Amino Acids (2012) 43:447–455
DOI 10.1007/s00726-011-1100-2

ORIGINAL ARTICLE

Predicting protein sumoylation sites from sequence features

Shaolei Teng · Hong Luo · Liangjiang Wang

Received: 5 April 2011 / Accepted: 20 September 2011 / Published online: 7 October 2011

Abstract Protein sumoylation is a post-translational modification that plays an important role in a wide range of cellular processes. Small ubiquitin-related modifier (SUMO) can be covalently and reversibly conjugated to the sumoylation sites of target proteins, many of which are implicated in various human genetic disorders. The accurate prediction of protein sumoylation sites may help biomedical researchers to design their experiments and understand the molecular mechanism of protein sumoylation. In this study, a new machine learning approach has been developed for predicting sumoylation sites from protein sequence information. Random forests (RFs) and support vector machines (SVMs) were trained with the data collected from the literature. Domain-specific knowledge in terms of relevant biological features was used for input vector encoding. It was shown that RF classifier performance was affected by the sequence context of sumoylation sites, and 20 residues with the core motif ΨKXE in the middle appeared to provide enough context information for sumoylation site prediction. The RF classifiers were also found to outperform SVM models for predicting protein sumoylation sites from sequence features. The results suggest that the machine learning approach gives rise to more accurate prediction of protein sumoylation sites than the other existing methods. The accurate classifiers have been used to develop a new web server, called seeSUMO (http://bioinfo.ggc.org/seesumo/), for sequence-based prediction of protein sumoylation sites.

Keywords Protein sumoylation site prediction · Random forests · Support vector machines · Biological features · SeeSUMO

Introduction

Post-translational modifications regulate protein functions, and orchestrate a variety of cellular processes. Protein sumoylation, an important reversible post-translational modification, is essential for many eukaryotic cellular processes, including DNA damage recovery regulation, subcellular transport, transcription factor transactivation, protein stability, cell cycle progression and chromosome segregation (Zhao 2007). Small ubiquitin-related modifier (SUMO) can be covalently attached to and detached from specific lysine residues in target proteins (Geiss-Friedlander and Melchior 2007). Many sumoylated proteins, including huntingtin, DJ-1, ataxin-1 and tau, play key roles in disease states. For instance, the stability and correct targeting of huntingtin are controlled by sumoylation, and any alternations of the process may cause Huntington's disease (Steffan et al. 2004). Sumoylation is also involved in the pathogenesis of Parkinson's disease, Alzheimer's disease, neuronal intranuclear inclusion disease, amyotrophic lateral sclerosis, spinobulbar muscular atrophy, spinocerebellar ataxia type 1 and several human cancers (Sarge and Park-Sarge 2009).

Only one or a few lysine residues in a protein may be involved in sumoylation. It is rather difficult and time-consuming to identify the sumoylated lysine among many

Electronic supplementary material The online version of this article (doi:10.1007/s00726-011-1100-2) contains supplementary material, which is available to authorized users.

S. Teng · H. Luo · L. Wang (✉)
Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA
e-mail: liangjw@clemson.edu

L. Wang
J.C. Self Research Institute of Human Genetics, Greenwood Genetic Center, Greenwood, SC 29646, USA

candidate lysine residues through experimental approaches. Accurate computational prediction of protein sumoylation sites can help biologists better design their experiments and interpret the experimental data. A core consensus motif ΨKXE has been identified for sumoylation sites, in which Ψ represents an aliphatic amino acid (I, V, L, A, P or M), K is the sumoylation site, X indicates any amino acid, and E is glutamic acid. Extended sumoylation motifs have also been reported (Martin et al. 2007), such as negatively charged amino acid-dependent sumoylation motif [NDSM: ΨKXE + downstream cluster of (D/E)] (Yang et al. 2006), phosphorylation-dependent sumoylation motif (PDSM: ΨKXEXXSP) (Hietakangas et al. 2006) and SUMO-acetyl switch (ΨKXEP) (Stankovic-Valentin et al. 2007). These findings suggest that the sequence flanking the core motif (ΨKXE) may also contribute to the specific recognition of the sumoylation sites. Moreover, it is noteworthy that some sumoylation sites do not follow the above motifs, and not all lysine residues matched to these motifs are sumoylated. It is still challenging to accurately predict the true sumoylation sites recognized by the cellular machinery.

Accurate prediction of sumoylation site could help understand the mechanism of protein sumoylation underlying human genetic disorders. Several computational methods have been reported for predicting sumoylation sites. A statistical method used by the SUMOpre web server (Xu et al. 2008) can predict sumoylation sites at the overall accuracy of 96.71% and Matthews correlation coefficient (MCC) of 0.6364 in cross-validation tests. Xue et al. (2006) developed the SUMOsp 1.0 web server, which used the group-based phosphorylation scoring (GPS) algorithm with the pattern recognition strategy MotifX for sumoylation site prediction. SUMOsp 2.0 (Ren et al. 2009) was developed with the upgraded GPS algorithm. It has been shown that SUMOsp 2.0 reached better predictive performance than SUMOsp 1.0. However, the previous studies did not utilize the domain-specific knowledge for classifier construction.

