In silico Prediction of Active RNA Genes in Legumes

Jiayu Wen

April 2007

A thesis submitted for the degree of Doctor of Philosophy of the Australian National University
Abstract

Accumulating evidence suggests that non-coding RNAs (ncRNAs) play key roles in gene regulation and may form the basis of an inter–gene communication system. MicroRNAs are a class of small non-coding RNAs found in both plants and animals that regulate the expression of other genes. Identification and analysis of microRNAs enhances our understanding of the important roles that microRNAs play in this complex regulatory network. The work presented in this thesis constitutes the first large-scale prediction and characterization of both ncRNAs and miRNAs in the model legume *Medicago truncatula* and *Lotus japonicus*, and provides a basis for further research on elucidating ncRNA function in legume genomics.

Thus far the search for novel ncRNA genes is hampered by a lack of statistically significant sequence features for predicting ncRNA genes. However, many ncRNAs are synthesized similar to mRNAs and can be detected through screening of polyA-rich EST or cDNA libraries. In this thesis, I 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 *M. truncatula* and *L. japonicus*. This process identified sets of 673 *M. truncatula* and 1637 *L. japonicus* mRNA-like RNA transcripts that appear not to encode proteins. Further computational analysis showed that many of these ncRNA candidates only had discernable homologues in closely related plants. By using a machine
learning approach, I showed that they differ substantially from protein coding genes and non-transcribed regions in their base and oligonucleotide compositions as well as in aspects of secondary structure asymmetry; the features presented in this thesis indicative of non-coding RNAs may 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 *M. truncatula* ncRNA set was highly expressed relative to mRNA in the root nodule tissue. Analysis of promoter regions suggests an elevated level of bidirectional transcription, whereby one of the divergent gene partners is non-coding.

I designed PCR primers for the *M. truncatula* ncRNA set to validate their expression through collaborations. So far the initial experiment with a qRT-PCR assay showed that 93% experimentally tested primers successfully amplified expected regions, demonstrating that the ncRNA sets consist of genuine transcribed genes.

It is often difficult to perform large scale validation of miRNA expressions that are predicted from genomic regions. Expressed sequences provide an alternative resource to facilitate identification of miRNAs and their targets. I developed a computational pipeline to scan for miRNA genes from the identified ncRNA transcripts and intronic regions of *M. truncatula* and *L. japonicus* genomic sequences. The data is represented in the database–MIRATdb (MiRNA And Target gene Data Base). It provides detailed information on the sequences of the predicted miRNAs, their precursors, and potential target genes. It also details sequence source information such as the EST library, tissue category, and number of EST clones. Information regarding miRNA conservation in other species, functional classification of target genes, and clusters of similar miRNAs are further provided.
The web interface to the database provides researchers with the ability to narrow their search for miRNAs and target genes of interest by using a variety of filters.
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARC</td>
<td>Australian Research Council</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under receiver operating characteristic Curves</td>
</tr>
<tr>
<td>BAC</td>
<td>bacterial artificial chromosome</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>bp</td>
<td>base pair(s)</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary deoxyribonucleic acid</td>
</tr>
<tr>
<td>cdRNA</td>
<td>coding RNA</td>
</tr>
<tr>
<td>CILR</td>
<td>ARC Centre of excellence for Integrative Legume Research</td>
</tr>
<tr>
<td>EST</td>
<td>Expressed Sequence Tag</td>
</tr>
<tr>
<td>ET</td>
<td>Expressed Transcripts</td>
</tr>
<tr>
<td>GI(s)</td>
<td>TIGR Gene Index sequence(s)</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>GTD</td>
<td>G and T nucleotide density differences (in the folded local structures)</td>
</tr>
<tr>
<td>ΔGTD</td>
<td>G and T nucleotide density differences between the two complementary strands</td>
</tr>
<tr>
<td>IRD</td>
<td>inverted repeat density</td>
</tr>
<tr>
<td>ΔIRD</td>
<td>inverted repeat density differences between the two complementary strands</td>
</tr>
<tr>
<td>kb</td>
<td>kilobase (1000 bp)</td>
</tr>
<tr>
<td>kcal</td>
<td>kilocalorie</td>
</tr>
<tr>
<td>KDE</td>
<td>Kernel Density Estimate</td>
</tr>
<tr>
<td>LjGI</td>
<td><em>Lotus japonicus</em> Gene Index</td>
</tr>
</tbody>
</table>
MtGI  
Medicago truncatula Gene Index

Mb  
megabase (1000 kb)

MFE  
Minimum Free Energy

MFED  
MFE density

MIRATdb  
microRNA and target database

miRNA  
microRNA

miRNA*  
The sequence regions that pairs with the miRNA in the miRNA precursor.

mol  
mole

mRNA  
messenger RNA

NCBI  
The National Center for Biotechnology Information

ncRNA  
non-coding RNA

ntDNA  
on-transcribed DNA

ORF  
Open Reading Frames

pri-miRNA  
primary miRNA transcript

qRT-PCR  
Quantitative Real Time Polymerase Chain Reaction

RACE  
Rapid Amplification of cDNA ends

RISC  
RNA induced silencing complex

RNA  
Ribonucleic Acid

ROC  
Receiver Operating Characteristic

rRNA  
ribosomal RNA

RT-PCR  
Real Time Polymerase Chain Reaction

siRNA  
small interfering RNA

SLURM  
Simple Linux Utility for Resource Management

snoRNA  
small nucleolar RNA

snRNA  
small nuclear RNA

SSR  
Simple Sequence Repeats
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>TC</td>
<td>Tentative Consensus (sequences)</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription Factor</td>
</tr>
<tr>
<td>TIGR</td>
<td>The Institute for Genomic Research</td>
</tr>
<tr>
<td>Tm</td>
<td>melting temperature</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcription Start Site</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated Region</td>
</tr>
</tbody>
</table>
To my masters, beloved parents, and my loves

给我的上师，我的父母，和我爱的人
Statement of Originality

I declare that this dissertation is my own original work. It does not contain any material previously submitted for a degree to this or any other university or written by another person where due reference is not made in the text.

Jiayu Wen
Acknowledgement

I would like to thank my principle supervisor Dr. Georg Weiller for critical guideline, ideas, supports, and encouragement throughout my past three and half year studies. Without him, this dissertation would not have been possible.

I thank my associate supervisor Dr. Michael Djordjevic and advisor Professor Peter Gresshoff for their advice and encouragements. I also appreciate the ARC legume research centre for providing me this study opportunity.

I acknowledge my fellow colleagues Dr. Brian Parker for great collaborations with papers and statistical advice; I have learned lots from him. I also thank Dr. Tancred Frickey for many valuable discussions. I appreciate Professor John Campbell for helping me with the papers.

Special thanks to

my masters for all their spirit supports and encouragements whenever I need most.

My dear Dad, you had been always supportive and wanted to see your children getting high education; this is for you from your daughter. Mom, I am grateful that your loves and cares are always there for me no matter what.

Rob and Dodo for being special ones in my life, the no. 1 fans, all the supports, and great help with this dissertation.

my friends Wei, Di, Tina, Arvin, Giel, Rosi, and many other friends for sharing the enjoyable time together.
# Contents

1 Introduction

1.1 Are we still living in “the RNA world”?

1.2 What are non-coding RNAs?

   1.2.1 SnoRNAs

   1.2.2 miRNAs and siRNAs

   1.2.3 mRNA-like ncRNAs

1.3 A ncRNA-mediated regulatory network

1.4 Why study plant ncRNAs? and why legumes?

1.5 Aim of this study

1.6 Methodology

1.7 Significance of the study

1.8 Thesis structure

2 A review of ncRNA studies

2.1 Why ncRNAs were largely ignored in the past?

2.2 Challenges to detect ncRNAs

2.3 Review of experimental and computational approaches

   2.3.1 Experimental methods

   2.3.2 Computational approaches

2.4 Current ncRNA databases
CONTENTS

2.4.1 Rfam ................................................. 17
2.4.2 ASRP ................................................. 18
2.4.3 RNAdb ............................................... 18
2.4.4 NONCODE ............................................. 18
2.4.5 The noncoding RNA (ncRNA) database ................. 19
2.4.6 Other ncRNA databases ................................ 19

3 Computational identification of mRNA-like ncRNAs in legumes 21

3.1 Prediction of ncRNA candidates ............................. 21
  3.1.1 The strategy of ncRNA prediction ...................... 21
  3.1.2 The ncRNA prediction pipeline ....................... 24
  3.1.3 Summary ............................................. 32

3.2 Analysis of ORF distribution in putative ncRNA set ........ 34
  3.2.1 Strategy ............................................. 34
  3.2.2 Methods ............................................. 34
  3.2.3 Results and Discussion .............................. 35

3.3 Conservation of putative ncRNAs in other species .......... 35

3.4 Ranking putative ncRNAs .................................. 37

4 Characteristic comparisons of the ncRNA, mRNA, and ntDNA
  datasets .................................................. 41

4.1 The purpose of the characteristic comparison studies .. 41

4.2 Construction of comparison datasets ....................... 42

4.3 Univariate analysis of base composition biases ........... 43
  4.3.1 Strategy ............................................. 43
  4.3.2 Methods ............................................. 43
  4.3.3 Results and Discussion .............................. 44

4.4 Univariate analysis of RNA secondary structures .......... 49
# CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4.1</td>
<td>Strategy</td>
<td>49</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Methods</td>
<td>51</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Results and Discussion</td>
<td>52</td>
</tr>
<tr>
<td>4.5</td>
<td>Multivariate feature analysis using SVM</td>
<td>54</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Strategy</td>
<td>54</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Methods</td>
<td>56</td>
</tr>
<tr>
<td>4.5.3</td>
<td>Results and discussion</td>
<td>58</td>
</tr>
<tr>
<td>4.6</td>
<td>Comparison of transcript expressions</td>
<td>66</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Strategy</td>
<td>66</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Methods</td>
<td>66</td>
</tr>
<tr>
<td>4.6.3</td>
<td>Results and Discussion</td>
<td>67</td>
</tr>
<tr>
<td>4.7</td>
<td>Examination of bidirectional transcription</td>
<td>69</td>
</tr>
<tr>
<td>4.7.1</td>
<td>Strategy</td>
<td>69</td>
</tr>
<tr>
<td>4.7.2</td>
<td>Methods</td>
<td>72</td>
</tr>
<tr>
<td>4.7.3</td>
<td>Results and Discussion</td>
<td>72</td>
</tr>
<tr>
<td>5</td>
<td>RT-PCR primer design for <em>Medicago truncatula</em> putative ncRNA genes</td>
<td>77</td>
</tr>
<tr>
<td>5.1</td>
<td>Strategy</td>
<td>77</td>
</tr>
<tr>
<td>5.2</td>
<td>Methods</td>
<td>78</td>
</tr>
<tr>
<td>5.3</td>
<td>Results and Discussion</td>
<td>80</td>
</tr>
<tr>
<td>6</td>
<td>Prediction of miRNAs and their target genes</td>
<td>83</td>
</tr>
<tr>
<td>6.1</td>
<td>Background</td>
<td>83</td>
</tr>
<tr>
<td>6.2</td>
<td>Strategy</td>
<td>87</td>
</tr>
<tr>
<td>6.3</td>
<td>Methods</td>
<td>88</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Source sequences</td>
<td>88</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Prediction criteria derived from the known miRNAs</td>
<td>89</td>
</tr>
</tbody>
</table>
### CONTENTS

6.3.3 Prediction pipeline of miRNAs and their targets 89  
6.3.4 Assessment of miRNA prediction 94  
6.4 Results and Discussion 96  
6.4.1 Data statistics 96  
6.4.2 Exonic and intronic miRNAs 98  
6.4.3 Advantages and disadvantages of using expressed sequences as source sequences 100

7 Characterization of the predicted miRNAs and targets 101  
7.1 Aim 101  
7.2 Results and Discussion 101  
7.2.1 Taxonomic conservation of miRNAs 101  
7.2.2 Clustering miRNAs 104  
7.2.3 Classifying miRNA target function 105

8 MIRATdb – legume putative miRNA and target database 109  
8.1 Aim 109  
8.2 Implementation 110  
8.3 Database interface 110

9 Implementations of programming source code 115  
9.1 Perl scripts 115  
9.2 Web-based application 120  
9.3 Statistical computing using R 123  
9.4 Usage of high performance Linux clusters 125  
9.4.1 High performance clusters 125  
9.4.2 Specification of RSBS GIG Linux cluster 125  
9.4.3 Uses of the linux cluster in this project 126
List of Figures

1.1 ncRNAs regulate gene expression in many cellular processes . . . 3
1.2 The ncRNA-mediated regulatory network . . . . . . . . . . . . . 6

3.1 EST constructions in TIGR and Unigene . . . . . . . . . . . . . . 23
3.2 The ncRNA prediction strategy . . . . . . . . . . . . . . . . . . . 24
3.3 Flowchart of the computational pipeline used to identify mRNA-like ncRNAs in M. truncatula and L. japonicus . . . . . . . . . . . 26
3.4 Relative overlap calculations . . . . . . . . . . . . . . . . . . . . 30
3.5 Relative overlap density distributions in M. truncatula and L. japonicus plotted using Kernel Density Estimate (KDE) . . . . . . . . 31
3.6 The bootstrapped sampling distribution for the longest ORF . . . 36
3.7 ncRNA conservation across other species . . . . . . . . . . . . . . 38

4.1 Density plots of base compositions for M. truncatula . . . . . . . 46
4.2 Density plots of base compositions for L. japonicus . . . . . . . 47
4.3 The plot for inverted repeat scores when compared to MFEs . . . 53
4.4 Density plots of both global (ΔMFED) and local (ΔIRD, ΔGTD) . 55
4.5 Plots of a three dimensional feature clouds for M. truncatula and L. japonicus . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59
4.6 The ROC curves for ncRNA, mRNA, and ntDNA classification for M. truncatula . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 64
LIST OF FIGURES

4.7 The ROC curves for ncRNA, mRNA, and ntDNA classification for
L. japonicus .................................................. 65
4.8 The distributions of EST isolation frequency for coding and non–
coding genes in M. truncatula and L. japonicus ................. 68
4.9 Comparison of tissue expression levels in M. truncatula ....... 70
4.10 Comparison of tissue expression levels in L. japonicus ......... 71
4.11 Distance distributions of gene pairs in M. truncatula and L. japon-
icus) ............................................................ 74

5.1 The flowchart of RT-PCR primer design pipeline ............... 81

6.1 Current models of miRNA biogenesis in plants and animals . 85
6.2 Flowchart of the computational pipeline of miRNA and target pre-
diction ............................................................... 91
6.3 Bootstrap frequency distribution for the miRNAs predicted from
shuffled sequences .................................................. 95
6.4 Statistics of using different miRNA::target duplex MFE thresholds
obtained in M. truncatula and L. japonicus ...................... 99

7.1 Clustering of 21mer putative miRNAs .......................... 106

8.1 Screenshot of the miRNA details table .......................... 113
8.2 Screenshot of the target details table ........................... 114

9.1 miRNA query interface .......................................... 120
9.2 The cluster model ................................................. 125
9.3 Distributed memory approach ................................... 127

A.1 MIRATdb schema ................................................. 141
# List of Tables

3.1 Purification summary of *Medicago truncatula* and *Lotus japonicus* ncRNAs and peptide genes ........................................ 33

3.2 Summary of the longest ORF distributions in ncRNA set and the randomized same sequence set ..................................... 35

3.3 ncRNA Ranks ........................................................................... 39

4.1 Base composition bias among the ncRNA, mRNA, and ntDNA datasets for *M. truncatula* .................................................. 45

4.2 Base composition bias among the ncRNA, mRNA, and ntDNA datasets for *L. japonicus* ..................................................... 45

4.3 Comparisons of RNA secondary structures between the ncRNA, mRNA, and ntDNA datasets for *M. truncatula* .................. 54

4.4 Comparisons of RNA secondary structures between the ncRNA, mRNA, and ntDNA datasets for *L. japonicus* .................. 56

4.5 Classification based on base composition biases and RNA structure stand asymmetry using SVM for *M. truncatula* ............ 60

4.6 Classification based on base composition biases and RNA structure stand asymmetry using SVM for *L. japonicus* ............ 61

6.1 Summary of miRNA and target prediction ................................. 97
7.1 Summary of miRNA conservation and clustering . . . . . . . . . . 103
7.2 Functional analysis of predicted miRNA target genes versus all
genes in Medicago truncatula . . . . . . . . . . . . . . . . . . . . . 107
7.3 Functional analysis of predicted miRNA target genes versus all
genes in Lotus japonicus . . . . . . . . . . . . . . . . . . . . . . . 108
7.4 Known miRNA target function class associated with the putative
miRNAs and targets . . . . . . . . . . . . . . . . . . . . . . . . . . . 108
Chapter 1

Introduction

1.1 Are we still living in “the RNA world”?

Many biologists hypothesize that life on the earth existed in “the RNA world” before the modern cells were formed. That is, RNA carried genetic information and catalyzed chemical reactions before the more efficient and chemically stable DNA and protein usurped these roles (Gesteland et al., 1998). Traditionally, we believe that the “central dogma” (DNA to RNA to protein) represents the whole picture of genetic transmission. That is, present-day RNAs are only the evolitional remains of “the RNA world” and merely function as DNA templates. Until recently therefore, the study of RNAs has been eschewed in favor of the more popular DNAs and proteins.

