Analysis of N-terminal Acetylation data with Kernel-Based Clustering

Ying Liu
Department of Computational Biology, School of Medicine
University of Pittsburgh

yil43@pitt.edu

1 Introduction

N-terminal acetylation is one of the most common protein modifications in eukaryotes, occurring on approximately 80%—90% of the cytoplasmic mammalian proteins and 50% of yeast proteins (Polevoda *et al*., 2003). In my previous work, I have trained a SVM classifier to classify the acetylated residues. A yeast data set obtained from (Kiemer *et al*., 2005) was used as the training set. I also tested the model on a mammalian protein data set extracted from the Uniprot which containes 77 mammalian proteins with the maximum similarity of 80%. The polypeptide sequences were first truncated to their N-terminal 40 residues, then I extracted patterns with a sliding window of several amino acids. Sparse coding scheme (Blom *et al*., 1996) was used for translating the amino acids to 0-1 vectors as input to the model. Then support vector machine (SVM) was used as the training model and classifier.

In the traditional scheme of dealing with such problems (N-terminal modification), position 1 is usually used as the target residue; but here I also tried another experiment using position 2 as the target residue. The reason why I am doing this can be explained from the observation of Figure 1, from which we see that all positive examples begin with either "M" (methionine) or "X" (empty). Thus the information about the N-terminal methionine cleavage has been encoded into the patterns if we use position 2 as the target residue. A comparison was made between these two experiments both using SVM but different pattern extraction schemes. By doing so I found there is a significant increase in prediction accuracy (Figure 2). So the effect brought by the new scheme is two-fold: first, it contains N-terminal location information; second, it contains N-terminal methionine cleavage information, which may affect the acetylation motif to some extent, as previous research indicates that methionine cleavage occurs ahead of acetylation in time (Polevoda *et al*., 2003). But this hypothesis remains to be verified. In this paper, I would like to explore this problem with unsupervised procedures.

2 Related Work

[1] Lars Kiemer, Jannick Dyrløv Bendtsen, Nikolaj Blom. (2005) NetAcet: prediction of N-terminal acetylation sites. *Bioinformatics*, 21(7): 1269-1270.

[2] Bernhard Schölkopf & Alexander J. Smola. (2002) Learning with Kernels. Cambridge, MA: MIT Press.

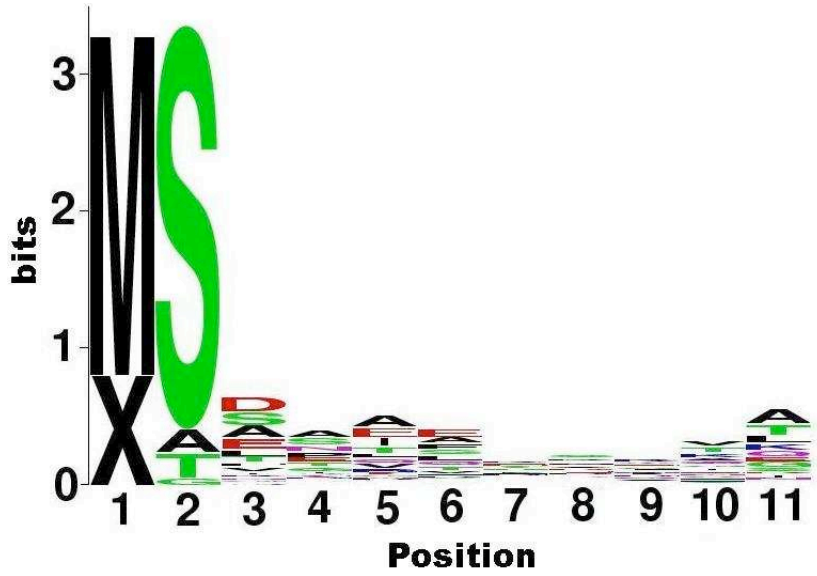

Figure 1: Shannon information (Shannon, 1948) sequence logo of 57 acetylation sites, in the format of extracted patterns. Acetylation is reported on Position 2 in the logo.

[3] 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.

