Biochimica et Biophysica Acta 1834 (2013) 1461–1467

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

hCKSAAP_UbSite: Improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties

Zhen Chen [a,1], Yuan Zhou [a,1], Jiangning Song [b,c,*], Ziding Zhang [a,**]

[a] State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China
[b] National Engineering Laboratory for Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China
[c] Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, VIC 3800, Australia

ARTICLE INFO

Article history:
Received 2 February 2013
Received in revised form 28 March 2013
Accepted 9 April 2013
Available online 19 April 2013

Keywords:
Ubiquitination site
Post-translational modification
Composition of k-spaced amino acid pairs
Amino acid physicochemical properties

ABSTRACT

As one of the most common post-translational modifications, ubiquitination regulates the quantity and function of a variety of proteins. Experimental and clinical investigations have also suggested the crucial roles of ubiquitination in several human diseases. The complicated sequence context of human ubiquitination sites revealed by proteomic studies highlights the need of developing effective computational strategies to predict human ubiquitination sites. Here we report the establishment of a novel human-specific ubiquitination site predictor through the integration of multiple complementary classifiers. Firstly, a Support Vector Machine (SVM) classier was constructed based on the composition of *k*-spaced amino acid pairs (CKSAAP) encoding, which has been utilized in our previous yeast ubiquitination site predictor. To further exploit the pattern and properties of the ubiquitination sites and their flanking residues, three additional SVM classifiers were constructed using the binary amino acid encoding, the AAindex physicochemical property encoding and the protein aggregation propensity encoding, respectively. Through an integration that relied on logistic regression, the resulting predictor termed hCKSAAP_UbSite achieved an area under ROC curve (AUC) of 0.770 in 5-fold cross-validation test on a class-balanced training dataset. When tested on a class-balanced independent testing dataset that contains 3419 ubiquitination sites, hCKSAAP_UbSite has also achieved a robust performance with an AUC of 0.757. Specifically, it has consistently performed better than the predictor using the CKSAAP encoding alone and two other publicly available predictors which are not human-specific. Given its promising performance in our large-scale datasets, hCKSAAP_UbSite has been made publicly available at our server (http://protein.cau.edu.cn/cksaap_ubsite/).

1. Introduction

Ubiquitination is a post-translational modification process where a residue (i.e. lysine for most cases) is covalently attached with single or multiple ubiquitin(s) [1,2]. This modification on one hand tags proteins to be degraded through the proteolytic system [3], on the other hand regulates a wide spectrum of biological processes [4] including but not limited to transcription [5], endocytosis [6] and cell cycle [7]. Despite the vital role of ubiquitination modification, the number of publicly available ubiquitination site prediction server is still very limited. UbiPred uses the mean values of the selected physicochemical properties of amino acids as input to a Support Vector Machine (SVM) classifier [8]. UbPred is a yeast-centric ubiquitination site predictor which employs extensive sequence, structural and evolutionary features [9]. In our previous work, we developed a yeast ubiquitination site predictor based on the composition of *k*-spaced amino acid pair (CKSAAP) encoding [10]. The CKSAAP encoding, originally termed as the collocated amino acid pair encoding [11], has been proposed to solve a number of protein structure-related classification problems including the prediction of flexible/rigid region [12], protein crystallization ability [11], protein structural class [13] and membrane protein type [14]. This encoding has been recently exploited to develop a variety of post-translational modification site predictors [10,15–17]. The developed CKSAAP_UbSite predictor achieved an overall performance improvement in comparison to several other predictors on multiple datasets [10].

Here, we report the development of a new predictor specifically designed for human ubiquitination site prediction. The reason for doing this is three folds. First, given the intriguing relationship between ubiquitination and key human health topics like cancer [18],

Abbreviations: CKSAAP, composition of *k*-spaced amino acid pairs; SVM, Support Vector Machine; ROC, receiver-operating-characteristic; AUC, area under ROC curve; AUC01, the relative area under ROC curve limiting an up to 10% false positive rate

* Correspondence to: J. Song, Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, VIC 3800, Australia. Tel.: +61 3 99029304.

** Corresponding author. Tel.: +86 10 62734376.
E-mail addresses: jiangning.song@monash.edu (J. Song), zidingzhang@cau.edu.cn (Z. Zhang).

[1] These authors contributed equally to this work.

1570-9639/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbapap.2013.04.006

virus infection [19] and inflammation [20], there is currently no human-centric ubiquitination site prediction server. Indeed, many users of our previous yeast-specific server expressed urgent desire for such a server by even requesting for the prediction outputs on human proteins through our yeast-specific predictor. This urgent need to develop human-specific ubiquitination site prediction server has motivated this study. Second, all of the aforementioned methods were trained on relatively small datasets with no more than 500 ubiquitination sites, because of the limited data availability at that time when they were established. It has been found that predictors trained on a dataset of limited size and coverage often failed to identify novel ubiquitination sites [9,21]. Therefore, predictors trained on a large proteomic dataset are deemed essential in order to better characterize the underlying *bona fide* ubiquitination motifs at the proteome scale. Recent breakthrough of proteomic techniques has resulted in a rapid growth of ubiquitination site data by orders of magnitudes [22–24], providing unprecedented opportunities for further improvement of ubiquitination predictors. Third and utmost, the sequence context of human ubiquitination sites differs significantly from the yeast counterpart (see the sequence logos [25] in Fig. 1 for an intuitive illustration). Based on preliminary tests, it has been observed that prediction of human ubiquitination sites based on yeast-centric predictors has often ended with a frustrating and dissatisfying accuracy [10,21,26].

