Amino Acids (2012) 42:1387–1395
DOI 10.1007/s00726-011-0835-0

ORIGINAL ARTICLE

Prediction of lysine ubiquitination with mRMR feature selection and analysis

Yudong Cai · Tao Huang · Lele Hu · Xiaohe Shi · Lu Xie · Yixue Li

Received: 25 June 2010 / Accepted: 11 January 2011 / Published online: 26 January 2011

Abstract Ubiquitination, one of the most important post-translational modifications of proteins, occurs when ubiquitin (a small 76-amino acid protein) is attached to lysine on a target protein. It often commits the labeled protein to degradation and plays important roles in regulating many cellular processes implicated in a variety of diseases. Since ubiquitination is rapid and reversible, it is time-consuming and labor-intensive to identify ubiquitination sites using conventional experimental approaches. To efficiently discover lysine-ubiquitination sites, a sequence-based predictor of ubiquitination site was developed based on nearest neighbor algorithm. We used the maximum relevance and minimum redundancy principle to identify the key features and the incremental feature selection procedure to optimize the prediction engine. PSSM conservation scores, amino acid factors and disorder scores of the surrounding sequence formed the optimized 456 features. The Mathew's correlation coefficient (MCC) of our ubiquitination site predictor achieved 0.142 by jackknife cross-validation test on a large benchmark dataset. In independent test, the MCC of our method was 0.139, higher than the existing ubiquitination site predictor UbiPred and UbPred. The MCCs of UbiPred and UbPred on the same test set were 0.135 and 0.117, respectively. Our analysis shows that the conservation of amino acids at and around lysine plays an important role in ubiquitination site prediction. What's more, disorder and ubiquitination have a strong relevance. These findings might provide useful insights for studying the mechanisms of ubiquitination and modulating the ubiquitination pathway, potentially leading to potential therapeutic strategies in the future.

Keywords Ubiquitination · Maximum relevance and minimum redundancy (mRMR) · Incremental feature selection (IFS) · Nearest neighbor algorithm (NNA)

Y. Cai and T. Huang contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-011-0835-0) contains supplementary material, which is available to authorized users.

Y. Cai (✉) · L. Hu
Institute of Systems Biology, Shanghai University, Shanghai 200444, People's Republic of China
e-mail: cai_yud@yahoo.com.cn

T. Huang · Y. Li (✉)
Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China
e-mail: yxli@sibs.ac.cn

T. Huang · L. Xie · Y. Li
Shanghai Center for Bioinformation Technology, Shanghai 200235, People's Republic of China

Y. Cai
Centre for Computational Systems Biology, Fudan University, Shanghai 200433, People's Republic of China

X. Shi
Singapore Bioimaging Consortium, Agency for Science, Technology and Research, Singapore 138667, Singapore

Introduction

In the post-genomic era, knowledge of post-translational modifications (PTMs) of proteins is crucial for understanding the dynamic proteome and various signaling pathways or networks in cells (Aguilar and Wendland 2003; Saghatelian and Cravatt 2005; Herrmann et al. 2007; Hicke and Dunn 2003; Welchman et al. 2005). One of the most important and universal post-translational modifications,

protein ubiquitination is a rapid and reversible biochemical process in which an iso-peptide bond forms covalently between the C-terminal double-glycine carboxy group of a ubiquitin protein and the ε-amino group of lysine residues of a substrate protein (Pickart 2001). Ubiquitination regulates a variety of biological processes, such as signal transduction, cell division/mitosis, apoptosis, and endocytosis (Sun and Chen 2004; Reinstein and Ciechanover 2006; Hoeller et al. 2006; Hicke 2001). An aberrance of the ubiquitin–proteasome system (UPS) is associated in numerous pathological diseases, such as inflammatory diseases, neurodegenerative disorders, and cancers (Hoeller et al. 2006; Reinstein and Ciechanover 2006).

Identification of ubiquitinated proteins sites is one of the greatest challenges in gaining a full understanding of the regulatory roles of ubiquitination regulation and the molecular mechanism of the ubiquitin system. It is time-consuming and labor-intensive to use conventional experimental approaches to identify the potential ubiquitinated proteins sites, such as site-mutagenesis (Lin et al. 2005), antibodies of Ub (anti-Ub) (Gentry et al. 2005), and high-throughput mass-spectrometry (MS) (Kirkpatrick et al. 2005). Therefore, it is convenient and efficient to use in silico algorithms in prediction of ubiquitination sites.

