Journal of Computational Intelligence in Bioinformatics
ISSN: 0973-385X Volume 2 Number 2 (2009) pp. 103–107

http://www.ripublication.com/jcib.htm

SUMOylation Site Prediction using Support Vector Machines

Sarabjot Singh Pabla[1], Simarjot Singh Pabla and Hetalkumar Panchal[2]

G.H. Patel P.G.Dept of Computer Science and Technology.
Sardar Patel University, Vallabh Vidyanagar, Gujarat – India.
E-mail: pablasarabjot@gmail.com and 2 swamihetal@gmail.com

Abstract

SUMOuylation has been identified as a crucial post-translational modification responsible for many if not all cellular processes. A considerable amount of information pertaining to this process is still elusive. The available information can still be used in development of in-silico procedures to guide the prediction of SUMOuylation sites in protein substrates. Since a highly accurate prediction system is essential to guide efficient experimental designs. We put forth a SUMOuylation site prediction program using support vector machines for discriminating SUMOylating substrates from non-SUMOylating substrates. The algorithm uses manually curated data of experimentally verified SUMOuylation sites as training data. As the number of experimentally verified proteins increases, the accuracy and efficiency of the program will increase accordingly. The web interface for the tool is still under development, meanwhile the stand alone version of the program can be requested from the corresponding author at the given email address.

Keywords: SUMOuylation, Site Prediction, Support Vector Machines and Bioinformatics.

Introduction

SUMO (Small Ubiquitin-like Modifier) proteins constitute a family of proteins that covalently binds and subsequently disengages from other proteins in cells, thus altering their function. SUMOylation is a type of post-translational modification central to various cellular processes like transcriptional regulation, nuclear-cytosolic transport, maintaining stability of proteins, apoptosis, succession through the cell cycle and response to stress [1]. SUMO proteins are comparable to ubiquitin, and SUMOylation follows an enzymatic cascade similar to that in

ubiquitination. Unlike ubiquitin, SUMO is not used to mark proteins for elimination. Activated SUMO is formed when the last four amino acids of the C-terminus have been cleaved off. It permits the creation of an isopeptide bond between an acceptor lysine on the target protein and C-terminal glycine residue of SUMO. SUMOylation of proteins has many outcomes. The most common and best studied are nuclear-cytosolic transport, protein stability and transcriptional regulation. Usually, only a small portion of a given protein is SUMOlated and this alteration is rapidly inverted by deSUMOylating enzymes. RanGAP1 when modified by SUMO-1, leads to its transport from cytosol to nuclear pore complex [2][3]. Similarly in hNinein, SUMO modification results in its trafficking from the centrosome to the nucleus [4]. Many times transcriptional regulators undergoing SUMO modification results in inhibition of transcription [5]. SUMO proteins being smaller in size; are about 100 amino acids in length and 12 kDa in mass.

The support vector machine (SVM) algorithm is a classification algorithm that offers high degree of performance in a wide range of application domains which includes object recognition, handwriting recognition, face detection, text categorization, and speaker identification [6]. During the past three years, SVMs have been applied very broadly within the field of computational biology, to various problems like pattern recognition which encompasses recognition of translation start sites, microarray gene expression analysis, functional classification of promoter regions, protein remote homology detection, prediction of protein-protein interactions, and peptide identification from mass spectrometry data.

Implementation

SUMOylation and Sequence conservation

The identification of a consensus SUMO-acceptor site was possible after mapping acceptor Lys residues in just a few SUMO targets, which included RanGAP1 [7][8], PML20, Sp100 [9], inhibitor of nuclear factor-κB (IκBα) [10], p53 and c-Jun [11]. The SUMO-acceptor site was shown to be ΨKxE (in which Ψ is an aliphatic branched amino acid and x is any amino acid). This motif is crucial for predicting whether a protein undergoes the process of SUMOylation or not.

Data Preparation

Our dataset consists of manually curated protein sequences obtained via usage of "SUMO" and related keywords in Pubmed. Since we had all the swiss-prot IDs, it would have been a very tedious task to extract every sequence one by one from the database. To overcome this lengthy procedure we designed a supplementary program "FEST (**F**asta **E**xtractor for **S**wiss-prot/**T**remble)" to automate this task.