Domain-specific knowledge in terms of relevant biological features can be used to enhance classifier performance for predicting DNA-binding residues and protein stability changes upon amino acid substitutions (Teng et al. 2010; Wang and Brown 2006a; Wang and Brown 2006b). For example, the predictive performance of DNA-binding site prediction could be significantly improved by using biochemical features (Wang and Brown 2006a) and evolutionary information (Ahmad et al. 2004) for input vector encoding. In this study, we have developed a new approach for sequence-based prediction of protein sumoylation sites using random forests (RFs) and support vector machines (SVMs). The biological knowledge in terms of 40 sequence features were used for input encoding. It was found that the RF classifier performance was affected by sequence context of sumoylation sites. The results obtained in this study indicate that the RF classifiers achieved better predictive performance than the SVM classifiers and previous predictors. To make our classifiers publicly accessible to the biological research community, we have developed a new web server called seeSUMO (freely available at http://bioinfo.ggc.org/seesumo).

Methods

Data

We collected 457 experimentally verified sumoylation sites in 263 proteins, by searching the research articles in NCBI PubMed using ‘SUMO’ and ‘sumoylation’ as keywords (Supplementary Table 1). This dataset contained all the instances used by SUMOpre (Xu et al. 2008), including 268 sumoylation sites in 159 proteins from research articles reported before August 10, 2006. The other 189 sumoylation sites have been manually collected from research articles published between August 10, 2006 and June 1, 2010. The amino acid sequences of these proteins were extracted from the SwissProt database. In order to remove redundancy in the dataset, the blastclust program in the BLAST software package (http://blast.ncbi.nlm.nih.gov/) was used for clustering analysis with a 25% sequence identity threshold, and ClustalX (Larkin et al. 2007) was used for multiple sequence alignment of the sequences in each cluster. The redundant sumoylation sites were manually removed from the dataset. The final dataset contains 9,952 lysine residues in 247 proteins, including 425 non-redundant sumoylation sites used as positive data instances and 9,527 non-sumoylated lysine sites used as negative data instances.

To compare the predictive performance of our classifiers with previous predictors, the final dataset was divided into two subsets. The training dataset included 377 sumoylation sites and 8,237 non-sumoylated lysine residues in 221 proteins from publications before January 2010. The remaining 48 sumoylation sites and 1,290 non-sumoylation sites in 26 proteins reported after January 2010 were used as the test dataset for classifier evaluation and comparison.

Sequence logos

Protein sequence logos (http://www.cbs.dtu.dk/~gorodkin/appl/plogo.html) was used to generate the sequence logo for visualizing the sequence pattern of sumoylation motifs. The 28 residues with the core motif ΨKXE in the middle of 388 known sumoylation sites was used as the inputs, and the frequencies of residues at each position were measured in bits of information as described in previous studies (Gorodkin et al. 1997; Schneider and Stephens 1990). The height of residue k at position i (d_{ik}) is proportional to its frequency relative to the expected frequencies, which is computed as follows:

$$d_{ik} = \frac{q_{ik}/p_k}{\sum_l q_{il}/p_l} I_i \tag{1}$$

where q_{ik} represents the fraction of residue k at position i, and p_k indicates the priori amino acid distribution, which was set to the amino acid composition obtained from UniProtKB/Swiss-Prot Release 57.15 in this study. I_i is the information content of position i as described below:

$$I_i = \sum_{k \in A} I_{ik} = \sum_{k \in A} q_{ik} \log_2 \frac{q_{ik}}{p_k} \tag{2}$$

where A is the set of residues including gaps.

Biological features

Forty biological features, including ten biochemical features, seven structural features, nine thermodynamic features, six empirical features and eight other biological features, selected from Protscale (Gasteiger et al. 2005) and AAindex (Kawashima and Kanehisa 2000), were used to encode each amino acid residue in a data instance (Supplementary Table 2). These features represent different types of biological knowledge such as biochemical properties, structural information, protein stability, folding energy, etc. For example, the biochemical feature, polarity (P), represents the dipole–dipole intermolecular interactions between the positively and negatively charged residues, and the structural feature, conformational parameter for alpha-helix (A), indicates the tendency of an amino acid to form the secondary structures, alpha-helix. Some of these features were used for predicting DNA-binding residues and protein stability changes upon amino acid substitutions in previous studies (Teng et al. 2010; Wang and Brown 2006a; Wang and Brown 2006b).

Evolutionary information

It was shown that utilizing the evolutionary information in terms of position-specific scoring matrix (PSSM) scores could improve the performance of RFs for DNA-binding site prediction (Ahmad and Sarai 2005). The PSSM scores generated by PSI-BLAST in this study indicated how well each position of a sequence was conserved among its homologues. The protein sequences downloaded from UniProtKB/Swiss-Prot (http://www.pir.uniprot.org/, release 57.15) were used as the reference database, and PSI-BLAST was run for three iterations with the E value threshold set to 1e−5.

