Amino Acids (2014) 46:853–862
DOI 10.1007/s00726-013-1645-3

ORIGINAL ARTICLE

RUBI: rapid proteomic-scale prediction of lysine ubiquitination and factors influencing predictor performance

Ian Walsh · Tomás Di Domenico · Silvio C. E. Tosatto

Received: 7 September 2013 / Accepted: 11 December 2013 / Published online: 23 December 2013

Abstract Post-translational modification of protein lysines was recently shown to be a common feature of eukaryotic organisms. The ubiquitin modification is regarded as a versatile regulatory mechanism with many important cellular roles. Large-scale datasets are becoming available for *H. sapiens* ubiquitination. However, using current experimental techniques the vast majority of their sites remain unidentified and *in silico* tools may offer an alternative. Here, we introduce Rapid UBIquitination (RUBI) a sequence-based ubiquitination predictor designed for rapid application on a genome scale. RUBI was constructed using an iterative approach. At each iteration, important factors which influenced performance and its usability were investigated. The final RUBI model has an AUC of 0.868 on a large cross-validation set and is shown to outperform other available methods on independent sets. Predicted intrinsic disorder is shown to be weakly anti-correlated to ubiquitination for the *H. sapiens* dataset and improves performance slightly. RUBI predicts the number of ubiquitination sites correctly within three sites for ca. 80 % of the tested proteins. The average potentially ubiquitinated proteome fraction is predicted to be at least 25 % across a variety of model organisms, including several thousand possible *H. sapiens* proteins awaiting experimental characterization. RUBI can accurately predict ubiquitination on unseen examples and has a signal across different eukaryotic organisms. The factors which influenced the construction of RUBI could also be tested in other post-translational modification predictors. One of the more interesting factors is the influence of intrinsic protein disorder on ubiquitinated lysines where residues with low disorder probability are preferred.

Keywords Ubiquitination · Ubiquitylation · Lysine · Machine learning · Intrinsic disorder

Introduction

Post-translational modifications (PTMs) contribute to the complexity of an organism, bestowing multiple protein functions on a single encoding gene (Hunter 2007). Lysine ubiquitination is a reversible PTM found in all eukaryotic cells. After translation, a protein can be modified by covalent bonding of ubiquitin, a small and highly conserved regulatory protein. The enzymatic process for ubiquitination involves a three-step sequential process between the E1, E2 and E3 enzymes (Glickman and Ciechanover 2002). The bonding can be a single ubiquitin molecule (mono-ubiquitination) or multiple chains (poly-ubiquitination) resulting in a wide variety of cellular processes. One of the earliest functional associations was proteasomal degradation (Chau et al. 1989). However, currently the ubiquitin system is regarded as a more versatile regulatory mechanism (Sun and Chen 2004). For example, lysine 63-linked poly-ubiquitin chain is involved in both DNA repair and endocytosis (Chen and Sun 2009).

The RUBI web server and stand-alone version for Linux are available from the URL: http://protein.bio.unipd.it/rubi/.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-013-1645-3) contains supplementary material, which is available to authorized users.

I. Walsh · T. Di Domenico · S. C. E. Tosatto (✉)
Department of Biology, University of Padua, Viale G. Colombo
3, 35131 Padua, Italy
e-mail: silvio.tosatto@unipd.it

I. Walsh
e-mail: ian.walsh@bio.unipd.it

Mono-ubiquitination can also modify a protein to perform various functions ranging from membrane transport to transcriptional regulation (Hicke 2001). In contrast, deregulation of ubiquitin is implicated in cancer (Hoeller et al. 2006) and neurogenerative disorders (Bingol and Sheng 2011). Given these important functions, targeting the multi-functional role of ubiquitination and its pathway can be of massive therapeutic benefit (Nalepa et al. 2006; Wong et al. 2003).

In vitro tools are difficult to develop for ubiquitination because the modification is large (ca. 8 kDa) and the modified protein has a rapid turnover (Peng et al. 2003). Recently, more robust and accurate techniques, such as global mass spectrometry, are now able to process thousands of sites (Udeshi et al. 2013). However, these experimental techniques often vary, capturing different protein properties, and experimental detection is hampered by the diversity of the type of ubiquitin chain (Ikeda and Dikic 2008). Computational tools are thus still needed to fill the gap. To guide in vitro experiments, these tools should be fast, with high specificity (minimal false-positive rate) and significant sensitivity (true-positive rate). Efficient and highly specific predictors are also needed for hypothesis testing. For example, ubiquitination cross talks with other PTMs resulting in a sophisticated coding scheme for different protein functions (Lonard and O'Malley 2007); hence combinations of ubiquitination and other PTM in silico tools may shed light on this phenomenon.

Computational prediction of phosphorylation is well studied and may be used to describe the problem of ubiquitination data shortage by analogy. Phosphorylation prediction methods have been initially constructed in a general (Blom et al. 1999) and later enzyme-specific manner (Blom et al. 2004). However, data deficiency becomes an issue when applying an enzyme-specific approach. Moreover, if the data are split further based on individual organisms (assuming sites are different across species), the data become even more deficient, making highly specific prediction tools difficult to implement. This usually results in two computational prediction modes, either enzyme-specific but organism-independent or enzyme-independent but organism-specific tools. In this work we focus on an enzyme-independent, but human-specific tool.

