Journal of Computational Intelligence in Bioinformatics
ISSN: 0973-385X Volume 4 Number 1 (2011) pp. 29–35
© Research India Publications
http://www.ripublication.com/jcib.htm

Determination of Protein Subcellular Locations Using Support Vector Machines

M.G. Nirmal[1], K. Varun Gopal[2] and P.K. Krishnan Namboori[2]

Computational Chemistry Group
[1]Amrita School of Biotechnology,
AMRITA Vishwa Vidyapeetham University, Kollam- 690525, India
[2]Computational Engineering and Networking,
AMRITA Vishwa Vidyapeetham University, Coimbatore-641105, India
E-mail: n_krishnan@cb.amrita.edu

Abstract

Most of the proteins in a eukaryotic cell are synthesized in the cytoplasm. Newly synthesized proteins are targeted to the exact subcellular compartments and perform their biological roles. Thus, computational methods for predicting protein subcellular locations are valuable tools for obtaining functional properties from the amino acid sequence information. The subcellular location of a protein is closely associated to its function. Thus, computational prediction of subcellular locations of a protein from its amino acid sequence details would help in annotation and functional prediction of protein coding genes. A machine learning approach based on support vector machines (SVMs) has been developed. 12 subcellular locations in eukaryotic cells: chloroplast, extracellular medium, Golgi apparatus, lysosome, mitochondrion, nucleus, peroxisome, plasma membrane, cytoplasm, cytoskeleton, endoplasmic reticulum and vacuole were considered. SVM learning algorithm was used to extract sequence features from the training data set of proteins, whose subcellular locations are classified into 12 groups. The validity of using different SVM kernel functions and parameters, and also using diverse sequence properties represented by the compositions of amino acids and amino acid pairs were examined. The Subcellular Locations of the proteins were successfully predicted using the technique and the results attained through 5-fold cross-validation tests showed an enhancement in the prediction accuracy.

Keywords: Subcellular locations, Support Vector Machines, kernel functions.

Introduction

Each and every protein synthesized is generally sent to a particular part of the cell. A major portion of the cell biology is the examination of molecular mechanisms by which proteins are translocated to different parts inside cells or secreted from cells [1]. Eukaryotic cell has a membrane that envelops the cell, separates its interior from its environment, regulates what moves in and out (selectively permeable), and maintains the electric potential of the cell [2]. Inside the membrane, a salty cytoplasm takes up most of the cell volume. All cells possess DNA, the hereditary material of genes, and RNA, containing the information necessary to build various proteins such as enzymes, the primary machinery of the cell [3]. And this protein is transferred to various subcellular location based on its function in the cell [4]. Most of the proteins are synthesized in the ribosomes of the cells. The process is generally known as protein translation [5]. Some proteins which have to be transported to the membranes, membrane proteins, are initially transported into the endoplasmic reticulum (ER) after synthesis and further modification in the Golgi apparatus.

From the Golgi apparatus, membrane proteins move to the plasma membrane, to other subcellular organelles or can be secreted from the cell. The ER and Golgi apparatus can be considered as the membrane protein synthesis organelles and the membrane protein processing organelles, respectively. There is a constant or a semi-constant flow of proteins through these organelles [6]. ER and Golgi-associated proteins correlate with other proteins but they remain in their particular organelle. The motor proteins present in the cell transports membrane protein-containing vesicles along cytoskeletal regions to distant parts (Figure 1).

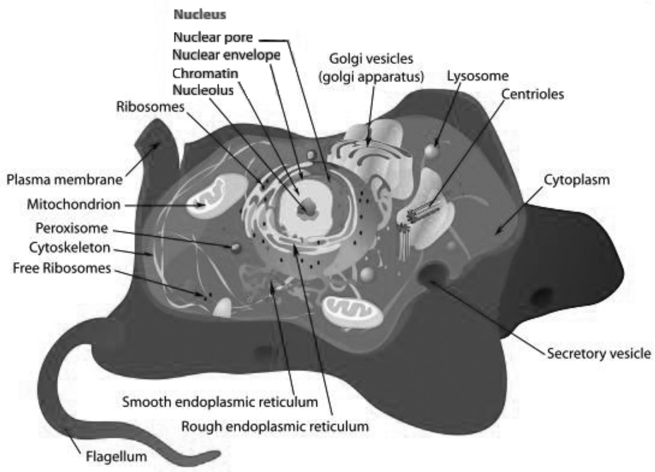

Figure 1: A typical Eukaryotic cell representation.

