Predicting Protein Subcellular Localization using PsePSSM and Support Vector Machines

Eric Y.T. Juan, J.H. Jhang, W.J. Li

Department of Computer Science and Engineering
National Taiwan Ocean University, Taiwan

Abstract

The prediction of protein subcellular localization (PSL) has been an important field of research. Many prediction systems nowadays have been developed that support the need. Most of these systems however focus on the development of new methods to describe a protein, which in turn can increase the prediction performance. In this paper,we propose a novel prediction system, an evolutionary algorithm based support vector machine for the prediction of PSL (ESVM-PSL) which aims to increase the prediction performance by optimizing SVMs. We apply ESVM-PSL to a set of proteins with jackknife validation. The prediction accuracy of SVMs is effectively increased from 48.9% to 67.5%. Our proposed method is also competitive with the previous systems in terms of prediction accuracy.

Keywords: subcellular localization, evolutionary, support vector machines (SVMs), jackknife validation

1. Introduction

During the process of cell activation, many corresponding proteins will target to correct subcellular sites and play their biological role. Since the localization is vital to the functionality of the cell, the different targeted location with intended subcellular site may lead to different genetic effects, such as cancer and other genetic disease [1]. Therefore, the prediction of PSL becomes crucial for obtaining functional clues from proteins.

Although, for the time being, many biotechnology tools, such as microscope and dispersive x-ray system may reveal to us the subcellular location of certain proteins and its quaternary structure [2] and furthermore, leads to the understanding of its functionality and characteristic. However, these methods are often expensive and time consuming. Therefore, it is a central task to find an economic, efficient and reliable protein analysis research method. During the last decade, with the advancing of biotechnology, a bountiful protein data was extracted and many protein databases were established. Since then, the computational prediction of PSL has become a vibrant field of study [3, 5, 6, 7, 9, 12], especially in recent years. The achievement of computational analysis provides a profound help to problems, such as drug design [13, 14].

Since amino acid composition (AAC) only requires the composition percentage of primary structure sequence, the retrieval of AAC of proteins becomes much easier than retrieving gene ontology data. Its result to AAC only computes the percentage of sequence composition. Though the information other than AAC is exponentially accumulated, as most of the primary structure of protein sequence stays unchanging. AAC still has a certain discrimination power.

Many researches have been made to analyze

protein composition and classify protein. These methods can be classified into two major categories: composition based [3, 5] and similarity based [6]. OET-kNN [12] is one of the well performed tools. It combines multi-kNN classifier and PseAA preprocessing technique. It is quite a breakthrough when compared with former classifiers [18, 19, 12].

In this paper, we build a system, ESVM-PSL, which combines support vector machines (SVMs) [20, 21] and evolutionary algorithm with a pre-process technique PsePSSM [6], in order to classify proteins more effectively.

2. Method

This section will introduce the pre-process method: PsePSSM [6]. PsePSSM will generate a feature vector, which will be inputted into SVMs that has been optimized by evolutionary algorithm, and thus predicts PSL. Through the procedure we will construct our prediction system.

2.1. Encode Proteins to feature vectors

PsePSSM employed in the study is proposed by K. C. Chou [6]. PsePSSM takes a FASTA format protein data as input and generates feature vectors with sequencial order information and protein composition. These feature vectors generated by PsePSSM will be used in classifiers for training and making predictions. PsePSSM is generated by translating a Position Specific Scoring Matrix (PSSM), which needs to blast the protein FASTA file against the SwissProt database (http://www.ebi.ac.uk/swissprot/) for constructing through PSI-BLAST [22]. PSI-BLAST is considered the best protein sequence similarity searching algorithm so far. It supports multiple iterations and allows E-value threshold setting to filter out the protein sequences with lower similarities. In this paper, we set the parameter of PSI-BLAST with 3 iterations, E-value equal to 0.001 and BLOSUM62 for scoring matrix. The constructed PSSM format is shown in formula 1.

$$M_{PSSM} = \begin{bmatrix} E_{1,1} & E_{1,2} & \cdots & E_{1,20} \\ E_{2,1} & E_{2,2} & \cdots & E_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ E_{i,1} & E_{i,2} & \cdots & E_{i,20} \\ \vdots & \vdots & \vdots & \vdots \\ E_{L,1} & E_{L,2} & \cdots & E_{L,20} \end{bmatrix} \quad (1)$$

This size of the matrix, M_{PSSM} , is L* 20, in which L is the length of the amino acid sequence. In the matrix, $E_{i,j}$ indicates the probability of ith posited amino acid against 20 types of amino acid assigned by BLOSUM62 within the i^{th} row and j^{th} column. Next, we use PsePSSM [6] to transform M_{PSSM} of a single protein into a feature vector $P^{\alpha}_{PsePSSM}$, as formula 2 shown.

