Predicting SUMOylation Sites

Denis C. Bauer[1], Fabian A. Buske[1], and Mikael Bodén[1]

Institute for Molecular Bioscience, University of Queensland, Brisbane, Qld. 4072
Australia
d.bauer@imb.uq.edu.au,

Abstract. Recent evidence suggests that SUMOylation of proteins plays a key regulatory role in the assembly and dis-assembly of nuclear sub-compartments, and may repress transcription by modifying chromatin. Determining whether a protein contains a SUMOylation site or not thus provides essential clues about a substrate's intra-nuclear spatial association and function.
Previous SUMOylation predictors are largely based on a degenerate and functionally unreliable consensus motif description, not rendering satisfactory accuracy to confidently map the extent of this essential class of regulatory modifications. This paper embarks on an exploration of predictive dependencies among SUMOylation site amino acids, non-local and structural properties (including secondary structure, solvent accessibility and evolutionary profiles).
An extensive examination of two main machine learning paradigms, Support-Vector-Machine and Bidirectional Recurrent Neural Networks, demonstrates that (1) with careful attention to generalization issues both methods achieve comparable performance and, that (2) local features enable best generalization, with structural features having little to no impact. The predictive model for SUMOylation sites based on the primary protein sequence achieves an area under the ROC of 0.92 using 5-fold cross-validation, and 96% accuracy on an independent hold-out test set. However, similar to other predictors, the new predictor is unable to generalize beyond the simple consensus motif.

1 Introduction

SUMOylation is a post-translational modification attaching a small ubiquitin-like modifier (SUMO) covalently to a target protein. It has been shown that SUMO plays an important role in many essential biological functions, such as preserving the integrity and function of intra-nuclear compartments, chromatin organization and ultimately gene regulation [14, 9]. By modifying histones, dynamically competing with acetylation and ubiquitylation, SUMOylation appears to play a pivotal role in repressing transcription. Dysfunction of the SUMOylation pathway is related to several neurodegenerative diseases in human, such as Huntington's disease [5]. The significance of the SUMO conjugation system is further underscored by the apparent conservation through evolution among eukaryotic organisms.

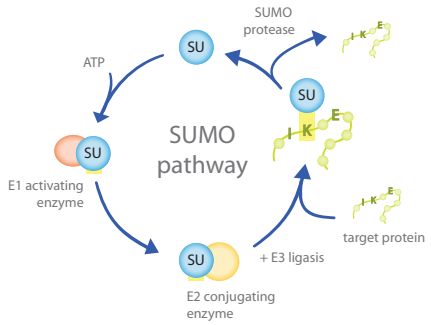


Fig. 1. SUMOylation pathway. The figure shows the role of the involved proteins in the SUMOylation pathway. E1 activates SUMO in an ATP requiring process. E2 attaches SUMO to the Lysine in the target protein, supported by E3. SUMO-protease removes SUMO from the protein, now free to be re-used in another cycle.

The SUMOylation pathway comprises four proteins: E1 activating enzyme, E2 conjugating enzyme, E3 ligase and SUMO-protease (illustrated in Fig. 1). E1 prepares SUMO for binding to the target protein (the substrate). E2 and E3 interact directly (in a concerted fashion) with the substrate at the SUMOylation site, usually conforming to the consensus motif, ΨKxE (where Ψ is a large hydrophobic residue, K is Lysine and E is Glutamic acid). E2 and E3 mediate the binding between SUMO and the central Lysine [11]. Finally, the SUMO-protease disassociates SUMO from the target protein.

Unfortunately, the motif is an unreliable predictor. Some substrates are modified on sites not matching the consensus motif [8]. Furthermore, not every consensus site in a protein is modified by SUMO. It has been suggested that there are additional factors, such as the appropriate presentation of the substrate sequence and protein sub-cellular location, which determine whether modification is completed [8].

