International Journal of Innovative
Computing, Information and Control
Volume **6**, Number **11**, November **2010**

ICIC International ©2010 ISSN 1349-4198
pp. 5093–5103

USING ELMAN NETWORKS ENSEMBLE FOR PROTEIN SUBNUCLEAR LOCATION PREDICTION

Junwei Ma, Wenqi Liu and Hong Gu

Faculty of Electronic Information and Electrical Engineering
Dalian University of Technology
Dalian 116023, P. R. China
junweima@yahoo.com.cn; { wqliu; guhong }@dlut.edu.cn

Received July 2009; revised November 2009

Abstract. *Knowledge of nuclear-protein-localizations plays a very important role in understanding the biochemical processes of the nucleus. With the avalanche of protein sequences generated in the post-genomic era, an automated method is badly needed to annotate the subnuclear locations of numerous newly found nuclear protein sequences in a short time. In this paper, a novel approach is developed for predicting the protein subnuclear locations. Firstly, a powerful ensemble classifier based on the Elman networks is proposed. Secondly, the protein samples are represented by amphiphilic pseudo amino acid (PseAA) composition, which can incorporate a considerable amount of sequence-order effects. Thirdly, six different algorithms are adopted to consider the diversity of the base ensemble, and 18 Elman networks are subsequently obtained by using three different node numbers of neurons in the hidden layers. Lastly, as a demonstration, identifications are performed for 9 subnuclear locations in 714 nuclear proteins. The accuracy rates, obtained in both a re-substitution test and a jackknife test, are significantly higher than those achieved by other classifiers. It is anticipated that the proposed approach may become a useful tool to reduce the huge gap between the number of gene sequences in databases and the number of gene products which have been functionally characterized.*
Keywords: Protein subnuclear location, Elman networks ensemble, Amphiphilic PseAA amino acid composition, Re-substitution test, Jackknife test

1. **Introduction.** Eukaryotic cells consist of two major parts, the nucleus and the cytoplasm. The nucleus is a highly complex organelle that forms a package for cells and their corresponding regulatory factors [1]. It guides life process of cells by performing the following functions: (1) directing cellular reproduction; (2) controlling a cell's differentiation during the development of the organism; and (3) regulating the metabolic activities of the cell.

In addition to the genetic material, a nucleus contains many proteins located at different compartments, which are called subnuclear locations. Information of the subnuclear locations of these proteins is important because it provides useful clues about their functions, also helps understand how and in what kind of microenvironments they interact with each other and with other molecules as well [2]. Furthermore, compartmentalization of cell nucleus is highly related to several nuclear processes, and has potential influence of cancer-related alternatives on gene express. Mis-localized nuclear proteins can lead to human genetic disease and cancer [1]. Therefore, accurately predicting protein subnuclear localization is crucial for understanding genomic regulation and life process of the cell.

Although the subnuclear localization can be determined by conducting various experiments, it is time-consuming and costly to acquire such information alone by experiments.