Improving Protein Localization Prediction Using Amino Acid Group Based Physichemical Encoding

Jianjun Hu* and Fan Zhang

Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, 29208, USA
{jianjunh,zhangf}@cec.sc.edu

Abstract. Computational prediction of protein localization is one common way to characterize the functions of newly sequenced proteins. Sequence features such as amino acid (AA) composition have been widely used for subcellular localization prediction due to their simplicity while suffering from low coverage and low prediction accuracy. We present a physichemical encoding method that maps protein sequences into feature vectors composed of the locations and lengths of amino acid groups (AAGs) with similar physichemical properties. This high-level modular representation of protein sequences overcomes the shortcoming of losing order information in the commonly used AA composition and AA pair composition encoding. When applied with SVM classifiers, we showed that AAG based features are able to achieve higher prediction accuracy (up to 20% improvement) than the widely used AA composition and AA pair composition to differentiate proteins of different localizations. When AAGs and AA composition encoding combined, the prediction accuracy can be further improved thus achieving synergistic effect.

Keywords: Physical encoding, protein subcellular location prediction, AA index, Support Vector Machines, amino acid groups.

1 Introduction

Determination of subcellular locations of a protein experimentally [1;2] or computationally [3-6] can greatly help to infer its function. Due to its simplicity, automated prediction of subcellular localization has been routinely used to annotate protein sequences and dozens of algorithms have been developed [3-6]. These algorithms employ a variety of supervised machine learning techniques including neural networks [7;8], nearest neighbor classifier, Markov models, Bayesian networks [9], expert rules, meta-classifiers [10], and the support vector machines [11-13]. While algorithm variation can tune up the prediction performance, a more critical factor for accurate prediction is to extract effective features for inferring the subcellular location of a protein. A variety of information has been used as features for subcellular prediction as discussed below.

* Correspondence Author.

S. Rajasekaran (Ed.): BICoB 2009, LNBI 5462, pp. 248–258, 2009.