Send Orders for Reprints to reprints@benthamscience.net

Current Genomics, 2019, *20*, 389-399

RESEARCH ARTICLE

Computational Prediction of Ubiquitination Proteins Using Evolutionary Profiles and Functional Domain Annotation

Wangren Qiu[1,#], Chunhui Xu[2,#], Xuan Xiao[1] and Dong Xu[2,3,*]

[1]*Computer Department, Jingdezhen Ceramic Institute, Jingdezhen 333046, China;* [2]*Informatics Institute, University of Missouri, Columbia, MO 65201, USA;* [3]*Department of Electrical Engineering and Computer Science, Bond Life Sciences Center, University of Missouri, Columbia, MO 65201, USA*

ARTICLE HISTORY

Received: April 26, 2019
Revised: July 14, 2019
Accepted: August 29, 2019

DOI: 10.2174/1389202919666191014091250

Abstract: ***Background*:** Ubiquitination, as a post-translational modification, is a crucial biological process in cell signaling, apoptosis, and localization. Identification of ubiquitination proteins is of fundamental importance for understanding the molecular mechanisms in biological systems and diseases. Although high-throughput experimental studies using mass spectrometry have identified many ubiquitination proteins and ubiquitination sites, the vast majority of ubiquitination proteins remain undiscovered, even in well-studied model organisms.

***Objective*:** To reduce experimental costs, computational methods have been introduced to predict ubiquitination sites, but the accuracy is unsatisfactory. If it can be predicted whether a protein can be ubiquitinated or not, it will help in predicting ubiquitination sites. However, all the computational methods so far can only predict ubiquitination sites.

***Methods*:** In this study, the first computational method for predicting ubiquitination proteins without relying on ubiquitination site prediction has been developed. The method extracts features from sequence conservation information through a grey system model, as well as functional domain annotation and subcellular localization.

***Results*:** Together with the feature analysis and application of the relief feature selection algorithm, the results of 5-fold cross-validation on three datasets achieved a high accuracy of 90.13%, with Matthew's correlation coefficient of 80.34%. The predicted results on an independent test data achieved 87.71% as accuracy and 75.43% of Matthew's correlation coefficient, better than the prediction from the best ubiquitination site prediction tool available.

***Conclusion*:** Our study may guide experimental design and provide useful insights for studying the mechanisms and modulation of ubiquitination pathways. The code is available at: https://github.com/Chunhuixu/UBIPredic_QWRCHX

Keywords: Ubiquitination, machine learning, random forest, protein annotation, subcellular localization, functional domain.

1. INTRODUCTION

As a well-known post-translational modification (PTM), ubiquitination is crucial in proteome dynamics and various signaling pathways in the cells [1]. Ubiquitination is an enzymatic PTM in which ubiquitin (a small regulatory protein) [2] is attached to a lysine residue of the targeting protein [3]. Ubiquitination marks proteins for degradation through the proteasome [4], alters their cellular location [5], and regulates protein interactions [5]. It is involved in signal transduction [6], apoptosis, endocytosis, gene transcription, DNA repair, and replication, intracellular trafficking, virus budding [3], cellular transformation, immune response, and inflammatory response [7].

Due to its importance and complexity, the identification of ubiquitination proteins and ubiquitination sites is highly valuable. However, experimental identification is time-consuming and expensive [8] particularly because the ubiquitination process is dynamic, rapid and reversible [9-11]. Hence, computational predictions become an important and practical alternative. A number of computational methods were developed based on the traditional machine learning method for predicting lysine ubiquitination sites, including Radivojac's UbPred [12], Cai's mRMR model [13], Zhao's ensemble classifier [14], and Chen's CKSAAP approach [15], but they can only identify ubiquitination sites with limited accuracies. It is not practical to predict ubiquitination proteins through these methods since the false-positive rates would be too high to be useful. Recently, deep learning

*Address correspondence to this author at the Department of Electrical Engineering and Computer Science, Bond Life Sciences Center, University of Missouri, Columbia, MO 65201, USA; Tel: 573 - 882 - 2299;
E-mail: xudong@missouri.edu

[#]*These authors contributed equally to this work.*

1389-2029/19 $58.00+.00

method tools, such as MusiteDeep-Capsule [16] have become competitive due to the advancing of computing resources, our results show good performance even when compared with it, also, it has a better performance than three machine learning-based tools, UbiProber [17], UbiSite [18] and PDM-PUB [19]. To the best of our knowledge, so far no computational method has been developed to predict whether an uncharacterized protein is able to be ubiquitinated or not. The present study was initiated in an attempt to address this problem for the first time. If it can be predicted whether an uncharacterized protein can be ubiquitinated or not, it will be valuable for the prediction itself and helpful for identifying ubiquitination sites. Although this problem has not been addressed, some explorations have been made on a related study, *i.e.* to predict whether a protein can be phosphorylated or not [20, 21]. A method has been presented for identifying human phosphorylated proteins by incorporating evolutionary information into a general pseudo amino acid composition (PseAAC) model through a grey system [21-23]. It is believed that the formulation and approach can be also used to predict ubiquitination proteins. Ubiquitination is much less frequent than phosphorylation, as ubiquitination is a much more "expensive" biological operation than phosphorylation. The smaller group of ubiquitination proteins may have even stronger common features than phosphorylation proteins so that a whole sequence-based method may work better in predicting ubiquitination proteins. Furthermore, other gene features will be used, such as Gene Ontology (GO) [24], a structured repository of concepts (GO Terms) related to gene functions for the prediction, none of which was used in predicting phosphorylation sites.

In this study, a novel computational method, was developed to predict ubiquitination proteins for a query amino acid sequence on the basis of its evolutionary information through a grey system model [25] and K Nearest Neighbour (KNN) scores calculated with the fuzzy distance by using its Functional Domain Annotation (FDA) and subcellular localization. There are two major feature sets in this study: one set includes 80 sequence grey model features extracted from the sequence evolution information and another contains the features calculated by KNN scores based on FDAs. To thoroughly evaluate the proposed model, it was trained and tested with different datasets and cross-validations methods. In addition, the distribution of the above-mentioned features in predicted ubiquitination proteins was analyzed and it provided some hypotheses for distinguishing ubiquitination proteins from non-ubiquitination ones.

2. MATERIALS AND METHODS

2.1. Benchmark Dataset

The dataset used was extracted from Uniprot at http://www.ebi.ac.uk/uniprot [26]. The version of protein data used in the current study was released on May 2017. The positive dataset containing 1906 known ubiquitination proteins was generated through the following queries in the UniProt advanced search: "annotation: (type:crosslnk ubiquitin) length: [50 TO *] AND reviewed: yes." Three hundred of 1906 proteins were separated as an independent test dataset so that the remaining 1606 positive proteins were kept in training and validation. For the negative dataset, we started from all reviewed proteins (~550,000 totals) and performed a filtering process by using CD-HIT-2D [27] with a threshold of 70%, after this step, there were 320,096 negative proteins left. To conduct balanced training, these samples were randomly taken to form three negative datasets, in which the number of samples was the same as the given positive dataset. At this time, a 300 proteins negative independent dataset was randomly selected and isolated for testing. There is no overlap between training and testing datasets. For the annotation information, we extracted the 8 types UniProt annotations of 'Subcellular localization (SL) [28]' and FDAs of 'GO [29]' , 'Pfam [30]', 'Smart [31]', 'PROSITE [32]', 'SUPFAM [33]', 'InterPro [34]', and 'PRINTS [35]' for all the proteins in the datasets. SL was reorganized by the UniProt build-in hierarchical subcellular localization table.