It should be noted that, however, our human ubiquitination site prediction server (termed hCKSAAP_UbSite) is not a simple CKSAAP predictor re-trained on the human datasets. In fact, we improved the accuracy by further incorporating three informative amino acid pattern and property encoding schemes. Briefly speaking, we found that the binary encoding, which directly reflect the position-specific amino acid pattern surrounding the ubiquitination site, is complementary to the position-independent CKSAAP encoding. By further integrating two groups of amino acid properties, namely, the selected AAindex [27] physicochemical features and residue aggregation propensity, our method can achieve an even better performance when evaluated on both cross-validation test and large-scale independent test. In the following sections we first provide technical details about how the predictor was established. The performance assessment, discussion about encoding schemes and server implementation will be subsequently presented.

2. Materials and methods

Briefly, the hCKSAAP_UbSite predictor was constructed based on the integration of four SVM classifiers' prediction results. Each SVM classifier was trained with a specific set of features, which is described in Section 2.2. A summary of the computational framework of our method is available in Fig. 2.

2.1. Datasets

We collected experimentally validated ubiquitination sites by retrieving Uniprot entries or searching recent literature (see Table S1 for the complete list of literature-derived sites). To further enlarge the dataset, ubiquitination sites identified from two proteomic assays [23,24] were also incorporated. We then removed the redundant sequences using the Blastclust program (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html) with a 30% identity cutoff. The whole dataset was composed of 9537 ubiquitination sites (positive samples) from 3852 proteins. About one third of the dataset (3419 sites from 1352 proteins) was split out as the independent testing dataset and the remainder was used to train the predictor. Finally, we randomly chose equal number of non-ubiquitination sites as negative samples with one restriction in the training dataset that the sequence distance between a negative sample to any ubiquitination site in the same protein should not be smaller than 50 amino acids. This restriction is

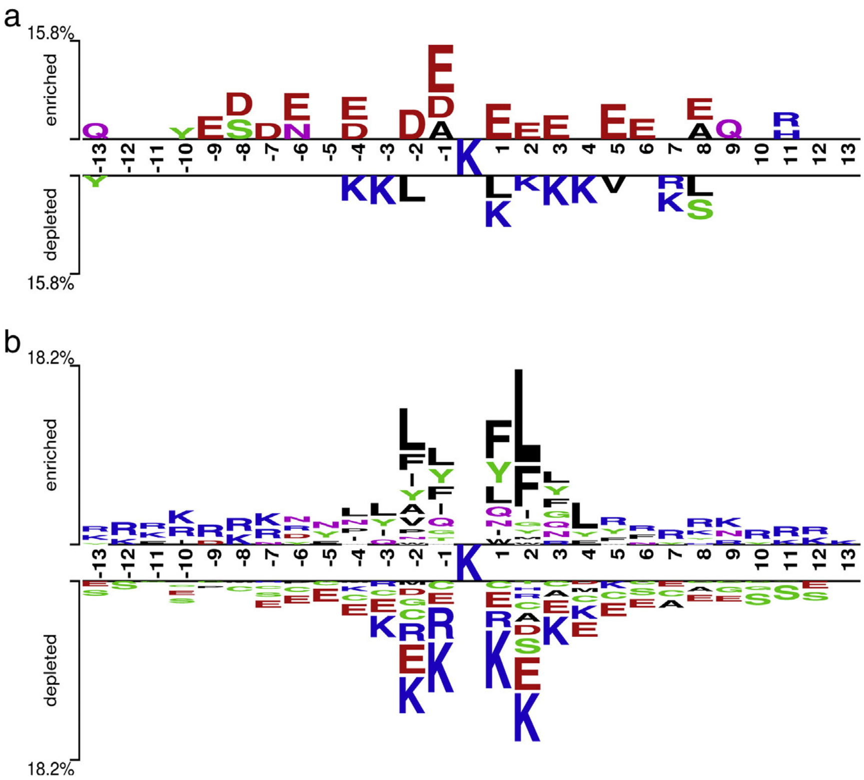


Fig. 1. Sequence logos of (a) yeast dataset and (b) human dataset. The yeast sequence logo is adapted from our previous publication [10] with permission. The significantly enriched or depleted residues at individual positions surrounding the ubiquitination sites are illustrated. These pictures were rendered using Two Sample Logo server (http://www.twosamplelogo.org/) with default settings, with the exception that the human dataset's sequence logo was enlarged for clarity.