In this work, we developed a new computational method to predict lysine-ubiquitination. Specifically, we used a machine learning approach (Nearest Neighbor Algorithm) combined with feature selection (IFS based on mRMR, Peng et al. 2005a). Twenty-six parameters were used to describe each amino acid of the lysine site and its surrounding ones (from −10 to +10). The 26 parameters can be broken down into 3 categories: 20 position-specific scoring matrices (PSSM) conservation scores, 5 amino acid factors and 1 disorder score. The PSSM score quantifies the conservation status of each site in the protein sequence (Altschul et al. 1997). Amino acid factors were defined by Atchley et al. (2005) through multivariate statistical analyses on AAIndex (Kawashima and Kanehisa 2000) to produce five amino acid factors that reflected polarity (AAFactor 1), secondary structure (AAFactor 2), molecular volume (AAFactor 3), codon diversity (AAFactor 4), and electrostatic charge (AAFactor 5). Disorder score (Peng et al. 2006) represents the disorder status of each site in the protein sequence. Under physiological conditions, disordered regions in proteins do not have fixed three-dimensional structures, but they play various roles in signaling and regulation.

This study focuses on the computational identification of lysine (K) ubiquitination. The Mathew's correlation coefficient (MCC) of lysine (K) ubiquitination site predictions was 0.142 on training set evaluated by jackknife cross-validation and 0.139 on independent test set. The following features distinguish our study from previous ubiquitination prediction models (Radivojac et al. 2010; Tung and Ho 2008): (1) a larger benchmark dataset was used, (2) the feature set was much smaller and more compact, (3) jackknife cross-validation and independent test were used to evaluate effectively and objectively the performance of our classifier, (4) the applied prediction model nearest neighbor algorithm was much simpler and faster than SVM (Tung and Ho 2008) or random forest (Radivojac et al. 2010), both of which could have easily introduced overfitting problems, and (5) on independent test our model has better performance than two existing predictors: UbiPred and UbPred. Our analysis shows that the conservation of amino acid at and around the lysine site plays important roles in ubiquitination site prediction. It also shows that biochemical and physicochemical properties of amino acids in the flanking sequences are important for the ubiquitination process. Interestingly, disorder and ubiquitination have a strong relevance.

Materials and methods

Dataset

The ubiquitinated protein sequences we used for training comes from SysPTM (Li et al. 2009). Peptides containing lysine (K) were extracted as our training samples. According to Tung's work (Tung and Ho 2008), the best window size for ubiquitination site prediction is 21. So we adopted their window size and represent each lysine-ubiquitination site with a peptide fragment consisting of 21 residues with 10 residues upstream and 10 residues downstream of the lysine (K). The original dataset downloaded from SysPTM has 514 lysine-ubiquitination sites from 349 proteins. After removing the redundancy of the 349 protein sequences against homology bias using the program cd-hit (Li and Godzik 2006), we obtained 273 distinct sequences among which the sequence identity was lower than 0.6. We randomly selected 12 proteins to form the independent test set and the left 261 proteins to construct the training set. Since the number of ubiquitinated lysine sites and non-ubiquitinated lysine sites were highly imbalanced, we randomly selected three times negative samples (non-ubiquitinated lysine fragments) to match the positive ones (ubiquitinated lysine fragments) in the training set. In the independent test set, we retained all the positive and negative samples to make it close to real situation. There were 364 positive samples and 1,092 negative samples in the training set; meanwhile in the independent test set, there were 14 positive samples and 267 negative samples. The benchmark dataset we used was larger than Tung's 157 ubiquitination sites (Tung and Ho 2008) or Radivojac's 272 ubiquitinated fragments (Radivojac et al. 2010). Both the positive and negative lysine samples for training and independent test can be found in Dataset S1.

Feature construction

PSSM conservation scores

Evolutionary conservation usually indicates important biology function. If an amino acid at a particular site of a protein is conserved, it may locate in a functionally important region of the protein.

Position-specific iterated (PSI)-BLAST (Altschul et al. 1997) can measure the residue conservation in a given location. Each residue can be represented by a 20-dimensional vector which stands for the probabilities of conservation against mutations to 20 different kinds of amino acids. Position-specific scoring matrix (PSSM) (Ahmad and Sarai 2005) is a matrix in which each row is such a 20-dimensional vector. The rows of matrix correspond to the residues in the protein sequence. If a residue is conserved according to PSI-BLAST, it is likely to be biologically important and ubiquitinated. In this study, we encoded the conservation status of each amino acid in the protein sequence with PSSM conservation score. The program "blastpgp" (PSI-BLAST) downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast was used to calculate the PSSM conservation score with three iterations (−j 3) and e-value threshold for inclusion in multipass model 0.0001 (−h 0.0001).

Amino acid factors

Atchley et al. (2005) did multivariate statistical analyses on AAIndex (Kawashima and Kanehisa 2000) which is a database of various physicochemical and biochemical properties of amino acids, to produce five multidimensional patterns of attribute covariation reflecting polarity (AA-Factor 1), secondary structure (AAFactor 2), molecular volume (AAFactor 3), codon diversity (AAFactor 4), and electrostatic charge (AAFactor 5). These five transformed scores (called "amino acid factors" here) have been used to successfully solve several difficult biology problems, such as deleterious non-synonymous SNP identification (Huang et al. 2010b) and B cell epitopes prediction (Rubinstein et al. 2009). Here, we used these five amino acid factors to encode each amino acid in the lysine fragment.