2.2. Incorporate Extracted Features into the General Pseudo amino Acid Composition

It is known that most traditional machine-learning algorithms, such as Neural Network [36], Covariant Discriminant [37], Support Vector Machine [38], K Nearest Neighbor [39], and Random Forest [40], can only handle vector but not sequence samples. To formulate a biological sequence of a variable length into a discrete model or a vector, yet still considerably keep its sequence pattern or inherent characteristics, researchers formulated the protein sequence or peptides using pseudo amino acid composition (PseAAC) [21], encoding method [41] or other approaches [42]. Here, a model following the general form of PseAAC [43] has been proposed, which formulates a protein $\mathbf{P}$ as (Eq. 1):

$$\mathbf{P} = [\mathbf{P_1}\ \mathbf{P_2}\ \cdots\ \mathbf{P_u}\ \cdots\ \mathbf{P_\Omega}]^{\mathbf{T}} \quad (1)$$

where T is a transpose operator, the subscript $\boldsymbol{\Omega}$ is an integer, and its value as well as the components $\mathbf{P_1}$, $\mathbf{P_2}$, … depend on the extraction of the desired information from the amino acid sequence of $\mathbf{P}$ as described below.

2.3. Vectorization of Sequence Profile through a Grey System Model

From the evolutionary viewpoint, all the protein sequences have been evolved from a very limited number of ancestral species. Their evolution involves mutations of single residues, as well as insertions and deletions of residues, gene duplication, and gene fusion. With these changes accumulated for a long period of time, many similarities between the original and evolved amino acid sequences have gradually disappeared, but they may still share some common features, such as belonging to the same type of protein [44], residing in a same subcellular location [45], or having a similar biological function [46]. It is assumed that ubiquitination proteins have evolutionary relationships that are reflected in some common attributes encoded in sequence profiles, *i.e.* the Position Specific Scoring Matrix (PSSM), as described below. The sequence profile by a $L\times 20$ matrix as $\mathbf{P}$ is given as:

$$\mathbf{P}_{\mathbf{PSSM}}^{(\mathbf{0})} = \begin{bmatrix} m_{1,1}^{(0)} & m_{1,2}^{(0)} & \cdots & m_{1,20}^{(0)} \\ m_{2,1}^{(0)} & m_{2,2}^{(0)} & \cdots & m_{2,20}^{(0)} \\ \vdots & \vdots & \vdots & \vdots \\ m_{L,1}^{(0)} & m_{L,2}^{(0)} & \cdots & m_{L,20}^{(0)} \end{bmatrix} \quad (2)$$

where L is the number of residues of the protein, $m_{i,j}^{(0)}$ represents the profile score in the i-th ($i = 1,2,\cdots,L$) position of the protein having amino acid type j ($j = 1,2,\cdots,20$) during the evolution. Here, the numerical codes 1, 2, …, 20 are used to denote the 20 natural amino acid types according to the alphabetical order of their single-character codes. The $L\times20$ scores in Eq. 2 were generated by using PSI-BLAST [47] to search the UniProtKB/Swiss-Prot database through three iterations with 0.001 as the E-value cutoff for protein **P**. Following the same approach as done by Lin *et al.* [48], PSSM was encoded to a fixed-size vector with a total of $\Omega = 3\times20 = 60$ quantities by using the first-order grey model, or a total of $4\times20 = 80$ quantities by the second-order model. Thus, the components of a given protein sample can be represented by Eq. 3 or Eq. 3' (which is called SeqEvo Descriptor), for the first-order and the second-order models, respectively.

$$\boldsymbol{P}_{PSSM-Grey}^{(60)} = [a^1, b_1^1, b_2^1, a^2, b_1^2, b_2^2, \cdots, a^{20}, b_1^{20}, b_2^{20}]^{\mathbf{T}} \quad (3)$$

$$\boldsymbol{P}_{PSSM-Grey}^{(80)} = [a^{1\prime}, b_1^{1\prime}, b_2^{1\prime}, b_3^{1\prime}, a^{2\prime}, b_1^{2\prime}, b_2^{2\prime}, b_3^{2\prime}, \cdots, a^{20\prime}, b_1^{20\prime}, b_2^{20\prime}, b_3^{20\prime}]^{\mathbf{T}} \quad (3')$$

where $\boldsymbol{a^j}, \boldsymbol{b_1^j}, \boldsymbol{b_2^j}$ are the parameters of the first-order model for the $\boldsymbol{j}$th amino acid ($\boldsymbol{j = 1,2,\cdots 20}$); $\boldsymbol{a^{j\prime}}, \boldsymbol{b_1^{j\prime}}, \boldsymbol{b_2^{j\prime}}, \boldsymbol{b_3^{j\prime}}$ are the parameters of the second-order model for the $\boldsymbol{j}$th amino acid ($\boldsymbol{j = 1,2,\cdots 20}$).

2.4. KNN Score based on FDA and Subcellular Localization

A set of complementary FDAs and subcellular localization were used as features of ubiquitination proteins, including (1) UniProt annotations of subcellular localization and FDAs of 'GO', 'Pfam', 'Smart', 'PROSITE', 'SUPFAM', 'InterPro', and 'PRINTS', as well as GO annotations with categories of molecular function, biological process, and cellular component; (2) Pfam, a large collection of protein families generated using hidden Markov models; (3) Smart [31], a collection of protein domains and domain architectures; (4) PROSITE [32], a database of protein families, domains and functional sites; (5) SUPFAM, a database of protein structural and functional annotation; (6) InterPro, a resource of protein families, domains and important sites; (7) PRINTS, a collection of sequence "fingerprints" and protein families. Studies have shown that ubiquitination proteins often share the same subcellular localization [49]. Hence, subcellular localizations were also used as a feature for predicting ubiquitination proteins. These features were used based on a KNN algorithm as follows:

Step 1. For a query protein sequence, find its $\boldsymbol{k}$ nearest neighbors in the whole set, including positive and negative samples, according to local sequence similarity. For proteins $\boldsymbol{p}$ and $\boldsymbol{q}$, let:

$$\mathrm{FDA}_j(\boldsymbol{p}) = \{N_1^{p,j}, N_2^{p,j}, \cdots, N_{n_p}^{p,j}\}$$

$$\mathbf{FDA}_j(\boldsymbol{q}) = \{N_1^{q,j}, N_2^{q,j}, \cdots, N_{n_q}^{q,j}\}$$

Which represents the j-th feature of FDA of $\boldsymbol{p}$ and $\boldsymbol{q}$, respectively. j =1,2,…,7,8 represents 'GO', 'Pfam', 'Smart', 'PROSITE', 'SUPFAM', 'InterPro', 'PRINTS' or 'subcellular localization', respectively, and the distance $\mathbf{Dist}_j(\boldsymbol{p},\boldsymbol{q})$ between $\boldsymbol{p}$ and $\boldsymbol{q}$ is defined as follows in Eq. (4):

$$\mathrm{Dist}_j(\boldsymbol{p},\boldsymbol{q}) = 1 - \frac{\|\mathrm{FDA}_{\boldsymbol{p}}(\mathrm{j}) \cap \mathrm{FDA}_{\boldsymbol{q}}(\mathrm{j})\|}{\|\mathrm{FDA}_{\boldsymbol{p}}(\mathrm{j}) \cup \mathrm{FDA}_{\boldsymbol{q}}(\mathrm{j})\|} \quad (4)$$

where $\cup$ and $\cap$ represent the "union" and "intersection" in the set theory, and $\|\ \ \|$ is the operator acting on the set therein to count the number of its elements.

Step 2. A corresponding KNN feature is then extracted by calculating the KNN score, represented by the percentage of positive neighbors (ubiquitination proteins) in its $\boldsymbol{k}$ nearest neighbors.

