論文の内容の要旨

応用生命工学　　専攻
平成１６年度博士課程　進学
氏　　名　曹　　巍
指導教員名　清水　謙多郎

論文題目　Post-translational lipid modification prediction by using Support Vector Machine
（サポートベクタマシンを用いた翻訳後脂質修飾予測に関する研究）

1. Introduction

Post-translational modifications are vital to protein structure and function for many immature proteins in which the processing of the initial translation products includes elimination of sequences of the protein and attachment of other biochemical functional groups to extend the range of functions of the proteins. In this thesis, two kinds of lipid modification were studied, namely Glycosylphosphatidylinositol (GPI) anchored and Myristoylated lipid modifications. The former one has been spotlighted as an important means for protein post-translational modifications; it has been widely studied since the existence of GPI anchor was accepted in the mid 1980s. In GPI lipid modification, the COOH-terminal signal sequence of precursor proteins is cleaved and GPI moiety added (the new COOH-terminus known as ω-site). The latter one is an irreversible protein post-translational modification found in animals, plants, fungi and viruses. A myristoyl group is covalently attached via an amide bond to the amino group of an N-terminal glycine residue of a nascent polypeptide catalyzed by the N-myristoyltransferase (NMT). However, as currently practiced, identification of these two modifications for proteins faces all sorts of limitations of experimental techniques. With the number of protein sequences uninterruptedly increasing in the existing protein sequence database, methods of identification and prediction of these two modifications of protein sequences

have been receiving more attention in the field of computational biology. It is also this growth of the number of identified protein sequences that makes theoretical analysis and prediction possible.

As far as the status quo of GPI-(like)-anchored proteins prediction is concerned, many research groups have made great efforts. These methods employed could be roughly classified into two categories: (1) ones based on statistical analysis of amino acid composition around ω-site, Big-PI and DGPI, (2) ones based on techniques of machine learning, such as K-nearest-neighbor method employed in PSORT-II and the Kohonen self-organizing map used in GPI-SOM. GPI-SOM, an unsupervised learning method, performs better than DGPI or Big-PI. Although experimental results show importance of hydrophobicity of the COOH-terminus of GPI-(like)-anchored proteins, prediction accuracy of the GPI-(like)-anchored proteins by solely using hydrophobicity scale (~83%) is not as good as being expected. It is also desirable to identify precisely and reliably myristoylated proteins for protein functional annotations in the proteome-wide, especially when experimental measurements are unavailable. There are four prediction schemes based on protein sequences alone (these are available online). However, they still have certain limitations. The first one, PS00008 of PROSITE, has not been updated since 1989, and it is reported that produces a great number of not only false positive but false negative predictions since a small dataset was used to construct it. A taxon-specific scheme advocated by Maurer-Stroh et al. and denoted as NMT predictor gives an ambiguous prediction "twilight zone". Despite Boisson, Giglione and Meinnel (BGM) attempted to modify threshold parameter and improve identification for plant protein sequences, ambiguous prediction results (i.e. twilight zone) are still unsolved. Bologna et al. suggested a rule-based model using average output scores generated from 25 neural networks and Podell et al. put forward a plant-specific hidden markov model. However, the former one needs much more samples to optimize the rule set and the latter one is also taxon-specific. In this thesis, Support Vector Machine (SVM), as a new method, was used to identify GPI-(like)-anchored and myristoylated lipid modifications of proteins. The predictors trained by using SVM show higher performance under 5-fold cross validation test for performance assessment.

2. Methods

As a supervised learning algorithm, SVM, developed by Vapnik and his coworkers has outstanding performance, and it has been successfully applied to many aspects of computational biology. In present work, 1-norm soft margin SVM was employed. With respect to a dichotomic classification problem, the basic idea behind SVM is to map feature vectors by which each sample in a training dataset is represented into a high

dimensional feature space and then construct an optimal separating plane so called hyperplane in this space. Subsequently, a boundary of the margin between positive and negative samples is maximized for giving good generalization properties. The decision boundary is used for classification of unknown samples. To overcome the dimension disaster in computation caused by mapping, kernel functions are proposed for implicit mapping of input data. Here, radial basic function (RBF) was chosen as the kernel function for implicitly mapping input vectors into the high dimensional feature space. To optimize parameters, a regularization parameter C and a parameter γ of RBF function, a population based stochastic optimization technique, a modified version of Particle Swarm Optimization(PSO), was implemented.