$$P^{\alpha}_{PsePSSM} = [T', H'] \quad (2)$$

$$T' = \left[\overline{T}_1, \cdots, \overline{T}_{20}\right] \quad (3)$$

$$H' = [H^{\alpha}_1, \cdots, H^{\alpha}_{20}] \quad (4)$$

$$\overline{T_j} = \frac{1}{L}\sum_{i=1}^{L} T_{i,j} \quad (5)$$

$T_{i,j}$, as formula 5 shows, is the element, within i^{th} row and j^{th} column, of matrix M_T normalized from M_{PSSM}. In M_T, the average score of ith amino acid against 20 type amino acid equals to 0. The higher the score of the amino acid to a specific type, the higher the mutation rate to the certain type, or vice versa. The normalization method is represented as formula 6 shown.

$$T_{i,j} = \frac{E_{i,j} - mean_i}{STD_i} \quad (6)$$

$$mean_i = \frac{1}{20}\sum_{k=1}^{20} E_{i,k} \quad (7)$$

$$STD_i = \sqrt{\frac{\sum_{u=1}^{20}(E_{i,u} - mean_i)}{20}} \quad (8)$$

$$H_j^{\alpha} = \frac{1}{L-\alpha} \sum_{i=1}^{L-\alpha} [T_{i,j} - T_{(i+\alpha),j}]^2 \qquad (9)$$

In formula 6, $E_{i,j}$ is the element of M_{PSSM} generated from PSI-BLAST (as formula 1 shown). In formula 4, H_j^{α} stands for the relationship of an amino acid and the α^{th} amino acid afterward. The notation is written in formula 9. H_j^{α} provides us a representation of the hidden sequencial order information in real number.

PsePSSM transforms amino acid sequence into a 40-dimension feature vector $P_{PsePSSM}^{\alpha}$. This feature vector effectively records the composition of a single protein without losing the implicit information of sequence order.

2.2. Support Vector Machines

In this paper, we employ support vector machines (SVMs) as the classfier. SVMs is a statistic based and effective supervised learning classification algorithm [21].

Standard SVMs is a binary classifier. When it comes to multiple class problem, two major extended SVMs are developed: one-against-rest and one-against-one. Since the prediction of PSL is a multi-class problem, we employ a multi-label and well-known classifier, LIBSVM [20, 24, 25], with one-against-one classification. LIBSVM can use different kernel types and class penalty parameters. Thus, a specific protein localization prediction can be manually adjusted to provide a higher accuracy.

2.3. Optimize SVM by evolutionary algorithm

Evolutionary algorithm [26] is a heuristic algorithm based on the simulation of nature evolution. Evolutionary computation employs the concept of reproduction, mutation, recombination and selection. By following Darwin's theory of evolution, survival of the fittest, the fitness of the species gradually increases to fit the environment. When it comes to a problem, we can treat the solution to the problem as an individual and define a fitness function as the environment. Through the concept of evolution, less fitted individual will be eliminated and the fitness of the whole species gradually increases. In other words, the quality of solution increases and further optimized to an approximation solution to the problem. Therefore, we can use an evolutionary algorithm to optimize LIBSVM. In this work, we employ the Particle Swarm Optimization (PSO) as the optimization tool for LIBSVM.

3. Results

In this section, we will use a set of 714 protein sequences as our experimental data. Since multiple subcellular localization protein data have been filtered, protein sequences are non-redundant. The following section will introduce the details of these data.

3.1. Data sets

In PSL prediction problem, Nuc-Ploc [27] uses two protein descriptors: PseAA and PsePSSM, to design a highly accurate system. For a fair comparison, we also use the dataset, Nuclear Protein 714 (NP714), to predict 9 subcellular localization sites to validate the prediction accuracy of our ESVM-PSL. The subcellular localization sites and the amounts of NP714 are listed in Table 1.

Subcellular location	Number of protein
Chromatin	99
Heterochromatin	22
Nuclear envelope	61
Nuclear matrix	29
Nuclear pore complex	79
Nuclear speckle	67
Nucleolus	307
Nucleoplasm	37
Nuclear PML body	13
Overall	714

Table 1: The subcellular localizations and numbers of 714 nuclear proteins.

Prediction Systems	Hit Rate	Overall Accuarcy
AAC on ProtLoc	261/714	36.6%
AAC on SVM	349/714	48.9%
PsePSSM on LIBSVM	349/714	48.9%
PseAA on OET-KNN	397/714	55.6%
PseAA & PsePSSM on Ensemble-classifier	481/714	67.4%
PsePSSM on ESVM-PSL	482/714	67.5%

Table 2: Accuracy of subcellular location prediction on 714 nuclear proteins.

3.2. Performance

Table 2 shows the experimental results on 714 nuclear proteins. Jackknife test method is used to estimate the accuracy of several prediction systems. A simple discrete representation of protein samples is based on AAC which is used in both ProtLoc and SVM. The prediction accuracy is higher when SVM is adopted in the prediction system. The performance remains the same when we use LIBSVM combined with another representation of protein samples, PsePSSM.

A substantial improvement can be achieved after we optimized LIBSVM using PSO in our ESVM-PSL. We successfully raise the prediction accuracy of LIBSVM from 48.9% to 67.5%. Moreover, our proposed ESVM-PSL outperforms OET-kNN: 55.6% [12] and is competitive with Ensemble classifier: 67.4% [27] in terms of prediction accuracy.