Step 3. To take advantage of different properties of neighbors with various similarity cutoffs, Steps 1 and 2 were repeated for different $\boldsymbol{k}$ values to obtain multiple features for the ubiquitination protein predictor. In this study, based on empirical trials, by default, $\boldsymbol{k}$ was chosen to be 0.1%, 0.4%, 0.7 %, …, 14.5% and 14.8%; then the number of features is 50, *i.e.* 50 KNN scores were extracted as features for predicting ubiquitination proteins. For the j-th member of FDA, the protein **P** can be formulated as (Eq. 5):

$$\boldsymbol{P}_{\mathbf{FDA}_j} = [\boldsymbol{\varphi_1(j)}, \boldsymbol{\varphi_2(j)}, \cdots, \boldsymbol{\varphi_K(j)}\ ]^{\boldsymbol{T}} \quad (5)$$

where $\boldsymbol{\varphi_1(j)}, \boldsymbol{\varphi_2(j)}\ldots\ \boldsymbol{\varphi_{50}(j)}$ are the ratios of positive neighbors to the whole samples at 0.1%, 0.4%...14.8% of the training data set size, respectively. Hence, a query protein sequence can be formulated with seven 50-dimension vectors, *i.e.*, $\boldsymbol{P}_{\mathbf{FDA}} = [\boldsymbol{P}_{\mathbf{FDA_1}}, \boldsymbol{P}_{\mathbf{FDA_2}}, \ldots, \boldsymbol{P}_{\mathbf{FDA_7}}]$by using the FDA database and a 60- or 80- dimension vector for each $\boldsymbol{P}_{\mathbf{FDA}_k}$, *i.e.*, $\boldsymbol{P}_{\boldsymbol{PSSM-Grey}}^{(\mathbf{60})}$ or $\boldsymbol{P}_{\boldsymbol{PSSM-Grey}}^{(\mathbf{80})}$. These digital representations are used as the input of query protein for the prediction model.

2.5. Algorithm

Random Forest has been used as the main classifier of the predictor. The workflow (Fig. **1**) illustrates how our classifier works. In the proposed model, the first step is to input the query amino acid sequence with its FDAs. The next step is to generate two sets of features of a given protein as described above, where the annotation features are encoded into a distance matrix based on the KNN-score extraction, and PSI-BLAST was used to generate the PSSM and then transform into the SeqEvo Descriptor. In the last step, two types of features are assembled to enter the machine learning classifier as input for training.

2.6. Method Evaluation

To evaluate the prediction performance of our method, a 5-fold cross-validation test was performed following several widely-accepted measurements: (1) overall accuracy (ACC),

the ratio of true positive and true negative sample among all the samples; (2) Mathew's correlation coefficient or MCC; (3) sensitivity (SN), the ratio between true positive and positive samples; and (4) specificity (SP), the ratio between true negative and negative samples; (5) precision (Pre), the ratio of true positive among the sum of true positive and false positive. As mentioned in Section 2.1, there were 3 sets of negative data, and 3 sets of training data were constructed whose positive datasets were the same. Then training and 5-fold cross-validation were performed 3 times, then all these measurements were calculated from the average of 3 training sets. Furthermore, Receiver Operating Characteristic (ROC) curves were calculated and plotted based on specificities and sensitivities. The Areas under ROC curves (AUCs) were also calculated based on the trapezoidal approximation.

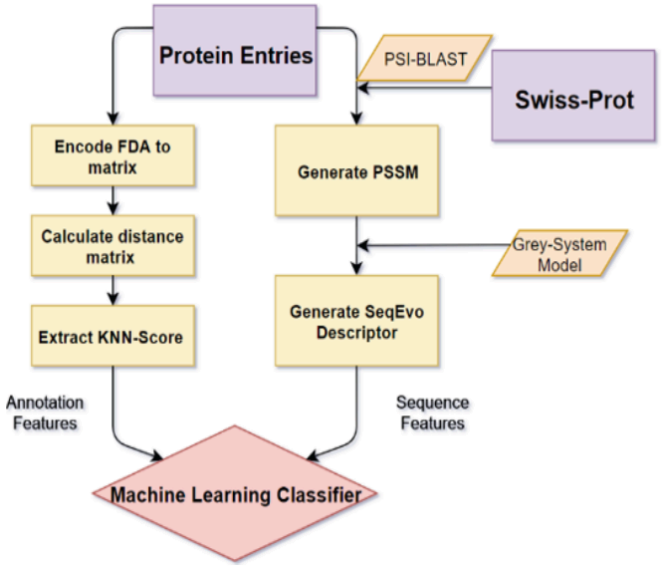

Fig. (1). Flowchart of our algorithm approach. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

3. RESULTS

3.1. Investigating the Performances of PSSM-Grey Feature

As mentioned above in the introduction, the PSSM-Grey feature is mainly based on the evolutionary conservation of proteins. To determine whether ubiquitination proteins and non-ubiquitination proteins have distinct evolutionary conservation patterns, the error bar of their four components $\boldsymbol{a^{j\prime}}$, $\boldsymbol{b_1^{j\prime}}$, $\boldsymbol{b_2^{j\prime}}$, and $\boldsymbol{b_3^{j\prime}}$ (Equation 3') was compared between ubiquitination proteins and non-ubiquitination proteins (Fig. **S1**). The result indicates that the ubiquitination proteins and non-ubiquitination proteins have some different patterns in the four parameters, but not very significant. $\boldsymbol{P_{PSSM-Grey}^{(60)}}$ and $\boldsymbol{P_{PSSM-Grey}^{(80)}}$ were applied in three general machine learning algorithms (Table **1**). The performance of SVM and RF was better than KNN, but there is a little difference between SVM and RF. The performance of $\boldsymbol{P_{PSSM-Grey}^{(60)}}$ was overall similar to that of $\boldsymbol{P_{PSSM-Grey}^{(80)}}$.

The GO enrichment network of the training dataset A, B, C indicates the positive datasets of *H. sapiens*, followed by *M. musculus* and *A. thaliana*, D, E, F indicate the negative datasets with the same order of species. The network was generated using Cytoscape [44], packaged in Metascape, with p-value < 0.01, minimum count 3, and enrichment factor > 1.5. A Kappa score of 4 was used as the similarity metric when performing hierarchical clustering on the enriched terms and then sub-trees with similarity > 0.3 were considered a cluster. Each node represents an enriched cluster and colored by its cluster ID as shown in the legend. The edge indicates the number of shared proteins between two-term nodes.

3.2. GO Enrichment Analysis

To confirm the classification results, the gene set GO enrichment analysis was performed using Metascape [50]. Here, the analysis of 1906 positive data on three species: *Homo sapiens* (393 GO terms), *Mus musculus* (390 GO terms) and *Arabidopsis thaliana* (277 GO terms) has been perfomed. For 10,000 negative dataset, the numbers of GO annotations was 1059, 1034 and 639, respectively. It shows that in all three species, most ubiquitination proteins belong to a small number of biological annotation terms with small p-values, which indicates that annotations could be useful features for our machine learning approach. Figs. (**2**) and (**3**) show that for *H. sapiens* and *M. musculus*, the positive datasets are more centered in the same functional GO term groups, with many linked edges this may be due to the commonness of mammals. The negative datasets for these two species are clustered in independent groups with fewer edges. Such a pattern is less obvious in *A. thaliana*.

Table 1. Performance comparison of PSSM-Grey by a 5-fold cross-validation.

%	$\boldsymbol{PSSM-Grey\,(80)}$				$\boldsymbol{PSSM-Grey\,(60)}$			
	ACC	MCC	SN	SP	ACC	MCC	SN	SP
KNN	79.99	60.52	86.63	73.35	80.43	61.37	86.79	74.08
SVM	88.68	77.60	84.75	92.61	87.62	75.28	86.00	89.24
RF	86.21	72.44	87.24	85.19	86.19	72.41	87.24	85.15
Average	84.96	70.19	86.20	83.72	84.75	69.68	86.68	82.82

Note: The abbreviations in the table are: Accuracy (ACC), Matthews Correlation Coefficient (MCC), Sensitivity (SN) and Specificity (SP).