Algorithm

The algorithm isolates potential SUMOylation peptides with lysine (K) residue neighbored by "n" residues upstream and "n" residues downstream [12]. This assumption is based on the hypothesis that the biochemical properties of a SUMOylation site mainly depend on the neighboring amino acids. In our program, we

use n=7, which is validated by the prediction performance. A raw protein sequence is entered by the user into the text area provided in program input window [Figure 1]. Program searches for all lysine residues in the sequence and notes their respective positions. A fifteen residue fragment is extracted from all the positions. Four hundred and twenty four features of each of these fragments are calculated. Support vector machine then classifies all the fragments into two classes i.e. positive or negative. All the predicted positive fragments are then displayed along with their positions in the output [Figure 2].

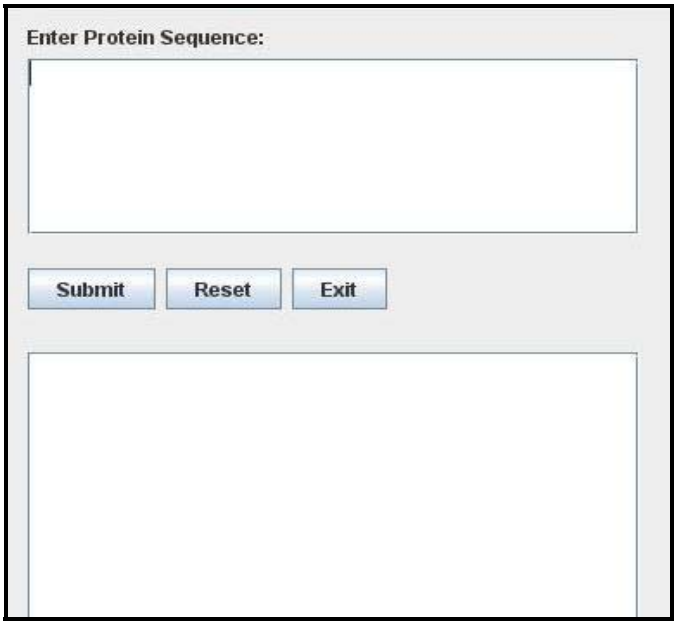

Figure 1: Input window of our prediction system.

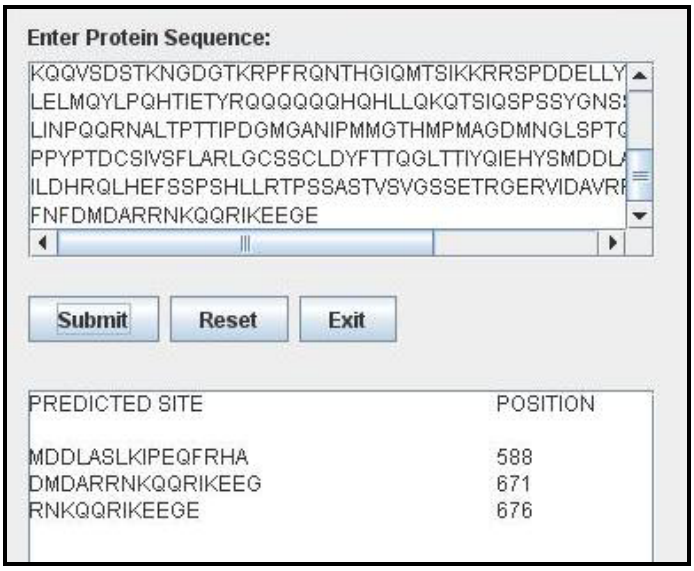

Figure 2: Output window of our prediction system.