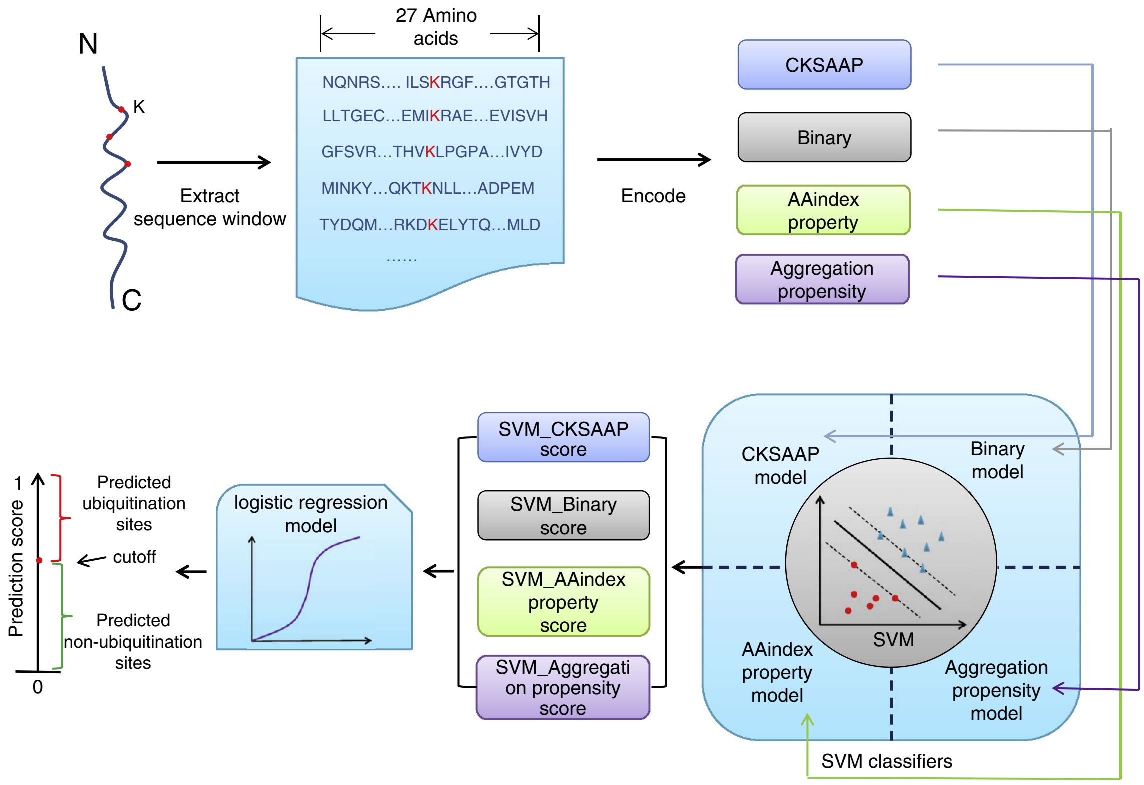


Fig. 2. The computational framework of our predictor. First, a 27-amino acid sequence window centered on each ubiquitinated/non-ubiquitinated lysine was extracted from the protein. The sequence window was then encoded in four different fashions described in Section 2.2. These four groups of encoded features were separately input into SVM models to generate four independent sets of SVM prediction scores. Finally, the prediction scores of four SVM classifiers were integrated via logistic regression. A certain cutoff was applied to the combined prediction score to distinguish the ubiquitination site.

applied to avoid overlapping between the surrounding sequence window of a ubiquitination site and that of a non-ubiquitination site. Note that there is no such restriction in the testing dataset and the performance assessment through the independent test is therefore not over-estimated. The training and testing datasets are available from the Download page of our server (http://protein.cau.edu.cn/cksaap_ubsite/download/DatasetForhCKSAAP_UbSite.rar).

2.2. Feature encoding

For each ubiquitinated or non-ubiquitinated lysine, a sequence window that contains the central lysine and its ±13 flanking residues was extracted. This window size was previously optimized in the yeast dataset [10]. Our preliminary test on human dataset also showed that this window size is optimal for our baseline encoding (i.e., the CKSAPP encoding). It is possible that the central lysine is located near N- or C-terminus of a protein sequence. In such a case, a truncated sequence window was used for CKSAAP encoding. However, for the other encodings, the size of the sequence window was fixed to 27 residues and the missing positions were filled with residue "X"s in this study.

2.2.1. CKSAAP encoding

A sequence window can be represented as a combination of multiple k-spaced amino acid pairs [10], for example, "ExxE", whose space number k is equal to 2. We calculated the composition of each possible k-spaced amino acid pair i by the following equation:

$$CKSAAP[i = 1, 2, \ldots, (k_{max} + 1) \times 400] = N_i/(W{-}k{-}1) \quad (1)$$

where N_i is the count of the k-spaced amino acid pair i and W is the window size. The maximum space taken into consideration (k_{max}) was optimized to be 5, resulting in a 2400-dimensional feature vector.

2.2.2. Binary encoding

We encoded amino acid at each position using a 21-dimonsional binary vector (20 amino acid plus the aforementioned gap-filling residue "X"). For example, residue A was encoded as the vector (1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0). The central lysine was omitted for encoding, resulting in a total 546-dimensional feature vector.

2.2.3. Amino acid property encodings

The primary physicochemical properties were extracted from the AAindex database [27]. An amino acid at each position can be represented as a feature vector that contains 531 raw values of the AAindex amino acid properties. The number of AAindex amino acid properties used is slightly smaller than the total number of AAindex entries, because those entries with missing or uncertain information were removed before the calculation. We note that the total dimension of features would be prohibitively high if all of these position-specific amino acid properties were considered. Therefore, 200 most informative AAindex features were selected via the minimum redundancy maximum relevance feature selection approaches (see related references [21,28] for details). To avoid over-fitting, the feature selection was performed based solely on the training samples during cross-validation tests.

We also introduced another derivative amino acid property, i.e., the aggregation propensity, to depict the physiochemical properties of ubiquitination sites and their sequence neighbors. Similar to AAindex, the aggregation propensity score for each residue in the sequence window was encoded as an individual feature. Thus the dimension of the aggregation propensity feature vector should be 27. We used the

FoldAmyloid tool [29] to predict the aggregation propensity score for each residue in a protein.