Computational prediction tools are only recently emerging. UbiPred (Tung and Ho 2008) uses a support vector machine (SVM) approach to predict ubiquitination sites on 105 mainly *S. cerevisiae* proteins. Another method, UbPred (Radivojac et al. 2010), based on the random forest algorithm was trained on a dataset of motifs from *S. cerevisiae*. The algorithm is fast allowing for large-scale analysis and is used to show the enrichment in various molecular functions and a preference for proteins with very short half-lives. Two more recent methods (Chen et al. 2011; Cai et al. 2012) highlight the use of a sophisticated encoding scheme and feature selection, respectively. Recently, data and predictors have become available for mammalian sites. An updated version of Chen et al. (2011) was retrained for *H. sapiens* sites (Chen et al. 2013b). UbiProber (Chen et al. 2013a) incorporates both general and species-specific ubiquitination sites using key motif positions and amino acid residue features. It was constructed on three species, *H. sapiens*, *M. musculus* and *S. cerevisiae,* showing interesting performance in particular for cross species predictions. All of these methods were motif based, meaning patterns of *N* amino acids surrounding each lysine were used for learning.

A recent independent assessment of PTM predictors (Schwartz 2012) has shown that 9 out of 11 predictors behave worse than random on unseen data. In this paper, we try to determine factors which may influence the performance, not as a criticism of other methods but simply as appropriate factors for the model being constructed. We have developed a novel method, RUBI, trained on a large dataset of over 10,000 experimentally determined *H. sapiens* ubiquitination sites (Wagner et al. 2011). Rapid UBIquitination detection (RUBI) is a fast predictor aimed at proteomic-scale applications with high specificity and significant sensitivity. We show that it accurately mimics the results of high-throughput mass spectrometry techniques on unseen data.

Methods

Datasets

Lysine ubiquitination is a binary classification problem. However, while the positive set can be defined easily from high-throughput datasets, negative lysines become a thorny issue (Schwartz et al. 2009). Assignment of negative cases can only be tentative, as new experimental evidence may reveal them to be ubiquitinated. For this reason they are often called background lysines. Performance on PTM classification depends on the amount of redundancy in the underlying data and the availability of high-quality annotated data (Blom et al. 2004). Here, 11,054 positive *H. sapiens* motifs were taken directly from high-resolution mass spectrometry data on 4,273 proteins (Wagner et al. 2011), where for each site position in sequence and local context of length, 13 was given. The experimental technique uses immuno-enrichment by anti-di-glycine antibody and we try to reproduce this technique. The learning set was constructed from this data, while the independent set was constructed from another database; their construction is described in the following and summarized in Table 1. The most clear observation is the massive imbalance of

Table 1 Distribution of residues and lysines in the datasets

Dataset	Proteins	Residues	Lysines	Positives	Background
Motif40[a]	3,705	2,014,272	154,944	9,237	145,686
Seq40[a]	3,705	2,460,023	154,944	9,237	145,686
PUB[b]	1,264	900,370	52,766	2,563	50,204
CST[b]	2,103	1,687,456	108,022	3,768	104,254
LTP[b]	91	70,103	4,500	201	4,299

Number of proteins, residues and lysines in the datasets. Positives and background refer to the lysine classification. The main independent dataset is published high-throughput mass spectrometry data from the literature (PUB)

[a] Learning sets

[b] Independent sets

positives to background lysines when using full sequences (e.g., ratio 1:16 for Seq40).

Seq40

Seq40 considers full human protein sequences taken from Wagner et al. (2011). Each sequence was reduced to a maximum pairwise identity of 40 % with CD-HIT (Li and Godzik 2006). For this data learning proceeded on the entire sequence (see "Machine learning"). When testing performance, positive lysines were experimentally validated ones, while the background was defined as those lysines not found ubiquitinated in the same study. A similar positive and background extraction was done (Cai et al. 2012; Chen et al. 2013a).

Motif40

Motif40 considers motifs taken from Wagner et al. (2011). Motifs surrounding lysines of length 13 were extracted directly from Seq40. The diversity at the motif level will be analyzed in the results as it is an important factor. For these data, learning proceeds on the site motifs (see "Machine learning"). The ratio of positives to negatives is identical to Seq40.

Independent benchmark

This was constructed from Phosphosite plus, an interactive database of manually curated PTMs (Hornbeck et al. 2012). Upon model construction, it contained 39,037 ubiquitination sites mainly for *M. musculus* and *H. sapiens*. CD-HIT was used to remove pairs of sequences with more than 40 % pairwise sequence identity to each other and to the training data. After this, only 22 out of 2,563 positive sites (sites of 13 residues surrounding lysines) had ≥9 residues in common with the training positives. Phosphosite plus annotates ubiquitination in three sub-classes, derived from different experimental protocols. Low-throughput (LTP) data are taken from the literature. High-throughput mass spectrometry data can be distinguished as either taken from multiple sources in the literature (PUB) or unpublished and generated at Cell Signaling Technology, Inc. (CST). Our main independent performance (generalization) will be evaluated on PUB. However, some interesting observations about CST and LTP will also be shown.

