Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 358 (2007) 136–139
BBRC
www.elsevier.com/locate/ybbrc

Predicting the protein SUMO modification sites based on Properties Sequential Forward Selection (PSFS)

Boshu Liu [b,c,1], Sujun Li [b,1], Yinglin Wang [c], Lin Lu [b], Yixue Li [b,*], Yudong Cai [a,*]

[a] CAS-MPG Partner Institute for Computational Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, China
[b] Bioinformatics Center, Key Lab of Systems Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, China
[c] School of Software, Shanghai Jiao Tong University, China

Received 5 April 2007
Available online 23 April 2007

Abstract

Protein SUMO modification is an important post-translational modification and the optimization of prediction methods remains a challenge. Here, by using Support Vector Machines algorithm (SVM), a novel computational method was developed for SUMO modification site prediction based on Sequential Forward Selection (SFS) of hundreds of amino acid properties, which are collected by Amino Acid Index database (http://www.genome.jp/aaindex). Our method also compares with the 0/1 system, in which the 20 amino acids are represented by 20-dimensional vectors ($A = 00000000000000000001$, $C = 00000000000000000010$ and so on). The overall accuracy of leave-one-out cross-validation for our method reaches 89.18%, which is higher than 0/1 system. It indicated that the SUMO modification prediction process is highly related to the amino acid property and this approach here provide a helpful tool for further investigation of the SUMO modification and identification of sumoylation sites in proteins. The software is available at http://www.biosino.org/sumo.

Keywords: Prediction; SUMO; SVM; AAindex; Sequential Forward Selection (SFS); Sumoylation

Post-translational modification is a crucial biological process presented in cell signaling, death or localization. And many kinds of modification occur in the cell process. Ubiquitin and ubiquitin-like modifiers represent a very special example as the modifier is a small polypeptide, and is usually attached to lysine side chains of the protein [1]. Small ubiquitin-related modifier (SUMO) has been shown to present in all eukaryotes and be covalently conjugated to many kinds of cellular proteins [1–3]. Sumoylation participates in many important cellular processes such as transcriptional regulation [4], transcription factor activity [5], signal transduction [6] and so on. Many sumoylation modification occurs at a consensus motif ΨKxE (where Ψ represents L, I, V or F) [3,7,8], However, previous work proved that about 23% real sumoylation sites do not match this consensus motif [9]. The SUMO site specificity recognition still remains a challenge.

A primary limitation in conducting the site recognition mainly concerns feature selection. In previous prediction studies, the 0/1 system was mainly choose to represent amino acids or nuclear acid [10,11] without estimation of amino acid properties or selected a subset of properties manually. There are many kinds of properties, like molecular weight and isoelectric point, to represent the amino acid numerically besides the 0/1 systems. Actually, Amino Acid Index Database [12,13] (AAindex, http://www.genome.jp/aaindex) collects hundreds of published amino acid indices representing the different physicochemical and biological properties of amino acids. Protein sumoylation is still not fully understood by researchers. In particular, it

* Corresponding authors.
E-mail addresses: yxli@sibs.ac.cn (Y. Li), cyd@picb.ac.cn (Y. Cai).
[1] These authors contributed equally to this work.

0006-291X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2007.04.097

is not known which properties are important for this process. Through comprehensive analysis on these hundreds of amino acid properties, more information about the sumoylation process could be acquired and also avoid artificial matter. Here, we present a prediction work, adopting a Sequential Forward Selection (SFS) method based on hundreds of amino acid properties to investigate those properties that are important to sumoylation in the prediction and moreover, develop a prediction system based on a SVM algorithm and feature selection to help find the protein sumoylation sites. The software is available at http://www.biosino.org/sumo and the source code is also available upon request.

Materials and methods

Data preparation. The positive dataset comes from SUMOsp [9]. Twelve-mer peptides data (six residues upstream and six residues downstream of the sumoylation sites) were extracted from this dataset. Twelve-mer peptides were used in building models because this length is the most optimal in our testing process. The redundant peptides were filter manually from this original dataset and a non-redundant dataset with 227 peptides was finally used. The negative dataset, which has 226 peptides, was randomly selected from the non-sumoylation lysine peptides in the sumoylated proteins (see Supplementary material 1).