Random forests

The use of 40 biological features and evolutionary information for input vector encoding gives rise to a large number of input variables, especially with a large window size. Considering the relatively small number of positive instances (experimentally identified sumoylation sites) available for this study, this might result in model overfitting. To avoid model overfitting, the RF learning algorithm was used in this study. A typical RF model is made up of many independent decision trees constructed using bootstrap samples from the training data. During tree construction, m variables out of all the n input variables ($m \ll n$) are randomly selected at each node, and the tree nodes are split using the selected m variables. For classifying a data instance, a RF classifier combines the votes made by the decision trees, and gives the most popular class as the output of the ensemble. Because of the random feature selection, RFs have the capability of handling a large number of input variables and avoid model overfitting.

In this study, the RF algorithm is implemented using the randomForest package in R. The number of variables selected to split each node (*mtry*) was set to 6, and the number of trees to grow (*ntree*) was set to 1,000. Other values of the *mtry* and *ntree* parameters for training were also examined, but did not result in significant improvement of classifier performance.

Support vector machine training

Support vector machine classifiers were also constructed, and compared with RF classifiers for protein sumoylation site prediction. The SVM method is a data-driven approach for binary classification. The SVM learning algorithm can be described by four basic concepts, including the separating hyperplane, the maximum-margin hyperplane, the soft margin, and the kernel function (Noble 2006). For a linear classification, data instances are represented as n-dimensional vectors, and an $(n - 1)$ dimensional hyperplane is used to separate the positive instances from the negative ones. Non-linear classifications are generally used for the analysis of complex biological data. In these cases, a kernel function can be used to measure the distance between data points in a higher dimensional space, which allows the SVM algorithm to fit the maximum-margin hyperplane in the transformed space. The SVMlight software package (http://svmlight.joachims.org/) was utilized to construct the SVM classifiers using the radial basis function (RBF) kernel in this study.

In this study, 40 biological features were used to build the SVM models. However, the features used for classifier construction might contain redundant or correlated information. Thus, feature selection was performed to choose an optional subset of relevant features for constructing simple, efficient models. The five relevant features were selected by RFs, and then used to construct SVM classifiers.

Classifier evaluation

The predictive performance of classifier was evaluated by tenfold cross-validation. The whole dataset was randomly distributed into ten equal-sized portions. In each of the ten iterations, the classifier was trained using nine of the ten portions and tested using the remaining portion. Since the dataset was imbalanced with only 4% of lysine residues as sumoylation sites, the positive instances of training data were replicated to get the approximately equal number with the negative instances. However, the positive instances in the test data were not replicated. The prediction results made for the test data instances in all the ten iterations were combined and evaluated by various performance measures, including accuracy (AC), sensitivity (SN), specificity (SP), strength (ST) and MCC:

$$\text{Accuracy (AC)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (3)$$

$$\text{Sensitivity (SN)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (4)$$

$$\text{Specificiy (SP)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad (5)$$

$$\text{Strength (ST)} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \qquad (6)$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \qquad (7)$$

where TP is the number of true positives; TN the number of true negatives; FP the number of false positives; and FN is the number of false negatives. For imbalanced datasets, the accuracy alone could be misleading. Thus, sensitivity, specificity and their average (strength) were also computed from prediction results. MCC was used to measure the correlation between predictions and the actual class labels. However, different trade-offs of sensitivity and specificity may give rise to different MCC values for a classifier.

The receiver operating characteristic (ROC) curve (Swets 1988) is probably the most robust approach for classifier evaluation and comparison. In the present study, the ROC curve was generated by varying the output threshold of a RF classifier and plotting the true positive rate (sensitivity) against the false positive rate (1 − specificity) for each threshold value. Since the ROC curve of an accurate classifier is close to the left-hand and top borders of the plot, the area under the curve (AUC) can be used as a reliable measure of classifier performance (Bradley 1997). The range of AUC value is 0.5 (random guessing) to 1 (perfect classifier).

Results and discussion

Sequence patterns of protein sumoylation sites

Protein sumoylation sites are often identified with the consensus motif ΨKXE, where Ψ represents an aliphatic amino acid (I, V, L, A, P or M), and X indicates any residue. However, 159 (~35%) of 457 known sumoylation sites in this study do not contain the core motif (ΨKXE), whereas the dataset contains 228 non-sumoylated lysine residues that match this motif. To visualize the sequence patterns in sumoylation sites and their flanking sequences, the sequence logo was generated using the 28-residue sequences from 388 experimentally identified sumoylation sites (Fig. 1). The result suggests that certain positions outside of the core motif (ΨKXE), such as the positions −7, 1, 3, 4 and 9, may contain some information for the specific recognition of sumoylation sites in the cell. For instance, the most abundant residue at positions 1 is proline (P), which agrees well with the SUMO-acetyl switch (ΨKXEP). Interestingly, glutamine (Q), methionine (M) and threonine (T) appear to be more abundant than any other residues in the X position of the core motif ΨKXE, suggesting that there may be subtle amino acid preference at the X position.