Machine learning

Two different machine learning approaches were used to develop RUBI, SVMs (Cortes and Vapnik 1995) and bi-directional recurrent neural networks (BRNNs) (Baldi et al. 1999). SVMs were tested with four kernel functions: linear (1 parameter), polynomial (3 parameters), radial basis (2 parameters) and sigmoid (2 parameters). A grid search was used to determine the best SVM parameters for each kernel and the C parameter. We split the data into ten random folds. Each split consisted of 80 % for training parameters, 10 % for validation and 10 % for testing on unseen data. The validation set was used as an overfitting flag since generalization on unseen data can be measured. Three values were used to flag overfitting: low C parameter, large margin (i.e., low $\|w\|$ in SVM formulation, see Cortes and Vapnik 1995) and sensitivity at 5 % false-positive rate. The former two are commonly known to affect generalization and the latter gives a measure of the generalization with a low false-positive count. The input vectors were the encoded sequence motifs of length M centered on lysines ($M = 13$ here). Assuming the central lysine is redundant information, the input vector for the SVM was $[X(i-6), \ldots, X(i-1), X(i+1), \ldots, X(i+6)] \in \mathbb{N}^{252}$ where $X() \in \mathbb{N}^{21}$ and i was the location of the central lysine. Each of the 21 components in $X()$ was sorted alphabetically by the amino acid symbol. One-hot encoding was employed, i.e., a component was set to 1 for the corresponding amino acid symbol (e.g., 1st position set to 1 for alanine and so on) and the rest set to 0. When the motif was extracted from the N and C termini (i.e., $i < 7$ and $i > N - 7$, respectively), the 21st position was set to one. After extensive testing the radial basis kernel outperformed the other alternatives (data not shown), and only results for this will be described throughout the paper.

BRNNs can be likened to an ensemble of three neural networks, learning the N-terminal sequence context, the sequence and the C-terminal sequence context, respectively (Baldi et al. 1999). This important local context surrounding the lysine was learned and stored as hidden layers using two specialized neural networks. It is important to note that BRNNs capture local context and are unable to capture the entire sequence. Where regular neural networks

Table 2 Method definitions used in the paper

Acronym	Sequence	Additional
SVM-M40	Motif only	
BRNN-S40	All residues	
BRNN-S40+ali	All residues	Alignment
RUBI: best+disorder	All residues	Disorder

and SVMs use a sliding window of predetermined size, BRNNs learn this context information through the recursive dynamics of the network. This reduces the number of parameters and extracts information implicitly from the surrounding local context. An identical BRNN was implemented as described in Walsh et al. (2012) and Pollastri and McLysaght (2005), with one-hot encoding of the amino acids as above.

Models

Several method variants were tested to develop RUBI, as summarized in Table 2. The main distinction was between motif-based SVMs or full sequence-based BRNNs with additional input. The background dataset was constructed with all residues present in the set of full sequences excluding known ubiquitination sites. This possibly included some lysines which can be ubiquitinated, but for which experimental evidence is missing, i.e., experimental false negatives. The sequence information was then encoded either as a local motif or the full sequence. For the full sequence, learning proceeded on each residue, including non-lysines. For the motifs learning proceeded on the 13 motif fragment only. When calculating performance on unseen data, only lysines were considered.

Two additional information sources were tested: multiple sequence alignments and predicted intrinsic disorder. Multiple sequence alignments were calculated using three rounds of PSI-BLAST (Altschul et al. 1997) with options (b) 3,000, (e) 0.001, (h) 1e−10 on the UniRef90 database. The frequency of each amino acid and gap frequencies in the multiple sequence alignment at a given position along the sequence was used instead of one-hot encoding. This has been previously shown to be useful for secondary structure prediction (Rost and Sander 1993). For intrinsic disorder, the probability of disorder at a sequence position was derived from ESpritz (Walsh et al. 2012) and encoded as a new component to the input vector.

Model construction was an iterative process where experiments such as site distribution, machine learning technology (SVM and BRNN) and alignments were tried and poor performers discarded. Finally, with the best surviving model, particular attention was paid to protein disorder and its relationship to ubiquitination. RUBI performance for each model was estimated using tenfold cross-validation.

Both SVM and BRNN produce a prediction score and the threshold producing optimal decisions often depends on the ratio of positive to negative examples. For all models, the decision threshold was determined at low false-positive rate on the training set. This decision threshold can thus be considered an extra learning parameter.

Comparison to other methods and availability

To compare RUBI to the state of the art, we tried to download and/or request from the authors executables for other published methods. Some methods (Chen et al. 2011; Cai et al. 2012) are available only as a server for single-input sequences and had to be discarded. UbiPred (Tung and Ho 2008) and UbPred (Radivojac et al. 2010) could be installed locally and will be used for comparison. Two recently published methods, UbiProber and hCKSAAP_UbSite, provided their training and testing data (Chen et al. 2013a, b). The UbiProber and hCKSAAP_UbSite comparisons were important because they allowed us to compare our method to recent state-of-the-art human predictors. Four common accuracy measures are used in analogy to our previous work on disorder (Walsh et al. 2012), sensitivity (sens), specificity (spec), Matthews correlation coefficient (MCC) and area under the sensitivity vs false-positive rate curve (AUC).

RUBI is available both as a Linux executable for download and a web server able to process thousands of sequences at a time from the homepage. All datasets used to throughout the paper can be found together with server, executable and online documentation from url http://protein.bio.unipd.it/rubi/.

Results

The goal for developing RUBI was to find a model with high sensitivity at high specificity. High specificity (or low false-positive rate) is vital for confident determination of ubiquitination sites and for any algorithm aiding in vivo experiments (Schwartz 2012). The first concern was overestimation of performance due to similarity between fragments surrounding each lysine. To this end, we started by analyzing the distribution of lysine local context in the data and proceeded iteratively.

