

# Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites

Xiang Chen<sup>1</sup>, Jian-Ding Qiu<sup>1,2,\*</sup>, Shao-Ping Shi<sup>1,3</sup>, Sheng-Bao Suo<sup>1</sup>, Shu-Yun Huang<sup>1</sup> and Ru-Ping Liang<sup>1</sup>

<sup>1</sup>Department of Chemistry, Nanchang University, Nanchang 330031, People's Republic of China, <sup>2</sup>Department of Chemical Engineering, Pingxiang College, Pingxiang 337055, People's Republic of China and <sup>3</sup>Department of Mathematics, Nanchang University, Nanchang 330031, People's Republic of China

Associate Editor: Jonathan Wren

## ABSTRACT

**Motivation:** Systematic dissection of the ubiquitylation proteome is emerging as an appealing but challenging research topic because of the significant roles ubiquitylation play not only in protein degradation but also in many other cellular functions. High-throughput experimental studies using mass spectrometry have identified many ubiquitylation sites, primarily from eukaryotes. However, the vast majority of ubiquitylation sites remain undiscovered, even in well-studied systems. Because mass spectrometry-based experimental approaches for identifying ubiquitylation events are costly, time-consuming and biased toward abundant proteins and proteotypic peptides, *in silico* prediction of ubiquitylation sites is a potentially useful alternative strategy for whole proteome annotation. Because of various limitations, current ubiquitylation site prediction tools were not well designed to comprehensively assess proteomes.

**Results:** We present a novel tool known as UbiProber, specifically designed for large-scale predictions of both general and species-specific ubiquitylation sites. We collected proteomics data for ubiquitylation from multiple species from several reliable sources and used them to train prediction models by a comprehensive machine-learning approach that integrates the information from key positions and key amino acid residues. Cross-validation tests reveal that UbiProber achieves some improvement over existing tools in predicting species-specific ubiquitylation sites. Moreover, independent tests show that UbiProber improves the areas under receiver operating characteristic curves by ~15% by using the *Combined* model.

**Availability:** The UbiProber server is freely available on the web at <http://bioinfo.ncu.edu.cn/UbiProber.aspx>. The software system of UbiProber can be downloaded at the same site.

**Contact:** [jdqiu@ncu.edu.cn](mailto:jdqiu@ncu.edu.cn)

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on November 21, 2012; revised on April 13, 2013; accepted on April 21, 2013

## 1 INTRODUCTION

The 2004 Nobel Prize in Chemistry was awarded to Aaron Ciechanover, Avram Hershko and Irwin Rose for their discovery of ubiquitin-mediated protein degradation in eukaryotes.

Numerous studies subsequent to their original work showed that the ubiquitin–proteasome system (UPS) plays a critical role in regulating a variety of biological processes, such as the cell cycle and division, the immune response, inflammation and signal transduction (Reinstein and Ciechanover, 2006). Further analysis proposed that ubiquitin-mediated ubiquitylation was a ubiquitous post-translational modification in eukaryotes. Ubiquitylation has a conserved proteasome system in which the conjugation of ubiquitin to substrates usually involves three steps (Pickart and Eddins, 2004). Ubiquitin is first activated by a ubiquitin-activating enzyme (E1), then conjugated to a ubiquitin-conjugating enzyme (E2) and finally transferred to a substrate molecule with the assistance of a ubiquitin-ligase enzyme (E3), forming an isopeptide bond with an internal lysine of the target protein. In the UPS, ubiquitylation regulates a variety of biological processes. Thereby, aberrations of UPS function in all of the above-mentioned processes have been implicated in the pathogenesis of human diseases, ranging from inflammatory neurodegenerative diseases, to various forms of malignancies (Mani and Gelmann, 2005; Reinstein and Ciechanover, 2006).

To further decipher the mechanism of the UPS or other regulatory roles of ubiquitylation at the molecular level, an initial but crucial step is to identify ubiquitylation substrates and the corresponding ubiquitylation sites (Haglund and Dikic, 2005). Researchers have used several experimental methods to purify ubiquitylation sites, such as the use of affinity-tagged ubiquitin (Tomlinson *et al.*, 2007), ubiquitin antibodies and ubiquitin-binding proteins (Kim *et al.*, 2011; Xu *et al.*, 2010) and high-throughput mass spectrometry (MS) technique (Peng *et al.*, 2003; Wagner *et al.*, 2011). Unfortunately, experimental identification of ubiquitylation sites in a global manner is still a difficult, expensive and time-consuming task (Hitchcock *et al.*, 2003; Ikeda and Dikic, 2008). Hence, *in silico* prediction of ubiquitylation events can be highly valuable in many cases. Tung and Ho (2008) designed the first ubiquitylation sites predictor (UbiPred) using a support vector machine (SVM) with 31 informative physicochemical features selected from the published amino acid indices (Kawashima and Kanehisa, 2000). Subsequently, Radivojac *et al.* (2010) proposed a random forest-based predictor called UbPred, in which 586 sequence attributes were used as the input feature vectors. Lee *et al.* (2011) developed a method, UbSite, using an efficient radial basis function (RBF) network to identify protein ubiquitylation sites.

\*To whom correspondence should be addressed.

CKSAAP\_UbSite was constructed by Chen *et al.* (2011) for identifying ubiquitylation sites by using the composition of  $k$ -space amino acid pairs surrounding a query site with the assistance of a SVM. Recently, Cai *et al.* (2012) presented a method based on multi-sequence features and the nearest neighbor algorithm. These computational methods can be divided into two categories: methods of ubiquitylation site prediction for single species and multispecies. UbPred and CKSAAP\_UbSite fall into the first category. The latter category includes UbiPred, UbSite and the method of Cai *et al.* More information about these tools or methods is given in Supplementary Table S1.