The above observations suggest that the flanking sequences of protein sumoylation sites have some subtle patterns. However, these patterns may not be modeled completely by consensus motifs or sequence logos, which do not consider the dependence among neighboring residues. Thus, a machine learning approach has been developed in

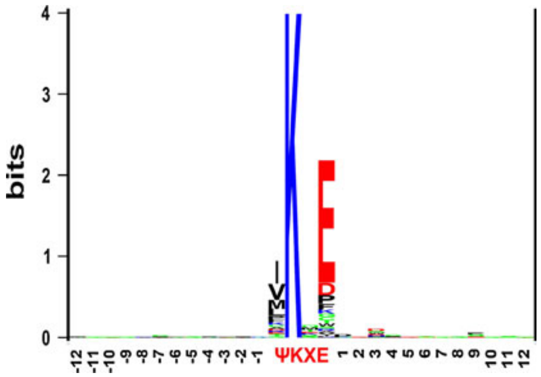


Fig. 1 The sequence logo of the protein sumoylation motif (ΨKXE) and its flanking residues

this study to model the sequence patterns of protein sumoylation sites.

Effect of sequence context on classifier performance

We first constructed RF classifiers using the 40 biological features for input vector encoding. The RF classifiers were trained with data instances of various window sizes. The results suggest that RF classifier performance was affected by the sequence context of sumoylation sites (Table 1). The classifier constructed with KX (window size $w = 2$) gave predictive performance with the prediction strength (ST) = 57.07%, MCC = 0.0590 and ROC AUC = 0.6107. The classifier performance was improved significantly when the core motif ΨKXE ($w = 4$) was used for input encoding. The classifier gave the prediction strength at 82.04% with MCC = 0.5379 and AUC = 0.9024. When the neighboring residues of the core motif were used to construct the classifiers, the predictive performance was further improved. For example, the classifier constructed with ΨKXE+/−5 ($w = 14$) achieved the highest MCC at 0.6786. Since the dataset was imbalanced with only 4% of lysine residues as the sumoylation sites, the ROC AUC is probably the most reliable performance measure for the present study. The classifier using the 20 residues with the core motif ΨKXE in the middle (ΨKXE+/−8, $w = 20$) reached the highest ROC AUC at 0.9200. The classifier also shows the highest overall accuracy at 97.68% with 56.00% sensitivity and 99.50% specificity, and high MCC = 0.6711. Thus, this RF classifier is considered as the best classifier in Table 1.

The ROC analysis for investigating the effect of sequence context information on RF classifier performance has been shown in Fig. 2. The classifier constructed with ΨKXE is clearly better than the classifier constructed with KX. Furthermore, the classifier using 20 residues with the core motif in the middle (ΨKXE+/−8) appears to be slightly better than the classifier constructed with ΨKXE. The results suggest that the context information in the flanking sequences may be useful for sumoylation site prediction.

RF versus SVM classifiers

Support vector machines have been widely used for biological pattern classification. In this study, we constructed SVM classifiers using the 40 biological features, and compared their ROC AUC values with those of the RF classifiers over various window sizes. As shown in Fig. 3, the RF classifiers using the 40 features (RF40) achieved comparable performance measures over various window sizes with the highest AUC at $w = 20$ (ΨKXE+/−8). However, SVM classifiers using the 40 features (SVM40) showed significantly degraded performances with large window sizes. For example, the AUC value of SVM40 decreased from 0.8090 to 0.5254 when the window size was increased from $w = 4$ (ΨKXE) to $w = 8$ (ΨKXE+/−2). Thus, the SVM classifiers did not achieve the same level of predictive performance as the RF classifiers. The possible explanation is that some of the 40 features may contain redundant or correlated information for sumoylation site prediction, which may have caused the degradation of SVM classifier performance.

To enhance the predictive performance of SVM classifiers, feature selection was performed using RFs. Five highly relevant features selected by RFs, including polarity (*P*), conformational parameters for alpha-helix (*A*) and coil

Table 1 Effect of sequence context on predictive performance of random forest classifiers

Sequence context	AC (%)	SN (%)	SP (%)	ST (%)	MCC	ROC AUC
KX	61.80	51.89	62.25	57.07	0.0590	0.6107
ΨKXE	95.27	**67.57**	96.52	**82.04**	0.5379	0.9024
ΨKXE+/−1	97.28	61.35	98.90	80.13	0.6489	0.9038
ΨKXE+/−2	97.42	59.73	99.12	79.42	0.6582	0.9145
ΨKXE+/−3	97.45	60.00	99.15	79.58	0.6638	0.9172
ΨKXE+/−4	97.54	60.28	99.20	79.74	0.6688	0.9074
ΨKXE+/−5	97.63	60.00	99.31	79.65	**0.6786**	0.9103
ΨKXE+/−6	97.57	57.78	99.34	78.56	0.6668	0.9048
ΨKXE+/−7	97.54	54.86	99.40	77.13	0.6508	0.9188
ΨKXE+/−8	**97.68**	56.00	99.50	77.75	0.6711	**0.9200**
ΨKXE+/−9	97.59	51.71	99.60	75.66	0.6522	0.9133
ΨKXE+/−10	97.57	47.65	99.70	73.67	0.6340	0.9149
ΨKXE+/−11	97.46	43.82	99.75	71.79	0.6115	0.9136
ΨKXE+/−12	97.43	42.35	**99.79**	71.07	0.6056	0.9124