Site diversity

Given that there may be similar sites in our data, we examined a local context of length 13 surrounding each lysine in Seq40. Between positive and background, 10 % of the positive sites had high similarity, i.e., ≥9 residues in common with the negative sites (see supplementary Fig. S1). From a pattern matching point of view, the local context of the positive and negative lysines intersect considerably, making the discrimination problem difficult. In other words the classification is not easily separable and, moreover, non-linear algorithms should be used to separate data. Within each set of positive and background lysines, there could be pairs of common sites. This can result in algorithm learning more about these sites at the expense of others. Moreover, similar sites will overestimate performance because the testing fold may contain similar data in the learning fold of the cross-validation. The positive set was quite diverse with only 1.1 % sharing similarity with another site ≥9 residues (see supplementary Fig. S2 left). The negative set was also diverse with 4.5 % having ≥9 residues common, but this reduced dramatically after nine residues in common (see supplementary Fig. S2 right). To have good generalization, it is vital for any learning algorithm to assess the diversity of the local context as opposed to the full sequence diversity. We conclude that the Seq40 set was indeed diverse to a sufficient level and decided to leave it intact to maximize data size.

BRNN improvement over standard motif-based approaches

Standard motif-based approaches must define a sequence window a priori (13 residues in this work). However, the BRNN architecture allows learning a dynamic window which potentially captures a greater local context and reduces the likelihood of overfitting due to a decrease in parameters. It effectively takes the full sequence as input and tries to capture local context dynamically. These reasons, with perhaps others, allow for an increased performance of the BRNN over the SVM trained on the same data (see Table 3). The MCC almost doubled from 0.152 to 0.295 and AUC increased by 12 % points when compared with SVM-M40 and BRNN-S40. It was also important to note that the ratio of positive to background lysines was approximately 1:16. However, the results changed very little when balancing the ratio (AUC for SVM-M40 0.740, 0.724, 0.734, 0.729 and 0.740 for ratios 1:1, 1:4, 1:8, 1:12 and 1:16 respectively). We conclude that training with the BRNN and its full sequence representation to be best. Next, we examine if conservation information in the form of multiple sequence alignments improved performance.

Table 3 Cross-validation results

Method	MCC	Spec	Sens	AUC
SVM-M40	0.152	0.951	0.199	0.740
BRNN-S40	0.295	0.947	0.375	0.860
BRNN-S40+ali	0.133	0.947	0.187	0.707

SVM and BRNN and alignment cross-validation performance on the Seq40 set. 5 % FPR thresholds found on training folds since it is a parameter

Conservation

Table 3 shows a simple conservation encoding degraded performance by over 11 % points in AUC with a similar trend for MCC (BRNN-S40 vs BRNN-S40+ali). Ubiquitination was previously found to be unconserved across different species (Chen et al. 2011), suggesting that using many species (UniRef90) for our alignment sequence database might need to be revised. Perhaps, more sensitive alignments based only on eukaryotic sequences should be considered in the future. It was the goal of this work to construct a fast proteomic-scale predictor with good performance and, due to no improvements using simple PSI-BLAST alignments, the basic amino acid encoding was retained. At this stage, training with the BRNN and a simple amino acid encoding was found to be the best technique. Next, we examined the phenomenon of intrinsic disorder and ubiquitination for our *H. sapiens* dataset.

Intrinsic disorder and ubiquitination

There have been different views on whether intrinsically disordered regions/proteins are involved in ubiquitination, which may be due to the datasets available for each analysis. For example, observations in *S. cerevisiae*-based data suggested ubiquitination to be correlated with disorder (Radivojac et al. 2010; Cai et al. 2012). Recently, it was noted that *S. cerevisiae*-disordered proteins are highly ubiquitinated after heat-shock treatment (Ng et al. 2013). Other works based on larger-scale analysis report a weak correlation between structure and ubiquitination (Hagai et al. 2011; (Wagner et al. 2011). All these views need to be investigated in detail by collecting many more datasets, but for simplicity we restricted our analysis to our available data.

Disordered regions can be predicted from sequence with good accuracy (Monastyrskyy et al. 2011). Our accurate and fast method, ESpritz (Walsh et al. 2012), was used to predict if each amino acid in Seq40 was disordered. Table 4 proves that disorder as a feature produced a statistically significant gain of 1.6 % in MCC and almost 1 % point in AUC (RUBI-5 % FPR vs BRNN-S40 in Table 3; Student's test p value <0.01). In fact, as a baseline predictor it was observed that disorder probability alone can be used to find ubiquitination with an AUC of 0.584 (see supplementary Table S1). The final predictor (hereafter termed RUBI) was trained using a BRNN with a simple amino acid encoding plus disorder probability. Table 4 also shows the behavior of RUBI at two strict false-positive rates in the cross-validation.

Table 4 Cross-validation performance using disorder feature

Method	MCC	Spec	Sens	AUC
RUBI-5 % FPR	0.311	0.949	0.389	0.868
RUBI-1 % FPR	0.211	0.990	0.127	0.868

Experiment with disorder probability for each residue. 1 % and 5 % FPR thresholds found on training folds since it is a parameter

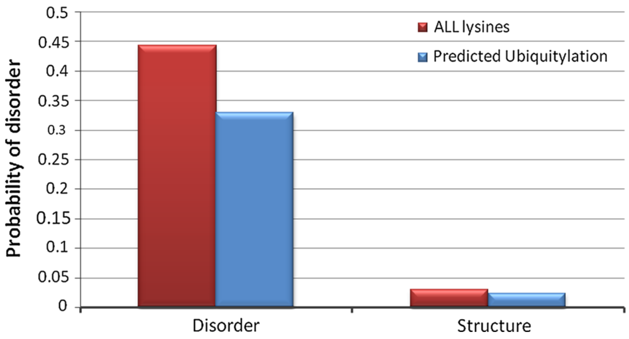


Fig. 1 Disorder probability for all and ubiquitinated lysines. Mean disorder probability for residues classified in disorder or structure by ESpritz X-ray (5 % FPR decision). Predictions performed on the PUB data with ESpritz X-ray disorder probabilities. In total there were 405 ubiquitination sites predicted in disordered regions and 3,799 sites in structured regions. Both differences are statistically significant (*p* value <<0.001, unpaired Wilcoxon rank sum test)