Despite the availability of various ubiquitylation site prediction methods, they have limitations when applied to whole proteomes. The first limitation of ubiquitylation site prediction is accuracy. The overall performance of the several aforementioned predictors is still not fully satisfactory, and there is still room to improve the predictive accuracy. The second limitation is that the ubiquitylation relationships among different species are not discussed in these methods. Additionally, there is no systematic species-specific model to maximize the predictor performance of ubiquitylation sites. The third limitation is the training data deficiency and homology bias that lead to lack of statistical significance. Furthermore, some of these tools were only published as a method and did not provide a user-friendly web service, and this makes predictions a laborious or impossible task. Even those tools with web services either do not support stringency adjustment or only support several predefined stringency levels from which a user can choose that may not meet every user's requirement. User control of the prediction stringency is important, especially for large-scale predictions, because typically a user is interested only in predictions above a certain confidence threshold.

To address the limitations of the existing methods, we developed a new tool known as UbiProber, which is specifically designed to predict both general and species-specific ubiquitylation sites. Reliable and large-scale experimental ubiquitin proteomics data from multiple species were collected from several sources and used to train the ubiquitylation site prediction models. Three sets of features [ $k$  nearest neighbor (KNN) feature, physicochemical property (PCP) and amino acid composition (AAC)] were extracted from the training data and combined using a SVM to make predictions. The KNN features capture the local sequence similarity around sites that are ubiquitylated by the same enzyme or enzyme family regardless of whether the enzyme-substrate interactions are known. PCPs and AACs reflect the biochemical environment of the regions surrounding ubiquitylation sites, and these regions play various roles in the structure and function of a protein. Additionally, to extract the meaningful information and enhance the overall accuracy of the predictor, the information gain (IG) method (Chen *et al.*, 2009) was first used to choose some key positions and key amino acid residues to optimize each feature set. Furthermore, we discussed the relationship of ubiquitylation in different species. Our analysis shows the following: (i) the ubiquitylation patterns are conserved across different species; (ii) some key positions and key amino acid residues are essential for improving the prediction performance of a ubiquitylation model; and (iii) the physicochemical properties of residues in the flanking sequences are important for the ubiquitylation process.

Finally, the software system and web service of UbiProber were implemented in .Net 4.0 framework and are freely available to academic researchers. In UbiProber, we have trained the general prediction model for a eukaryotic proteome and species-specific prediction models for three species-specific proteomes. In addition, UbiProber supports continuous stringency adjustment to meet the various confidence requirements of users. Cross-validation tests and comparisons with other tools show that UbiProber performs better at general prediction and is comparable with existing methods at species-specific predictions. We believe that UbiProber could serve as a powerful and complementary tool for *in vivo* or *in vitro* ubiquitylation site identification. Additionally, the combination of computational analyses with experimental verification could greatly speed up our systematic understanding of the mechanisms and dynamics of ubiquitylation.

## 2 METHODS

### 2.1 Data collection and preprocessing

Ubiquitylation data for three species, *Homo sapiens*, *Mus musculus* and *Saccharomyces cerevisiae* from several sources including PhosphoSite (December 23, 2012), UniProtKB/Swiss-Prot (December 23, 2012), UbiProt (Chernorudskiy *et al.*, 2007), SCUD (Lee *et al.*, 2008) and two large-scale proteomics studies (Kim *et al.*, 2011; Radivojac *et al.*, 2010) were collected (Fig. 1). Although there are exceptions (Supplementary Table S2), ubiquitin is usually attached to lysine residues of proteins (Shimizu *et al.*, 2010). Therefore, we only considered lysine ubiquitylation. Supplementary Table S3 lists the ubiquitylation data sources and statistics for different sources for each species. The original dataset contained 11 547 proteins covering different species; of these proteins, 8044 are from *H.sapiens*, 3355 are from *M.musculus* and 208 are from *S.cerevisiae*. After removing the redundant protein sequences to eliminate homology bias using the program cd-hit (Li and Godzik, 2006), we obtained 6323 *H.sapiens*, 2342 *M.musculus* and 145 *S. cerevisiae* distinct proteins among which the sequence identity was <40%. Afterward, we randomly selected 20 proteins from the datasets of three species to form the independent test sets, and the remaining 8024, 2322, 125 proteins were used to construct the training set, respectively.

We truncated these sequences to 27-residue long symmetrical windows ( $-13 \sim +13$ ) with lysine residues at the center to characterize the ubiquitylation sites, and obtained 25 194 *H.sapiens* sequence segments, 5348

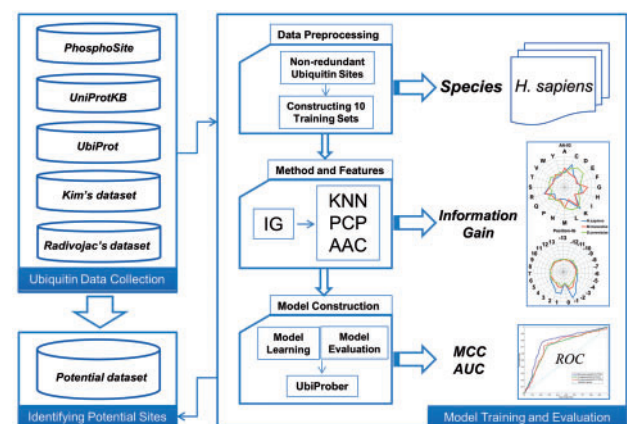


Fig. 1. Work flow of UbiProber

*M.musculus* sequence segments and 175 *S.cerevisiae* sequence segments for the positive training sets. Redundant segments with sequence similarity >40% were also removed again by using the cd-hit algorithm, and the final positive training datasets contained 18 085 *H.sapiens*, 3986 *M.musculus* and 121 *S.cerevisiae* symmetrical segments. At present, there is no uniform standard for constructing a negative sample. In this work, it would be difficult to prove definitively that a particular lysine residue is not ubiquitylated under any conditions. However, almost all researchers (Cai *et al.*, 2012; Chen *et al.*, 2011; Lee *et al.*, 2011; Radivojac *et al.*, 2010) of ubiquitylation prediction use a strategy in which any lysine residue that is not marked by any ubiquitylation information on the same protein is a non-ubiquitylation site. In addition, Radivojac *et al.* (2010) concluded that this strategy did not significantly influence the prediction performance for a model of yeast. In fact, although ubiquitylation occurs frequently in cells, the amount of ubiquitylation sites is still small compared with the large amount of peptide sequences. Thus, non-ubiquitylation sites should dominate negative samples. Therefore, we adopted this strategy for choosing negative training sets. Because the number of ubiquitylation sites and non-ubiquitylation sites were highly imbalanced, we randomly chose negative samples (non-ubiquitylation segments) for 10 times to match the number of positive ones (ubiquitylation segments) in the training sets. In the independent test sets, we retained the entire positive and the same number of negative samples to make it close to the training situation. There were 61, 45 and 43 positive samples for *H.sapiens*, *M.musculus* and *S.cerevisiae*, respectively. Following the evaluation by 5-fold cross-validation, the trained model with the highest accuracy was further evaluated using an independent test. The independent test sets were used to test not only the proposed method but also the previously proposed ubiquitylation prediction tools, UbiPred, UbPred and CKSAAP\_UbSite.

