

# In silico identification and characterization of mRNA-like noncoding transcripts in *Medicago truncatula*

Jiayu Wen<sup>1</sup>, Brian J. Parker<sup>2</sup> and Georg F. Weiller\*<sup>1</sup>

<sup>1</sup>Australian Research Council (ARC) Centre of Excellence for Integrative Legume Research and Bioinformatics Laboratory, Research School of Biological Sciences, Australian National University, Canberra, Australia

<sup>2</sup>Statistical and Machine Learning Group, National ICT Australia (NICTA) and Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia

Email: Georg F. Weiller\* - georg.weiller@anu.edu.au;

\*Corresponding author

## Abstract

---

Accumulating evidence suggests that ncRNAs play key roles in gene regulation and may form the basis of an inter-gene communication system. Many ncRNAs are synthesized similar to mRNAs and can be detected through screening of polyA-rich EST or cDNA libraries. We developed a computational pipeline to screen EST and genomic sequence data for those transcribed genes with limited protein coding potential and applied this pipeline to the model legume *Medicago truncatula*. This process identified a set of 503 mRNA-like transcripts that appear not to encode proteins. Further computational analysis showed that many of these ncRNA candidates share structural similarities to known ncRNAs and that they clearly differ from protein coding genes and non-transcribed regions in their base and oligonucleotide compositions, as well as in aspects of secondary structure. By using a machine learning approach, we showed that the distinctive ncRNA features presented in this study can be used to discriminate most ncRNAs and may thus be useful for improving ncRNA prediction. Computational analysis of EST isolation frequencies in various plant tissues showed that the expression levels and expression profiles of the putative ncRNAs and mRNAs differ — most interestingly, the putative ncRNAs are highly expressed relative to mRNAs in the root nodule tissue and conserved only in closely related plants. The work presented here constitutes the first large-scale prediction and characterization of ncRNAs in legumes, and provides a basis for further research on elucidating ncRNA function in legume genomics.

Keywords: ncRNA, mRNA-like ncRNA, EST, *Medicago truncatula*, model legume, SVM, feature classification

---