ELSEVIER

Expert Systems with Applications 30 (2006) 73–81

Expert Systems with Applications

www.elsevier.com/locate/eswa

Splice site prediction using support vector machines with a Bayes kernel

Ya Zhang [a], Chao-Hsien Chu [a,*], Yixin Chen [b], Hongyuan Zha [c], Xiang Ji [d]

[a] School of Information Sciences and Technology, 301 K IST Building, Pennsylvania State University, University Park, PA 16802, USA
[b] Department of Computer Science, University of New Orleans, New Orleans, LA 70148, USA
[c] Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802, USA
[d] NEC Laboratories America, Cupertino, CA 95014, USA

Abstract

One of the most important tasks in correctly annotating genes in higher organisms is to accurately locate the DNA splice sites. Although relatively high accuracy has been achieved by existing methods, most of these prediction methods are computationally extensive. Due to the enormous amount of DNA sequences to be processed, the computational speed is an important issue to consider. In this paper, we present a new machine learning method for predicting DNA splice sites, which first applies a Bayes feature mapping (kernel) to project the data into a new feature space and then uses a linear Support Vector Machine (SVM) as a classifier to recognize the true splice sites. The computation time is linear to the number of sequences tested, while the performance is notably improved compared with the Naive Bayes classifier in terms of classification accuracy, precision, and recall. Our classification results are also comparable to the solution quality obtained by the SVMs with polynomial kernels, while the speed of our proposed method is significantly faster. This is a notable improvement in computational modeling considering the huge amount of DNA sequences to be processed.

Keywords: Splice site prediction; SVM; Support vector machines; Bayes classifier; Machine learning; Splice Site Prediction Using Support Vector Machines with Bayes Kernel

1. Introduction

The advances in sequencing technologies have resulted in a large amount of DNA sequence information and therefore a dramatic increase in the size of genetic and genomic databases. The genome sequence information is produced as sequences of base pairs. However, no real knowledge of how the genome works is revealed unless different regions of the genome and their functions are characterized. Therefore, an important goal in bioinformatics is to accurately annotate the genome sequence information within an acceptable timeframe. Many computational efforts have recently been explored for predicting gene structures (Burge & Karlin, 1997) from DNA sequences and aiding the extensive analysis of the genome sequences, including recognizing translation initiation site of genes (Zien, Ratsch, Mika, Scholkopf, Lengauer & Muller, 2000), discovering transcriptional factor binding sites in promoter sequences (Lim, Sim, Chung, & Park, 2003), and identifying DNA splice sites (Jones & Watkins, 2000; Mache & Levi, 2000; Weber, 2001).

Gene expression in eukaryotes starts with the transcription of DNA sequences into mRNA sequences, followed by the processing of pre-mRNAs to mature mRNAs, and then the translation of mRNAs to proteins. Splicing is one of the primary post-processing steps of pre-mRNAs in eukaryotes. During splicing, the introns, the non-coding regions of genes, are removed from the primary transcripts, and the exons, the coding regions, are joined to form a continuous sequence that specifies a functional polypeptide (See Fig. 1 for illustration). The 5′ side of the intron is a donor splice site and 3′ side is an acceptor splice site. As most eukaryotic genes contain introns, many of which interrupt an exon within a codon, an important part of gene prediction in eukaryotes is therefore to predict splice sites.

This paper focuses on the problem of identifying DNA splice sites. Locating splice sites is an interesting problem to address because of the special structure in sequences around splice sites. The residual pairs *GT* and *AG* are often indicative of donor and acceptor splice sites. However, this canonical *GT–AG* rule does not always hold. Thus, it is natural to model the prediction of splice sites as a binary classification problem,

* Corresponding author. : Tel.: +1 814 865 4446; fax: +1 814 865 6426. Current address: Currently on leave at Singapore Management University.
E-mail addresses: yzhang@ist.psu.edu (Y. Zhang), chu@ist.psu.edu (C.-H. Chu), yixin@cs.uno.edu (Y. Chen), zha@cse.psu.edu (H. Zha), xji@sv.nec-labs.com (X. Ji).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.09.052

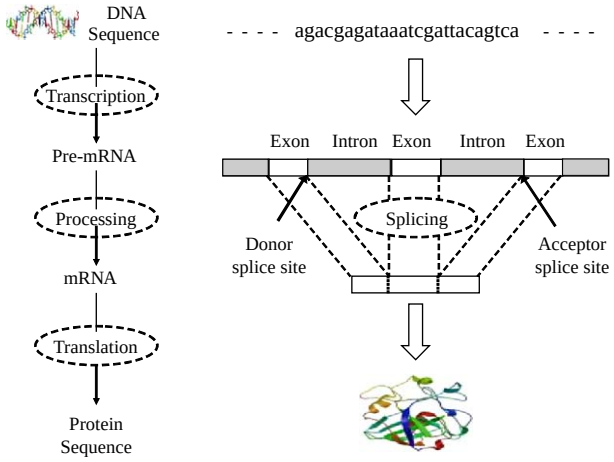

Fig. 1. The process of gene expression.

using DNA sequences with experimentally confirmed splice sites as positive training examples and those DNA sequences with *GT–AG* structure but confirmed *not* to be real splice sites as negative training examples.