2.3. Predictor training and testing

Each group of feature vectors (i.e. CKSAAP, Binary, AAindex and residue aggregation propensity) was used as input to train individual SVM classifiers. We used SVM-light software package (http://svmlight.joachims.org/) to build the SVM classifiers. The parameters *c* and *gamma* of each SVM classifier were optimized separately during cross-validation tests utilizing grid search strategy. The optimized cost parameters *c* of each classifier were 2, 8, 32 and 2, respectively; while the optimized *gamma* parameters of each classifier were 2^3, 2^{-5}, 2^{-15} and 2^{-7}, respectively. The final prediction score *P* of our hCKSAAP_UbSite predictor was deduced through a logistic regression approach, which integrated the outputs of the four individual SVM classifiers as follows:

$$\log\left(\frac{P}{1-P}\right) = \sum_{i=1}^{4} b_i S_i + a \tag{2}$$

where the coefficient b_i of each SVM output S_i and the constant term a were deduced through the regression process. By definition, the output *P* denotes the probability of the central lysine to be ubiquitinated. The generalized linear model (i.e. the glm function) in R (http://www.R-project.org/) was used to generate the logistic regression model.

Five-fold cross-validation tests were performed to preliminarily assess the performance of the SVM classifiers and the hCKSAAP_UbSite predictor. In these tests, each time 20% of the data were selected as the testing fold. The testing fold's SVM outputs were predicted by the SVM classifiers trained with the rest of data (the training fold). After the SVM outputs of the whole training dataset were predicted via cross-validation, we generated five logistic regression models. Each logistic regression model was generated with the SVM outputs of one training fold and utilized to integrate its corresponding non-overlapping testing fold's SVM outputs into final prediction scores. We noted possible overestimation of the performance through cross-validation tests. Therefore, the predictor was also strictly benchmarked on the independent testing dataset. The SVM outputs of the testing dataset were predicted by the four SVM classifiers trained with the whole training dataset. And the final prediction scores were deduced through the logistic regression model which was generated with the whole training dataset's SVM outputs. The results shows no prominent performance decrease for the hCKSAAP_UbSite on the independent testing dataset (see Section 3.2 for details), a significant overestimation via cross-validation test turns out to be unlikely.

We plotted receiver-operating-characteristic (ROC) curves by varying the thresholds. Two measurements extracted from the ROC curve, i.e. the total area under ROC curve (AUC) and the relative area under ROC curve limiting an up to 10% false positive rate (AUC01), were exploited for robust performance evaluation. For both measurements, 1 implies perfect performance. A value of 0.5 and 0.05 indicate random prediction for AUC and AUC01, respectively.

3. Results and discussion

3.1. Construction of hCKSAAP_UbSite by integrating amino acid pattern and properties

CKSAAP encoding is a sequence-based encoding that describes the spectrum of spaced amino acid pair surrounding a given functional site. Our previous yeast-specific ubiquitination site predictor CKSAAP_UbSite is based on this encoding [16]. To scrutinize whether a SVM classifier using this encoding alone could also effectively predict human ubiquitination sites, we retrained the SVM classifier using the comprehensive human training dataset described in Section 2.1. After parameter optimization, this SVM classifier could reach an AUC of 0.735 through five-fold cross validation on this balanced dataset (Table 1), indicating that the overall performance is acceptable. Because the CKSAAP encoding was also proven to be useful for prediction of yeast ubiquitination sites [10], we attempted to interrogate possible biological implications of this encoding by analyzing the most informative *k*-spaced amino acid pairs in yeast and human datasets. The top 50 informative *k*-spaced amino acid pairs were selected via the minimum redundancy maximum relevance feature selection approaches [21,28] (Table S2). As depicted in Figure S2, the residue usage of the most informative *k*-spaced amino acid pairs are not uniformly distributed for both yeast and human (Kolmogorov–Smirnov test, $P = 0.010$ and $P = 0.024$). But the residue usages between these two groups of *k*-spaced amino acid pairs seem not to be totally different (Kolmogorov–Smirnov test, $P = 0.81$), indicating that only a few residues exhibit prominent bias of usage. An intuitive and simple reason of such bias could be attributed to the dramatic divergent evolution of the ubiquitination E3 ligase enzymes from yeast to human [2,30]. However, apart from this, we speculate that some other related factors also exist. An example is the overrepresentation of the acidic residue (D and E) pairs for yeast ubiquitination sites. This may be partially explained by the residue composition near the catalytic core of the ubiquitination E2 conjugating enzymes. Indeed, some of the yeast ubiquitination E2 enzymes allocate more positively charged residues in proximity to their catalytic core (e.g. the Fig. 3 in [31]), thereby favoring acidic residues around the yeast ubiquitination sites. Another example is the enrichment of amino acid pairs composed of one or more hydrophobic residues (e.g. L, F and Y) in the vicinity of the human ubiquitination sites. This preference can be correlated with the frequent occurrence of human ubiquitination sites in the folded protein domains [32]. An alternative explanation is related to the proteasome activity. Biochemical assays have revealed strong catalytic activity of the human proteasome at hydrophobic residues of the substrates [33]. Considering the close relationship between ubiquitination and proteasomal degradation [3], it is possible that the allocation of some hydrophobic residues around the human ubiquitination sites might offer advantages of facilitating the protein degradation. A third example is the significant depletion of cysteine amino acid pairs around the yeast and human's ubiquitination sites. Recent proteomic experiments have revealed a considerable quantity of cysteine ubiquitination sites in yeast proteins [34], indicating that the co-occurrence of cysteine ubiquitination and lysine ubiquitination is possible. It is therefore plausible that cysteine residues are selected against in the vicinity of the ubiquitinated lysine to avoid direct competition, especially in human whose ubiquitination system is larger and more complicated [30]. Although the above implications could be interesting, we wish to emphasize that all of the above speculations should be carefully examined with biochemical assays in the near future.

