BRIEFINGS IN BIOINFORMATICS. VOL 16. NO 4. 640–657 doi:10.1093/bib/bbu031
Advance Access published on 10 September 2014

Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features

Zhen Chen, Yuan Zhou, Ziding Zhang and Jiangning Song

Submitted: 3rd July 2014; Received (in revised form): 18th August 2014

Abstract

Protein ubiquitination is one of the most important reversible post-translational modifications (PTMs). In many biochemical, pathological and pharmaceutical studies on understanding the function of proteins in biological processes, identification of ubiquitination sites is an important first step. However, experimental approaches for identifying ubiquitination sites are often expensive, labor-intensive and time-consuming, partly due to the dynamics and reversibility of ubiquitination. *In silico* prediction of ubiquitination sites is potentially a useful strategy for whole proteome annotation. A number of bioinformatics approaches and tools have recently been developed for predicting protein ubiquitination sites. However, these tools have different methodologies, prediction algorithms, functionality and features, which complicate their utility and application. The purpose of this review is to aid users in selecting appropriate tools for specific analyses and circumstances. We first compared five popular webservers and standalone software options, assessing their performance on four up-to-date ubiquitination benchmark datasets from *Saccharomyces cerevisiae*, *Homo sapiens*, *Mus musculus* and *Arabidopsis thaliana*. We then discussed and summarized these tools to guide users in choosing among the tools efficiently and rapidly. Finally, we assessed the importance of features of existing tools for ubiquitination site prediction, ranking them by performance. We also discussed the features that make noticeable contributions to species-specific ubiquitination site prediction.

***Keywords:** protein ubiqutination; bioinformatics; tool development; species-specific ubiquitination sites; sequence analysis; feature selection*

INTRODUCTION

Protein ubiquitination is one of the most important reversible post-translational modifications (PTMs) [1]. Conjugation of ubiquitin (Ub) to lysine residues of a target protein is regulated by the sequential activity of Ub-activating (E1), Ub-conjugating (E2) and Ub-ligating (E3) enzymes. Ubiquitination varies in the number of added Ubs, either single

Corresponding authors: Ziding Zhang, State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China. Tel: +86-10-62734376; E-mail: zidingzhang@cau.edu.cn; Jiangning Song, Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, VIC 3800, Australia Tel: +61-3-99029304; E-mail: jiangning.song@monash.edu

Zhen Chen received his PhD in Bioinformatics at State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University in 2014. His research interests include protein bioinformatics, machine learning and next-generation sequencing analysis.

Yuan Zhou is a PhD student at State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University. Since his graduation from Capital Normal University, China, he has been involved in several bioinformatics projects to analyze protein function, structures and interaction networks.

Ziding Zhang is a Professor at State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University. His research interests are protein bioinformatics and plant molecular systems biology.

Jiangning Song is a Senior Research Fellow at the Monash Bioinformatics Platform, Faculty of Medicine, Monash University, Australia. He is also a Principal Investigator at the Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences (CAS). His research interests are bioinformatics, systems biology, machine learning, systems pharmacology and enzyme engineering.

Ub or poly-Ub chains [2–4]. Ubiquitination is involved in regulating a variety of fundamental cellular processes [5, 6], including protein degradation, gene transcription, DNA repair and replication, intracellular trafficking and virus particle budding [7]. Accumulating experimental evidence suggests that changes in the ubiquitination system are closely related to cellular transformation, immune response and inflammatory responses [8].

The Ub-proteasome system targets 80% of the proteins of a eukaryotic cell for degradation [9]. However, the complete repertoire of ubiquitylated substrates and their corresponding ubiquitination sites remains to be fully characterized. Current experimental methods for identifying ubiquitination sites include site-directed mutagenesis [10] and mass spectrometry [11, 12]. With efficient purification methods such as affinity-tagged Ub, Ub antibodies or Ub-binding proteins, mass spectrometry is particularly suitable for high-throughput identification of ubiquitination sites [11, 12]. However, protein ubiquitination is a rapid and reversible PTM, so large-scale identification of ubiquitylated proteins and their ubiquitination sites is often expensive, labor-intensive and time consuming.

In parallel with experimental identification of ubiquitination sites, computational prediction of potential ubiquitination sites has become a useful strategy for complete proteome annotation, prioritization of candidate ubiquitination substrates and hypothesis-driven experimental design. Almost all proposed computational methods formulate ubiquitination site prediction as a binary classification problem, classifying each candidate lysine as either a ubiquitination or a non-ubiquitination site. These methods can predict new ubiquitination sites by learning the features of the sequence context of experimentally verified ubiquitination sites via classification algorithms. The input for a ubiquitination site predictor is generally a sequence fragment with a central lysine (K) of interest followed by *n* flanking residues on each side (i.e., the window size for the sequence fragment is $2n+1$). An appropriate scheme to encode the sequence fragment is required for the prediction algorithm. Finally, a predictor is constructed or trained using statistical or machine-learning algorithms to predict potential ubiquitination sites in other proteins or the entire proteome.

A number of computational methods have recently been developed. Tung and Ho (2008) developed the first tool for ubiquitination site prediction, named UbiPred [13], which used a Support Vector Machine (SVM) with 31 informative physicochemical features selected from published amino acid indices [14]. Radivojac *et al.* proposed a Random Forest-based predictor called UbPred, which uses 586 sequence attributes as the input feature vector [15]. Lee *et al.* (2011) designed UbSite [16], using an efficient radial basis function (RBF) kernel for identifying ubiquitination sites. We recently developed CKSAAP_UbSite [17], which is a SVM-based predictor that considers composition of *k*-spaced amino acid pairs surrounding potential ubiquitination sites. In 2012, Cai *et al.* proposed a ubiquitination site predictor based on a nearest-neighbor algorithm [18]. They performed incremental feature selection and characterized key components from an initial set of 541 features to improve prediction performance. More recently, Chen *et al.* (2013) presented a new tool termed UbiProber [19], which was specifically designed for large-scale predictions of both general and species-specific ubiquitination sites. Almost at the same time, our group developed the hCKSAAP_UbSite tool [20] by integrating the outputs of four different types of predictors. More information about these methods is in Table 1. These methods differ in the training and test datasets used, the ratio of positive versus negative samples, the sliding window size and algorithms chosen, the types of sequence or structural descriptors employed, and whether the prediction models are general or species-specific. Other notable differences among these methods include implementation as webservers or standalone software, support of batch predictions, ability to adjust prediction stringency thresholds and computational efficiency.

Despite the availability of various ubiquitination site prediction tools, an important issue is comprehensively evaluating the performance and comparing the strengths and weaknesses of the tools. A comprehensive performance evaluation of these methods will enable a better understanding of the pros and cons of the methods and assist users in choosing prediction tools for their particular circumstances. In addition, from a practical biological perspective, a systematic comparative analysis of the most important determining features of lysine ubiquitination considered by the tools will further our understanding of the underlying mechanisms of ubiquitination conjugation to target proteins. This timely comparison will facilitate bioinformaticians in identifying research directions and problems that require urgent attention, which will inform future development of better tools.

In this review, we analyzed and compared five popular webservers or standalone tools using four

Table 1: Summary of ubiquitination site prediction tools compared in this study

Tool	Species	Webserver	Algorithm	Option of batch prediction	Adjustment of prediction thresholds	Stand-alone software & Platform	Technical framework of the software	Dataset size (Ubiquitination sites/ proteins)	Ratio of positive to negative samples	Window size	Time for processing a sequence
UbiPred	Multi-species	http://iclab.life.nctu.edu.tw/ubipred/	SVM	Yes (Maximum 100 protein sequence with FASTA format at once)	No	–	–	157/105	1:1	21	Within a second
UbPred	*Saccharomyces cerevisiae*	http://www.ubpred.org/	Random Forest	No	High/Median/ Low	Windows/ Linux	Shell script(depend on MATLAB Compiler)	265/201	1:1	25	10 seconds
CKSAAP_ UbSite	*Saccharomyces cerevisiae*	http://protein.cau.edu.cn/cksaap.ubsite/	SVM	No	High/Low	Linux	PERL (Depend on SVM-Light)	263/203	1:1	27	One minute
UbSite	Multi-species	No server	SVM	–	–	–	–	385/301	1:1	41	–
mRMR_ Ub Site	Multi-species	No server	Nearest Neighbor algorithm	–	–	–	–	378/273	1:3	21	–
UbiProber	Multi-species and single species	http://bioinfo.ncu.edu.cn/ubiprober.aspx	SVM	Yes (Maximum data size)	Continuous adjustment	Windows	C# (.NET 4.0 framework)	22 192/8750	1:1	27	Within 10 seconds
CKSAAP_ UbSite	*Homo sapiens*	http://protein.cau.edu.cn/cksaap.ubsite/	SVM	No	High/Low	Linux	C++	6118/2500	1:1	27	Two minutes

current benchmark datasets for four species: *Saccharomyces cerevisiae*, *Homo sapiens*, *Mus musculus* and *Arabidopsis thaliana*. Our purpose was providing practical and informative insights about accurate protein ubiquitination site prediction. In particular, we aimed to determine: (i) whether a universal best predictor exists that can be used to predict ubiquitination sites across different species; (ii) if not, which predictor provides the best species-specific performance; (iii) whether the predictive power of existing tools can be improved; and (iv) the most important features that contribute to prediction of ubiquitination sites. To address these issues, we performed a comparative analysis by collecting four large-scale benchmark datasets from recent experimental studies and extracting features shown to be useful for prediction in previous studies. We systematically assessed the performance of three different machine-learning/statistical methods: Naïve Bayes, Random Forest and SVM with four window sizes (21, 25, 27 and 41). We assessed the statistical significance and predictive power of individual and combined features and discussed their relative importance and contribution to the identification of ubiquitination sites.

