A Novel Method for N-terminal Acetylation Prediction

Ying Liu[1]* and Yuanlie Lin[2]

[1] *School of Software, Tsinghua University, Beijing 100084, China;* [2] *Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China.*

The NetAcet method has been developed to make predictions of N-terminal acetylation sites, but more information of the data set could be utilized to improve the performance of the model. By employing a new way to extract patterns from sequences and using a sample balancing mechanism, we obtained a correlation coefficient of 0.85, and a sensitivity of 93% on an independent mammalian data set. A web server utilizing this method has been constructed and is available at http://166.111.24.5/acetylation.html.

Key words: N-terminal acetylation, support vector machine (SVM)

Introduction

N-terminal acetylation is one of the most common protein modifications in eukaryotes, occurring on approximately 80%-90% of the cytosolic mammalian proteins (*1*, *2*). Previously, much work has been done to make predictions based on the data available. The latest achievement in this field is the NetAcet method (*3*).

The NetAcet method was based on a yeast dataset (*1*, *2*) and the Yeast Protein Map (YPM) resource (*4*). Only substrates reported to be acetylated by N-acetyltransferase A (NatA) were extracted. After redundancy reduction, there were finally 57 positive and 72 negative sequences.

In NetAcet, sequences were first truncated to their N-terminal 40 residues. Then, patterns were extracted with a window size of seven amino acids, with position 1 being the target residue. Only negative examples with either serine, threonine, alanine, or glycine in the first position of the window were used, as the other types were trivial.

An artificial neural network was trained using 3-fold cross-validation, with the extracted patterns as its training set. Since the number of negative examples was much greater than that of positive ones (there were 57 positive examples but more than 1,000 negative ones), 57 negative examples were randomly selected from the overall negative data set. Along with the 57 positive examples, they composed the input to the model. A Matthews correlation coefficient (MCC; ref. *5*) of 0.69 was obtained from the model, with a sensitivity of 75% and a specificity of 92%. On an independent test set of mammalian N-acetylated protein extracted from Uniprot (*6*), it achieved a sensitivity of 74% on acetylated serines.

Results

After improving the experiment, we obtained a Matthews correlation coefficient of 0.85. This reflects a sensitivity of 86% and a specificity of 97%. In NetAcet, the corresponding values are 0.69, 75% and 92% (Table 1). The specificity on negative examples with a serine residue at position 2 is 98%. That is about 38% higher than NetAcet.

Table 1 Performance Comparison of the Two Methods

Performances	NetAcet	SVM
MCC	0.69	0.85
Sensitivity	75%	86%
Specificity	92%	97%
Specificity on non-acetylated serines	60%	98%
Sensitivity on N-acetylserine of mammalian data	74%	93%

We also tested the model on a mammalian protein data set extracted from the Uniprot. By using the FtDescription (Feature) option from Sequence Retrieval System, we extracted 260 mammalian proteins reported to have N-acetylated serine. By using the *Decrease redundancy* program provided by ExPASy, we obtained 77 mammalian proteins that have the

*** Corresponding author.**
E-mail: yingliu03@mails.tsinghua.edu.cn

maximum similarity of 80%. We tested the training model on this data set and obtained a sensitivity of 93% (72 were found with acetylated serine).

Discussion

The sensitivity obtained from the cross validation is 86%, which is 7% lower than that from the mammalian data (93%). That is because there are only acetylated serines in the mammalian data. Furthermore, for other types of substrates (threonine, alanine, and glysine), we obtained a much lower performance with the same model, which we attribute to the inadequacy of positive examples of other types of acetylated residues. The method presented here greatly improves the prediction performance of N-acetylation of N-acetyltransferase A. The experiment results convince us that N-terminal methionine cleavage has a profound effect on N-terminal acetylation. This relationship will become clearer if more data are available to enable further statistical analysis.

Methods

We employed the same data set used by NetAcet as our training set, but we made improvements in the way to extract patterns. Furthermore, we used the support vector machine (SVM) as the training model.

The data set is composed of yeast proteins. As previous studies indicated, removal of N-terminal methionine is an essential function in yeast (*1*, *2*). Moreover, methionine excision occurs before N-terminal acetylation, and it also takes place at N-terminal. So we are encouraged by these facts to assume that the pattern of the acetylated site is more or less relative to the methionine cleavage at N-terminal of the sequence. Interestingly, the information contained in the positive data set is consistent with our hypothesis to a certain degree. The acetylated site is either located at N-terminal or rightly next to the N-terminal methionine. If all the information can be encoded into the patterns, the model will be able to perform better in classification.

So we extracted patterns like this. In addition to subsequent residues following an acetylated site, we included one more residue ahead of each acetylated site. If the acetylated residue is located first at N-terminal, we use a symbol "X" to represent the residue ahead of it. We found that all positive examples begin with either "M" or "X" (Figure 1). Thus the information about the N-terminal methionine cleavage has been encoded into the patterns that we have extracted.

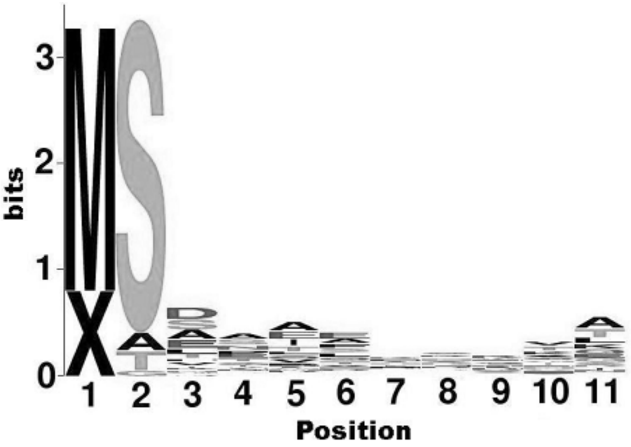

Fig. 1 Shannon information (*7*) sequence logo (*8*) of 57 acetylation sites in the positive samples, in the format of extracted patterns. The height of each letter is made proportional to its frequency, and the letters are sorted so that the most common one is on top. Acetylation is reported on Position 2 in the logo. Position 1 is either methionine (M) or empty (X). Position 2 is mostly occupied by S, which means that our positive samples are primarily composed of acetylated serines.

With regard to negative examples, patterns were also extracted as described above, with the target residue at position 2. However, in order to balance the information added to positive examples, we no longer select negative examples in a completely random way, as NetAcet did. Instead, we collected all the negative examples that began with "M" or "X" (the number amounts to 40), and made them "fixed negative examples". Then we selected the other negative examples randomly from the ones left in order to form a training set of 171 examples, which will be the input to the model.

Finally, the SVM model was trained using 3-fold cross-validation. Sparse coding was used for translating the amino acids to data input to the model (*9*). In our experiment several window lengths have been tried. In the optimal case (when the window length is 5), the RBF kernel was used with the optimal parameters $\gamma = 0.14$, C = 1.1.

Acknowledgements

The authors would like to thank Lars Kiemer, the creator of NetAcet, for his valuable assistance.

References

1. Polevoda, B. and Sherman, F. 2000. N$^{\alpha}$-terminal acetylation of eukaryotic proteins. *J. Biol. Chem.* 275: 36479-36482.
2. Polevoda, B. and Sherman, F. 2003. N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins. *J. Mol. Biol.* 325: 595-622.
3. Kiemer, L., *et al.* NetAcet: prediction of N-terminal acetylation sites. *Bioinformatics*. In press.
4. Perrot, M., *et al.* 1999. Two-dimensional gel protein database of *Saccharomyces cerevisiae* (update 1999). *Electrophoresis* 20: 2280-2298.
5. Matthews, B.W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405: 442-451.
6. Apweiler, *et al.* 2004. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 32: D115-119.
7. Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Tech. J.* 27: 379-423, 623-656.
8. Schneider, T.D. and Stephens, R.M. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18: 6097-6100.
9. Blom, N., *et al.* 1996. Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks. *Protein Sci.* 5: 2203-2216.

This work was supported by the National Natural Science Foundation of China (No.10371063) and the National Key Technologies R&D Program of China (No. 2004ba711a21).