Some proteins that are synthesized in the cytoplasm contain structural features that help them to move to mitochondria or the nucleus [7]. Some mitochondrial proteins are made inside mitochondria and are coded for by mitochondrial DNA. In plants, chloroplasts also synthesize some cell proteins [8, 9]. Computational methods for predicting protein subcellular locations are valuable tools for obtaining functional clues from the amino acid sequence information [10, 11, 12].

SVMs are a new generation of machine learning algorithms, which is gaining popularity in the analysis of biological problems such as gene, cell, protein and tissue classifications from microarray expression data, protein fold identification and protein secondary structure prediction, as well as the protein localization prediction [13, 14]. Here in this research work an SVM learning algorithm is used to extract sequence features from the training data set of proteins, whose subcellular locations are classified into 12 groups. Specifically, we examine the validity of using different SVM kernel functions and parameters, and also using different sequence properties represented by the compositions of amino acids and amino acid pairs.

Theoretical Background

The aim of machine learning is to build systems that can adapt to their environment and to learn from experience. Support Vector Machines (SVMs) probably represent the greatest known paradigm of this class of algorithms. The general class of algorithms following from this procedure is known as ‘kernel methods’ or ‘kernel machines’. They utilize the mathematical techniques mentioned earlier in order to attain maximum flexibility, generality and performance, in terms of both generalization and computational cost.

SVM construct a hyper plane that separates two classes which extends to multiclass problems [15]. Support vector machines map input vectors to a higher dimensional space where a maximal separating hyper plane is constructed. Two parallel hyper planes are constructed on each side of the hyper plane that separates the data. The separating hyper plane is the hyper plane that maximizes the distance between the two parallel hyper planes. An assumption is made that the distance between these parallel hyper planes the better the generalization error of the classifier will be. A SVM algorithm tries achieving maximum separation between the classes as shown in Figure 2.

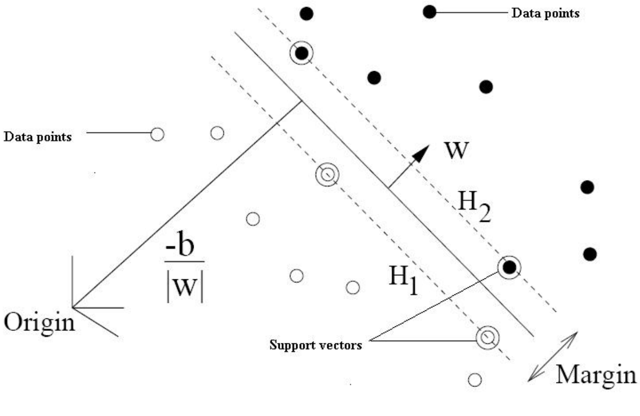

Figure 2: Mapping of input vectors by SVM.

Separating the classes with large margin minimizes a bound on expected generalization error. By ‘minimizing generalization error’, it means that when new data points with unknown class values arrive for classification, the chance of making error in the prediction based on learned classifier or the hyper plane should be minimum. The two planes parallel to the classifier and which passes through one or more points in the datasets is called bounding planes, including the planes which maximizes this margin [16].

SVMs can be trained with a parameter called complexity parameter(C), also known as capacity parameter, for the purpose of regulating overfitting. Choosing a decision boundary, that is extremely partial towards the training set and does not generalize well is called overfitting. For efficient classification, it is mandatory to choose the optimum complexity parameter. It determines the trade-off between choosing large-margin classifier and the amount by which wrongly classified samples are tolerated [17]. A larger C value means that more priority is given to minimize the amount of misclassification than to finding a wide margin model. There are many other parameters associated with the kernel, apart from the C value.

