Computational Biology and Chemistry xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

SpliceVec: Distributed feature representations for splice junction prediction

Aparajita Dutta[a], Tushar Dubey[b], Kusum Kumari Singh[b], Ashish Anand[a,*]

[a] Department of CSE, Indian Institute of Technology, Guwahati, India
[b] Department of BSBE, Indian Institute of Technology, Guwahati, India

ARTICLE INFO

Article history:
Received 8 March 2018
Accepted 12 March 2018
Available online xxx

Keywords:
Splice junction
Feature representation
Alternative splicing
Distributed representation

ABSTRACT

Identification of intron boundaries, called splice junctions, is an important part of delineating gene structure and functions. This also provides valuable insights into the role of alternative splicing in increasing functional diversity of genes. Identification of splice junctions through RNA-seq is by mapping short reads to the reference genome which is prone to errors due to random sequence matches. This encourages identification of splicing junctions through computational methods based on machine learning. Existing models are dependent on feature extraction and selection for capturing splicing signals lying in the vicinity of splice junctions. But such manually extracted features are not exhaustive. We introduce distributed feature representation, *SpliceVec*, to avoid explicit and biased feature extraction generally adopted for such tasks. SpliceVec is based on two widely used distributed representation models in natural language processing. Learned feature representation in form of SpliceVec is fed to multilayer perceptron for splice junction classification task. An intrinsic evaluation of SpliceVec indicates that it is able to group true and false sites distinctly. Our study on optimal context to be considered for feature extraction indicates inclusion of entire intronic sequence to be better than flanking upstream and downstream region around splice junctions. Further, SpliceVec is invariant to canonical and non-canonical splice junction detection. The proposed model is consistent in its performance even with reduced dataset and class-imbalanced dataset. SpliceVec is computationally efficient and can be trained with user-defined data as well.

1. Introduction

In most of the eukaryotes, a protein coding gene is interrupted by intervening sequences called introns. These introns are removed from the exonic sequences via splicing process, which occurs cotranscriptionally (Shomron and Levy, 2009). The process generates mature substrates to be translated into proteins. Splicing involves snipping of introns at exon-intron and intron-exon junctions, called *splice sites* or *splice junctions*, and ligation of exons. Identification of splice junctions is therefore an important part of delineating gene structure and functions. Gene identification as well as its structural annotation has become an important problem in bioinformatics due to the abundance of sequenced genomes made available by advanced sequencing technologies like RNA-seq. RNA-seq, in recent times, has also provided meaningful insights into the role of alternative splicing (AS) in increasing functional diversity of genes. AS is a regulated process which alternatively skips or joins exons, or parts of exons or introns to form variety of proteins during translation. Recent estimates by RNA-seq suggest that more than 90% of multi-exon genes in human body undergo alternative splicing (Lu et al., 2012), thus making the identification of splice junctions all the more crucial. Identification of splice junctions from RNA-seq involves mapping of millions of short reads to the reference genome. However, multiple potential match of the short read on the reference genome makes sequence mapping less reliable (Li and Dewey, 2011). Also, the existing alignment based methods (Trapnell et al., 2009; Au et al., 2010) consider detection of splice sites based on only canonical splicing patterns (GT at donor site and AG at acceptor site) thereby missing important non-canonical splicing patterns (Lee and Yoon, 2015).

Abbreviations: AS, alternative splicing; CNN, convolutional neural network; t-SNE, stochastic neighbor embedding; MLP, multilayer perceptron; nt, nucleotides; bp, base pairs; NLP, natural language processing; CBOW, continuous bag of words; DBOW, distributed bag of words; DM, distributed memory; ReLU, rectified linear units; LSVM, linear support vector machines.

* Corresponding author.
E-mail addresses: d.aparajita@iitg.ernet.in (A. Dutta), tdiitg@gmail.com (T. Dubey), kusumsingh@iitg.ac.in (K.K. Singh), anand.ashish@iitg.ernet.in (A. Anand).

https://doi.org/10.1016/j.compbiolchem.2018.03.009
1476-9271/

The vast availability of annotated sequences makes it possible to create large enough training datasets for supervised learning algorithms to predict splice sites. The correct prediction of splice sites is facilitated by the identification of relationships and dependencies among the nucleotides around the splice sites. This is motivated by the observation that the splicing signals are most likely to reside in the vicinity of splice sites (Akerman and Mandel-Gutfreund, 2007). The learning algorithms need a set of features for training the model. Creating an optimized set of features that best represent the dataset has always remained a challenge for splice site prediction. The presence or absence of certain nucleotide sequences close to the splice sites were considered as features for splice site prediction for a long time (Brunak et al., 1991; Reese et al., 1997; Pertea et al., 2001; Degroeve et al., 2004; Huang et al., 2006; Baten et al., 2006; Sonnenburg et al., 2007). Since all such features were not known, there have been constant efforts to improve or refine features as well as include more relevant features by taking into account recent experimental observations. The large number of features may not only contain many irrelevant features but may also adversely affect classifier performance due to its high dimension. This led to many efforts for obtaining the optimum feature set through feature selection (Degroeve et al., 2002; Islamaj et al., 2006; Saeys et al., 2004, 2002). But the set of features obtained were not exhaustive.