MATERIALS AND METHODS

Benchmark datasets for assessing method performance

Experimentally verified ubiquitination site datasets for *S. cerevisiae*, *H. sapiens*, *M. musculus* and *A. thaliana* were collected from five large-scale proteomics studies [21–25] and named *S.dataset* (*S. cerevisiae*), *H.dataset* (*H. sapiens*), *M.dataset* (*M. musculus*) and *A.dataset* (*A. thaliana*). Because *H.dataset* was from two large-scale proteomic studies, we considered only proteins overlapping in the two studies. We removed sequence redundancy in the datasets using the Blastclust program (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html) with a 40% identity cutoff. Experimentally identified ubiquitylated lysine residues were regarded as positive samples. An equal number of negative samples were randomly selected from the remaining lysine residues. We note that it would be difficult to identify absolute negatives that could not be ubiquitylated under any conditions. In addition, the remaining residues could contain ubiquitination sites that have not yet been experimentally verified. However, the amount of ubiquitination sites is very small compared with the number of non-ubiquitination sites, which should dominate the negative samples. Hence, we presumed the chance of including true positives in the negative samples was relatively small. The numbers of ubiquitylated proteins and ubiquitination sites for each dataset are in Table 2.

Dataset partition for assessing the importance and contribution of individual features

To assess the predictive power and relative importance of individual features, each of the four datasets (i.e. *S.dataset*, *H.dataset*, *M.dataset* and *A.dataset* in Table 2) was divided into training and independent testing datasets. That is, some proteins (~30–40% of the total dataset, dependent on the computational burden of model training) were randomly separated as the independent testing dataset. Using the above procedure, we ensured that the positive and negative samples in each dataset were from the same set of proteins, with the same ratio of 1:1 between the positive to negative samples in the training and independent testing datasets. Supplementary Table S1 summarizes the training and independent datasets used for each species. Considering that *H.dataset* had more ubiquitination sites, we further reduced the data size to save the computational time required in model construction. Thus, a subset of proteins was randomly selected from *H.dataset* to keep the number of ubiquitination sites comparable to *M.dataset*.

Prediction methods under assessment

Our main criterion for including a method in the comparison analysis is that such method has been implemented as either a web-server or a stand-alone software. Five methods were analyzed: UbiPred [13], UbPred [15], CKSAAP_UbSite [17], hCKSAAP_UbSite [20] and UbiProber [19]. Amongst these five methods, UbPred (specific for *S. cerevisiae*), CKSAAP_UbSite (specific for *S. cerevisiae*) and hCKSAAP_UbSite (specific for *H. sapiens*) are species-specific predictors, while UbiPred can be

Table 2: Statistics of the four benchmark datasets used for assessing the performance of different methods

Dataset	Number of ubiquitylated proteins	Number of ubiquitination sites
S.dataset	418	820
A.dataset	167	204
M.dataset	4040	13 973
H.dataset	3657	32 756

used to predict ubiquitination sites for multiple species. UbiProber predictions can be either general (UbiProber_Combined) or species-specific (e.g. UbiProber_H.sapiens, UbiProber_M.musculus or UbiProber_S.cerevisiae). More information about these methods is in Table 1.

Performance assessment of individual features

To examine the importance and contribution of individual features, we extracted features used for ubiquitination site prediction in previous studies and grouped them into 10 feature types. These features are summarized in Table 3. It is worth mentioning that AAC encoding was constructed using the local windows of length $w_{in} \in (3, 7, 11, 21, 27, 31, 41)$. Fifteen different k values (i.e., $k = 3, 5, 7, 9, 11, 15, 21, 31, 41, 51, 61, 71, 81, 91, 101$) were adopted in the KNN encoding. We used FoldAmyloid [30] and VSL2 [31, 32] to predict the aggregation propensity and disorder scores for each residue in a protein. For each residue, the aggregation propensity score in the sequence window was encoded as an individual feature, while the disorder score was averaged within $w_{in} \in (3, 7, 11, 21, 27, 31, 41)$. The PSSM profile was obtained by running PSI-BLAST against the NCBI nr database with parameters -*h* of 0.0001 and–*j* of 3. The 42 outputs (20-dimensional PSSM vector, 20-dimensional weighted observed percentages and 2-dimensional relative weight of gapless real matches to pseudocounts) in each PSSM row were averaged over $w_{in} \in (3, 7, 11, 21, 27, 31, 41)$ for each lysine residue.

The capability of features to predict ubiquitination sites (formulated as a binary classification problem) was assessed using three algorithms: Naïve Bayes, Random Forest and SVM. We chose these three algorithms considering that they have been previously used for developing ubiquitination site prediction servers. Naïve Bayes and Random Forest algorithms were implemented via the Weka (version 3–6-9) package [33] and 1000 trees were built using the Random Forest algorithm. For SVM, we used SVM-light (http://svmlight.joachims.org/) and selected the RBF as the kernel function to build the models. To improve SVM performance, two parameters (regularization parameter C and width parameter γ) were preliminarily optimized through a grid search strategy.

Five-fold cross-validation test and independent tests were conducted to assess the performance and importance of individual features. During assessment, we also examined the impact of the window size on the predictive power of features. Four window sizes (21, 25, 27 and 41) were adopted and assessed according to previous studies of ubiquitination site prediction.

Performance evaluation metrics

To comprehensively assess the predictive performance of the methods, we plotted receiver operating characteristic (ROC) curves [34, 35] by varying prediction thresholds. Two performance measures based on the ROC curve, total area under ROC curve (AUC) and relative area under ROC curve with limiting up to a 10% false positive rate (AUC01)

Table 3: Individual features used in previous ubiquitination site prediction methods

Feature type	Biological interpretation	Citation
AAC (amino acid content)	The amino acid composition of the sequence fragments surrounding ubiquitination sites.	[15, 16, 19]
AGG (aggregation opensity)	The aggregation propensity of the sequences surrounding ubiquitination sites.	[20]
AAindex	Based on the AAindex database [14], AAindex encoding reflects the physicochemical properties of the sequences surrounding ubiquitination sites.	[13, 18, 20]
BLOSUM62	The BLOSUM62 matrix is adopted to represent the protein primary sequence information, which reflects the similarity of two sequence fragments.	[16]
Charge-hyd (charge/ hydrophobicity ratio)	The charge and hydrophobicity ratio of the sequences surrounding ubiquitination sites.	[15, 26]
CKSAAP (composition of *k*-spaced amino acids pairs)	The CKSAAP encoding reflects the short range interactions of residues within the sequences surrounding ubiquitination sites.	[17, 20, 27–29]
Binary	The binary encoding reflects the position-specific information of the amino acids surrounding ubiquitination sites.	[15, 18]
Disorder	The predicted disorder information of the residues surrounding ubiquitination sites.	[15, 18]
KNN	The KNN encoding implies the clustering information of sequences surrounding ubiquitination sites.	[19]
PSSM (position-specific scoring matrix)	The PSSM reflects the evolutionary information of the amino acids surrounding the ubiquitination sites.	[15, 16, 18]

were calculated for robust performance evaluation. Generally, an AUC value closer to 1 and an AUC01 value close to 0.1 indicate better performance.

Statistical tests

We performed statistical tests to evaluate the significance of performance differences between all pairs of prediction methods and all pairs of individual features. This analysis determined the likelihood that a given method or feature performed significantly better than another one. The bootstrap test, originally proposed by Hanley and McNeil [36], was adopted to compare the paired AUCs or AUC01s. Taking the comparison of the paired AUCs as an example, the following formula was used:

$$D = \frac{AUC1 - AUC2}{Sd(AUC1' - AUC2')} \quad (1)$$

where *AUC*1 and *AUC*2 are the two original AUCs, while *AUC*1' and*AUC*2' are the bootstrap resampled AUCs and *Sd* denotes the standard deviation. We computed *Sd(AUC*1'*-AUC*2'*)* with 100 bootstrap replicates. In each replicate, the original measurements were resampled with replacement and the corresponding new ROC curves were plotted. Therefore, the resampled *AUC*1' and *AUC*2', and their difference (i.e., *AUC*1'-*AUC*2') were computed. Because *D* approximately follows a normal distribution, the *P*-value could be readily calculated. We performed the bootstrap tests using the pROC [37] package of R (http://www.r-project.org/) by comparing pairs of ROC curves in terms of AUC orAUC01. For all comparisons, $P \leq 0.05$ indicated significant difference in the predictive abilities between two compared predictors or features.