Supported by several genome sequencing projects, we found that, however, only tiny parts of the sequence encode proteins in the higher eukaryotes (MouseGenomeSequencingConsortium, 2002; Rubin et al., 2000; Venter et al., 2001). More specifically, studies assessing levels of non-protein-coding sequences across different species showed that organisms become more advanced if they have more non-protein-coding DNAs. For example, it is revealed that less than
25% of prokaryote genomes consisted of non-protein-coding DNAs, that more than 50% of more advanced animal and plant genomes consisted of non-protein-coding DNAs, and that whereas up to 98% of human genome sequence consisted of non-protein-coding DNA (Mattick, 2004). Furthermore, humans and mice are 99% similar in their protein-coding genes (MouseGenomeSequencingConsortium, 2002), and individual humans are 99% identical in their protein-coding genes (Venter et al., 2001). More still, a large proportion of mammalian genomes are transcribed (Frith et al., 2005). It has therefore been surmised that the individual variations within or between species may rely on non-protein-coding elements (Mattick, 2001). Thus, we logically wonder what the rest of non-protein coding transcripts (non-coding RNAs or ncRNAs) do if they do not translate to proteins. The answer to this question is that they are either junk transcripts or they have meaningful functions. If the latter holds true, we may have missed something essential in our previous understandings of genetic mechanisms. Indeed, the recent discovery of various ncRNAs has shed the light on the later case and therefore provide evidence that we may still live in the RNA world — a modern RNA world.

1.2 What are non-coding RNAs?

Non-coding RNAs (ncRNAs) are characterized as functional RNAs that are transcribed but are not translated to proteins and that therefore do not require long open reading frames (ORFs). ncRNAs have a variety of sizes, ranging from small (i.e., 20 bp microRNAs) to very large (i.e., more than 10 kb imprinting RNAs) (Storz, 2002). Aside from well-known transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and small nuclear RNAs (snRNAs), a variety of functional ncRNAs have been shown to play key roles in a number of cellular processes including chromatin
1.2. WHAT ARE NON-CODING RNAs?

Evidence has shown that ncRNAs play key roles in chromatin modification, transcription initiation regulation, mRNA synthesis, protein synthesis, and post-translational process regulation. The most significant findings are small nucleolar RNAs (snoRNAs), microRNAs (miRNAs), small interfering RNAs (siRNAs), and mRNA-like ncRNAs.

1.2.1 SnoRNAs

SnoRNAs base paired with rRNA modification sites guide rRNA modifications of either 2'-O-methylated nucleotides or pseudouridines (Alberts et al., 2002). Research shows that snoRNAs can target not only rRNAs, but also other types of ncRNAs, including tRNAs, snRNAs, and mRNAs (Bachellerie et al., 2002).
The majority of snoRNAs are processed from introns of either protein-coding or non-protein-coding transcripts (Bachellerie et al., 2002).

1.2.2 miRNAs and siRNAs

The recent discovered classes of ncRNAs are miRNAs (microRNAs) and siRNAs. Thus far, research shows that they have very similar mechanisms. Both miRNAs and siRNAs have short lengths of approximately 19–25 bp. Both of them are often associated with many proteins thought to be adaptors for proteins that are required to recognize other sequences (Bartel, 2004; Kim, 2005a). Both of them target genes by imperfect or nearly perfect base pairings and, in turn, lead to gene silencing.

However, miRNAs and siRNAs differ in their origins, loci where they target, and conservation levels (Bartel, 2004; He and Hannon, 2004). Mature miRNAs are processed from hairpin–shaped miRNA precursors which are located in both exons of non-coding genes and introns of coding and non-coding genes (Kim, 2005a). SiRNAs, however, are processed from long double strand RNAs that are derived from transposons, viruses, or mRNAs (Kim, 2005a). MiRNAs target different loci from which they are generated, whereas siRNAs target the same loci from which they are generated (Kim, 2005a). For that reason, miRNAs target endogenous mRNAs through one of three modes: mRNA cleavage, translational repression, or RNA-directed DNA methylation (Floyd and Bowman, 2005), while siRNAs prevent invasion by targeting viruses and transposons through RNA cleavage or alteration of chromatin structures (Floyd and Bowman, 2005). The known miRNAs are often conserved in related species whereas siRNAs are varied.

It is worth mentioning that although differences exist in the current modes of miRNAs and siRNAs, much more remains to be discovered. Take the recently
discovered trans-acting siRNAs (tasiRNAs) - a subclass of siRNAs for example. TasiRNAs, which act like miRNAs, have been shown to be derived from introns of non-coding genes and silence endogenous mRNAs by trans-acting (Kim, 2005b).

1.2.3 mRNA-like ncRNAs

One class of ncRNAs acts like mRNAs in that they are spliced, polyadenylated, and possibly 5’ capped termed as “mRNA-like ncRNAs” (Erdmann et al., 2000). Compared to tiny microRNAs and small interfering RNAs, these ncRNAs are usually longer in size ranging from hundreds of base pairs to more than 10k base pairs (Storz, 2002). Examples of mRNA-like ncRNAs include mouse air RNA required for gene imprinting (Sleutels et al., 2002), yeast meiRNA involved in meiosis control (Ohno and Mattaj, 1999; Yamashita et al., 1998) and mammalian XIST RNAs required for X chromosome inactivation (Kelley and Kuroda, 2000).

1.3 A ncRNA-mediated regulatory network

Single stranded RNAs are characterized by their folding into intricate structures by not only base pairings, but also tertiary interactions (Alberts et al., 2002). This enables RNAs to interact with proteins, DNAs, and other RNAs (Mattick and Gagen, 2001). Therefore, small RNAs are excellent candidates to play regulatory roles in the cell.

Many ncRNAs such as snoRNAs and miRNAs are encoded in transcription units and mainly derived from introns of the coding genes or introns and exons of the non-coding genes (Kim, 2005a). Mattick pointed out that the “central dogma” only shows parts of the picture in higher eukaryotic genetics and suggests that ncRNAs may form the basis of an as yet unknown inter-gene communication system which is posited to be parallel to protein synthesis to govern gene
CHAPTER 1. INTRODUCTION

Figure 1.2: The ncRNA-mediated regulatory network
ncRNAs, parallel to protein synthesis, are derived from introns of the coding and the non-coding genes or exons of the non-coding genes. These ncRNAs govern gene regulation at many levels and therefore may form an inter-gene communication system. The figure is taken from (Mattick and Makunin, 2005).

regulation (Mattick, 2001, 2003, 2004; Mattick and Gagen, 2001; Mattick and Makunin, 2005). Although proteins are the major elements involved in cellular processes, they require a variety of small RNAs that serve as recognition sites or catalysts to control gene expressions at almost every step in the gene expression pathway from DNA to RNA to protein (Storz, 2002). RNA-mediated regulatory controls, together with proteins, form a very complex gene network in higher organisms (Mattick, 2001, 2003, 2004; Mattick and Gagen, 2001; Mattick and Makunin, 2005). Figure 1.2 demonstrates a proposed ncRNA-mediated network.

A striking example is miRNAs. Research shows that some of the miRNAs appear in clusters thereby suggesting that they may be co-expressed (Baskerville and Bartel, 2005; Sempere et al., 2003). MiRNAs and their target genes form a complex regulatory network with one miRNA targeting many different genes and one gene targeted by many different miRNAs.
1.4 Why study plant ncRNAs? and why legumes?

The study of ncRNAs in plants lags behind the bacterial, yeast, and animal systems. Indeed, plant miRNAs were discovered 10 years later than those in animals. However, ncRNAs are often expressed in specific tissues and fundamental differences in the anatomy of plants and animals suggest that plant ncRNAs may have unique regulatory roles compared to animal ncRNAs. For instance, RNA molecules are transported in plants via a unidirectional vascular system (Lucas et al., 2001; Yoo et al., 2004) that is not present in animals. Within the plant kingdom the legumes have unique anatomical structures in their capability to interact with soil bacteria. This plant-microbial symbiosis makes root and nitrogen-fixing nodule special organs for legume plant development. Therefore, legume plants are an ideal tool to study meristem development as nodules are the only optional meristem, whereas mutants are lethal in other meristems (e.g., shoot meristem, root meristem). Accordingly, the development of nodules in response to bacterial stimuli is an excellent model to study ncRNA and peptide signalling involved in meristem development.

The discovery of novel legume ncRNAs is needed also because the model plant *Arabidopsis* is not a suitable plant for legume studies given its poor microsynteny with legumes at the genomic level (Frugolia and Harrisb, 2001). Until recently, legume studies have been hampered by the genomic complexity. Two legume plants *Medicago truncatula* and *Lotus japonicus* have been defined as model legumes due to their small, diploid (n=8 and n=6, respectively) genomes (both of them are 500 Mbps), and rapid life cycles. Although the sequencing of these two legumes is incomplete, the number of sequences present in databases now allows me to identify ncRNAs in this study.
1.5 Aim of this study

The aim of this study is fourfold. First, to design a computational approach to identify ncRNAs from two model legume plants Medicago truncatula and Lotus japonicus. Second, to understand the distinctive features of ncRNAs compared to protein-coding genes and non-transcribed sequences. Third, to identify miRNAs and their target genes as well as to provide the community a candidate miRNA database. Finally, to provide further confirmation of the hypothesis that ncRNAs play major roles in the gene regulatory system, especially in meristem development.

1.6 Methodology

The strategy of ncRNA identification is to develop a computational pipeline that uses the publicly available ESTs (Expressed Sequence Tags) and genomic data to screen for all transcribed genes that do not encode proteins. Genomic sequences are used for gene prediction as EST represents only partial gene sequences and therefore are insufficient to predict gene structures. Several characteristics of predicted ncRNAs are then contrasted with two datasets: protein-coding and non-transcribed sets. As miRNA precursors are predominantly located in the exons and introns of non-coding genes and the introns of coding genes (Kim, 2005a), I analyzed the set of ncRNAs and intronic regions for their possibility to form miRNA precursor structures.

1.7 Significance of the study

The present study is the first systematic ncRNA study conducted on legume plants. It is expected that the results of this study will contribute to RNA
research in plants, especially in legumes. The results will also be integrated with corresponding biochemical studies being done elsewhere in the “ARC Centre of Excellence for Integrative Legume Research”, which has funded this project.

1.8 Thesis structure

This thesis is composed of 10 chapters.

Chapter 1 “Introduction” provides a general background of the study as well as discusses topics such as the importance of studying ncRNAs, the aim of the study, the methodology that was employed, and the significance of the study.

Chapter 2 “A review of ncRNA studies” first discusses the reasons that ncRNAs were largely ignored in the past and challenges involved in detecting ncRNAs. It then introduces a review on both experimental and computational approaches to discover ncRNAs. Extant ncRNA databases are also introduced.

Chapter 3 “Computational identification of mRNA-like ncRNAs in legumes” describes the identification of “mRNA-like” ncRNAs in two model legumes Medicago truncatula and Lotus japonicus using a computational pipeline.

Chapter 4 “Characteristic comparisons of ncRNA, mRNA, and ntDNA datasets” investigates ncRNA features that are distinguished from coding RNAs and non-transcribed sequences. Comparisons discussed include base composition biases, global and local RNA secondary structures, expression profiles, and the possibility of bidirectional transcription.

Chapter 5 “RT-PCR primer design for M. truncatula putative ncRNA genes” talks about methods to design RT-PCR primers for the predicted M. truncatula ncRNAs, which will be used to test the expression of ncRNAs experimentally.

Chapter 6 “Prediction of miRNAs and their targets” describes a computa-
ditional pipeline designed to identify miRNAs and their corresponding targets in two model legumes.

Chapter 7 “Characterization of the predicted miRNAs and targets”. Topics include the clustering of highly similar miRNAs, miRNA conservation in other plants, classification of miRNA target genes, and assessment of miRNA predictions.

Chapter 8 “MIRATdb – legume putative miRNA and target database” talks about the construction of a legume database interface for putative miRNAs and targets.

Chapter 9 “Implementations of programming source code” gives examples of four types programming codes used in this study.

Chapter 10 “Conclusion and future work” summarizes the findings of this research, discusses the contributions and limitations of this study, and gives recommendations for the future research.

“References” lists the reference in this thesis.

“Appendices” gives MIRATdb database schema.
Chapter 2

A review of ncRNA studies

In this chapter, I will first discuss why ncRNAs have remained unnoticed for such a long time. Then the current challenges of ncRNA prediction will be discussed. Next, I will review both the computational and biochemical approaches that have been designed to detect ncRNAs. Finally, current major ncRNA databases will be introduced.

2.1 Why ncRNAs were largely ignored in the past?

If ncRNAs are as abundant as we see now, why were they largely ignored in the past? There are two main reasons. One reason is that more attention was paid to proteins than small RNAs. More specifically, it was believed that proteins represented all genes and so another layer of gene products — small non-coding RNAs — were neglected. Both computational gene finding programs and biochemical techniques were designed to detect proteins and study proteomics to the exclusion of ncRNAs. Most gene prediction programs (Burge and Karlin, 1997; Lukashin and Borodovsky, 1998; Salamov and Solovyev, 2000) use signals
such as open reading frames (ORF), codon bias, splice site signals, polyA signals, and start/stop codons, which are only suitable for predicting protein coding genes rather than ncRNAs. The second reason is that ncRNAs often escaped the scope of genetic, biochemical, and molecular methods (Eddy, 2001). Unlike proteins, however, ncRNAs are not usually detectable from gels and do not show significant phenotypes when mutations occur (Eddy, 2001; Mattick, 2004). Hence, protein analytic proteome – based experiments often miss non-protein-coding genes. Indeed, several early discovered ncRNAs, such as line-4 and let-7, were merely accidental findings.

2.2 Challenges to detect ncRNAs

ncRNA prediction is currently in an incipient form. Both experimental and computational approaches have a number of difficulties in detecting ncRNAs. First, ncRNAs lack significant features, such as long open reading frames (ORF) and codon biases, on which most of the available gene prediction methods rely. Second, searching transcripts that do not encode long ORFs is not sufficient to prove that those transcripts do not encode proteins. Different types of start codons, for instance, may lead to erroneous ORF predictions. Furthermore, the frame shifts can easily create or miss a stop codon and therefore create an incorrect ORF length. Third, research has shown that even RNA secondary structure, its signature, is not sufficient to distinguish RNAs from random sequences alone (Rivas and Eddy, 2000). Furthermore, designing proper expression-based experiments to detect novel ncRNAs is difficult because ncRNAs often undergo specific developmental stages or they are in specific cell lines and tissues. For examples, the well-studied microRNAs line-4 and let-7, which control developmental timing in worm larvae, are temporally expressed (Lee et al., 1993; Reinhart et al., 2000);
and the brain specific RNAs (Bsr RNA, BC1 RNA, and BC200 RNA) are only expressed in nervous systems (Komine et al., 1999; Skryabin et al., 1998) Finally, ncRNAs have a variety of types, origins, functions, and pathways, but we know little about them. For instance, ncRNAs may be transcribed by different RNA polymerases and processed in different regions (intron, intergenic, exon of protein-coding, and non-coding genes) (Mattick, 2003, 2004; Mattick and Gagen, 2001; Mattick and Makunin, 2005). Even for the known ncRNAs, which only make up a very small part of the whole ncRNA network, we still do not know their exact mechanisms. Taking miRNA as an example, despite widespread efforts to study miRNAs, we still have not completely worked out their mechanisms. Due to the above problems, the detection of ncRNAs has become a big challenge to both bioinformaticians and biologists.

