Mol Divers (2010) 14:81–86
DOI 10.1007/s11030-009-9149-5

FULL-LENGTH PAPER

Protein sumoylation sites prediction based on two-stage feature selection

Lin Lu · Xiao-He Shi · Su-Jun Li · Zhi-Qun Xie · Yong-Li Feng · Wen-Cong Lu · Yi-Xue Li · Haipeng Li · Yu-Dong Cai

Received: 30 December 2008 / Accepted: 21 April 2009 / Published online: 27 May 2009

Abstract Protein sumoylation is one of the most important post-translational modifications. Accurate prediction of sumoylation sites is very useful for the analysis of proteome. Though the putative motif $\Psi K X E$ can be used, optimization of prediction models still remains a challenge. In this study, we developed a prediction system based on feature selection strategy. A total of 1,272 peptides with 14 residues from SUMOsp (Xue et al. [8] Nucleic Acids Res 34:W254–W257, 2006) were investigated in this study, including 212 substrates and 1,060 non-substrates. Among the substrates, only 162 substrates comply to the motif $\Psi K X E$. First, 1,272 substrates were divided into training set and test set. All the substrates were encoded into feature vectors by hundreds of amino acid properties collected by Amino Acid Index Database (AAIndex, http://www.genome.jp/aaindex). Then, mRMR (minimum redundancy–maximum relevance) method was applied to extract the most informative features. Finally, Nearest Neighbor Algorithm (NNA) was used to produce the prediction models. Tested by Leave-one-out (LOO) cross-validation, the optimal prediction model reaches the accuracy of 84.4% for the training set and 76.4% for the test set. Especially, 180 substrates were correctly predicted, which was 18 more than using the motif $\Psi K X E$. The final selected features indicate that amino acid residues with two-residue downstream and one-residue upstream of the sumoylation sites play the most important role in determining the occurrence of sumoylation. Based on the feature selection strategy, our prediction system can not only be used for high throughput prediction of sumoylation sites but also as a tool to investigate the mechanism of sumoylation.

Keywords Prediction · Protein sumoylation · mRMR · AAIndex · Nearest Neighbor Algorithm · Leave-one-out cross-validation · Bioinformatics

Electronic supplementary material The online version of this article (doi:10.1007/s11030-009-9149-5) contains supplementary material, which is available to authorized users.

Lin Lu and Xiao-He Shi contributed equally to this study.

Y.-D. Cai
Institute of System Biology, Shanghai University, 99 Shang-Da Road, 200244 Shanghai, China

L. Lu
Department of Biomedical Engineering, Shanghai Jiao Tong University, 200240 Shanghai, China

Z.-Q. Xie · H. Li (✉) · Y.-D. Cai (✉)
CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 200031 Shanghai, China
e-mail: lihaipeng@picb.ac.cn; cyd@picb.ac.cn

Y.-X. Li
Life Science and Technology, School of Shanghai Jiao Tong University, 200240 Shanghai, China

X.-H. Shi
Institute of Health Science, Shanghai Institute for Biological Science, Chinese Academy of Science, 225 South ChongQing Road, 200025 Shanghai, China

Y.-L. Feng · W.-C. Lu
Department of Chemistry, College of Sciences, 99 Shang-Da Road, 200444 Shanghai, China

S.-J. Li · Y.-X. Li (✉)
Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 200031 Shanghai, China
e-mail:yxli@sibs.ac.cn

Introduction

Protein post-translational modifications are very crucial to the analysis of an organism's proteome because the diversity and functional breadth of the organism's proteome will be greatly expanded by a variety of post-translational modifications [1]. Small ubiquitin-related modifier (SUMO) family proteins can covalently attach to other proteins as an important type of post-translational modifications called sumoylation [2]. SUMO may modify proteins that participate in crucial biological processes, for example, transcriptional regulation [3] and signal transduction [4]. Protein sumoylation has also been reported to have great relationship with many diseases, such as type-1 diabetes [5] and Parkinson's disease [6].