A compelling distinction of our system is that ESVM-PSL uses only one discrete descriptor, PsePSSM, and one classifier, LIBSVM, while Ensemble classifier uses two representations, PseAA and PsePSSM, and fuses 1000 basic individual classifiers.

4. Discussion

Nowadays, a number of methods have been developed that support prediction of PSL [7, 9, 3, 8, 5]. Our experiment results reveal that the accuracy of prediction systems can be further optimized according to the characteristics of data with the aid of evolutionary optimization. With the usage of evolutionary algorithm, ESVM-PSL outperforms OET-kNN [12] and is also competitive with Ensemble classifier [27] in prediction accuracy. Moreover, in comparison with standard SVMs, the prediction accuracy of our proposed method substantially increases.

5. Acknowledgment

The research is supported by National Taiwan Ocean University through grant CMBB-97-P-B-529002H2.

References

[1] R. D. Phair and T. Misteli. High mobility of proteins in the mammalian cell nucleus. *Nature*, 404:604–609, April 2000.

[2] R. Yuste. Fluorescence microscopy today. *Nature Methods*, 2:902–904, 2005.

[3] H.B. Shen and K.C. Chou. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. *Analytical Biochemistry*, 373:386–388, 2008.

[4] K.Nakai and P. Horton. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci*, 24:34–35, 1999.

[5] J. Y. Yang M. Q. Yang T. Habib, C. Zhang and Y. Deng. Supervised learning method for the prediction of subcellular localization of proteins using amino acid and amino acid pair composition. *BMC Genomics*, 9:S16, 2008.

[6] K. C. Chou and H. B. Shen. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating

evolution information through Pse-PSSM. *Biochem Biophys Res Comm*, 360:339–345, 2007.

[7] A. Lo J. K. Hwang T. Y Sung C.Y. Su, H. S.Chiu and W. L. Hsu. Protein subcellular localization prediction based on compartment-specific features and structure conservation. *BMC Bioinformatics*, 8:330, 2007.

[8] B. F. Buxton J. J. Ward, L. J. McGuffin and D. T. Jones. Secondary structure prediction with support vector machines. *Bioinformatics*, 19:1650–1655, 2003.

[9] A. Krishnan J. Wang, W. K. Sung and K. B. Li. Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines. *BMC Bioinformatics*, 6:174, 2005.

[10] K. Wang M. Ester G. E. Tusnady I. Simon S. Hua K. deFays C. Lambert K. Nakai J. L. Gardy, C. Spencer and F. S. L. Brinkman. PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Research*, 31:3613–17, 2003.

[11] F. Chen S. Rey C. J. Walsh M. Ester J. L. Gardy, M. R. Laird and F. S. L. Brinkman. PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, 21:617–623, 2005.

[12] H. B. Shen and K. C. Chou. Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. *Biochem. Biophys. Res. Comm.*, 337:752–56, 2005.

[13] J. W. Yang G. Lubec, L. Afjehi-Sadat and J. P. P. John. Searching for hypothetical proteins: Theory and practice based upon original data and literature. *Progress in Neurobiology*, 77:90–127, 2005.

[14] K. C. Chou. Structural Bioinformatics and its Impact to Biomedical Science. *Curr Med Chem*, 11:2105–2134, 2004.

[15] K. C. Chou. Prediction of protein cellular attributes using pseudo-amino-acid-composition. *PROTEINS: Structure, Function, and Genetics*, 43:246–255, 2001.

[16] K. C. Chou and H. B. Shen. Large-scale plant protein subcellular location prediction. Journal of Cellular Biochemistry. *Journal of Cellular Biochemistry*, 100:665–678, 2007.

[17] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet.*, 25:25–29, 2000.

[18] S. W. Ho S. F. Hwang W. L. Huang, C. W. Tung and S. Y. Ho. ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC Bioinformatics*, 9:80, 2008.

[19] K. C. Chou and H. B. Shen. Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. *Protein Engineering, Design, and Selection*, 20:39–46, 2007.

[20] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–397, 1995.

[21] S. Kotsiantis. Supervised Machine Learning: A Review of Classification Techniques. *Informatica Journal*, 31:249–268, 2007.

[22] A. A. Schaffer J. Zhang Z. Zhang W. Miller S. F. Altschul, T. L. Madden and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402, 1997.

[23] T.L. Madden S. Shavirin J.L. Spouge Y.I.Wolf E.V. Koonin and S.F. Altschul A.A. Schaffer, L. Aravind. Improving the accuracy of PSI-BLAST protein database searches with composition-based statisticsand other refinements. *Nucleic Acids Res.*, 29:2997–3005, 2001.

[24] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector ma-*

chines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[25] C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[26] J. H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. The MIT Press, 1975.

[27] H. B. Shen and K. C. Chou. Nuc-PLoc : a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. *Protein Engineering, Design & Selection*, 20:561–567, 2007.