Although disorder probability improves RUBI's performance we still have not answered the question whether it correlated with structure or intrinsic disorder. A simple statistical test was carried out on the independent set (PUB) as it contains multiple experimental sources. The Wilcoxon rank sum test was calculated for both disordered and structured lysines on the 1,264 PUB proteins. The mean disorder probabilities of ubiquitinated lysines were compared with those of all lysines (i.e., the control set). Figure 1 confirms a statistical significance (*p* value <<0.001) for the structure–ubiquitination relationship. There were 405 ubiquitination sites predicted in disordered regions and 3,799 sites in structured regions. When a lysine is predicted in a disordered region, its probability of disorder when ubiquitinated is significantly lower (0.442 vs 0.329). When a lysine is predicted in a structured region, the probability of disorder is slightly but significantly lower when ubiquitinated (0.031 vs 0.025). In conclusion, our measurements showed that ubiquitination sites had a preference for structured regions on this data source.

Ubiquitination content

The main concern in the literature is the determination of the actual lysine sites, but little attention is paid to how many sites are contained per protein. To test the accuracy of RUBI beyond the mere recognition of individual sites and to check for systematic errors, we investigated the distribution of predicted vs experimental ubiquitination sites in the dataset. The distribution of experimental ubiquitination sites on the Seq40 dataset indicates that proteins with up to five experimental ubiquitination sites account for 91 % of the dataset, with those with a single site accounting for 49 % alone and the rest decaying exponentially (see supplementary Fig. S3). The results in the cross-validation, shown in Fig. 2, indicate that RUBI predicts the number of ubiquitination sites correctly, especially for proteins with fewer sites, ranging from ca. 25 % (for mono-ubiquitination) down to ca. 10 % (for ≥5 sites). Allowing up to three over- or underpredictions yields an accuracy of over 80 %. These results suggest that RUBI performs well across a wide range of proteins and the number of predicted ubiquitination sites roughly corresponds to the experimental sites.

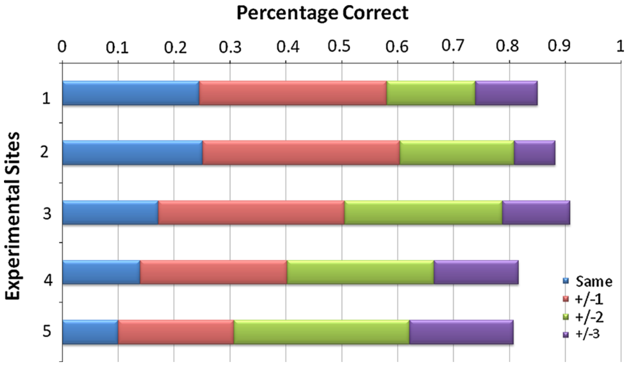


Fig. 2 Ubiquitination content performance. Percentage of correctly predicted ubiquitination sites (*x* axis) is shown with one through five experimentally ubiquitinated sites (*y* axis). The *colored bars* correspond to the same number of predicted and experimental sites and a difference of up to three predictions more or less than the experimental sites (color figure online)

Independent testing

RUBI was retrained on all 3,705 sequences in Seq40 and tested on an independent set. In addition, UbiPred (Tung and Ho 2008) and UbPred (Radivojac et al. 2010) were compared to our model. Table 5 shows the performance of UbPred, UbiPred and our final model on the full sequences in the PUB independent set. To demonstrate the different levels of confidence, thresholds at 5 and 1 % false-positive rates were used. The AUC for RUBI is 0.745, proving that the method behaves well on unseen data. The specificity and sensitivity remain high with only 10.9 % false positives (specificity 0.891) producing approximately 31 % true-positive rate (sensitivity 0.306). At a higher confidence level (RUBI-1 FPR %), as expected specificity was increased at the expense of sensitivity.

Table 5 Performance on the PUB independent set

Method	MCC	Spec	Sens	AUC
RUBI-5 FPR %	0.131	0.891	0.306	0.745
RUBI-1 FPR %	0.053	0.990	0.035	0.745
UbPred	0.024	0.830	0.211	0.599
UbiPred	0.062	0.653	0.537	0.592

5 % FPR thresholds found on Seq40. UbPred (medium confidence, 5 %FPR) motifs of length 29. UbiPred default 0.5 score decision, full sequences

On first inspection there was a good signal, albeit with a decrease compared to the cross-validation results in Table 4, which may be explained by different high-throughput experimental conditions (see next section). Despite this, RUBI still generalized well on different high-throughput mass spectrometry techniques. Both UbPred and UbiPred performed poorly on our dataset. This was to be expected, as performance was tested on humans but both were trained on yeast data.

While working on RUBI, two new predictors appeared: UbiProber (Chen et al. 2013a) and hCKSAAP_UbSite (Chen et al. 2013b). UbiProber and hCKSAAP_UbSite reported good performance when detecting *H. sapiens* sites, thus allowing us to compare fairly. In Table 6, RUBI was directly compared with the independent sets of both together with UbPred and UbiPred. For *H. sapiens*-based predictions, RUBI holds its performance, outperforming all predictors except UbiProber but still remaining comparable (UbiProber vs RUBI 0.782 vs 0.758). Table 6 also demonstrates RUBI has a good signal on other species such as *M. musculus* (AUC 0.616) and on the more distant species *S. cerevisiae* with AUC 0.750. The high *S. cerevisiae* result is particularly interesting because our model was trained on *H. sapiens* only, showing the possibility for cross species prediction.

Ideally, to avoid intersection of training and independent sets, RUBI should be retrained on the exact training and independent split proposed by each method. To ensure this, RUBI was retrained on the *H. sapiens* sets from UbiProber and hCKSAAP_UbSite. Table 7 shows that RUBI outperforms all others with respect to AUC (UbiProber vs RUBI 0.782 vs 0.818 and hCKSAAP_UbSite vs RUBI 0.757 vs 0.820) for *H. sapiens* detection. Table 7 is encouraging for future versions of RUBI which can be easily retrained and updated with the latest experimental data.

Experimental annotation is vital

It is commonly held that ubiquitination sites differ by species, but not much is known about differences in the experimental technique used to detect them. The Phosphosite plus database (Hornbeck et al. 2012) offers three styles of annotation allowing fluctuations in RUBI's performance to be measured if the annotation strategy is different. The three sets, PUB, CST and LTP (see "Datasets"), are considered separately as ROC curves in Fig. 3. Clearly, there was a substantial difference in the techniques used for experimental annotation, with AUCs of 0.745, 0.605 and 0.491 for PUB, CST and LTP, respectively.

Table 7 AUC testing performance on previously published *H. sapiens* data in UbiProber and hCKSAAP_UbSite

Method	UbiProber splits	Method	hCKSAAP_ UbSite splits
RUBI retrained	0.818	RUBI retrained	0.820
UbiProber *H. sapiens*[a]	0.782	hCKSAAP_UbSite[a]	0.757
UbiPred[a]	0.586	UbiPred[a]	0.560
UbPred[a]	0.596	UbPred[a]	0.497

RUBI model was trained on our original training set, but tested on sets from UbiProber and hCKSAAP_ubisite

[a] AUC taken directly from the corresponding publication. Table columns split into two *H. sapiens* data sets from UniProber and hCKSAAP_UbSite. Each row shows different predictors on the different data split

Table 6 AUC testing performance on previously published data in UbiProber and hCKSAAP_UbSite

UbiProber data				hCKSAAP_UbSite data	
Method	*H. sapiens*	*M. musculus*	*S. cerevisiae*	Method	*H. sapiens*
RUBI	0.758	0.616	0.750	RUBI	0.888
UbiProber *H. sapiens*[a]	0.782	0.838	0.899	hCKSAAP_UbSite[a]	0.757
UbiPred[a]	0.586	0.462	0.404	UbiPred[a]	0.560
UbPred[a]	0.596	0.644	0.736	UbPred[a]	0.497

RUBI model was trained on our original training set, but tested on sets from UbiProber and hCKSAAP_ubisite

[a] AUC taken directly from the corresponding publication. Table columns split into the two data sets UniProber and hCKSAAP_UbSite and further into species when available. Each row shows different predictors on the different data splits