RESULTS AND DISCUSSION

Performance comparison of different prediction methods

No ubiquitination site predictor is universally best

Four species-specific datasets were used to assess the performance of five popular ubiquitination site prediction methods. To assess predictive capacity, we examined overall performance using AUC value and analyzed performance at high specificity using AUC01 value, which reflected the practical utility of a predictor in real-life applications. The larger the AUC01 value, the more ubiquitination sites were identified at a low false-positive rate. Figure 1 shows the ROC curves of the five compared methods based on the four species-specific datasets.

For the *S.dataset*, UbPred and CKSAAP_UbSite were the best-performing predictors with an AUC of 0.630 for UbPred and 0.616 for CKSAAP_UbSite (Figure 1A). In terms of statistical significance, the *P*-values for AUC and AUC01 between UbPred and CKSAAP_UbSite were 0.3276 and 0.8557, respectively, which were larger than 0.05, indicating the performance difference between the two methods was not significant. By contrast, AUC value was 0.609 for UbiProber_H.sapiens and 0.608 for UbiProber_Combined. Nevertheless, UbPred (AUC01 = 0.016) and CKSAAP_UbSite (AUC01 = 0.015) had higher AUC01 values than UbiProber_H.sapiens (AUC01 = 0.007) and UbiProber_Combined (AUC01 = 0.007). Indeed, the *P*-values showed that the AUC01 differences were significant (e.g., the *P*-value for UbPred versus UbiProber_H.sapiens is 1.063×10^{-8}; Table 4), indicating that UbPred and CKSAAP_UbSite are more practical for predicting potential ubiquitination sites for *S. cerevisiae*. These results were not surprising because both UbPred and CKSAAP_UbSite were developed specifically for *S. cerevisiae*, while UbiProber_H.sapiens and UbiProber_Combined were not. We also noted that current ubiquitination site predictors could achieve acceptable but not highly satisfactory performance when predicting yeast ubiquitination sites. A possible reason is that the number of known yeast ubiquitination sites is not sufficient enough to fully exploit the sequence pattern of yeast ubiquitination sites.

We found that hCKSAAP_UbSite had the best performance for predicting ubiquitination sites on both *H.dataset* (AUC = 0.662) and *M.datase*t (AUC = 0.677), followed by UbiProber_H.sapiens and UbiProber_Combined (Figure 1B and C). The latter two tools had AUC values of 0.637 and 0.637 for *H.dataset,* and 0.662 and 0.658 for *M.dataset*, respectively. The overall performance of both UbiProber_H.sapiens and UbiProber_Combined was close to that of hCKSAAP_UbSite. However, the comparison of AUC01 values in Supplementary Tables S2 and S3 suggested that hCKSAAP_UbSite had a significant higher AUC01 than UbiProber_H.sapiens and UbiProber_Combined.

All predictors did not achieve satisfactory performance for ubiquitination site prediction when

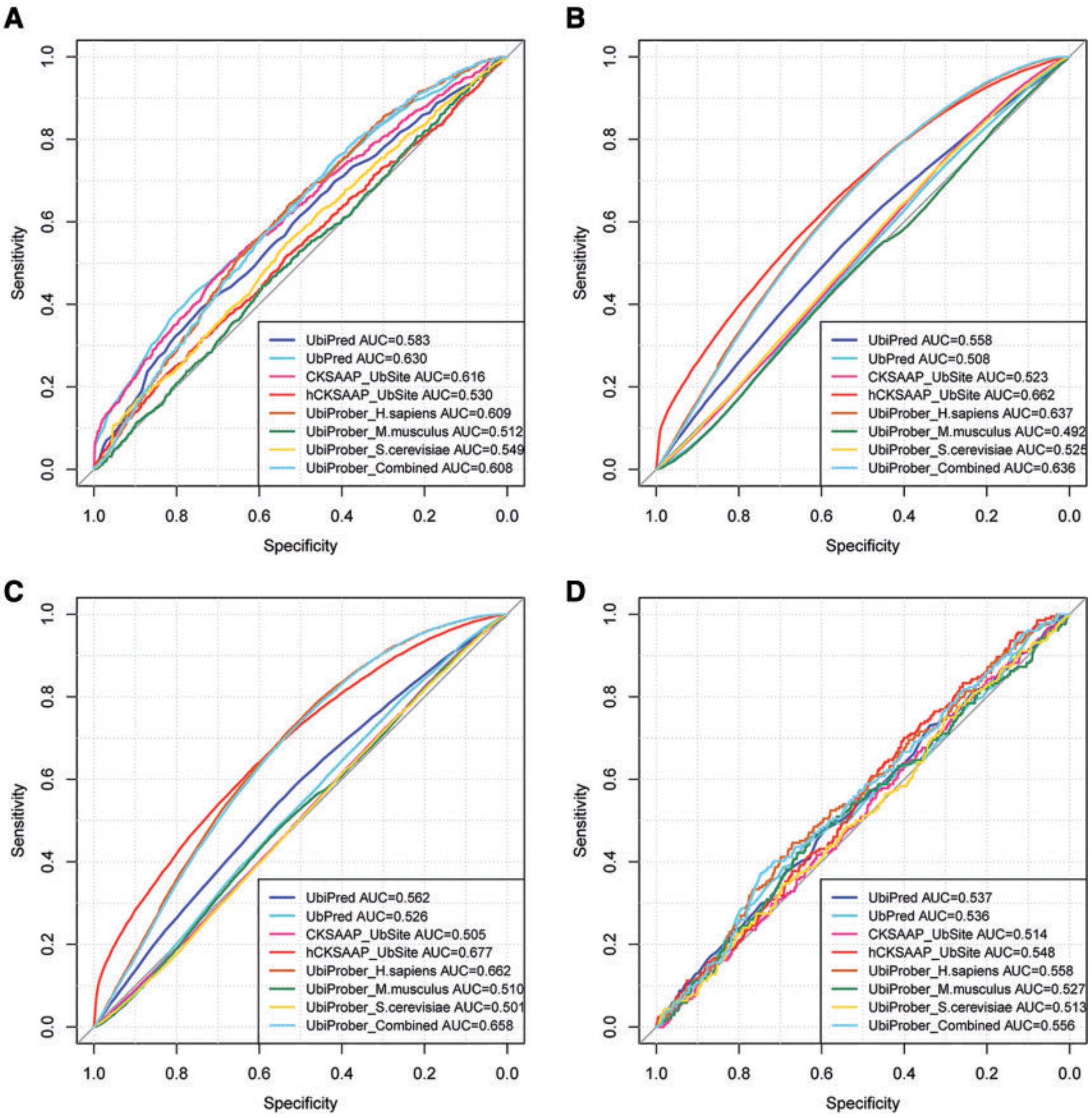

Figure 1: ROC curves of the five compared methods based on the four benchmark species-specific datasets. The performances of all the tools on the *S.dataset*, *H.dataset*, *M.dataset* and *A.dataset* are shown in panels (**A–D**), respectively. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

evaluated using the *A.dataset* (Figure 1D). On this dataset, the highest AUC value was 0.558 for UbiProber_H.sapiens. This prediction performance was close to random prediction, as 0.5 AUC and 0.005 AUC01 indicate random prediction. The reason might be ascribed to the fact that the training datasets of the five predictors contained no ubiquitination data from *A. thaliana*. We performed statistical tests to examine the significance of performance differences between all pairs of predictors (Supplementary Table S4). All *P*-values between any two predictors were larger than 0.05, proving the null hypothesis of no significant performance difference for all predictors in predicting *A. thaliana* ubiquitination sites.

We concluded that currently, no universal best predictor exists for predicting ubiquitination sites across all species. Both the results in Figure 1 and our statistical tests showed that UbPred and CKSAAP_UbSite are good choice for users to predict ubiquitination sites in *S. cerevisiae*. The hCKSAAP_UbSite tool had the best performance for *M. musculus* and *H. sapiens*. However, none of the five predictors were suitable for predicting ubiquitination sites for *A. thaliana*, indicating the need for a ubiquitination site predictor specific for *A. thaliana*.