To date, three specialized SUMOylation site predictors have been published. SUMOplot[1] is commercial. SUMOsp [17] combines two algorithms originally designed for phosphorylation site prediction (the scoring-based function GPS [19] and an iterative statistical approach MotifX [13]). SUMOpre [16] is based on a probabilistic method that optimizes the entropy of the motif.

An immediate application for *in silico* prediction is to determine the putative SUMOylation sites in the four core histones of *S. cerevisiae*. Nathan *et al.* [10] demonstrated that H2A, H2B, H3 and H4 are frequently SUMOylated. However, Nathan and colleagues were only able to experimentally identify the exact location of a fraction of the expected sites. The SUMOylation sites of the histones do not conform to the consensus motif. SUMOpre and SUMOsp both fail to predict a single SUMOylation site in protein sequences of the four histones. This exemplifies the need for a SUMOylation site predictor that captures dependencies beyond the consensus motif.

It is understood that SUMOylation site recognition by E2 and E3 depends mainly on the amino acid composition in the immediate neighbourhood of the central Lysine. However, it is unclear (1) if there are relevant dependencies be-

[1] `http://www.abgent.com/doc/sumoplot`

tween central residues and surrounding residues not captured by simpler models, (2) if the site's structural presentation influences binding and the computational recognition of it, and (3) if sequence conservation can be used to improve the recognition of functional sites. In this study, we investigate the ability of two machine learning techniques in predicting SUMOylation sites. Support-Vector-Machine (SVM) [15] and Bidirectional Recurrent Neural Networks (BRNN) [2] have both been successful in incorporating a range of dependencies into biological sequence models. To evaluate the contribution of dependencies putatively relevant to SUMOylation, we explore a range of features and functions for presenting our data to these machine learning algorithms.

SVMs use kernels to map samples into a high dimensional feature space to find the best separating decision hyperplane between the two classes (by maximizing the margin between them). In this study we investigate standard vector-based kernels as well as sequence-adapted kernels, including the string P-kernel [7] and the local alignment kernel [12], all acting on a fixed sequence-window around Lysine residues.

In the BRNN, the sequence input is instead fed iteratively into a network of interconnected nodes with feedback connections incorporating a trace of past sequence inputs. A BRNN is thus capable of accounting for sequence information beyond that of a current input (here coming from both a downstream and an upstream direction). The BRNN uses a gradient-based learning algorithm [2], which involves updating network "weights" to minimize the difference between predicted and target values.

We investigate the usefulness of secondary structure (SS) and solvent accessibility (SolvAcc) for SUMOylation site recognition. Unfortunately, experimentally resolved structures are available for only a fraction of known SUMOylated proteins, hence both SS and SolvAcc are obtained from predictors. We use the continuum secondary structure predictor, CSSP (with a reported $Q_3 = 77\%$) [3] and the solvent accessibility predictor, ASAP (with a correlation coefficient of 0.69) [18].

The present paper is organized as follows. First, we give an overview of the SUMOylation sites and analyse their distribution in our dataset. Second, we investigate the abilities of the different machine learning approaches when applied directly on the primary data and then with additional features. In the last section, we compare the best model with previous predictors, SUMOplot, SUMOsp and SUMOpre.

2 Methods

2.1 Dataset

This study uses the dataset of Xu *et al.* [16] only containing proteins with at least one SUMOylation site. Using the same strategy as Xu *et al.* for dividing the data results in 144 proteins used for training and testing, and 14 proteins set aside for final validation.

The 144 proteins contain a total of 241 validated SUMOylation sites, which collectively form the positive class. Roughly 68% of the SUMOylation sites contain the consensus motif. The set of 5,741 Lysines which are not modified by SUMO form the negative class. The 13 proteins in the hold-out set contain 27 sites of which 48% match the consensus. Noteworthy, the resulting dataset is strongly unbalanced and could bias the method to prioritize the larger (negative) class. Steps are taken to investigate any effects of this imbalance.