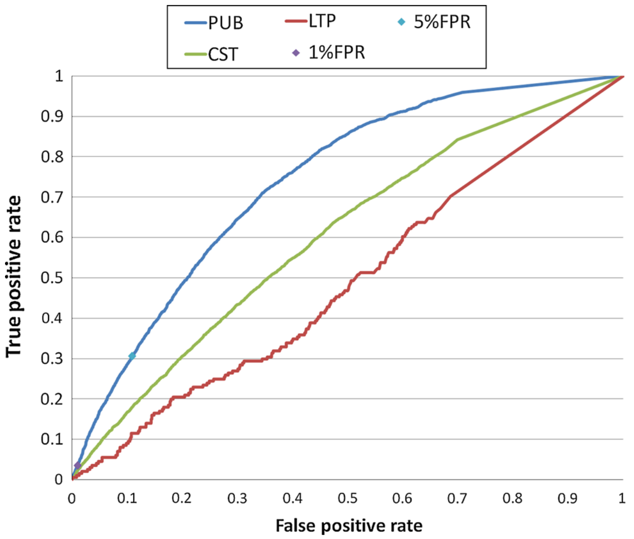


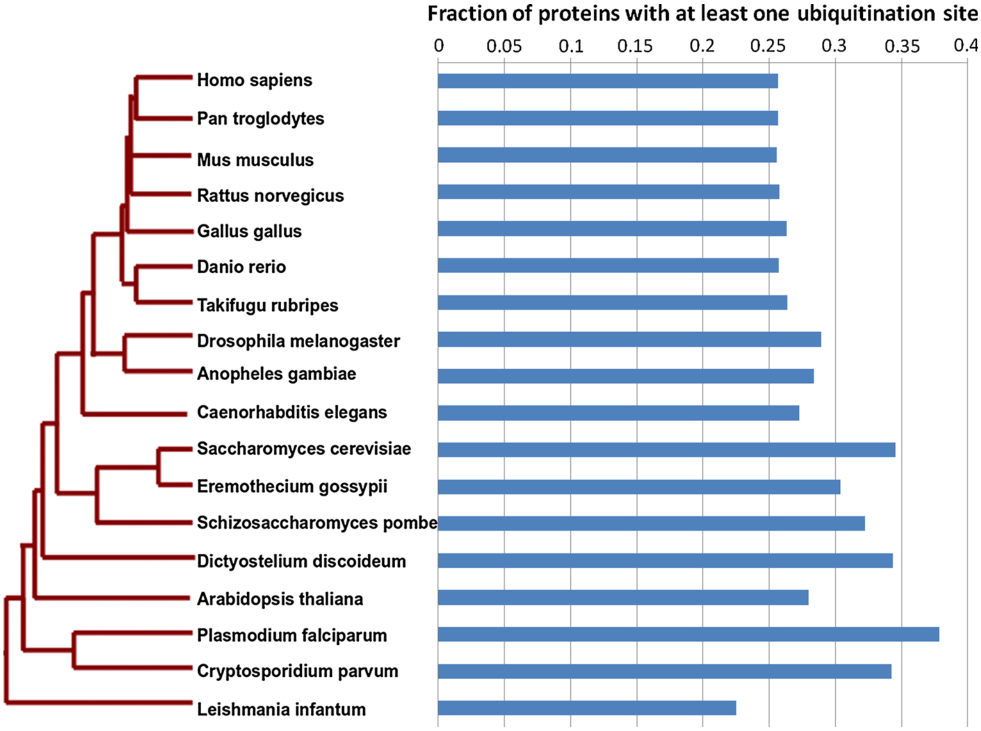
Fig. 3 Receiver operating characteristic (ROC) on different annotation styles. ROC curves are shown for different subsets of the independent dataset. Low-throughput (LTP) data are shown in *red*, while public high-throughput data (PUB) are shown in *blue* and unpublished Cell Signalling, Inc. data (CST) in *green*. *Purple* and *cyan diamonds* show the 1 and 5 % FPR on the PUB curve respectively (color figure online)

We can postulate that the sequence surrounding the lysine was different in each category. It may even be argued that rather than species variation, differences were due to the experimental technique used on each species. However, more experiments are needed to verify this point. Our model closely resembles the public high-throughput mass spectrometry data in the literature (see blue curve in Fig. 3; Table 5 for the PUB set performance and in Tables 6, 7 for further proof). This was to be expected as it was of most similar category to the training data. It is interesting that the LTP category is predicted poorly, suggesting that our model can only mimic high-throughput experiments. In addition, CST annotations were predicted to be substantially worse than the published literature annotations (PUB). It is important to note that we are by no means stating any experimental technique is incorrect, but merely trying to find the technique our model captures best.

Proteome-scale predictions

To make RUBI viable for high-throughput usage on a proteomic scale, it was benchmarked for speed on a standard Linux server. The CPU time (single core) per protein was found to be 0.42 s on average or ca. 7 min per thousand proteins (see supplementary Fig. S4), making it a very efficient predictor which can be easily applied on many genomes. We therefore applied RUBI at 1 % FPR to predict the distribution of high-confidence ubiquitination sites across a variety of genomes. The results for a representative set of model organisms are shown in Fig. 4. The average fraction of proteins with at least one ubiquitination site was found to be at least 25 % for all organisms, except *Leishmania infantum*. The fraction was very constant among higher eukaryotes, but raised and fluctuated among simpler organisms such as fungi and parasites. While the exact proportions will need further investigation, it is

Fig. 4 Fraction of predicted ubiquitinated proteins per model organism. The phylogenetic tree is shown together with the ratio of proteins with at least one ubiquitination site predicted by RUBI at 1 % FPR for each genome



interesting to note that RUBI predicts many possible novel high-confidence ubiquitination targets in the *H. sapiens* genome.

Conclusion

In this paper we have presented RUBI, a novel method for ubiquitination site prediction from sequence. The method was trained on one of the largest currently available experimental *H. sapiens* data and was shown to be robust and accurate across a wide range of conditions. The final predictor was constructed in an iterative manner and some factors influencing its performance were illustrated. The factors which boosted RUBI's performance included: (i) the sequence representation imposed by the machine learning algorithm and (ii) intrinsic disorder. Other performance factors, such as local lysine sequence distribution and addition of conservation from sequence alignments, were also analyzed. The former must be calculated to ensure site diversity in learning, while the latter degraded performance. Each factor could potentially aid in the development of other post-translational modification predictors. In addition to evaluating RUBI on individual sites, we also attempted to measure if it could detect the amount of ubiquitination per protein. We believe this is the first such measurement. The best model found in this work can be retrained on different datasets in a matter of days. As higher-quality data become available, the RUBI server will undergo systematic updates. Protein structure (non-disordered regions) was found to correlate with ubiquitination for the datasets used in this work. Intrinsic disorder analysis separating ubiquitination into different data sources (e.g., *S. cerevisiae* vs *H. sapiens* or different experimental techniques) might produce different correlations. We plan to integrate RUBI into our database of disorder annotations for proteins MobiDB (Di Domenico et al. 2012), thus allowing these correlations to be calculated easily. RUBI has a good generalization ability and signal across different eukaryotic organisms. It is also fast enough to enable the first genome-wide comparison of ubiquitination sites, which suggests the existence of thousands of possible ubiquitination sites awaiting experimental validation.

Acknowledgments The authors are grateful to members of the Biocomputing UP Lab for insightful discussions.

Conflict of interest The authors declare that they have no conflict of interest.