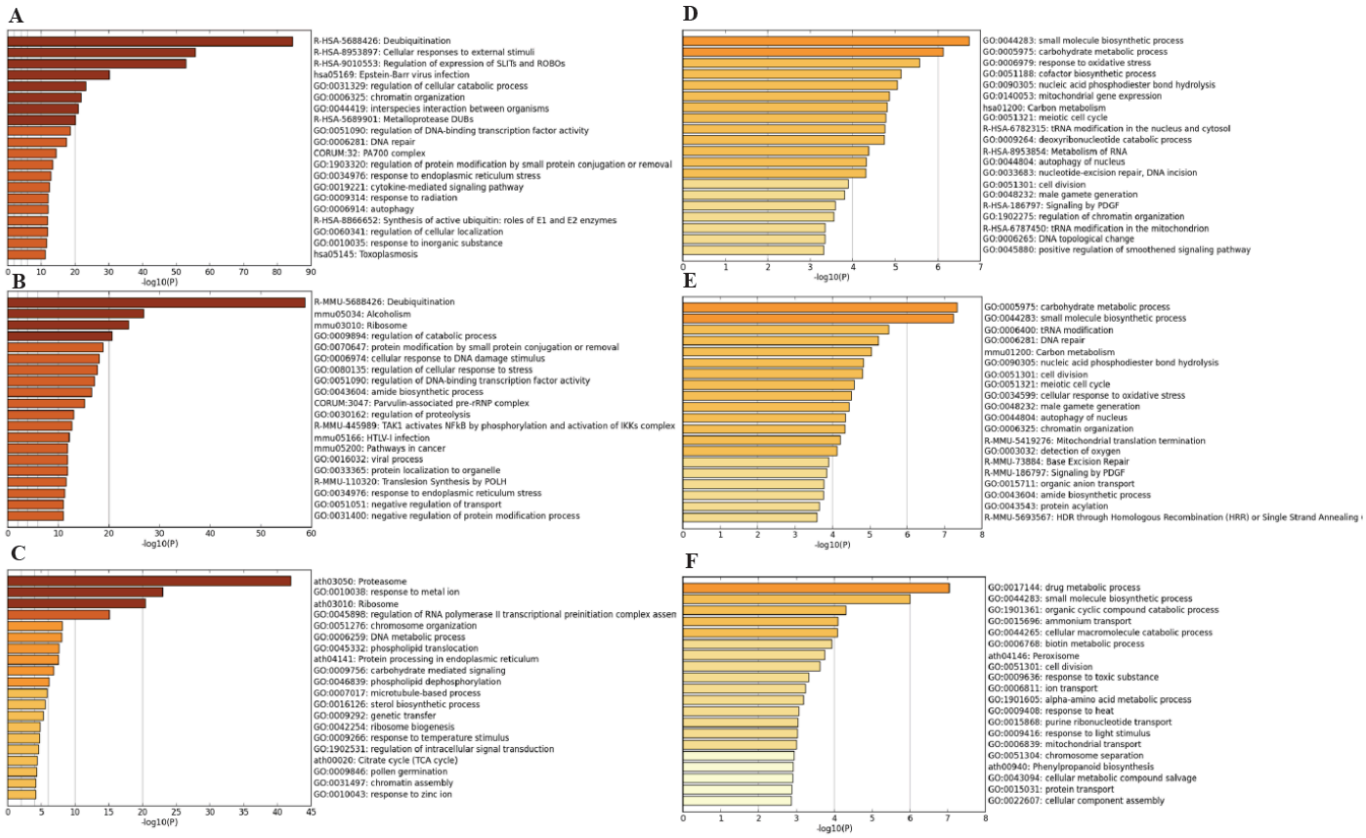

Fig. (2). GO enrichment analysis of the training dataset. **A**, **B**, **C** indicate the positive datasets of *H. Sapiens*, followed by *M. Musculus* and *A. Thaliana*, **D**, **E**, **F** indicate the negative datasets with same order of species. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

3.3. Subcellular Localization Enrichment Analysis

To study the relationships between ubiquitination proteins and subcellular localization, the enrichment analysis of subcellular localization for this dataset has been performed (Fig. **4**). As shown in Fig. (**4**), 49.8% of the positive data were labeled with the nucleus location, and 44.3% were localized in the cytoplasm, which is significantly different from the negative data (8.1% and 26.5%, respectively). Also, it is noted that 94.4% positive proteins have more than two localization annotations and 61.1% for negative proteins, especially, for positive data which is localized in the nucleus, around half of them (460 of 950) share other localizations. Some earlier studies have shown that the ubiquitin-related enzymes are highly localization- specific [24], and hence the subcellular localization could be an informative feature to predict ubiquitination proteins.

3.4. Investigating the Performances of KNN Score of Features

It was found that 5,184 GO terms were involved in the training dataset, of which 3,012 appeared in the set of ubiquitination proteins, 3,565 appeared in non-ubiquitination proteins, and only 1,393 GO terms were shared by both positive and negative datasets. Hence, the functional properties of the two groups are significantly different, as consistently shown in Figs. (**2** and **3**). Following this idea, the KNN scores of ubiquitination proteins were compared with those of non-ubiquitination proteins on all the FDA features (Fig. **S2**). Overall, ubiquitination proteins gained obvious larger KNN scores which are greater than 0.5 (*i.e.*, with significant information content as prediction feature; the larger, the more significant) on GO and subcellular localization, and a slightly larger score greater than 0.5 in the Smart, SUPFAM, and InterPro.

Specifically, for ubiquitination proteins, the average KNN scores of GO with different sizes of nearest neighbors were within 0.5 - 0.8, and for non-ubiquitination proteins, the average KNN scores were within 0.2 - 0.4. For subcellular localization, the average KNN scores of ubiquitination proteins were in the range of 0.5 - 0.7, while those of non-ubiquitination proteins fluctuated around 0.4. For Smart, SUPFAM, and InterPro, there was no clear gap between the ubiquitination proteins and non-ubiquitination proteins, especially with the growth of KNN cutoffs. Subsequently, the eight types of features were tested on the three datasets with KNN, RF and SVM algorithms on the training dataset, and the mean performance of these three algorithms is listed in Table **2**. The best results of accuracy for RF, SVM, and KNN are 0.88 (using InterPro), 0.85 (using Pfam) and 0.85 (using Pfam), respectively. The Random Forest algorithm has the best performance on all the features of accuracy, where the accuracy is between 0.70 - 0.88. Hence, Random Forest has been selected as our classifier.

3.5. Performance of the Proposed Model

Since the combined features generated a high-dimensional vector output, the Relief method [25] can be used to rank the values of the underlying features. To