Artificial neural networks (Acır & Güzeli, 2004), Bayesian classifiers (Stockwell, 1993), and SVMs (Min & Lee, 2005; Shin, Lee, & Kim, 2005) are important expert systems that have been applied to solve real world problems. Several of these expert systems have been applied to many interesting bioinformatics problems. For example Wang, Kuo, Chen, Hsiao, and Tsai (2005) built a knowledge sharing system for protein families (KSPF) using sequence pattern data mining and knowledge management. In this paper, we focus on the problem of recognizing true splice sites. Table 1 summarizes selected models used in predicting splice sites and their references. Although relatively high accuracy has been achieved with the methods currently available, almost all of the existing methods are computationally very demanding.

Table 1
Machine learning methods for splice sites prediction

Methods/models	Refs.
Statistical methods	
-Logit linear model	Brendel & Kleffe, 1998
-Quadratic discriminant analysis	Zhang & Luo, 2003
-Naïve Bayes classifier	Degroeve, De Baets, Van de Peer, & Rouz, 2002
Decision trees:	
-Maximal dependence decomposition (MDD)	Burge & Karlin, 1997
-MDD with Markov model	Pertea, Lin, & Salzberg, 2001
-C 4.5 induction tree	Patterson, Yasuhara, & Ruzzo, 2002
Artificial neural networks	
-Percepton	Weber, 2001
-Multi-layer Backpropagation	Mache & Levi, 2000; Reese, Eeckman, Kulp, & Haussler, 1997; Sonnenburg, Ratsch, Jagota, & Muller, 2002.
SVMs	
-Linear kernels	Degroeve *et al*., 2002
-Polynomial kernels	Jones & Watkins, 2000; Patterson *et al*., 2002; Zien *et al*., 2000

Consequently, splice site prediction continues to be a major bottleneck in gene annotation.

In this study, we employ a linear SVM, which is computationally less extensive than SVMs with polynomial kernels, to recognize true splice sites. However, the DNA sequence information is given as strings while the SVM classifier can only take numerical inputs. Thus, the very first step is to encode or map the DNA sequences into numbers. A widely used encoding method is sparse encoding, where each letter in the DNA sequence is represented in four bits. But with this encoding method, the sequence data are in general linearly inseparable by SVMs. Instead, a novel mapping/encoding method derived from Bayes' rule is used to project the data into a new feature space where the true splice sites and the false splice sites can then be classified by linear SVMs. An advantage of the Bayes encoding method is that it takes into consideration the natural mutations in the DNA sequences with a probabilistic encoding framework. Experimental results have shown that the performance of our proposed method is comparable to that of SVMs with polynomial kernels in terms of accuracy, precision and recall, while the speed of our method is significantly faster. The computation time is linear to the number of sequences tested, while the performance is notably improved compared to the Naive Bayes classifier in terms of accuracy, precision and recall. Considering the overwhelming amount of DNA sequences that needs to be processed, the increased speed of our method is a very desirable property.

The rest of this paper is organized as follows. In Section 2, we give an introduction to SVMs. In Section 3, the Naive Bayes classifier is explained, and the Bayes feature mapping method is explored. In Section 4, we describe our experiment with splice site prediction and our theoretical analysis of the proposed method. In Section 5, the experiment results are presented. Finally, we give the conclusion in Section 6.

2. Support vector machines

Support Vector Machines (Vapnik, 1998) are powerful pattern recognition techniques that have been successfully applied to many machine learning tasks such as classification (Scholkopf, Burges, & Smola, 1999) and regression (Smola & Scholkopf, 2004). They have outperformed many other machine learning methods such as artificial neural networks and *k*-nearest neighbors and attracted a great deal of attention from the machine learning community because of many needed properties, including good generalization performance, robustness in the presence of noise, ability to deal with high dimensional data, and fast convergence. Classification problems are very common in bioinformatics and many of them involve high-dimensional and noisy data, with which SVMs are known to perform well. Applications of SVMs in bioinformatics include but are not limited to protein structure prediction (Hu, Pan, Harrison, & Tai, 2004), protein/gene function classification (Cai, Han, Ji, Chen, & Chen, 2003), protein subcellular localization prediction (Hua & Sun, 2001), splice site prediction (Degroeve et al., 2002; Jones & Watkins,