In contrast to the overall performance of the CKSAPP encoding, the accuracy is likely to be decreased when the false positive rate is limited to be low (i.e., AUC01 = 0.169, see also Table 1). The CKSAAP encoding was capable of robustly discovering the cryptic information about the possible motifs around the ubiquitination sites irrespective of the exact positions where the motifs are located.

Table 1
The performance of single and combined classifiers.

Predictor	AUC	AUC01
CKSAAP	0.735	0.169
CKSAAP + Binary	0.761	0.206
CKSAAP + Binary + AAindex	0.766	0.216
hCKSAAP_UbSite (CKSAAP + Binary + AAindex + Aggregation propensity)	0.770	0.226

However, discarding position-specific information on the other hand might reduce the sensitivity of a predictor. Indeed, the CKSAAP-based classifier could predict only 30.3% of the ubiquitination sites at the 90% specificity level, evaluated by the cross-validation test (Fig. S1). To examine if there is any position-specific information about the human ubiquitination sites that was omitted by CKSAAP, we generated sequence logos [25] to depict the sequence pattern around human ubiquitination sites. In addition to the significant divergence of the amino acid usage propensity between yeast and human ubiquitination sites, the overall distribution of informative residues for these two groups of sites is significantly different (Fig. 1). The yeast ubiquitination sites are characterized by widespread glutamic acid (E) residues across the upstream -9 and downstream $+8$ positions, while the amino acid preference in human ubiquitination sites seems to differ from one position to another. For example, the arginine (R) residues are depleted at the nearest positions from the human ubiquitination sites but become enriched at relatively distal positions (Fig. 1). Therefore, it is of particular interest to examine the possibility of boosting the predictor if the position-specific amino acid pattern encoding was introduced.

Consequently, the binary encoding-based SVM classifier was incorporated to the predictive framework. After integrating such a position-specific encoding, a major augment of performance was observed (Table 1). Specifically, an 18% increase of AUC01 (from 0.169 to 0.206) indicates a significant improvement in the predictor's sensitivity. Although this binary encoding is position-specific, we would like to emphasize that it is not equivalent to the position-specific scoring matrix. The latter reflects the evolutionary information of functional sites rather than the sequence pattern of the sites themselves. In fact, results of our cross-validation tests suggest that even profile-CKSAAP [13], a sophisticated method integrating CKSAAP encoding with the position-specific scoring matrix, failed to significantly improve the original CKSAAP classifier (data not shown).

To better exploit position-specific information, a third SVM classifier trained with amino acid property encoding, namely positional AAindex value, was established and added to the predictive framework. To avoid use of excessive number of features, only 200 most informative features were considered. Interestingly, all of these top features came from positions -2, -1, $+1$ and $+2$ to the central lysine. This highlights the importance of the proximal residues to regulate local physiochemical properties and control the specificity of ubiquitination. As the physicochemical properties were utilized by UbiPred in a distinct fashion [8], we here compared the informative properties demonstrated by Tung et al. [8] and those proposed by us. In our human training dataset, the most informative features contain polarity, principal component I, atom-based hydrophobic moment and helix termination parameter at position $j + 1$ (AAindex IDs: ZIMJ680103, SNEP660101, EISD860102 and FINA910104), all of which are highly discriminative at 4 out of 27 positions in the local sequence window. Due to the differences in the datasets and feature selection methods, the top features reported by the two studies are largely not overlapping with each other. For example, although the average reduced distance for side chain (AAindex ID: MEIH800102) has been shown to be important for ubiquitination site prediction [8], this feature was not included in our top feature list. Nevertheless, there is a limited agreement between the two sets of informative features. For example, the linker propensity (AAindex ID: GEOR030108) was suggested to be informative in both studies. A more comprehensive list of the top informative amino acid properties selected based on our dataset is provided in Table S3.

We are also aware of the fact that some derivative residue properties, such as secondary structure, solvent accessibility and disorder propensity have been previously proposed to predict ubiquitination sites [9,21,26]. Therefore, we predicted the secondary structure of each position in the sequence window by PSIPRED [35] and encoded this information as an 81-dimensional (3×27) binary feature vector. The SVM classifier trained with this feature showed an unsatisfactory performance (AUC = 0.588), indicating that the secondary structure is not very informative for predicting human ubiquitination sites. Similar results have been obtained for solvent accessibility and disorder propensity when these features were encoded in binary fashion (data not shown). In contrast, classifiers that exploited the residue aggregation propensity score showed an unexpected moderate performance (AUC = 0.676). Moreover, incorporation of the output of this SVM classifier as the fourth component of the predictor could further improve the performance, especially in terms of AUC01 (Table 1). We calculated the mean residue aggregation propensity at each position of the sequence window, and found that ubiquitinated lysines and residues in the proximity tend to have higher aggregation propensity than the non-ubiquitinated counterparts (paired t-test, $P = 1.53 \times 10^{-8}$). To the best of our knowledge, no direct association between the residue aggregation propensity and ubiquitination has been previously reported. Here, we provide two hypotheses as to how this association is established. The first one is that, as a major function of the ubiquitination is to tag the misfolded or aggregated polypeptides in the cell to be degraded through proteolytic or autophagic pathways [3,23,36], organisms tend to allocate the ubiquitination sites onto aggregation prone regions to better monitor the *in vivo* aggregation of polypeptides. An alternative explanation, however, is not directly related to the aggregation phenomenon. Instead, we note the fact that ubiquitination is often catalyzed by a protein complex through direct protein–protein interaction (for example, see Tian et al. [37]). We thus speculate that ubiquitination sites have a tendency to be localized on or close to the protein–protein interaction interface whose aggregation propensity is naturally higher [38]. Whether these hypotheses hold or not, however, should be tested through experimental investigations and is beyond the scope of this study.