SUMO is usually linked to substrates containing a consensus motif ΨKXE (Ψ is a hydrophobic amino acid) [7]. However, more and more experiments have shown that such motif is not the only determinant to the substrates recognition [8,9]. In order to explore a more universal motif, which can promote our understanding of sumoylation mechanism and guide our biological experiment designs, a prediction system called SSPFS (Sumoylation Site Prediction base on Features Selection) was developed in our study. The dataset we used comes from SUMOsp [8], in which only 162 substrates complied to the motif ΨKXE.

Our SSPFS prediction system is based on feature selection strategy. Peptides were first transformed into feature vectors according to hundreds of different physiochemical and biological amino acid properties collected in Amino Acid Index Database [10,11]. The mRMR (minimum redundancy–maximum relevance) method [12] was used to extract formative features from those feature vectors. Nearest Neighbor Algorithm (NNA) [13] was used to produce prediction models, including candidate prediction models during the process of feature selection and the final optimal prediction model. Tested by LOO cross-validation, the optimal prediction model reaches the accuracy of 84.4% for the training set and 76.4% for the test set. Especially, 180 substrates were correctly predicted, which was 18 more than using the motif ΨKXE. Analysis of optimal features indicates that amino acid residues with two-residue downstream and one-residue upstream of the sumoylation sites play the most important role in determining the occurrence of sumoylation. Based on feature selection strategy, our prediction system can not only be used for high throughput prediction of sumoylation sites but also as a tool to investigate the mechanism of sumoylation. The prediction software is upon request.

Materials and methods

Data preparation

The protein sequences we used for training comes from SUMOsp [8]. Peptides containing Lysine (K) were extracted as our training samples [7]. Considering the computational complexity, our sample peptides consisted of 14 residues with seven residues upstream and seven residues downstream of the Lysine (K). The real sumoylation peptides were assigned as positive samples, while the left ones were assigned as the negative. After removing redundancy, we attained 212 positive samples and 5,435 negative samples. Among the 212 positive samples there were 162 substrates complying to the motif ΨKXE. As the robustness of the prediction model would be greatly weakened by the quantitative unbalance between the positive and negative samples (approximately 1:25), we only randomly chose $212 * 5$ negative samples for our study. Then, all the 1,272 samples were divided into training set (see the supplementary material I) and test set (see the supplementary material II). The training set contained 191 positive samples and 954 negative samples, while the test set contained 21 positive samples and 106 negative samples. The ratio between the training set and the test set was approximately 1:9.

Feature construction

The Amino Acid Index Database (AAIndex) [10,11] is a dataset that contains hundreds of various physiochemical and biological amino acid properties. Each index represents a kind of amino acid properties which are presented in the form of numeric matrixes. The AAIndex release 8.0 containing 562 indices was used to encode our peptide samples. Only 506 of them were chosen, excluding indices containing missing values and ambiguous annotations. Hence, a 14-residue peptide sample could be represented by a $506 * 14$ dimension vector. Each dimension in the vector was regarded as a feature. As a result, the total number of our candidate features is 7,084. See the supplementary material III for the 506 amino acid indices.

The conversion for a peptide sample to a feature vector can be computed by Eq. 1:

$$P = (f_0, f_1, \ldots, f_i, \ldots, f_n) \quad (i = 0, 1, \ldots, 7,083), \quad (1)$$

where i can be computed as $i = (\text{position}_{\text{residue}} \times 506) + \text{index}_{\text{residue}}$

In a reversed way, feature i can be easily mapped back to the position and index of amino acid property by the following equations:

$$\text{position}_{\text{residue}} = i/506, \quad \text{index}_{\text{residue}} = i\%506$$

Nearest Neighbor Algorithm (NNA)

In our study, NNA [13] was used to produce the prediction models. The basic idea of NNA is to classify samples based on their nearest neighbors. The "near" level of two samples is measured by the similarity between them. NNA works well on samples with extremely large dimension and are not easily influenced by the quantitative unbalance between positive and negative samples.