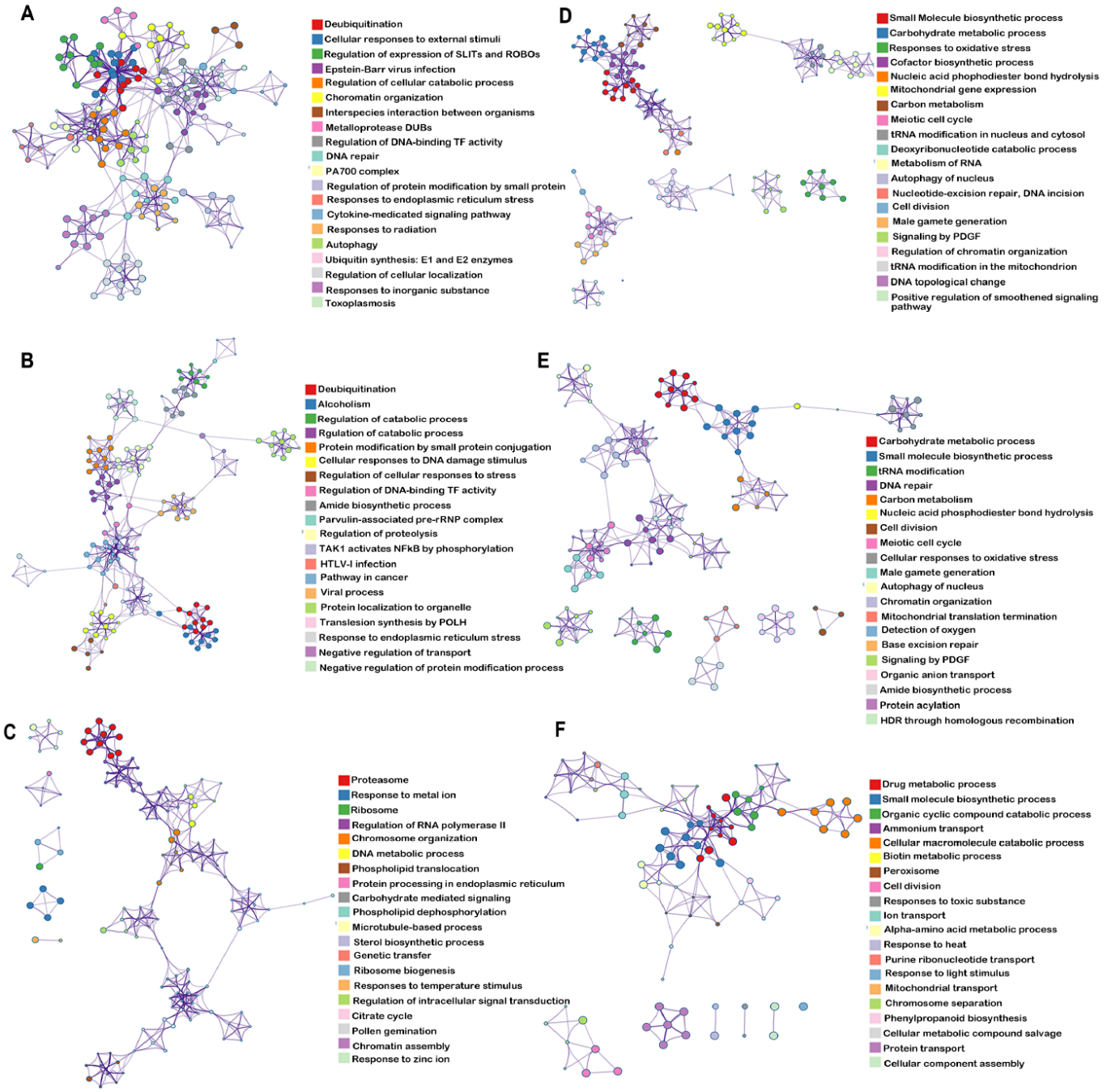

Fig. (3). GO enrichment analysis network visualization of the training dataset. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

evaluate the performance of our method for different features, 5-fold cross-validation has been performed (Tables **3** and **4** for training and testing, respectively). In general, the evaluation result was the best when all features were included reaching the accuracy of 87.71% and MCC of 75.43%. The performance of the proposed models was further illustrated by the ROC analysis (Fig. **5**), especially using AUC (Area Under the Curve). The AUC value is a number between 0 and 1, and the greater the AUC value, the better is the predictor. The AUC value of the proposed model is 0.8507 (Go), 0.8509 (PFAM), 0.8502 (SMART), 0.8495 (PROSITE), 0.8513 (SUPFAM), 0.8501 (INTERPRO), 0.8497 (PRINTS), 0.8476 (Subcellular localization), 0.9396 (GreyPssm) and 0.9598 (All).

3.6. Testing Data Performance and Comparison with Ubiquitination Site Prediction

A balanced independent test dataset was used to evaluate our model in comparison with ubiquitination site prediction tools. The results were based on five-fold cross-validation, with the model including all features together for our tool. The results were compared with a deep learning ubiquitination site prediction tool, MusiteDeep-Capsule [16]; a machine learning approach based tool, UbiProber [17]; an SVM

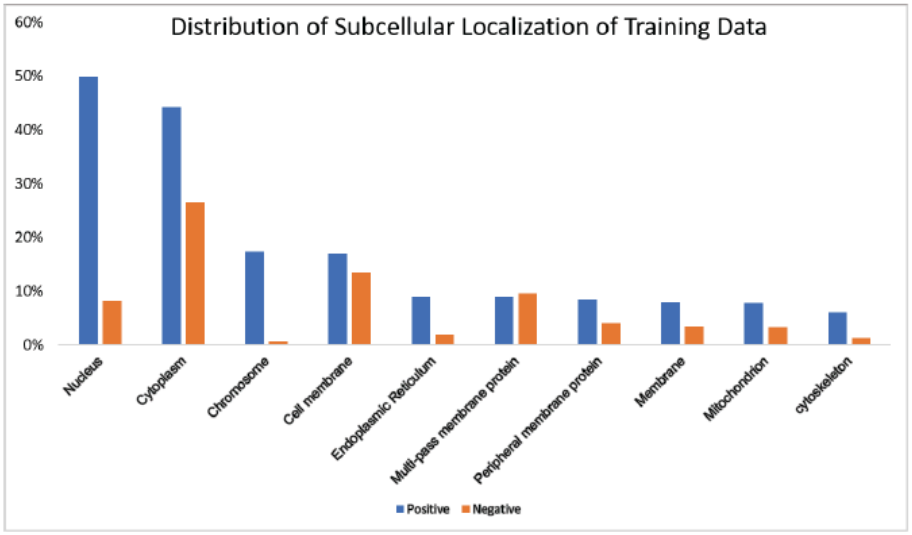

Fig. (4). Distribution of subcellular localization of the training data. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Table 2. A comparison of eight features with different algorithms.

%	ACC			MCC			SN			SP		
Algorithm	**RF**	**SVM**	**KNN**	**RF**	**SVM**	**KNN**	**RF**	**SVM**	**KNN**	**RF**	**SVM**	**KNN**
GO	82	81	80	64	62	60	83	79	79	81	83	81
Pfam	87	85	85	75	70	70	83	80	83	92	90	87
Smart	73	72	70	50	50	42	50	49	60	95	95	81
PROSITE	79	78	76	60	59	55	64	61	67	94	94	86
SUPFAM	75	73	74	52	48	50	60	56	60	90	90	88
InterPro	88	83	83	77	67	67	86	79	81	91	88	86
PRINTS	70	70	69	49	49	45	41	41	43	99	99	95
SL*	80	78	77	59	57	54	76	78	76	83	79	77

Note: The abbreviations in the table are: Accuracy (ACC), Matthews Correlation Coefficient (MCC), Sensitivity (SN) and Specificity (SP). Three algorithms, Random Forest (RF), Support Vector Machine (SVM) and KNN (K-Nearest Neighbor) were applied. * indicates Subcellular Localization (SL).

based tool, UbiSite [18]; and a Bayesian Discriminant Method based tool, BDM-PUB [19]. Since existing prediction tools are designed for site prediction, their site prediction results were transformed to ubiquitinated protein results by using the following strategy: if any site from a given protein was predicted as 'positive' or 'ubiquitinated', the whole protein was labeled as 'positive' as well; if multiple sites were predicted as positive, the max predicted score was picked from them for generating the Receiver Operator Curve (ROC). Their pre-trained models were used with default parameters to conduct the prediction comparison with our method. For MusiteDeep-Capsule (another in-house tool), the same test dataset was used to train the models and the same independent test dataset was used to perform the comparison. For other tools, they do not provide customized model training, therefore their pre-trained model was used to predict for the same testing dataset. The results are shown in Fig. (**6**), in which our tool showed a better performance than other tools; in particular, our tool has a significantly lower false-positive discovery rate and a higher true positive discovery rate than other tools.

4. DISCUSSION

In order to detect ubiquitination proteins, a method was developed based on the Random Forest algorithm using the sequence conservation information, as well as the information of 'GO', 'Pfam', 'Smart', 'PROSITE', 'SUPFAM', 'InterPro', 'PRINTS' and subcellular localization of the query protein. The features only incorporate the sequence conservation using a grey system model and KNN scores based on protein annotation databases. This method achieved an overall accuracy of 90.03%, MCC of 80.13%, Sn of 87.94%, Sp of 92.13% and Precision of 91.78%, which indicates that this method reflects the sequence patterns well, containing

Table 3. A comparison of eight features performance in the training data.