Our hCKSAAP_UbSite builds upon the integration of the four aforementioned SVM classifiers using logistic regression. It should be noticed that the overall performance may also be attributed to the reasonable integration of all the four predictors. Indeed, casual integration of four classifiers by summing their prediction scores did not improve the predictor as much as the logistic regression in the cross-validation test (AUC = 0.749). According to the final logistic regression model, the output scores of CKSAAP-based SVM classifier have the highest *Z*-score ($Z = 21.7$), indicating that it has the highest discriminative capability. We used McKelvey and Zavoina's Pseudo R^2 [39] to estimate how much variation is explained by individual sets of prediction scores or their combination. The four sets of prediction scores taken together accounted for about 30.5% of the variation. The best-fit logistic regression using the outputs of CKSAAP-based SVM only reported that about 23.0% variation was explained. This justifies the usage of CKSAAP as the baseline classifier in our predictor while emphasizing on the indispensible importance of the other three classifiers.

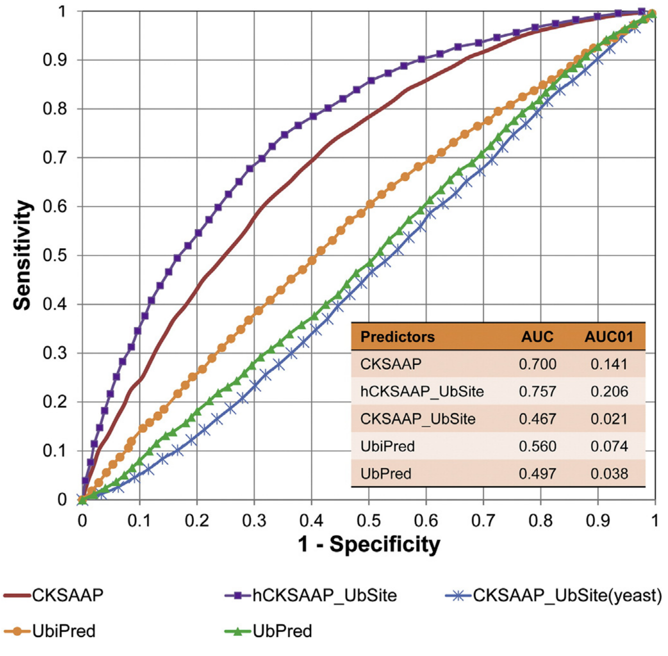
3.2. Performance assessment of hCKSAAP_UbSite via independent test

We have demonstrated that hCKSAAP_UbSite could achieve a promising prediction performance in the 5-fold cross-validation test. To objectively evaluate our predictor, we further tested our method on an independent dataset. As described, this independent dataset has 3419 ubiquitination sites and an equal number of non-ubiquitination as negative samples. The size of our independent dataset is more than 10 folds larger than the independent datasets used in previous studies [10,21] which guarantees a much more robust performance assessment. Moreover, because the proteins in the independent dataset were randomly selected from a non-redundant group of proteins and no homologous information has been exploited by our predictor, the performance is less likely to be over-estimated due to the presence of sequence homology.

The result for hCKSAAP_UbSite of the independent test generally agreed well with that of the cross-validation tests, with a slight decrease to some extent (Fig. 3 and Table S4). In addition, we are aware of a noticeable decrease of the CKSAAP-based SVM classifier's performance. One possible reason is that the high dimension of the CKSAAP feature vector makes this SVM classifier prone to be over-fitting through model optimization. Nevertheless, a considerable increase of the overall performance after the integration of the other three classifiers was observed again. The final predictor hCKSAAP_UbSite achieved an AUC as high as 0.757 on this independent test. Note that the gap between the performance assessed by cross-validation test (Table 1) and that assessed by independent test (Fig. 3 and Table S4) was also narrowed down via this integration, implying the necessity of integrating all four classifiers to achieve the robust prediction of ubiquitination sites. Finally, we also tested the performance of our yeast ubiquitination site predictor CKSAAP_UbSite [10] on this human dataset. As expected, the yeast-centric predictor did not effectively predict human ubiquitination sites (Fig. 3 and Table S4). Similar results were obtained when the other two ubiquitination site predictors UbPred and UbiPred were evaluated on this independent test. None of these predictors was specifically designed to predict human ubiquitination sites, and none of them produced predictions with satisfying accuracy (Fig. 3 and Table S4). These results confirm again that construction of a human-specific predictor is necessary and crucial.

3.3. Server implementation

To facilitate the users, hCKSAPP_UbSite has been integrated into our existing CKSAAP_UbSite server (http://protein.cau.edu.cn/cksaap_ubsite/). Neither registration nor license acquisition is required for academic usage of this server. Users have options to choose which predictor (yeast- or human-specific) to be used when submitting their queries. For prediction analysis of protein sequences from some distinct groups of organisms like plants, the users should be very careful because either predictor may not provide satisfactory predictions in these distinct organisms.