We use the following distance to compute the similarity level between two sample vectors.

$$D(P_x, P_y) = 1 - \frac{P_x \cdot P_y}{\|P_x\| \cdot \|P_y\|}, \tag{2}$$

where $P_x \cdot P_y$ is the dot product of vectors P_x and P_y. $\|P_x\|$ and $\|P_y\|$ are their modulus, respectively. The smaller the $D(P_x, P_y)$ is, the more similar the two vectors are. Especially, when $D(P_x, P_y) = 0$, the two vectors are identical.

Based on the measurement defined in Eq. 2, a target sample P_t will be designated to the same class as its nearest neighbor P_n which has the smallest $D(P_n, P_t)$ with it.

$$D(P_n, P_t) = \text{Min}\{D(P_1, P_t), D(P_2, P_t), \ldots, D(P_i, P_t), \ldots, D(P_N, P_t)\} \quad (i \neq t), \tag{3}$$

where N represents the total number of the training set.

Leave-one-out cross-validation (LOO)

In statistical prediction, LOO is deemed the most objective way to evaluate the performance of the prediction model [14–16]. In our study, different candidate feature sets will lead to different prediction models. LOO is used to compute the classification accuracy for different models and help to produce the classification accuracy curve.

Feature selection

The extraction of the informative features from all the candidate features is not an easy job. Hypothesizing that there are N candidate features, there will be $(2^N - 1)$ possible feature subsets. In our study, $(2^{7,084} - 1)$ $(N = 7{,}084)$ is such a large number that it is impossible to make a full evaluation for all the feature subsets with the current computational source because the increment of time consumed is in an exponential scale.

Hence, we developed a feature selection strategy to solve such time-consuming problem. First, mRMR method is used to pre-evaluate all the candidate features. Then, an incremental feature selection method was applied to extract the most informative features.

mRMR method [12]

mRMR stands for minimal-redundancy–maximal-relevance criterion for feature selection. It was developed by Peng et al. [12]. Maximal relevance means to select features which have the strongest correlation with the target class, while minimal redundancy means to make the redundancy existing among the already selected features as little as possible. mRMR method could evaluate features according to both the maximal relevance and minimal redundancy. Details of the mRMR method can be found in reference [12].

Incremental feature selection

The mRMR program would output the result as follows:

$$S = \{f_0, f_1, \ldots, f_i, \ldots, f_{N-1}\} \quad (N = 7{,}084), \tag{4}$$

where i represents the feature order determined by the mRMR method.

According to the mRMR method, higher order features were better than the lower ones. For example, f_a satisfied minimal-redundancy–maximal-relevance criterion much more than f_b, if $a < b$. Only the feature orders were not enough, and we needed to know how many and what features should be selected. In this study, incremental feature selection was applied to determine the final optimal feature subset.

Construct $N(N = 7{,}084)$ sequential feature subsets by incrementally adding features from S defined in Eq. 4,

$$\begin{aligned} S_0 &= \{f_0\} \\ S_1 &= \{f_0, f_1\} \\ &\vdots \\ S_i &= \{f_0, f_1, \ldots, f_i\} \\ &\vdots \\ S_{N-1} &= \{f_0, f_1, \ldots, f_{N-1}\}. \end{aligned} \tag{5}$$

NNA was used to produce the prediction model with features within the same sequential feature subsets. Then, LOO was used to evaluate the prediction models and produce a classification accuracy curve. The feature subset corresponding to the apex of the accuracy curve would be selected as the optimal feature set.

Results

mRMR result

At the first stage, all the candidate features were evaluated by mRMR program. In our study, the mRMR program was downloaded from the website http://research.janelia.org/peng/proj/mRMR/index.htm. Since the type of our data was