Hydrophobicity is an important physico-chemical characteristic of amino acids. For example, hydrophobic residues prefer to be in a non-aqueous environment. A list of values for hydrophobicity measurement of amino acids is called hydrophobicity scale, such as Kyte-Doolittle scale. The hydrophobicity plot is such that a window of a given size slides along the protein sequence from N-terminus to COOH-terminus (one residue at a time in present work), and the mean value within the window is placed in the numerical sequence at each time. For example, given 60 residues taken from COOH-terminus of a protein sequence and window size of 9 residues, the protein sequence descriptor for representing this protein sequence generated by the hydrophobicity plot consists of 52 elements, i.e. a 52-Dimensional vector.

3. GPI-(like)-anchored proteins prediction

The sliding window algorithm, i.e., hydrophobicity plot, was used to obtain a numerical representation of amino acid sequences, and the Kyte-Doolittle scale was employed for delineating the hydrophobic character of 20 standard amino acids (others are set to zero). To eliminate noises in the numerical dataset, feature selection was conducted by using an elongation simulation and a deletion simulation.

Under 5-fold cross validation test, the SVM classifier shows not only accuracy (96%) ; the area under receiver operating characteristic curve (AUC) value of 0.97 is also close to the ideal value 1.00, given that 60 residues were counted starting from COOH-terminus as input sequence length, the window size for hydrophobicity plot was 9 residues and the SVM parameters were optimized by PSO.

4. Myristoylated proteins prediction

The results from myristoylated protein sequence analysis show a motif that has three regions, positions 1-6 for fitting the binding pocket, positions 7-10 for interacting with the surface of N-myristoyltransferase at the mouth of the catalytic cavity and positions 11-17 for containing a hydrophilic linker. The SVM classifiers were trained by using 17

residues which were counted starting from N-terminus as the input length. We used the following properties of 17 residues for training: hydrophobicity (hydrophobicity plot with Kyte-Doolittle scale and the window size of 3 residues) with one of physical property patterns (preference of protein secondary structure, relative stability and geometry property). All prediction accuracies of trained SVM classifiers under 5-fold cross validation test are over 98% and the corresponding AUC values are also over 0.96.

5. Summary

In this work, a new and simple method is presented for the identification of two kinds of protein post-translational modification, GPI-(like)-anchored (occurring on COOH-terminus) and myristoylated (occurring on N-terminus) lipid modifications. By only using hydrophobicity scale, it was reported that prediction accuracy of GPI-anchored proteins is ~83%. The new method improves the prediction accuracy by 13% (i.e. 96%) With respect to myristoylated protein prediction, compared with the previous predictor trained by using neural network method, prediction accuracy increased by 4% and reached 98%. Furthermore, while three of the four previously proposed schemes are taxon-specific, the new method proposed is not limited to be taxon-specific. The computational efficiency and remarkable generalization ability of our method will be helpful for proteomic-wide proteins post-translational modification annotation.

[1] W. Cao, K. Shimizu, Identification of GPI-(like)-Anchored Proteins by Using SVM. Proc. 1st IMSCCS06, 2:711–715, 2006.

[2] W. Cao, K. Sumikoshi, T. Terada, S. Nakamura, K. Shimizu, Insight of the Signal Motif of GPI-(like)-anchored Proteins by Using SVM Proc. BIOCOMP'06, 541-546, Las Vegas, USA 2006.

[3]W. Cao, S. Nakamura, K. Shimizu, Developing SVM classifier for GPI-(like)-anchored proteins prediction. The 17th International Conference on Genome Informatics (GIW), 2006