Predictors	AUC	AUC01
CKSAAP	0.700	0.141
hCKSAAP_UbSite	0.757	0.206
CKSAAP_UbSite	0.467	0.021
UbiPred	0.560	0.074
UbPred	0.497	0.038

Fig. 3. The ROC curves of several predictors on the independent test. The predictors include the CKSAAP-based SVM classifier (CKSAAP), the predictor integrating four SVM classifiers (hCKSAAP_UbSite), our previous yeast ubiquitination site predictor (CKSAAP_UbSite) and two publicly available ubiquitination site predictors (UbiPred and UbPred). The UbiPred prediction results were obtained by submitting the independent test set directly to its online server. The downloadable version of UbPred was used because a direct submission of such a large-scale test set to the UbPred server will result in a prohibitively heavy burden. The AUC and AUC01 of the corresponding predictors are also provided.

After a raw- or FASTA-formatted sequence is submitted to our server, the user will be redirected to the result webpage where the prediction score of individual SVM classifier and the final prediction score will be presented. Fig. S3 provides a sample screenshot of the result webpage. Generally, it takes 1 to 3 min to predict the ubiquitination site for a protein sequence shorter than 1000 amino acids. But there is no need to bookmark this webpage or keep it open because a hyperlink to this result page will be sent to the user's E-mail address when the task is accomplished. A prediction result is usually kept for one month in our server. As a public server, we have also made recent prediction results accessible in the job list.

4. Concluding remarks

We have presented a new ubiquitination site predictor specifically developed to predict human ubiquitination sites. Different from our previous work [10], in this study, the SVM classifiers utilizing various encodings of ubiquitination sites and their sequence neighbors have been integrated into a logistic regression framework. The features used in our method include a variety of important aspects of amino acid patterns and propensities. To the best of our knowledge, one such feature, residue aggregation propensity, is proposed as an indicator of ubiquitination site for the first time. As a result, our novel predictor has consistently achieved a better performance and robustness compared with the predictor using the CKSAAP encoding alone. Due to this considerable performance improvement on our datasets, we have made hCKSAAP_UbSite freely available as a component of our public ubiquitination site prediction server.

Given the significant difference in amino acid preference between the sequence neighbors of human ubiquitination sites and yeast counterparts, a yeast ubiquitination site predictor usually fails to predict human ubiquitination sites with high accuracy. However, it has been noticed that human ubiquitination sites identified by different high-throughput proteomic screens also do not well agree with each other (see for example [40]). Therefore, it is possible that our method can be further improved with the increasing quality of proteomic data in the near future.

Acknowledgements

We thank Xin Miao at China Agricultural University for assistance in the preliminary data analysis. This work was partially supported by the National Key Basic Research Program of China (2009CB918802), the National Natural Science Foundation of China (31070259 and 61202167), the grants from the National Health and Medical Research Council of Australia (NHMRC), the Hundred Talents Program of the Chinese Academy of Sciences (CAS), the Knowledge Innovative Program of CAS (KSCX2-EW-G-8) and Tianjin Municipal Science & Technology Commission (10ZCKFSY05600). JS is an NHMRC Peter Doherty Fellow and a Recipient of the Hundred Talents Program of CAS.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bbapap.2013.04.006.

References

[1] O. Kerscher, R. Felberbaum, M. Hochstrasser, Modification of proteins by ubiquitin and ubiquitin-like proteins, Annu. Rev. Cell Dev. Biol. 22 (2006) 159–180.
[2] T. Gao, Z. Liu, Y. Wang, H. Cheng, Q. Yang, A. Guo, J. Ren, Y. Xue, UUCD: a family-based database of ubiquitin and ubiquitin-like conjugation, Nucleic Acids Res. 41 (2013) D445–D451.
[3] M.H. Glickman, A. Ciechanover, The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction, Physiol. Rev. 82 (2002) 373–428.