Redundancy reduction of sequence similarity is not performed. Standard redundancy reduction targets the overall sequence similarity within a dataset and does not reduce the similarity of the relatively short SUMOylation sites.

When a numerical encoding is required (e.g. when using vector-based kernels), each amino acid in the sequence is represented by a one-hot bit vector ("plain") or the position-specific score profile produced by psi-Blast [1] for the protein ("profile"). The "plain" encoding is neutral in that no similarities are incorporated *a priori*. The "profile" encoding reflects the evolutionary divergence between homologous proteins, making available information about sequence conservation. Such "profiles" have found great utility for predicting structural features from sequence. In either case, the full sequence is represented by concatenating the position-specific vectors.

We apply CSSP (using default setting) to predict the secondary structure from primary sequence. The secondary structure is represented by the probability of a residue to adopt each of the three considered classes (helix, sheet, coil). ASAP provides predicted residue-wise relative solvent accessibility (using default settings). The predicted value is normalized to range between zero and one (with one indicating a maximally exposed residue). In either case, each residue-wise prediction is concatenated to the "plain" or "profile" encoding.

2.2 Cross-validation and Evaluation

We evaluate every predictor configuration using 5-fold cross validation, where the dataset is randomly divided into five subsets. All but one of the five are used for training with the remaining one used for testing. This routine is repeated until all five subsets have been used for testing exactly once. In most cases, each evaluation is then repeated five times, with averages and standard deviations reported. To evaluate the performance we compare the predictions with the known positives and report on the correlation coefficient (CC), the sensitivity (SN), specificity (SP), and, the area under the ROC (AUC) (see e.g. [6] for standard definitions). Only the AUC is not influenced by the arbitrary setting of a specific classification threshold and we thus use this as the primary measure. The large number of negatives makes it easy to reach high specificity by simply predicting all but a few certain as negatives. We do revert to CC, SN and SP to discuss specific issues and to compare with previous results.

Finally, trying a large number of configurations and selecting parameter values on basis of test results will impart some selection bias. We therefore report and rely on results for the hold-out set, which has not influenced any predictor settings.

2.3 BRNN

A BRNN first centered on a particular position in a sequence (in our case this is always a Lysine). Then, in an iterated fashion it processes w_n residues on the N-terminal side and w_c residues on the C-terminal side, from both flanks and working towards the centre (in steps of w_n and w_c residues, respectively). The hidden nodes in this network are divided into two "wheels", serving as feedback modules in the N-terminal and C-terminal direction, respectively. Each wheel is equipped with a specified number of nodes, effectively controlling the trace of input from the flanks. The influence decays with the distance to the centered Lysine.

Tuning the internal weights of the BRNN is an iterative process, requiring many passes through the training set. With an independent test set left for the final evaluation, we monitor the performance on the cross-validation test set of each fold and stop training when the performance starts to deteriorate.

The unbalanced dataset could potentially also compromise performance. In addition to the original training set we create a balanced set by sampling positive and negative training data with equal probability. However, during testing all positives and negatives from the test set in the particular fold are evaluated.

2.4 SVM and Kernels

To train the SVM we extract a sequence window covering w_n residues towards the N-terminus of the protein and w_c residues to the C-terminus surrounding every Lysine in the dataset. To account for the imbalance of the dataset, we evaluate the influence of class-specific soft margin parameters, C_+ and C_-, for positives and negatives, respectively.

Apart from window size and C-values, the performance also depends on the choice of the kernel. Here, we evaluate five different kernels, the three standard kernels: linear, radial basis function (RBF) and polynomial kernel, all requiring numerical input ("plain" or "profile" encoding) and two sequence based kernels which operate directly on the sequence data in the window.