%	ACC	MCC	SN	SP	Precision	Recall
1 GO	82.18	64.39	83.29	81.08	81.52	83.29
2 Pfam	87.35	75.01	82.88	91.83	91.03	82.88
3 Smart	72.62	50.45	50.49	94.75	90.59	50.49
4 PROSITE	78.69	60.24	63.50	93.89	91.23	63.50
5 SUPFAM	75.01	52.30	60.43	89.59	85.34	60.43
6 InterPro	88.42	76.91	86.25	90.59	90.16	86.25
7 PRINTS	70.07	49.06	41.31	98.83	97.24	41.31
8 SL	79.64	59.49	76.00	83.28	82.02	76.00
9 PSSM	86.19	72.40	87.34	85.04	85.37	87.34
Feature(1-8)	89.74	79.53	87.85	91.63	91.30	87.85
Feature(1-9)	90.13	80.34	87.99	92.28	91.93	87.99

Note: The abbreviations in the table are: Accuracy (ACC), Matthews Correlation Coefficient (MCC),
Sensitivity (SN) and Specificity (SP). "Feature(1-8)" indicates that the first 8 features were applied, and "Feature(1-9)" means that all features were applied.

Table 4. A comparison of eight features performance in the test data.

-	ACC	MCC	SN	SP	Precision	Recall
1 GO	77.45	57.40	91.63	63.27	71.51	91.63
2 Pfam	82.39	65.06	78.10	86.67	85.49	78.10
3 Smart	72.61	47.70	56.80	88.43	83.13	56.80
4 PROSITE	78.14	56.86	71.05	85.23	82.79	71.05
5 SUPFAM	72.71	45.78	66.80	78.63	75.81	66.80
6 InterPro	80.16	60.40	81.76	78.56	79.31	81.76
7 PRINTS	70.78	47.54	46.54	95.03	90.37	46.54
8 SL	74.35	49.35	80.46	68.24	72.06	80.46
9 PSSM	85.72	71.45	86.21	85.23	85.40	86.21
Feature(1-8)	86.27	72.57	87.25	85.29	85.58	87.25
Feature(1-9)	87.71	75.43	87.91	87.52	87.57	87.91

Note: The abbreviations in the table are: Accuracy (ACC), Matthews Correlation Coefficient (MCC),
Sensitivity (SN) and Specificity (SP). "Feature(1-8)" indicates that the first 8 features were applied, and "Features(1-9)" means that all features were applied.

the ubiquitination sites. Since our method could do the prediction without relying on sequence profiles, it can scan a batch of unknown proteins very efficiently. In addition, our method showed better performance than the existing tools for the protein level prediction of ubiquitination. The user may apply our predictor to select potential candidates before doing the site prediction or the lab work.

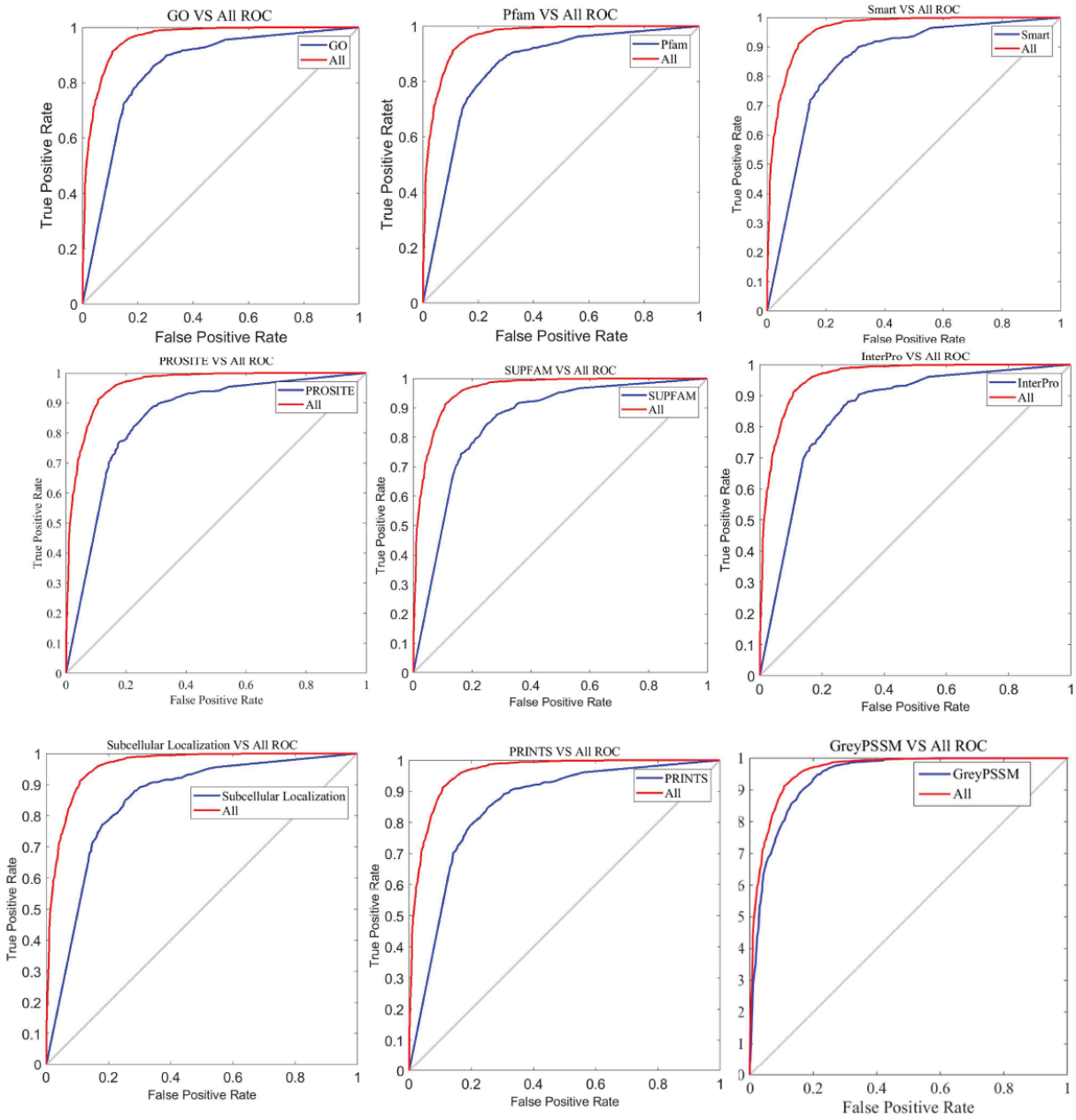

Fig. (5). ROC curves to show the performance of proposed models. The blue curve indicates the model with single feature and the red curve indicates the model includes all 9 features. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

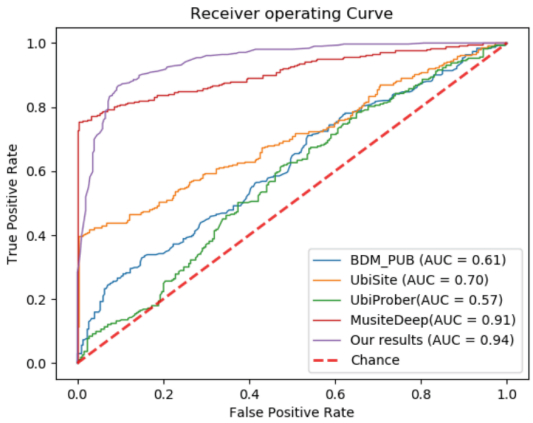

Fig. (6). ROC curves to show the performance comparison with other prediction tools. AUC indicates the area under the curve. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

CONCLUSION

Our study may guide experimental design and provide useful insights for studying the mechanisms and modulation of ubiquitination pathways. The comparison results indicate that we have an advantage in ubiquitination prediction at the protein level. It may improve the sensitivity when conducting the ubiquitination site prediction if our method is applied first to remove the false positive samples. In addition, it may help accelerate the expensive and time-consuming process of identifying ubiquitination proteins with known annotations.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The data supporting the findings of the article is available in the GitHub at: https://github.com/Chunhuixu/UBIPredic_QWRCHX.