Disorder score

Under physiological conditions, disordered regions in proteins do not have fixed three-dimensional structures, but they play various roles in signaling and regulation by multiple binding of proteins and high-specificity low-affinity interactions (Sickmeier et al. 2007). In this study, we encoded the disorder status of each amino acid in the protein sequence with disorder score calculated by VSL2 (Peng et al. 2006). The VSL2 predictors can accurately identify disordered regions of any length, especially the short disordered regions. The disorder scores of lysine site and its surrounding amino acids formed the features of disorders.

The feature space

The lysine (K) ubiquitination site was encoded by 20 PSSM conservation scores and 1 disorder score, in total 21 features. Each of its surrounding amino acids (10 residues upstream and 10 residues downstream) was encoded by 26 features, including 20 PSSM conservation scores, 5 amino acid factors, and 1 disorder score. Overall, each sample was represented by $26 \times 20 + 21 = 541$ features.

mRMR method

In this study, we applied the maximum relevance and minimum redundancy (mRMR) method (Peng et al. 2005b) to analyze the importance of different features. Each feature can be ranked based on its relevance to target variable, and the ranking process is able to consider the redundancy of these features at the same time. A "good" feature is defined as one that has the best trade-off between minimum redundancy within the features and maximum relevance to target variable. Mutual information (MI), which measures the mutual dependence of two variables, is used to quantify both relevance and redundancy in this method. MI is defined as following

$$I(X,Y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

where X, Y are vectors, $p(x, y)$ is the joint probabilistic density, $p(x)$ and $p(y)$ are the marginal probabilistic densities. Given M data points drawn from the joint probability distribution (x_i, y_i), $i = 1,\ldots, M$, the joint and marginal densities can be estimated by the Gaussian kernel estimator as following (Beirlant et al. 1997; Qiu et al. 2009)

$$p(x,y) = \frac{1}{M}\sum\frac{1}{2\pi h^2}e^{-\frac{1}{2h^2}((x-x_i)^2+(y-y_i)^2)} \tag{2}$$

$$p(x) = \frac{1}{M}\sum\frac{1}{\sqrt{2\pi h^2}}e^{-\frac{1}{2h^2}(x-x_i)^2} \tag{3}$$

$$p(y) = \frac{1}{M}\sum\frac{1}{\sqrt{2\pi h^2}}e^{-\frac{1}{2h^2}(y-y_i)^2} \tag{4}$$

h is a tuning parameter that controls the width of the kernels.

Let Ω denote the whole feature set, while Ω_s denotes the already-selected feature set which contains m features and Ω_t denotes the to-be-selected feature set which contains n features. Relevance D of the feature f in Ω_t with the target c can be calculated by:

$$D = I(f, c). \tag{5}$$

And redundancy R of the feature f in Ω_t with all the features in Ω_s can be calculated by:

$$R = \frac{1}{m}\sum_{f_i \in \Omega_s} I(f, f_i). \tag{6}$$

To obtain the feature f_j in Ω_t with maximum relevance and minimum redundancy, Eqs. 5 and 6 are combined with the mRMR function:

$$\max_{f_j \in \Omega_t}\left[I(f_j, c) - \frac{1}{m}\sum_{f_i \in \Omega_s} I(f_j, f_i)\right](j = 1, 2, \ldots, n). \tag{7}$$

For a feature set with $N(=m + n)$ features, the feature evaluation will continue N rounds. After these evaluations, we will get a feature set S by mRMR method:

$$S = \left\{f_1', f_2', \ldots, f_h', \ldots, f_N'\right\}. \tag{8}$$

The feature index h indicates the importance of feature. The better a feature is, the smaller its index h will be.

Nearest neighbor algorithm

We used nearest neighbor algorithm (NNA) to build the prediction model. NNA makes its decision by calculating similarities between the test sample and all the training samples. In our study, the distance between vector $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots, b_n)$ is defined as follows (Qian et al. 2006; Huang et al. 2009, 2010a; Cai et al. 2010):

$$D(A, B) = 1 - \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}. \tag{9}$$

In NNA, the query vector will be designated to the same class of its nearest neighbor in training set with known classes which has the smallest distance.

Jackknife cross-validation and independent test

We used the jackknife cross-validation (Li et al. 2007; Cai et al. 2009; Huang et al. 2008) to evaluate the performance of our classifier on training set. With jackknife cross-validation, every sample is tested by the predictor trained with all the other samples. Besides the jackknife cross-validation on training set, we also did independent test of our model. Since the positive and negative samples are highly imbalanced in training set and independent test set, the Matthews's correlation coefficient (MCC) (Baldi et al. 2000) was used to evaluate the prediction performance and defined as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FP}) \times (\text{TN} + \text{FN})}} \tag{10}$$

where TP, TN, FP and FN stand for true positive, true negative, false positive and false negative, respectively.

Taking both sensitivity and specificity into account, MCC is considered as a balanced measure in dealing with imbalanced data (Baldi et al. 2000; Han et al. 2008).