Table 4: Performance difference of the five considered tools for *S.dataset* estimated by bootstrap test

	UbiPred	UbPred	CKSAAP.UbSite	hCKSAAP.UbSite	UbiProber.S.cerevisiae	UbiProber.M.musculus	UbiProber.H.sapiens	UbiProber.Combined
UbiPred	–	0.0012	0.0259	0.0010	0.0143	3.961×10^{-6}	0.04932	0.09289
UbPred	3.890×10^{-5}	–	0.3276	1.388×10^{-12}	4.425×10^{-10}	$<2.200 \times 10^{-16}$	0.1183	0.1213
CKSAAP.UbSite	3.695×10^{-7}	0.8557	–	5.061×10^{-9}	1.786×10^{-5}	8.090×10^{-16}	0.5874	0.5631
hCKSAAP.UbSite	0.2511	1.177×10^{-7}	6.490×10^{-7}	–	0.1921	0.1728	1.056×10^{-8}	4.882×10^{-7}
UbiProber.S.cerevisiae	0.5379	1.062×10^{-7}	1.829×10^{-7}	0.6148	–	0.0019	2.506×10^{-6}	3.409×10^{-6}
UbiProber.M.musculus	1.470×10^{-6}	$<2.200 \times 10^{-16}$	4.656×10^{-14}	0.0002	3.374×10^{-5}	–	1.695×10^{-14}	2.529×10^{-14}
UbiProber.H.sapiens	0.0789	1.063×10^{-8}	3.443×10^{-8}	0.4592	0.1528	0.0144	–	0.7451
UbiProber.Combined	0.0841	1.953×10^{-12}	3.258×10^{-11}	0.4685	0.2280	0.0139	0.8066	–

The upper right matrix is the paired *P*-values for AUC and the bottom left is the paired *P*-values for AUC0l.

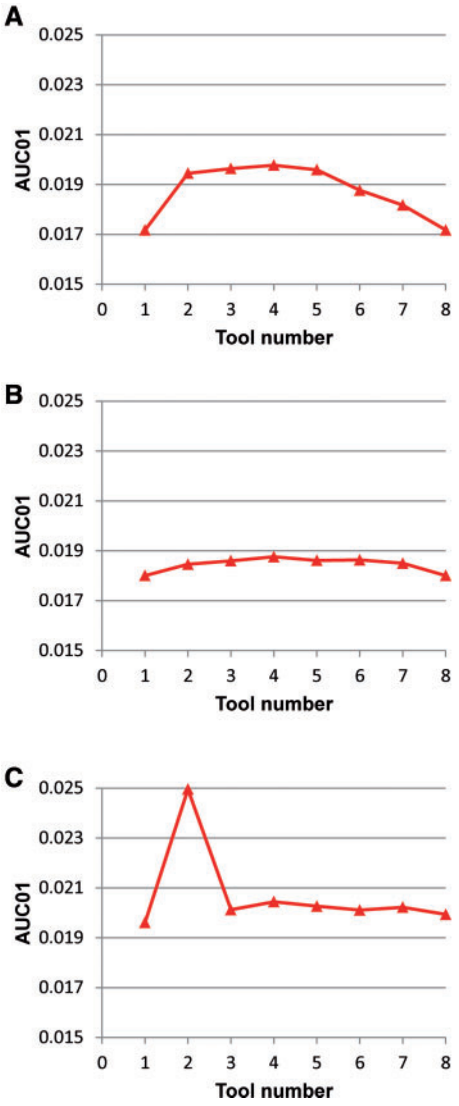

Figure 2: The performance of the combined predictors. The x-axis is the number of predictors, the y-axis represents the AUC0l values of combined predictors. The performances of the combined predictors on *S.dataset*, *H.dataset* and *M.dataset* are shown in panels (**A–C**), respectively.

Integration of individual predictors significantly improves prediction accuracy

We investigated whether the integration of the predictors that we compared improved performance. The outputs of different individual predictors were combined using a logistic regression approach. Each of the four datasets was divided into two subsets: one for training the logistic regression model and one to be the test dataset to evaluate performance of the combined predictor. We performed experiments by enumerating all possible combinations of the compared predictors and examining the performance of the resulting k-predictors (where k is the number of the combined individual predictors, $2 \leq k \leq 8$). The final prediction score P of the combined predictor was defined as:

$$\log\left(\frac{P}{1-P}\right) = \sum_{i=1}^{k} b_i S_i + \alpha \qquad (2)$$

where the coefficient b_i of the prediction score S_i, and the constant term α were deduced via a regression process, and k, which denotes the number of individual predictors in the combined predictor, varied from 2 to 8. According to the definition, the final prediction score P is the probability that the residue of interest is a ubiquitination site. The generalized linear model (i.e. the *glm* function) in R (http://www.R-project.org/) was used to generate the logistic regression model (see Supplementary Table S5 for the optimal regression formula).

We built combined predictors by integrating k single predictors. For each k, we mainly focused on the combined predictor that achieved the highest AUC01 value, as predictive ability at high specificity is considered most important in practical applications. By integrating different individual predictors, the combined predictors further improved performance compared with the best individual predictors (i.e., $k = 1$) (Figure 2). For *S.dataset* (Figure 2A), the highest AUC01 value was reached when $k = 4$ (i.e., when UbPred, CKSAAP_UbSite, UbiPred and UbiProber_S.cerevisiae were combined), and AUC01 value improved from 0.017 to 0.019. For *H.dataset* (Figure 2B), when $k = 4$ (combining hCKSAAP_UbSite, UbiProber_H.sapiens, UbiProber_M.musculus and CKSAAP_UbSite), we achieved the highest AUC01 (AUC01 = 0.0187). For *M.dataset* (Figure 2C), the highest AUC01 value achieved by a single predictor was 0.020. In contrast, when combining two predictors (UbiProber_S.cerevisiae and UbiProber_M.musculus), the AUC01 reached 0.025. We did not discuss the corresponding performance on the *A.dataset*, because no single predictor was found to be suitable for predicting ubiquitination sites in this species. Our results showed that the predictive ability of existing tools could be further improved through a simple logistic regression integration approach. In this way, the combined predictor could harness the advantages of different individual predictors. In practical applications, users could better determine if a particular lysine residue in a protein sequence is more or less likely to be a ubiquitination site by combining the results from several predictors. Nonetheless, we recommend that bioinformatics researchers develop novel meta or consensus predictors to improve prediction accuracy.

Species-specific sequence patterns challenge current ubiquitination site predictors

The lack of universal ubiquitination site predictor could be partly explained by differences in sequence patterns around ubiquitination sites in the four datasets. These patterns can be visualized using two-sample logo representation [38], which identifies and displays significant differences in position-specific amino acid compositions between two sets of multiple sequence alignments (i.e., ubiquitination sites versus non-ubiquitination sites). In the graphical output of two-sample logo, the upper section displays a set of amino acids enriched around ubiquitination sites, the lower section displays a set of amino acids depleted around ubiquitination sites, while the middle section displays consistent residues. Figure 3 shows two sample logos of four datasets. The sequence pattern of *A.dataset* is more widely scattered than the pattern for the other three datasets. The *A.dataset* sequence pattern is difficult to depict because it has only two types of hydrophobic residues [39] (I and L) enriched in more than three positions (−9, −5, +7, +10; −4, +2, +3 and +10). A common feature of the sequence patterns in the other three datasets is that positively charged residues [40] (H, K and R) are significantly depleted at varying positions from −6 to +6 surrounding the ubiquitination sites. However, this preference was not observed in *A. dataset*. Accordingly, the poor performance for *A. thaliana* of current ubiquitination site predictors, which were trained using data from the other three species, could have been expected. In addition, compared with *M.dataset* and *H.dataset*, a remarkable difference for *S.dataset* is that negatively charged [40] amino acids (D and E) appeared to be more frequently distributed around the ubiquitination sites (−3, −2, −1, +1, +3 and +6). Therefore, species-specific predictors are recommended as the first choice for predicting *S. cerevisiae* ubiquitination sites (Figure 1A). Sequence patterns for *M.dataset* and *H.dataset* closely resembled each other. For example, the hydrophobic residues (A, G, F, I, L and V) were enriched at positions from −4 to +4 around the ubiquitination sites, and the positively charged residue (R) was also significantly enriched in the flanking regions on both sides