Representation of the amino acid in the dataset. For a predictor, it is recommended to represent the dataset in the form of numbers. Two encoding systems were then selected to represent the dataset:

Amino acid property in the form of the numeric matrix exists in AAindex [12,13]. For example, amino acid could be represented in property 'Transfer free energy to surface' [14], in which $A = -0.2$, $L = -2.46$ and so on. In this work, AAindex (more details could be found on its web http://www.genome.jp/aaindex.) release 8.0 was used, in which there are total 516 indices. As there are some missing values exist in the indices description, 492 indices without missing values were used to encode the dataset.

0/1 system: the 20 amino acids are represented by 20-dimensional vectors ($A = 00000000000000000001$, $C = 00000000000000000010$ and so on).

Support Vector Machine (SVM). Recently, SVM (Support Vector Machine) is widely applied in the field of biological sciences [10,11,15–17]. It is a machine learning technique based on statistical theory. The basic idea of SVM algorithm is mapping the input vectors into a higher dimension and then constructs a hyper-plane to separate these vectors into different classes with the maximal margin and the least error. More details about SVM could be found in Vapnik's [18–20] and other publications [21–23]. In this paper, the SVMLight [22] was used in this work.

Property Sequential Forward Search (PSFS) procedure based on SVM. As similar to the previous work [9], we also think that the site sumoylation process depends on the upstream and downstream amino acid physical and chemistry properties. There are many kinds of amino acid properties collected by AAindex [12,13]. And these properties number will become lager and larger. There is always redundancy information in the great number of properties, the hypothesis could be proposed that a small number of these properties can be combined to form an effective classifier. As there are 2^{492} kinds of combinations in these properties, we cannot test this combination exhaustively. Hence, consideration in practice has to be given to implement searching procedure for reducing the number of properties. In this work, we consider a Sequential Forward Selection (SFS) procedure according to the rule of accuracy acquired by leave-one-out cross-validation based on SVM algorithm. This method mainly performed in three steps and is illustrated in the following Pseudocode:

```
1. (Initialization)
   (1) P_k = {∅};
       k = 0;
   (2) J_0 = 0;
   (3) S_M = {X_1, X_2 ... X_M}
       M is the number of total properties;
       S is the original feature set.

2. (Forward Inclusion)
   (1) For each x left in the S_{M-k} select the x with the greatest J
       according to accuracy evaluated by SVM;
   (2) J_{k-1} = J
       if J_{k-1} > J_k, then
       P_{k+1} = P_k + X;
       S_{M-k-1} = S_{M-k} - X;
       k = k + 1;
       if k <= M, then
              goto 2. (1);

3. (Output)
       P_c = {P_1, P_2 ... P_c}
```

P is the selected feature set; k is the number of selected feature; J is accuracy acquired by leave-one-out cross-validation based on SVM algorithm; c is final the number of selected features; p is the selected feature.

First of all, each of the properties in the AAindex database was used to encode the dataset. Thus, the dataset could be transformed to 492 different data matrix. The performance of these data matrix were evaluated by accuracy acquired by leave-one-out cross-validation based on SVM. The best property, which has the highest performance, was selected into the selected property pool directly. The remaining properties will be used to co-encoding the dataset with the selected property pool one-by-one and the performance will be evaluated again. If the performance is the highest and raised or equal to the original selected property pool, this property will be added to the pool. Otherwise, this property will be regarded as the useless property to the current property pool. This process will be repeated until the accuracy begins to descend. This method demands a very aggressive process which would discard the vast majority of features and fewer properties will be remaining in the model. In summary, Property Sequential Forward Selection (PSFS) method is starting from the empty properties set, then sequentially add the property that results in the highest accuracy when combined with the properties that have already been selected. This PSFS procedure used the greedy selection strategy under the property monotonic assumption. In the final model building process, the properties in the property pool will be selected to build the effective classifier.

Leave-one-out cross-validation. Cross-validation is a method for error rate estimation. The leave-one-out cross-validation is regarded as the most objective evaluation method and hence adopted by more and more researchers. In our experiments, leave-one-out cross-validation was used to evaluate the performance of the classifiers.

Performance measurement. After prediction, the dataset contains the four parts: true positives (TP), false positives (FP), false negatives (FN), true negatives (TN). The accuracy (AC) could be formed as following: $AC = (TP + TN)/(TP + FP + FN + TN)$. In this work, we will use the accuracy (AC) to estimate the performance of the classifiers.