[4] Blom N., et al. (1996). Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks. *Protein Sci.*, 5: 2203-2216.

3 Method: k-Means Clustering with Kernel Tricks

The basic idea is to use clustering on the dataset and see how the pattern extraction method will influence the unsupervised procedure. Here we use k-means clustering, where k is simply 2. The general k-means clustering uses Euclidean distance as the distance measure, but it is obviously inappropriate for the sparse coding scheme. In this case the Euclidean distance will become the square root of Hamming distance, which causes evident information loss. Furthermore, because I used SVM as the training model, and the distance measure in SVM is represented in terms of kernel function, thus it is advisable that kernels also be used in the clustering procedure.

Kernels can be regarded as generalized dot products (Bernhard *et al*., 2002), denoted as $k(\cdot,\cdot)$. When a kernel is introduced, a non-linear feature space is implicitly constructed along with a mapping from the original space to the new one. Let this mapping be denoted Φ, thus each sample x in the original space is mapped into the feature space as $\Phi(x)$. Given the kernel function, we can compute the dot product of x and x' in the feature space without explicitly solving $\Phi(x)$ and $\Phi(x')$, i.e.,

$$k(x,x') = \langle \Phi(x), \Phi(x') \rangle.$$

Basically, all the computation here concerning the distance can be done using kernel.

By definition, the Euclidean distance between two samples is the norm of the difference between them:

$$d(x,x') = \|x - x'\| = \sqrt{\langle x - x', x - x' \rangle}.$$

For comparison only, the square root can be cast away,

$$\begin{aligned} d^{\,2}(x,x') &= \|x - x'\|^2 = \langle x - x', x - x' \rangle \\ &= \langle x, x \rangle + \langle x', x' \rangle - 2\langle x, x' \rangle. \end{aligned}$$

thus we only need to know the dot product to solve the distance.

Similarly, we can compute the distance in the feature space with kernels alone, that is

$$d^{\,2}(\Phi(x), \Phi(x')) = k(x, x) + k(x', x') - 2k(x, x').$$

In k-means clustering, a key step is to compute the center of each cluster in each iteration, and compute the distance between each sample and each center. We can do this with kernels as well:

$$\langle \Phi(x), \frac{1}{m} \sum_{i=1}^{m} \Phi(x_i) \rangle = \frac{1}{m} \sum_{i=1}^{m} k(x, x_i).$$

Where $\frac{1}{m} \sum_{i=1}^{m} \Phi(x_i)$ is the center of a cluster containing x_i, $i = 1, \ldots, m$.

Then here comes the problem of kernel selection. First of all, in order to be consistent with my previous work, I use the RBF kernel. The RBF kernel has the form

$$k(x, x') = \exp\left(-\gamma ||x - x'||^2\right).$$

I also employ exactly the same parameter as in the previous SVM experiment, in which $\gamma = 0.14$. Note that for the RBF kernel the dot product of two identical samples is a constant 1, i.e., $k(x, x) = 1$. Thus

$$d^{\,2}(\Phi(x), \Phi(x')) = 2 - 2k(x, x').$$

Finally, I would try another interesting kernel. My new kernel will be based on the substitution matrix Blosum62 used in the sequence alignment. The substitution matrix is computated based on rigorous statistical theories and contains scores for all possible exchanges of one amino acid with another. The equation for calculating a score $s(a, b)$ for aligning two residues a and b is :

$$s(a, b) = \frac{1}{\lambda} \log \frac{p_{ab}}{f_a f_b}$$

where p_{ab} is the probability that we expect to observe residues a and b aligned in homologous sequence alignments. f_a and f_b are background frequencies: the probabilities that we expect to observe amino acids a and b on average in any protein sequence. Positive scores mean conservative substitutions, and negative scores indicate nonconservative substitutions. The substitution probability comes from the homologous aspects of the residues of the proteins, intuitively it stands for some kind of "distance" between different residues. Then it is natural to consider constructing a kernel from the substitution matrix. However, to construct a kernel, it is required that the Gram matrix be positive definite. Unfortunately Blosum62 itself is not, so we add a constant(say, 4) to each element in the matrix to make it positive definite. And then we can use this matrix as the Gram matrix of the kernel.

