The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

The Prediction of Protein Post-translational Modification Sites

Naiyang Deng[1] Ling-Yun Wu[2] Xiaobo Wang[1]

[1]College of Science, China Agricultural University, Beijing 100083, China
[2]Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

The post-translational modification (PTM) of proteins is a common biological mechanism for regulating protein functions and modulating all aspects of cellular life. Far from being mere "decorations", PTM of a protein can determine its activity state, localization, turnover, and interactions with other proteins. Due to the great importance of PTM for biological function, the high-throughput experimental technologies have been developed. However, they are usually expensive and time-consuming. Therefore, the computational method to predict PTM sites is desired.

Since many PTM sites already have been verified by experiments, the prediction problem can be transferred into a binary classification problem. Our approach is to solve this classification problem by support vector machine (SVM). In fact, considering prediction of PTM sites in an amino acid sequence of a protein, a general computational model is proposed in [1]. The training datasets are constructed in the following ways: The training datasets include a positive sample set and a negative one. Note that a positive sample means an experimentally verified PTM site. So we manually select the substrate PTM proteins with the PTM sites from the specialized databases such as Phospho.ELM, and get their sequence fragments including the corresponding verified sites in central with a fixed length by using the sliding window technology. Then the fragments are transferred into vectors in Euclidian space by some encoding schemes. Every vector is a positive sample. Thus, the positive sample set is constructed. For the negative sample set, we consider the same proteins used above. From its sequence, take its sites except the ones which have been verified to be PTM sites. Then the corresponding fragments are transferred into vectors. Every vector obtained is a negative sample and the negative sample set is constructed.

After getting the training dataset, the SVM with linear kernel function is applied and a decision function is generated. For five PTM sites including the sites of phosphorylation, glycosylation, palmitoylation, acetylation and methylation, the general computational model is tested on the training datasets by the 10-fold cross validation and some independent testing datasets. The results are satisfactory, particularly for phosphorylation and glycosylation sites prediction. However, for the prediction of palmitoylation, acetylation and methylation sites, the results can be improved further by introducing different encoding schemes and different kernels, see [2], [3], and [4] respectively. According to our preliminary experiments, generally speaking, our models are significantly superior to the existing models.

The methods are implemented as several standalone programs with flexible command line options. They are freely and publicly available for PTM researchers. For any type of the above PTM sites, when the user inputs a protein, the corresponding software will show him/her whether this protein has PTM sites and where they are. In addition, the software allows the user to use their own training dataset by providing the corresponding sequence fragments. At last, the parameters such as C and cutoff are allowed to be adjusted by the users although the default values are suggested.

References

[1] Wang, X.B., Wang, Y.C., Tian, X.J., Shao, X. J., Wu, L.Y., Deng, N.Y. (2009) Prediction of post-translational modification sites from sequences with kernel methods. In submission.

[2] Wang, X.B., Wu, L.Y., Wang, Y.C., Deng, N.Y. (2009) Prediction of palmitoylation site using the composition of K-spaced amino acid pairs. *Protein Engineering, Design, and Selection*, in press.

[3] Xu, Y., Wang, X.B., Wu, L.Y., Deng, N.Y. (2009) Prediction of N^{ε}-acetylation on internal Lysine based on sequence information. In submission.

[4] Wang, X.B., Wu, L.Y., Deng, N.Y. (2009) Prediction of methylation site using the composition of K-spaced amino acid pairs. In submission.