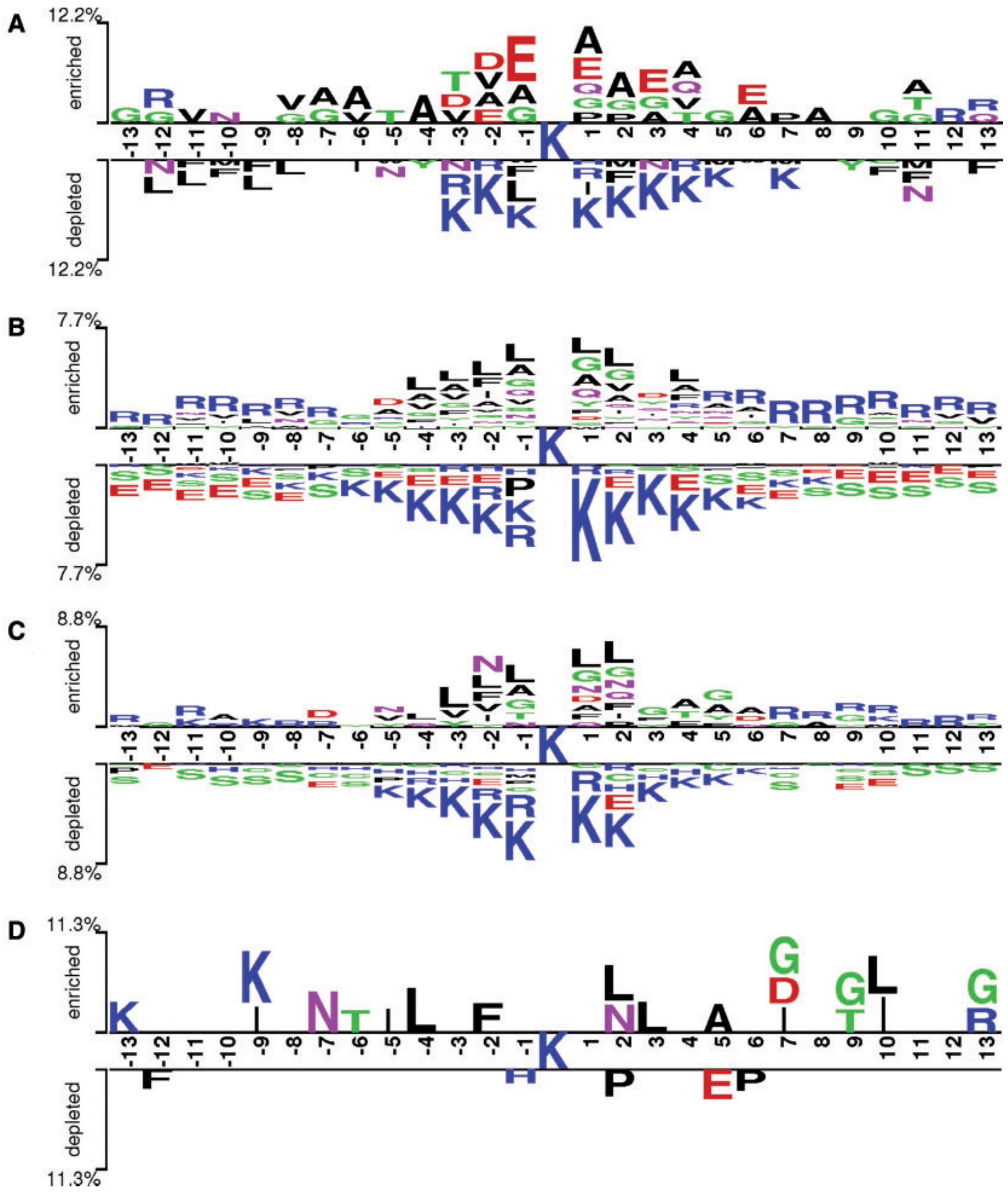

Figure 3: The Two-Sample-Logo representation of position-specific residue composition surrounding the ubiquitination sites and non-ubiquitination sites, based on *S.dataset* (**A**), *H.dataset* (**B**), *M.dataset* (**C**) and *A.dataset* (**D**). The two logo graphs were generated using the web server http://www.twosamplelogo.org/. Only residues significantly enriched or depleted (*t*-test, $P < 0.05$) flanking the centred ubiquitination sites (upstream 13 residues and downstream 13 residues) are shown.

from positions −13 to −7 and 7 to 13. This phenomenon might be because mice are evolutionarily closer to human than to plants or yeast. Thus, a *H. sapiens* ubiquitination site predictor could be used to predict ubiquitination sites for *M. musculus*.

We concluded that the ubiquitination sites from different species were surrounded by distinct sets of characteristic amino acids with different physiochemical properties. This phenomenon was especially clear when studying two distally related species. The physiochemical properties of specific amino acids around lysine residues might therefore be used as informative features for building models to predict ubiquitination sites, as suggested in previous studies [13, 15, 18, 20]. Also, given the contrasting difference in the sequence patterns of the four datasets, we suggest that users should initially consider using species-specific predictors to

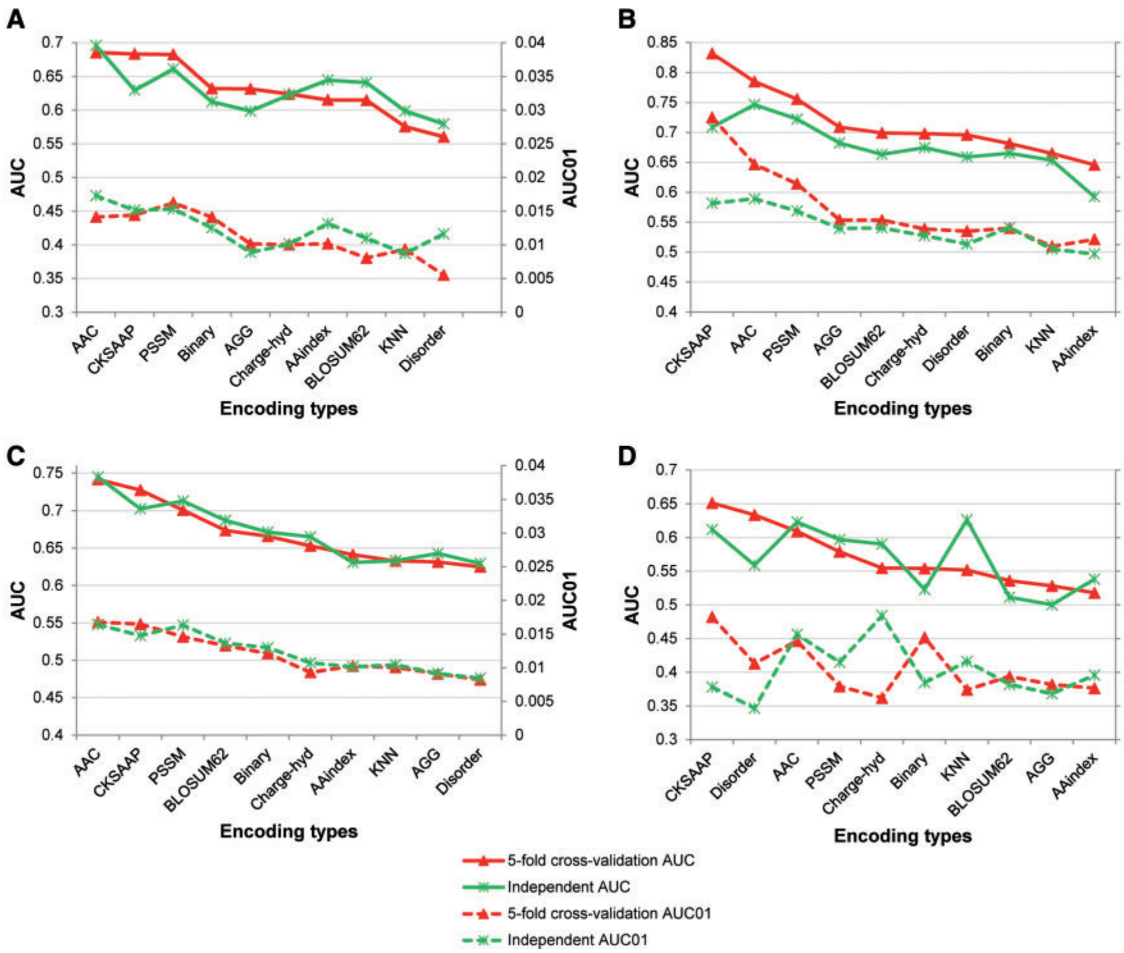

Figure 4: The performance of individual features on the four benchmark datasets. The features were sorted in the order of their AUC values in the 5-fold cross-validation test. The solid lines denote the AUC values, while the dashed lines are the AUC0l values. The performances of the individual features on *S.dataset, H.dataset*, *M.dataset* and *A.dataset* are given in panels (**A**), (**B**), (**C**) and (**D**), respectively. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

predict potential ubiquitination sites in different species.

Comparison of different tools from a user's perspective

We further compared different tools, assessing developed webservers or standalone software from a user's perspective. In particular, we compared these aspects: (i) whether the webserver supported batch prediction; (ii) whether the method had standalone software to implement its algorithm; and (iii) limitations of the webserver or software. Our comparison is summarized in Table 1. Among the tools we investigated, UbiPred was the only one that did not provide a standalone version to implement its algorithm. Its webserver supports batch prediction, accepting a maximum of 100 FASTA protein sequences at a time, and the submitted jobs are generally completed within a few seconds. The prediction output for a potential ubiquitination site includes four items: residue position, sequence fragment, annotation ('Y'or "N") and prediction score.