## 2.2 Feature extraction

**KNN Features** To take advantage of the cluster information of local sequences for predicting ubiquitylation sites, we used the local sequence around a possible ubiquitylation site in a query protein and extracted features from similar sequences in both the positive and negative sets by a KNN algorithm as follows:

- (I) For a query site (possible ubiquitylation site), find its KNNs in the positive and negative sets, respectively, according to the local sequence similarity. For two local sequence  $s_1$  and  $s_2$ , the distance  $\text{Dist}(s_1, s_2)$  is defined as

$$\text{Dist}(s_1, s_2) = 1 - \frac{\sum_{i=-p}^p \text{Sim}(s_1(i), s_2(i))}{2p + 1} \quad (1)$$

where  $p$  denotes the number of flanking residues from the central site in the protein sequence fragment, and  $i$  denotes the position of an amino acid in the target sequence segment. Sim, the amino acid similarity matrix, is derived from the BLOSUM62 substitution matrix (Henikoff and Henikoff, 1992) as

$$\text{Sim}(a, b) = \frac{M(a, b) - \min\{M\}}{\max\{M\} - \min\{M\}} \quad (2)$$

where  $a$  and  $b$  are two amino acids,  $M$  is the substitution matrix and  $\max\{M\}$  and  $\min\{M\}$  represent the largest/smallest number in the matrix, respectively.

- (II) The corresponding KNN feature is then extracted as follows:
  - (i) Form a set of neighbors, known as the Comparison set, by combining the positive and negative sets;
  - (ii) Calculate the average distances from the query sequences to the Comparison set;
  - (iii) Sort the neighbors by the distances, and pick the KNNs;
  - (iv) Calculate the KNN score, which is the percentage of positive neighbors (ubiquitylation sites) in its KNNs; and
  - (v) To obtain

multiple features, choose  $k$  as different values of the size of the training dataset. In UbiProber, five KNN scores were extracted as features for ubiquitylation prediction.

**PCP** AAindex (Kawashima and Kanehisa, 2000) is a database of amino acid indices and amino acid mutation matrices. Currently, 544 PCPs can be retrieved from the database of version 9.0. An amino acid index is a set of 20 numerical values representing various PCPs of amino acid. The structure and function of proteins are largely dependent on various properties of each of the 20 amino acids.

PCP has been successfully applied to predict such protein modification as phosphorylation (Moses *et al.*, 2007), acetylation (Lee *et al.*, 2010) and sulfation (Niu *et al.*, 2010). Moreover, Tung and Ho (2008) and Radivojac *et al.* (2010) have shown that the feature set of PCPs is the best type of feature for predicting ubiquitylation sites. Consequently, some informative PCPs were selected in our work by comparing the prediction accuracy of all PCPs. The comparison results are displayed in Supplementary Figures S1–S3. Here, the top six PCPs were selected and defined as informative features for the prediction model. The informative PCPs are shown in Supplementary Table S4.

**AAC** Lee *et al.* (2011) and Radivojac *et al.* (2010) analyzed the AAC of the sequences surrounding ubiquitylation sites in their works and found that the AAC can be a useful features for ubiquitylation site prediction. Hence, in this work, AAC was used in the same manner as the researchers described. We calculated the amino acid frequencies in the sequence surrounding the query site (the site itself is not counted). There are 20 types of amino acids, and thus 20 frequencies are calculated, the sum of which is 1.

## 2.3 Feature optimization

All positions and amino acid residues in a protein are not equally important. Some are essential for the proper structure and function of a protein. In addition, there are also differences for homologous proteins in different species (Whisstock and Lesk, 2003). Although mammalian proteins show a high degree of sequences similarity, there are still differences in protein function among different species (Gabaldon and Huynen, 2004). Here, a well-established feature selection method, IG (Chen *et al.*, 2009), was first used to distinguish the importance of different positions and different residues for ubiquitylation sequence segments. The calculation process is described in the Supplementary Materials. Using this method, we ranked list of IG scores for 27 positions and 20 amino acid residues. Note that the larger the IG scores of positions or residues are, the more important the positions or residues are. Therefore, the positions and residues were sorted from high to low using the IG score, after which we trained the model only using the top  $n$  positions and the top  $m$  residues. In this way, we chose key positions and key residues for optimizing the features. The optimized processes were performed using the following procedures: (i) Key positions were used in KNN features; when we calculated the KNN score, Equation (1) was calculated using only the key positions [that is to say  $i$  equal to key positions in Equation (1)]. (ii) Key positions were used in the PCP features and key residues were used in the AAC features; only the features of the key positions were extracted after calculating the PCP features, and only the features of the key residues were extracted after calculating the AAC features.

## 2.4 Model training and evaluation

We used the extracted features to train the ubiquitylation prediction models. As illustrated in Figure 1, the model construction procedure consisted of two separate procedures: model learning and model evaluation.

**SVM Learning** As a machine-learning method of binary classification, SVM aims to find a rule that best maps each member of a training



set to the correct classification (Vapnik, 1999), and SVM has been used for diverse prediction/classification tasks related to protein bioinformatics. Using the feature encoding of the key positions and key amino acid residues, the SVM was trained to distinguish ubiquitylation and non-ubiquitylation sites in UbiProber. The implemented SVM algorithm was LIBSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>) and the applied kernel function was the RBF. To maximize the performance of the SVM algorithm, the grid search method was applied to tune the parameters.