2.3 Review of experimental and computational approaches for identifying ncRNAs

2.3.1 Experimental methods

ncRNAs have traditionally been studied on the wet bench individually. Several groups have recently carried out large-scale ncRNA studies experimentally by cDNA (Complementary Deoxyribonucleic Acid) cloning and high-density arrays. One such study was undertaken by Tom Gingeras and colleagues (Kapranov et al., 2001). They isolated and purified polyadenylated transcripts from 11 human cell lines and extracted the transcripts of short length (less than 200 bp). Meanwhile, they made three Affymetrix microarrays to cover human chromosomes 11 and 12. cDNAs made from those transcripts were subsequently hybridized to the probes on the arrays. Their results showed that about 60% of the transcripts did not map
onto any known exons thereby suggesting that the transcripts were non-coding RNAs.

Tom Gineras and colleagues recently improved their techniques by combining rapid amplification of cDNA ends (RACE) and high resolution tiling arrays to further study the human transcriptome (Cheng et al., 2005). These new techniques allowed them to cover a larger range of human transcripts. Their elegant work has not only generated polyadenylated transcripts but also non-polyadenylated transcripts.

In a similar strategy, together with transcriptome analysis, other groups have also identified many non-coding transcripts outside annotated genes in human and other species (Bertone et al., 2004; Cawley et al., 2004; Marker, 2002; Okazaki, 2002). Those findings indicate that a large proportion of the genome is transcribed and many of these transcripts tend to be non-coding.

The cDNA cloning and the tiling array techniques have some limitations, however. First, they may not detect small RNAs, like miRNAs, because it is difficult to clone very short RNAs and purify RNAs from cDNA library contaminations and rRNAs (Frith et al., 2005; Lagos-quintana et al., 2003; Mattick and Makunin, 2005). In addition, tiling arrays do not tell which strand is transcribed and do not show clear exon boundaries (Frith et al., 2005). These techniques will become more useful as microarray and cloning technology continues to improve.

2.3.2 Computational approaches

Given the wealth of genomic data and new computational tools and methods that are becoming available, computational approaches will become powerful tools for ncRNA discovery. Several computational methods have been developed to identify ncRNAs although they are still in the early stage. Those approaches can
be categorised into three classes: comparative genomic approaches, transcript screening approaches, and special ncRNA predictions. Here I give a brief review of these approaches.

**Comparative genomic approach**

The evolutionally related ncRNAs may conserve RNA secondary structures across different species, especially in closely related species. The human “ultraconserved elements” are striking examples. Those “ultraconserved elements” are located in exonic or intronic regions and are more than 95% identical with other animals including mice, dogs, and chickens, which indicates that some of those regions may contain ncRNAs (Bejerano et al., 2004). Motivated by the knowledge of RNA structure conservation, several groups are using genomic comparative analysis to search for novel ncRNAs that are structurally homologous to known ncRNAs. The easiest and fastest way is to use BLAST (Altschul et al., 1997). However, the BLAST homologous search is based on consensus sequences rather than consensus structures and so is not a suitable program for finding related RNAs which share structure similarities. The first program that used this approach was QRNA (Rivas and Eddy, 2001). It was designed to search for mutation patterns by pairwise aligning two structurally homologous sequences. QRNA has been used for identifying ncRNAs in bacteria and yeast. Another two programs, DDBRNA (Bernardo et al., 2003) and MSARI (Coventry et al., 2004), have also employed conserved structures to detect ncRNAs by multiple alignments instead of pairwise alignments. A recent program, RNAZ (Washietl et al., 2005), combined conserved structures and folding minimum free energy (MFE) stabilities to detect ncRNAs.

The main weakness associated with extant comparative genomic based ncRNA prediction programs is that they are not suitable for detecting novel ncRNAs in complex higher eukaryotes (Numata et al., 2003; Washietl et al., 2005). Further,
most of these programs need information of RNA secondary structure conserva-
tion across species which may not be available for many identified ncRNAs.

**Transcript screening approach**

The principle of this approach is to mine transcripts such as ESTs or cDNAs that
do not have coding properties. One class of ncRNAs acts like mRNAs in that they
are spliced, polyadenylated, and possibly 5’ capped and are termed “mRNA-like
ncRNAs” (Erdmann et al., 2000). Compared to the tiny microRNAs and small
interfering RNAs, these ncRNAs are usually longer in size ranging from hundreds
of base pairs to more than 10k base pairs (Storz, 2002). Therefore, we can detect
those mRNA-like ncRNAs through the screening of polyA-rich EST or cDNA
libraries. MacIntosh and colleges first used this approach to identify 19 ncRNA
and peptide genes from Arabidopsis ESTs (MacIntosh et al., 2001). One group
searched for ncRNAs from RIKEN mouse full-length cDNAs and characterized
ncRNA candidates by whether they had corresponding ESTs and transcriptional
signals like CpG islands and polyA signals (Numata et al., 2003). Another group
(Tupy et al., 2005) screened ncRNAs that did not contain long ORFs from both
ESTs and cDNAs in the *Drosophila melanogaster* genome and used RT-PCR
assays and Northern Blot to verify the ncRNA expression.

The transcript based approach is a feasible computational approach to detect
novel ncRNAs on a large scale even without the aid of known ncRNAs. Further-
more, we can apply this method, together with ncRNA characteristic analysis, to
almost any organism. Finally, this approach can take advantage of using publicly
available EST and cDNA data. The limitation of this approach, however, is that
it can not detect non-polyadenylated ncRNAs. Nonetheless, this limitation may
be resolved when non-polyadenylated transcripts are cloned and sequenced.
2.4. CURRENT NCRNA DATABASES

Special ncRNA predictions

Several specialized ncRNA methods were designed to find special types of ncRNA, such as tRNAscan-SE which was designed to find tRNAs (Lowe and Eddy, 1997), snoscan for snoRNAs (Lowe and Eddy, 1999), and SRP scan for SRP RNAs (Regalia et al., 2002). Several systematic searches for miRNAs (miRNA prediction method will be reviewed in Chapter 6) and antisense RNA genes have also been carried out (Dahary et al., 2005; Kiyosawa et al., 2003; Yelin et al., 2003).

2.4 Current ncRNA databases

Several non-coding RNA databases that comprise the known ncRNAs are available on-line.

2.4.1 Rfam

Rfam (Griffiths-Jones et al., 2005), which is analogous to Pfam (Bateman et al., 2004), is a comprehensive non-coding RNA family database. Rfam classifies the known RNAs into families by consensus RNA secondary structures and uses multiple sequence alignments of covariance models to represent those families (Griffiths-Jones et al., 2005). The current version of Rfam (release 7.0) includes 503 RNA families. Rfam is available at \url{http://www.sanger.ac.uk/Software/Rfam/} and \url{http://rfam.wustl.edu/}. One can use the potential ncRNA sequences to search for homologies to known ncRNAs on the Rfam web server. One can also download and install Rfam covariance models on the local machines, and then use a software package, INFERNAL (Griffiths-Jones et al., 2005), to batch search homologous ncRNAs against the Rfam database locally. INFERNAL...
NAL, similar to HMMER (Bateman et al., 2004) in Pfam, also allows us to make customized consensus RNA structure profiles and align more sequences to these profiles to create new structural multiple alignments (Griffiths-Jones et al., 2005).

2.4.2 ASRP

ASRP (Arabidopsis Small RNA Project Database) (Gustafson et al., 2005) contains Arabidopsis small RNAs including miRNAs, siRNAs, and tasi-RNAs collected from both in-house cloning analysis and public available resources. The current version 3.0 of the database contains 1,920 unique sequences. The database is available at http://asrp.cgrb.oregonstate.edu.

2.4.3 RNAdb

RNAdb is a mammalian ncRNA database. It covers more than 800 ncRNAs that were discovered experimentally in animals (Pang, 2005). RNAdb is available at http://research.imb.uq.edu.au/rnadb/.

2.4.4 NONCODE

NONCODE (Liu, 2005) collects all types of ncRNAs from the literature except for tRNAs and rRNAs. It also provides the ncRNA information on their functions, chromosome locations, and sources. The current release (v1.0) consists of 5,339 unique sequences from 861 organisms. NONCODE can be accessed at http://noncode.bioinfo.org.cn/.
2.4. CURRENT NCRNA DATABASES

2.4.5 The noncoding RNA (ncRNA) database

This database lists the known ncRNAs that are grouped into vertebrates, plants, bacteria, and insects. It does not include microRNAs and snoRNAs. It is available at http://biobases.ibch.poznan.pl/ncRNA/)

2.4.6 Other ncRNA databases

Aside from the comprehensive ncRNA databases listed above, a number of specialized ncRNA databases are also available on-line, such as miRNA and target database miRBase (Griffiths-Jones et al., 2006), rRNA database (Wuyts, 2001), uRNA database (Zwieb, 1997), SRP database (Gorodkin et al., 2001), and others (Cannone, 2002; Klosterman et al., 2002; Knudsen et al., 2001; Szymanski et al., 2002; van Batenburg et al., 2001; Williams, 2002).
Chapter 3

Computational identification of mRNA-like ncRNAs in legumes

3.1 Prediction of ncRNA candidates

3.1.1 The strategy of ncRNA prediction

A transcript selection method was designed to screen all EST sequences and then exclude members of other classes of RNAs. The rationale for identifying ncRNAs from EST collections is summarized in the following three points. First, many ncRNA genes are polyadenylated and should therefore be present in EST libraries (MacIntosh et al., 2001). Second, requiring candidate ncRNA sequences to be present in EST libraries ensures that they are transcribed and thus avoid many potential false predictions. Third, there is a considerable amount of publicly available ESTs including additional information such as the tissue of expression, reisolation frequencies, and microarray data which assist in ncRNA characterization and future functional analysis. There are, however, two major disadvantages associated with EST sequences. First, the EST sequences usually have a high
error rate. Second, ESTs represent only partial gene sequences which are often insufficient to predict their coding status. The analysis of the genomic regions that correspond to an EST avoids both limitations and increases prediction accuracy. I therefore retrieved the genomic sequence regions corresponding to ESTs, and subsequently restricted the analyses to ESTs for which genomic sequences are available. Although this greatly reduces the number of ncRNA candidates, it greatly enhances the accuracy of prediction. I have generally aimed at avoiding false positive prediction at the expense of completeness.

It is well known that ESTs have high redundancies, high error rates (∼1/100nt), and frame shift errors due to their single-pass readings from the 5’ ends or the 3’ ends of cDNA clones (Cerutti, 2003). However, the quality of EST data can be greatly improved by clustering and assembling them into consensus sequences because minimizing redundancies can correct errors. Currently, there are two main ways to access EST data: one is through TIGR Gene Indices (http://www.tigr.org/tdb/tgi/) and another is through NCBI UniGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene). Methods of constructing ESTs for both TIGR and UniGene initially involve pre-processing to remove noisy sequences such as vector contaminations, repeat sequences, and low complexity sequences. TIGR also trims poly A and poly T tails. The two methods mainly differ in that TIGR clusters and assembles ESTs into consensus sequences termed Tentative Consensus (TC), whereas UniGene merely groups similar genes and alternatively spliced transcripts into clusters (Liang et al., 2000). Figure 3.1 demonstrates the EST construction procedures in TIGR and UniGene (Cerutti, 2003). The benefits of assembled consensus sequence in the TIGR database are that they largely reduce sequence redundancies, that they distinguish similar genes and alternatively spliced transcripts, and that they produce longer sequences. I therefore chose to use TIGR ESTs for the present work.
3.1. PREDICTION OF NCRNA CANDIDATES

Figure 3.1: EST constructions in TIGR and Unigene. TIGR clusters and assembles ESTs into consensus sequences whereas UniGene groups similar genes and alternatively spliced transcripts into clusters but does not assemble them into consensus sequences. The figure is taken from (Cerutti, 2003)
My methodology makes use of the genomic sequence regions that correspond to the ESTs as these were used to map ESTs and to predict genes as well as to assess the coding potentials. TIGR also provides genomic sequences in the form of assembled BAC sequences, which were used in this work. ESTs that overlapped with the predicted protein genes were removed and additional purification steps including the constraint of ORF length as well as sequence similarity to other proteins were used to exclude those remaining ESTs potentially encoding proteins. Figure 3.2 illustrates the ncRNA prediction strategy using both ESTs and genomic sequences.

3.1.2 The ncRNA prediction pipeline

Figure 3.3 gives a flowchart of the computational pipeline designed to detect putative ncRNAs and small peptide genes. The pipeline is initiated with the collection of both genomic data and TIGR Gene Index sequences (GIs). Several purifica-
3.1. PREDICTION OF NCRNA CANDIDATES

tion steps comprises the pipeline including that of mapping ESTs to BACs, gene prediction in BACs, the exclusion of ESTs possessing long ORFs, the exclusion of GIs homologous to known proteins, and the elimination of tRNAs and rRNAs. Each step is described in detail below.

Data collection

I used *M. truncatula* Gene Index (MtGI) sequences (Release 8.0, 19 January 2005) and *L. japonicus* Gene Index (LjGI) sequences (Release 3.0, 23 June 2004) from TIGR. The current *M. truncatula* release provides a total of 36,878 GI sequences (18,612 TCs, 18,238 Singleton ESTs, and 28 ETs) and is presumably the final release of this database. The current release of *L. japonicus* consists of 28,460 LjGI sequences (12,485 TCs, 15,919 Singleton ESTs, and 56 ETs). They are collectively referred to as GIs.

As *M. truncatula* and *L. japonicus* genome sequencing have yet to be completed, I used BAC (Bacterial Artificial Chromosomes) sequences. There are four groups currently working on *Medicago truncatula* genome sequencing and they submit their BAC sequences to Genbank. TIGR developed an automated *M. truncatula* sequence contig pipeline that retrieves those BAC sequences from Genbank daily and then, in a similar fashion to EST assemblies, clusters and assembles BACs into BAC Contigs. Sequences of *M. truncatula* these BAC Contigs were downloaded on 5 March 2005 (~129M bp) from TIGR ([http://www.tigr.org/tdb/e2k1/mta1/](http://www.tigr.org/tdb/e2k1/mta1/)). These ~129M bp of sequence cover about 26% of the *M. truncatula* genome.

*Lotus japonicus* BAC sequences (~269M bp, 20 May 2005, unpublished) were obtained through collaborations with the Kazusa DNA Research Institute. The genome sizes of *L. japonicus* are 472 M bp. Thus, the BAC sequences that I obtained represent approximately 58% of the *L. japonicus* genome.
Figure 3.3: Flowchart of the computational pipeline used to identify mRNA-like ncRNAs in *M. truncatula* and *L. japonicus*. The individual purification steps (blue box) and the resulting datasets (grey box) are shown. Each step is described in detail in the text.
Both downloaded GI and BAC sequences in FASTA-format were parsed by a Perl script and stored in a relational MySQL database for further analysis.

### Mapping GIs to BACs

Aligning EST sequences with genomic sequences differs from the conventional sequence alignments in that the special consideration must be given. That is, the sequence alignment should be robust to EST sequence error, and the alignments need to take into account the insertion of introns, including intron start and stop at the splice consensus sites GT and AG. There are several programs available to align EST/mRNA to the genomic sequences including est2genome (http://emboss.sourceforge.net/apps/est2genome.html), sim4 (Florea et al., 1998), and BLAT (Kent, 2002). I used est2genome, a part of the EMOBSS suite (Rice et al., 2000), because it produces all the information required for further analysis. Given that est2genome is based on the dynamic programming algorithm, it is notably compute-intensive. My benchwork estimated that it would take several years to complete the process of TIGR ESTs mapping to BACs on a Linux machine. This problem, however, could be circumvented by excluding the alignments of sequence with no similarities. Consequently, I initially mapped GIs to BACs with BLAST (version 2.1.10) (Altschul et al., 1997) to speed up the mapping process. The matched GI–BAC pairs were then further refined with the est2genome program.

As GIs are transcripts of genomic sequences, there is nearly identical sequence similarity between them. This allows me to apply a very stringent threshold of E-value $\leq 1E-20$ to the BLAST mapping of GIs to BACs. To implement the BLAST mapping, I installed the BLAST program on a Linux machine and developed a BioPerl program to run a BLAST search locally and extracted such information
as identifiers of GIs and BACs, E-value, score, identity, matched position, and sequence strand, and stored them into in a MySQL database.

Another Perl script “run-est2genome” extracted matched GI–BAC pairs with E-value \( \leq 1E-20 \) from the database and then further mapped them with the est2genome program. The est2genome parameters that I chose were: a match of 1, a mismatch of 3, a gap-penalty of 2, an intron-penalty of 20, a splice-penalty of 10, and a min-score of 10, which can correctly detect exons as short as 14 bp correctly (commented at \texttt{http://emboss.sourceforge.net/apps/est2genome.html}). GIs with \( \geq 95\% \) identity to BACs and \( \geq 90\% \) sequence overlaps with BACs were selected and stored in the database for further analysis. The run-est2genome ran parallel on a 20 node Linux Cluster to speed up the procedure. Examples of the implementation of the Bioperl scripts and high performance computing in this thesis are described in Chapter 9. A total of 27,164 (74\%) MtGIs and 12,431 (44\%) LjGIs that did not meet my mapping threshold were eliminated after this step.

\textbf{Gene prediction in BACs}

In order to choose the most accurate protein-coding gene prediction program available, I compared the performances of three commonly used gene predictors, GenScan (Burge and Karlin, 1997), GeneMark.hmm (Lukashin and Borodovsky, 1998), and Fgenesh (Salamov and Solovyev, 2000), both from literature and on my data.

The literature suggests that GenScan was primarily designed in order to predict gene structures on vertebrate genomes and to rely on organism-specific parameters so as to improve the accuracy of prediction on plant genomes. However, the current GenScan version includes no legume-specific parameters. GeneMark.hmm was deemed to be the most efficient tool to predict \textit{Arabidopsis} genes,
3.1. PREDICTION OF NCRNA CANDIDATES

based on the evaluation of several gene prediction programs (Pavy et al., 1999). Like GenScan however, it includes no legume training set. Further, it neither predicts transcription start (TSS) nor polyA sites.

To my knowledge, Fgenesh is the only gene predictor that has been trained on legumes. The program consistently predicted more genes than majority of the other methods according to both previous research (Yu et al., 2002) and my results. For example, a total of 18,158 \textit{M. truncatula} BAC sequence genes were predicted by GenScan, 18,408 by GeneMark.hmm, whereas 29,506 by Fgenesh. Also of importance was that Fgenesh performed much faster (at least 50 times) than the other two gene predictors.

I therefore chose Fgenesh to predict protein coding genes in the BAC contig sequences based on the aforementioned performance comparisons. The implementation and extraction of gene structure of exon positions, TSS and polyA sites, and strand were accomplished with a Perl script. In this step, 29,506 \textit{M. truncatula} and 86,134 \textit{L. japonicus} protein coding genes were predicted.

**Exclusion of the GIs overlapping the predicted genes**

An EST ideally contains only exons and matches exactly to a predicted gene. In reality, all sorts of matches result due to the fact that both the gene predictor and the program that performs EST mapping to genome could wrongly predict intron and exon boundaries, transcriptional start sites, and translational start sites. I therefore developed a measurement of “relative overlap” to determine whether a GI represents a predicted gene. I calculated the overlaps between GIs and the predicted exons including untranslated regions (UTRs); both of them are located on the BAC sequences. The overlaps, measured in base pairs, were then converted to a relative overlap in the range of 0–1 by dividing by the total GI length mapped to the genomic BAC sequences. Figure 3.4 illustrates how the relative overlap
Figure 3.4: Relative overlap calculations. First, overlap length between GIs and predicted genes in BACs were calculated, in the example, by summing up Overlap 1 + Overlap 2. Next, total length of the “GI-on-BAC” was computed (in the example, GI in BAC length = giCDE1 + giCDE2 + giCDE3). Finally, relative overlap was calculated by dividing overlap length by GI-on-BAC length. The prdCDE segments represent the predicted cDNA elements in BAC. giCDE represents cDNA elements of GI mapping to BAC.

was calculated. To determine the optimal cutoff value for separating noncoding transcripts from others, I plotted and compared the distributions of the relative overlap in both legumes (Figure 3.5). Both clearly show that the relative overlap can be divided into three ranges of ≤ 10%, ≥ 90%, and 10% – 90%. Consequently, I regard all ESTs with an relative overlap of equal or more than 90% as protein encoding transcripts. The GIs with an overlap size between 10% and 90% were inconclusive. Only GIs with a relative overlap of 10% or less, 853 MtGIs and 2,119 LjGIs were retained for further analysis (Figure 3.3).

Exclusion of the GIs with long ORFs

To discard the sequences that failed to be identified as protein-coding genes from the previous steps, I applied an ORF length filter to the remaining sequences. In doing so, I used the EMBOSS getORF program (Rice et al., 2000) to predict the longest possible ORF length for each transcript from either stop codon or the start of the sequence to either the stop codon or the end of the sequence in both strands, and then eliminated sequences that possessed ≥ 100 codons. The longest ORF length was calculated by taking from either stop codon or the start of the sequence
3.1. PREDICTION OF NCRNA CANDIDATES

Figure 3.5: Relative overlap density distributions in *M. truncatula* and *L. japonicus* plotted using Kernel Density Estimate (KDE)

to either the stop codon or the end of the sequence in both strands. The same value has been widely employed in other ncRNA prediction studies (MacIntosh et al., 2001; Numata et al., 2003; Tupy et al., 2005). I compared the threshold with the longest ORF distributions of the protein coding sequences - 92% of the distribution of protein coding sequences were excluded by this threshold, and so given the 53 genes that were excluded by this filter, theoretically about four *M. truncatula* and 38 *L. japonicus* coding genes would be expected to pass this filter.

A total of 802 MtGIs and 1,637 LjGIs fulfilled this constraint.

**Exclusion of the GIs homologous to known proteins**

In some cases, protein coding gene finders may miss genes with atypical features such as short exons. To limit the number of missed genes that may correspond to the ESTs, I Blast searched the remaining GIs for sequence similarities to any known, putative, and hypothetical proteins represented in Swiss-Prot (http://au.expasy.org/sprot/), trEMBL (http://au.expasy.org/sprot/),
Among 128 MtGI that matched sequences in those databases with an E-value of 1E-5 or less, ten of them matched to known proteins while others matched to sequences either annotated as hypothetical proteins or as predicted proteins. The remaining LjGI did not match given the threshold of E-value \(\leq 1E-5\). After applying this filter, 674 \(M.\ truncatula\) and 1,637 \(L.\ japonicus\) GIs were retained.

**Elimination of tRNAs and rRNAs**

As tRNAs and rRNAs are transcribed by PolII or PolIII, they are not polyadenylated, and as consequence should not be present in a polyA-rich EST library. Yet, it is possible that some of them made it into the EST libraries as these RNAs may be co-purified with polyA(+) RNAs. To exclude tRNAs and rRNAs, Blast searches to the nucleotide database (downloaded in April 2005 from ftp://ftp.ncbi.nlm.nih.gov/blast/db/) were used to identify GIs similar to tRNAs and rRNAs (E-value \(\leq 1E-5\)). In addition, tRNAscan-SE (version 1.21) (Lowe and Eddy, 1997) was also used to search for tRNAs. Taken together, only one sequence was found with this filter, confirming that tRNAs and rRNAs are not significantly present in EST libraries.

### 3.1.3 Summary

The ncRNA prediction pipeline began with the TIGR \(M.\ truncatula\) and \(L.\ japonicus\) Gene Index database which contains 36,878 and 28,460 GIs, respectively. In the analyzed releases of genomic sequences, \(L.\ japonicus\) possesses twice the sequenced BAC size as \(M.\ truncatula\) (269 Mbp vs 129 Mbp). This is reflected in my analysis showing that a higher proportion of LjGIs (16,029, 56\%) are associated with corresponding genomic sequences than MtGIs (9,714, 26\%).
3.1. PREDICTION OF NCRNA CANDIDATES

Table 3.1: Purification summary of *Medicago truncatula* and *Lotus japonicus* ncRNAs and peptide genes

<table>
<thead>
<tr>
<th>Computational analysis method</th>
<th>Cutoff</th>
<th>Remaining GIs in <em>M. truncatula</em></th>
<th>Remaining GIs in <em>L. japonicus</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial GI collections from TIGR</td>
<td></td>
<td>36,878</td>
<td>28,460</td>
</tr>
<tr>
<td>GIs mapping BACs (initial mapping using BLAST)</td>
<td>E-value ≤ 1E-20</td>
<td>15,297</td>
<td>22,364</td>
</tr>
<tr>
<td>GIs mapping BACs (second mapping using est2genome)</td>
<td>identity ≥ 95% and GI coverage ≥ 90%</td>
<td>9,714</td>
<td>16,029</td>
</tr>
<tr>
<td>GI overlapping predicted genes (gene prediction using Fgenesh)</td>
<td>overlap fraction ≤ 10%</td>
<td>853</td>
<td>2,119</td>
</tr>
<tr>
<td>Longest ORF length</td>
<td>≤100 codons</td>
<td>802</td>
<td>1,637</td>
</tr>
<tr>
<td>GI homologous to known proteins</td>
<td>E-value ≤ 1E-5</td>
<td>674</td>
<td>1,637</td>
</tr>
<tr>
<td>find tRNAs and rRNAs</td>
<td>E-value ≤ 1E-5 and tRNAs predicted by tRNAScan</td>
<td>673</td>
<td>1,637</td>
</tr>
</tbody>
</table>

After removing the GIs that were significantly overlapped with the predicted genes, 853 (2.3%) MtGIs and 2,119 (7.4%) were retained. A total of 674 MtGIs and 1,637 LjGIs survived both the longest ORF and homologues to known protein filters. The tRNA and rRNA search eliminated one sequence. Differences in both genomic and GI sequences resulted in 673 MtGIs and 1,637 LjGIs forming the sets of putative ncRNAs. Note that small peptides may have been missed by this pipeline and so the final set may include some small peptide RNAs. Table 3.1 summarizes the purification steps used to identify ncRNAs in *M. truncatula* and *L. japonicus*. 
3.2 Analysis of ORF distribution in putative ncRNA set

3.2.1 Strategy

Although sequences with discernable ORFs have been excluded, short peptide genes may cause the average ORF length of the ncRNA set to be longer than what would be expected by chance. To further analyze the coding potentials of the ncRNAs, I compared the longest ORF length of ncRNAs to that of randomized sequences.

3.2.2 Methods

To test whether ORF lengths of sequences in the ncRNA set differ from that of random sequences, I performed a hybrid bootstrap sampling test. The longest ORF for each sequence in the ncRNA set was computed by taking from either stop codon or the start of the sequence to either the stop codon or the end of the sequence in both strands. I then randomly shuffled the ncRNA sequences retaining the base composition and subsequently computed the longest ORF for each shuffled sequence. The ORF lengths from both sets were merged into one set. I then drew two pseudosamples from the merged set by sampling with replacement and recorded the mean differences of ORF length between them. After repeating this procedure 1000 times, I obtained a distribution (Figure 3.6) that was used to determine the probability of ORF length differences between actual and randomized sequences under the null hypothesis that two sequence sets do not possess significant ORF length differences, using a two-sided test. A R script (see Chapter 9) was written to perform the hybrid bootstrap sampling for the longest ORFs of the coding and non-coding RNAs.
### 3.3. CONSERVATION OF PUTATIVE NCRNAS IN OTHER SPECIES

Table 3.2: Summary of the longest ORF distributions in ncRNA set and the randomized same sequence set

<table>
<thead>
<tr>
<th></th>
<th>Min. (aa)</th>
<th>1st Quartile (aa)</th>
<th>Median (aa)</th>
<th>Mean (aa)</th>
<th>3rd Quartile (aa)</th>
<th>Max (aa)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>M. truncatula</em> (ncRNA set)</td>
<td>20</td>
<td>49</td>
<td>60</td>
<td>60</td>
<td>71</td>
<td>100</td>
</tr>
<tr>
<td><em>M. truncatula</em> (randomized set)</td>
<td>19</td>
<td>44</td>
<td>55</td>
<td>57</td>
<td>68</td>
<td>129</td>
</tr>
<tr>
<td><em>L. japonicus</em> (ncRNA set)</td>
<td>19</td>
<td>51</td>
<td>62</td>
<td>62</td>
<td>75</td>
<td>100</td>
</tr>
<tr>
<td><em>L. japonicus</em> (randomized set)</td>
<td>23</td>
<td>47</td>
<td>58</td>
<td>60</td>
<td>70</td>
<td>142</td>
</tr>
</tbody>
</table>

#### 3.2.3 Results and Discussion

The estimation for the significance of the measured ORF length as determined by the bootstrap samplings showed there to be no significant differences between the assessed ncRNAs and the randomized sequences (p = 0.88 for *M. truncatula* and p = 0.94 for *L. japonicus*). On average ncRNAs and randomized sequences have ORF length differences of only three amino acids for both legumes. Such a finding suggests that the small ORFs found in the ncRNA set are likely to have occurred by chance, and are very similar to those in the random sequences. This provides further evidence that ORFs encoded in the putative ncRNA set are indeed short in length and that their likelihood for encoding proteins is slight. Table 3.2 summarizes the longest ORF distributions in ncRNA set and the randomized same sequence set.

#### 3.3 Conservation of putative ncRNAs in other species

To analyze whether the predicted ncRNAs are conserved in other species, I searched for sequences that are similar to transcripts in other species using BLASTN. BLAST searches of the predicted 673 *M. truncatula* and 1,637 *L. japonicus*...
**Figure 3.6:** The bootstrapped sampling distribution for the longest ORF. The graphs demonstrate whether the ORF length occurs by chance. The distributions that are expected by chance for the mean difference of the longest ORF length between actual and randomized sequences are given for the ncRNA set. These distributions were obtained by 1000 iterations hybrid bootstrap sampling (see Methods). The red arrow points to the mean ORF difference of the actual sequences and the black arrow points to the critical value to reject the null hypothesis at the .05 level. For the ncRNA set, the mean ORF difference of the actual sequences falls approximately in the middle of the distribution of the bootstrapped mean ORF difference expected by chance, far from the rejection region (p = 0.88 for *M. truncatula* and p = 0.94 for *L. japonicus*), indicating that ncRNA ORFs occur by chance.
3.4. RANKING PUTATIVE NCRNAS

japonicus ncRNA genes against NCBI dbEST (http://www.ncbi.nlm.nih.gov/dbEST) found 126 M. truncatula and 435 L. japonicus ncRNAs matching to the sequences of other species with an E-value of 1E-5 or less (Figure 3.7). In M. truncatula, all of these ncRNAs match to other plants and only 4 ncRNAs match to other non-plant species, albeit with relatively high E-values. It further showed that 363 L. japonicus ncRNAs are homologous to sequences in other species, that 362 of them match to other plants, that five match to non-plants species, and that only one matches non-plants. This is consistent with the finding that the majority of Arabidopsis ncRNAs have been shown to be plant specific (MacIntosh et al., 2001). Of the 126 M. truncatula ncRNAs that are apparently homologous to other species, 117 (93%) match to those of other legumes, 90 (71%) are exclusive to legumes, and 27 (21%) occur in both legumes and other plants. Only 9 (7%) showed sequence similarities to non-legume plants. Even more pronounced, 353 out of 363 (97%) L. japonicus ncRNAs match to other legumes, 311 of them (86%) are exclusively conserved in legumes, 42 match to both legumes and non-legumes, and 13 match to non-legumes. This finding further suggests that ncRNA sequences tend to be conserved in closely related species only, indicating that some ncRNAs evolve rapidly and therefore lose sequence similarities with homologues in other taxa. It further suggests that many of the ncRNAs have legume-specific functions.