Meanwhile, sensitivity (S_n), specificity (S_p) and accuracy (ACC) defined as following were also calculated

$$\text{S}_n = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{S}_p = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{12}$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{13}$$

where TP, TN, FP and FN stand for true positive, true negative, false positive and false negative, respectively.

Incremental feature selection (IFS)

After mRMR gives the rank of features according to their importance, it is still unknown how many fore features in the list should be chosen. To identify the optimal number of features, incremental feature selection (IFS) (Huang et al. 2009, 2010a; Cai et al. 2010) was used.

An incremental feature selection is conducted for each of the independent predictor with the ranked features. Features in a set are added one by one from higher to lower rank. If one feature is added, a new feature set is obtained, then we get N feature sets where N is the number of features, and the ith feature set is:

$$S_i = \{f_1, f_2, \ldots, f_i\}(1 \le i \le N).$$

Based on each of the N feature sets, NNA predictors were constructed and tested by jackknife cross-validation on training set. With MCC of jackknife cross-validation calculated, we obtain an IFS table with the number of features and the performance of them. S_{optimal} is the optimal feature set that achieves the highest MCC.

Results

mRMR result

Using the program "mRMR" (Peng et al. 2005b) downloaded from http://penglab.janelia.org/proj/mRMR, we obtained the ranked mRMR list of 541 features. The smaller

index of feature indicates more important roles in discriminating positive samples from negative ones. The mRMR list was used in IFS procedure for feature selection and analysis.

IFS result

Based on the outputs of mRMR, we built 541 individual predictors for the 541 sub-feature sets to predict the lysine-ubiquitination sites. As described in the "Materials and methods", we tested the predictors with one feature, two features, three features, etc., and obtained the IFS result which can be found in Table S1.

Figure 1 shows IFS curve plotted based on Table S1. The highest MCC was 0.142 when 456 features were used. So these 456 features were considered as the optimal feature set of our classifier. The 456 optimal features were given in Table S2.

Independent test and comparison with other methods

We tested our model in an independent dataset in which there were 14 positive samples and 267 negative samples. The MCC of our method independent test was 0.139. Meanwhile, we also predicted the independent set with two existing ubiquitination site predictors: UbiPred (Tung and Ho 2008) and UbPred (Radivojac et al. 2010). The MCCs of UbiPred and UbPred on the same independent test set were 0.135 and 0.117, respectively. The performance of our model is better than both UbiPred and UbPred on the independent test set in which the positive and negative samples are highly imbalanced and close to real situation.

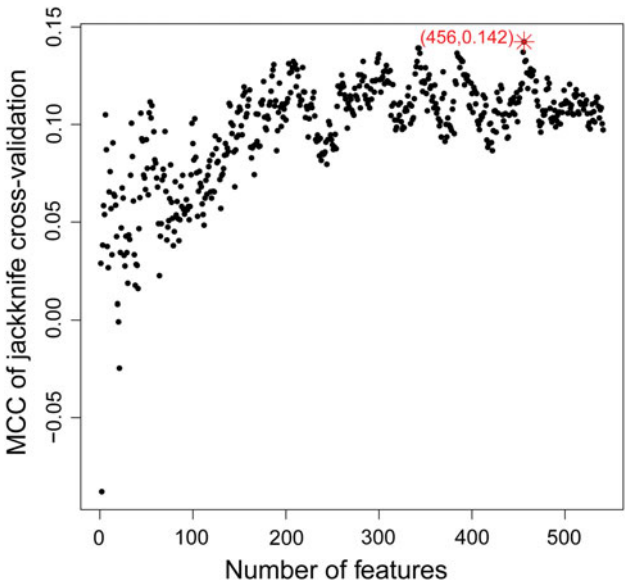

Fig. 1 The IFS curve of predictors. In the IFS curve, the *x*-axis is the number of features and the *y*-axis is the MCC of jackknife cross-validation. The highest MCC was 0.142 when 456 features were used. So these 456 features were considered as the optimal feature set of our classifier

The distribution of the optimized feature set

As described in the "Materials and methods", there were three kinds of features: PSSM conservation scores, amino acid factors and disorder scores. The number of each type of features in optimal feature set was investigated and shown in Fig. 2a. The number of each site of features in optimal feature set was shown in Fig. 2b. In the optimized 456 features, there were 100 amino acid factor features, 8 disorder score features and 348 PSSM conservation score features. This may suggest that conservation played important role for the ubiquitination site prediction. Similar evolutionary information exploited through position-specific scoring matrices (PSSMs) was also used in two previous prediction models of ubiquitylation (Radivojac et al. 2010; Tung and Ho 2008).

Since there were 348 PSSM conservation score features which count for a large proportion in the optimized 456 features, we investigated the number of each kind of amino acid of PSSM features (Fig. 3a) and the number of each site of PSSM features (Fig. 3b). The conservation of lysine site (AA11) was most important for the ubiquitination, and there were more PSSM conservation score features at nearby site AA7, AA8, AA9, AA12, AA14 and remote site AA1, AA18, AA19, AA21 than others. The importance of remote site explained why Tung found that the proper window size for ubiquitylation site prediction is 21 (Tung and Ho 2008). In addition, the conservation against mutations to 20 amino acids played different roles. Mutations to amino acids A, C, F, H, I, L, M, S, T, V, W and Y have more influence on ubiquitination than other kinds of mutations.

The number of amino acid factor features in the optimal feature set was 100, which means all amino acid factor features have been selected and all the five amino acid factors were equally important.

There were 8 disorder scores selected in the optimal feature set: the disorder scores at site AA6, AA7, AA8, AA9, AA10, AA14, AA17 and AA18. The disorder score of AA7 ranked first in the mRMR list. This indicated the disorder status of amino acid around the ubiquitination site could affect the ubiquitination process. It has been reported that disordered proteins have a greater proportion of predicted ubiquitination sites (Edwards et al. 2009). To better investigate the relationship between disorder and ubiquitination, we averaged the disorder scores at each site in ubiquitinated fragments and non-ubiquitinated fragments and compared them in Fig. 4. In Fig. 4, the red and blue dots were the mean of disorder scores at each site in ubiquitinated fragments and non-ubiquitinated fragments, respectively. The width of error bar represents the standard error of the mean. It is quite clear that the ubiquitinated

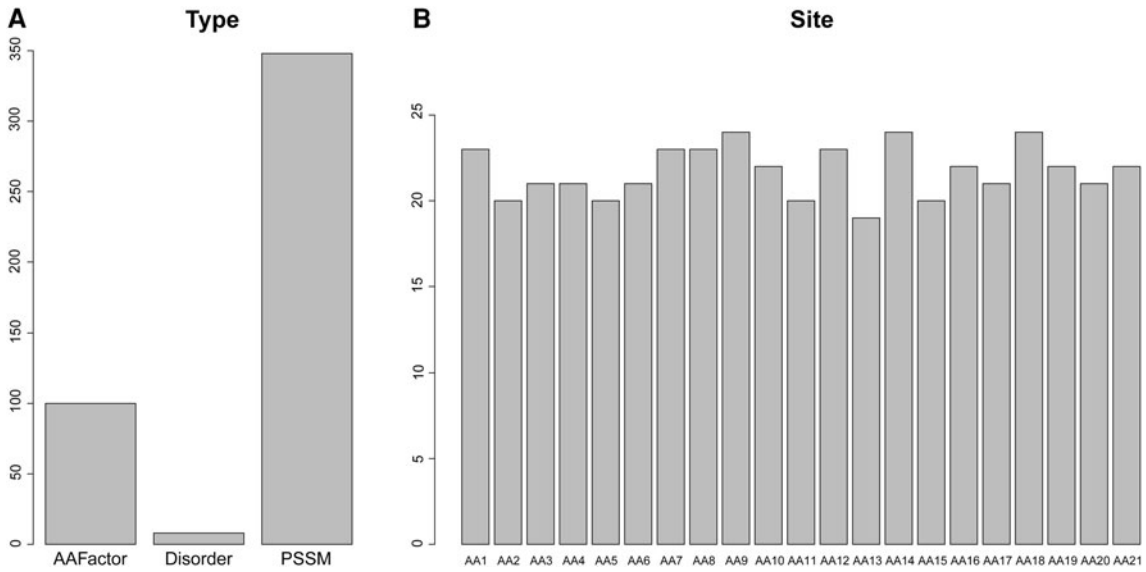

Fig. 2 The number of each type or each site of features in optimal feature set. **a** The number of each type of features in optimal feature set. There were 100 amino acid factor features, 8 disorder score features and 348 PSSM conservation score features. **b** The number of each site of features in optimal feature set. From 10 residues upstream to 10 residues downstream ("AA1", "AA2", …, "AA20", "AA21"), there were 23, 20, 21, 21, 20, 21, 23, 23, 24, 22, 20, 23, 19, 24, 20, 22, 21, 24, 22, 21 and 22 features, respectively

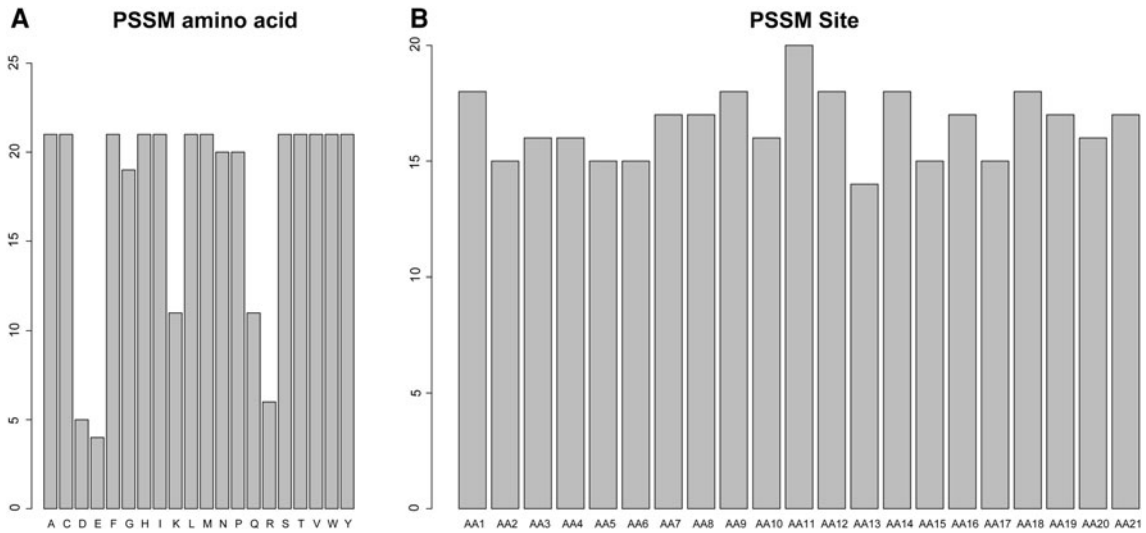

Fig. 3 The number of each type or each site of PSSM features in optimal feature set. **a** The number of each type of PSSM features in optimal feature set. **b** The number of each site of PSSM features in optimal feature set. The conservation of lysine site (AA11) was most important for the ubiquitination, and there were more PSSM conservation score features at nearby site AA7, AA8, AA9, AA12, AA14 and remote site AA1, AA18, AA19, AA21 than others

fragments and non-ubiquitinated fragments have very different disorder score pattern. The disorder score at each site in the ubiquitinated fragments is higher than the one in the non-ubiquitinated fragments.

Discussion

Proteins are targeted for degradation by the covalent ligation to ubiquitin, a small 76-amino acid residue protein. Ubiquitination of target substrates is a highly collaborative process involving a three-step cascade mechanism between the ubiquitin-activating enzyme (E1), ubiquitin-conjugating enzymes (E2), and ubiquitin ligases (E3) (Hershko and Ciechanover 1998).

Within the selected physicochemical property parameters, we show that polarity (AAFactor 1), secondary structure (AAFactor 2), molecular volume (AAFactor 3), codon diversity (AAFactor 4), and electrostatic charge (AAFactor 5) share similar role in protein ubiquitination

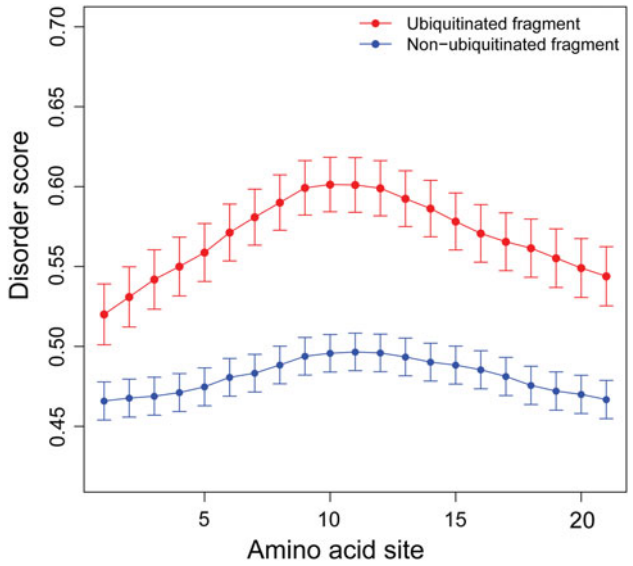

Fig. 4 The disorder scores at each site in ubiquitinated fragments and non-ubiquitinated fragments. The *upper* and *lower dots* were the mean of disorder scores at each site in ubiquitinated fragments and non-ubiquitinated fragments, respectively. The width of *error bar* represents the standard error of the mean

selection. The most pronounced feature of Ub sites is the abundance of charged and polar amino acids, especially negatively charged D and E, and the depletion of hydrophobic residues, such as L, I, F, and P around Ub sites (Nonaka et al. 2005; Radivojac et al. 2010). These parameters are highly related to electrostatic charge and amino acid composition in the adjacent sequence. The known E3 enzymes could be separated in two protein families: HECT domain and RING E3s. The crystal structures of these complexes reveal extraordinary specificity of interaction by a small set of loops at the end of the UbcH7 β-sheet (a subset of secondary structure) (Zheng et al. 2000; Huang et al. 1999). From these results, it is easier to understand how the presence of a few divergent surface residues could modulate the catalytic properties of ubiquitination. The similar positions of the three substrate-binding domains supported that RING E3s promote ubiquitin transfer by positioning the substrate in a manner such that the lysine is optimally E2 active size (Zheng et al. 2002; Schulman et al. 2000), spacing between the destruction motif and the ubiquitin-acceptor lysine residue as a parameter that affects the rate of substrate ubiquitination, further supporting the positioning model (Wu et al. 2003). These structure analyses emphasize the importance of secondary structure, molecular size or volume to the ubiquitination process.

