Original article

Recognition of DNA Splice Junction via Machine Learning Approaches

Chanin Nantasenamat¹, Thanakorn Naenna¹, Chartchalert Isarankura-Na-Ayudhya¹, Virapong Prachayasittikul¹

¹ Department of Clinical Microbiology, Faculty of Medical Technology, Mahidol University, Bangkok 10700
2 Department of Industrial Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom 73170, Thailand


ABSTRACT
Successful recognition of splice junction sites of human DNA sequences was achieved via three machine learning approaches. Both unsupervised (Kohonen’s Self-Organizing Map, KSOM) and supervised (Back-propagation Neural Network, BNN; and Support Vector Machine, SVM) machine learning techniques were used for the classification of sequences from the testing set into one of three categories: transition from exon to intron, transition from intron to exon, and no transition. The dataset used in this study is comprised of 1,424 DNA sequences obtained from the National Center for Bioinformatics Information (NCBI). Performance of the machine learning approaches were assessed by the construction of learning models from 1,000 sequences of the training set and evaluated on the 424 sequences of the testing set that is unknown to the learning model. Each sequence is a window of 32 nucleotides long with regions comprising -15 to +15 nucleotides from the dinucleotide splice site. Since the nucleotides (A, C, G, and T) are represented by four digit binary code (e.g. 0001, 0010, 0100, and 1000) the number of descriptors increased from 32 to 128. The performance of machine learning techniques in order of increasing accuracy are as follows SVM > BNN > KSOM, suggesting that SVM is a robust method in the identification of unknown splice site. Although KSOM gave lower prediction accuracy than the two supervised methods, it is fascinating that it was able to make such prediction based only on knowledge of the input whereas the supervised method requires that the output be known during training. It is expected that the Support Vector Machine method can provide a powerful computational tool for predicting the splice junction sites of uncharacterized DNA.

Keywords: DNA splice junction, machine learning, Kohonen’s Self-Organizing Map, Backpropagation Neural Network, Support Vector Machine

EXCLI Journal 2005;4:114-129