FUNDING

This work was partially supported by the National Nature Science Foundation of China (No. 31760315, 31560316), the Natural Science Foundation of Jiangxi Province, China (No. 20171BAB202020), China Postdoctoral Science Foundation Funded Project (Project No. 2017M612949). The Scientific Research plan of the Department of Education of JiangXi Province (GJJ180703). It was also partially supported by the US National Institutes of Health grants R21-LM012790. The funders had no role in study design and data collection.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

We like to thank Duolin Wang for helping prepare the result with MusiteDeep-Capsule.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Aguilar, R.C.; Wendland, B. Ubiquitin: Not just for proteasomes anymore. *Curr. Opin. Cell Biol.,* **2003**, *15*(2), 184-190. [http://dx.doi.org/10.1016/S0955-0674(03)00010-3] [PMID: 12648674]

[2] Welchman, R.L.; Gordon, C.; Mayer, R.J. Ubiquitin and ubiquitin-like proteins as multifunctional signals. *Nat. Rev. Mol. Cell Biol.,* **2005**, *6*(8), 599-609. [http://dx.doi.org/10.1038/nrm1700] [PMID: 16064136]

[3] Haglund, K.; Dikic, I. Ubiquitylation and cell signaling. *EMBO J.,* **2005**, *24*(19), 3353-3359. [http://dx.doi.org/10.1038/sj.emboj.7600808] [PMID: 16148945]

[4] Hoeller, D.; Hecker, C.M.; Dikic, I. Ubiquitin and ubiquitin-like proteins in cancer pathogenesis. *Nat. Rev. Cancer,* **2006**, *6*(10), 776-788. [http://dx.doi.org/10.1038/nrc1994] [PMID: 16990855]

[5] Jadhav, T.; Wooten, M.W. Defining an embedded code for protein ubiquitination. *J. Proteomics Bioinform.,* **2009**, *2*, 316. [http://dx.doi.org/10.4172/jpb.1000091] [PMID: 20148194]

[6] Reinstein, E.; Ciechanover, A. Narrative review: Protein degradation and human diseases: The ubiquitin connection. *Ann. Intern. Med.,* **2006**, *145*(9), 676-684. [http://dx.doi.org/10.7326/0003-4819-145-9-200611070-00010] [PMID: 17088581]

[7] Schwartz, A.L.; Ciechanover, A. The ubiquitin-proteasome pathway and pathogenesis of human diseases. *Annu. Rev. Med.,* **1999**, *50*, 57-74. [http://dx.doi.org/10.1146/annurev.med.50.1.57] [PMID: 10073263]

[8] Iconomou, M.; Saunders, D.N. Systematic approaches to identify E3 ligase substrates. *Biochem. J.,* **2016**, *473*(22), 4083-4101. [http://dx.doi.org/10.1042/BCJ20160719] [PMID: 27834739]

[9] Cai, B.; Jiang, X. Computational methods for ubiquitination site prediction using physicochemical properties of protein sequences. *BMC Bioinformatics,* **2016**, *17*, 116. [http://dx.doi.org/10.1186/s12859-016-0959-z] [PMID: 26940649]

[10] Cai, Y.; Jiang, X. Computational methods for ubiquitination site prediction using physicochemical properties of protein sequences. BMC Bioinformatics, 2016, 17, 116. [https://doi.org/10.1186/s12859-016-0959-z] [PMID: 26940649]

[11] 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.,* **2015**, *16*(4), 640-657. [http://dx.doi.org/10.1093/bib/bbu031] [PMID: 25212598]

[12] Radivojac, P.; Vacic, V.; Haynes, C.; Cocklin, R.R.; Mohan, A.; Heyen, J.W.; Goebl, M.G.; Iakoucheva, L.M. Identification, analysis, and prediction of protein ubiquitination sites. *Proteins,* **2010**, *78*(2), 365-380. [http://dx.doi.org/10.1002/prot.22555] [PMID: 19722269]

[13] Cai, Y.; Huang, T.; Hu, L.; Shi, X.; Xie, L.; Li, Y. Prediction of lysine ubiquitination with mRMR feature selection and analysis. Amino Acids, 2012, 42(4), 1387-1395. [http://dx.doi.org/10.1007/s00726-011-0835-0] [PMID: 21267749]

[14] Zhao, X.; Li, X.; Ma, Z.; Yin, M. Prediction of lysine ubiquitylation with ensemble classifier and feature selection. *Int. J. Mol. Sci.,* **2011**, *12*(12), 8347-8361. [http://dx.doi.org/10.3390/ijms12128347] [PMID: 22272076]

[15] 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. [http://dx.doi.org/10.1371/journal.pone.0022930] [PMID: 21829559]

[16] Wang, D.; Liang, Y.; Xu, D. Capsule network for protein post-translational modification site prediction. *Bioinformatics,* **2019**, *35*(14), 2386-2394.

[17] Chen, X.; Qiu, J.D.; Shi, S.P.; Suo, S.B.; Huang, S.Y.; Liang, R.P. Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics,* **2013**, *29*(13), 1614-1622. [http://dx.doi.org/10.1093/bioinformatics/btt196] [PMID: 23626001]

[18] Huang, C.H.; Su, M.G.; Kao, H.J.; Jhong, J.H.; Weng, S.L.; Lee, T.Y. UbiSite: Incorporating two-layered machine learning method with substrate motifs to predict ubiquitin-conjugation site on lysines. *BMC Syst. Biol.,* **2016**, *10*(Suppl. 1), 6. [http://dx.doi.org/10.1186/s12918-015-0246-z] [PMID: 26818456]

[19] Li, X.; Gao, X.; Ren, J.; Jin, C.; Xue, Y. *BDM-PUB: Computational prediction of protein ubiquitination sites with a Bayesian discriminant method*, **2009**.

[20] Qiu, W.R.; Sun, B.Q.; Xiao, X.; Xu, D.; Chou, K.C. iPhos-PseEvo: Identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC *via* grey system theory. *Mol. Informatics,* **2016**, *36*(5-6). doi: 10.1002/minf.201600010.

[21] Chou, K.C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteomics,*

2009, *6*(4), 262-274.
[http://dx.doi.org/10.2174/157016409789973707]

[22] Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins,* **2001**, *43*(3), 246-255.
[http://dx.doi.org/10.1002/prot.1035] [PMID: 11288174]

[23] Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L.; Chou, K.C. Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.,* **2015**, *43*(W1), W65-W71.
[http://dx.doi.org/10.1093/nar/gkv458] [PMID: 25958395]

[24] Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; Harris, M.A.; Hill, D.P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J.C.; Richardson, J.E.; Ringwald, M.; Rubin, G.M.; Sherlock, G. Consortium, G.O. The Gene Ontology Consortium. Gene ontology: Tool for the unification of biology. *Nat. Genet.,* **2000**, *25*(1), 25-29.
[http://dx.doi.org/10.1038/75556] [PMID: 10802651]

[25] Jones, D.T. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics,* **2007**, *23*(5), 538-544.
[http://dx.doi.org/10.1093/bioinformatics/btl677] [PMID: 17237066]

[26] The UniProt Consortium. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.,* **2017**, *45*(D1), D158-D169.
[http://dx.doi.org/10.1093/nar/gkw1099] [PMID: 27899622]

[27] Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* **2006**, *22*(13), 1658-1659.
[http://dx.doi.org/10.1093/bioinformatics/btl158] [PMID: 16731699]

