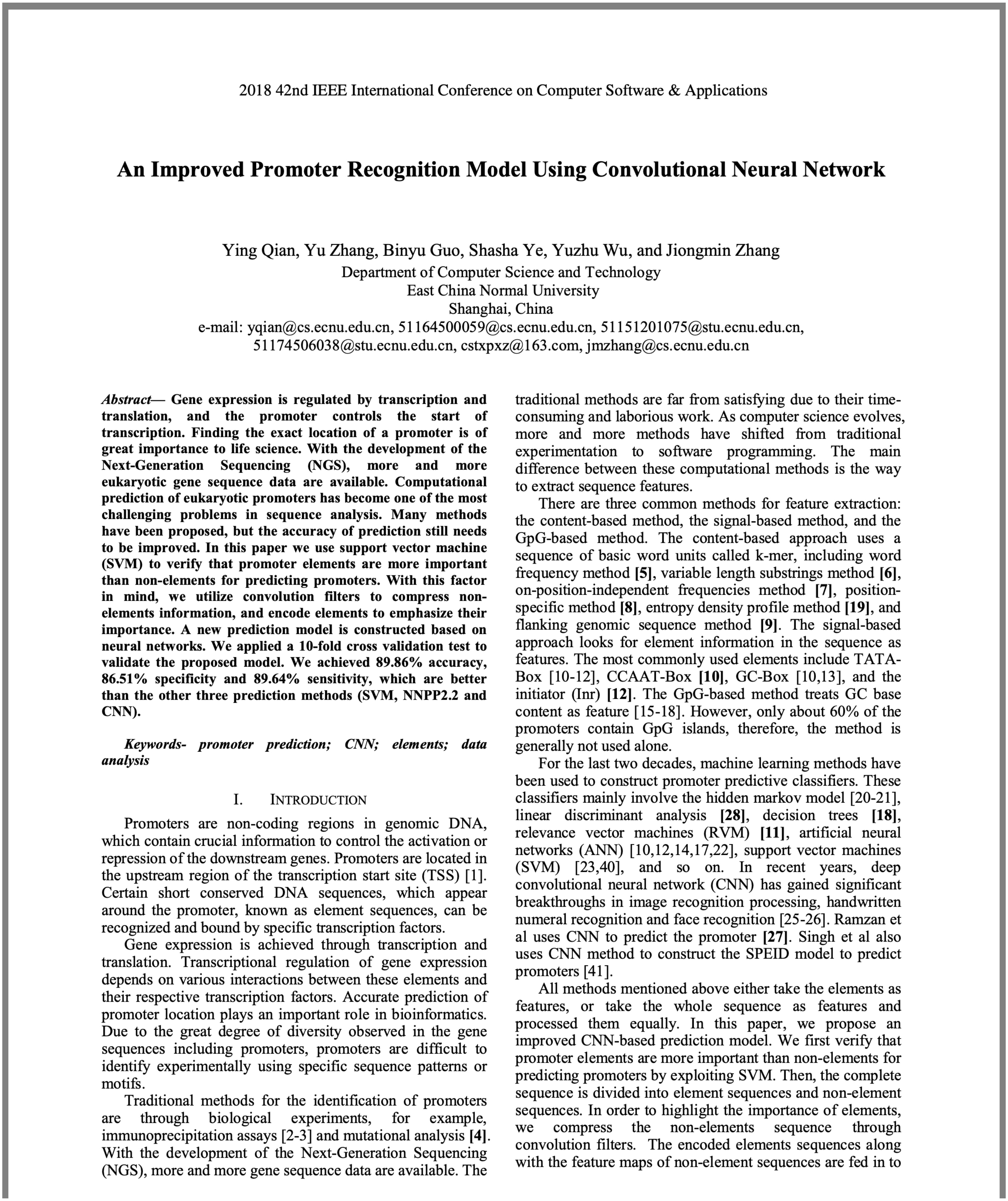

2018 42nd IEEE International Conference on Computer Software & Applications

An Improved Promoter Recognition Model Using Convolutional Neural Network

Ying Qian, Yu Zhang, Binyu Guo, Shasha Ye, Yuzhu Wu, and Jiongmin Zhang
Department of Computer Science and Technology
East China Normal University
Shanghai, China
e-mail: yqian@cs.ecnu.edu.cn, 51164500059@cs.ecnu.edu.cn, 51151201075@stu.ecnu.edu.cn,
51174506038@stu.ecnu.edu.cn, cstxpxz@163.com, jmzhang@cs.ecnu.edu.cn