Haussler proposed a string kernel known as the string P-kernel that probabilistically evaluates (by convolution) the similarity between sequences by exploring their alignment with all ancestral sequences [7]. Since we are only dealing with fixed-length ($N = w_n+1+w_c$) amino acid sequences without gaps, the string P-kernel is computed as $K_P(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{N} \sum_{\alpha \in A} P(\alpha) P(x_i|\alpha) P(y_i|\alpha)$ where A is the amino acid alphabet and $\mathbf{x}$ and $\mathbf{y}$ are the two amino acid patterns being evaluated. The prior and conditional probabilities of amino acids are taken from the data used to create the BLOSUM62 substitution matrix.

In contrast, the local alignment kernel compares two sequences by exploring *all* their alignments including those with gaps [12]. An alignment between the two sequences is quantified using an amino acid substitution matrix (here BLOSUM62) and a gap penalty setting (we use the default setting). The contribution of non-optimal alignments to the final score is controlled (we use $\beta = 0.1$

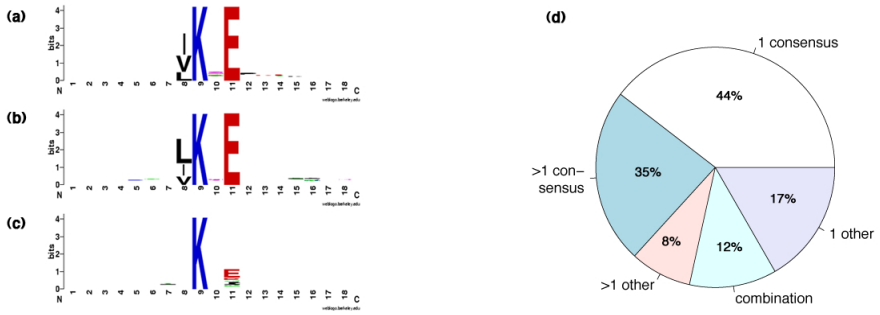


Fig. 2. Comparison between the sequence Logos of SUMOylated and non-SUMOylated sites as well as site distribution in the dataset. Panel **a** shows the sequence Logo created from 165 SUMOylated sites containing the consensus motif (positive class). Panel **b** shows the Logo of 88 non-SUMOylated sites which contain the consensus motif. Panel **c** shows the Logo of the remaining 76 non-consensus sites of the positive class. Panel **d** shows a pie diagram of the SUMOylation distribution in the dataset.

which implies that many local alignments influence the result). All kernels are normalized.

3 Results

3.1 Dataset Analysis

This section illustrates the discrepancy between the dominant consensus motif and alternative SUMOylation sites.

165 out of the 241 sites in the training set have the consensus motif of ΨKxE. The motif seems to be direction dependent, reading in direction of the C-terminus. However, there are four validated SUMOylation sites which show the reverse motif. As shown in Tab. 4 a simple regular expression parser for the consensus motif can achieve a CC of 0.68 – exceeding the 0.64 reported for SUMOpre – by identifying the 165 SUMOylated sites containing the consensus motif and missing 76. However, it wrongly predicts 88 sites to be SUMOylated. It should be noted that on a proteomic scale the dataset contains an unrealistically high proportion of SUMOylation sites so the estimates are optimistic.

The difficulty of discriminating between SUMOylated and non-SUMOylated sites on basis of the consensus is illustrated in Fig. 2**a**-**b** using sequence Logos of both positives and negatives that match the motif [4]. A Logo of the known SUMOylation sites not matching the consensus motif is shown in Fig. 2**c**. The central Lysine is still predominantly flanked by Glutamic acid (E) on the C-terminal side, however the N-terminal hydrophobic residue is missing.

Fig. 2**d** shows the distribution of consensus vs non-consensus SUMOylation sites in the dataset of 144 proteins. 56% of the proteins have a single SUMOylation site only of which two thirds are consensus sites. A similar ratio can be

Dataset	*hidden nodes*	*AUC (sd)*
unbalanced	2	0.923 (0.006)
unbalanced	10	0.919 (0.004)
unbalanced	20	0.914 (0.007)
balanced	2	0.895 (0.012)
balanced	10	0.906 (0.007)
balanced	20	0.906 (0.010)

Table 1. Overview of the performance of examined BRNN settings. Average area under the ROC (AUC) of different benchmark settings for BRNN (five times repeated).

observed for proteins, which contain more than one SUMOylation site. Only 12 proteins contain consensus as well as non-consensus sites. This indicates that there is no cascade effect, where the "strong" consensus site is SUMOylated first and then aids in the SUMOylation of "weaker" non-consensus sites.