**Performance Assessment** To evaluate the prediction performance of UbiProber, 5-fold cross-validation tests were performed. Receiver operating characteristic (ROC) curves were plotted based on specificity ( $Sp$ ) and sensitivity ( $Sn$ ), and the areas under the ROC curves (AUC) were calculated using the trapezoidal approximation. The AUC provides a measure of the classifier performance; the larger the value of the AUC is, the better the performance of the classifier is. In addition, the Matthew correlation coefficient ( $MCC$ ) was also used to evaluate the prediction performance. The  $MCC$  accounts for true and false positives and negatives and are usually regarded as a balanced measure that can be used even if the classes are of different sizes. The  $Sn$ ,  $Sp$  and  $MCC$  are defined as follows:

$$Sn = \frac{TP}{TP + FN} \quad (3)$$

$$Sp = \frac{TN}{TN + FP} \quad (4)$$

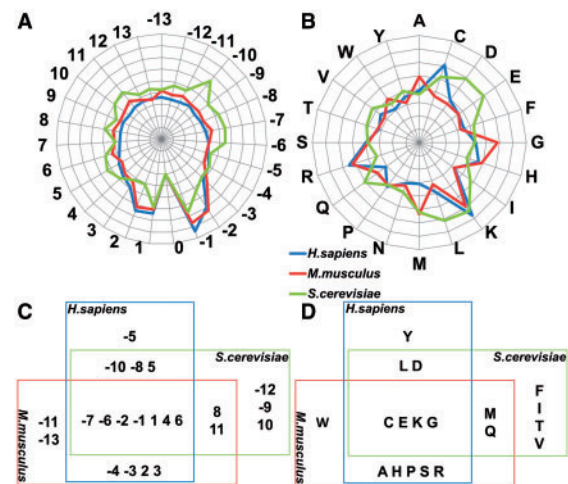
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \quad (5)$$

where  $TP$ ,  $TN$ ,  $FP$  and  $FN$  denote the number of true positives, true negatives, false positives and false negatives, respectively.

### 3 RESULTS AND DISCUSSION

#### 3.1 IG results

In this work, the IG method is first used to rank the importance of positions and amino acid residues. The IG scores of positions and residues for different species are calculated using the IG method, and the results are listed in Supplementary Table S5. This investigation focuses on the analysis of the similarities and differences of important positions and residues for different species. As shown in Figure 2A and B, there are differences in the importance of the same position and residue for different species. Taking the ubiquitylation proteins in *S.cerevisiae*, for example, the positions of -10, -7, -6, -1, 1, 5, 6, 11 and the residues of Cys, Asp, Glu, Ile, Leu, Lys are more important than other positions and residues. Actually, this result has been validated by several relevant reports. Catic *et al.* (2004) found similar results in which ubiquitylation sites in yeast proteins have a fold increase in Asp and Glu at positions -2 and -1 compared against the entire yeast proteome. Radivojac *et al.* (2010) reported that a feature of the ubiquitylation sequence of *S.cerevisiae* is the abundance of Asp and Glu. In addition, Xu *et al.*'s (2010) analysis showed that there is a subtle enrichment for mammalian proteins at the positions of -2, -1, 1 and 2, which is in agreement with our survey for *H.sapiens* and *M.musculus*. These results demonstrate compositional differences between ubiquitylated and non-ubiquitylated sequences at several specific positions and specific residues.



**Fig. 2.** Comparison of the importance of the different positions and amino acid residues from three species. (A) Radar of IG scores of positions around ubiquitylation sites. (B) Radar of IG scores of amino acid residues around ubiquitylation sites. (C) The numbers in the box indicate the positions of the ubiquitylation sequence segment. (D) The letters in the box indicate the amino acid residues of the ubiquitylation sequence segment

Moreover, we perceived that the different species have significant comparability after comparing the results for the top 15 positions and the top 12 residues of the three species according to the IG score ranking. As can be seen from Figure 2C, some positions (-7, -6, -2, -1, 1, 4, 6) are relatively more important for the ubiquitylation of eukaryotic substrates, and we proposed that the upstream positions are more important for the lysine residue to be ubiquitylated. Additionally, we can see that each of the species has unique key positions. There is a possibility that ubiquitin-related enzymes are species specific, although the reaction process is similar for eukaryotes (Pickart and Eddins, 2004). Interestingly, we also found that evolutionarily closer species have a more consistent trend in the key positions. For example, *H.sapiens* and *M.musculus* are more similar than *H.sapiens* and *S.cerevisiae* in the important positions. As can be seen from Figure 2D, the involvement of residues (Cys, Glu, Lys and Gly) in the interactional aspects of eukaryotic ubiquitylation are relatively important. Similarly, each species has unique key residues that are largely related to the amount of sample because the IG score and the total number of sequences are closely related; however, these results do not rule out the specificity of the ubiquitin-related enzymes. In a summary, there are many similarities in the ubiquitylated sequence preference of different species.

Based on the information contained in this study, the possible conclusions are that those key positions and key residues may be more important to form the proper structure for enzyme-substrate interactions. The key positions and key residues are conserved among different species. In fact, this finding is in agreement with some ubiquitylation site surveys for eukaryotic proteins in previous work (Haglund and Dikic, 2005; Peng *et al.*, 2003; Xu *et al.*, 2010). This conclusion also provides the theoretical basis to build a *Combined* model for general prediction later.

### 3.2 KNN scores results

The KNN feature has been successfully used in phosphorylation site prediction (Gao *et al.*, 2010). The KNN scores measures whether the local sequence surrounding a query site is more similar to the ubiquitylation sequence or to non-ubiquitylation sequences. A query sequence with a KNN score higher than 0.55 is more similar to the positive samples; a query sequence with a KNN score lower than 0.55 is more similar to the negative samples. The larger the KNN score is, the more similar the site is to some known ubiquitylation sites, and thus, the more likely it is to be a ubiquitylation site. Figure 3 compares the KNN scores between ubiquitylation and non-ubiquitylation sequences. (Due to the gap of data, we selected different  $k$  values and Comparison sets; Supplementary Table S6.) As can be seen, there is a significant difference among various values of  $k$  in the KNN score; ubiquitylation sequences have larger KNN scores than non-ubiquitylation sequences. For ubiquitylation, the average KNN scores with different sizes of nearest neighbors are within 0.55–0.80 for all three species. Therefore, the local sequences surrounding known ubiquitylation sites are more similar to their nearest neighbors in the positive set (excluding self-matches) on average as expected. Note that such similarities are not due to protein homology because the global sequence similarity between any two proteins in our non-redundant datasets is either insignificant or low. This finding confirms that ubiquitylation-related clusters exist in local sequences around ubiquitylation sites. For non-ubiquitylation sites, the average KNN scores are  $\sim 0.5$ , which means that overall the sequences in the negative set are not predominantly more similar to the nearest neighbors in either the positive or negative set. This result is not surprising because ubiquitylation-related sequence clusters are unlikely to exist in the negative set, and thus, the sequences in the negative set have a similar chance to find close neighbors in either the positive or negative set. In short, the KNN scores capture the cluster information in the local sequence around ubiquitylation

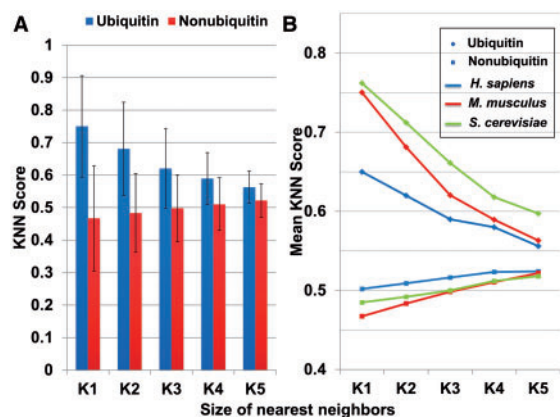


Fig. 3. Comparison of KNN scores between ubiquitylation sequences and non-ubiquitylation sequences. (A) Histogram of KNN scores of five thresholds (K1=0.00125, K2=0.0025, K3=0.005, K4=0.01, K5=0.02, Comparison set: 2000 positive and 2000 negative samples) in *M.musculus*. The error bars represent 90% confidence intervals. (B) Comparison of mean KNN scores between the ubiquitylation sequences and non-ubiquitylation sequences in three species. Note that these are the results after the IG processing

sites and hence distinguish them from the background. Therefore, KNN scores are suitable to be used as features for ubiquitylation site prediction.

KNN scores are effective features when used for predicting both general and species-specific ubiquitylation sites. For ubiquitylation site predictions, KNN scores can automatically capture the local sequence similarity between substrates of a common ubiquitin-related enzyme or ubiquitin-related enzyme family without relying on knowledge of enzymes–substrate interactions or enzymes-binding sequence motifs. Therefore, although most known ubiquitylation sites have no annotation about their corresponding enzymes, KNN scores can still use the inherent cluster information in them. Oftentimes, one enzyme corresponds to multiple local sequence motifs, and using a single sequence profile may not be as effective as KNN, which better handles diverse sequence clusters.

### 3.3 PCP results

From the previous work (Radivojac *et al.*, 2010; Tung and Ho, 2008), we knew that informative PCPs play an important role in ubiquitylation prediction. Here, a specific example of informative PCPs in *H.sapiens* was provided. As can be seen from Supplementary Table S4, the property ‘Positive charge’ (FAUJ880111) plays an important role in predicting ubiquitylation sites. We calculated the mean positive charge at each position (excluding the ubiquitylation site itself) in ubiquitylation sequences and non-ubiquitylation sequences and depicted them in Figure 4. It is clear that the ubiquitylation sequences and non-ubiquitylation sequences have different charge pattern. The positive charge of the ubiquitylation sequences at certain upstream positions (−13 ~ −4) and certain downstream positions (6 ~ 12) is higher than the charge in the non-ubiquitylation sequences. The positive charge in some positions (−2, −1, 1, 2 and 3) near the central sites and some downstream positions (5 and 13) of the non-ubiquitylation sequences are higher than that of the ubiquitylation sequences. This difference is more remarkable in the upstream and ubiquitylation sites nearby than in the downstream position. This result is entirely consistent with the previously mentioned IG results that also prove the importance and availability of the IG method, which can find key differences. In addition, we are also surprised to find that the ubiquitylation sequences contain more positively charged amino acid

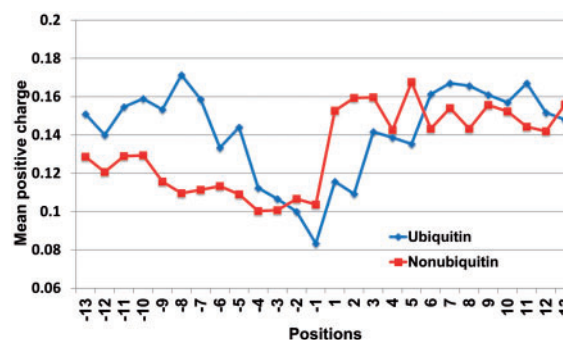
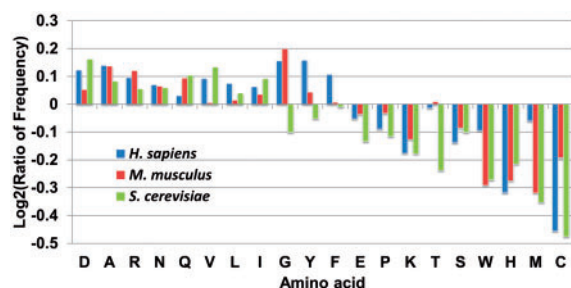


Fig. 4. The mean positive charge at each position in ubiquitylation sequences and non-ubiquitylation sequences



**Fig. 5.** Comparisons of AACs in positive and negative datasets. The vertical axis represents the  $\log_2$  ratio of amino acid frequencies surrounding ubiquitylation sites and to those surrounding non-ubiquitylation sites. The horizontal axis represents the 20 amino acids sorted in descending order by the mean  $\log_2$  ratio