4 Experiment

The first two experiments try to find answers to the following two questions:

1. whether the methionine cleavage is related to the acetylation motif;
2. whether the methionine cleavage information is consistent with the acetylation motif to affect acetylation.

Hereby a sample selection mechanism is introduced. For the first question, only positive samples are considered, and they are labeled with whether they begin with "M", namely, whether they retain the N-terminal methionine; for the second question, both positive and negative samples are considered, and they are labeled with whether they have been acetylated. However, in order to neutralize the location information, I focus on the negative samples that begin with "M" or "X", which generally means they are also at the N-terminus. This yields to about 95 samples altogether. The first two experiments both use RBF kernels. The third experiment The third experiment is performed on the data used in Experiment 2. It is primarily for exploratory purpose, I just want to see how this new kernel works on this problem.

	M	no_M
MCC	0.38	0.94
Sensitivity	71%	100%
Specificity	64%	94%
Specificity on all negative examples	58%	98%
Sensitivity on UNI-PROT data	81%	84%

Figure 2: Results obtained from SVM classification with different pattern extraction methods.

For the sake of comparison, I perform each of the experiment in two groups, each group corresponds to a different pattern extraction scheme. That is, one group uses position 1 as the target, the other uses position 2 as the target. I would like to see how much the result "agrees" with the label. Here I postulate that the clustering procedure could discriminate the natural tendency of the data. The level of "agreement" is measured by summing up the common part of clusters and label groups. And I will perform each set of experiments 10 times and calculate the average to avoid accidental outcome.

5 Results and Discussion

Figure 5, 6 and 7 illustrate the results we obtained. In each figure, different curve type corresponds to different pattern extraction scheme. In Figure 5, the dashed line does not vary with the window length and remains at a low level. Since the samples used in experiment 1 are all positive, it is inferred that without additional information, the acetylation motif does not contain information to discriminate whether N-terminal methionine is retained or not. And the reason for the variation of the solid curve is self-evident, as the first residue in the pattern agrees with the label.

In Figure 6, the solid curve is under the dashed curve most of the time, which suggests the methionine cleavage information somehow obscures the motif that discriminates acetylation. It is also notable that both curves reach their peak when the window length is 6, from which we can infer the approximate range of the acetylation motif is 5 residues after the target. Interestingly, this inference agrees with both the result obtained in (Liu *et al*., 2004) and the previously proposed supposition in (Polevoda *et al*., 2003).

6 Conclusion

From the result of the experiment, we conclude that N-terminal methionine cleavage is, if not uncorrelated, inconsistent with the acetylation motif to affect acetylation; and the improvement obtained in our previous work mainly results from the N-terminal location information brought by the new pattern extraction scheme. We also find that clustering in combination with kernel tricks is useful in exploring motif identification problems, as different kernels can be used in specific problems to retrieve more information from the data. Also, the new kernel seems working just fine on this problem. I am looking forward to more powerful kernels for biological problems.

	5		6		7		8		9		10		11	
	M	no_M	M	no_M	M	no_M	M	no_M	M	no_M	M	no_M	M	no_M
1	39	36	44	35	35	28	33	34	29	30	28	34	33	31
2	57	28	41	34	36	36	31	24	27	35	28	25	30	32
3	46	30	57	31	30	28	40	34	32	27	29	26	25	32
4	35	31	49	32	48	31	29	27	30	33	32	30	32	28
5	34	28	55	28	39	29	32	30	30	36	35	33	29	29
6	57	34	33	28	41	35	26	33	29	25	27	32	34	31
7	57	26	32	25	36	25	32	31	22	34	27	33	24	31
8	57	32	57	31	30	31	34	28	23	29	34	27	29	35
9	57	34	57	29	52	29	25	32	31	28	27	27	26	25
10	38	29	34	35	34	34	26	31	32	27	24	29	33	27

Figure 3: Experimental data of Experiment 1.