The UbPred webserver allows prediction of only a single sequence per user at a time, although users can download and install standalone versions of UbPred for both Linux and Windows operating systems to run batch prediction tasks. The UbPred output is in three columns: residue position, ubiquitination scores and predicted ubiquitination site annotation. Depending on the prediction score, three stringency thresholds of low, medium and high confidence are available. In general, for a typical protein sequence with around 500 amino acid residues, completing a single prediction task takes about 10 seconds.

Users of CKSAAP_UbSite and hCKSAAP_UbSite can submit protein sequences in RAW or FASTA format to webservers and select models for different species. The processing time for a protein sequence was 1–2 min, which was a bit longer than other tools. The prediction output contains three items: residue position, prediction score and ubiquitination site annotation. Prediction results are stored at the webservers for a month and users can query the results by searching through a job list or putting the job ID in a query box. Two thresholds, low and high confidence, are available. To make batch predictions, users can download standalone versions of CKSAAP_UbSite and hCKSAAP_UbSite (for Linux and available upon request). UbiProber does not use fixed cutoffs but allows users to adjust prediction stringency thresholds. Both the webserver and standalone software support batch prediction. The stand-alone version of UbiProber was implemented as a Windows application in the .NET4.0 framework using C# language. UbiProber had several specific requirements for input sequences: (i) the length of the FASTA header must be longer than nine characters; (ii) the protein sequence must be strictly formatted as 60 amino acids per line; and (iii) the maximal input size cannot be greater than 35 kilobytes.

Relative importance and predictive power of individual features

Overview

To assess the contribution and capabilities of individual features to the prediction of ubiquitination sites, we evaluated 10 types of features (encoding schemes) found to be useful in previous predictors. All features were directly computed or derived from protein sequence information. The predictive capabilities of the features were dependent on the sliding window size and the classification algorithm used to train the models. Therefore, we examined four window sizes (21, 25, 27 and 41) used in previous studies and three popular classification algorithms (Naïve Bayes, Random Forest and SVM). As a result, each feature had 12 values (4 window sizes and 3 classification algorithms) calculated for AUC and AUC01.

Figure 4 shows the best performance among the four window size types and three classification algorithms. These results were based on individual features and both 5-fold cross-validation and independent tests. In most cases, the results of 5-fold cross-validation tests were in accordance with the results from independent tests. No features performed well for *A.dataset*. This result might be attributed to two main reasons: (i) Although most of the considered features were used to predict ubiquitination sites for *S. cerevisiae*, *H. sapiens* and *M. musculus*, these features could not effectively describe the characteristics of sequence context surrounding *A. thaliana* ubiquitination sites; (ii) *A.dataset* was too small to efficiently represent the majority of ubiquitination sites in *A. thaliana*. Thus, machine learning-based predictors did not perform well on this limited dataset.

Important features for predicting ubiquitination sites

To assess the importance of the features, we ranked them based on AUC values (Supplementary Table S6). Only results achieved using the best window size and best classification algorithm were used for the ranking. In most cases, the AUC value ranking was consistent with the AUC01 value ranking. Moreover, we defined a significant contribution to prediction as a feature with an AUC value larger than 0.600 and an AUC01 value larger than 0.010. As a result, the top three most powerful features for *S.dataset* were AAC, PSSM and AAindex (Table 5). For *H.dataset* and *M.dataset*, the top three most powerful features were AAC, PSSM and CKSAAP. However, only AAC and KNN encoding schemes satisfied our criterion for the *A.dataset*. AAC encoding had the best performance in predicting ubiquitination sites across all four datasets, suggesting that AAC surrounding ubiquitination sites was discriminative and could be distinguished from non-ubiquitinated sites. For *S.dataset*, *H.dataset* and *M.dataset*, the PSSM encoding was another powerful feature for identifying ubiquitination sites. The results were understandable because mammalian ubiquitination sites are slightly more conserved than unmodified lysine residues [41]. CKSAAP encoding that describes short-range interactions of residues within a sequence or a sequence fragment [17] was also a powerful feature, achieving good performance on *M.dataset* and *H.dataset*.

Effects of window size and classification methods on prediction performance

The window size and the classification methods adopted differed from each other in previous studies. From Table 5, we concluded that the optimal window size and classification methods depended on the dataset and feature type. To further explore this possibility, we calculated the optimal window

Table 5: The ranking of the top three most powerful features according to their AUC values; the corresponding optimal window size and the classification algorithm are also listed[a]

	S.dataset					*H.dataset*					*M.dataset*					*A.dataset*				
Rank	**Encoding**	**AUC**	**AUC01**	**Window size**	**Algorithm**	**Encoding**	**AUC**	**AUC01**	**Window size**	**Algorithm**	**Encoding**	**AUC**	**AUC01**	**Window size**	**Algorithm**	**Encoding**	**AUC**	**AUC01**	**Window size**	**Algorithm**
1	AAC	0.6956	0.0173	41	SVM	AAC	0.7463	0.0168	41	RF	AAC	0.7445	0.0164	41	RF	AAC	0.6228	0.0156	27	RF
2	PSSM	0.6609	0.0153	41	RF	PSSM	0.7223	0.015	41	RF	PSSM	0.7127	0.0163	41	RF	KNN	0.6264	0.0116	27	RF
3	AAindex	0.6445	0.0131	41	RF	CKSAAP	0.7089	0.0161	41	SVM	CKSAAP	0.7025	0.0148	27	SVM	CKSAAP	0.6119	0.0078	25	NB

[a] SVM, RF and NB stands for Support Vector Machine, Random Forest and Naïve Bayes, respectively.

size and classification method for the four datasets using our criterion of AUC value larger than 0.600 and AUC01 value larger than 0.010 (Although these thresholds were relatively arbitrary, we found that the conclusion remained largely unchanged when different thresholds were applied). We found that 26 out of 40 test results (4 datasets times 10 features) met this criterion. For the optimal window size (Supplementary Figure S1A), the window size 41 had a much higher percentage (accounting for 53.85%) among all the four kinds of window sizes, followed by the window sizes 27, 25, and 21 which accounted for 30.77,11.54 and 3.85%, respectively. The results indicated that window size 41 was optimal for most of the considered features and use of distant sequence features could improve predictive performance [16]. For classification methods, Random Forest was the algorithm with the best-performance for the most informative features across the four datasets, followed by the SVM algorithm (Supplementary Figure S1B). This result suggested that algorithms such as Random Forest and SVM were especially suitable for higher-dimensional and complicated features and were more powerful in predicting ubiquitination sites than simple Naïve Bayes algorithm.

Prediction performance of models using combined features

We integrated the results of the features that showed good performance to determine if we could achieve optimal combination of the features through a logistical regression approach. Figure 5 shows prediction performance using feature combination. Combining the features led to 1–4% AUC improvement, depending on the dataset. The optimal combination of features is given in Table 6, and the corresponding machine learning algorithm, window size and regression weighting of each individual feature for the four datasets are listed Supplementary Table S7. However, a bootstrap test showed that not all improvements were statistically significant. For *S.dataset*, *H.dataset* and *M.dataset*, all *P*-values were <0.05 with the exception of AUC improvement for *S.dataset* (bootstrap *P*-value $= 0.073$), indicating a significant performance improvement for the combined features on the three datasets. For *A.dataset*, however, *P*-value of AUC difference between the best single feature and the combined features is 0.619. This result meant that combining features did not lead to a significant improvement, possibly