Results and discussion

According to the prediction procedure, 492 classifiers based on 492 indices in the AAindex were built in the first step. Leave-one-out cross-validation was adopted to evaluate these classifiers. The overall result was shown in Fig. 1 (detailed information see Supplementary material 2). Eighty-three classifiers based on different properties could

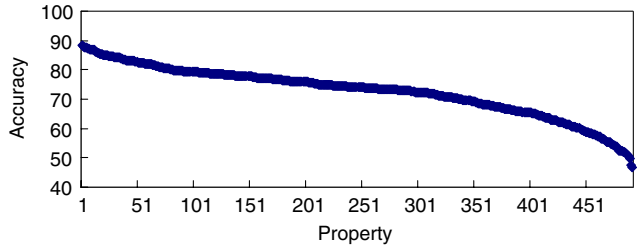

Fig. 1. The overall accuracy acquired by the total 492 classifiers. Eighty-three classifiers based on different properties reached 80% accuracy. Fifty-two classifiers based on different properties could not reach 60% accuracy. The highest accuracy is 88.3%.

reach 80% accuracy. Fifty-two classifiers based on different properties even could not reach 60% accuracy. The first seven properties which have the highest accuracy were listed in Table 1.The highest accuracy is 88.3%, which is the result of 'Retention coefficient in HPLC' property. The second highest accuracy is 'Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water'. The remaining five properties are mainly 'alpha-helix-, beta-strand'-related ones. These properties extracted by this process are highly similar to each other. From this result, we could infer that the hydrophobic-related property and the second structure of the peptides could contribute greatly to the SUMO prediction. However, some information about the properties interaction is still missed in this process.

After applied PSFS method illustrated in the pseudo-code applied, seven properties remain in the property pool. These properties are listed in the Table 2. The dataset were encoded by these seven properties at the same time and a model based on SVM was built. The performance of this model was also estimated by leave-one-out cross-validation and the accuracy was listed in the Table 3. This result was also compared with the 0/1 system. It is clear that our method's performance is higher than 0/1 system. One possible explanation could be that 0/1 system only represents the sequence information and ignore the importance in the different amino acid properties. As another comparison, the first seven properties which get the highest performance were also used to encode the dataset at the same time. Even the best seven properties were used at the same time; the accuracy could not be reached higher than our method. The explanation could be that the first seven properties are highly similar to each other but the properties selected by our method behave independently. They could be complementary to each other in the classify process and avoid missing information. Except the 'Retention coefficient in HPLC' and second structure properties like 'helix, sheet, and strand', important properties like 'flexibility parameters' and the 'Composition of amino acids in nuclear proteins' were also extracted. The former one could be explained as the post-translational modification is most likely to occur in flexible region in the protein. It has been proposed that nuclear localization signal (NLS) [8] in the protein confer the possibility to be sumoylated and this could be explained for the latter property. Through this comprehensive analysis process, important properties were selected and its performance could not be reached by 0/1 system and simple combination. Also this kind of process avoids missing information caused by artificial feature selection.

This result indicates that the sumoylation prediction is closely correlated with the amino acid property around the sumoylation site. And computational method developed in this work could be a powerful tool to investigate

Table 1
The first seven properties which get the highest performance

Id in AAindex	Description in AAindex	Accuracy (%)
MEEJ800101	Retention coefficient in HPLC, pH 7.4	88.3
WIMW960101	Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water	88.08
NAGK730101	Normalized frequency of α-helix	87.64
AURR980106	Normalized positional residue frequency at helix termini N1	87.64
AURR980101	Normalized positional residue frequency at helix termini N4′	87.64
NAGK730102	Normalized frequency of β-structure	87.2
GEIM800107	β-Strand indices for α/β-proteins	87.2

The accuracy that they reached was listed in the third column.

Table 2
Seven properties in the final property pool after applied PSFS method

Id in AAindex	Name in AAindex
QIAN880114	Weights for β-sheet at the window position of −6
MEEJ800101	Retention coefficient in HPLC, pH 7.4
VINM940104	Normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbours
GEOR030104	Linker property from 3-linker dataset
QIAN880102	Weights for α-helix at the window position of −5
CEDJ970105	Composition of amino acids in nuclear proteins (percent)
RACS820105	Average relative fractional occurrence in E0(i)

Table 3
The accuracy of the four prediction models

	0/1 system	The best property of the 492 classifiers	Combination of the first seven properties which have the best performance of 492 classifiers	PSFS method
AC	87.64%	88.3%	87.64%	89.18%

The accuracy evaluated by leave-one-out cross-validation of the four prediction models.

sumoylation process preference systematically. This software is available at http://www.biosino.org/sumo.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2007.04.097.

References

[1] S. Muller, C. Hoege, G. Pyrowolakis, S. Jentsch, SUMO, ubiquitin's mysterious cousin, Nat. Rev. Mol. Cell Biol. 2 (2001) 202–210.
[2] R.T. Hay, Protein modification by SUMO, Trends Biochem. Sci. 26 (2001) 332–333.
[3] E.S. Johnson, Protein modification by SUMO, Annu. Rev. Biochem. 73 (2004) 355–382.
[4] D.W. Girdwood, M.H. Tatham, R.T. Hay, SUMO and transcriptional regulation, Semin. Cell Dev. Biol. 15 (2004) 201–210.
[5] G. Gill, Post-translational modification by the small ubiquitin-related modifier SUMO has big effects on transcription factor activity, Curr. Opin. Genet. Dev. 13 (2003) 108–113.
[6] M. Liang, F. Melchior, X.H. Feng, X. Lin, Regulation of Smad4 sumoylation and transforming growth factor-beta signaling by protein inhibitor of activated STAT1, J. Biol. Chem. 279 (2004) 22857–22865.
[7] R.T. Hay, SUMO: a history of modification, Mol. Cell 18 (2005) 1–12.
[8] F. Melchior, M. Schergaut, A. Pichler, SUMO: ligases, isopeptidases and nuclear pores, Trends Biochem. Sci. 28 (2003) 612–618.
[9] Y. Xue, F. Zhou, C. Fu, Y. Xu, X. Yao, SUMOsp: a web server for sumoylation site prediction, Nucleic Acids Res. 34 (2006) W254–W257.
[10] Z. Qian, Y.D. Cai, Y. Li, A novel computational method to predict transcription factor DNA binding preference, Biochem. Biophys. Res. Commun. 348 (2006) 1034–1037.
[11] S. Li, B. Liu, R. Zeng, Y. Cai, Y. Li, Predicting *O*-glycosylation sites in mammalian proteins by using SVMs, Comput. Biol. Chem. 30 (2006) 203–208.
[12] S. Kawashima, M. Kanehisa, AAindex: amino acid index database, Nucleic Acids Res. 28 (2000) 374.
[13] S. Kawashima, H. Ogata, M. Kanehisa, AAindex: Amino Acid Index Database, Nucleic Acids Res. 27 (1999) 368–369.
[14] H.B. Bull, K. Breese, Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues, Arch. Biochem. Biophys. 161 (1974) 665–670.
[15] Y.D. Cai, P.W. Ricardo, C.H. Jen, K.C. Chou, Application of SVM to predict membrane protein types, J. Theor. Biol. 226 (2004) 373–376.
[16] P. Jia, T. Shi, Y. Cai, Y. Li, Demonstration of two novel methods for predicting functional siRNA efficiency, BMC Bioinformatics 7 (2006) 271.
[17] G.Q. Zhang, Z.W. Cao, Q.M. Luo, Y.D. Cai, Y.X. Li, Operon prediction based on SVM, Comput. Biol. Chem. 30 (2006) 233–240.
[18] E.B. Bernhard, M.G. Isabelle, N.V. Vladimir, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, ACM Press, Pittsburgh, Pennsylvania, United States, 1992.
[19] C. Corinna, V. Vladimir, Support-Vector Networks, Kluwer Academic Publishers, 1995, pp. 273–297.
[20] N.V. Vladimir, The Nature of Statistical Learning Theory, Springer-Verlag, New York, Inc., 1995.
[21] C. Kai-Min, K. Wei-Chun, S. Chia-Liang, W. Li-Lun, L. Chih-Jen, Radius Margin Bounds for Support Vector Machines with the RBF Kernel, MIT Press, 2003, pp. 2643–2681.
[22] J. Thorsten, Making large-scale support vector machine learning practical, in: Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, CA, USA, 1999, pp. 169–184.
[23] N. Zavaljevski, F.J. Stevens, J. Reifman, Support Vector Machines with Selective Kernel Scaling for Protein Classification and Identification of Key Amino Acid Positions, Bioinformatics 18 (2002) 689–696.