

# BioSPRINT: Classification of Intron and Exon Sequences Using the SPRINT Algorithm

Kevin Crosby  
Furman University

Paula Gabbert  
Furman University  
paula.gabbert@furman.edu

## Abstract

*An important problem for computer scientists as well as geneticists involves classifying particular items into common groups. This paper focuses on classifying sequences of DNA as either an intron or an exon. Insights from this classification can reduce the time needed for laboratory work to distinguish between introns and exons. Using a classification tree based on the SPRINT algorithm, sequences from the *Drosophila melanogaster* and the *Caenorhabditis elegans* genomes were used for training and testing. A large test sample error rate of 15% was shown for the *Drosophila melanogaster*, whereas the *Caenorhabditis elegans* was only 1.6%.*

## 1. Introduction

Prior work on classification of DNA sequences includes K-nearest neighbor techniques [1] and wavelet representations [2]. However, little work has focused on the classification of intron and exon sequences. The purpose of BioSPRINT is to successfully classify sequences of DNA as either introns or exons using decision tree based classifiers. Generically, these trees can be used to classify any type of data, not just genomic data. However, a problem with genomic data is that features must be extracted from the sequences before classification algorithms can be used. Once this is done, geneticists can use the algorithm as an aid to distinguish between introns and exons.

Sequence data from the DNA of the *Caenorhabditis elegans* (worm) and the *Drosophila melanogaster* (fruit fly) is formatted into specific numeric values according to the nucleotide patterns that make up the strand. In order to do this, the

percent composition of each nucleotide and the number of dinucleotide patterns are counted and classification relies on these features. BioSPRINT is run on a Silicon Graphics Onyx II system consisting of four processors with a shared memory architecture. Although the parallel environment is not directly used in the initial implementation of BioSPRINT, it can be used as a springboard for many extensions to the algorithm.

## 2. Sequence Classification

BioSPRINT first builds an initial tree using the Scalable PaRallelizable INduction of decision Trees (SPRINT) algorithm until each node only contains one class [3], and then the pruning method prunes upward according to the cost-complexity algorithm presented in *Classification and Regression Trees* (CART) [4]. Each internal node in the resulting tree identifies a feature for splitting the data (i.e. % Thymine > 27%), while terminal nodes are assigned a class (i.e. intron or exon), so DNA sequences can be dropped down the tree and classified as either intron or exon based on the split criteria.

Example strands of exons and introns are used to build the classification tree. Since classification tree methods rely on characteristics of the sequences, not the sequences themselves, the sequences are first analyzed for certain characteristics. Choosing which characteristics to use is fundamental to this problem. The initial implementation of BioSPRINT uses compositional features including the frequency of the dinucleotide subsequences (e.g. AA, AT, etc.) and the percent composition of each nucleotide. There are sixteen dinucleotide patterns and four percent composition attributes for a total of twenty features for each sequence of DNA.

After deciding which features to use for classification, there still exists the gap between the sequence nature of the string and the numerical representation needed for classification. To automate this, BioSPRINT includes a component to scan each sequence and calculate the specific number of dinucleotide sequences, as well as the percent composition to create a data set for classification.

A large portion of this data set is used to build an initial tree with pure terminal nodes. The Gini index is used to determine the best split of the data at each node using the compositional features extracted from the sequences. This index attempts to find splits that minimize misclassification error in the training data. Splitting of nodes stops when nodes are completely "pure" and contain *either* introns or exons (but not both).

After this initial tree is built, appropriate pruning is done using the remaining data for test sample estimation of error. Pruning removes subtrees of the large tree that do not provide sufficient information for classification. Pruning continues in a step-wise manner until only the single root node remains. At each step of pruning the test sample error is calculated. The pruned tree with the lowest error denotes the most accurate tree, and this tree can be used for classifying new sequences of DNA.

### 3. Results

The sequence data for the *Drosophila melanogaster* and the *Caenorhabditis elegans* was attained from GenBank [5]. There are roughly 200,000 intron sequences and 100,000 exon sequences. These sequences vary in length from about ten nucleotides up to thousands of nucleotides. For initial testing of the BioSPRINT algorithm, 2000 sequences were used to build the tree. Exactly half of these are introns, and the other half exons. This number was chosen because it is large enough to build an accurate tree, yet small enough for the algorithm to handle without using external files during classification.

For the *C. elegans*, BioSPRINT built a very accurate tree with twelve internal nodes. The overall test sample error for this tree was 1.6%. This tree could be used to provide information for wet lab experiments to determine whether a sequence is an intron or an exon. Since this error is so small, it leads to the conclusion that percent composition and dinucleotide sequences are fundamental features when determining introns and exons in the *C. elegans* genome. Specifically, the percentage of thymine is

split upon multiple times, showing that it plays an integral role in distinguishing between introns and exons for the *C. elegans*.

The tree built from the introns and exons of the *Drosophila melanogaster* is not as accurate as the one for the *C. elegans* and was much larger containing over 80 internal nodes. The test sample error for this dataset was 15%. This shows that the features (percent composition and dinucleotide sequences) used to classify the introns and exons of the *Drosophila melanogaster* were not useful indicators for introns and exons.

### 5. Conclusion and Future Work

As shown in the results section, the error minimizing tree created for this project from the *C. elegans* data makes for an accurate classifier, whereas the *Drosophila melanogaster* tree was not quite as accurate. Future work in this area will investigate two methods that could possibly enhance the classification capabilities for both of these organisms and lead to better results. First, additional features of the sequence will be investigated including positional features [2]. Another area for future work is the inclusion of additional training sequences by using external files during classification or using statistical analysis on all the sequences to create smaller uncertain learning sets that contain feature data and probability distributions on the class label.

### 6. References

- [1] Deshpande, M. and Karypis, G., "Evaluation of Techniques for Classifying Biological Sequence", Technical Report, TR 01-33, University of Minnesota, 2001.
- [2] Aggarwal, C. C., "On Effective Classification of Strings with Wavelets", *Proceedings of the Eighth SIGKDD Conference*, Alberta, Canada, 2002.
- [3] J. Shafer, R. Agrawal, and M. Mehta, "SPRINT: A Scalable Parallel Classifier for Data Mining", *Proceedings of the 22nd VLDB Conference*, Bombay, India, 1996, pp. 544-555.
- [4] Breiman, L., *Classification and Regression Trees*, Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, California, 1984.
- [5] NCBI HomePage (GenBank), National Center for Biotechnology Information, <http://www.ncbi.nlm.nih.gov/>, 2004.