P6: **Developing a New Bioinformatic Tool for Predicting Caspase3 Cleavage Motif Using Support Vector Machine**

Ruba Sultan[1], Hashem Tamimi and Yaqoub Ashhab

[1]Biotechnology Research Center, Palestine Polytechnic University, P.O-Box 198, Hebron, Palestine

Introduction

Caspase3 is a protein which belongs to cysteine protease family. It has a major role in programmed cell death (apoptosis) as well as other vital cellular processes. As a specific endopeptidase, caspase3 cleaves its substrates after aspartic acid residue 'D'. Although the presence of the amino acid D in the target sequence is a mandatory condition yet it is not enough for recognition and cleavage by this caspase.

Identification of caspase3 novel substrates is crucial to advance our understanding of the biological roles of this important enzyme. Experimental approaches to identify new caspase3 substrates such as site directed mutagenesis and proteomic analysis are laborious and expensive. Therefore, there is an increasing need to develop bioinformatic approaches to discover new caspase3 substrates.

Several computational methods have been developed to predict for cleavage motifs of endopeptidase and caspases in general. Nowadays machine learning approaches offer a more attractive solution for classification [1]. In this study we use Support Vector Machine (SVM) which is a well known powerful machine learning classifier [2]. SVM can find the largest margin between classes of data sets and therefore has a high ability to generalize its decision. The use of SVM has been shown as a potential technique in analyzing biological sequences and designing a strong pattern recognition tools.

Material and methods

We use the already known 150 caspase3 human substrates with identified cleavage positions [3]. The 150 mapped cleavage sites (14 amino acids peptides) of these proteins are used as a positive data set. While the negative data are 150 non-cleaved peptides (14 amino acids) extracted randomly and contained aspartic acid residue 'D' but outside the caspase3 cleaved site. The protein sequences are cleansed in order to extract 14 amino acid motifs that contain an aspartic acid residue ‘D’ at location 9 in addition to 5 amino acids before and 9 after.

Support vector machine (SVM) has been proposed with three different kernel methods. The first method uses “All non contiguous substrings" kernel [4]. This kernel handles the data as strings and produces a kernel matrix that represented the data in high dimentional space. This kernel compares every two stringS by counting their common substrings. The overall similarity is proportional to the number of common substrings. The second method uses a more general string kernel known as "Gap weighted subsequences kernel" [5]. This kernel takes into consideration that two strings are similar with tolerance to the presence of gabs between them. The third method uses four physico-chemical properties of each amino acid; Polarizability, Chou-fasman parameter of coil conformation, Residue accessible surface area in folded protein and amino acid size [6]. The values for each amino acid obtained from each physico-chemical property index are used as an input vector for the support vector machine. To allow data linearization in the third method, two numerical kernels is used; polynomial kernel and linear kernel.

Experiments

By using LIBSVM [7], a toolbox under Matlab environment, the SVM was trained on negative and positive data then it was tested to predict whether new protein sequences contain a potential caspase3 cleavage site or not. In the testing phase, leave-one-out cross validation method was used. In this method the training was applied to 299 sequences, while the single left out sequence was used to test accuracy. This process was repeated 300 times. The

success rate was calculated by counting how many times the predicted label (cleaved or uncleaved) by the SVM is equal to the expected label based on the following equation:

Results

After training and testing SVM, we have achieved the various success rates in the three different approaches. Using the first method (SVM with All-Non-Contiguous kernel) we achieved a success rate of 79.5% .While the success rate in the second method (SVM with Gap-Weighted kernel) was 80.2%. The third method (SVM with physico-chemical properties) has achieved a success rate of 71%.

SVM approach shows that predicting cleavage sites in proteins can be done with low cost and high efficiency, while doing this experimentally in the lab is very expensive and time consuming. Our methodology represents a prototypical model that can be easily adapted to other pattern recognition problems in biological sequences. Our approach can be a useful tool to perform screening on large scale proteomic data to discover new caspase3 substrates, which will provide more insight into the molecular mechanisms that regulate apoptosis and enhance our understanding of cancer biology and treatment.

References

Baldi P, Brunak S. Bioinformatics: The Machine Learning Approach. MIT Press, Cambridge. ISBN 0-262-02506-X(2001).

Wee L, Tan T, Ranganathan S, SVM- based prediction of caspase substrate cleavage sites(2006)

Ayaash M, Ashhab Y, Developing a new bioinformatics tool that predicts caspase -3 cleavage patterns (PICCT conference 2008.)

Darrin P. Lewis, Tony Jebara and William Stafford Noble. Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. Bioinformatics. 22(22):2753-2760, 2006.

Juho Rousu, John Shawe-Taylor, Efficient Computation of Gapped Substring Kernels on Large Alphabets, Journal of Machine Learning Research 6 (2005) 1323–1344

Amino acid Physical-chemical property Database[online]: http://www.rfdn.org/bioinfo/APDbase

Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm