A method to distinguish between lysine acetylation and lysine ubiquitination with feature selection and analysis

You Zhou\textsuperscript{a,1}, Ning Zhang\textsuperscript{b,1}, Bi-Qing Li\textsuperscript{c}, Tao Huang\textsuperscript{a,\textasteriskcentered 2}, Yu-Dong Cai\textsuperscript{c} and Xiang-Yin Kong\textsuperscript{a}\textsuperscript{**}

\textsuperscript{a}The Key Laboratory of Stem Cell Biology, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences and Shanghai Jiao Tong University School of Medicine, Shanghai 200031, P.R. China; \textsuperscript{b}Department of Biomedical Engineering, Tianjin Key Lab of BME Measurement, Tianjin University, Tianjin 300072, P.R. China; \textsuperscript{c}Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, P.R. China;

Communicated by Ramaswamy H. Sarma

(Received 28 September 2014; accepted 18 December 2014)

Lysine acetylation and ubiquitination are two primary post-translational modifications (PTMs) in most eukaryotic proteins. Lysine residues are targets for both types of PTMs, resulting in different cellular roles. With the increasing availability of protein sequences and PTM data, it is challenging to distinguish the two types of PTMs on lysine residues. Experimental approaches are often laborious and time consuming. There is an urgent need for computational tools to distinguish between lysine acetylation and ubiquitination. In this study, we developed a novel method, called DAUFSA (distinguish between lysine acetylation and lysine ubiquitination with feature selection and analysis), to discriminate ubiquitinated and acetylated lysine residues. The method incorporated several types of features: PSSM (position-specific scoring matrix) conservation scores, amino acid factors, secondary structures, solvent accessibilities, and disorder scores. By using the mRMR (maximum relevance minimum redundancy) method and the IFS (incremental feature selection) method, an optimal feature set containing 290 features was selected from all incorporated features. A dagging-based classifier constructed by the optimal features achieved a classification accuracy of 69.53%, with an MCC of .3853. An optimal feature set analysis showed that the PSSM conservation score features and the amino acid factor features were the most important attributes, suggesting differences between acetylation and ubiquitination. Our study results also supported previous findings that different motifs were employed by acetylation and ubiquitination. The feature differences between the two modifications revealed in this study are worthy of experimental validation and further investigation.

\textbf{Keywords:} ubiquitination; acetylation; post-translational modification; dagging; incremental feature selection

1. Introduction

Lysine acetylation and ubiquitination are two primary post-translational modifications (PTMs) in most eukaryotic proteins (Grant, 2001; Kessler, 2013). Since its discovery on histone proteins in the 1960s (Vidal, Gershey, & Allfrey, 1968), lysine acetylation has been one of the most studied modifications due to its roles in a wide repertoire of biological processes, including chromatin stability, cell cycle control, and energy metabolism (Lee et al., 2010; Rice & Allis, 2001; Shubassi, Robert, Vanoli, Minucci, & Foiani, 2012; Xiong & Guan, 2012). More than 80% of human proteins are acetylated at the N-terminus (Van Damme, Arnesen, & Gevaert, 2011). This reversible modification is catalyzed by lysine acetyltransferases (KATs, also known as histone acetyltransferases (HATs)) that transfer the acetyl group of acetyl-CoA to the epsilon-amino group of the lysine residue for both nuclear and cytoplasmic proteins (Yang & Seto, 2008). Lysine deacetylases (KDACs, also known as histone deacetylases (HDACs)) remove the acetyl groups (Cai & Lu, 2008). Accumulating evidence has shown that lysine acetylation is essential for cellular development, and its dysregulation can lead to serious diseases such as cancer (Allfrey, Faulkner, & Mirsky, 1964; Choudhary et al., 2009; Xiong & Guan, 2012).

Lysine ubiquitination is another reversible PTM that has been implicated in various cellular processes, such as protein degradation, DNA repair and replication, transcriptional regulation, and intracellular trafficking (Alonso & Friedman, 2013; Chen et al., 2011; Kessler, 2013; Micel, Tintel, Smith, & Eckhardt, 2013). It is a highly collaborative process, in which an isopeptide bond covalently forms between the C-terminus of the activated ubiquitin protein and the epsilon-amino group of the lysine

*Corresponding authors. Email: huangtao@sibs.ac.cn (T. Huang), caiyudong@staff.shu.edu.cn (Y.-D. Cai), xykong@sibs.ac.cn (X.-Y. Kong)

\textsuperscript{1}These authors contributed equally to this work.
residue of a substrate protein. The reaction is catalyzed by three enzymes acting in tandem: a ubiquitin-activating enzyme (E1), ubiquitin-conjugating enzyme (E2), and ubiquitin ligase (E3) (Haglund & Dikic, 2005; Radijovic et al., 2010). Multiple studies have demonstrated that ubiquitination is a significant and widespread PTM and is involved in diverse cellular activities. The aberration of this process is associated with numerous pathological diseases, including inflammatory diseases and cancers (Kessler, 2013; Micel et al., 2013).

Although previous research primarily focused on single PTMs, a growing number of studies have discovered the co-regulation of different types of PTMs within and between a number of proteins. This co-regulation is crucial for the control of important cellular processes and for understanding the cell’s protein regulatory networks (Kontaki & Talianidis, 2010; Minguez et al., 2012). The co-evolution of various PTMs also indicates their vital roles in regulating multiple protein functional states (Minguez et al., 2012; Minguez, Letunic, Parca, & Bork, 2013). As the same, lysine residue can undergo different PTMs, some interplay exists among them. For example, previous studies have revealed that a lysine residue in the RNase III enzyme Drosha has opposing roles during acetylation and ubiquitination (Tang et al., 2013); this effect was verified for the same residue through large-scale mass spectrometric analysis (Danielsen et al., 2011). Recent studies have identified crosstalk between acetylation and ubiquitination in cell cycle regulation (Liu et al., 2013). Considering their functional associations with chromatin regulation and protein synthesis, there is a need to understand the mechanisms of lysine acetylation and ubiquitination. Identification of protein acetylation and ubiquitination sites is fundamental for elucidating their modification dynamics and molecular principles.

The large-scale identification of acetylated and ubiquitylated proteins using traditional experimental approaches, including high-throughput mass spectrometry (Basu et al., 2009; Kirkpatrick, Denison, & Gygi, 2005), specific antibodies (Gentry, Worby, & Dixon, 2005), and Chip-on-Chip (Johnson et al., 2008), is often labor intensive and time consuming. Computational methods have been applied to predict potential modification sites in query proteins (Cai & Lu, 2008; Cai et al., 2012; Chen et al., 2011; Gnadt, Ren, Choudhary, Cox, & Mann, 2010; Lee et al., 2010; Li, Xue, Jin, Wang, & Yao, 2006; Li et al., 2009; Radijovic et al., 2010; Tung & Ho, 2008; Xu, Wang, Ding, Wu & Deng, 2010; Zhao, Li, Ma, & Yin, 2011). Most of these methods are carried out as support vector machine (SVM) classifiers based on sequence-derived features, providing convenience, and high efficiency. Li et al. introduced a coding scheme for protein sequences by coupling patterns into the SVM method to predict lysine acetylation, which achieved superior predictive accuracy (Li et al., 2009). Lee et al. incorporates solvent accessibility and physicochemical properties of proteins to identify acetylated sites, in which a two-stage SVM was applied to learn the computational models (Lee et al., 2010). Tung and Ho built a predictor for putative ubiquitylated sites using an SVM classifier based on mining informative physicochemical properties from protein sequences (Tung & Ho, 2008). Radijovic et al. developed a random forest-based predictor by utilizing sequence attributes as the input feature vector (Radijovic et al., 2010). Cai et al. proposed a nearest neighbor algorithm-based ubiquitation site predictor and analyzed key features that played important roles in the process of ubiquitination (Cai et al., 2012). Chen et al. developed an SVM-based predictor for ubiquitination sites by employing the composition of k-spaced amino acid pairs surrounding the modified site as input that was limited to the yeast proteome (Chen et al., 2011). Qiu et al. also proposed iUbq-Lys for lysine ubiquitination site prediction (Qiu, Xiao, Lin, & Chou, 2014). However, these methods in prior studies were developed only for predicting one type of PTM (either acetylation or ubiquitination) and revealing these modification characteristics, such as consensus motifs and catalytic mechanisms.

In this study, we developed an in silico framework to discriminate acetylation sites from ubiquitination sites. As shown by a recent publication (Chen, Feng, Lin, & Chou, 2013) and proposed by Chou in (2011), to provide a useful statistical analysis for a biological system, the following procedures must be clearly reported: (a) constructing or selecting a valid benchmark data-set to train and test; (b) formulating the biological samples with an effective mathematical expression that reflects their intrinsic correlation with the target to be analyzed; (c) introducing or developing a powerful algorithm to operate the computation; and (d) properly performing cross-validation tests to objectively evaluate the anticipated accuracy.

2. Methods
2.1. Data-set
The sequences of the acetylated and ubiquitylated proteins used in this study were downloaded from the UniProt database (http://www.uniprot.org/, release 2013_07). Data with non-experimentally modified residues were removed. We also removed proteins with sequence identities >40%. To remove homologous sequences from a benchmark data-set, a cut-off threshold of 25% was suggested in the literature such as (Lin, Fang, Xiao, & Chou, 2013), by which proteins having 25% or more than 25% sequence identity must be removed. However, the data-set of known ubiquitination sites is small. Therefore, in this study, we did not use such a stringent criterion, or too few ubiquitination sites would be left to perform this study. Instead, the threshold 40% was used in this study. The remaining 1801 proteins contained
2751 acetylation sites and 428 ubiquitination sites. The data-set is provided in Supporting Information S1.

We applied a sliding window strategy to extract positive and negative peptide samples. In this study, we adopted 21 as the window length; that is, we extracted 21-residue peptide segments centered either on an ubiquitinated lysine or on an acetylated lysine residue, with 10 residues upstream and 10 residues downstream from the centered lysine. Peptide segments with lengths less than 21 were complemented by the character ‘X’, denoting blanks in the peptides. Peptides with a centered, ubiquitinated lysine were regarded as positive samples, while peptides with a centered, acetylated lysine were negative. Accordingly, 428 positive and 2751 negative samples were extracted.

The data-set was unbalanced, with a greater number of acetylation samples. Therefore, we randomly split the set of acetylation samples into five parts without overlap, with each part having 550, 551, 550, 550, and 550 acetylation samples, respectively. In turn, all ubiquitination samples were combined with one part of the acetylation samples. At each epoch, one data-set was generated combining all the ubiquitination samples and one part of the acetylation samples. In total, five data-sets were obtained, which were named Data-set 1, Data-set 2, Data-set 3, Data-set 4, and Data-set 5, respectively.

2.2. Feature extraction

We used the following features to encode all 21-residue peptides, both for the positive and negative samples.

2.2.1. Features of PSSM conservation scores

In biological sequence analyses, the evolutionary conservation from multiple sequence alignments is widely considered important (Jiang et al., 2013). A conserved residue can be under strong selective pressure and hence, play a vital role in protein structure and function. In this study, the position-specific iterative BLAST (PSI-BLAST) (Altschul et al., 1997), a powerful sequence search method, was applied to calculate the conservation status of every residue in a peptide. PSI-BLAST searched the UniRef100 database (Release: 15.10, 03-Nov-2009) through three iterations with .0001 as the E-value cut-off. For one residue in a peptide, a 20-dimensional vector can be computed, reflecting probabilities of the residue against its mutations to the 20 standard native amino acids. All such 20-dimensional vectors for all 21 residues in a 21-mer peptide composed a $20 \times 21$ matrix, called a position-specific scoring matrix (PSSM). The $20 \times 21 = 420$ values in the matrix, called PSSM conservation scores, were used in this study as features in the construction of our classifiers.

2.2.2. Features of amino acid factors

Different amino acids have different physicochemical and biochemical properties. Proteins with different amino acid compositions, thus, have different physicochemical and biochemical properties, which could affect protein structure and function. Atchley, Zhao, Fernandes, and Druke (2005) performed multivariate statistical analyses on the AAindex (Kawashima & Kanehisa, 2000) database and generated five numerical patterns to reflect an amino acid’s physicochemical and biochemical properties. These were codon diversity, electrostatic charge, molecular volume, polarity, and secondary structure. Herein, we used five numerical scores, called amino acid factor features, as another type of feature to encode each residue in a 21-mer peptide to construct our classification model.

As the centered residue in a 21-mer peptide was always lysine, it was not necessary to incorporate the centered lysine. Only the 20 surrounding residues needed to be incorporated. Therefore, there were only $5 \times 20 = 100$ amino acid factor features for a 21-mer peptide.

2.2.3. Features of secondary structures

Protein secondary structures can play important roles in protein post-translational modifications (Shi et al., 2012). The peptide residue secondary structure states were employed as another feature type in the classifier construction. In this study, we used SSpro4 (Cheng, Randall, Sweredsoki, & Baldi, 2005), a commonly used secondary structure prediction method, to predict one of three different states (‘helix,’ ‘strand,’ or ‘other’) for every residue. To transform the secondary structure states to numeric features, the three states were denoted as a 3-bit binary value (‘100,’ ‘010,’ and ‘001,’ respectively). One 3-bit binary value could be regarded as three numeric features, although every feature value was either 0 or 1. For example, ‘100’ can be regarded as three numeric features, 1, 0, and 0. Therefore, there were $3 \times 21 = 63$ secondary structure features for a 21-mer peptide.

2.2.4. Features of solvent accessibilities

Residue solvent accessibilities were taken into account as previous studies demonstrated (Trouillas, Berges, & Houee-Levin, 2011) that they were important in residue modifications. We used SSpro4 (Cheng et al., 2005) to compute the solvent accessibilities of every residue in a 21-mer peptide. SSpro4 can provide information on whether each residue is in a ‘buried’ or ‘exposed’ state. We transformed the two different solvent accessibility states to 2-bit binary values: ‘buried’ as ‘10’ and
‘exposed’ as ‘01’. Therefore, there were $2 \times 21 = 42$ solvent accessibility features for a 21-mer peptide. These 42 peptide features were used in the model constructions.

2.2.5. Feature of disorder scores
Disordered regions are of great importance to residue modifications (Ferron, Longhi, Canard, & Karlin, 2006; Noivirt-Brik, Prilusky, & Sussman, 2009; Zhang, Li, Gao, Ruan, & Cai, 2012). If a protein region is devoid of stable structure or if it has a large number of conformations, it is called a ‘disordered region.’ The probabilities of each residue forming a disordered peptide structure were predicted by the software VSL2 (Peng, Radivojac, Vucetic, Dunker, & Obradovic, 2006). VSL2 provides a disorder score for each peptide residue. The higher the score, the more likely the residue forms a disordered structure. The disorder scores of all 21 residues in a 21-mer peptide were computed by VSL2 and were used in the construction of our classifiers.

Features utilized in this study are summarized in Table 1. For a 21-amino acid peptide, there were 420 PSSM conservation score features, 100 amino acid factor features, 63 secondary structure features, 42 solvent accessibility features, and 21 disorder score features. A total of 646 features were extracted for every 21-length peptide. This method was similar to that used in (Zhang et al., 2012) for predicting protein γ-carboxylation sites and that used in (Jiang et al., 2013) for predicting protein pyruvoyl-serine sites.

| Feature type            | Features                                           | Number |
|-------------------------|----------------------------------------------------|--------|
| PSSM conservation scores| 20-dimensional vector                              | 420    |
| Amino acid factors      | Polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge (only for surrounding sites, except the center) | 100    |
| Secondary structures    | Secondary structures: helix, strand, and other      | 63     |
| Solvent accessibilities | Solvent accessibilities: buried and exposed        | 42     |
| Disorder scores         | Disorder score reflecting the disorder status of the residue | 21     |
| Total                   |                                                    | 646    |

2.3. Feature selection
The 646 features are not equally important. Less important features should not be used to construct the classification model; otherwise, they can reduce the model performance. The mRMR (maximum relevance minimum redundancy) method (Peng, Long, & Ding, 2005; Li, Hu et al., 2012; Li, Huang, Liu, Cai, & Chou, 2012) was used to rank the importance of the 646 features, from most to least important. This method is based on the criteria of maximum relevance minimum redundancy. The former criterion selects features most related to the target, while the latter criterion excludes features containing redundant information among the already selected features. For details of the mRMR method, please refer to (Jiang et al., 2013; Li, Hu et al., 2012; Li, Huang, et al., 2012; Peng et al., 2005; Zhang et al., 2012).

By using this method, an ordered list was obtained ranking the 646 features from most to least important called the ‘mRMR table.’ In this table, features with smaller indices indicated that they had a better trade-off between the maximum relevance and the minimum redundancy and thus, were important features. Based on the ordered feature list, a series of classifiers can be constructed using different features. For example, using only the top feature from the list, a classifier can be constructed. By using the top two features from the list, another classifier can be constructed, and so on. The classifier created in the next round always contains one more feature from the ranked list following the previous one. In this procedure, features in the ranked feature list were added one by one from higher to lower rank. A new feature set was generated when another feature was added, until all features were finally added. For 646 features in the list, a total of 646 classifiers are constructed. This procedure is called the IFS (incremental feature selection) method (He et al., 2010; Huang et al., 2009). The 646 classifiers constructed would use the first feature, the first two features, the first three features, and so on, until all 646 features were used from the ranked feature list. From the 646 classifiers, we can select the best one to discriminate the two modifications that achieves the best performance. The features used by that classifier were regarded as composing the optimal feature set.

2.4. Prediction methods
In this study, dagging (Ting & Witten, 1997) was employed to construct classifiers. Dagging is a meta-classifier that creates a number of disjointed and
stratified folds out of the data and feeds each data fold into a single learning algorithm called the ‘base classifier.’ For a query sample, each of these classifiers provides an output. The final predicted result is the class with the most votes. In this study, ‘Dagging’ in Weka 3.6.4 (Hall et al., 2009) was implemented, and sequential minimal optimization (SMO) was used as the base classifier. SMO is an algorithm used for solving the quadratic programming problem for training SVMs. It was run with default parameters.

The method developed in this study was called DAUFSA (distinguish between lysine acetylation and lysine ubiquitination with feature selection and analysis).

### 2.5. Performance measurements

In statistical prediction, the following three cross-validation methods are often used to examine a machine learning approach: an independent data-set test, a subsampling test or $K$-fold cross-validation test, and a jackknife test. However, of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark data-set as elaborated in Chou (2011), and demonstrated by Equations (28)–(30) in the same review article (Chou, 2011). Accordingly, the jackknife test has been increasingly used and widely recognized by investigators to examine the various predictor qualities (Du, Jiang, He, Li, & Chou, 2006; Shen & Chou, 2010; Xu et al., 2014).

To reduce the computational time, we adopted the 10-fold cross-validation in this study, as previously used by many investigators with the prediction SVM engine. In the 10-fold cross validation, each of the five combined datasets is divided into 10 equal parts. Nine parts are used for training and the remaining one part for testing. This procedure is repeated 10 times in turn, and the final prediction accuracy is the average accuracy.

The following measurements were used in this study:

$$
\text{ACC} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \times 100\% \\
\text{Sensitivity} = \frac{TP}{(TP + FN)} \times 100\% \\
\text{Specificity} = \frac{TN}{(TN + FP)} \times 100\%
$$

(1) (2) (3)

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

(4)

In measurements 1–4, TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. Peptides with a centered, ubiquitinated lysine were regarded as positive samples, while peptides with a centered, acetylated lysine were regarded as negative samples.

Sensitivity is the percentage of positive, ubiquitylated samples that are correctly classified to be positive by the method. Specificity is the percentage of negative, acetylated samples that were correctly classified to be negative. Values <100% in sensitivity and specificity reflect the occurrence of false-negative and false-positive errors of the method, respectively.

Matthews correlation coefficient (MCC) (Matthews, 1975) is a single value, but robust a performance measurement. The MCC value ranges from $-1.0$ to $+1.0$, where $0$ represents a random correlation between the classified variables and the actual variables, $+1.0$ a perfect correlation, and $-1.0$ a perfect negative correlation (Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000; Petersen, Lundegaard, & Petersen, 2010; Vihinen, 2012). MCC takes both false-positive and false-negative errors into account and is generally deemed to be a balanced measurement, even if the classes vastly differ in size (Shi et al., 2012). For these reasons, MCC is more reliable than accuracy, as previously reported (Jiang et al. 2013; Shi et al., 2012; Zhang et al., 2012, 2014); thus, MCC was used throughout this study as the main evaluator.

### 3. Results

#### 3.1. The mRMR and IFS results

We generated five data-sets, each containing all the ubiquitination samples and one part of the five split data-sets of acetylation samples. For each data-set, an individual mRMR and IFS procedure was performed.

We obtained five mRMR tables, one for each data-set. In each table, the 646 features are ordered according to their importance for the classification. The five mRMR tables are provided as Supporting Information S2.

For each data-set, from the corresponding mRMR table, the ranked features were added one by one to generate 646 different feature sets, and accordingly, 646 individual classifiers were constructed. All classifier performances can be found in Supporting Information S3. We also plotted the MCC values of these classifiers in Figure 1 as IFS curves. From Figure 1 and the data in Supporting Information S3, it can be seen that performances of different classifiers adopting the same features differ slightly between data-sets. In data-set 1, the classifier adopting the top 137 features performed the best,
with an MCC of .3268. In data-set 2, the best classifier used 214 features, yielding an MCC of .3453. In data-set 3, the MCC reached the maximum .3598 when a classifier was constructed using the top 102 features. In data-set 4, the classifier adopting the top 39 features performed the best, with an MCC of .3353. In data-set 5, the best classifier used 66 features, yielding an MCC of .3853. We summarized the best MCC values and the corresponding measurements (i.e. SN, SP, and ACC) in Table 2. A total of 137, 214, 102, 39, and 66 features were regarded as composing the five optimal feature sets for the five data-sets, respectively. The five optimal feature sets details can be found in the mRMR tables in Supporting Information S2.

Table 2 indicates that the optimal features were capable of distinguishing between the two PTMs: lysine ubiquitination and lysine acetylation. Features in the selected optimal feature sets reflected differences in the properties and governing factors of the two types of PTMs. Analysis of the features may contribute to understand the mechanisms of the two modifications.

3.2. The combined optimal feature set

We combined the features in the five optimal feature sets excluding duplicates and obtained 290 optimal features, which can be found in Supporting Information S4. Our subsequent analysis was based on the assumption that if acetylation and ubiquitination used similar mechanisms, they would prefer similar factors that govern the formation of the modifications. By analogy, features that can optimally distinguish the two modifications were good candidates for analyzing the two modification mechanisms.

The type distributions of the five optimal feature sets computed from the five data-sets are depicted in Figure 2(A–E). Although the numbers of optimal features vary greatly between different data-sets (Table 2), the distributions are similar. Therefore, we can combine all the optimal features from the five data-sets to analyze the differences between acetylation and ubiquitination.

The distribution of the 290 optimal features and their corresponding selection ratios are depicted in Figure 2(F). A total of 192 of the 290 optimal features belong to the PSSM conservation score, 66 to the amino acid factor, 10 to the solvent accessibility, 18 to the secondary structure, and 4 to the disorder. The ratios of selected/total features for each feature type were 192/420 = 46% (PSSM conservation score); 66/100 = 66% (amino acid factor); 10/42 = 24% (solvent accessibility); 18/63 = 29% (secondary structure); and 4/21 = 19% (disorder). Interestingly, PSSM occupied more than half of the optimal features, followed by the amino acid factor, suggesting their dominant roles in discriminating acetylation from ubiquitination.

4. Discussion

4.1. Analysis of optimal features

4.1.1. PSSM conservation score features

As PSSM conservation score features accounted for most of the optimized features, we investigated the subtype distribution (i.e. the PSSM conservation scores) against mutations in 20 amino acids, with the results shown in Figure 3(A). Mutations to W (tryptophan) and C (cysteine) had the greatest effect (more than 15 features for each mutation). Previous studies have emphasized the

![Figure 1](image-url)  
Figure 1. The IFS curves for the five data-sets. Plots of the performances of the different classifiers, with MCC as the main measurement denoted on the y-axis. The number of classifier features used is denoted in the x-axis.
importance of W and C on their involvement in protein–protein interactions and structural integrity maintenance (Benezra, 1994; Chen et al., 2012; Komla-Soukha & Sureau, 2006; Meitzler, Hinde, Banfi, Nauseef, & Ortiz de Montellano, 2013). Given that acetylation and ubiquitination are both enzymatic and reversible modifications, it can be inferred that they occupied different modes of action, in which W and C behaved differently. Furthermore, ubiquitination is associated with an abundance of negatively charged and polar amino acids, and the depletion of hydrophobic residues around the modified sites (Radivojac et al., 2010), while acetylated sites prefer positively charged and hydrophobic residues (Hou et al., 2014; Huang et al., 2013; Yang, 2004). Consistent with our findings, charged amino acids and residues with different hydrophobic properties also had more PSSM features, especially K (lysine), H (histidine), and M (methionine), which ranked among the top 15 in our mRMR feature list. These results indicate the vital effect of these features on the differences between acetylation and ubiquitination.

4.1.2. Amino acid factor features

Another important factor that contributed to distinguishing acetylation and ubiquitination was amino acid factor features, although they had less contribution than PSSM. The number of each type in the 290 optimal feature set is depicted in Figure 3(C). There were few differences among them, suggesting that these properties had comparable impacts on the discrimination of the two PTMs. From our mRMR feature list, it can also be seen that almost half of the top 10 features were composed of codon diversity, molecular volume, secondary structure, electrostatic charge, and polarity, especially at sites 9 and 10 (Supporting Information S2). As both acetylation and ubiquitination had a bias for the amino acid utilization similar to the PSSM conservation score features, we propose that different characteristics of each amino acid exert diverse influences on the modified site selection. Additionally, it was reported that ubiquitylated sites did not tend to cluster, likely, due to the structural constraints that would prevent two or more close ubiquitin
molecules from attaching to the same substrate simultaneously (Radivojac et al., 2010). This stressed the importance of secondary structure as supported by our results.

4.1.3. Other features
An increasing number of studies have indicated that unstructured regions can make it easier for proteins to fit into modifying enzymes’ catalytic sites (Pang, Hayen, & Wilkins, 2007). Our recent work demonstrated that enzyme-catalyzed modifications, such as acetylation and ubiquitination, are preferentially located in disordered structures (Cai et al. 2012; Huang et al., 2013). Thus, non-regular coiled structures (‘other’) were not sufficient to make a distinction between acetylation and ubiquitination. In contrast, regular structures (‘strand’) were mainly responsible for this discrimination, as most features were selected from this type (Figure 3(B)). These findings could also explain why disordered features accounted for the lowest number of hits in the optimal feature set (Figure 2). Some studies have suggested that the structure of the histone tail may be altered by acetylation through increased alpha-helical content (Rice & Allis,

Figure 3. Subtype distributions of the 290 combined optimal features. Histogram to show (A) Distribution of subtypes of conservation score features; (B) Distribution of subtypes of secondary structure features; (C) Distribution of amino acid subtype factor features; and (D) Distribution of solvent accessibility subtype features.
Combined with Figure 3(B), the ‘helix’ structure accounted for the lowest number of features, and thus, it could be inferred that ubiquitination had the preference of ‘helix’ to a similar extent as acetylation. Figure 3(D) shows that exposed solvent accessibility sites had more features than buried sites, indicating their importance in distinguishing acetylation from ubiquitination. As mentioned above, acetylation preferred hydrophobic residues, which were absent from ubiquitylated sites. This finding suggested that site exposure might be a key rule for discerning acetylation and ubiquitination.

4.2. Site distribution of optimal features

To investigate whether there was a specific pattern around modified sites, we analyzed the site distribution of the 290 optimal features (Figure 4).

Strikingly, sites 8 and 9 stood out with the highest number of features, indicating their crucial roles in the discrimination of acetylation and ubiquitination. Meanwhile, many PSSM conservation score features also resided at sites 8 and 9, underlining the considerable importance of the conservation status of these two sites in distinguishing these two modifications. Figure 4 shows that amino acid factor features were evenly distributed among all sites except site 11, which was exactly modified, showing the variability among the surrounding sites between acetylation and ubiquitination.

Notably, the few disordered features resided at sites 8, 10, 13, and 19 (Figure 4). Although acetylation and ubiquitination preferentially occur in regions that are intrinsically disordered (Pang et al., 2007), they could still be separated by different structures at those sites. From Figure 4, it could also be concluded that adjacent sites seemed to be more important, regardless of what types and numbers of features existed, compared to sites more distant from the central modified lysine site (here referred to as site 11). In light of distinguishing acetylation from ubiquitination, more attention should be paid to sites that are located in close proximity to the central modification site.

4.3. Occurrence frequencies of amino acids

The compositional biases of 20 native amino acids around the acetylation and ubiquitination sites were illustrated by WebLogo (Crooks, Hon, Chandonia, & Brenner, 2004) (Figure 5). Amino acids surrounding the acetylated sites were most likely to be K and L, which was in line with previous findings (Huang et al., 2013; Lee et al., 2010; Vidali et al., 1968). In addition to the reported preference of D and E, L was vastly overrepresented in ubiquitinated sites. This result could explain why the conservation status of L was less important for discriminating between acetylation and ubiquitination because these sites only contained a few features that belonged to the PSSM category in our study (Figure 5). Instead, K occupied numerous PSSM conservation score features, which dominated the surrounding acetylated sites and was depleted around ubiquitylated sites. This result was also supported by our finding that the PSSM K conservation status at several specific sites (for example, sites 6, 8, 14, 16, 18, and 19) ranked high in the
It is the first time a computational method has been developed for distinguishing the two modification types. Analysis of the optimal DAUFSA features revealed some important feature differences and motif differences between the two types of PTMs. The algorithm developed here can be used in practice to study lysine acetylation and lysine ubiquitination. It also can be extended to the biological studies of differences among other PTM types, in addition to acetylation and ubiquitination.

Supplementary material
The supplementary material for this paper is available online at http://dx.doi.org/10.1080/07391102.2014.1001793.

Funding
This work was supported by grants from the National Basic Research Program of China [grant numbers 2011CB5010101, 2011CB5010102]; the National Natural Science Foundation of China [grant numbers 61401302, 31371335, 81171342, 81201148, 31050110432, 31150110577]; the Tianjin Research Program of Application Foundation and Advanced Technology [grant number 14JCQNJC09500], the Specialized Research Fund for the Doctoral Program of Higher Education of China [grant numbers 2013032120070, 20130031120001, 2012032120073]; the Seed Foundation of Tianjin University [grant numbers 60302064, 60302069]; and the Innovation Program of Shanghai Municipal Education Commission [grant number 12ZZ087] for the grant titled ‘The First-class Discipline of Universities in Shanghai.’

References
Allfrey, V. G., Faulkner, R., & Mirsky, A. E. (1964). Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. Proceedings of the National Academy of Sciences USA, 51, 786–794.
Alonso, V., & Friedman, P. A. (2013). Minireview: Ubiquitination-regulated G Protein-coupled receptor signaling and trafficking. Molecular Endocrinology, 27, 558–572.
Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research, 25, 3389–3402.
Atchley, W. R., Zhao, J. P., Fernandes, A. D., & Duke, T. (2005). Solving the protein sequence metric problem. Proceedings of the National Academy of Sciences USA, 102, 6395–6400.
Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. Bioinformatics, 16, 412–424.
Basu, A., Rose, K. L., Zhang, J., Beavis, R. C., Ueberheide, B., Garcia, B. A., ... Hake, S. B. (2009). Proteome-wide prediction of acetylation substrates. Proceedings of the National Academy of Sciences USA, 106, 13785–13790.
Benezra, R. (1994). An intermolecular disulfide bond stabilizes E2A homodimers and is required for DNA binding at physiological temperatures. Cell, 79, 1057–1067.

Figure 5. Amino acid occurrence frequencies surrounding the active lysine generated by WebLogo. Logo illustrations were generated based on all 21-residue peptides in our data-set, showing the occurrence frequencies of 20 amino acids surrounding the ubiquitinated lysine (a) or the acetylated lysine (b). N and C represent the N- and C-terminuses of the 21-residue peptides, respectively.

mRMR list (Supporting Information S2). These results aided in unraveling the significant contributions of these factors to the discrimination of these two modifications.

4.4. Guidance for experimental validations
Features selected by our computational method for weighing the differences between acetylation and ubiquitination sites may provide comprehensive knowledge of these two modification types, as well as useful clues for further experimental validations. We showed that the conservation status and physicochemical properties of amino acids at sites 8–10 played pivotal roles in the determination of whether a site is acetylated or ubiquitylated, which is in accordance with the findings of different motifs employed by acetylation and ubiquitination (Figure 5). Secondary structure features displayed great power in differentiating acetylation from ubiquitination due to diverse interaction modules of these two enzyme-catalyzed modifications (Figure 2), in which a regular structure (‘strand’) may aid in more efficient discrimination (Figure 3(B)). Our results revealed that exposed amino acids were more efficient in distinguishing acetylation from ubiquitination because they had more features than buried amino acids (Figure 3(D)). These results demonstrated feature differences that reflect governing factors between the two modifications that require experimental validation and further investigation.

5. Conclusion
In this study, we developed DAUFSA to distinguish between lysine acetylation and lysine ubiquitination and achieved an accuracy of 69.53% with .3853 MCC value.
Cai, Y., Huang, T., Hu, L., Shi, X., Xie, L., & Li, Y. (2012). Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino Acids*, 42, 1387–1395.

Cai, Y. D., & Lu, L. (2008). Predicting N-terminal acetylation based on feature selection method. *Biochemical and Biophysical Research Communications*, 372, 862–865.

Chen, Z., Chen, Y. Z., Wang, X. F., Wang, C., Yan, R. X., & Zhang, Z. (2011). Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PLoS ONE*, 6, e22930.

Chen, W., Feng, P. M., Lin, H., & Chou, K. C. (2013). iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Research*, 41, e68.

Chen, W., Li, L., Du, Z., Liu, J., Reitter, J. N., Mills, K. V., … Wang, C. (2012). Intramolecular disulfide bond between catalytic cysteines in an in vitro precursor. *Journal of the American Chemical Society*, 134, 2500–2503.

Cheng, J., Randall, A. Z., Sweredoski, M. J., & Baldi, P. (2005). SCRATCH: A protein structure and structural feature prediction server. *Nucleic Acids Research*, 33, W72–W76.

Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology*, 273, 236–247.

Choudhary, C., Kumar, C., Gnad, F., Nielsen, M. L., Rehan, M., Walther, T. C., … Mann, M. (2009). Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, 325, 834–840.

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: A sequence logo generator. *Genome Research*, 14, 1188–1190.

Danielsen, J. M., Sylvestersen, K. B., Bekker-Jensen, S., Szklarczyk, D., Poulsen, J. W., Horn, H., … Nielsen, M. L. (2011). Mass spectrometric analysis of lysine ubiquitination reveals promiscuity at site level. *Molecular and Cellular Proteomics*, 10(M110), 003590.

Du, Q. S., Jiang, Z. Q., He, W. Z., Li, D. P., & Chou, K. C. (2006). Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. *Journal of Biomolecular Structure & Dynamics*, 23, 635–640.

Ferron, F., Longhi, S., Canard, B., & Karlin, D. (2006). A practical overview of protein disorder prediction methods. *Proteins-Structure Function and Bioinformatics*, 65, 1–14.

Gent, M. S., Worby, C. A., & Dixon, J. E. (2005). Insights into Lafora disease: Malin is an E3 ubiquitin ligase that ubiquitinates and promotes the degradation of laforin. *Proceedings of the National Academy of Sciences USA*, 102, 8501–8506.

Gnad, F., Ren, S., Choudhary, C., Cox, J., & Mann, M. (2010). Predicting post-translational lysine acetylation using support vector machines. *Bioinformatics*, 26, 1666–1668.

Grant, P. A. (2001). A tale of histone modifications. *Genome Biology*, 2: REVIEWS0003.

Haglund, K., & Dikic, I. (2005). Ubiquitination and cell signaling. *The EMBO Journal*, 24, 3353–3359.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, 11, 10–18.

He, Z. S., Zhang, J., Shi, X. H., Hu, L. L., Kong, X. Y., Cai, Y. D. & Chou, K. C. (2010). Predicting drug–target interaction networks based on functional genes and biological features. *PLoS ONE*, 5, e9603.

Hou, T., Zheng, G. Y., Zhang, P. Y., Jia, J., Li, J., Xie, L., … Li, Y. X. (2014). LAcEP: Lysine acetylation site prediction using logistic regression classifiers. *PLoS ONE*, 9, e89575.

Huang, T., Cui, W. R., Hu, L. L., Feng, K. Y., Li, Y. X., & Cai, Y. D. (2009). Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles. *PLoS ONE*, 4, e8126.

Huang, G., Zhou, Y., Zhang, Y., Li, B. Q., Zhang, N., & Cai, Y. D. (2013). Prediction of carbamylated lysine sites based on the one-class k-nearest neighbor method. *Molecular BioSystems*, 9, 2729–2740.

Jiang, Y., Li, B. Q., Zhang, Y., Feng, Y. M., Gao, Y. F., Zhang, N., & Cai, Y. D. (2013). Prediction and analysis of post-translational pyruvoyl residue modification sites from internal serines in proteins. *PLoS ONE*, 8, e66678.

Johnson, D. S., Li, W., Gordon, D. B., Bhattacharjee, A., Curry, B., Ghosh, J., … Liu, S. X. (2008). Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Research*, 18, 393–403.

Kawashima, S., & Kaniehisa, M. (2000). AIndex: Amino Acid index database. *Nucleic Acids Research*, 28, 374–374.

Kessler, B. M. (2013). Ubiquitin – Omics reveals novel networks and associations with human disease. *Current Opinion in Chemical Biology*, 17, 59–65.

Kirkpatrick, D. S., Denison, C., & Gygi, S. P. (2005). Weighing in on ubiquitin: The expanding role of mass-spectrometry-based proteomics. *Nature Cell Biology*, 7, 750–757.

Komla-Soukha, I., & Sureau, C. (2006). A tryptophan-rich motif in the carboxyl terminus of the small envelope protein of hepatitis B virus is central to the assembly of hepatitis delta virus particles. *Journal of Virology*, 80, 4648–4655.

Komaki, H., & Talianidis, I. (2010). Cross-talk between post-translational modifications regulate life or death decisions by E2F1. *Cell Cycle*, 9, 8386–8387.

Lee, T. Y., Hsu, J. B., Lin, F. M., Chang, W. C., Hsu, P. C., & Huang, H. D. (2010). N-Ace: Using solvent accessibility and physicochemical properties to identify protein N-acetylation sites. *Journal of Computational Chemistry*, 31, 2759–2771.

Li, B. Q., Hu, L. L., Chen, L., Feng, K. Y., Cai, Y. D., & Chou, K. C. (2012a). Prediction of protein domain with mRMR feature selection and analysis. *PLoS ONE*, 7, e39308.

Li, B. Q., Huang, T., Liu, L., Cai, Y. D., & Chou, K. C. (2012b). Identification of colorectal cancer related genes with mRMR and shortest path in protein–protein interaction network. *PLoS ONE*, 7, e33393.

Li, S. L., Li, H., Li, M. F., Shyr, Y., Xie, L., & Li, Y. X. (2009). Improved prediction of lysine acetylation by support vector machines. *Protein & Peptide Letters*, 16, 977–983.

Li, A., Xue, Y., Jin, C. J., Wang, M. H., & Yao, X. B. (2006). Prediction of N-epsilon-acetylation on internal lysines implemented in Bayesian Discriminant Method. *Biochemical and Biophysical Research Communications*, 350, 818–824.

Lin, W. Z., Fang, J. A., Xiao, X., & Chou, K. C. (2013). iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. *Molecular BioSystems*, 9, 634–644.

Liu, X., Xiao, W., Wang, X. D., Li, Y. F., Han, J., & Li, Y. (2013). The p38-interacting protein (p38IP) regulates G2/M progression by promoting alpha-tubulin acetylation via inhibiting ubiquitination-induced degradation of the acetyltransferase GCN5. *Journal of Biological Chemistry*, 288, 36648–36661.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405, 442–451.
Tang, X., Wen, S., Zheng, D., Tucker, L., Cao, L., Pantazatos, D., ... Ramaratnam, B. (2013). Acetylation of drosha on the N-terminus inhibits its degradation by ubiquitination. PLoS ONE, 8, e72503.

Ting, K. M. & Witten, I. H. (1997). Stacking bagged and bagged models. Fourteenth International Conference on Machine Learning, San Francisco, CA, 367–375.

Trouillas, P., Berges, J., & Houee-Levin, C. (2011). Toward understanding the protein oxidation processes: (OH)-O-center dot addition on tyrosine, phenylalanine, or methionine? International Journal of Quantum Chemistry, 111, 1143–1151.

Tung, C. W., & Ho, S. Y. (2008). Computational identification of ubiquitylation sites from protein sequences. BMC Bioinformatics, 9, 310.

Van Damme, P., Arnesen, T., & Gevaert, K. (2011). Protein alpha-N-acetylation studied by N-terminomics. FEBS Journal, 278, 3822–3834.

Vidal, G., Gershey, E. L., & Allfrey, V. G. (1968). Chemical studies of histone acetylation. The distribution of epsilon-N-acetyllysine in calf thymus histones. Journal of Biological Chemistry, 243, 6361–6366.

Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics, 13, S2.

Wang, X., Moore, S. C., Laszczak, M., & Ausio, J. (2000). Acetylation increases the alpha-helical content of the histone tails of the nucleosome. Journal of Biological Chemistry, 275, 35013–35020.

Xiong, Y., & Guan, K. L. (2012). Mechanistic insights into the regulation of metabolic enzymes by acetylation. The Journal of Cell Biology, 198, 155–164.

Xu, Y., Wang, X. B., Ding, J., Wu, L. Y., & Deng, N. Y. (2010). Lysine acetylation sites prediction using an ensemble of support vector machine classifiers. Journal of Theoretical Biology, 264, 130–135.

Xu, R., Zhou, J., Liu, B., He, Y., Zou, Q., Wang, X., & Chou, K. C. (2014). Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach. Journal of Biomolecular Structure & Dynamics, published online ahead of print, doi:10.1080/07391102.2014.968624, 1–11.

Yang, X. J. (2004). Lysine acetylation and the bromodomain: a new partnership for signaling. BioEssays, 26, 1076–1087.

Yang, X. J., & Seto, E. (2008). Lysine acetylation: Codified crosstalk with other posttranslational modifications. Molecular Cell, 31, 449–461.

Zhang, N., Li, B. Q., Gao, S., Ruan, J. S., & Cai, Y. D. (2012). Computational prediction and analysis of protein γ-carboxylation sites based on a random forest method. Molecular BioSystems, 8, 2946–2955.

Zhang, N., Zhou, Y., Huang, T., Zhang, Y. C., Li, B. Q., Chen, L., & Cai, Y. D. (2014). Discriminating between lysine sumoylation and lysine acetylation using mRMR feature selection and analysis. PLoS ONE, 9, e107464.

Zhao, X., Li, X., Ma, Z., & Yin, M. (2011). Prediction of lysine ubiquitylation with ensemble classifier and feature selection. International Journal of Molecular Sciences, 12, 8347–8361.