[28] Nakai, K.; Horton, P. PSORT: A program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.,* **1999**, *24*(1), 34-36.
[http://dx.doi.org/10.1016/S0968-0004(98)01336-X] [PMID: 10087920]

[29] Harris, M.A.; Clark, J.; Ireland, A.; Lomax, J.; Ashburner, M.; Foulger, R.; Eilbeck, K.; Lewis, S.; Marshall, B.; Mungall, C.; Richter, J.; Rubin, G.M.; Blake, J.A.; Bult, C.; Dolan, M.; Drabkin, H.; Eppig, J.T.; Hill, D.P.; Ni, L.; Ringwald, M.; Balakrishnan, R.; Cherry, J.M.; Christie, K.R.; Costanzo, M.C.; Dwight, S.S.; Engel, S.; Fisk, D.G.; Hirschman, J.E.; Hong, E.L.; Nash, R.S.; Sethuraman, A.; Theesfeld, C.L.; Botstein, D.; Dolinski, K.; Feierbach, B.; Berardini, T.; Mundodi, S.; Rhee, S.Y.; Apweiler, R.; Barrell, D.; Camon, E.; Dimmer, E.; Lee, V.; Chisholm, R.; Gaudet, P.; Kibbe, W.; Kishore, R.; Schwarz, E.M.; Sternberg, P.; Gwinn, M.; Hannick, L.; Wortman, J.; Berriman, M.; Wood, V.; de la Cruz, N.; Tonellato, P.; Jaiswal, P.; Seigfried, T.; White, R.; Gene Ontology, C. The Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.,* **2004**, *32*(Database issue), D258-D261.
[PMID: 14681407]

[30] Bateman, A.; Birney, E.; Durbin, R.; Eddy, S.R.; Finn, R.D.; Sonnhammer, E.L. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.,* **1999**, *27*(1), 260-262.
[http://dx.doi.org/10.1093/nar/27.1.260] [PMID: 9847196]

[31] Letunic, I.; Copley, R.R.; Schmidt, S.; Ciccarelli, F.D.; Doerks, T.; Schultz, J.; Ponting, C.P.; Bork, P. SMART 4.0: Towards genomic data integration. *Nucleic Acids Res.,* **2004**, *32*(Database issue), D142-D144.
[http://dx.doi.org/10.1093/nar/gkh088] [PMID: 14681379]

[32] Sigrist, C.J.; Cerutti, L.; de Castro, E.; Langendijk-Genevaux, P.S.; Bulliard, V.; Bairoch, A.; Hulo, N. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.,* **2010**, *38*(Database issue), D161-D166.
[http://dx.doi.org/10.1093/nar/gkp885] [PMID: 19858104]

[33] Pandit, S.B.; Bhadra, R.; Gowri, V.S.; Balaji, S.; Anand, B.; Srinivasan, N. SUPFAM: A database of sequence superfamilies of protein domains. *BMC Bioinformatics,* **2004**, *5*, 28.
[http://dx.doi.org/10.1186/1471-2105-5-28] [PMID: 15113407]

[34] Hunter, S.; Apweiler, R.; Attwood, T.K.; Bairoch, A.; Bateman, A.; Binns, D.; Bork, P.; Das, U.; Daugherty, L.; Duquenne, L.; Finn, R.D.; Gough, J.; Haft, D.; Hulo, N.; Kahn, D.; Kelly, E.; Laugraud, A.; Letunic, I.; Lonsdale, D.; Lopez, R.; Madera, M.; Maslen, J.; McAnulla, C.; McDowall, J.; Mistry, J.; Mitchell, A.; Mulder, N.; Natale, D.; Orengo, C.; Quinn, A.F.; Selengut, J.D.; Sigrist, C.J.; Thimma, M.; Thomas, P.D.; Valentin, F.; Wilson, D.; Wu, C.H.; Yeats, C. InterPro: The integrative protein signature database. *Nucleic Acids Res.,* **2009**, *37*(Database issue), D211-D215.
[http://dx.doi.org/10.1093/nar/gkn785] [PMID: 18940856]

[35] Attwood, T.K.; Coletta, A.; Muirhead, G.; Pavlopoulou, A.; Philippou, P.B.; Popov, I.; Romá-Mateo, C.; Theodosiou, A.; Mitchell, A.L. The PRINTS database: A fine-grained protein sequence annotation and analysis resource--its status in 2012. *Database (Oxford),* **2012**, *2012*, bas019.
[http://dx.doi.org/10.1093/database/bas019] [PMID: 22508994]

[36] Mcculloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biol.,* **1990**, *52*(1-2), 99-115.
[http://dx.doi.org/10.1007/BF02459570] [PMID: 2185863]

[37] Chou, K.C.; Elrod, D.W. Bioinformatical analysis of G-protein-coupled receptors. *J. Proteome Res.,* **2002**, *1*(5), 429-433.
[http://dx.doi.org/10.1021/pr025527k] [PMID: 12645914]

[38] Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.,* **1995**, *20*(3), 273-297.
[http://dx.doi.org/10.1007/BF00994018]

[39] Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory,* **1967**, *13*(1), 21-27.
[http://dx.doi.org/10.1109/TIT.1967.1053964]

[40] Ho, T.K. The random subspace method for constructing decision forests. *IEEE T Pattern Anal.,* **1998**, *20*(8), 832-844.
[http://dx.doi.org/10.1109/34.709601]

[41] Zhang, Z.H.; Wang, Z.H.; Zhang, Z.R.; Wang, Y.X. A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. *FEBS Lett.,* **2006**, *580*(26), 6169-6174.
[http://dx.doi.org/10.1016/j.febslet.2006.10.017] [PMID: 17069811]

[42] Xiao, X.; Lin, W.Z. Application of protein grey incidence degree measure to predict protein quaternary structural types. *Amino Acids,* **2009**, *37*(4), 741-749.
[http://dx.doi.org/10.1007/s00726-008-0212-9] [PMID: 19037711]

[43] Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.,* **2011**, *273*(1), 236-247.
[http://dx.doi.org/10.1016/j.jtbi.2010.12.024] [PMID: 21168420]

[44] Chou, K.C.; Shen, H.B. MemType-2L: A web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun.,* **2007**, *360*(2), 339-345.
[http://dx.doi.org/10.1016/j.bbrc.2007.06.027] [PMID: 17586467]

[45] Chou, K.C.; Shen, H.B. Recent progress in protein subcellular location prediction. *Anal. Biochem.,* **2007**, *370*(1), 1-16.
[http://dx.doi.org/10.1016/j.ab.2007.07.006] [PMID: 17698024]

[46] Chou, K.C. Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.,* **2004**, *11*(16), 2105-2134.
[http://dx.doi.org/10.2174/0929867043364667] [PMID: 15279552]

[47] Schäffer, A.A.; Aravind, L.; Madden, T.L.; Shavirin, S.; Spouge, J.L.; Wolf, Y.I.; Koonin, E.V.; Altschul, S.F. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.,* **2001**, *29*(14), 2994-3005.
[http://dx.doi.org/10.1093/nar/29.14.2994] [PMID: 11452024]

[48] Lin, W.Z.; Fang, J.A.; Xiao, X.; Chou, K.C. Predicting secretory proteins of malaria parasite by incorporating sequence evolution information into pseudo amino acid composition *via* grey system model. *PLoS One,* **2012**, *7*(11), e49040.
[http://dx.doi.org/10.1371/journal.pone.0049040] [PMID: 23189138]

[49] Beers, E.P.; Moreno, T.N.; Callis, J. Subcellular localization of ubiquitin and ubiquitinated proteins in *Arabidopsis thaliana*. *J. Biol. Chem.,* **1992**, *267*(22), 15432-15439.
[PMID: 1322398]

[50] Huang, W.; Sherman, B.T.; Lempicki, R.A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.,* **2009**, *37*(1), 1-13.
[http://dx.doi.org/10.1093/nar/gkn923] [PMID: 19033363]