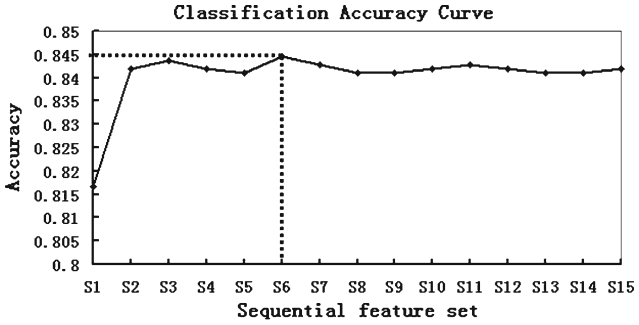

Fig. 1 The accuracy of prediction models based on sequential feature sets from S_1 to S_{15}

continuous, which was not compatible with mRMR program, we chose the parameter $t = 1$ to discretize our data to three categorical states according to the equation mean $\pm (t \cdot \text{std})$ (where mean is the mean value and std is the standard deviation). During process of the mRMR program, feature selected at each round was recorded and ranged in a list. The feature which satisfied the minimal-redundancy–maximal-relevance criterion more would attain higher order in the mRMR result list. See the supplementary material IV.

Classification accuracy curve

In the incremental feature selection, a classification accuracy curve would be plotted with the help of NNA [13] and LOO [14]. NNA was used to combine the selected features into a prediction model, while LOO was used to evaluate the performance of prediction models and generate the classification accuracy curve (See Fig. 1).

The curve showed that the S_6 could achieve the highest accuracy 84.4%. S_6 was defined as Eq. 5 and contained the first six features presented in the mRMR result list (See Table 1).

Comparison among three feature subsets

Table 2 showed that if all the features were used without feature selection, then the prediction accuracy was only 66.4% for the training set. After applying the feature selection procedure, the prediction accuracy rose to 84.4%. Especially, the first two features could achieve a high prediction accuracy of 81.6%, which was very close to the final optimal prediction accuracy of 84.4%.

Training set and test set

Tested by LOO cross-validation, the optimal prediction model reaches the accuracy of 84.4% for the training set and 76.4% for the test set. When applying the optimal prediction model to the whole data set (containing 212 substrates), 180 substrates were correctly predicted, which was 18 more than only using the motif ΨKXE.

Discussion

Feature construction based on Amino Acid Index

Using as many candidate features as possible might guarantee useful features which would not be missed, because most features were not understood before. In our study, 506 amino acid indices collected in the Amino Acid Index Database (AAIndex) [10,11] were selected for peptide coding. Hence, every peptide sample containing only 14 amino acid residues can be represented by as many as 7,084 features. Compared to other coding methods (for example, the 0/1 system), more amino acid properties were covered by using AAIndex. Furthermore, the Amino Acid Index Database is still growing as more and more new amino acid indices are being published.

On the other hand, since each feature element within the vector was represented by position and amino acid index as defined in Eq. 1, it was easy for us to map the selected features back to their corresponding positions and amino acid property indices, which makes a further biological analysis feasible. For example, hypothesize that the 2008th feature was included as one of the informative features. As a result, the 490th ($2008\%506 = 490$) index of amino acid residue at position 3($2008/506 = 3$) might have a relationship with the recognition of sumoylation sites.

Feature selection system

Table 2 indicated that if all the features were used, the prediction accuracy was only 66.4%, suggesting that some redundant and useless features within the original feature set should be discarded. In our prediction system, the mRMR method was the beginning, which was used for feature pre-evaluation. The features satisfied minimal-redundancy–maximal-relevance criterion well, which would be given high evaluation by mRMR method. From Table 2 we can see that the feature subset S_2, which contained the first two features in the mRMR result list, could be used to construct a prediction model achieving 81.6% prediction accuracy indicating that the evaluations provided by mRMR were reliable and efficient. Also, due to the mRMR method application, the time consumed by our feature selection system was reduced to a polynomial scale.

Final optimal features set

First two features

The first two features in the optimal feature subset could combine to achieve the accuracy of 81.6% which was close to the optimal accuracy 84.4%. They were also the first two features in the mRMR result list. See Table 1 for details.