3.2 Performance of BRNN

The optimal parameter setting of the BRNN was determined empirically. The window size of $w_n = 1$ and $w_c = 3$ has the highest AUC. Smaller windows give worse accuracy while larger windows do not bring any improvements. Tab. 1 summarizes the performance of several settings of hidden nodes, and on balanced and unbalanced presentation of data.

The performance is rather even across all settings. The BRNN performs slightly better when trained on the unmodified, unbalanced dataset. Increasing the number of hidden nodes appears to only decrease accuracy – suggesting that the site is simple to represent. The simplest topology with one hidden node in each wheel, trained without compensating for the class imbalance provides the best result with an average AUC of 0.93 (henceforth referred to as BRNN^{Best}). The BRNN^{Best} model contains 125 parameters to be optimized during training on approx. 4,800 samples. We do not observe a trend to overfit, which indicates a sufficient amount of training samples.

3.3 Performance of SVMs

In this section the performance of several SVM-settings are evaluated. Kernel, C-values and window size are problem specific and thus determined empirically. Fig. 3 exemplifies the influence of the choice of window size, as well as C-values for the linear, RBF and string P-kernel respectively. The optimal window sizes agree with the information content visualized in the Logos (Fig. 2): while there seems to be some conservation towards the C-terminus the performance drops when more than three residues are included towards the N-terminus. The best C-values for the linear kernel put equal weight for the negative and positive classes. For the RBF and the string P-kernel there seems to be specific C-value pairs, which perform better than others.

Tab. 2 summarizes the performance of the best setting for each kernel in terms of C-values, window sizes and kernel specific parameters.

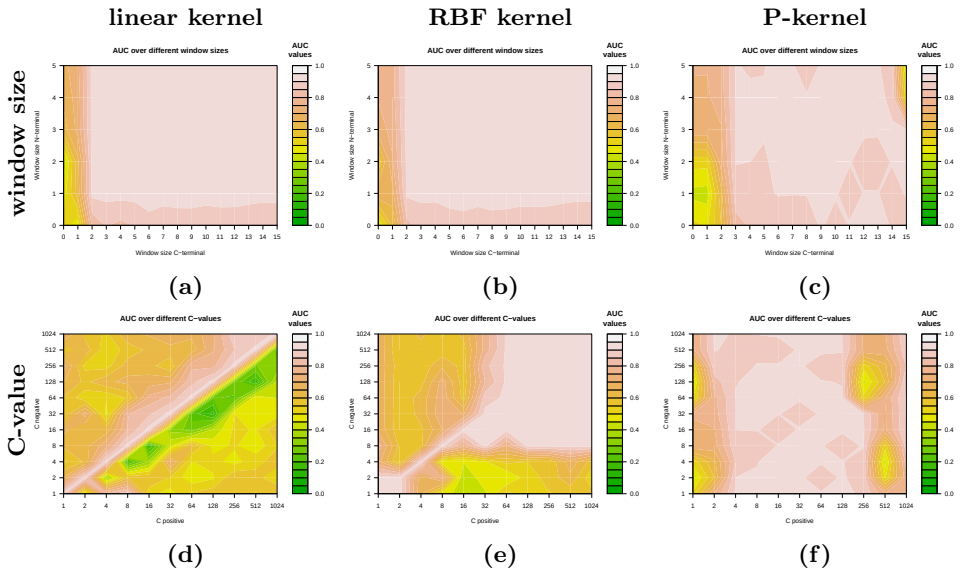


Fig. 3. Performance of different SVM settings. Each panel exemplifies the AUC on the test set for different configurations with varying window sizes (upper panels) and C-values (lower panels) for the linear, RBF and P-kernel respectively.