The highest value in each column is in bold

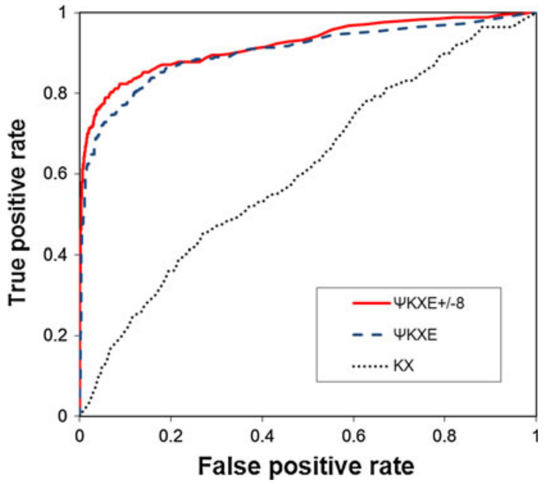


Fig. 2 ROC curves to show the effect of context information for sumoylation site prediction

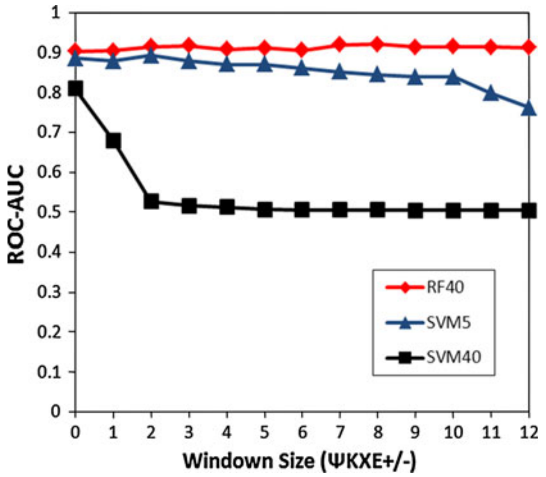


Fig. 3 Performance comparisons of random forest (RF) and support vector machine (SVM) classifiers

(*C*), short and medium range non-bonded energy per residue (Er) and free energy in alpha-helical conformation (Ea), were used to construct the SVM classifiers (SVM5). As shown in Fig. 3, the ROC AUC values of the SVM5 classifiers were higher than those of the SVM40 classifiers over various window sizes. For example, the SVM5 classifier constructed using eight residues (ΨKXE+/−2, $w = 8$) achieved the highest AUC value of 0.8917 over various window sizes (Fig. 3), which was significantly higher than the AUC value of the SVM40 classifier (AUC = 0.5254 at $w = 8$). This classifier reached the prediction strength at 78.42% (58.38% sensitivity and 98.46% specificity) and MCC = 0.5902 (Table 2). The SVM5 classifier constructed with ΨKXE+/−2 was regarded as the best SVM classifier in this study.

Table 2 Comparison of random forest and support vector machine classifiers constructed with ΨKXE+/−2 ($w = 8$)

Features	AC (%)	SN (%)	SP (%)	ST (%)	MCC	ROC AUC
RF40	**97.42**	**59.73**	99.12	**79.42**	**0.6582**	**0.9145**
SVM5	96.73	58.38	98.46	78.42	0.5902	0.8917
SVM40	95.66	0.00	**99.99**	49.99	-0.0023	0.5254

The highest value in each column is in bold

However, the RF40 classifiers still outperformed the SVM5 models (Fig. 3). As shown in Table 2, The RF40 classifier constructed using eight residues (ΨKXE+/−2) achieved the prediction strength of 79.42% with MCC = 0.6582 and AUC = 0.9145, which were higher than those of the SVM5 classifier in the same window size. Thus, the RF algorithm appears to be better for predicting protein sumoylation sites from sequence features. The possible explanation is that RFs can handle a large number of input variables and avoid model overfitting. The feature-encoded input vector has a large number of variables, especially with a large window size. For example, when 20 residues ($w = 20$) are used for classifier construction, the number of input variables is 800 for classifiers using 40 features and 100 for classifiers using five features. The large number of input variables, together with the small number of positive instances, may lead to model overfitting.

Use of evolutionary information

Evolutionary information in terms of PSSM scores was previously shown to improve classifier performance (Ahmad and Sarai 2005; Pu et al. 2007; Wang et al. 2010). To determine whether or not sumoylation site prediction could be further improved by using evolutionary information, the PSSM scores of 20 residues with the core motif ΨKXE in the middle were used to construct the RF classifiers. The scores in a PSSM represent how well each position of a protein sequence was conserved among its homologues. As shown in Table 3, the RF classifier constructed with PSSMs (PSSM, Table 3) reached the prediction strength of 51.96% with MCC = 0.1566 and AUC = 0.8672. By using both PSSMs and the 40

Table 3 Effect of evolutionary information on protein sumoylation site prediction

Features	AC (%)	SN (%)	SP (%)	ST (%)	MCC	ROC AUC
PSSM	95.90	4.00	**99.91**	51.96	0.1566	0.8672
Bio	**97.68**	**56.00**	99.50	**77.75**	**0.6711**	**0.9200**
Bio + PSSM	97.56	50.00	99.64	74.82	0.6443	0.9181

The highest value in each column is in bold biological features for input vector encoding, the RF classifier (Bio + PSSM, Table 3) gave a relatively high classifier performance (74.82% prediction strength, MCC = 0.6443 and AUC = 0.9181). However, these performance measures were not significantly different from those of the RF classifier using the biological features only (Bio, Table 3).