Table 1 Final optimal feature set

No.	Position	Amino acid property
1	9	Hydropathy scale based on self-information values in the two-state model
2	7	Normalized frequency of beta-turn
3	7	alpha-CH chemical shifts
4	7	Hydrophobicity factor
5	9	Negative charge
6	7	Relative preference value at N'
7	9	Loss of side chain hydropathy by helix formation

Table 2 The accuracy of three prediction models based on three different feature sets

Feature set	All features (%)	S_1 (%)	Final feature set (%)
Accuracy	66.4	81.6	84.4

All features means that the feature set is containing all the 7,084 features. S_1 is the second sequential feature subset. Final feature set contains the optimal features in Table 1. They are all combined by NNA[13] and evaluated by LOO[14]. The prediction accuracies are all for the training set

Since our sample peptides consisted of 14 residues with seven residues upstream and seven residues downstream of the Lysine (K), positions 9 and 7 are corresponding to locations two-residue downstream and two-residue upstream of the Lysine (K), which can be interpreted as $X_{2\text{nd}}KXX_{1\text{st}}$. That motif $X_{2\text{nd}}KXX_{1\text{st}}$ has been confirmed by the previous study [1,7]. In the SUMO conjugation pathway, SUMO-conjugating enzyme (Ubc9) plays the most important role in recognizing the substrates [17]. Furthermore, $X_{2\text{nd}}KXX_{1\text{st}}$ is the region that binds to Ubc9 most closely in the tertiary structure [18] (see Fig. 2). When compared to the widely used motif ΨKXE [1,7] mentioned in the introduction, it is understandable that mRMR places these two features at top two places in the result list. Maximal relevance may be illustrated as high correlation to motif ΨKXE, while minimum redundancy may be interpreted as different positions and different biological properties between these two features.

Five complemental features

Besides the first two features, there were five features left in the optimal feature subset which account for 2.8% accuracy enhancement. Such five features could not be found in the first 500 features of Max-relevance result list indicates their weak correlation with the class variable. However, they occupy top places in the Max-relevance–Min-redundancy result list suggesting that they have low redundancy against the first two features and must contain some information in determining the sumoylation sites that the first two features lack. Therefore, they are selected to optimize the classification accuracy.

Table 1 showed that these five features also belonged to either position 9 or position 7 which was consistent with the observation that amino acid residues which are two-residue downstream and one-residue upstream of the Lysine (K) play the most important role in determining the sumoylation sites [1,7]. On the other hand, their amino acid property tables refer to many kinds of physiochemical and biological properties, reflecting the fact that sumoylation is actually a

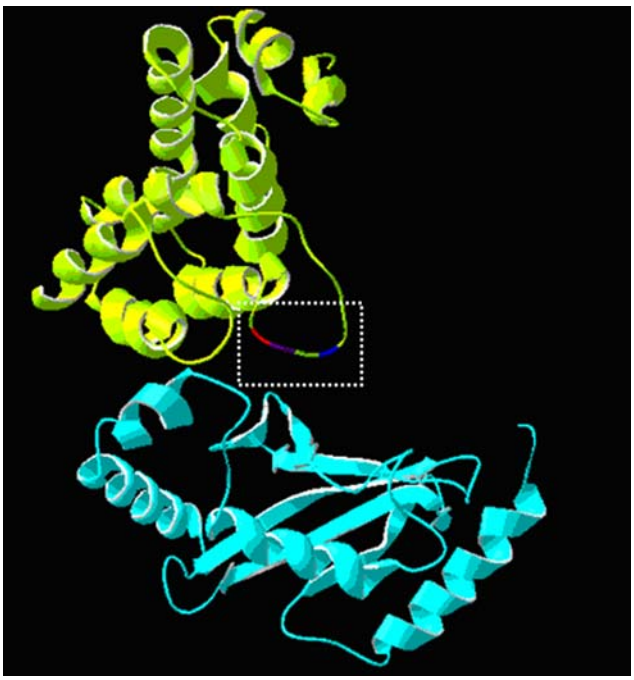

Fig. 2 The tertiary structure of SUMO-conjugating enzyme (Ubc9) binding to the substrate. The crystal data is attained from the work of Bernier–Villamor using the GanGAP1 as the substrate [18]. SUMO (*cerulean part*) is conjugated to the RanGAP1 (*green part*). And according to our sample definition, sumoylation site Lysine (K) is the purple point while position 9 and 7 are the *blue* and *red point*

complex process, and there must be many other unknown factors in the determination of sumoylation sites. Therefore, these complemental features can be used as clues for biologists to expand their knowledge of sumoylation mechanism.