Computational Methodology

All protein sequences were collected from SWISS-PROT database [18]. Eukaryotic proteins with exact sub cellular locations based on organism classification (OC) were identified from the comments and notes in the sequence information data which were experimentally proved. The proteins having multiple sub cellular location were not included in this dataset.

Protein sequence from 12 different sub cellular locations: nuclear, plasma membrane, cytoplasm, extracellular, mitochondria, chloroplast, peroxisomal, endoplasmic reticulum, lysosomal, vacuolar, cytoskeleton and golgi apparatus were collected. Sequence with high degree of similarity were identified by all to all sequence similarity search using ALIGN. The sequences having similarity higher than 80% were eliminated.

The total number of proteins in the final data set was 7589 for the 12 subcellular locations. The data set without restriction of organisms were constructed. The number of different organisms in the data set was 709. The top ranking five were 1027 yeast proteins, 1006 human proteins, 592 mouse proteins, 570 rat proteins and 309 worm proteins.

As there were 12 different classes in the datasets, *SVM multiclass* package were used. *SVM multiclass* is an implementation of the multi-class Support Vector Machine. The algorithm is based on Structural SVMs and it is an instance of *SVMstruct*. For linear kernels, *SVM multiclass V2.12* is very fast and runtime scales linearly with the number of training examples. Non-linear kernels are not really supported in the case of multi class SVMs. *SVMmulticlass* consists of a learning module (svm_multiclass_learn) and a classification module (svm_multiclass_classify). The classification module can be used to apply the learned model to classify new examples. As each protein sequence is made of amino acids, the feature values of each data set were selected as the number of amino acids, number of dipeptides and also the number of tripeptides present in the protein sequence. Each element in the amino acid composition feature Vector denotes the presence frequency of an amino acid. The representation of amino acid order information cannot observe from the feature vector. Thus our technique considers the amino acid order information along with sequence as well as the global information about amino acid sequences. So they are represented as features 1- 20.

Dipeptide composition and tripeptide composition were used to encapsulate the global information about each protein sequence, which gives a fixed pattern length of 400(20*20) in the case of dipeptides and 8000(20*20*20) for tripeptides. These representations encompassed the information of amino acid composition along with local order of amino acids. The feature values for each data set were extracted from each sequence. As a result altogether 8420 features were extracted from the sequences.

The prediction performance was examined by the 5-fold cross validation test, in which the data set of 7589 proteins for the 12 subcellular locations was divided into five subsets of roughly equal size. This means that the data was divided into training and test data in five different ways. After training the SVMs with a group of four subsets, the performance of the SVMs was tested using the fifth subset. This process is repeated for five times so that each and every subset is once used as the test data and the results are evaluated. The machine is trained by varying the C value – trade-off between training error and margin value on different data sets and prediction accuracy of the results are compared.

Results and Discussion

The entire work was divided into three phases. In the first phase, five datasets were created consisting of percentage composition of each amino acid (20 features for each instance of protein sequence). This dataset was then subjected to SVM using linear classifier with different C values. After analyzing the results in the first phase, an accuracy of about 56 % was obtained. To improve the accuracy the number of features were increased. The percentage composition of dipeptide was also included along with percentage composition of amino acid. This set was considered in the second phase. In the second phase, the dataset was created including the percentage composition of dipeptide along with percentage composition of each aminoacid (420 features for each instance of protein sequence).

After analyzing the result in the second phase, an accuracy of about 58 % was obtained. To improve the accuracy the number of features were again increased. The percentage composition of tripeptide was also included along with percentage composition of amino acid and dipeptides. This dataset was considered in the third phase. In this dataset, each instance of protein sequence had 8420 features. The results in the third phase were analyzed and an accuracy of about 86% was obtained. It has been found out as the number of features increased the prediction accuracy has also been increased. So it is assumed that the feature tripeptide composition was found to be playing a major role in classifying the proteins based on location.

The accuracy of each test set data for various values of C values are shown in

Table 1. Each cell in the table shows the value of accuracy for a particular value of C in a particular data set. From the table, it is clear that the maximum value of prediction accuracy is obtained from DataSet 2 for c=35 with a maximum prediction accuracy of 86.5%.

