Amino Acid Features for Prediction of Protein-Protein Interface Residues with Support Vector Machines

Minh N. Nguyen[1], Jagath C. Rajapakse[1,2], and Kai-Bo Duan[1]

[1] BioInformatics Research Centre, School of Computer Engineering, Nanyang Technological University, Singapore
[2] Singapore-MIT Alliance, Singapore
{nmnguyen,asjagath,askbduan}@ntu.edu.sg

Abstract. Knowledge of protein-protein interaction sites is vital to determine proteins' function and involvement in different pathways. Support Vector Machines (SVM) have been proposed over the recent years to predict protein-protein interface residues, primarily based on single amino acid sequence inputs. We investigate the features of amino acids that can be best used with SVM for predicting residues at protein-protein interfaces. The optimal feature set was derived from investigation into features such as amino acid composition, hydrophobic characters of amino acids, secondary structure propensity of amino acids, accessible surface areas, and evolutionary information generated by PSI-BLAST profiles. Using a backward elimination procedure, amino acid composition, accessible surface areas, and evolutionary information generated by PSI-BLAST profiles gave the best performance. The present approach achieved overall prediction accuracy of 74.2% for 77 individulal proteins collected from the Protein Data Bank, which is better than the previously reported accuracies.

1 Introduction

The knowledge of protein-protein interaction is valuable for understanding mechanisms of diseases of living organisms and for facilitating discovery of new drugs. The identification of interface residues has many applications such as drug design, protein mimetic engineering, elucidation of molecular pathways [1,2], and understanding of disease mechanisms [3]. Proper identification of the residues at interfaces helps guiding of the processes of docking to build the structural models of protein-protein complexes [4].

Many computational techniques are available in the literature to predict protein-protein interface residues based on different characteristics of known protein-protein interaction sites [4,5,6,7,8,9]. Neural networks use residues in a local neighborhood or a window as inputs to predict protein-protein interface residues at a particular location of an amino acid sequence by finding an appropriate non-linear mapping. Protein-protein interaction sites are predicted from a neural network with sequence profiles of neighboring residues and solvent

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 187–196, 2007.

exposures as inputs [6]. Fariselli *et al.* implemented a neural network method for predicting protein-protein interaction sites based on the information of evolutionary conservation and surface disposition [7]. Ofran and Rost proposed a neural network to predict interaction sites from local sequence information [8]. Chen and Zhou introduced a consensus neural network that combines predictions from multiple models with different levels of accuracy and coverage [4]. Input information derived from single sequences has been used by support vector machine (SVM) for predicting protein-protein interface residues [9]. Recently, Yan *et al.* proposed a two-stage classifier consisting of an SVM and a Bayesian network classifier that identifies interface residues primarily on the basis of sequence information [5].

Despite the existence of many approaches, the current success rates of existing approaches to protein-protein interface residue prediction are insufficient for practical applications; further improvement of the accuracy is necessary. Most of the existing techniques use conventional orthogonal encoding or information derived directly from amino acid sequences as inputs to predict protein-protein interaction residues. The biochemical properties of each amino acid residue have not exploited systematically in the prediction.

In this paper, we investigate into various features of amino acids previously used by various researchers and select the optimal combination to predict the residues at protein-protein interactions by using SVM which has a strong foundation in statistical learning theory [10]; its generalization capability is optimal compared to other statistical or machine learning methods in solving many biological problems [11]. SVMs are powerful and generally applicable tools in predicting protein secondary structure [12,13,14], relative solvent accessibility [15], accessible surface areas of amino acids [16], and cancer classification [17,18]. Apart from amino acid composition, we investigate hydrophobic characters of amino acids [19], secondary structure propensity of amino acids [20], accessible surface areas [16], and the evolutionary information generated by PSI-BLAST profiles into an encoding schema. We begin with all known features and reduce using a backward elimination procedure to find the optimal features.

The present approach achieves a substantial improvement of the prediction accuracy from 2.2% to 8.2% on a dataset of 77 individulal proteins collected from the Protein Data Bank compared to the previously reported best prediction accuracies with SVM methods [5,9]. Results of our experiments with the proposed encoding schema confirmed that the accessible surface areas and the evolutionary information of amino acids along the sequences can detect different sequence features to enhance the accuracy of protein-protein interaction residue prediction.

2 Feature Selection

The protein sequences are converted into feature vectors constructed from amino acid composition, hydrophobic characters of amino acids, secondary structure propensity of amino acids, accessible surface areas, and the evolutionary information