.com/10.1002/0471238961.1301191913151518.a01.pub2
24. Kannicht C, Fuchs B. Post-translational Modifications of proteins. Mol Biomethods Handb Second Ed. 2008;427–49.
25. Qiu WR, Xiao X, Lin WZ, Chou KC. IUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. J Biomol Struct Dyn [Internet]. 2015 Aug 3 [cited 2021 Jan 5];33(8):1731–42. Available from: https://pubmed.ncbi.nlm.nih.gov/25248923/
26. Wang J-R, Huang W-L, Tsai M-J, Hsu K-T, Huang H-L, Ho S-Y. ESA-UbiSite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives. Available from: https://academic.oup.com/bioinformatics/article/33/5/661/2849433
27. Tung CW, Ho SY. Computational identification of ubiquitylation sites from protein sequences. BMC Bioinformatics [Internet]. 2008 Jul 15 [cited 2021 Jan 5];9(1):1–15. Available from: https://link.springer.com/articles/10.1186/1471-2105-9-310
28. Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, Heyen JW, et al. Identification, analysis, and prediction of protein ubiquitination sites. Proteins Struct Funct Bioinforma. 2010;78(2):365–80.
29. Qiu WR, Xiao X, Lin WZ, Chou KC. IUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. J Biomol Struct Dyn [Internet]. 2015 Aug 3 [cited 2021 Feb 2];33(8):1731–42. Available from: https://www.tandfonline.com/doi/abs/10.1080/07391102.2014.968875
30. Wang XB, Wu LY, Wang YC, Deng NY. Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs. Protein Eng Des Sel. 2009;22(11):707–12.
31. Chen Z, Zhou Y, Song J, Zhang Z. HCKSAAP-UbSite: Improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. Biochim Biophys Acta - Proteins Proteomics. 2013 Aug 1;1834(8):1461–7.
32. Ju Z, Gu H. Predicting pupylation sites in prokaryotic proteins using semi-supervised self-training support vector machine algorithm. Anal Biochem [Internet]. 2016;507:1–6. Available from: http://dx.doi.org/10.1016/j.ab.2016.05.005
33. Nguyen VN, Huang KY, Huang CH, Lai KR, Lee TY. A New Scheme to Characterize and Identify Protein Ubiquitination Sites. IEEE/ACM Trans Comput Biol Bioinforma. 2017 Mar 1;14(2):393–403.
34. Chen Z, Zhou Y, Zhang Z, Song J. Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features. Available from: http://protein.cau.
35. Lee T-Y, Chen S-A, Hung H-Y, Ou Y-Y, Uversky V. Incorporating Distant Sequence Features and Radial Basis Function Networks to Identify Ubiquitin Conjugation Sites. 2011; Available from: www.plosone.org
36. Cui X, Yu Z, Yu B, Wang M, Tian B, Ma Q. UbiSitePred: A novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal

Chou's pseudo components. Chemom Intell Lab Syst [Internet]. 2019;184:28–43. Available from: https://doi.org/10.1016/j.chemolab.2018.11.012
37. OpenML [Internet]. [cited 2021 Feb 2]. Available from: https://www.openml.org/a/estimation-procedures/7
38. Self-consistency test of the algorithm. Fraction of references from the... | Download Scientific Diagram [Internet]. [cited 2021 Feb 2]. Available from: https://www.researchgate.net/figure/Self-consistency-test-of-the-algorithm-Fraction-of-references-from-the-stem-cell_fig1_7946314
39. Jackknife Test - an overview | ScienceDirect Topics [Internet]. [cited 2021 Jan 26]. Available from: https://www.sciencedirect.com/topics/nursing-and-health-professions/jackknife-test
40. Cai B, Jiang X. Computational methods for ubiquitination site prediction using physicochemical properties of protein sequences. BMC Bioinformatics [Internet]. 2016;17(1):1–12. Available from: http://dx.doi.org/10.1186/s12859-016-0959-z
41. Cai Y, Huang T, Hu L, Shi X, Xie L, Li Y. Prediction of lysine ubiquitination with mRMR feature selection and analysis. Amino Acids [Internet]. 2012 Apr 26 [cited 2021 Feb 2];42(4):1387–95. Available from: https://link.springer.com/article/10.1007/s00726-011-0835-0
42. An Introduction to Support Vector Machines (SVM) [Internet]. [cited 2021 Feb 1]. Available from: https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/
43. Limongelli I, Marini S, Bellazzi R. PaPI: Pseudo amino acid composition to score human protein-coding variants. BMC Bioinformatics [Internet]. 2015 Dec 12 [cited 2021 Feb 1];16(1):123. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0554-8
44. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins Struct Funct Genet. 2001;43(3):246–55.
45. statistical moment - Google Search [Internet]. [cited 2021 Feb 2]. Available from: https://www.google.com/search?q=statistical+moment&oq=statistical+moment&aqs=chrome..69i57j0l5j0i20i263i395j69i60.5630j1j7&sourceid=chrome&ie=UTF-8
46. Technology I, Aziz KA, Arabia S. PREDICTION OF SAUDI ARABIA SARS-COV 2 DIVERSIFICATIONS IN PROTEIN STRAIN AGAINST CHINA STRAIN. 2020;8(1):64–73.
47. Ju Z, Wang SY. Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition. Gene [Internet]. 2018 Jul 20 [cited 2021 Feb 2];664:78–83. Available from: https://pubmed.ncbi.nlm.nih.gov/29694908/
48. Khan YD, Rasool N, Hussain W, Khan SA, Chou KC. iPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC. Anal Biochem. 2018 Jun 1;550:109–16.
49. Gerig G. Lecture: Shape Analysis Moment Invariants. Cs 7960 [Internet]. 2010;38. Available from: http://www.sci.utah.edu/~gerig/CS7960-S2010/handouts/CS7960-AdvImProc-MomentInvariants.pdf
50. Khan YD, Khan SA, Ahmad F, Islam S. Iris recognition using image moments and k-Means algorithm. Sci World J. 2014;2014.
51. Akmal MA, Rasool N, Daanial Khan Y. Prediction of N-linked glycosylation sites using

position relative features and statistical moments. 2017; Available from: https://doi.org/10.1371/journal.pone.0181966
52. Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell [Internet]. 1998 [cited 2021 Feb 1];20(8):832–44. Available from: http://ect.bell-labs.com/who/tkh/publications/papers/df.pdf
53. Evaluation Metrics Definition | DeepAI [Internet]. [cited 2021 Feb 2]. Available from: https://deepai.org/machine-learning-glossary-and-terms/evaluation-metrics
54. Machine Learning Classifier: Basics and Evaluation | by James Le | Data Notes | Medium [Internet]. [cited 2021 Feb 2]. Available from: https://medium.com/cracking-the-data-science-interview/machine-learning-classifier-basics-and-evaluation-44dd760fea50
55. Evaluation Metrics Machine Learning [Internet]. [cited 2021 Feb 2]. Available from: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/
56. What is Cross Validation in Machine learning? Types of Cross Validation [Internet]. [cited 2021 Feb 11]. Available from: https://www.mygreatlearning.com/blog/cross-validation/
57. Jackknife - an overview | ScienceDirect Topics [Internet]. [cited 2021 Feb 2]. Available from: https://www.sciencedirect.com/topics/mathematics/jackknife

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ClassificationFinalset1.ipynb
- Classificationfinalset2.ipynb
- ClassificationFinalset3.ipynb
- dataset1.txt
- dataset2.txt
- dataset3.txt