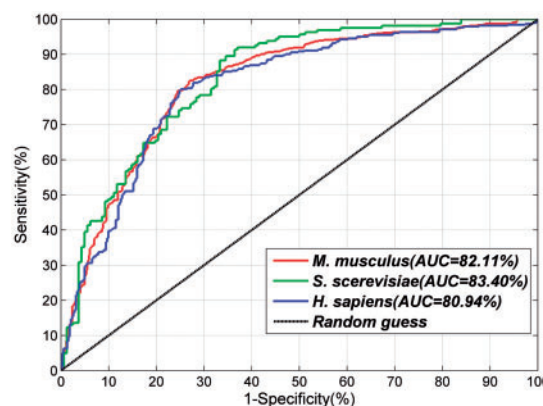
residues (Lys, Arg) than do the non-ubiquitylation sequences in the subsequent AAC results (Fig. 5). In short, this investigation fully proves that informative PCPs can contribute to distinguishing ubiquitylation sequences and non-ubiquitylation sequences.

### 3.4 AAC results

In this section, we will study the difference between the AAC surrounding ubiquitylation sites and those of non-ubiquitylation sites. In Figure 5, from left to right, the amino acids vary from being enriched to being depleted in the surrounding sequences of ubiquitylation sites. With slight variations among different species, the overall trends are similar. For ubiquitylated sequences, hydrophobic residues, Ala, Val, Leu and Ile, are enriched in the region surrounding ubiquitylation sites, whereas hydrophilic residues, Cys, His, Thr, Ser and Lys, are depleted. This result has been confirmed in a ubiquitylation site survey (Xu *et al.*, 2010). Furthermore, the charged amino acids Arg and Asp are enriched in the ubiquitylated sequences, and the uncharged residues Cys, Met, Trp, Thr, Ser and Lys are depleted. This investigation also implies that the different compositions of amino acids surrounding ubiquitylation sites and non-ubiquitylation sites justify the use of amino acid frequencies as a feature for ubiquitylation site prediction.

### 3.5 Using IG optimize the model

Three feature sets (KNN, PCP and AAC) as well as their combined feature sets were evaluated before using the IG method by means of 5-fold cross-validation. The results showed that the prediction performance was enhanced through combined features, as displayed in Supplementary Table S7. To further improve the prediction performance, these features were optimized based on the above-mentioned IG optimization method. In this work, we trained the model only using the top  $n$  ( $n = 7 \sim 20$ , 27) positions and the top  $m$  ( $m = 9 \sim 20$ ) amino acids. The results are displayed in Supplementary Figure S4. For comparison, we chose key positions and key amino acid residues (Supplementary Table S5) for different species that would be used to optimize model. The prediction performance of the optimized features was also evaluated as shown in Supplementary Table S7. The predictive accuracy for the models trained with the optimized features improved compared with the accuracy using the non-optimized features, as expected. This result indicates that



**Fig. 6.** ROC curves of UbiProber predictions of *H.sapiens*, *M.musculus* and *S.cerevisiae*. Each curve represents the average sensitivities and specificities for different thresholds over 5-fold cross-validation

the key positions and amino acid residues used in optimizing the models can promote the performance of ubiquitylation sites prediction. Consequently, the models combined with key positions and amino acid residues were applied to implement a novel and high-performance tool for identifying lysine ubiquitylation.

### 3.6 Species-specific Ubiquitylation site prediction of UbiProber

The proposed UbiProber predictor trained and tested three models of *H.sapiens*, *M.musculus* and *S.cerevisiae* based on the combined features and the IG method. To evaluate the performance of UbiProber for species-specific ubiquitylation site prediction, we performed a 5-fold cross-validation test in each species. Sensitivities at different specificity levels in each cross-validation run were calculated according to Equations (3) and (4). By taking different thresholds, we then plotted the ROC curves and calculated the AUCs, as shown in Figure 6.

The three models, *H.sapiens*, *M.musculus* and *S.cerevisiae* achieved AUCs of 80.94, 82.11 and 83.40%, respectively. We knew that the most confident predictions are those with high specificities. From the ROC curves, for the *H.sapiens*, *M.musculus* and *S.cerevisiae* models, the prediction sensitivities reached 37, 43 and 46% at the 90% specificity level, respectively, and 22, 22.5 and 27.5% at the 95% specificity level, respectively. Interestingly, from the ROC curves, the predictions for *S.cerevisiae* performed significantly better than those for *H.sapiens*. This result can be explained by the fact that KNN features in *S.cerevisiae* show stronger patterns in distinguishing positive and negative data, as shown in Figure 3.

### 3.7 Cross-species prediction of general Ubiquitylation sites

The performance of cross-species predictions using UbiProber was evaluated against three species. An independent test dataset was built for evaluation by randomly selecting 20 proteins from the non-redundant dataset of each species. UbiProber has built three prediction models for every species. We also built a combined prediction model using the data by combining the



**Table 1.** Cross-species prediction performance of UbiProber

Test	Model			
	<i>H.sapiens</i>	<i>M.musculus</i>	<i>S.cerevisiae</i>	Combined
<i>H.sapiens</i>	<b>78.23/67.25<sup>a</sup></b>	72.32/51.51	65.86/42.75	76.54/62.59
<i>M.musculus</i>	83.77/80.95	<b>88.55/76.87</b>	84.89/76.32	81.82/75.69
<i>S.cerevisiae</i>	89.87/81.78	81.98/68.54	<b>99.25/88.98</b>	88.35/80.91

<sup>a</sup>The two numbers in each cell represent the AUC (%) and MCC (%). The training model with the highest AUC for each test dataset is bolded.

three training datasets and running the non-redundant dataset building procedure with an identity threshold of 40%. The tests were then performed by submitting all ubiquitylation sequence segments, and the same number of non-ubiquitylation sequence segments in each of the three test datasets to each of the four prediction models. The AUCs and MCCs were calculated, as shown in Table 1. Note that there was no overlap between the training sets and the independent test sets.