Once the optimal parameter setting is determined, all kernels seem to be able to recognize SUMOylation sites quite accurately, since the average AUC is around 0.92 (with no statistical significant difference using t-test p-value < 0.05). The RBF and string P-kernel achieve the highest average AUC (both at 0.923) and have the same predictive power as the BRNN, albeit with a smaller standard deviation.

We choose the SVM with RBF kernel ($w_c = 6, w_n = 4, C_+ = 2, C_- = 1$) as our final predictor. Though not statistical significantly better its performance is more robust than the BRNN approach. Compared to the string P-kernel the RBF-kernel is much faster to train and test. We refer to the SVM-RBF kernel as SUMOsvm.

3.4 Assessing Enhanced Input Data and Multi-SVM Architecture

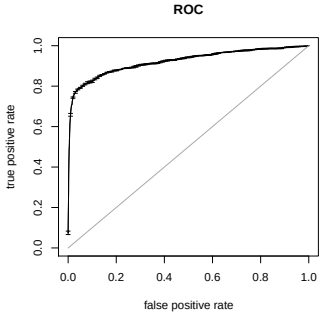
In this section we evaluate the impact of incorporating structural features and evolutionary information into the predictor, as well as combining several kernels into one "committee"-like SVM.

The results from the extended input features are summarized in Tab. 3. We observe no performance increase when incorporating secondary structure or solvent accessibility. The small increase using psi-Blast profiles is not statistically significant.

A multi-SVM committee yields no observable performance increase. An improvement in performance due to a committee-style prediction is expected only

ML method	*AUC* (*sd*)	*parameter settings*				
		w_c	w_n	C_+	C_-	*method specific*
RBF kernel	0.923 (0.001)	6	4	2	1	$\sigma = 0.014$
String P-kernel	0.923 (0.004)	6	4	8	1	$\gamma = 0.1$, BLOSUM62
BRNN^{Best}	0.923 (0.006)	3	1			hidden nodes=2
Linear kernel	0.920 (0.004)	6	1	2	2	
Polynomial kernel	0.920 (0.004)	12	1	4	1	$order = 3$
Local alignment kernel	0.913 (0.002)	3	2	1	1	$\beta = 0.1$, BLOSUM62

Table 2. Overview of the performance of the examined machine learning methods. The methods are ordered according to the average AUC, achieved by the different kernels and BRNN^{Best}. Each SVM and BRNN is represented by its best performing parameter setting regarding test error, 5-fold CV, five times repeated.



encoding	*AUC*
Plain	0.92
Profile	0.93
SS	0.92
SolvAcc	0.92
SS+SolvAcc	0.92

(a) ROC of SUMOsvm **(b) Enriched encoding**

Table 3. Influence of evolutionary and structural features on the performance of SUMOsvm. Panel **a**: average ROC for SUMOsvm using plain encoding (five times repeated). Panel **b**: Performance of SUMOsvm with additional features input. The structural features are secondary structure (SS), solvent accessibility (SolvAcc) or both. Evolutionary features are psi-Blast profiles.

when the kernels deliver qualitatively different predictions. This is not the case here as we observe at least 90% of the false predictions are shared amongst the majority of all kernels.

3.5 Comparison and Discussion

In this section we compare SUMOsvm with the previously reported SUMOylation site predictors. In Tab. 4, we show the testing error measured on the 144 proteins during cross-validation and the prediction error on the 14 proteins in the hold-out set. To obtain the hold-out error, we perform a voted prediction of the SVMs trained during the 5-fold cross-validation. The performance measures from the other methods are obtained from the original publications.