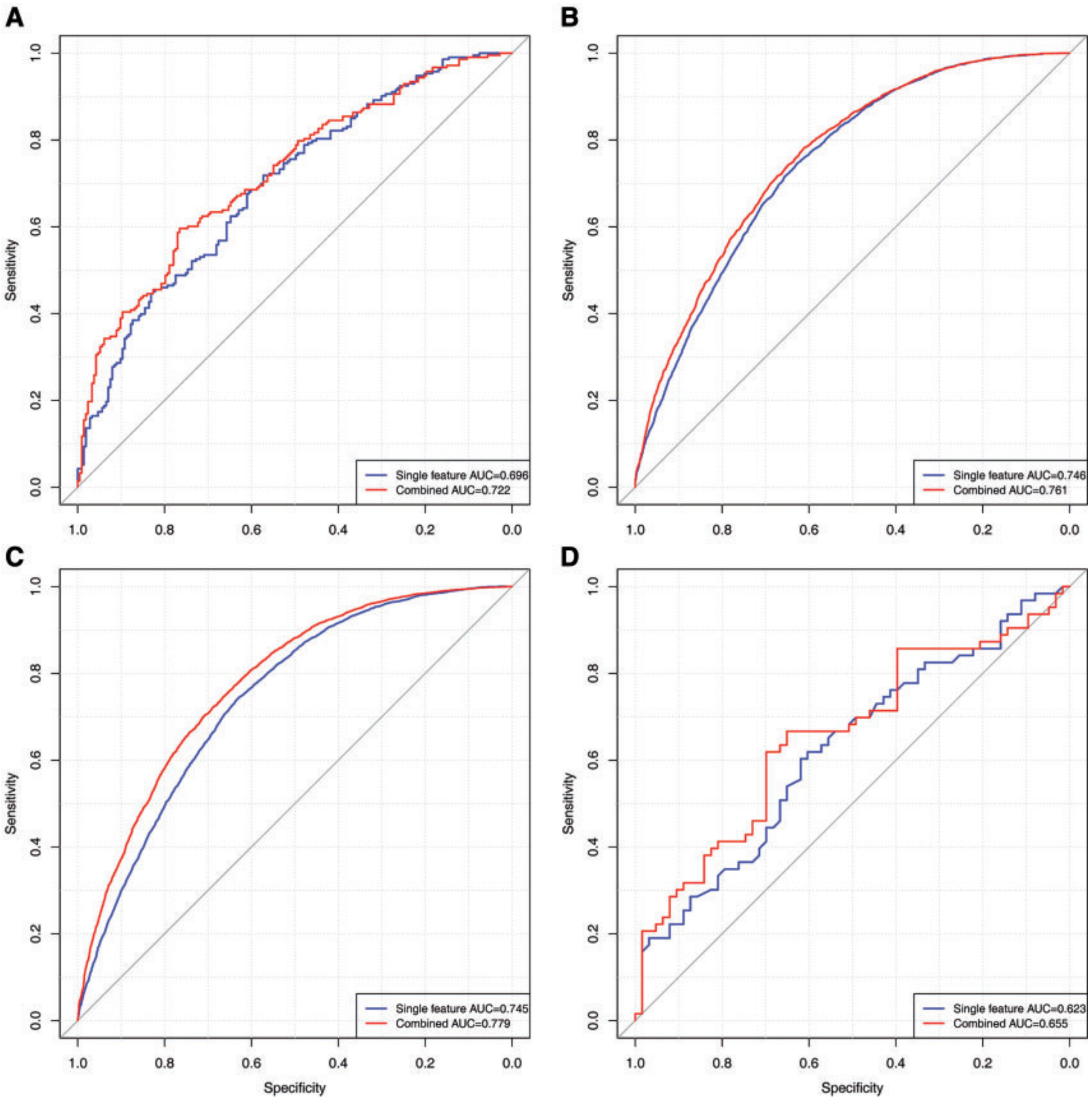

Figure 5: ROC curves of the best-performing single features and the combined optimal features. The performances of the best single feature and the combined optimal features on *S.dataset* (**A**), *H.dataset* (**B**), *M.dataset* (**C**) and *A.dataset* (**D**) are shown. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

due to the poor predictive ability of the individual features and the small *A.dataset* sample size.

Future perspectives

Successful prediction of ubiquitination sites requires careful predictor choice

The first ubiquitination site predictor UbiPred was developed five years ago [13]. By deriving useful features from a multispecies dataset, UbiPred had been expected to be suitable for predicting ubiquitination sites for a wide spectrum of eukaryotes. However, our results indicated that UbiPred achieved acceptable performance only for predicting yeast ubiquitination sites, which was the source of most of its training data (Figure 1A). The same result is seen for the second multispecies ubiquitination site predictor, UbiProber_combined [19]. Its training set was mainly a large number of mammalian ubiquitination sites and this predictor had much better predictive power when tested on mammalian datasets (Figure 1B and C). Because these two general ubiquitination site predictors cannot be widely applied, species-specific predictors, if available, are preferred. However, as shown by independent tests, few species-specific predictors had relatively satisfactory performance. UbPred and CKSAAP_UbSite outperformed other predictors on the yeast dataset, while hCKSAAP_UbSite had the best performance when used to predict mammalian ubiquitination sites. A reliable ubiquitination site

Table 6: The optimal combinations of the features for the four datasets

Dataset	Single feature			Combined features		
	Type	AUC	AUC01	Type	AUC	AUC01
S.dataset	AAC	0.696	0.018	AAC+PSSM+Charge-hyd+BLOSUM62+AGG	0.722	0.026
H.dataset	AAC	0.746	0.017	AAC+CKSAAP+PSSM+Charge-hyd+AGG	0.761	0.021
M.dataset	AAC	0.745	0.017	AAC+KNN+CKSAAP+PSSM+Charge-hyd+ Binary+BLOSUM62+AGG	0.779	0.023
A.dataset	AAC	0.623	0.017	CKSAAP+PSSM+Disorder+Binary+AGG	0.655	0.020

predictor dedicated to *Arabidopsis*is yet to be developed.

Developing a novel species-specific ubiquitination site predictor is not easy and straightforward. Among many other factors, better encoding of flanking sequences remains an important issue to be addressed. Through the ongoing efforts of bioinformaticians, a number of ubiquitination site-related sequence features have been proposed. We tested the performance of several representative features used in previous studies. The most striking result was that simple amino acid composition together with optimal window size and the appropriate machine learning algorithm achieved prediction performance comparable to other sophisticated encoding schemes (Figure 4). Another interesting result concerning feature combination was that a model's performance did not always increase when more features were considered. The top four or five features usually achieved maximum performance in combination. These results highlighted the importance of the design of feature-encoding schemes and the identification of novel sequence features. Researchers must also optimize sliding window size to extract sequence context surrounding ubiquitination sites and determine machine learning algorithm suitable for training the models.

Ubiquitination site prediction: thirst for new data

The last decade has witnessed a rapid accumulation of high-throughput ubiquitination data, especially for humans. As a consequence, characteristic sequence patterns surrounding human ubiquitination sites can be clearly determined. In comparison, a few years ago, the sequence logo of human ubiquitination sites was sparse, and the overrepresentation and underrepresentation of specific residues were not clear. The current sequence logo, as exemplified in this review (Figure 3), contains enriched information about site-specific amino acid preference. As expected, the improvement of sequence logo representation is in accordance with the increased sensitivity of recently developed predictors (Figure 1).

The question of how available experimental data limit the predictor performance is answered in Figure 5, which shows that the best single feature achieved a cross-validation AUC higher than 0.75 on the human dataset (containing ~7000 sites). However, the corresponding single best feature reached only a cross-validation AUC of 0.65 for the *Arabidopsis* dataset (containing ~200 sites). Thus, the greatest chance for further improvement of non-mammalian ubiquitination site predictors appeared to depend on the availability of new high-throughput ubiquitination data. Non-mammalian ubiquitination site predictors will benefit from new high-throughput data by correcting potential bias and enhancing robustness. However, even current mammalian ubiquitination site data are far from perfect. In many cases, little knowledge about ubiquitin linkage (e.g., monoubiquitination, K63-/K48-linked polyubiquitination) and catalyzing enzymes can be obtained from high-throughput technology. The ubiquitination system is complicated, so current ubiquitination site datasets are likely to be mixtures of different groups of ubiquitination sites. We expect that novel, sizable data about catalyzing enzymes and ubiquitin linkage will propel the development of powerful ubiquitination site predictors. Therefore, we recommend that developers update and calibrate ubiquitination site predictors based on new data to ensure the competitiveness and performance of their predictors.

Apart from newly collected ubiquitination data, related information such as the presence of other PTM sites and structural propensities could also be helpful for further improving predictors. For example, Bork and collaborators showed that functional crosstalk between PTMs is ubiquitous in eukaryotic proteomes and several pairs of PTM sites including ubiquitination and phosphorylation sites exhibit a strong tendency toward co-occurrence and co-evolution [42]. Phosphorylation is the most well-studied PTM type and several accurate phosphorylation site predictors are available. Therefore, interrogating and encoding co-occurrence or co-evolution with a phosphorylation site as novel and potentially useful features is likely to facilitate the prediction of ubiquitination sites. Our laboratory has recently analyzed informative structural features associated with ubiquitination sites [43] and found that several novel structural propensities of ubiquitination sites such as protrusion index and centrality tend to be complementary to sequence pattern. Our analysis also indicated that the integration of structural propensities and sequence pattern could further improve the prediction performance of ubiquitination sites.

CONCLUSION

We used four species-specific datasets (*S.dataset*, *H.dataset*, *M.dataset* and *A.dataset*) from five recent large-scale proteomic studies to assess currently available ubiquitination site prediction methods that provide webservers or standalone software to implement their algorithms. Using these benchmark datasets, we comprehensively compared method performances. We discussed the advantages and disadvantages of the webservers and stand-alone software from different aspects to guide users to choose tools that best suit their purposes. Finally, we tested the major features used in existing ubiquitination site predictors and ranked the features according to their contribution to the predictive performance of species-specific ubiquitination sites. We also evaluated the predictive abilities of combinations of features to identify the optimal combination that led to the overall best performance.