The ROC curves of the three RF classifiers are compared in Fig. 4. The results confirm that classifier performance is not improved by adding the evolutionary information to the biological features for input encoding. The possible explanation is that the PSSM, which is designed for PSI-BLAST searches, may not capture the evolutionary information for sumoylation site prediction. Another possibility is that the 40 biological features may already contain the evolutionary information necessary for predicting sumoylation sites.

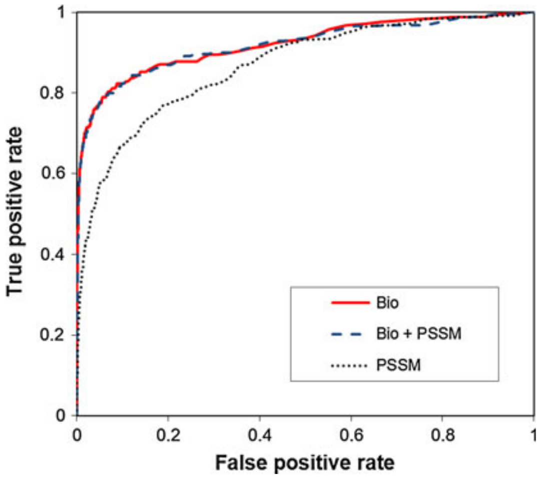


Fig. 4 ROC curves to show the effect of evolutionary information on classifier performance

Comparison with previous studies

The existing computational methods for protein sumoylation site prediction include SUMOplot (http://www.abgent.com//tools/toSumoplot), SUMOsp2 (http://sumosp.biocuckoo.org/online.php) and SUMOpre (Xu et al. 2008). The datasets used in these previous studies are smaller than the dataset used in the present work. We manually collected additional instances of experimentally identified sumoylation sites from the latest publications. To further demonstrate the improved performance of our classifiers, the most accurate RF classifier (ΨKXE+/−8, Table 1) and SVM classifier (SVM5, Table 2) have been compared with the previous classifiers, SUMOplot and SUMOsp2, using an independent test dataset with 48 sumoylation sites reported after January 2010. SUMOplot predicts the probability of sumoylation sites based on the SUMO consensus sequence and hydrophobicity, whereas SUMOsp2 (Ren et al. 2009) uses two pattern recognition strategies (GPS and MotifX) for sumoylation site prediction. The two types of prediction in SUMOplot were low (motifs with low probability) and high (motifs with high probability), whereas the three levels of stringency in SUMOsp2 were low, medium and high. The corresponding thresholds of classifier output in our approach (seeSUMO) were set to −0.2 (low), 0 (medium) and 0.2 (high).

As shown in Table 4, the overall accuracy (AC), specificity (SP) and MCC of our SVM classifier (seeSUMO-SVM) and RF classifier (seeSUMO-RF) are considerably higher than those of SUMOsp2 and SUMOplot in the low-threshold predictions. SUMOsp2 with its medium threshold gave the prediction strength at 68.31% (43.75% sensitivity and 92.87% specificity) and MCC = 0.2449. Our SVM classifier achieved a similar level of performance with 67.36% prediction strength, 41.67% sensitivity, 93.06% specificity and MCC = 0.2361. The RF classifier with the medium threshold reached higher performance with 71.60% prediction strength, 51.16% sensitivity, 92.04% specificity and MCC = 0.2639. For the high-threshold predictions, the overall accuracy and MCC of our RF classifier are also higher than those of SUMOsp2 and SUMOplot. It is noteworthy that the MCC values of our RF classifier are the highest in any level of threshold predictions. Therefore, the performance of the RF classifier developed in this study compares favorably with SUMOsp2 and SUMOplot for protein sumoylation site prediction.

SUMOpre (Xu et al. 2008) uses a statistical method for predicting protein sumoylation sites. It was not included in the direct comparisons because a web-based tool was not available for the classifier. However, our classifier uses a larger dataset and shows better predictive performance. For example, the dataset used by SUMOpre (Xu et al. 2008) contains 268 sumoylation sites, and the predictor reached 96.71% overall accuracy and MCC = 0.6364. In the present study, the dataset includes 377 sumoylation sites, and the best RF classifier (ΨKXE+/−8, Table 1) achieved 97.68% overall accuracy and MCC = 0.6711.