3.4 Ranking putative ncRNAs

In the future, we plan to experimentally validate the predicted ncRNAs via collaborations. Due to the cost and time involved in biochemical experiments, many experimentalists are interested in only a few of the best candidates to perform these tests. It is therefore helpful to provide rankings of these ncRNA transcripts.
Figure 3.7: The proportion of ncRNAs conservation in (A) plants and non-plants and (B) legumes and non-legumes are given in Venn diagrams to assist experimental design. I ranked all the predicted ncRNAs according to the following criteria: (1) identity and coverage of the GI that matches the BAC, (2) relative overlap size, (3) absolute minimum free energy density, (4) inverted repeat density, (5) potential ORF length, and (6) GI annotation from TIGR. These criteria were developed as characteristics to distinguish ncRNA from other types of sequences; the first four criteria were defined through the ncRNA prediction and the characterization of ncRNAs (see section 3.1 and Chapter 4). The calculation of the longest ORF were described in 3.2. Given that the longest ORF criterion provides only an approximate estimate of coding potentials for an assessed sequence, the Diogenes (Crow and Retzel, 2005) and BESTORF (www.softberry.com) programs were employed to render a more accurate coding estimation. Both report the likelihood of protein coding potentials for a given short sequence based on codon usage. The ranked 673 *M. truncatula* ncRNAs and 1,637 *L. japonicus* ncRNAs are summarized in Table 3.3.
### 3.4. RANKING PUTATIVE NCRNAS

Table 3.3: ncRNA Ranks

<table>
<thead>
<tr>
<th>Rank</th>
<th>Num of ncRNAs (M. truncatula)</th>
<th>Num of ncRNAs (L. japonicus)</th>
<th>GI / BAC match (identity/coverage)</th>
<th>Overlap predicted genes</th>
<th>Diogenes(^a) ORF length</th>
<th>BestORF(^b) length</th>
<th>Longest ORF(^b)</th>
<th>Stability of secondary structure(^c)</th>
<th>Inverted repeats density(^d)</th>
<th>TIGR GI annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank 1</td>
<td>16</td>
<td>0</td>
<td>≥96 / ≥96%</td>
<td>no</td>
<td>No ORF</td>
<td>≥100</td>
<td>≤100</td>
<td>≥0.03</td>
<td>≥0.9</td>
<td>no</td>
</tr>
<tr>
<td>Rank 2</td>
<td>15</td>
<td>29</td>
<td>≥88 / ≥96%</td>
<td>no</td>
<td>No ORF</td>
<td>≥100</td>
<td>≤150</td>
<td>≥0.01</td>
<td>≥0.4</td>
<td>no</td>
</tr>
<tr>
<td>Rank 3</td>
<td>30</td>
<td>0</td>
<td>≥97 / ≥96%</td>
<td>no</td>
<td>No ORF</td>
<td>≤100</td>
<td>≤150</td>
<td>≥0.03</td>
<td>≥0.9</td>
<td>no</td>
</tr>
<tr>
<td>Rank 4</td>
<td>51</td>
<td>209</td>
<td>≥95 / ≥96%</td>
<td>no</td>
<td>No ORF</td>
<td>≥100</td>
<td>≤150</td>
<td>≥0.01</td>
<td>≥0.4</td>
<td>no</td>
</tr>
<tr>
<td>Rank 5</td>
<td>117</td>
<td>209</td>
<td>≥96 / ≥96%</td>
<td>≤10%</td>
<td>≤100</td>
<td>≤100</td>
<td>≤200</td>
<td>150</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>Rank 6</td>
<td>116</td>
<td>305</td>
<td>≤96 / ≥96%</td>
<td>≤10%</td>
<td>≤150</td>
<td>≥100</td>
<td>≤200</td>
<td>150</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>Rank 7</td>
<td>148</td>
<td>307</td>
<td>≥96 / ≥96%</td>
<td>≤10%</td>
<td>≤150</td>
<td>≥100</td>
<td>≤200</td>
<td>150</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>Rank 8</td>
<td>95</td>
<td>258</td>
<td>≥96 / ≥96%</td>
<td>≤10%</td>
<td>≤150</td>
<td>≥100</td>
<td>≤200</td>
<td>150</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>Rank 9</td>
<td>81</td>
<td>320</td>
<td>≥96 / ≥96%</td>
<td>≤10%</td>
<td>≤150</td>
<td>≥100</td>
<td>≤200</td>
<td>150</td>
<td>300</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)Diogenes, the ORF prediction program. Available from [http://analysis.ccgib.umd.edu/diogenes/index.html](http://analysis.ccgib.umd.edu/diogenes/index.html)


\(^c\)The absolute MFED difference between the two complementary strands

\(^d\)The absolute IRD difference between the two complementary strands
Chapter 4

Characteristic comparisons of the ncRNA, mRNA, and ntDNA datasets

4.1 The purpose of the characteristic comparison studies

Non-coding RNAs are an emerging class of transcripts with intriguing characteristics. This thesis sets an initial stage toward uncovering the function of this type of RNA in legumes. The characterization of ncRNAs will not only help provide evidence to distinguish them from other types of molecules, it may also aid in future prediction and functional analyses. Hence, I further characterized the ncRNA sequences and explored sequence feature characteristics for ncRNA genes by comparing ncRNAs with coding RNAs (mRNAs) and with non-transcribed DNA sequences (ntDNA).

I examined a set of features for their ability to distinguish this putative ncRNA
set, including features for base composition biases and RNA structure strand asymmetry, and analyze them individually and in combination using Support Vector Machine classification. I further investigated the tissue-expression patterns of the putative ncRNA set by computational analysis of the EST isolation frequencies in various plant tissues. Finally, I assess the possibility that ncRNAs regulate nearby coding genes through bidirectional transcription.

4.2 Construction of comparison datasets

I constructed the two other datasets as controls for characteristic comparison as follows. (1) mRNA dataset: the pipeline for identifying ncRNAs included a step that determined the match (degree of overlap) between GIs and predicted genes (see Figure 1). We used the 2,311 GIs that had more than 90% overlap with Fgenesh predicted genes to construct the mRNA dataset. (2) ntDNA dataset: although it can never be said with certainty that a sequence is not transcribed, for comparison purposes I constructed a dataset of sequences with a low probability of transcription. I selected all genomic regions in the BAC contig sequences of at least 5.5 kb in which no GIs mapped and no gene could be predicted. The central 800 bp of those regions were extracted. These sequences were checked for sequence homologies with sequences in known non-coding RNA databases including tRNAs (GtRNAdb (Lowe and Eddy, 1997)), rRNAs (the European ribosomal RNA database (Lowe and Eddy, 1997)), snoRNAs (Plant snoRNA database (Brown et al., 2003)), miRNAs (miRBase (Griffiths-Jones et al., 2006)), and other ncRNAs (NONCODE (Liu, 2005)), as well as the structural similarities to the Rfam known ncRNAs (Griffiths-Jones et al., 2005). The 606 *M. truncatula* and 819 *L. japonicus* regions form the ntDNA dataset, respectively.

In the sequence and structure feature comparison analysis, two subsets of the
full mRNA set matched to the ncRNA set were generated in order to minimize confounding by ORF length and sequence length differences in the comparisons. For both legumes, the first subset was the mRNA transcripts with a longest ORF length \( \leq 100 \) codons. For \textit{M. truncatula}, the second subset was a subset of size 1007 which was approximately matched to have the same EST sequence length distribution as the ncRNA set. This was achieved by generating a histogram with bins of width 100 nt of the ncRNA set length distribution and then random sampling the corresponding bins in the mRNA set, in the ratio of approximately 2:1. For \textit{L. japonicus}, the full mRNA set was taken as the EST length matched set as it has the similar EST sequence length distribution as ncRNA set.

4.3 Univariate analysis of base composition biases

4.3.1 Strategy

RNA intrinsic structural constraints may affect di-nucleotide base compositions \cite{29} and cause deviation from approximately \( A\% = T\% \) and \( C\% = G\% \). Previous studies have suggested that the strand asymmetry features \( (A-T)\% \) and \( (G-C)\% \), and the base composition features \( (G+C)\% \) and \( \rho_{CG} \) may serve as indicators of ncRNAs \cite{Klein et al., 2002; Schattner, 2002} although these studies have compared ncRNAs only to genomic averages. I therefore investigated these base composition biases among the ncRNA, mRNA, and ntDNA sets.

4.3.2 Methods

The ncRNA set was compared against a subset of the mRNA set matched for EST sequence length. The discriminability of the features was analyzed with re-
ceiver operating characteristic (ROC) curves. ROC curves plot true positive rates (sensitivity) vs false positive rates (1 - specificity) over the full range of possible classification thresholds so they have the advantage of showing the performance over the full range of classification costs (see Figure 4.6 for examples).

The area under the ROC curve (AUC) is a measure of the discriminability of the classes using the given features and classifier, and varies from 0.5 for non-distinguishable classes to 1.0 for perfectly separable classes: a value of 0.9–1.0 indicates excellent, 0.8–0.9 good, 0.7–0.8 moderate, 0.6–0.7 poor, and 0.5–0.6 no useful discrimination. The AUC can be interpreted as the probability that two random individuals from two classes will be ranked correctly, and is invariant to changes in class proportions (unlike accuracy).

Mono-nucleotide differences and di-nucleotide frequencies were calculated according to the formulae (Burge et al., 1992): \( (G+C)\% = 100 \times (f_G + f_C) \), \( (G-C)\% = 100 \times \frac{(f_G - f_C)}{(f_G + f_C)} \), \( (A-T)\% = 100 \times \frac{(f_A - f_T)}{(f_A + f_T)} \), and \( \rho_{XY} = \frac{f_{XY}}{(f_X \times f_Y)} \), where \( f_X \), \( f_{XY} \), and \( f_{XYZ} \) represented mono-, di-, and tri-nucleotide frequencies. The calculation of oligo-nucleotide frequencies was based on formulae presented above. They were applied to the transcribed strands of ncRNA and mRNA sequences. Given that no strand can be assumed for ntDNA sequences, the formulae were applied to a given sequence strand of ntRNA sequences. Shannon entropy measure (Shannon, 1948) was used to calculate base compositional entropy (H), \( H = - \sum_{i=1}^{n} p_i \log_2 p_i \) where \( n = 4 \).

### 4.3.3 Results and Discussion

I observed several substantial base composition biases consistent with previously reported results (Klein et al., 2002; Schattner, 2002) (Table 4.1 – 4.2 and Figure 4.1 – 4.2).
4.3. UNIVARIATE ANALYSIS OF BASE COMPOSITION BIASES

Table 4.1: Base composition bias among the ncRNA, mRNA, and ntDNA datasets for *M. truncatula*

<table>
<thead>
<tr>
<th>Base composition biases</th>
<th>Mean (standard deviation)</th>
<th>AUC (standard error)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ncRNA set</td>
<td>mRNA set</td>
</tr>
<tr>
<td>(A-T)%</td>
<td>7.23 (12.27)</td>
<td>1.57 (10.05)</td>
</tr>
<tr>
<td>(G+C)%</td>
<td>1.72 (19.46)</td>
<td>5.00 (16.19)</td>
</tr>
<tr>
<td>(G+T)%</td>
<td>31.78 (3.94)</td>
<td>41.25 (3.53)</td>
</tr>
<tr>
<td>ρ_{CG}</td>
<td>0.49 (0.31)</td>
<td>0.51 (0.24)</td>
</tr>
<tr>
<td>ρ_{TA}</td>
<td>0.79 (0.14)</td>
<td>0.60 (0.11)</td>
</tr>
<tr>
<td>Entropy</td>
<td>1.88 (0.05)</td>
<td>1.96 (0.04)</td>
</tr>
</tbody>
</table>

*The base compositions were calculated on the transcribed strands for the sequences in the ncRNA and mRNA sets, and on the ncRNA/mRNA curve.
*Standard error was calculated by Hanley-McNeil estimates.
*p-value<0.05, **p-value<0.005, ***p-value<0.0005; p-value of AUC were calculated by Mann-Whitney-Wilcoxon test.

Table 4.2: Base composition bias among the ncRNA, mRNA, and ntDNA datasets for *L. japonicus*

<table>
<thead>
<tr>
<th>Oligonucleotide base compositions biases</th>
<th>Mean (standard deviation)</th>
<th>AUC (standard error)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ncRNA set</td>
<td>mRNA set</td>
</tr>
<tr>
<td>(A-T)%</td>
<td>-9.87 (14.38)</td>
<td>0.21 (11.67)</td>
</tr>
<tr>
<td>(G+C)%</td>
<td>6.14 (17.81)</td>
<td>0.71 (19.78)</td>
</tr>
<tr>
<td>(G+T)%</td>
<td>36.51 (4.77)</td>
<td>46.60 (5.76)</td>
</tr>
<tr>
<td>ρ_{CG}</td>
<td>0.42 (0.27)</td>
<td>0.55 (0.32)</td>
</tr>
<tr>
<td>ρ_{TA}</td>
<td>0.76 (0.17)</td>
<td>0.54 (0.16)</td>
</tr>
<tr>
<td>Entropy</td>
<td>1.91 (0.05)</td>
<td>1.97 (0.03)</td>
</tr>
</tbody>
</table>

*The base compositions were calculated on the transcribed strands for the sequences in the ncRNA and mRNA sets, and on the random ncRNA/mRNA curve.
*Standard error was calculated by Hanley-McNeil estimates.
*p-value<0.05, **p-value<0.005, ***p-value<0.0005; p-value of AUC were calculated by Mann-Whitney-Wilcoxon test.
Figure 4.1: Density plots of base compositions for *M. truncatula*
Mean values are indicated as dashed lines
4.3. UNIVARIATE ANALYSIS OF BASE COMPOSITION BIASES

Figure 4.2: Density plots of base compositions for *L. japonicus*
Mean values are indicated as dashed lines
(A-T)% shows a large negative mean value of -7.23 in the *M. truncatula* ncRNA set and -9.87 in *L. japonicus* ncRNA set, compared with small mean (A-T)% values observed in the ntDNA and mRNA sets, setting them apart from the ncRNAs. These biases may have been caused by selection for RNA G-U pairings in the transcribed strands. By itself, this feature gives an AUC of approximately 0.7 against both mRNAs and ntDNAs for both legumes. A-T and G-C biases are only expressed in transcribed sequences so it is possible that the relatively high proportion of G and T nucleotides in transcribed sequences reflects the usage of these nucleotides in RNA structures. Similarly, a higher mean (G-C)% value of 6.14 was observed in the ncRNA set than in the mRNA set (0.71) in *L. japonicus*. Yet, the mean (G-C)% value in *M. truncatula* showed a contrasted result from that in *L. japonicus* with 1.72 for the ncRNA set and 5.00 for the mRNA set. It is unclear what caused the difference of (G-C)% in these two closely related legumes. (G-C)%, however, showed no substantial differences among the three classes, with a maximum AUC of only approximately 0.65.

The average (G+C)% value of the mRNA set (41.25% in *M. truncatula* and 46.6% in *L. japonicus*) was greatly elevated whereas those of the ncRNAs and non-transcribed DNAs were close to the whole genome value of 33.3%. This feature gave an AUC ≥ 0.93 for discriminating mRNAs from both the ncRNA and ntDNA sets.

Also, the frequency of the CG dinucleotide, ρCG, is lower than expected in all three datasets. Under-representation of CG is usually attributed to cytidine methylation of this di-nucleotide which increases the mutation rate of CG/GC to TG/CA or CA/GT (reviewed in (Karlin et al., 1998)). I found that CG suppression is most noticeable in transcribed sequences (both ncRNAs and mRNAs), which are somewhat lower than in the non-transcribed sequences (ntDNAs) in both species. By itself, it gives an AUC of 0.7 in distinguishing the *L. japoni-
4.4. UNIVARIATE ANALYSIS OF RNA SECONDARY STRUCTURES

cus ncRNA set from the ntDNA set, but has a lower AUC of approx. 0.68 in discriminating ntDNAs from mRNAs and ncRNAs in *M. truncatula*.

I also found that the dinucleotide TA is significantly under-represented in all three datasets, but particularly in the mRNA set. Three possible explanations have been offered to explain the suppression of TA in coding sequences: 1) the avoidance of cleavage of UA by ribonucleases, 2) the low usage of TAY codons tyrosine (*Y* = C or T), and 3) a paucity of stop codons (UAA and UAG) (Beutler et al., 1989; Burge et al., 1992). The TA feature gives an AUC of approximately ≥ 0.84 when comparing mRNA against both ncRNAs and ntDNAs in both legumes.

Base compositional entropy also showed excellent discrimination of mRNA from others (AUC ≥ 0.88)

4.4 Univariate analysis of RNA secondary structures

4.4.1 Strategy

Current opinion differs as to whether ncRNAs can be recognized by their secondary structures. The algorithms have been developed to use thermodynamic considerations to compute a secondary structure with minimum free energy (MFEs) for an RNA sequence (Mathews et al., 1999). Although examples have been shown where the MFEs of ncRNAs have not differed from the random sequences (Rivas and Eddy, 2000), MFE has been successfully used to recognize some ncRNAs (Washietl et al., 2005). In particular, micro-RNA precursors have lower MFEs than is expected by chance (Bonnet et al., 2004b). It has also been suggested that the number of alternative theoretical structures for a given sequence region is indicative for secondary structure formation *in vivo* (Steele
et al., 2006). It can be assumed that the function of ncRNAs invokes structural features. However, this may not be enough to distinguish ncRNAs from mRNAs as RNA structures are also important for mRNA function. In contrast, there is no apparent reason for ntDNA sequences to evolve stable RNA structures, although simple sequence motifs found predominantly in ntDNAs may increase the number and stability of theoretical structures formed by these sequences. I therefore examined whether the ncRNA sequences reported here differ from mRNA and ntDNA sequences in their propensities to form RNA structures by comparing both global and local structure formations among the three datasets.