Abstract— **Gene expression is regulated by transcription and translation, and the promoter controls the start of transcription. Finding the exact location of a promoter is of great importance to life science. With the development of the Next-Generation Sequencing (NGS), more and more eukaryotic gene sequence data are available. Computational prediction of eukaryotic promoters has become one of the most challenging problems in sequence analysis. Many methods have been proposed, but the accuracy of prediction still needs to be improved. In this paper we use support vector machine (SVM) to verify that promoter elements are more important than non-elements for predicting promoters. With this factor in mind, we utilize convolution filters to compress non-elements information, and encode elements to emphasize their importance. A new prediction model is constructed based on neural networks. We applied a 10-fold cross validation test to validate the proposed model. We achieved 89.86% accuracy, 86.51% specificity and 89.64% sensitivity, which are better than the other three prediction methods (SVM, NNPP2.2 and CNN).**

Keywords- promoter prediction; CNN; elements; data analysis

I. INTRODUCTION

Promoters are non-coding regions in genomic DNA, which contain crucial information to control the activation or repression of the downstream genes. Promoters are located in the upstream region of the transcription start site (TSS) [1]. Certain short conserved DNA sequences, which appear around the promoter, known as element sequences, can be recognized and bound by specific transcription factors.

Gene expression is achieved through transcription and translation. Transcriptional regulation of gene expression depends on various interactions between these elements and their respective transcription factors. Accurate prediction of promoter location plays an important role in bioinformatics. Due to the great degree of diversity observed in the gene sequences including promoters, promoters are difficult to identify experimentally using specific sequence patterns or motifs.

Traditional methods for the identification of promoters are through biological experiments, for example, immunoprecipitation assays [2-3] and mutational analysis **[4]**. With the development of the Next-Generation Sequencing (NGS), more and more gene sequence data are available. The traditional methods are far from satisfying due to their time-consuming and laborious work. As computer science evolves, more and more methods have shifted from traditional experimentation to software programming. The main difference between these computational methods is the way to extract sequence features.

There are three common methods for feature extraction: the content-based method, the signal-based method, and the GpG-based method. The content-based approach uses a sequence of basic word units called k-mer, including word frequency method **[5]**, variable length substrings method **[6]**, on-position-independent frequencies method **[7]**, position-specific method **[8]**, entropy density profile method **[19]**, and flanking genomic sequence method **[9]**. The signal-based approach looks for element information in the sequence as features. The most commonly used elements include TATA-Box [10-12], CCAAT-Box **[10]**, GC-Box [10,13], and the initiator (Inr) **[12]**. The GpG-based method treats GC base content as feature [15-18]. However, only about 60% of the promoters contain GpG islands, therefore, the method is generally not used alone.

For the last two decades, machine learning methods have been used to construct promoter predictive classifiers. These classifiers mainly involve the hidden markov model [20-21], linear discriminant analysis **[28]**, decision trees **[18]**, relevance vector machines (RVM) **[11]**, artificial neural networks (ANN) [10,12,14,17,22], support vector machines (SVM) [23,40], and so on. In recent years, deep convolutional neural network (CNN) has gained significant breakthroughs in image recognition processing, handwritten numeral recognition and face recognition [25-26]. Ramzan et al uses CNN to predict the promoter **[27]**. Singh et al also uses CNN method to construct the SPEID model to predict promoters [41].

All methods mentioned above either take the elements as features, or take the whole sequence as features and processed them equally. In this paper, we propose an improved CNN-based prediction model. We first verify that promoter elements are more important than non-elements for predicting promoters by exploiting SVM. Then, the complete sequence is divided into element sequences and non-element sequences. In order to highlight the importance of elements, we compress the non-elements sequence through convolution filters. The encoded elements sequences along with the feature maps of non-element sequences are fed in to