Table 1: Prediction Accuracy after 5-Fold Cross Validation in the Third Phase (%).

c-value	*DataSet 1*	*DataSet 2*	*DataSet 3*	*DataSet 4*	*DataSet 5*
1	72.30	73.75	68.60	63.20	73.95
2	75.15	76.60	75.50	74.95	75.25
5	74.95	76.70	75.65	75.65	75.75
10	75.50	77.00	76.50	76.80	75.75
15	78.55	79.65	80.80	79.80	77.00
20	79.40	79.35	81.60	79.05	77.75
25	80.55	81.10	82.70	81.80	78.75
30	85.60	84.10	82.65	82.00	80 .65
35	81.50	86.50	84.85	80.25	78.15
40	75.30	76.10	76.70	76.70	76.05
50	67.40	65.85	66.70	66.35	66.05
100	65.50	64.45	61.30	62.90	63.95
200	53.60	52.80	50.35	50.30	50.60

Conclusion

Machine learning technique helps in making prediction of subcellular locations of the proteins effectively. SVM, which is considered as a common bioinformatics prediction tool gives 86% prediction accuracy for the data set. SVM learning algorithm was also used to extract sequence features from the training data set of proteins, whose subcellular locations are classified into 12 groups. The validity of using different SVM kernel functions and parameters, and also using different sequence features characterized by the compositions of amino acids and amino acid pairs was examined. The Subcellular Locations of the proteins were successfully predicted. Different sets of SVMs were trained to predict the class of a given protein depending upon its amino acids and amino acid pairs. Results attained through 5-fold cross-validation tests showed a progress in the prediction accuracy of 86%.

References

[1] Satir, P; Christensen, ST; Søren T. Christensen. "Structure and function of mammalian cilia". Histochemistry and Cell Biology (Springer Berlin / Heidelberg) 129 (6): 687–693, 2008

[2] Alberts B, Johnson A, Lewis J. et al. Molecular Biology of the Cell, 4e. Garland Science. 2002

[3] Orgel LE. "The origin of life--a review of facts and speculations". Trends Biochem Sci 23 (12): 491–5, 1998

[4] Nelson, David L.; Cox, Michael M. Lehninger Principles of Biochemistry (4th ed.). New York: W.H. Freeman, 2005

[5] Lynsey Peterson, "Mastering the Parts of a Cell". Lesson Planet., 2010.

[6] Wilson, Edmund B. The cell in Development and Inheritance (second ed.). New York: The Macmillan Company, 1900.

[7] The universal nature of biochemistry, by Norman R. Pace, PNAS, January 30, 2001, vol. 98 , no. 3, 805-808

[8] Albert Frey-Wyssling: Concerning the concept "Organelle". Experientia 34, 547, 1978.

[9] Cormack, Introduction to Histology, Lippincott, 1984

[10] Cedano, J., Aloy, P., P'erez-Pons, J. A. & Querol, E. Relation between amino acid composition and cellular location of proteins. J. Mol. Biol 266, 594-600, 1997

[11] Kuo-Chen Chou, Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics (Erratum: ibid., 2001, Vol.44, 60) 43, 246-255, 2001

[12] Mundra, P., Kumar, M., Kumar, K. K., Jayaraman, V. K. & Kulkarni, B. D. Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. Pattern Recognition Letters 28, 1610-1615, 2007

[13] Du, P., Cao, S. & Li, Y. SubChlo: predicting protein subchloroplast locations with pseudo-amino acid composition and the evidence-theoretic K-nearest neighbor (ET-KNN) algorithm. Journal of Theoretical Biolology 261, 330-335, 2009

[14] Li, F. M. & Li, Q. Z. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein & Peptide Letters 15, 612-616, 2008

[15] Ovidiu Ivanciuc, Applications of Support Vector Machines in Chemistry, In: Reviews in Computational Chemistry, Volume 23, pp. 291–400, 2007

[16] Thorsten Joachims, Learning to Classify Text Using Support Vector Machines. Dissertation, Kluwer, 2002.

[17] David Meyer, Friedrich Leisch, and Kurt Hornik. The support vector machine under test. Neurocomputing 55(1-2): 169-186, 2003.

[18] Bairoch,A. and Apweiler,R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res., 28, 45–48, 2000.