It is important to note the structure differences in the two complementary strands because RNA secondary structures typically include G-U base pairs, whereas the corresponding C-A nucleotides of the complementary strand do not pair. As a consequence, the secondary structures formed depend on the strand transcribed. Thus, sequences that have evolved functional RNA structures should have done so predominantly on the transcribed strand. The difference of potential secondary structures in the two complementary strands may be used as a measure of RNA structure evolution. This measure is not affected by simple sequences motifs or other sequence peculiarities that would otherwise equally affect both strands.

I examined whether the ncRNA sequences reported here differ from mRNA and ntDNA sequences in their propensities to form RNA structures by comparing both global and local structure formations among the three datasets. In this section I evaluate these features individually—in the next section I examine them in multivariable combinations.
4.4. UNIVARIATE ANALYSIS OF RNA SECONDARY STRUCTURES

4.4.2 Methods

The RNA global secondary structures

To compare the most stable global structures that could be formed by the ncRNA, mRNA, and ntDNA sequences, I used MFOLD (version 3.2) (Mathews et al., 1999) to compute RNA global secondary structure formations and MFEs for both sequence strands of the GIs in these datasets. MFE values were normalized to the length of sequences yielding the MFE densities (MFED). MFED differences (∆MFED) between the transcribed strand and the non-transcribed strand were calculated.

The RNA local secondary structures

The number of inverted repeats able to form stem loop structures was used as the measure of potential alternative local structures. The mlocalS program (http://hto-13.usc.edu/software/seqaln/index.html) which implements the Smith-Waterman local alignment algorithm was used in order to find inverted repeats. Every sequence was compared with its own reverse copy and a scoring matrix was defined for scoring RNA base pairings. To explore the scoring parameters, I tested different scoring matrices: (a) mismatch 4, gap opening (go) 5, and gap extension (ge) 4, (b) mismatch 4, go 6, and gp 5, (c) mismatch 4, go 6, and ge 4, (d) mismatch 5, go 4, and ge 4, and for all GC = 3, AU = 2, GU = 1. The alignment scores obtained with these parameters were then compared with the MFEs obtained from MFOLD using a test set of short sequences with imperfect inverted repeats. The parameter yielding the highest correlations (correlation = 0.73) between inverted repeat scores and MFEs among four different scoring matrixes were chosen as the inverted repeat scoring matrix. The scatter matrix plot is given in Figure 4.3. Each sequence was aligned against its own reverse
copy using the scoring matrix of GC = 3, AU = 2 as well as GU = 1, mismatch = 4, a gap opening penalty of 5 and a gap extension penalty of 4. The score is calculated by summing the values of each match, the penalties of each mismatch and the large penalties of any gaps. Inverted repeats were then calculated on both strands with a cutoff score of $\geq 13$ and a loop length of 3–100 bp. The total score for each GI sequence was then normalized by the sequence length to give the inverted repeat density (IRD). IRD differences ($\Delta$IRD) between the transcribed strand and the non-transcribed strand were calculated. The number of G-T pairings in the folded structures was counted for both strands. For each sequence, G-T differences between the complementary strands were then calculated and normalized by sequence length (GTD). G and T nucleotide density differences ($\Delta$GTD) between the transcribed strand and the non-transcribed strand were computed.

**Implementation**

I developed several Bioperl scripts that run on the Linux Cluster in order to calculate the MFE and the inverted repeats on both sequence strands of the ncRNA, mRNA, and ntDNA sets, and stored results in a MySQL database table. SQL database queries were subsequently used to calculate the MFED and IRD strand differences on the sequence sets described above.

**4.4.3 Results and Discussion**

Table 4.3 – 4.4 gives the differences between transcribed and non-transcribed strands for both global ($\Delta$MFED) and local ($\Delta$IRD, $\Delta$GTD) structures (see Methods). As no direction of transcription can be assumed for ntDNAs, a given sequence strand was selected as the transcribed strand. The results, however,
Figure 4.3: The plot for inverted repeat scores when compared to MFEs
y axis represents inverted repeat scores with a mismatch of 4, a gap opening of
5, and a gap extension of 5. x axis is the minimum free energies calculated on
the same sequences as inverted repeat calculations.
Table 4.3: Comparisons of RNA secondary structures between the ncRNA, mRNA, and ntDNA datasets for *M. truncatula*

<table>
<thead>
<tr>
<th>Global secondary structures (kcal/mol-bp)</th>
<th>Mean (standard deviation)</th>
<th>AUC(\delta) (standard error(\delta))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ncRNA set</td>
<td>mRNA set</td>
</tr>
<tr>
<td>ΔMFED(\ast)</td>
<td>-0.01 (0.04)</td>
<td>-0.01 (0.04)</td>
</tr>
<tr>
<td>∆IRD(\ast)</td>
<td>0.25 (0.66)</td>
<td>0.12 (0.56)</td>
</tr>
<tr>
<td>ΔGTD(\ast)</td>
<td>0.09 (0.23)</td>
<td>0.07 (0.21)</td>
</tr>
</tbody>
</table>

\*Minimum free energy density (MFED) differences between the transcribed strand and the non-transcribed strand
\*Inverted repeat density (IRD) differences between the transcribed strand and the non-transcribed strand
\*G and T nucleotide density (GTD) differences (in the folded local structures) between the transcribed strand and the random strand was selected as the transcribed strand.

\(\科学技术\) Area under Receiver Operating Characteristic Curves (ROC) curve
\*Standard error was calculated by Hanley-McNeil estimates
\*p-value<0.05, **p-value<0.005, ***p-value<0.0005; p-value of AUC were calculated by Mann-Whitney-Wilcoxon test

showed no substantial differences between the three classes when these features were used individually, with a maximum AUC of only approximately 0.69 in distinguishing transcribed from non-transcribed sequences in both legumes. The large variances of the features lead to poor discriminability even though there are, in fact, some differences between the mean values which do reach statistical significance. The feature distributions (Figure 4.4) also show the three classes are largely overlapped based upon these features individually.

### 4.5 Multivariate analysis of RNA structure strand asymmetry and base composition bias features using SVM

#### 4.5.1 Strategy

In the previous sections, I have explored and compared features based on base composition biases and RNA structure strand asymmetry among the three
Figure 4.4: Density plots of both global ($\Delta$MFED) and local ($\Delta$IRD, $\Delta$GTD) for *M. truncatula* (left column) and *L. japonicus* (right column).
Table 4.4: Comparisons of RNA secondary structures between the ncRNA, mRNA, and ntDNA datasets for *L. japonicus*

<table>
<thead>
<tr>
<th></th>
<th>ncRNA set</th>
<th>mRNA set</th>
<th>ntDNA set</th>
<th>ncRNA/mRNA</th>
<th>ncRNA/ntDNA</th>
<th>mRNA/ntDNA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Global secondary structures (kcal/mol-bp)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔMFED*</td>
<td>-0.022 (0.04)</td>
<td>-0.002 (0.05)</td>
<td>-0.002 (0.04)</td>
<td>0.59 (0.010)**</td>
<td>0.67 (0.011)**</td>
<td>0.52 (0.012)**</td>
</tr>
<tr>
<td><strong>Local secondary structures (scores/bp)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔIRD*</td>
<td>0.421 (0.69)</td>
<td>0.014 (0.77)</td>
<td>0.021 (0.57)</td>
<td>0.64 (0.009)**</td>
<td>0.69 (0.011)**</td>
<td>0.51 (0.012)</td>
</tr>
<tr>
<td>ΔGTD*</td>
<td>0.152 (0.25)</td>
<td>0.002 (0.28)</td>
<td>-0.011 (0.19)</td>
<td>0.65 (0.010)**</td>
<td>0.69 (0.011)**</td>
<td>0.51 (0.012)</td>
</tr>
</tbody>
</table>

*Minimum free energy density (MFED) differences between the transcribed strand and the non-transcribed strand

*Inverted repeat density (IRD) differences between the transcribed strand and the non-transcribed strand

*G and T nucleotide density (GTD) differences (in the folded local structures) between the transcribed strand and the non-

*Standard error was calculated by Hanley-McNeil estimates

*p-value<0.05, **p-value<0.005, ***p-value<0.0005; p-value of AUC were calculated by Mann-Whitney-Wilcoxon test

classes—ncRNAs, mRNAs, and ntDNAs. Such analysis was conducted by univariate feature comparison among these three classes described above. Given that univariate feature comparison may not be powerful enough to distinguish among all classes, I investigated the results further by combining these features. To make such an assessment, I employed a support vector machine (SVM) classifier (Chang and Lin, 2001) to test whether the combination of these features can distinguish between classes. SVM possesses the capacity to map each multi-variable input onto a high dimensional and non-linear feature space by a kernel function that generates an optimal hyperplane to separate classes.

### 4.5.2 Methods

To perform SVM classification, I employed nine features that were categorized into base composition biases, base composition asymmetry, and RNA structure stand asymmetry. The features in base composition biases include the dinucleotide frequency of ρCG and ρTA, (G+C)%, and base compositional entropy. The features in base composition asymmetry include mono-nucleotide differences of (A-T)% and (G-C)%. The features in RNA structure stand asymmetry are
4.5. MULTIVARIATE FEATURE ANALYSIS USING SVM

comprised of both global (ΔMFED) and local (ΔIRD and ΔGTD) structure differentials between two complementary strands. Note that ORF length has been used to select the ncRNAs and mRNA classes in the computational pipeline, and so it is not used as a feature in this later classification stage to avoid selection biases: a feature used to select the classes will, of course, show some differences between the classes. Also it is possible that some of the other features do have correlations with the features used by the gene prediction algorithms, and so some remaining selection bias cannot be completely ruled out. The feature calculations were applied to ncRNA, mRNA, and ntDNA sequences (for ntDNAs the given sequence strand was selected as the transcribed strand). Those features were then standardized to z-scores \( z = (m - \mu)/s \) prior to SVM input, where \( \mu \) is the mean and \( s \) is the standard deviation. To avoid inflated accuracies due to imbalance of class sizes in the dataset, the sizes of the sets were made approximately equal by randomly removing samples from the larger mRNA and ntDNA sets.

The SVM classifier was trained on a randomly selected 2/3 of the data and evaluated on the remaining 1/3 as an independent test set. This process was repeated 50 times and the results averaged. To evaluate the discriminability of the classes, the ROC curves were calculated as well as the associated area under the ROC curve (AUC). The ROC curves were averaged over the 50 repetitions. Sensitivity, specificity, and accuracy are also used to evaluate classifications. Sensitivity is a statistical measure of how well a binary classification test correctly identifies a condition,

\[
Sensitivity = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}
\]

The specificity is a statistical measure of how well a binary classification test correctly identifies the negative cases,

\[
Specificity = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}
\]
CHAPTER 4. CHARACTERISTIC COMPARISONS

Accuracy is also used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition,

\[
\text{Accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}
\]

A radial basis kernel was used with the SVM and a one-against-one strategy used for multi-class classification. To optimise the parameters of the SVM which are: C, the regularization cost, and \( \gamma \), the kernel width, a grid search over the parameters was done with 10-fold cross validation used to choose the best parameter settings. Note that the parameter optimisation was repeated on each training set in the 50 repetitions. The parameter ranges searched were: \( C = 1 \) to \( 100 \), \( \gamma = 0.0025 \) to \( 0.015 \).

4.5.3 Results and discussion

To visualize the combinations of features, we first embedded the full set of nine features into the best fitting three-dimensional space by (metric) multidimensional scaling (using Euclidean distance). The plots (Figure 4.5) indicates that the three classes separated into three distinct regions.

I next used SVM classification with a repeated holdout of an independent test set (1/3 of the data) to obtain the accuracy rate, ROC curves and AUC for classification (See Methods). The ncRNA set was compared with the subset of the mRNA set matched for sequence length, and with the ntDNA set (see Section 4.2). The results are shown in Table 4.5 – 4.6 and Figure 4.6 – 4.7.

Base composition features

The base composition features dinucleotide frequencies TA and CG, (G+C)%, and base compositional entropy were combined and, as expected from the univariate analysis above, the discrimination between mRNA and others was excel-
Figure 4.5: Plots of a three-dimensional feature clouds for *M. truncatula* and *L. japonicus*

All features of three classes (ncRNA, mRNA, and ntDNA) were projected into a three-dimensional space by multi-dimensional scaling.
Table 4.5: Classification based on base composition biases and RNA structure stand asymmetry using SVM for *M. truncatula*

<table>
<thead>
<tr>
<th>Classes</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC (standard error)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base composition features:</strong> $\rho_{TA} + \rho_{CG} + (G+C)% +$ entropy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRNA/mRNA</td>
<td>94%</td>
<td>93%</td>
<td>94%</td>
<td>0.97 (0.002)</td>
</tr>
<tr>
<td>ncRNA/ntDNA</td>
<td>72%</td>
<td>66%</td>
<td>78%</td>
<td>0.79 (0.004)</td>
</tr>
<tr>
<td>mRNA/ntDNA</td>
<td>96%</td>
<td>96%</td>
<td>96%</td>
<td>0.99 (0.000)</td>
</tr>
<tr>
<td>ncRNA/mRNA/ntDNA</td>
<td>77%</td>
<td>--</td>
<td>--</td>
<td>0.92</td>
</tr>
<tr>
<td><strong>Base composition asymmetry features:</strong> (A-T)% + (G-C)%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRNA/mRNA</td>
<td>66%</td>
<td>63%</td>
<td>70%</td>
<td>0.72 (0.005)</td>
</tr>
<tr>
<td>ncRNA/ntDNA</td>
<td>74%</td>
<td>62%</td>
<td>85%</td>
<td>0.79 (0.004)</td>
</tr>
<tr>
<td>mRNA/ntDNA</td>
<td>66%</td>
<td>54%</td>
<td>79%</td>
<td>0.73 (0.003)</td>
</tr>
<tr>
<td>ncRNA/mRNA/ntDNA</td>
<td>52%</td>
<td>--</td>
<td>--</td>
<td>0.75</td>
</tr>
<tr>
<td><strong>Structure strand asymmetry features:</strong> $\Delta MFED + \Delta IRD + \Delta GTD$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRNA/mRNA</td>
<td>49%</td>
<td>42%</td>
<td>58%</td>
<td>0.50 (0.008)</td>
</tr>
<tr>
<td>ncRNA/ntDNA</td>
<td>67%</td>
<td>55%</td>
<td>81%</td>
<td>0.72 (0.005)</td>
</tr>
<tr>
<td>mRNA/ntDNA</td>
<td>67%</td>
<td>55%</td>
<td>78%</td>
<td>0.72 (0.003)</td>
</tr>
<tr>
<td>ncRNA/mRNA/ntDNA</td>
<td>45%</td>
<td>--</td>
<td>--</td>
<td>0.65</td>
</tr>
<tr>
<td><strong>Combination of base composition asymmetry and structure stand asymmetry</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRNA/mRNA</td>
<td>62%</td>
<td>60%</td>
<td>65%</td>
<td>0.68 (0.006)</td>
</tr>
<tr>
<td>ncRNA/ntDNA</td>
<td>72%</td>
<td>61%</td>
<td>83%</td>
<td>0.80 (0.005)</td>
</tr>
<tr>
<td>mRNA/ntDNA</td>
<td>71%</td>
<td>63%</td>
<td>78%</td>
<td>0.78 (0.003)</td>
</tr>
<tr>
<td>ncRNA/mRNA/ntDNA</td>
<td>52%</td>
<td>--</td>
<td>--</td>
<td>0.75</td>
</tr>
<tr>
<td><strong>All features combination</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRNA/mRNA</td>
<td>94%</td>
<td>92%</td>
<td>95%</td>
<td>0.98 (0.001)</td>
</tr>
<tr>
<td>ncRNA/ntDNA</td>
<td>75%</td>
<td>68%</td>
<td>82%</td>
<td>0.84 (0.004)</td>
</tr>
<tr>
<td>mRNA/ntDNA</td>
<td>97%</td>
<td>97%</td>
<td>96%</td>
<td>0.99 (0.000)</td>
</tr>
<tr>
<td>ncRNA/mRNA/ntDNA</td>
<td>79%</td>
<td>--</td>
<td>--</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Discriminative power measured in AUC are colored in red.

Note that the expected values for random pairwise classification are 50% and AUC 0.5, and for the three class classification at approximately 33% and AUC 0.5

-- not applicable
4.5. MULTIVARIATE FEATURE ANALYSIS USING SVM

Table 4.6: Classification based on base composition biases and RNA structure stand asymmetry using SVM for *L. japonicus*

<table>
<thead>
<tr>
<th>Classes</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC (standard error)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base composition features:</strong> $p_{TA} + p_{CG} + (G+C)% +$ entropy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRNA/mRNA</td>
<td>90%</td>
<td>88%</td>
<td>92%</td>
<td>0.95 (0.001)</td>
</tr>
<tr>
<td>ncRNA/ntDNA</td>
<td>75%</td>
<td>76%</td>
<td>74%</td>
<td>0.82 (0.003)</td>
</tr>
<tr>
<td>mRNA/ntDNA</td>
<td>94%</td>
<td>96%</td>
<td>98%</td>
<td>0.99 (0.000)</td>
</tr>
<tr>
<td>ncRNA/mRNA/ntDNA</td>
<td>77%</td>
<td>--</td>
<td>--</td>
<td>0.92</td>
</tr>
<tr>
<td><strong>Base composition asymmetry features:</strong> $(A-T)% + (G-C)%$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRNA/mRNA</td>
<td>69%</td>
<td>64%</td>
<td>74%</td>
<td>0.74 (0.003)</td>
</tr>
<tr>
<td>ncRNA/ntDNA</td>
<td>71%</td>
<td>68%</td>
<td>75%</td>
<td>0.78 (0.003)</td>
</tr>
<tr>
<td>mRNA/ntDNA</td>
<td>57%</td>
<td>44%</td>
<td>71%</td>
<td>0.60 (0.006)</td>
</tr>
<tr>
<td>ncRNA/mRNA/ntDNA</td>
<td>52%</td>
<td>--</td>
<td>--</td>
<td>0.71</td>
</tr>
<tr>
<td><strong>Structure strand asymmetry features:</strong> $\Delta MFED + \Delta IRD + \Delta GTD$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRNA/mRNA</td>
<td>62%</td>
<td>64%</td>
<td>60%</td>
<td>0.67 (0.004)</td>
</tr>
<tr>
<td>ncRNA/ntDNA</td>
<td>66%</td>
<td>62%</td>
<td>71%</td>
<td>0.73 (0.003)</td>
</tr>
<tr>
<td>mRNA/ntDNA</td>
<td>62%</td>
<td>44%</td>
<td>79%</td>
<td>0.66 (0.003)</td>
</tr>
<tr>
<td>ncRNA/mRNA/ntDNA</td>
<td>48%</td>
<td>--</td>
<td>--</td>
<td>0.69</td>
</tr>
<tr>
<td><strong>Combination of base composition asymmetry and structure strand asymmetry</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRNA/mRNA</td>
<td>71%</td>
<td>64%</td>
<td>78%</td>
<td>0.78 (0.006)</td>
</tr>
<tr>
<td>ncRNA/ntDNA</td>
<td>72%</td>
<td>65%</td>
<td>79%</td>
<td>0.79 (0.005)</td>
</tr>
<tr>
<td>mRNA/ntDNA</td>
<td>65%</td>
<td>50%</td>
<td>80%</td>
<td>0.70 (0.003)</td>
</tr>
<tr>
<td>ncRNA/mRNA/ntDNA</td>
<td>56%</td>
<td>--</td>
<td>--</td>
<td>0.76</td>
</tr>
<tr>
<td><strong>All features combination</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRNA/mRNA</td>
<td>91%</td>
<td>89%</td>
<td>92%</td>
<td>0.96 (0.001)</td>
</tr>
<tr>
<td>ncRNA/ntDNA</td>
<td>78%</td>
<td>76%</td>
<td>80%</td>
<td>0.86 (0.002)</td>
</tr>
<tr>
<td>mRNA/ntDNA</td>
<td>94%</td>
<td>95%</td>
<td>92%</td>
<td>0.98 (0.000)</td>
</tr>
<tr>
<td>ncRNA/mRNA/ntDNA</td>
<td>80%</td>
<td>--</td>
<td>--</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Discriminative power measured in AUC are colored in blue.

Note that the expected values for random pairwise classification are 50\% and AUC 0.5, and for the three class classification at approximately 33\% and AUC 0.5

-- not applicable
lent (AUC $\geq 0.95$ for both legumes), but gave a moderate discrimination between ncRNAs and ntDNAs (AUC 0.79 for *M. truncatula* and AUC 0.77 for *L. japonicus*).

**Asymmetry features**

Although not discriminative individually, when combined as a multivariate feature set the global ($\Delta$MFED) and local ($\Delta$IRD and $\Delta$GTD) structure asymmetry features were complementary. In *M. truncatula*, results showed a moderate discrimination between the transcribed and non-transcribed sequences (ncRNA/ntDNA AUC 0.72; mRNA/ntDNA AUC 0.72). However, it could not distinguish mRNA and ncRNA (AUC 0.50). In *L. japonicus*, the combination of structure asymmetry features also gave a moderate discriminability between the ncRNA and ntDNA sets (AUC 0.73), but the mRNA set showed poor discrimination from the ncRNA and ntDNA sets (AUC approx 0.66).

The base composition asymmetry features (A-T)\% and (G-C)\% showed a moderate discriminative power for three pairwise classifications in both legumes (AUC between 0.72 and 0.79), with the exception that the *Lotus japonicus* mRNA and ntDNA sets are not distinguishable (AUC 0.6). The base composition asymmetries were in fact a relatively stronger feature than the structure asymmetry features, suggesting that the structure asymmetry features examined here may be correlated with the base composition asymmetry features.

When the base composition and structure asymmetry features were combined, both legumes showed increased discriminability power of classification between the three data sets.

In summary, these results show that strand asymmetries do exist between the transcribed (ncRNA and mRNA) and non-transcribed (ntDNA) sequences, to a
lesser extent between ncRNAs and mRNAs, and can help in distinguishing these classes.

**Combined feature results**

For distinguishing mRNA from the ncRNAs and ntDNAs, the base composition features were the strongest and the combined AUC for all features for ncRNA/mRNA (AUC 0.98 for *M. truncatula* and AUC 0.96 for *L. japonicus*) and mRNA/ntDNA classification (AUC 0.99 for *M. truncatula* and AUC 0.98 for *L. japonicus*). The combination of all features also give a good discriminability between ncRNA and ntDNA (AUC 0.84 for *M. truncatula* and AUC 0.86 for *L. japonicus*).

The multi-class AUC (Hand and Till, 2001) were performed similar in both legumes: AUC was ≥ 0.92 using all base composition features, ≥ 0.75 using all asymmetry features, and ≥ 0.93 using all features. The multi-class accuracy rates was ≥ 79% in both legumes using all features; the three pairwise classifications of ncRNA/mRNA, ncRNA/ntDNA, and mRNA/ntDNA gave accuracies of ≥91%, ≥75%, and ≥93%, respectively.

**Matched ORF length mRNA subset**

A separate comparison was also performed of the ncRNA set against a short-ORF subset of the mRNA set with longest ORF length ≤ 100 codons to approximately match the ncRNA ORF length. This showed comparable discriminability results to the above comparison against the lengthmatched mRNA set, providing further evidence that the ncRNA set is substantially different from short-ORF mRNAs. These results therefore provide supportive evidence that the ncRNAs identified here possess different nucleotide and structure features to exert their functions from that of mRNAs and that of non-transcribed sequences.
Figure 4.6: The ROC curves for ncRNA, mRNA, and ntDNA classification for *M. truncatula*

The ROC curves plot the true positive (sensitivity) versus false positive (1-specificity) for four pairwise classifications: ncRNA/mRNA, ncRNA/ntDNA, mRNA/ntDNA, and ncRNA/mRNA with the longest ORF 100 codons or less. Five feature combinations were demonstrated: (i) base composition features $\rho_{TA}, \rho_{CG}, (G+C)\%$, and entropy, (ii) base composition asymmetry features: $(A-T)\%$ and $(G-C)\%$, (iii) structure strand asymmetry features: $\Delta MFED$, $\Delta IRD$, and $\Delta IRD$, (iv) combination of (ii) and (iii), and (v) all features combinations. The ROC curves were averaged over the 50 repetitions, and standard errors are shown in error bars.
Figure 4.7: The ROC curves for ncRNA, mRNA, and ntDNA classification for *M. truncatula*

The ROC curves plot the true positive (sensitivity) versus false positive (1-specificity) for four pairwise classifications: ncRNA/mRNA, ncRNA/ntDNA, mRNA/ntDNA, and ncRNA/mRNA with the longest ORF 100 codons or less. Five feature combinations were demonstrated: (i) base composition features $\rho$TA, $\rho$CG, (G+C)% and entropy, (ii) base composition asymmetry features: (A-T)% and (G-C)%,(iii) structure strand asymmetry features: $\Delta$MFED, $\Delta$IRD, and $\Delta$IRD, (iv) combination of (ii) and (iii), and (v) all features combinations. The ROC curves were averaged over the 50 repetitions, and standard errors are shown in error bars.
4.6  Comparison of transcript expressions

4.6.1  Strategy

Although the predicted ncRNAs resemble mRNAs in that they are transcribed by RNA polymerase II, possibly capped, and polyadenylated, they may have different expression abundance and tissue expression patterns. To make this analysis, I compared features of the ncRNA and mRNA sets based on (i) the number of EST clones corresponding to each GI, and (ii) tissue-expression profiles.

4.6.2  Methods

Information regarding the number of EST clones assembled into each GI together with GI tissue libraries were obtained from TIGR.

I counted the number of EST clones that belong to the ncRNA set and mRNA set in the different tissue libraries separately. The clone numbers were then normalized by the total number of EST clones in each of the corresponding tissues to generate an EST clone fraction comparable between tissues within data sets.

To compare tissue expression between the ncRNA and mRNA sets, ratios of ncRNA and mRNA tissue expression level were calculated by dividing the number of EST clones in the ncRNA set by the number of EST clones in the mRNA set in each of the corresponding tissues. The expected value of the ratios would be the total number of EST clones in the ncRNA set divided by the total number of EST clones in the mRNA set. The ratios for each tissue were then standardized by the expected value of the ratios so that the expected ratio was 1.0.

A Chi-squared test was performed to test whether any overall statistically significant differences in ncRNA/mRNA expression ratios between tissue types existed. A permutation test (1000 iterations) on the ratios of the ncRNA and
mRNA sets on each tissue type was performed to obtain per-tissue p-values (multiple hypothesis correction are not applied).

4.6.3 Results and Discussion

Comparisons of expression abundance

TIGR GIs used in this thesis were processed by clustering and assembling EST clones into tentative consensus sequences (TCs). Thus, the number of EST clones assembled into each GI may indicate the expression abundance for a given GI. I compared the expression abundance of ncRNAs and mRNAs by counting the number of composite EST clones that were assembled into each GI. The results showed that *M. truncatula* ncRNAs have an average of 1.5 ESTs per GI, whereas mRNAs have an average 2.7 ESTs per GI. This indicates that predicted ncRNAs tend to be expressed at a lower level compared to mRNAs. In addition, about 83% of the 673 *M. truncatula* putative ncRNAs are singletons whereas only 55% of 2311 mRNAs are singletons. However, for *L. japonicus*, ncRNAs show only slightly lower EST numbers per GI (1.79) than that in mRNAs (1.83), and about 71% of the 1637 ncRNAs are singletons comparable to 72% of the 2338 miRNAs are singletons. It is not clear why the contrasting results of *M. truncatula* and *L. japonicus*. This trend can be observed in the distributions of EST clones for each GI that are represented in Figure 4.8.

Comparisons of expression profiles

I further estimated the expression levels for each tissue type based on TIGR *M. truncatula* and *L. japonicus* Gene Index EST tissue information (see Methods).

In *M. truncatula*, the ratios of ncRNA to mRNA expression in the various tissues showed marked differences (Figure 4.9C). Most noticeably, in root nod-
Figure 4.8: The distributions of EST isolation frequency for coding and non-coding ncRNAs in M. truncatula and L. japonicus EST isolation frequency distribution. For each GI, underlying assembled ESTs were counted. The x axis gives the number of ESTs per GI. The y axis gives the proportion of GIs for ncRNAs (black bar) and mRNAs (light grey bar).

ules there was a substantially increased expression of the candidate ncRNA set compared with the mRNA set (3.9 fold). Seedlings also showed a large increase (2.4 fold). Leaf, pod, and stem, by contrast, showed a reduced ratio (0.5, 0.5, and 0.6 fold, respectively). A chi-squared test confirmed a statistically significant overall difference in tissue expression (p-value ≤ 2.2e-16), and permutation tests showed that these tissue differences reached statistical significance (p-values ≤ 0.05). These results suggest that ncRNAs play a differing role in the various tissues: in particular, nodule shows a major relative increase in ncRNA expression (Figure 4.9A), suggesting an important role for ncRNA in this tissue. (Figure 4.9B) shows that stem, pod, and shoot have relatively high expression of mRNA compared with other tissues.

In L. japonicus, the ratios of ncRNA to mRNA expression in roots and nodules also showed differences (1.6 fold) but to a lesser extent compared to M. truncatula, and showed an increased expression in seeds and pods (3.2 fold). By
contrast, mRNAs are predominately expressed in young plants. (Figure 4.10A). The differing results of *M. truncatula* and *L. japonicus* is unclear, but it may be related to the fact that *M. truncatula* Gene Index set is more complete than that of *L. japonicus*.

### 4.7 Examination of bidirectional transcription

#### 4.7.1 Strategy

It has been suggested that ncRNAs derived from the introns of primary transcripts may serve to signal the expression of the overlaying gene and thus function in the regulation of gene expression (Mattick, 1994, 2001). Some divergently expressed genes could serve a similar purpose. These divergently expressed genes share the same promoter regions but are expressed in opposite directions termed as bidirectional genes. It has been shown that bidirectional gene pairs are co-regulated (Trinklein et al., 2004; Williams and Bowles, 2004). It is therefore conceivable that expression of ncRNA genes could function to signal the expression of their partner genes and influence the regulation of the other genes. Several studies have previously examined bidirectional transcription of the protein-coding gene pairs in the human genome (Adachi and Lieber, 2002; Takai and Jones, 2003; Trinklein et al., 2004), but none have considered ncRNA genes. I have examined bidirectional promoter usage in ncRNA genes compared to mRNAs in *M. truncatula* and *L. japonicus*. 
Figure 4.9: The graphs show the transcript fraction in libraries associated with each tissue type for the (A) ncRNA and (B) mRNA sets. The ncRNA/mRNA ratios are shown in (C). The ratios were standardized so that the expected ratio is 1.0 (white line in the graph, refer to Methods for calculations of transcript fraction and ncRNA to mRNA ratios). Actual numbers of EST clones are given inside bars. White line represents the standardized expectation ratio. p-values for the ratios of ncRNA to mRNA deviate significantly from 1.0 for each tissue type are shown.
Figure 4.10: The graphs show the transcript fraction in libraries associated with each tissue type for the (A) ncRNA and (B) mRNA sets. The ncRNA/mRNA ratios are shown in (C). The ratios were standardized so that the expected ratio is 1.0 (white line in the graph, refer to Methods for calculations of transcript fraction and ncRNA to mRNA ratios). Actual numbers of EST clones are given inside bars. p-values for the ratios for each tissue type are shown.
4.7.2 Methods

GI transcriptional orientations

I obtained GI transcriptional orientations from TIGR (personal communication with Razvan Sultana) with 25,391 known MtGI (about 69% of total MtGIs) and 27,541 LjGI known orientations (about 97% of total LjGIs). TC transcriptional orientations determined by TIGR are based on the orientations of underlying ESTs (5' ESTs, 3' ESTs, and ETs), the orientations of the homologous proteins, and the polyA/T in the ESTs.

Bidirectional gene pairs and gene distances

For “head-to-head” and “head-to-tail” orientations, I calculated distances between GIs and their nearest predicted neighbour genes for TCs that have at least one underlying 5'EST and for 5’singleton ESTs. “Tail-to-tail” distances were calculated for TCs containing at least one 3’EST and for 3’singleton ESTs. The transcription start site and polyA site of predicted protein coding genes were based on the gene prediction information obtained by Fgenesh. Kernel density estimation was used for the distance distributions.