Conclusion

In our study, we developed a computational system to predict the sumoylation sites based on feature selection strategy. Our results showed that the prediction system could not only achieve high performance at low computational time but also provide much useful information for biological analysis. Therefore, our prediction system is capable of high throughput identification of sumoylation site and can also be used as a tool for investigation of sumoylation mechanism. The prediction software is upon request.

Acknowledgements This study was funded by "National Basic Research Program of China (973)":2 006CB910700, 2004CB518606, and 2003CB715901; "National High-Tech R & D Program (863)": 2006AA02Z334; Funding of CAS: KSCX2-YW-R-112; and Shanghai Leading Academic Discipline Project(J50101).

References

1. Mann M, Jensen ON (2003) Proteomic analysis of post-translational modifications. Nat Biotechnol 21:255–261. doi:10.1038/nbt0303-255
2. Johnson ES (2004) Protein modification by SUMO. Annu Rev Biochem 73:355–382. doi:10.1146/annurev.biochem.73.011303.074118
3. Girdwood DW, Tatham MH, Hay RT (2004) SUMO and transcriptional regulation. Semin Cell Dev Biol 15:201–210. doi:10.1016/j.semcdb.2003.12.001
4. Liang M, Melchior F, Feng XH, Lin X (2004) Regulation of Smad4 sumoylation and transforming growth factor-beta signaling by protein inhibitor of activated STAT1. J Biol Chem 279:22857–22865. doi:0.1074/jbc.M401554200
5. Li M, Guo D, Isales CM, Eizirik DL, Atkinson M, She JX, Wang CY (2005) SUMO wrestling with type 1 diabetes. J Mol Med 83:504–513. doi:10.1007/s00109-005-0645-5
6. Shinbo Y, Niki T, Taira T, Ooe H, Takahashi-Niki K, Maita C, Seino C, Iguchi-Ariga SM, Ariga H (2006) Proper SUMO-1 conjugation is essential to DJ-1 to exert its full activities. Cell Death Differ 13:96–108. doi:10.1038/sj.cdd.4401704
7. Hay RT (2005) SUMO: a history of modification. Mol Cell 18: 1–12. doi:10.1016/j.molcel.2005.03.012
8. Xue Y, Zhou F, Fu C, Xu Y, Yao X (2006) SUMOsp: a web server for sumoylation site prediction. Nucleic Acids Res 34:W254–W257. doi:10.1093/nar/gkl207
9. Harder Z, Zunino R, McBride H (2004) Sumo1 conjugates mitochondrial substrates and participates in mitochondrial fission. Curr Biol 14: 340–345
10. Kawashima S, Kanehisa M (2000) Amino acid index database. Nucleic Acids Res 28:374
11. Kawashima S, Ogata H, Kanehisa M (1999) Amino acid index database. Nucleic Acids Res 27:368–369. doi:10.1093/nar/27.1.368
12. Peng HC, Long FH, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238. doi:10.1109/TPAMI.2005.159
13. Cai YD, Chou KC (2006) Predicting membrane protein type by functional domain composition and pseudo-amino acid composition. J Theor Biol 238:395–400. doi:10.1016/j.jtbi.2005.05.035
14. Chou KC, Zhang CT (1995) Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349. doi:10.3109/10409239509083488
15. Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17:729–738. doi:10.1023/A:1020713915365
16. Cai YD (2001) Is it a paradox or misinterpretation? Proteins 43:336–338. doi:10.1002/prot.1045
17. Lin D, Tatham MH, Yu B, Kim S, Hay RT, Chen Y (2002) Identification of a substrate recognition site on Ubc9. J Biol Chem 277:21740–21748. doi:10.1074/jbc.M108418200
18. Bernier-Villamor V, Sampson DA, Matunis MJ, Lima CD (2002) Structural basis for E2-mediated SUMO conjugation revealed by a complex between ubiquitin-conjugating enzyme Ubc9 and RanGAP1. Cell 108:345–356. doi:10.1016/S0092-8674(02)00630-X