[4] V. Kirkin, I. Dikic, Role of ubiquitin- and Ubl-binding proteins in cell signaling, Curr. Opin. Cell Biol. 19 (2007) 199–205.
[5] Y. Zhang, Transcriptional regulation by histone ubiquitination and deubiquitination, Genes Dev. 17 (2003) 2733–2740.
[6] K. Haglund, S. Sigismund, S. Polo, I. Szymkiewicz, P.P. Di Fiore, I. Dikic, Multiple monoubiquitination of RTKs is sufficient for their endocytosis and degradation, Nat. Cell Biol. 5 (2003) 461–466.
[7] L.M. DeSalle, M. Pagano, Regulation of the G1 to S transition by the ubiquitin pathway, FEBS Lett. 490 (2001) 179–189.
[8] C.W. Tung, S.Y. Ho, Computational identification of ubiquitylation sites from protein sequences, BMC Bioinf. 9 (2008) 310.
[9] P. Radivojac, V. Vacic, C. Haynes, R.R. Cocklin, A. Mohan, J.W. Heyen, M.G. Goebl, L.M. Iakoucheva, Identification, analysis, and prediction of protein ubiquitination sites, Proteins 78 (2010) 365–380.
[10] Z. Chen, Y.Z. Chen, X.F. Wang, C. Wang, R.X. Yan, Z. Zhang, Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs, PLoS One 6 (2011) e22930.
[11] K. Chen, L. Kurgan, M. Rahbari, Prediction of protein crystallization using collocation of amino acid pairs, Biochem. Biophys. Res. Commun. 355 (2007) 764–769.
[12] K. Chen, L.A. Kurgan, J. Ruan, Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs, BMC Struct. Biol. 7 (2007) 25.
[13] K. Chen, L.A. Kurgan, J. Ruan, Prediction of protein structural class using novel evolutionary collocation-based sequence representation, J. Comput. Chem. 29 (2008) 1596–1604.
[14] K. Chen, Y. Jiang, L. Du, L. Kurgan, Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs, J. Comput. Chem. 30 (2009) 163–172.
[15] Y.Z. Chen, Y.R. Tang, Z.Y. Sheng, Z. Zhang, Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs, BMC Bioinf. 9 (2008) 101.
[16] X.B. Wang, L.Y. Wu, Y.C. Wang, N.Y. Deng, Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs, Protein Eng. Des. Sel. 22 (2009) 707–712.
[17] X. Zhao, W. Zhang, X. Xu, Z. Ma, M. Yin, Prediction of protein phosphorylation sites by using the composition of k-spaced amino acid pairs, PLoS One 7 (2012) e46302.
[18] M.H. Lee, R. Zhao, L. Phan, S.C. Yeung, Roles of COP9 signalosome in cancer, Cell Cycle 10 (2011) 3057–3066.
[19] H.J. Sharifi, A.M. Furuya, C.M. de Noronha, The role of HIV-1 Vpr in promoting the infection of nondividing cells and in cell cycle arrest, Curr. Opin. HIV AIDS 7 (2012) 187–194.
[20] F. Tokunaga, K. Iwai, LUBAC, a novel ubiquitin ligase for linear ubiquitination, is crucial for inflammation and immune responses, Microbes Infect. 14 (2012) 563–572.
[21] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, Y. Li, Prediction of lysine ubiquitination with mRMR feature selection and analysis, Amino Acids 42 (2012) 1387–1395.
[22] P.C. Chen, C.H. Na, J. Peng, Quantitative proteomics to decipher ubiquitin signaling, Amino Acids 43 (2012) 1049–1060.
[23] S.A. Wagner, P. Beli, B.T. Weinert, M.L. Nielsen, J. Cox, M. Mann, C. Choudhary, A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles, Mol. Cell. Proteomics 10 (2011), (M111 013284).
[24] J.M. Danielsen, K.B. Sylvestersen, S. Bekker-Jensen, D. Szklarczyk, J.W. Poulsen, H. Horn, L.J. Jensen, N. Mailand, M.L. Nielsen, Mass spectrometric analysis of lysine ubiquitylation reveals promiscuity at site level, Mol. Cell. Proteomics 10 (2011), (M110 003590).
[25] V. Vacic, L.M. Iakoucheva, P. Radivojac, Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments, Bioinformatics 22 (2006) 1536–1537.
[26] X. Zhao, X. Li, Z. Ma, M. Yin, Prediction of lysine ubiquitylation with ensemble classifier and feature selection, Int. J. Mol. Sci. 12 (2011) 8347–8361.
[27] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, AAindex: amino acid index database, progress report 2008, Nucleic Acids Res. 36 (2008) D202–D205.
[28] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, J. Bioinform. Comput. Biol. 3 (2005) 185–205.
[29] S.O. Garbuzynskiy, M.Y. Lobanov, O.V. Galzitskaya, FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence, Bioinformatics 26 (2010) 326–332.
[30] W. Li, M.H. Bengtson, A. Ulbrich, A. Matsuda, V.A. Reddy, A. Orth, S.K. Chanda, S. Batalov, C.A. Joazeiro, Genome-wide and functional annotation of human E3 ubiquitin ligases identifies MULAN, a mitochondrial E3 that regulates the organelle's dynamics and signaling, PLoS One 3 (2008) e1487.
[31] M. Sadowski, R. Suryadinata, X. Lai, J. Heierhorst, B. Sarcevic, Molecular basis for lysine specificity in the yeast ubiquitin-conjugating enzyme Cdc34, Mol. Cell. Biol. 30 (2010) 2316–2329.
[32] T. Hagai, A. Azia, A. Toth-Petroczy, Y. Levy, Intrinsic disorder in ubiquitination substrates, J. Mol. Biol. 412 (2011) 319–324.
[33] J.L. Harris, P.B. Alper, J. Li, M. Rechsteiner, B.J. Backes, Substrate specificity of the human proteasome, Chem. Biol. 8 (2001) 1131–1141.
[34] L.M. Starita, R.S. Lo, J.K. Eng, P.D. von Haller, S. Fields, Sites of ubiquitin attachment in *Saccharomyces cerevisiae*, Proteomics 12 (2012) 236–240.
[35] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292 (1999) 195–202.
[36] P.K. Kim, D.W. Hailey, R.T. Mullen, J. Lippincott-Schwartz, Ubiquitin signals autophagic degradation of cytosolic proteins and peroxisomes, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 20567–20574.
[37] W. Tian, B. Li, R. Warrington, D.R. Tomchick, H. Yu, X. Luo, Structural analysis of human Cdc20 supports multisite degron recognition by APC/C, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 18419–18424.
[38] S. Pechmann, E.D. Levy, G.G. Tartaglia, M. Vendruscolo, Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 10159–10164.
[39] R.D. McKelvey, W. Zavoina, A statistical model for the analysis of ordinal level dependent variables, J. Math. Sociol. 4 (1975) 103–120.
[40] M.J. Emanuele, A.E. Elia, Q. Xu, C.R. Thoma, L. Izhar, Y. Leng, A. Guo, Y.N. Chen, J. Rush, P.W. Hsu, H.C. Yen, S.J. Elledge, Global identification of modular cullin-RING ligase substrates, Cell 147 (2011) 459–474.