The comparison with other methods for predicting SUMOylation sites is complicated by the use of different validation methods. For SUMOpre, only three

Method	*Validation*	*AUC*	*CC*	*SN*	*SP*	*AC*
SUMOsvm	CV	0.92	0.67	0.62	0.99	0.97
	hold-out	-	0.56	0.44	0.99	0.96
RegularExp	training	NA	0.68	0.69	0.99	0.98
	hold-out	-	0.54	0.48	0.99	0.97
SUMOpre	CV*	0.87	0.64	0.74	0.98	0.97
	hold-out*	-	0.66	0.54	1.0	0.97
SUMOsp	CV	0.73	0.26	0.83	0.93	0.93
	hold-out	-	0.37	0.61	0.93	0.91
SUMOplot	training	NA	0.48	0.80	0.93	0.90
	hold-out	-	0.35	0.57	0.93	0.91

Table 4. Performance overview of the existing predictors and SUMOsvm. The values for the area under the ROC (AUC), correlation coefficient (CC), sensitivity (SN), specificity (SP) and accuracy (AC) are obtained from the original publications of SUMOpre and SUMOsp. The threshold chosen for SUMOsp was 18. *Though reported by Xu *et al.* as CV and hold-out error, the values are understand to be training error because "self-consistency test was used as the testing strategy" [16].

different test protocols are used: self-consistency (where "the SUMOylation state for each motif in the entire dataset is predicted by the rules derived from the same dataset" [16]), K-fold cross-validation and Leave-one-out cross-validation (which is identical to K-fold CV when K equals the size of the dataset minus one). The hold-out set is inspected only in the context of these protocols (all of which involve training on this set).

The AUC is not explicitly reported for SUMOpre, but here estimated from their ROC curve. Sensitivity and specificity are altered by simply changing the classification threshold. The threshold setting similarly affects the correlation between observed and predicted sites. We thus assume that all reported results are achieved when the threshold is the best possible.

SUMOsvm is not significantly better than the previously published methods, which in turn are not more powerful than a simple regular expression scan with `[LVI]K.E`. Neither the motif-flanking residues nor structural features appear to aid prediction. This begs the question how non-consensus sites are processed by SUMO.

One hypothesis is that sites are SUMOylated by different means (corresponding to different SUMOylation pathways). We would then expect that SUMOylation sites of proteins group in accordance with shared means. To identify such groups, we performed a kernel hierarchical cluster analysis, where the distances in the feature space (as seen by the RBF kernel) are used to generate a distance map between the different sites. The resulting map of the SUMOylation sites in the hold-out set is shown in Fig. 3.5. The correctly predicted sites (all conform to the consensus motif) are clustered and form the largest entity. There is only one other cluster formed containing a putative KxK motif in the hold-out set.

To investigate if SUMOylation binding is species or compartment dependent, we extracted all proteins in the dataset that belong to human and are localized

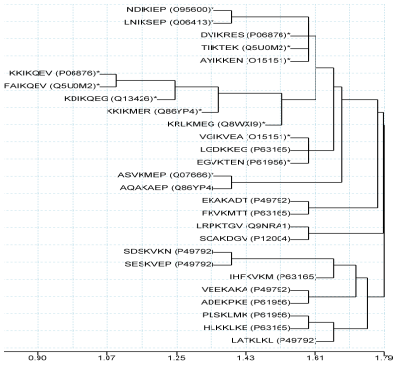


Fig. 4. Hierarchical clustering of the SUMOylation sites in the hold-out set. We use the RBF-kernel with $w_c = w_n = 3$ to obtain the hierarchical clustering plot. Sites SUMOsvm predicts correctly are marked with *.

to the nucleus. If the SUMOylation pathway is species and/or compartment dependent, one would expect to see a correlation of either with sequence motif. However, a similar fraction of the consensus motif appears amongst human nuclear proteins as in the original set, and no alternative motifs were obvious when Logos were used from this smaller group of binding sites. Also, no performance gain could be observed when retraining on this subset.