For all test datasets, the prediction results from the model trained using data from the same species performed the best (considering only AUC). For each of the three independent test datasets, the performances did not have large variations using different models trained on data from different species. As a possible explanation, although ubiquitin-conjugating enzymes and their substrates vary in different species, the biophysical mechanism of enzyme–substrate binding remains the same. The results suggest that UbiProber and the associated prediction models can be used for cross-species predictions of general ubiquitylation sites, which is especially useful when the ubiquitylation data in a species of interest is not sufficient for training a prediction model. Interestingly, for cross-species predictions, there is no apparent evidence that using a model trained on data from an evolutionarily closer species would perform better. As an example, for the *M.musculus* test dataset, the predictions from the *S.cerevisiae* model performed better than those from other models, except the *M.musculus* model itself. Given the small test data size, this may not be statistically significant. The performance variations in various models may be partially due to different quantities and qualities of ubiquitylation data in different species.

3.8 Comparison with other general prediction tools

To further evaluate the performance of general ubiquitylation site prediction by UbiProber, we compared it with three existing web service tools, UbiPred, UbPred and CKSAAP\_UbSite. Because different training data and methods were used in these tools, the prediction performance varies greatly among them. To avoid any bias, here, we applied the same independent test dataset (containing the test data of three species) as that used in the cross-species prediction evaluation. Sequences of these 20 proteins of each species dataset and sequences of these 60 proteins of the *combined* dataset were submitted to UbiPred, UbPred and CKSAAP\_UbSite for prediction. As a comparison, the AUC results comparing the predictions of the different tools are shown in Table 2. UbiProber achieves significantly higher AUC

**Table 2.** Comparison of ubiquitylation prediction performances of UbiPred, UbPred, CKSAAP\_UbSite and UbiProber

Independent test datasets	Method			
	<i>UbiPred</i>	<i>UbPred</i>	<i>CKSAAP_UbSite</i>	<i>UbiProber</i>
<i>H. dataset</i>	58.62 <sup>a</sup>	59.62	53.57	<b>76.54</b>
<i>M. dataset</i>	46.16	64.40	62.73	<b>81.82</b>
<i>S. dataset</i>	40.36	73.57	42.56	<b>88.35</b>
<i>Combined<sup>b</sup></i>	52.26	66.33	53.58	<b>83.65</b>

Note: The method with the highest AUC for each test dataset is in bold format.

<sup>a</sup>AUCs (%) were calculated and compared.

<sup>b</sup>The combined dataset consists of the *H. dataset*, the *M. dataset* and the *S. dataset*.

values than those of the other tools. Depending on our *Combined* model, UbiProber improved the AUC by 16.92, 17.42, 14.78 and 17.32% in the *H. dataset*, *M. dataset*, *S. dataset* and *Combined* dataset, respectively. The data source of the experimental ubiquitylation sites for the above three existing tools was collected from UbiProt, which mainly stores only a few hundred yeast ubiquitylation proteins. However, UbiProber integrated the experimental ubiquitylation sites from multiple resources (Fig. 1), which accumulated the ubiquitylation data from thousands of proteins from multiple species. These resources could partially explain why the UbiProber outperforms other predictors. In UbiPred and CKSAAP\_UbSite, the low AUC can be partly attributed to the use of single feature, which cannot fully extract features for ubiquitylated sequences. In contrast, UbiProber integrated three features to ensure the complete extraction of the sequence information. In addition to the AUC value comparison, we also calculated the MCC, sensitivity and specificity for comparisons with other tools, as shown in Supplementary Table S8. The results show that the above three tools can achieve high specificity, but sacrificing sensitivity would therefore result in a low MCC. UbiProber offers good sensitivity as well as high specificity, which also illustrates that our method is superior to the current methods. Note that, when performing the comparisons, we used a prediction model that was trained from a dataset excluding the protein sequences in the independent test dataset. However, for UbiPred, UbPred and CKSAAP\_UbSite, some of the test proteins might have been included in their training processes, and thus, the prediction performances may be biased favorably toward these tools in the comparisons. This possibility means that the performance improvement of UbiProber over these tools might be underestimated.

3.9 Software and web implementation

We have developed a stand-alone software system and web server, UbiProber, to implement the described ubiquitylation site prediction method. Currently, the software system and web server of UbiProber are implemented as a Windows application in the .NET 4.0 framework using the C# language. The web server and software system always share the same function. Supplementary Figure S5 shows an example of the prediction interfaces of the software system and web server. The software

system and web server, along with their user manual, are freely available at <http://bioinfo.ncu.edu.cn/UbiProber.aspx>.

### 3.10 Limitations and future work

Our method obtained a satisfactory success rate on large data sources, totaling 11 547 proteins, and also indicates that the key position and key amino acid residue features are promising for annotating ubiquitylation events in proteomes. In spite of this, our method still has certain limitations, which are common to almost all current prediction tools. First, although computational predictions indicate the possibilities that query sites can/cannot be ubiquitylated, our predicted results have not been correlated to different cell states or tissue conditions. We are trying to develop some novel models of ubiquitylation for specific cellular conditions, which are sparse in current ubiquitylation studies. Second, the ubiquitylation sites used in the training data were mostly identified by MS methods, which may have an inherent bias in terms of representing the global ubiquitylation events and hence may affect the prediction performance. As techniques such as electron transfer dissociation and alternative proteases are helping to resolve technology limitations, more complete ubiquitylation datasets will be released. We will adapt our program and prediction models as the new data become available. Third, a data limitation is that we have only labeled positive data, but we have not labeled negative data (i.e. it would be difficult to prove definitively that a particular lysine residue is not ubiquitylated under any conditions.). Therefore, if some of them are predicted to be ubiquitylation sites, we do not know whether they are false positives. For future work, we will try to develop some novel methods that do not require a negative data or only require few solid negative data (such as semi-supervised learning). In this way, it is expected that the bias caused by the negative data can also be reduced.