The major observations from our analysis are first, that no universal best predictors are currently available for predicting ubiquitination sites for all four species that we investigated. In particular, none of the existing predictors was suitable for predicting ubiquitination sites in *A. thaliana*. Second, although the performance of existing predictors on *S. cerevisiae*, *M. musculus* and *H. sapiens* datasets was acceptable, there is room to further improve prediction performance by combining different predictors through simple approaches such as logistic regression. Third, AAC encoding generally performed the best in predicting ubiquitination sites across all four datasets and combination of features could result in improved performance in most cases.

Finally, we emphasize that the main purpose of this analysis was not to simply rank different predictors through a rigorous performance comparison. Rather, we would also like to mention common issues that this new and promising field must deal with in the near future from the perspectives of both users and developers. For instance, the performance of the predictors tested in this analysis was less impressive than the performance reported in the original papers, implying performance overestimation may exist in previous studies. To avoid such performance overestimation, more extensive data collection and more careful feature representation are required in further development of ubiquitination site predictors. Fortunately, as more ubiquitination sites are experimentally verified, standard training and testing datasets are likely to be available to the community in the near future. These data will help developers construct and benchmark their methods more reliably and assist users in obtaining more reasonable explanations of prediction results. We therefore anticipate that better ubiquitination site prediction methods and tools with improved performance will continue to emerge as increasing amounts of ubiquitination data and rapidly evolving computational techniques become available.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- A number of computational tools have recently been developed for predicting protein ubiquitination sites. However, these tools have different training datasets, prediction algorithms, functionality and features, complicating their utility and application.
- We systematically assessed the performance of current ubiquitination site prediction tools using novel independent testing datasets from *S. cerevisiae*, *H. sapiens*, *M. musculus* and *A. thaliana*.The results implied that these tools could achieve somewhat acceptable but not highly satisfactory accuracies, and notably, there is no universal best predictor suitable for predicting ubiquitination sites in all species.
- For the users from experimental biologist communities, species-specific prediction tools, if available, are of best choice to predict ubiquitination sites in the closely relative species. To conduct proteome-wide prediction, a tool implementing batch prediction should be preferred.
- For the computational tool developers, as we have extensively evaluated the predictive abilities of individual features and their combinations, an optimal combination of features will be a good strategy to construct more powerful tools in the future.

Acknowledgments
We thank Professor Yu Xue (Huazhong University of Science and Technology, China) for helpful discussions on this work.

FUNDING

This work was financially supported by grants from the State Education Ministry of China (20120008110014), the National Natural Science Foundation of China (61202167, 61303169, 11250110508, 31350110507), National Health and Medical Research Council of Australia (NHMRC) (490989) and the Hundred Talents Program of the Chinese Academy of Sciences (CAS). J.S. is an NHMRC Peter Doherty Fellow and recipient of the Hundred Talents Program of CAS.

References

1. Hagai T, Levy Y. Ubiquitin not only serves as a tag but also assists degradation by inducing protein unfolding. *Proc Natl Acad Sci USA* 2010;**107**:2001–6.
2. Dikic I, Wakatsuki S, Walters KJ. Ubiquitin-binding domains - from structures to functions. *Nat Rev Mol Cell Biol* 2009;**10**:659–71.
3. Hershko A, Ciechanover A. The ubiquitin system. *Annu Rev Biochem* 1998;**67**:425–79.
4. Pickart CM, Eddins MJ. Ubiquitin: structures, functions, mechanisms. *Biochim Biophys Acta* 2004; **1695**:55–72.
5. Hicke L. Protein regulation by monoubiquitin. *Nat Rev Mol Cell Biol* 2001;**2**:195–201.
6. Pickart CM. Ubiquitin enters the new millennium. *Mol Cell* 2001;**8**:499–504.
7. Haglund K, Dikic I. Ubiquitylation and cell signaling. *EMBO J* 2005;**24**:3353–9.
8. Schwartz AL, Ciechanover A. The ubiquitin-proteasome pathway and pathogenesis of human diseases. *Annu Rev Med* 1999;**50**:57–74.
9. Herrmann J, Lerman LO, Lerman A. Ubiquitin and ubiquitin-like proteins in protein regulation. *Circ Res* 2007;**100**: 1276–91.
10. Gentry MS, Worby CA, Dixon JE. Insights into Lafora disease: malin is an E3 ubiquitin ligase that ubiquitinates and promotes the degradation of laforin. *Proc Natl Acad Sci USA* 2005;**102**:8501–6.
11. Peng J, Schwartz D, Elias JE, *et al*. A proteomics approach to understanding protein ubiquitination. *Nat Biotechnol* 2003; **21**:921–6.
12. Tomlinson E, Palaniyappan N, Tooth D, *et al*. Methods for the purification of ubiquitinated proteins. *Proteomics* 2007;**7**: 1016–22.
13. Tung CW, Ho SY. Computational identification of ubiquitylation sites from protein sequences. *BMC Bioinformatics* 2008;**9**:310.
14. Kawashima S, Pokarowski P, Pokarowska M, *et al*. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 2008;**36**:D202–5.
15. Radivojac P, Vacic V, Haynes C, *et al*. Identification, analysis, and prediction of protein ubiquitination sites. *Proteins* 2010;**78**:365–80.
16. Lee TY, Chen SA, Hung HY, *et al*. Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. *PLoS One* 2011;**6**: e17331.
17. Chen Z, Chen YZ, Wang XF, *et al*. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PLoS One* 2011;**6**:e22930.
18. Cai Y, Huang T, Hu L, *et al*. Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino Acids* 2012;**42**:1387–95.
19. Chen X, Qiu JD, Shi SP, *et al*. Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics* 2013; **29**:1614–22.
20. Chen Z, Zhou Y, Song J, *et al*. hCKSAAP_UbSite: Improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. *Biochim Biophys Acta* 2013;**1834**:1461–7.
21. Starita LM, Lo RS, Eng JK, *et al*. Sites of ubiquitin attachment in *Saccharomyces cerevisiae*. *Proteomics* 2012;**12**: 236–40.
22. Kim DY, Scalf M, Smith LM, *et al*. Advanced proteomic analyses yield a deep catalog of ubiquitylation targets in Arabidopsis. *Plant Cell* 2013;**25**:1523–40.
23. Wagner SA, Beli P, Weinert BT, *et al*. Proteomic analyses reveal divergent ubiquitylation site patterns in murine tissues. *Mol Cell Proteomics* 2012;**11**:1578–85.
24. Mertins P, Qiao JW, Patel J, *et al*. Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nat Methods* 2013;**10**:634–637.
25. Udeshi ND, Svinkina T, Mertins P, *et al*. Refined preparation and use of anti-diglycine remnant (K-epsilon-GG) antibody enables routine quantification of 10,000s of ubiquitination sites in single proteomics experiments. *Mol Cell Proteomics* 2013;**12**:825–31.
26. Uversky VN, Gillespie JR, Fink AL. Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins* 2000;**41**:415–27.
27. Chen YZ, Tang YR, Sheng ZY, *et al*. Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. *BMC Bioinformatics* 2008;**9**:101.
28. Chen K, Kurgan L, Rahbari M. Prediction of protein crystallization using collocation of amino acid pairs. *Biochem Biophys Res Commun* 2007;**355**:764–9.
29. Chen K, Kurgan LA, Ruan J. Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. *BMC Struct Biol* 2007;**7**:25.
30. Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics* 2010;**26**:326–32.
31. Peng K, Radivojac P, Vucetic S, *et al*. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 2006;**7**:208.

32. Obradovic Z, Peng K, Vucetic S, *et al*. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* 2005;**61(Suppl 7)**: 176–82.

33. Hall M, Frank E, Holmes G, *et al*. The WEKA data mining software. *ACM SIGKDD Explor Newslett* 2009;**11**: 10.

34. Centor RM. Signal detectability: the use of ROC curves and their analyses. *Med Decis Making* 1991;**11**:102–6.

35. Gribskov M, Robinson NL. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput Chem* 1996;**20**:25–33.

36. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983;**148**:839–43.

37. Robin X, Turck N, Hainard A, *et al*. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;**12**:77.

38. Vacic V, Iakoucheva LM, Radivojac P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* 2006;**22**: 1536–7.

39. Eisenberg D. Three-dimensional structure of membrane and surface proteins. *Annu Rev Biochem* 1984;**53**:595–623.

40. Fauchere JL, Charton M, Kier LB, *et al*. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int J Pept Protein Res* 1988;**32**:269–78.

41. Hagai T, Toth-Petroczy A, Azia A, *et al*. The origins and evolution of ubiquitination sites. *Mol Biosyst* 2012;**8**: 1865–77.

42. Minguez P, Parca L, Diella F, *et al*. Deciphering a global network of functionally associated post-translational modifications. *Mol Syst Biol* 2012;**8**:599.

43. Zhou Y, Liu S, Song J, *et al*. Structural propensities of human ubiquitination sites: accessibility, centrality and local conformation. *PLoS One* 2013;**8**:e83167.