4 Conclusion

We developed a SUMOylation site predictor, SUMOsvm, based on support vector classification and the RBF kernel. Several other configurations performed equally well including models based on alternative kernels and the bidirectional recurrent neural network. However, in the comparison to previously published SUMOylation site predictors we found that neither SUMOsvm nor the previously published methods are significantly better than a simple regular expression scanner.

The disappointing result is particularly noteworthy because we presented SUMOsvm with sequence data which were enriched with predicted structural features (secondary structure and relative solvent accessibility) and evolutionary information (psi-Blast profiles).

No predictor to date is able to identify the SUMOylation sites in the four core histones of yeast–a group of proteins which are known to be regulated by SUMO but for which we still have only partial understanding of actual sites involved. All predictors tend to rely on the consensus motif that describe a majority of known SUMOylated sites but do not include the sites on the histones. Until more of the SUMOylation pathway is uncovered, SUMOylation site prediction from the current paucity of data remains challenging.

Acknowledgment

The authors would like to thank Jialin Xu and colleagues for providing us with the SUMOylation dataset.

References

1. S. F. Altschul, T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped bLAST and pSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.
2. P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15:937–946, 1999.
3. M. Bodén, Z. Yuan, and T. L. Bailey. Prediction of protein continuum secondary structure with probabilistic models based on NMR solved structures. *BMC Bioinformatics*, 7:68, 2006.
4. G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Res*, 14(6):1188–1190, Jun 2004.
5. V. Dorval and P. E. Fraser. SUMO on the road to neurodegeneration. *Biochim Biophys Acta*, 1773(6):694–706, Jun 2007.
6. T. Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine Learning*, 2004.
7. D. Haussler. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, Department of Computer Science, University of California, Santa Cruz, CA 95064, 1999.
8. R. T. Hay. SUMO: a history of modification. *Mol Cell*, 18(1):1–12, Apr 2005.
9. P. Heun. SUMOrganization of the nucleus. *Curr Opin Cell Biol*, 19(3):350–355, Jun 2007.
10. D. Nathan, K. Ingvarsdottir, D. E. Sterner, G. R. Bylebyl, M. Dokmanovic, J. A. Dorsey, K. A. Whelan, M. Krsmanovic, W. S. Lane, P. B. Meluh, E. S. Johnson, and S. L. Berger. Histone sumoylation is a negative regulator in saccharomyces cerevisiae and shows dynamic interplay with positive-acting histone modifications. *Genes Dev*, 20(8):966–976, Apr 2006.
11. M. S. Rodriguez, C. Dargemont, and R. T. Hay. SUMO-1 conjugation in vivo requires both a consensus modification motif and nuclear targeting. *J Biol Chem*, 276(16):12654–12659, Apr 2001.
12. H. Saigo, J. Vert, N. Ueda, and T. Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689, Jul 2004.
13. D. Schwartz and S. P. Gygi. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol*, 23(11):1391–1398, Nov 2005.
14. T. H. Shen, H.-K. Lin, P. P. Scaglioni, T. M. Yung, and P. P. Pandolfi. The mechanisms of PML-nuclear body formation. *Mol Cell*, 24(3):331–339, Nov 2006.
15. V. Vapnik. *Statistical learning theory*. Wiley, 1998.
16. J. Xu, Y. He, B. Qiang, J. Yuan, X. Peng, and X. Pan. A novel method for high accuracy sumoylation site prediction from protein sequences. *BMC Bioinformatics*, 9:8, 2008.
17. Y. Xue, F. Zhou, C. Fu, Y. Xu, and X. Yao. SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res*, 34(Web Server issue):W254–W257, Jul 2006.
18. Z. Yuan and B. Huang. Prediction of protein accessible surface areas by support vector regression. *Proteins*, 57(3):558–564, Nov 2004.
19. F. Zhou, Y. Xue, G. Chen, and X. Yao. GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem Biophys Res Commun*, 325(4):1443–1448, Dec 2004.