Boundary Classification Learning from Splice-junction Gene Sequences Using Machine Learning Algorithms

Authors: *Kamalanathan S. Govindan, Wenting Guo*

Abstract

Exons and introns have important roles in molecular biology, as splicing is the editing of the nascent precursor messenger RNA transcript, and after splicing, introns are removed and exons are joined together. The problem for us is, when given a sequence of DNA, we can identify the boundaries between the introns and exons, and classify the two types of boundaries: exon-intron site known as donors and intron-exon site, known as acceptors.

The dataset is from UCI Machine learning repository, which all examples were taken from Genbank 64.1 (ftp site: genbank.bio.net). We will use cross-validation method to randomly select 1000 samples each time from the full 3190 data. We will use decision tree ID3 and C4.5 algorithm, and compare the accuracy rate of three different ensemble methods for constructing the tree: Bagging, Boosting, and Randomization. The algorithms will be implemented using python. It is expected that the result of C4.5 should perform better than ID3.

Keywords

Exon–intron splice sites, DNA sequence, Machine learning

Introduction

Genetic information of an organism is stored in the genes, the functional subunits of the genome, arranged in the strands of the DNA double helix in the nucleus. This information is transcribed from DNA into a messenger RNA (mRNA) template by a process called transcription. However, before the mRNA can be translated into proteins, non-coding portions of the sequence, called introns, must be removed and protein-coding parts, called exons, joined by RNA splicing to produce a mature mRNA.

Scientists have discovered alternative patterns of pre-mRNA splicing that produced different mature mRNAs containing various combinations of exons from a single precursor mRNA.

Alternative splicing therefore is a process by which exons or portions of exons or noncoding regions within a pre-mRNA transcript are differentially joined or skipped, resulting in multiple protein isoforms being encoded by a single gene. This mechanism increases the informational diversity and functional capacity of a gene during post-transcriptional processing and provides an opportunity for gene regulation.

Alternative splicing generates a tremendous amount of proteomic diversity in humans and significantly affects various functions in cellular processes, tissue specificity, developmental states, and disease conditions. Thus, finding the intron-exon and exon-intron boundaries are important for further studying of RNA splicing process, and gene expression.

Neelam Goel, Shailendra Singh, Trilok Chand Aseri (2015) suggest that second order markov model is an effective pre-processing approach. This approach when combined with support vector machine provides better classification accuracy in predicting splice sites.

Alessandra Lumini , Loris Nanni, (2006) applied hierarchical multiclassifier (HM) architecture,

whose results show a drastically error reduction with respect to the performance of methods proposed in the literature.
Ho and Rajapakse (2003) provide a hybrid approach consisting of two first- and two-second-order Markov chain models at the first stage and of a three-layer neural network at the second stage. Jacob Engelbrecht (1991) applied artificial neural network to predict human mRNA donor and acceptor sites from DNA. Salvatore Rampone (June 10, 1998) used BRAIN learning algorithm to recognize the splice junctions on DNA sequences, and refined the classification also by neural network. Mukund Deshpande and George Karypis(2002) that the SVM-based approaches are able to achieve higher classification accuracy compared to the more traditional sequence classification algorithms such as Markov model based techniques and K -nearest neighbor based approaches. There have been a lot studies about DNA using decision trees, but it seems that there's not much study of finding donors and acceptors using decision tree algorithm.

Methodology

Decision tree: ID3, C4.5, Ensemble method: Bagging, Boosting, and Random Forest.
Data preprocessing. Though most of the features are filled with one of the letter: A,T,C,G, there are several unique letters representing the ambiguity: D means A or G or T; N means A or G or C or T; S means C or G; R means A or G.

The heart of the project lies in the selection of features for the given sequences. The idea is to select the relevant features for all the sequences to apply the machine learning algorithms to be able to classify a give sequence into an acceptor, donor or neither. The possible features can be frequency of k-mers in the sequence, k ranging from 1 to 60. Our attempt is to essentially figure out the features that are most relevant to the prediction. Using dimensionality reduction techniques like Principal Components iteratively can help in reducing the number of dimensions and finding the maximum variance and spread in data.

References

An Improved Method for Splice Site Prediction in DNA Sequences Using Support Vector Machines
Neelam Goel, Shailendra Singh, Trilok Chand Aseri (2015)

Classification of splice-junction sequences via weighted position specific scoring approach
Efendi Nasibova, Sezin Tunaboylua(2010)

Identifying splice-junction sequences by hierarchical multiclassifier
Alessandra Lumini , Loris Nanni, (2006)

Prediction of human mRNA donor and acceptor sites from the DNA sequence
Jacob Engelbrecht(1992)

Recognition of splice junctions on DNA sequences by BRAIN learning algorithm
Salvatore Rampone (June 10, 1998)

Evaluation of Techniques for Classifying Biological Sequences
Mukund Deshpande and George Karypis. (2002)

Splice site detection with a higher-order Markov model implemented on a neural network
Ho and Rajapakse (2003)

Splice Site Prediction Using Artificial Neural Networks
Øystein Johansen, Tom Ryen, Trygve Eftesøl, Thomas Kjosmoen, and Peter Ruof (2008)