The relationship between ubiquitination and protein disorder is complex and remains unclear, but researchers have observed that the percentage of residues predicted as possible ubiquitination sites increases with increasing amounts of disorder (Edwards et al. 2009). A large proportion of disordered proteins are highly expressed in many tissues (Edwards et al. 2009). These proteins may have a higher chance of degradation, as they are likely to have a higher density of ubiquitination sites.

Although much knowledge about ubiquitination has been accumulated to date, it is difficult to assume that all substrates carry a similar preexisting structure before they bind to the components of the ubiquitination machinery. Here, we examine sequence and structural preferences of all available ubiquitination sites and show that they have selected physicochemical property parameters. Regulated protein targeting and turnover through the ubiquitin–proteasome system underlies a host of critical physiological and pathological states in humans. The ability to modulate the individual steps in the ubiquitination pathway offers potential therapeutic strategies in the future.

Conclusion

A novel sequence-based predictor was developed for identifying the ubiquitination at lysine site. With the IFS feature selection procedure based on mRMR analysis, the predictor achieved an MCC of 0.142 by jackknife cross-validation test on benchmark dataset. In independent test, the MCC of our predictor was 0.139, higher than the existing ubiquitination site prediction tools UbiPred and UbPred. Our analysis shows that the conservation of amino acid at and around lysine plays important roles in ubiquitination site prediction. It also shows that electrostatic charge, molecular volume, secondary structure, codon diversity, and polarity of amino acids in the flanking sequences are important for the ubiquitination process. Interestingly, disorder and ubiquitination have a strong relevance. Although the results reported here are quite encouraging, the present study is merely a preliminary one. Further investigation is needed to clarify the predicted relationship between conservation, disorder and ubiquitination.

Acknowledgments The authors acknowledge Yvonne Poindexter at the Vanderbilt University Cancer Biostatistics Center for her editing. This work was supported by grants from National High-Tech R&D Program of China (863 Program) (2006AA02Z334, 2007DFA31040), China National Key Projects for Infectious Disease (2008ZX10002-021), National Basic Research Program of China (2006CB910700), National Natural Science Foundation of China (Grant No. 31070752) and Key Research Program (CAS) (KSCX2-YW-R-112).

References

Aguilar RC, Wendland B (2003) Ubiquitin: not just for proteasomes anymore. Curr Opin Cell Biol 15(2):184–190

Ahmad S, Sarai A (2005) Pssm-based prediction of DNA binding sites in proteins. BMC Bioinform 6:33. doi:10.1186/1471-2105-6-33

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402 (pii:gka562)

Atchley WR, Zhao J, Fernandes AD, Druke T (2005) Solving the protein sequence metric problem. Proc Natl Acad Sci USA 102(18):6395–6400. doi:10.1073/pnas.0408677102

Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16(5):412–424

Beirlant J, Dudewicz E, Gyorfi L, Meulen Evd (1997) Nonparametric entropy estimation: an overview. Int J Math Stat Sci 6(1):17–39

Cai Y, He J, Li X, Lu L, Yang X, Feng K, Lu W, Kong X (2009) A novel computational approach to predict transcription factor DNA binding preference. J Proteome Res 8(2):999–1003. doi:10.1021/pr800717y

Cai YD, Huang T, Feng KY, Hu L, Xie L (2010) A unified 35-gene signature for both subtype classification and survival prediction in diffuse large B cell lymphomas. PLoS One 5(9). doi:10.1371/journal.pone.0012726

Edwards YJ, Lobley AE, Pentony MM, Jones DT (2009) Insights into the regulation of intrinsically disordered proteins in the human proteome by analyzing sequence and gene expression data. Genome Biol 10(5):R50. doi:10.1186/gb-2009-10-5-r50

Gentry MS, Worby CA, Dixon JE (2005) Insights into lafora disease: Malin is an e3 ubiquitin ligase that ubiquitinates and promotes the degradation of laforin. Proc Natl Acad Sci USA 102(24):8501–8506

Han L, Wang Y, Bryant SH (2008) Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in pubchem. BMC Bioinform 9:401. doi:10.1186/1471-2105-9-401

Herrmann J, Lerman LO, Lerman A (2007) Ubiquitin and ubiquitin-like proteins in protein regulation. Circ Res 100(9):1276–1291

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

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

Hicke L, Dunn R (2003) Regulation of membrane protein transport by ubiquitin and ubiquitin-binding proteins. Annu Rev Cell Dev Biol 19:141–172

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

Huang L, Kinnucan E, Wang G, Beaudenon S, Howley PM, Huibregtse JM, Pavletich NP (1999) Structure of an e6ap–ubch7 complex: insights into ubiquitination by the e2–e3 enzyme cascade. Science 286(5443):1321–1326

Huang T, Tu K, Shyr Y, Wei CC, Xie L, Li YX (2008) The prediction of interferon treatment effects based on time series microarray gene expression profiles. J Transl Med 6:44. doi:10.1186/1479-5876-6-44

Huang T, Cui W, Hu L, Feng K, Li YX, Cai YD (2009) Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles. PLoS ONE 4(12):e8126. doi:10.1371/journal.pone.0008126

