

# Prediction of *trans*-Splicing Sites Using a Genetic Algorithm

Saria Awadalla <sup>1</sup> Juan E. Ortiz <sup>2</sup> Shuba Gopal <sup>3</sup>

**Keywords:** *trans*-splicing, computational prediction, machine learning, genetic algorithms

## 1 Introduction.

*Trans*-splicing is an unusual process in which two separate RNA strands are spliced together to yield a mature mRNA. The process is most common in the Kinetoplastida, a family of eukaryotic parasites. While much effort has been focused on computationally predicting *cis*-splicing sites [4], a rigorous and thorough analysis of *trans*-splicing has not yet been undertaken. Yet, the two processes may have common mechanisms [3], and it is therefore possible that by considering the signals involved in *trans*-splicing new insights can be gained regarding RNA splicing in all eukaryotes.

The canonical *trans*-splicing signal is believed to be composed of four elements, of which a poly-pyrimidine (C,T-rich) tract is the easiest to discern computationally. The challenge with identifying these tracts is that they are not highly conserved but are variable both in composition and in sequence length. Here, we present a combination of statistical techniques and machine learning to develop a method capable of identifying *trans*-splicing regions and pinpointing the specific site of splicing in *Leishmania major*. Our method is 89% accurate, with a sensitivity of 86% and a specificity of 92%.

## 2 Software and files.

In our set of known *trans*-splicing regions, it was observed that the AG dinucleotide that serves as the splice acceptor site is often isolated from other potential splice acceptor AGs by long stretches of non-AG dinucleotides. In fact, the inter-AG distance alone can be used to identify nearly 60% of true *trans*-splicing sites. To improve on this predictive capacity, we included a measure of nucleotide composition within inter-AG segments. We evaluated the first-order transition probabilities for the inter-AG segments. We developed two matrices for scoring segments based on nucleotide composition: one derived from a previous analysis by linear discriminant analysis [2], a second optimized by a steady-state genetic algorithm (GA). This machine learning approach allows us to optimize a multivariate function while avoiding a complex combinatorial problem.

The GA utilizes a steady-state breeding algorithm with uniform cross-over. These features, in addition to a relatively high mutation rate (1% - 5%), provided a good mechanism for avoiding convergence onto local optima. The fitness function scored each candidate *trans*-splicing site using a weighted transition frequency matrix as well as generating a weighted score for the distance from the previous AG. Using the scoring function from these matrices

---

<sup>1</sup>Department of Biological Sciences, Rochester Institute of Technology, Rochester, NY. E-mail: ssa6996@rit.edu

<sup>2</sup>Department of Computer Science, Rochester Institute of Technology, Rochester, NY. E-mail: juan.e.ortiz@gmail.com

<sup>3</sup>Department of Biological Sciences, Rochester Institute of Technology, Rochester, NY. E-mail: sxgsbi@rit.edu

in combination with a log-transformed measure of the inter-AG segment length yielded a highly accurate predictor of *trans*-splicing sites. This is shown in Table 1.

### 3 Figures and tables.

	<b>Known Splice Sites</b> (Total: 136)	<b>Known Non-Splice Sites</b> (Total: 136)
<b>Predicted Splice Site</b>	True positive 117	False positive 11
<b>Predicted Non-Splice Site</b>	False negative 19	True negative 125
	<b>Sensitivity: 0.86</b>	<b>Specificity: 0.92</b>
	<b>Accuracy: 0.89</b>	

Table 1: Performance of the method is shown in this table. The test data consisted of 136 known *trans*-splicing sites from *L. major*, and an equivalent number of randomly selected inter-AG segments from known coding regions. We chose coding regions as the best model for true negatives because *trans*-splicing should never occur within a protein-coding region. There are no known instances of *cis*-splicing in *L. major*, so we can be reasonably confident that these inter-AG segments serve no role in either splicing process.

### 4 References and bibliography.

#### References

- [1] Blumenthal, T., Evans, D., Link, C. D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W. L., Duke, K., Kiraly, M., and Kim, S. K. 2002. A global analysis of *Caenorhabditis elegans* operons. *Nature*, 417:851854.
- [2] Gopal, S., Cross, G., and Gaasterland, T. 2003. An organism-specific method to rank predicted coding regions in *Trypanosoma brucei*. *Nucl. Acids. Res.*, 31:58775885.
- [3] Liang, X.-h., Haritan, A., Uliel, S., and Michaeli, S. 2003. *trans* and *cis* splicing in Trypanosomatids: Mechanism, Factors, Regulation. *Euk. Cell*, 2:830840.
- [4] Rogic, S., Mackworth, A. K., and Ouellette, B. F. 2001. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.*, 11:817832.