Results

The factors used to determine the prediction performance of our program are sensitivity (Sn), specificity (Sp) and accuracy (Ac) [Table 1]. Sensitivity measures the positive predictions while specificity measures negative predictions. Also, accuracy ascertains the correct prediction ratio. We used various kernel functions and their parameters to achieve maximum accuracy while keeping low false positive rate. We have also employed a correlation coefficient (CC) to evaluate the prediction system. CC values range between -1 to 1, and the nearer to 1 a CC is, the more precise the prediction is. Positive data set consists of known SUMOylation sites, whereas other lysine residues in known in sumoylated proteins are considered as negative.

$$Sn = \frac{TP}{(TP + FN)} \qquad Sp = \frac{TN}{(TN + FP)}$$

$$Ac = \frac{(TP + TN)}{(TP + FP + TN + FN)} \qquad CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{[(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)]}}$$

Where TP= True Positive, TN = True Negative, FN = False Negative and FP = False Positive

Table 1: Prediction performance of Our Program, SUMOsp and SUMOplot

Predictor	**Ac (%)**	**Sn (%)**	**Sp (%)**	**CC (%)**
SUMOsp	92.71	83.68	93.08	0.5012
SUMOplot	89.94	79.50	93.31	0.4825
Our Program	***90.34***	***85.72***	***92.29***	***0.4987***

Discussion and Conclusion

False positive rate of this tool can be further reduced if BLAST score is used. The predicted positive fragments can be compared with the known positive fragments and a high similarity will enable the program to set a reliable cut off value. This cut off value will then be used to filter out false positive fragments. This feature is not yet included in to program due to our limited understanding of integrating BLAST into our algorithm. Also our program is standalone application whose performance has been tuned for running offline. Our best efforts are underway to develop an online version so that it more easily accessible to the scientific community at large.

Although experimental validation is inevitable in SUMOylation research, computational prediction systems can greatly aid in designing workflows for such experiments. Also, it would considerably decrease the amount of time previously required to identify SUMOylation sites invivo or invitro. Computational approaches like these can be useful for performing various forms of analyses which include large scale proteome studies.

References

[1] Hay RT (Apr 2005). "SUMO: a history of modification". *Mol. Cell* 18 (1): 1–12.

[2] Matunis MJ, Coutavas E, Blobel G (Dec 1996). "A novel ubiquitin-like modification modulates the partitioning of the Ran-GTPase-activating protein RanGAP1 between the cytosol and the nuclear pore complex". *J Cell Biol.* 135 (6 Pt 1): 1457–70.

[3] Mahajan R, Delphin C, Guan T, Gerace L, Melchior F (Jan 1997). "A small ubiquitin-related polypeptide involved in targeting RanGAP1 to nuclear pore complex protein RanBP2". *Cell* 88 (1): 97–107.

[4] Cheng TS, Chang LK, Howng SL, Lu PJ, Lee CI, Hong YR (Feb 2006). "SUMO-1 modification of centrosomal protein hNinein promotes hNinein nuclear localization".*Life Sci.* 78 (10): 1114–20.

[5] Gill G (Oct 2005). "Something about SUMO inhibits transcription". *Curr Opin Genet Dev.* 15 (5): 536–41.

[6] Terry Furey, Nello Cristianini, Nigel Duffy, Michel Schummer, David Bednarski, David Haussler (2000) “Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data” Bioinformatics, 16(10): 906-914.

[7] Mahajan, R., Gerace, L. & Melchior, F (1998). Molecular characterization of the SUMO- 1 modification of RanGAP1 and its role in nuclear envelope association.*J. Cell Biol.* 140, 259–270.

[8] Matunis, M. J., Wu, J. & Blobel, G (1998). SUMO-1 modification and its role in targeting the Ran GTPaseactivating protein, RanGAP1, to the nuclear pore complex. *J. Cell Biol.* 140, 499–509.

[9] Kamitani, T. *et al* (1998). Identification of three major sentrinization sites in PML. *J. Biol. Chem.* 273, 26675–26682.

[10] Sternsdorf, T., Jensen, K., Reich, B. & Will, H (1999). The nuclear dot protein Sp100, characterization of domains necessary for dimerization, subcellular localization, and modification by small ubiquitin-like modifiers. *J. Biol. Chem.* 274, 12555–12566.

[11] Desterro, J. M., Rodriguez, M. S. & Hay, R. T (1998). SUMO-1 modification of IκBα inhibits NF-κB activation. *Mol. Cell* 2, 233–239

[12] Yu Xue, Fengfeng Zhou, Chuanhai Fu, Ying Xu and Xuebiao Yao (2006). SUMOsp: a web server for sumoylation site prediction. Nucleic Acids Res. 34:254-257.