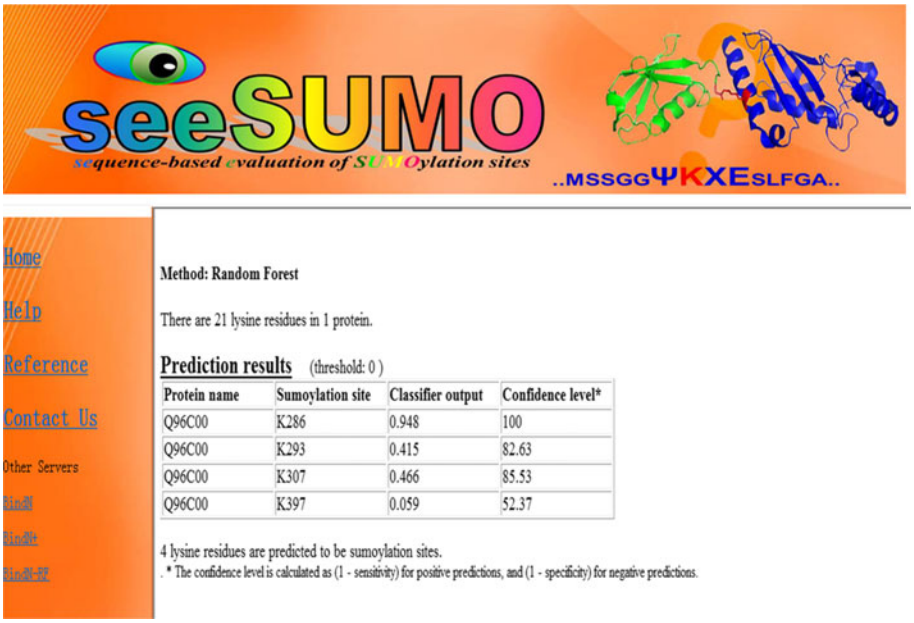
seeSUMO web server

To make our classifiers accessible to the biological research community, we have developed the seeSUMO web server (http://bioinfo.ggc.org/seesumo/). Users can enter an amino acid sequence or a batch of multiple sequences (up to 100 sequences) in the FASTA format, specify the methods, and input the proper threshold for prediction of protein sumoylation site. For prediction using the RF classifier, the system encodes the input sequences with the 40 biological features, and then calls the

Table 4 Comparison of classifier performance using an independent test dataset

Threshold	Methods	AC (%)	SN (%)	SP (%)	ST (%)	MCC
Low	SUMOplot	79.55	**68.75**	79.95	**74.35**	0.2196
	SUMOsp2	83.63	50.00	84.88	67.44	0.1753
	seeSUMO-SVM	89.48	47.92	91.04	69.48	0.2382
	seeSUMO-RF	**90.09**	53.49	**91.33**	72.41	**0.2644**
Medium	SUMOsp2	91.11	43.75	92.87	68.31	0.2449
	seeSUMO-SVM	**91.21**	41.67	**93.06**	67.36	0.2361
	seeSUMO-RF	90.70	**51.16**	92.04	**71.60**	**0.2639**
High	SUMOplot	91.69	**50.00**	93.24	**71.62**	0.2916
	SUMOsp2	94.24	39.58	**96.27**	67.93	0.3058
	seeSUMO-SVM	92.49	39.58	94.47	67.02	0.2528
	seeSUMO-RF	**94.36**	44.19	96.06	70.12	**0.3210**

The highest value for a given threshold in each column is in bold



Method: Random Forest

There are 21 lysine residues in 1 protein.

Prediction results (threshold: 0)

Protein name	Sumoylation site	Classifier output	Confidence level*
Q96C00	K286	0.948	100
Q96C00	K293	0.415	82.63
Q96C00	K307	0.466	85.53
Q96C00	K397	0.059	52.37

4 lysine residues are predicted to be sumoylation sites.

. * The confidence level is calculated as (1 - sensitivity) for positive predictions, and (1 - specificity) for negative predictions.

Fig. 5 Sample output from the seeSUMO web server

randomForest program of the R software package to classify the protein sumoylation sites using the most accurate RF model (ΨKXE+/−8, Table 1). For prediction using the SVM classifier, the system encodes the input sequences with the five highly relevant features, and then the best SVM classifier (SVM5, Table 2) constructed in this work is used to predict sumoylation sites in the query sequence. The seeSUMO web server will return the prediction results, including the protein name, potential sumoylated sites, classifier outputs and the prediction confidence levels (Fig. 5). The prediction confidence level is calculated as (1 − sensitivity) for positive predictions, and (1 − specificity) for negative predictions (Teng et al. 2010; Wang and Brown 2006a). An example output report is shown in Fig. 5 for zinc finger and BTB domain-containing protein 9 (ZBTB9). The lysine residue at position 286 is predicted to be a potential sumoylation site with a very high confidence level. This potential sumoylation site, but not the other sites, has recently been confirmed by the mass spectrometric analysis (Matic et al. 2010).

The seeSUMO web server is currently hosted using a Dell PowerEdge 6800 server computer. For single-sequence queries, the online prediction is very efficient. For batch jobs, seeSUMO will normally return prediction results within 1 min. When we tested seeSUMO with a batch of 100 protein sequences, the system response time was <1 min if the RF classifier was used, and about 15 s for prediction with the SVM classifier. The server website (http://bioinfo.ggc.org/seesumo/) contains help documents with detailed description of use cases of seeSUMO.