## 4 CONCLUSION

The annotation of protein ubiquitylation sites in proteomes is a crucial step in decoding the signaling networks in living cells. In recent years, tens of thousands of ubiquitylation sites in various species have been identified by large-scale MS-based studies. However, the vast majority of ubiquitylation sites, especially in non-mammal species, still remain undiscovered. Considering the limitations of MS-based experimental studies, a more practical and efficient approach will be *in silico* large-scale ubiquitylation site prediction. In this work, we have developed a novel computational method and computer program, UbiProber, for the highly specific prediction of ubiquitylation sites. Cross-validation tests and comparisons with other tools show that UbiProber performs better at general predictions and is comparable with existing methods at species-specific predictions. Unlike experimental approaches, computational predictions are capable of proteome-wide predictions without inherent technical biases. Furthermore, UbiProber could provide an even more powerful and cost-effective approach by combining experimental and computational methods iteratively, which could be especially useful for some hypothesis-driven experiments. In conclusion, UbiProber provides a unique tool for large-scale ubiquitylation site identification, and it is hoped that UbiProber will accelerate

the accumulation of our knowledge on protein ubiquitylation and hence help to explore the corresponding regulatory networks in living cells.

## ACKNOWLEDGEMENTS

We are thankful to the associate editor, Jonathan Wren, and two anonymous reviewers, whose suggestions have greatly improved the quality of this manuscript, and we also would like to thank Predrag Radivojac, Matthias Mann and Florian Gnäd for helpful discussions.

**Funding:** This work was supported by Program for New Century Excellent Talents in University (NCET-11-1002) and the National Natural Science Foundation of China (20605010 and 21175064).

**Conflict of Interest:** none declared.

## REFERENCES

- Cai, Y. *et al.* (2012) Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino Acids*, **42**, 1387–1395.
- Catic, A. *et al.* (2004) Preferred *in vivo* ubiquitination sites. *Bioinformatics*, **20**, 3302–3307.
- Chen, K. *et al.* (2009) Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. *J. Comput. Chem.*, **30**, 163–172.
- Chen, Z. *et al.* (2011) Prediction of Ubiquitination sites by using the composition of k-spaced amino acid pairs. *PLoS One*, **6**, e22930.
- Chernorudskiy, A. L. *et al.* (2007) UbiProt: a database of ubiquitylated proteins. *BMC Bioinformatics*, **8**, 126.
- Gabaldon, T. and Huynen, M. A. (2004) Prediction of protein function and pathways in the genome era. *Cell. Mol. Life Sci.*, **61**, 930–944.
- Gao, J. J. *et al.* (2010) Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol. Cell Proteomics*, **9**, 2586–2600.
- Haglund, K. and Dikic, I. (2005) Ubiquitylation and cell signaling. *EMBO J.*, **24**, 3353–3359.
- Henikoff, S. and Henikoff, J. G. (1992) Amino-acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.
- Hitchcock, A. L. *et al.* (2003) A subset of membrane-associated proteins is ubiquitinated in response to mutations in the endoplasmic reticulum degradation machinery. *Proc. Natl. Acad. Sci. USA*, **100**, 12735–12740.
- Ikeda, F. and Dikic, I. (2008) Atypical ubiquitin chains: new molecular signals. 'Protein modifications: Beyond the usual suspects' review series. *EMBO Rep.*, **9**, 536–542.
- Kawashima, S. and Kanehisa, M. (2000) AAindex: amino acid index database. *Nucleic Acids Res.*, **28**, 374–374.
- Kim, W. *et al.* (2011) Systematic and quantitative assessment of the Ubiquitin-modified proteome. *Mol Cell*, **44**, 325–340.
- Lee, T. Y. *et al.* (2010) N-Ace: using solvent accessibility and physicochemical properties to identify protein N-acetylation sites. *J. Comput. Chem.*, **31**, 2759–2771.
- Lee, T. Y. *et al.* (2011) Incorporating distant sequence features and radial basis function networks to identify Ubiquitin conjugation sites. *PLoS One*, **6**, e17331.
- Lee, W. C. *et al.* (2008) SCUD: *Saccharomyces cerevisiae* Ubiquitination database. *BMC Genomics*, **9**, 440.
- Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
- Mani, A. and Gelmann, E. P. (2005) The ubiquitin-proteasome pathway and its role in cancer. *J. Clin. Oncol.*, **23**, 4776–4789.
- Moses, A. M. *et al.* (2007) Clustering of phosphorylation site recognition motifs can be exploited to predict the targets of cyclin-dependent kinase. *Genome Biol.*, **8**, R23.
- Niu, S. *et al.* (2010) Prediction of tyrosine sulfation with mRMR feature selection and analysis. *J. Proteome Res.*, **9**, 6490–6497.
- Peng, J. M. *et al.* (2003) A proteomics approach to understanding protein ubiquitination. *Nat. Biotechnol.*, **21**, 921–926.



- Pickart,C.M. and Eddins,M.J. (2004) Ubiquitin: structures, functions, mechanisms. *BBA Mol. Cell Res.*, **1695**, 55–72.
- Radivojac,P. et al. (2010) Identification, analysis, and prediction of protein ubiquitination sites. *Proteins*, **78**, 365–380.
- Reinstein,E. and Ciechanover,A. (2006) Narrative review: protein degradation and human diseases: the ubiquitin connection. *Ann. Intern. Med.*, **145**, 676–684.
- Shimizu,Y. et al. (2010) Ubiquitylation of an ERAD substrate occurs on multiple types of amino acids. *Mol Cell*, **40**, 917–926.
- Tomlinson,E. et al. (2007) Methods for the purification of ubiquitinated proteins. *Proteomics*, **7**, 1016–1022.
- Tung,C.W. and Ho,S.Y. (2008) Computational identification of ubiquitylation sites from protein sequences. *BMC Bioinformatics*, **9**, 310.
- Vapnik,V.N. (1999) An overview of statistical learning theory. *IEEE. Trans. Neural Netw.*, **10**, 988–999.
- Wagner,S.A. et al. (2011) A proteome-wide, quantitative survey of *in vivo* ubiquitylation sites reveals widespread regulatory roles. *Mol. Cell Proteomics*, **10**, M111.013284.
- Whisstock,J.C. and Lesk,A.M. (2003) Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.*, **36**, 307–340.
- Xu,G. et al. (2010) Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. *Nat. Biotechnol.*, **28**, 868–873.