References

Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Baldi P, Brunak S, Frasconi P et al (1999) Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 15:937–946. doi:10.1093/bioinformatics/15.11.937

Bingol B, Sheng M (2011) Deconstruction for reconstruction: the role of proteolysis in neural plasticity and disease. Neuron 69:22–32. doi:10.1016/j.neuron.2010.11.006

Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294:1351–1362. doi:10.1006/jmbi.1999.3310

Blom N, Sicheritz-Pontén T, Gupta R et al (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics 4:1633–1649. doi:10.1002/pmic.200300771

Cai Y, Huang T, Hu L et al (2012) Prediction of lysine ubiquitination with mRMR feature selection and analysis. Amino Acids 42:1387–1395. doi:10.1007/s00726-011-0835-0

Chau V, Tobias JW, Bachmair A et al (1989) A multiubiquitin chain is confined to specific lysine in a targeted short-lived protein. Science 243:1576–1583

Chen ZJ, Sun LJ (2009) Nonproteolytic functions of ubiquitin in cell signaling. Mol Cell 33:275–286. doi:10.1016/j.molcel.2009.01.014

Chen Z, Chen Y-Z, Wang X-F et al (2011) Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. PLoS One 6:e22930. doi:10.1371/journal.pone.0022930

Chen X, Qiu J-D, Shi S-P et al (2013a) Incorporating key position and amino acid residue features to identify general and species-specific ubiquitin conjugation sites. Bioinformatics 29:1614–1622. doi:10.1093/bioinformatics/btt196

Chen Z, Zhou Y, Song J, Zhang Z (2013b) hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. Biochim Biophys Acta 1834:1461–1467. doi:10.1016/j.bbapap.2013.04.006

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. doi:10.1023/A:1022627411411

Di Domenico T, Walsh I, Martin AJM, Tosatto SCE (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. Bioinformatics 28:2080–2081. doi:10.1093/bioinformatics/bts327

Glickman MH, Ciechanover A (2002) The ubiquitin–proteasome proteolytic pathway: destruction for the sake of construction. Physiol Rev 82:373–428. doi:10.1152/physrev.00027.2001

Hagai T, Azia A, Tóth-Petróczy Á, Levy Y (2011) Intrinsic disorder in ubiquitination substrates. J Mol Biol 412:319–324. doi:10.1016/j.jmb.2011.07.024

Hicke L (2001) Protein regulation by monoubiquitin. Nat Rev Mol Cell Biol 2:195–201. doi:10.1038/35056583

Hoeller D, Hecker C-M, Dikic I (2006) Ubiquitin and ubiquitin-like proteins in cancer pathogenesis. Nat Rev Cancer 6:776–788. doi:10.1038/nrc1994

Hornbeck PV, Kornhauser JM, Tkachev S et al (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res 40:D261–D270. doi:10.1093/nar/gkr1122

Hunter T (2007) The age of crosstalk: phosphorylation, ubiquitination, and beyond. Mol Cell 28:730–738. doi:10.1016/j.molcel.2007.11.019

Ikeda F, Dikic I (2008) Atypical ubiquitin chains: new molecular signals. EMBO Rep 9:536–542. doi:10.1038/embor.2008.93

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

Lonard DM, O'Malley BW (2007) Nuclear receptor coregulators: judges, juries, and executioners of cellular regulation. Mol Cell 27:691–700. doi:10.1016/j.molcel.2007.08.012

Monastyrskyy B, Fidelis K, Moult J et al (2011) Evaluation of disorder predictions in CASP9. Proteins 79(suppl 10):107–118. doi:10.1002/prot.23161

Nalepa G, Rolfe M, Harper JW (2006) Drug discovery in the ubiquitin–proteasome system. Nat Rev Drug Discov 5:596–613. doi:10.1038/nrd2056

Ng AHM, Fang NN, Comyn SA et al (2013) System-wide analysis reveals intrinsically disordered proteins are prone to ubiquitylation after misfolding stress. Mol Cell Proteomics. doi:10.1074/mcp.M112.023416

Peng J, Schwartz D, Elias JE et al (2003) A proteomics approach to understanding protein ubiquitination. Nat Biotechnol 21:921–926. doi:10.1038/nbt849

Pollastri G, McLysaght A (2005) Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 21:1719–1720. doi:10.1093/bioinformatics/bti203

Radivojac P, Vacic V, Haynes C et al (2010) Identification, analysis and prediction of protein ubiquitination sites. Proteins 78:365–380. doi:10.1002/prot.22555

Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70 % accuracy. J Mol Biol 232:584–599. doi:10.1006/jmbi.1993.1413

Schwartz D (2012) Prediction of lysine post-translational modifications using bioinformatic tools. Essays Biochem 52:165–177. doi:10.1042/bse0520165

Schwartz D, Chou MF, Church GM (2009) Predicting protein post-translational modifications using meta-analysis of proteome scale data sets. Mol Cell Proteomics 8:365–379. doi:10.1074/mcp.M800332-MCP200

Sun L, Chen ZJ (2004) The novel functions of ubiquitination in signaling. Curr Opin Cell Biol 16:119–126. doi:10.1016/j.ceb.2004.02.005

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

Udeshi ND, Svinkina T, Mertins P et al (2013) Refined preparation and use of anti-diglycine remnant (K-ε-GG) antibody enables routine quantification of 10,000 s of ubiquitination sites in single proteomics experiments. Mol Cell Proteomics 12:825–831. doi:10.1074/mcp.O112.027094

Wagner SA, Beli P, Weinert BT et al (2011) A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. Mol Cell Proteomics. doi:10.1074/mcp.M111.013284

Walsh I, Martin AJM, Di Domenico T, Tosatto SCE (2012) ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 28:503–509. doi:10.1093/bioinformatics/btr682

Wong BR, Parlati F, Qu K et al (2003) Drug discovery in the ubiquitin regulatory pathway. Drug Discov Today 8:746–754. doi:10.1016/S1359-6446(03)02780-6