Conclusion

A new machine learning approach has been developed in this study for predicting protein sumoylation sites from protein sequence information. Domain-specific knowledge in terms of relevant biological features was used for input vector encoding. The results suggest that classifier performance is affected by the sequence context of sumoylation sites. The highest predictive performance (ROC AUC = 0.9200) has been achieved by the RF classifier using 20 residues with the core motif ΨKXE in the middle. Moreover, the RF classifiers were found to outperform SVM models on the imbalanced dataset. The classifiers developed in this study compare favorably in performance with the previous predictors for protein sumoylation site prediction. A web server, seeSUMO (http://bioinfo.ggc.org/seesumo/), has been developed to make our classifiers accessible to the biological research community.

Acknowledgments This work is supported by the CSREES/USDA, under project number SC-1700355. This is technical contribution number 5913 of the Clemson Experiment Station.

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

References

Ahmad S, Sarai A (2005) PSSM-based prediction of DNA binding sites in proteins. BMC Bioinforma 6:33

Ahmad S, Gromiha MM, Sarai A (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 20(4):477–486

Bradley A (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 30:1145–1159

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) The proteomics protocols handbook. Humana Press, Totowa

Geiss-Friedlander R, Melchior F (2007) Concepts in sumoylation: a decade on. Nat Rev Mol Cell Biol 8(12):947–956

Gorodkin J, Heyer LJ, Brunak S, Stormo GD (1997) Displaying the information contents of structural RNA alignments: the structure logos. Comput Appl Biosci 13(6):583–586

Hietakangas V, Anckar J, Blomster HA, Fujimoto M, Palvimo JJ, Nakai A, Sistonen L (2006) PDSM, a motif for phosphorylation-dependent SUMO modification. Proc Natl Acad Sci USA 103(1):45–50

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

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23(21): 2947–2948

Martin S, Wilkinson KA, Nishimune A, Henley JM (2007) Emerging extranuclear roles of protein SUMOylation in neuronal function and dysfunction. Nat Rev Neurosci 8(12):948–959

Matic I, Schimmel J, Hendriks IA, van Santen MA, van de Rijke F, van Dam H, Gnad F, Mann M, Vertegaal AC (2010) Site-specific identification of SUMO-2 targets in cells reveals an inverted SUMOylation motif and a hydrophobic cluster SUMOylation motif. Mol Cell 39(4):641–652

Noble WS (2006) What is a support vector machine? Nat Biotechnol 24(12):1565–1567

Pu X, Guo J, Leung H, Lin Y (2007) Prediction of membrane protein types from sequences and position-specific scoring matrices. J Theor Biol 247(2):259–265

Ren J, Gao X, Jin C, Zhu M, Wang X, Shaw A, Wen L, Yao X, Xue Y (2009) Systematic study of protein sumoylation: development of a site-specific predictor of SUMOsp 2.0. Proteomics 9(12):3409–3412

Sarge KD, Park-Sarge OK (2009) Sumoylation and human disease pathogenesis. Trends Biochem Sci 34(4):200–205

Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18(20):6097–6100

Stankovic-Valentin N, Deltour S, Seeler J, Pinte S, Vergoten G, Guerardel C, Dejean A, Leprince D (2007) An acetylation/deacetylation-SUMOylation switch through a phylogenetically conserved psiKXEP motif in the tumor suppressor HIC1 regulates transcriptional repression activity. Mol Cell Biol 27(7): 2661–2675

Steffan JS, Agrawal N, Pallos J, Rockabrand E, Trotman LC, Slepko N, Illes K, Lukacsovich T, Zhu YZ, Cattaneo E (2004) SUMO modification of Huntingtin and Huntington's disease pathology. Science 304(5667):100–104

Swets JA (1988) Measuring the accuracy of diagnostic systems. Science 240(4857):1285–1293

Teng S, Srivastava AK, Wang L (2010) Sequence feature-based prediction of protein stability changes upon amino acid substitutions. BMC Genomics 11(Suppl 2):S5

Wang L, Brown SJ (2006a) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 34(Web Server issue):W243–W248

Wang L, Brown SJ (2006b) Prediction of RNA-binding residues in protein sequences using support vector machines. Conf Proc IEEE Eng Med Biol Soc 1:5830–5833

Wang L, Huang C, Yang MQ, Yang JY (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst Biol 4(Suppl 1):S3

Xu J, He Y, Qiang B, Yuan J, Peng X, Pan XM (2008) A novel method for high accuracy sumoylation site prediction from protein sequences. BMC Bioinforma 9:8

Xue Y, Zhou F, Fu C, Xu Y, Yao X (2006) SUMOsp: a web server for sumoylation site prediction. Nucleic Acids Res 34(Web Server issue):W254–W257

Yang SH, Galanis A, Witty J, Sharrocks AD (2006) An extended consensus motif enhances the specificity of substrate modification by SUMO. EMBO J 25(21):5083–5093

Zhao J (2007) Sumoylation regulates diverse biological processes. Cell Mol Life Sci 64(23):3017–3033