Huang T, Shi XH, Wang P, He Z, Feng KY, Hu L, Kong X, Li YX, Cai YD, Chou KC (2010a) Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS One 5(6):e10972. doi:10.1371/journal.pone.0010972

Huang T, Wang P, Ye ZQ, Xu H, He Z, Feng KY, Hu L, Cui W, Wang K, Dong X, Xie L, Kong X, Cai YD, Li Y (2010b) Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties. PLoS One 5(7):e11900. doi:10.1371/journal.pone.0011900

Kawashima S, Kanehisa M (2000) Aaindex: amino acid index database. Nucleic Acids Res 28(1):374 pii:gkd029

Kirkpatrick DS, Denison C, Gygi SP (2005) Weighing in on ubiquitin: the expanding role of mass-spectrometry-based proteomics. Nat Cell Biol 7(8):750–757

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659. doi:10.1093/bioinformatics/btl158

Li S, Liu B, Cai Y, Li Y (2007) Predicting protein n-glycosylation by combining functional domain and secretion information. J Biomol Struct Dyn 25(1):49–54

Li H, Xing X, Ding G, Li Q, Wang C, Xie L, Zeng R, Li Y (2009) Sysptm: a systematic resource for proteomic research on post-translational modifications. Mol Cell Proteomics 8(8):1839–1849. doi:10.1074/mcp.M900030-MCP200

Lin DH, Sterling H, Wang Z, Babilonia E, Yang B, Dong K, Hebert SC, Giebisch G, Wang WH (2005) Romk1 channel activity is regulated by monoubiquitination. Proc Natl Acad Sci USA 102(12):4306–4311

Nonaka T, Iwatsubo T, Hasegawa M (2005) Ubiquitination of alpha-synuclein. Biochemistry 44(1):361–368

Peng H, Long F, Ding C (2005a) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238

Peng H, Long F, Ding C (2005b) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238. doi:10.1109/TPAMI.2005.159

Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z (2006) Length-dependent prediction of protein intrinsic disorder. BMC Bioinform 7:208. doi:10.1186/1471-2105-7-208

Pickart CM (2001) Mechanisms underlying ubiquitination. Annu Rev Biochem 70:503–533

Qian Z, Cai YD, Li Y (2006) A novel computational method to predict transcription factor DNA binding preference. Biochem Biophys Res Commun 348(3):1034–1037

Qiu P, Gentles AJ, Plevritis SK (2009) Fast calculation of pairwise mutual information for gene regulatory network reconstruction. Comput Methods Programs Biomed 94(2):177–180. doi:10.1016/j.cmpb.2008.11.003

Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, Heyen JW, Goebl MG, Iakoucheva LM (2010) Identification, analysis, and prediction of protein ubiquitination sites. Proteins 78(2):365–380. doi:10.1002/prot.22555

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

Rubinstein ND, Mayrose I, Pupko T (2009) A machine-learning approach for predicting B cell epitopes. Mol Immunol 46(5):840–847. doi:10.1016/j.molimm.2008.09.009

Saghatelian A, Cravatt BF (2005) Assignment of protein function in the postgenomic era. Nat Chem Biol 1(3):130–142

Schulman BA, Carrano AC, Jeffrey PD, Bowen Z, Kinnucan ER, Finnin MS, Elledge SJ, Harper JW, Pagano M, Pavletich NP (2000) Insights into scf ubiquitin ligases from the structure of the skp1–skp2 complex. Nature 408(6810):381–386

Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK (2007) Disprot: the database of disordered proteins. Nucleic Acids Res 35(Database issue):D786–D793. doi:10.1093/nar/gkl893

Sun L, Chen ZJ (2004) The novel functions of ubiquitination in signaling. Curr Opin Cell Biol 16(2):119–126

Tung CW, Ho SY (2008) Computational identification of ubiquitylation sites from protein sequences. BMC Bioinform 9:310. doi:10.1186/1471-2105-9-310

Welchman RL, Gordon C, Mayer RJ (2005) Ubiquitin and ubiquitin-like proteins as multifunctional signals. Nat Rev Mol Cell Biol 6(8):599–609

Wu G, Xu G, Schulman BA, Jeffrey PD, Harper JW, Pavletich NP (2003) Structure of a beta-trcp1-skp1-beta-catenin complex: destruction motif binding and lysine specificity of the scf(beta-trcp1) ubiquitin ligase. Mol Cell 11(6):1445–1456

Zheng N, Wang P, Jeffrey PD, Pavletich NP (2000) Structure of a c-cbl-ubch7 complex: ring domain function in ubiquitin-protein ligases. Cell 102(4):533–539

Zheng N, Schulman BA, Song L, Miller JJ, Jeffrey PD, Wang P, Chu C, Koepp DM, Elledge SJ, Pagano M, Conaway RC, Conaway JW, Harper JW, Pavletich NP (2002) Structure of the cul1-rbx1-skp1-f boxskp2 scf ubiquitin ligase complex. Nature 416(6882):703–709