	5		6		7		8		9		10		11	
	M	no_M	M	no_M	M	no_M	M	no_M	M	no_M	M	no_M	M	no_M
1	84	87	62	82	31	85	54	42	52	50	48	48	37	55
2	79	87	86	86	37	84	47	52	46	48	42	52	50	53
3	50	86	83	85	84	84	46	47	53	48	48	39	49	39
4	50	84	81	86	50	78	71	84	60	41	45	47	44	48
5	86	86	89	88	83	53	80	67	42	44	46	50	44	47
6	84	86	84	85	65	85	83	43	57	56	36	36	44	45
7	87	79	50	89	66	83	55	77	53	46	53	47	46	42
8	85	67	82	86	50	84	62	79	48	48	50	53	52	50
9	50	82	84	85	53	83	44	82	42	45	44	45	50	51
10	83	79	84	84	86	83	61	55	46	50	45	54	49	47

Figure 4: Experimental data of Experiment 2.

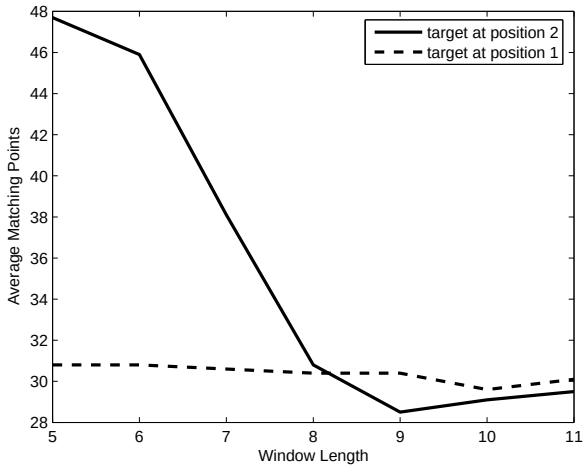

Figure 5: Result of Experiment 1. Number of average matching points between Clusters(using RBF kernel) and methionine cleavage label groups.

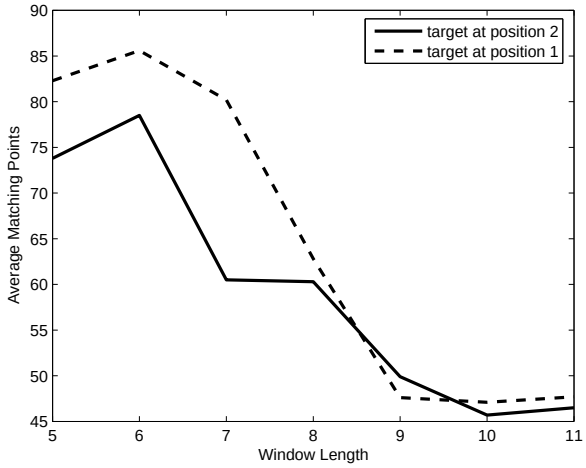

Figure 6: Result of Experiment 1. Number of average matching points between Clusters (using RBF kernel) and acetylation label groups.

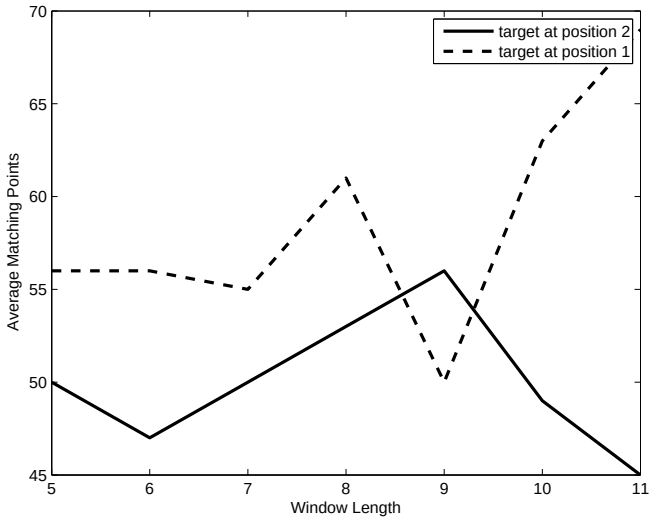

Figure 7: Result of Experiment 3. Number of average matching points between Clusters (using Blosum62 derived kernel) and acetylation label groups.