This scenario motivated the adoption of an approach that can represent features specific to splice junctions based on the splicing signals without any manual extraction and selection of features. Rather, the model itself will capture motifs from the biological sequences that act as splicing signals for splice site selection. Lee et al. proposed a deep Boltzmann machine based methodology for splice junction prediction (Lee and Yoon, 2015). Recently, Zhang et al. have employed a deep convolutional neural network (CNN), named *DeepSplice* (Zhang et al., 2016), that learns features that characterize the true and decoy splice junctions. They have predicted novel splice junctions based on these features and obtained state-of-the-art performance of 96% accuracy. A distributed representation of biological sequences was proposed by Asgari and Mofrad (2015) and Kimothi et al. (2016) for protein family classification task where a prediction accuracy of more than 99% was achieved.

This paper introduces a novel approach for distributed feature representation of splice junctions by embedding it in an n-dimensional feature space. Each dimension in the feature space represents one feature of the corresponding splice junction. This embedding, named *SpliceVec*, is in the form of n-dimensional continuous distributed vector representation. The embeddings are learned by a shallow neural network using unsupervised data. We explored two variants of SpliceVec, namely *genome based SpliceVec (SpliceVec-g)* and *splicing-context based SpliceVec (SpliceVec-sp)* for feature representation of splice junctions. We evaluate the quality of SpliceVec in both intrinsic and extrinsic tasks. For the intrinsic evaluation, we visually inspect two dimensional representation of true and false splice sites using Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008). We evaluate SpliceVec on splice junction classification task for the extrinsic evaluation. In contrast to the recent deep learning methods, we use simple multilayer perceptron (MLP) as a classifier. We name this model as *SpliceVec-MLP*.

Our results and contributions can be summarized as follows:

- We propose two variations (SpliceVec-g and SpliceVec-sp) for feature representation of splice sites. SpliceVec outperforms state-of-the-art methods by 2.42–18.86% in terms of accuracy for splice site prediction.
- We explored the optimal sequence length that best captures the splicing signals for improving the prediction results. We find that inclusion of entire intronic sequence significantly boosts the predictive power of the classifier.
- The proposed feature representations are more robust in handling reduced training samples. SpliceVec maintains an accuracy above 99% even with a 60% reduction of training samples whereas the accuracy of its counterpart drops by about 6%.
- SpliceVec is more consistent in its performance with class-imbalanced data making it more suitable for the real time scenario where number of pseudo sites are several times more than that of true splice sites.
- SpliceVec-MLP identified non-canonical splice junctions (junctions not comprising of the consensus dimer GT or AG at donor or acceptor site respectively) with 100% accuracy indicating that our feature representations are invariant to both canonical and non-canonical splice junctions.
- SpliceVec-MLP can be deployed in both CPU and GPU environment. SpliceVec-MLP, being 12.94 times computationally faster than the state-of-the-art model, contributes as a suitable option for classification of the abundant annotated sequences available these days by high-throughput sequencing technologies.

2. Methods

The proposed approach can be divided into two stages: the feature representation stage that generates a distributed representation for each splice junction based on either of the two frameworks, namely *word2vec* and *doc2vec*, and classification of splice junctions using MLP. We shall discuss all these in the following subsections. An overview of the proposed approach is shown in Fig. 1.

2.1. Data

We have used the latest release of GENCODE annotation (Harrow et al., 2012) (version 26), based on human genome assembly version *GRCh38*, for extracting true and false splice junctions. This version was released in March 2017. We extracted 294,576 splice junctions from the protein coding genes. Based on these splice junctions, we observed that an intron length varies from 1 to as much as 1,240,200 nucleotides (nt). A recent study suggests that the shortest known eukaryotic intron length is 30 base pairs (bp) belonging to the human *MST1L* gene (Piovesan et al., 2015). Introns shorter than 30 bp are usually accounted to sequencing errors in genomes. Based on this study, we considered only those introns whose length were greater than 30 bp. This reduced the number of splice junctions to 293,889. We selected 293,889 false splice junctions by randomly searching for splice site consensus sequences GT and AG which were not annotated as splice junctions. This was considered as a necessary condition for selection of false splice junctions because more than 98% of splice junctions are canonical, that is, they contain the consensus dinucleotide GT at the donor site and AG at the acceptor site (Burset et al., 2000). In DeepSplice, the length of false splice junctions was considered to be lying between 10 and 300,000 nt, the reason of which was not clear. So, instead, we considered only those false splice junctions whose length was not less than 30 nt and not more than 1,240,200 nt with both donor and acceptor splice sites lying in the same chromosome. We considered this range to mimic the scenario of true introns.

2.2. Distributed representation

Vector representation of a word or a sentence is an integral part of many natural language processing (NLP) tasks. Although local representations, like N-grams and bag-of-words, have been

Please cite this article in press as: A. Dutta, et al., SpliceVec: Distributed feature representations for splice junction prediction, Comput. Biol. Chem. (2018), https://doi.org/10.1016/j.compbiolchem.2018.03.009