4.7.3 Results and Discussion

I considered two possible types of gene pairs: 1) a ncRNA paired with either a predicted protein coding genes (ncRNA/mRNA) or another ncRNA (ncRNA/nncRNA), and 2) a mRNA paired with a predicted gene (mRNA/mRNA). For each, I computed the distance between the 5’end of the GIs to the transcription start sites (TSS) of the nearest predicted genes that were located on the opposite strands (“head-to-head”). In both M. truncatula and L. japonicus, the resultant distances for both cases (Figure 4.11) showed a central peak at approximately 3
4.7. EXAMINATION OF BIDIRECTIONAL TRANSCRIPTION

kb, but ncRNA data suggests a bimodal distribution with an additional smaller peak at ca 100 bp. As divergent pairs that possess distances of less than 300 bp are likely to share the same promoter regions (Adachi and Lieber, 2002), I conclude that the smaller peak represents a relative abundance of bidirectional genes, while the central peak represents the average distance of genes transcribed in opposite distances. Here, 10% out of the 123 and 9% out of the 76 divergent pairs in the ncRNA set have less than 300 bp “head-to-head” distances compared to only 1.5% out of 423 pairs and 3% out of 529 pairs in the mRNA set in M. truncatula and L. japonicus, respectively. This finding explains the pronounced differences in the graph given in Figure 4.11. For the distances \( \leq 1 \) kb in M. truncatula, 16.2% of divergent genes involve ncRNAs (22 ncRNA/mRNA and 1 ncRNA/ncRNA) and 12.1% involve mRNAs. In L. japonicus, 22% ncRNAs and 16% mRNAs divergently paired with the predicted proteins. This suggests that bidirectional transcription is more prevalent among ncRNA genes than in mRNA genes even though the direction of transcription is known of only some of the ncRNAs. The distances between genes arranged “head-to-tail” gave a single peak distribution in both the ncRNA and the mRNA sets in both legumes (data not shown). The comparisons of the “tail-to-tail” arrangements do not show the left small peaks in L. japonicus, and were inconclusive for M. truncatula as the 3’end is only known for few GIs (data not shown).

To further understand the role that ncRNAs play in bidirectional transcription I examined the functions of their divergent protein coding gene partners that are separated by head-to-head distances of less than 1k bps. The functions of the protein coding partners of ncRNAs are involved in protein kinase activity, flavonoid biosynthesis, F-box protein, RNA-dependent DNA replication, and defence response.

Previous reports (Trinklein et al., 2004; Williams and Bowles, 2004) suggest
Figure 4.11: Distance distributions of gene pairs in *M. truncatula* and *L. japonicus*

The density of the distance distribution of bidirectional genes is given for (A) mRNAs and (B) ncRNAs. Bidirectional gene distances were calculated between 5’GIs and the TSS of the nearest predicted gene partners.
4.7. EXAMINATION OF BIDIRECTIONAL TRANSCRIPTION

that many bidirectional coding genes are co-expressed or anti-regulated. Do divergently expressed ncRNAs regulate their coding gene partner? Alternatively, might they report the expression of their neighbour coding genes to other genes that need to be coregulated? If bidirectional non-coding and coding pairs are not co-regulated, for what reasons do they maintain bidirectional gene transcription?

My results enable the establishment of biochemical expression experiments for bidirectional transcription to address these issues.

The data available for this analysis are not very numerous as their identification depends upon the direction of transcription and the transcription start sites, and both are unknown for many GIs. However, the relative abundance of bidirectional arrangement is still significant and further supported by the fact that the same observation was made in both *M. truncatula* and *L. japonicus*. When more EST or BAC data are available, I will be able to further assess the insight of gene arrangements with ncRNA involvements.
Chapter 5

RT-PCR primer design for *Medicago truncatula* putative ncRNA genes

5.1 Strategy

Although the microarray technique has been widely used to analyze the expression levels of large numbers of genes simultaneously, it cannot measure low levels of gene expression. Quantitative real time polymerase chain reaction (qRT-PCR) is an ideal method to quantify low abundance RNAs at a particular time, at a particular cell, or tissue type. The qRT-PCR technique has the additional advantage of being cheaper and easier to setup than the microarray technique. The downside of this highly sensitivity technique is that it often produces unexpected products caused by unspecific primers to non-target sites (Wang and Seed, 2003). Thus, designing specific primers at the genomic level is an essential step to produce efficient qRT-PCR amplification.

As ncRNAs are often expressed at a low level, qRT-PCR is therefore a suit-
able method to monitor non-coding RNA gene expression experimentally. In collaboration with Dr. Michael Udvardi at the Samuel Nobel Foundation, the expressions of the predicted ncRNAs will be experimentally validated using qRT-PCR. I therefore computationally designed qRT-PCR primers for the predicted ncRNAs to prepare such expression experiments.

To ensure adequate specificity of PCR amplicons, the primer design method described here is to carefully choose the key parameters including melting temperature, primer length, GC content, 3’ end sequences, primer self-interaction, and cross-interaction. The designed primers for a gene also should not match to any other gene on the genome to reduce cross-reactivity.

5.2 Methods

The important design considerations are essential for specific amplification. (1) Melting temperature is the temperature for dissociating DNA duplex. A low melting temperature fails to dissociate DNA duplex to a single strand, yet a high temperature may result in the latter annealing. (2) Primer length. Primers cannot easily bind to target sites at the annealing temperature if the primer is too long. Yet a short primer adversely affects amplification specificity. (3) GC content. A widely accepted GC percentage of primers is 40% - 60% . (4) GC Clamp. A strong GC base pairing at the 3’ end aids 3’ end stability. (5) Both a self-interaction of primer (self-dimer) and a cross-interaction between forward and reverse primers (cross-dimer) should be avoided as they largely reduce the product yield. (6) Low complexity regions such as simple and mono-nucleotide repeats should be avoided as they cause mispriming. (7) 3’ end stability. The five 3’ bases, as determined by the maximum ΔG, are essential to produce stable
5.2. METHODS

bindings (8) Cross match. The designed primers for a gene should not match to any other genes on the genome. (9) Product length for high throughput analysis.

I first used Dust (ftp://www.ncbi.nlm.nih.gov/pub/tatusov/dust) and Repeat Masker (Smit et al., 2004) to mask the repeat regions of 673 ncRNA sequences. To ensure that amplicons performed with both high specificity and efficiency during the RT-PCR reaction, a set of stringent parameters were applied: (1) melting temperatures (Tm) of 60 ± 2 °C, (2) a primer length of 20 - 24, (3) a GC content of 45% - 55%, and (4) a PCR amplicon length of 60 - 150 bp. These parameters have been used to successfully design RT-PCR primers in Arabidopsis transcription factors (Czechowski et al., 2004). Three additional parameters were also applied including (5) a poly nucleotide of no more than 4, (6) exclusion of the front and end 5 bp, and (7) a GC clamp of at least 1. (8) a maximum of self-complementarity (≤ 8) and 3' end complementarity (≤ 3) between forward and reverse primers and the maximum stability of the five 3' base primings (≥ 9). These criteria were employed during the first round primer design. As these stringent criteria may fail to identify suitable primers to some ncRNA transcripts, a second round of primer search was added. I applied the relaxed criteria to the second round primer design with the primer length changed to 19-25 and the GC content changed to 40% - 60%. Criteria 6) and 7) were excluded in the second round.

To ensure that the primers perform with adequate specificity at genomic level, I searched the pairs of primers that were selected by eprimer3 (Rice et al., 2000) to both M. truncatula BAC and TIGR GI sequences. Both forward and reverse primers that matched (E-value ≤ 1E-3) elsewhere in BACs or GIs were discarded.
5.3 Results and Discussion

The computational pipeline that I developed for designing primers on 673 putative ncRNAs in *Medicago truncatula* is given in Figure 5.1 and described in further detail in the Methods section.

The first step involved masking the repeats given that low complexity regions will cause primer cross-reactivity. A pair of primers on each ncRNA was then picked up by the primer design program eprimer3. A set of stringent criteria were applied to be eprimer3 input parameters (see Methods for the first round criteria). In total, 466 ncRNAs found primers after this step and were subsequently BLAST searched against *M. truncatula* BACs. Primers pairs that matched elsewhere on the BACs were eliminated. As the *M. truncatula* genome is incomplete, to reduce the chance of unspecific amplicon fraction, I also discarded the primer pairs that are matched elsewhere on the TIGR Gene Index (GI). In total, 416 GIs encoding ncRNAs found primers through this procedure. Those ncRNA sequences that did not find primers were processed through the second round by applying relatively relaxed criteria (see Methods for the second round criteria) and the aforementioned BLAST search procedure. Another 159 ncRNAs found primers through this round. In all, 575 ncRNA genes have designed primer pairs.

Of the 288 primers experimentally tested so far only ten failed to amplify the expected regions and eight produced an additional fragment (data not shown).
Figure 5.1: The flowchart of RT-PCR primer design pipeline. See text for details.
Chapter 6

Prediction of miRNAs and their target genes

6.1 Background

MicroRNAs (miRNAs) are a class of small non-coding RNAs of approximately 20~24 bps in length that negatively regulate the expression of other genes. The first miRNAs to be identified were the lin-4 and let-7 families which are involved in C. elegan larval development (Lee et al., 1993; Reinhart et al., 2000). They were originally referred to as the small temporal RNAs (stRNAs) and have so far formed the largest family of small RNAs.

Like mRNAs and mRNA-like ncRNAs, many pri-miRNAs are transcribed by RNA Polymerase II, capped, and polyadenylated (reviewed in (Kim, 2005a)). Although miRNAs were originally believed to be processed only from non-coding RNA transcripts in intergenic regions, in animals, they have also been found in introns (reviewed in (Kim, 2005a)). The same may hold true for plants although most extant plant miRNA research has looked for miRNAs in intergenic regions only.
Mature miRNAs are processed from hairpin shaped miRNA precursors. In plants, these small ∼21 nucleotide miRNAs target other transcripts by forming nearly perfect base pairings rather than the imperfect matches found in animals. Such a formation silence genes through one of three modes: mRNA cleavage, translational repression, or RNA-directed DNA methylation (reviewed in (Floyd and Bowman, 2005; Millar and Waterhouse, 2005)). It has also been suggested that miRNAs play key regulatory roles in plants, including development, response to environmental stimulus, hormone signaling, and plant defence (reviewed in (Zhang et al., 2006)).

The finding of miRNAs in both plants and animals indicates that miRNAs existed before the common ancestor of plants and animals diverged. Figure 6.1 (Bartel, 2004) demonstrates proposed models of miRNA biogenesis in plants and animals: 1) miRNA genes are transcribed by RNA polymerase II (Pol II) to generate pri-miRNAs. 2) In plants, a dicer–like protein, DCL1, makes the first cut to produce pre-miRNAs; however the similar procedure is done through Drosha RNase III endonuclease in animals. 3) and 4) In plants, DCL1 or an as of yet unidentified enzyme makes the second cut to produce miRNA and miRNA* (the sequence regions that pairs with the miRNA in the miRNA precursor) before miRNA leaves the nucleus. In contrast, animal pre-miRNAs are exported from nucleus to cytoplasm by exportin5 before the second cut is done by dicer to produce the miRNA:miRNA* duplex. 5) One strand of this duplex is selected as mature miRNA to be loaded into RISC while another strand is degraded. The choice of strand depends on which 5’ end is less tightly paired. 6) microRNA directs the RISC to down-regulate gene expression by one of three modes described above: mRNA cleavage, translational repression, or RNA-directed DNA methylation. The choice of the first two modes relies on the degree of complementary matches between miRNAs and their targets. That is, mRNA cleavage occurs
6.1. BACKGROUND

Each step is described in detail in the text.

if the miRNA::target has near perfect matches; translational repression occurs otherwise (Bartel, 2004).

Although plant and animal miRNAs show several differences in the current proposed models of miRNA biogenesis, some examples of miRNAs have been studied more closely than others. Because the miRNA mechanism has not yet been fully understood, more detailed studies need to be conducted to reveal the biogenesis and mechanism of miRNAs.

Both experimental and computational approaches have been taken to discover miRNAs. Experimentally, direct cloning approaches have been used to discover miRNAs through isolating and cloning small RNAs from biological samples. The
cloning experiments have identified a number of miRNAs in both animals and plants (Lagos-quintana et al., 2001; Lau et al., 2001; Sunkar and Zhu, 2004; Xie et al., 2005) that expand our understanding of miRNAs. However, although the direct cloning approach has discovered a number of miRNAs, this approach is limited to the detection of highly expressed miRNAs. Bioinformatic miRNA detection approaches have been developed to complement this limitation. Several bioinformatic algorithms have initially been developed to identify miRNAs in animals (Lai et al., 2003; Lim et al., 2003a,b). However, they can not be directly applied to identify plant miRNAs because plant miRNAs tend to have longer and more variable loops than animal miRNA genes. Several plant miRNA detection algorithms have therefore been developed (Adai et al., 2005; Bonnet et al., 2004a; Jones-Rhoades and Bartel, 2004; Wang et al., 2004). Most of these approaches to the study of plants have focused on identifying evolutionarily conserved miRNAs from genomic sequences in Arabidopsis thaliana and Oryza sativa (Bonnet et al., 2004a; Jones-Rhoades and Bartel, 2004; Wang et al., 2004). Yet, given that transcription information is not available for many genomic regions, these approaches have difficulties with large scale validation of miRNA expression. Alternatively, if we were able to predict miRNAs from expressed sequences, the identified miRNA host gene expression and associated information would provide us with additional data for further miRNA functional analysis.

Although miRNA prediction approaches to identify conserved miRNAs across species would reduce the false positive prediction, they would exclude miRNAs that have recently diverged or changed rapidly. Given that prior research has tended to adopt such a focus, it is reasonable to speculate that many less-conserved miRNAs have yet to be discovered. This appears a reasonable assumption as non-coding RNAs, in general, tend to be less conserved than proteins.

Few miRNAs have been reported in legumes and are mainly homologues of
known miRNAs in other plants (Dezulian et al., 2005). The identification of novel legume miRNAs is necessary as legumes participate in symbiotic nitrogen fixation by the formation of root nodules, a process in which miRNAs are likely to be involved, and which can only be studied in legumes. The Integrative Legume Research Centre (CILR) where I was pursuing my Ph.D degree is interested in miRNA roles in legume development, in particular, meristem development. This motivated me to computationally predict and analyze miRNAs and targets in model legumes to assist further miRNA functional analysis in legumes.

6.2 Strategy

A computational approach for detecting miRNAs and their targets was employed in this study to (1) find all possible potential candidates and (2) to screen out the low probability ones based upon the core features of known plant miRNAs.

Two common ways to find potential miRNA candidates are the hairpin-structure orientated and the target orientated. Hairpin-structure orientated miRNA prediction approaches have been widely used (Bonnet et al., 2004a; Jones-Rhoades and Bartel, 2004; Lim et al., 2003b; Wang et al., 2004) and involve the search for evolutionarily conserved hairpin structures between two closely related species. Once found, hairpins are then examined by several purification filters to reduce false positive predictions. Finally, a search for target sites associated with the predicted miRNAs is conducted. The target-orientated miRNA prediction approaches (Adai et al., 2005) involve the use of target transcripts to search for potential miRNA matches. The matched segments are then assessed in their capacity to form hairpin structures.

The present study has followed the target-orientated prediction approach. Subsequent to finding potential matches between miRNA and their targets, sev-
eral purification filters were applied to screen out low complexity sequences as well as to check hairpin structure and target binding formation. This approach has two benefits: (i) any potential targets are identified along with predicted miRNAs; and (ii) it can be implemented in a single organism and so is not a search methodology that is restricted to conserved miRNAs across species as is often the case in the hairpin structure approaches.

6.3 Methods

6.3.1 Source sequences

The target-orientated approach requires two source sequence sets, one for predicting miRNAs and one for target sequence prediction. As miRNAs are predominantly either located in the exons of non-coding genes or in introns (Kim, 2005a), I searched for miRNAs in both types of sequences. The *M. truncatula* and *L. japonicus* putative ncRNA genes previously identified (see Chapter 3) were used for predicting miRNAs in ncRNA exons. Intronic regions were determined by mapping ESTs onto genomic sequences and the genomic sequences intervening the ESTs at splice consensus sites were extracted. In total, I used 673 *M. truncatula* ncRNA genes with an average length of 450 bp, and 1,637 *L. japonicus* ncRNAs with an average length of 412 bp, for predicting miRNAs from ncRNA exons. Further, 19,784 *M. truncatula* introns with an average length of 350 bp and 9,192 *L. japonicus* introns with an average length of 328 bp were retrieved for predicting miRNAs from intronic regions.

TIGR *M. truncatula* and *L. japonicus* Gene Index (GI) sequences (ftp://ftp.tigr.org/pub/data/tgi/) consisting of a total of 36,878 and 28,460 as-
6.3. METHODS

sembled GI transcripts, respectively, were obtained for predicting miRNA target sites.

6.3.2 Prediction criteria derived from the known miRNAs

I analyzed features of 751 known plant miRNAs downloaded from miRBase (Griffiths-Jones, 2004) to define our prediction criteria. The examined features include GC content, sequence entropy, simple sequence repeats, and precursor minimum free energies (MFE). As known miRNAs have extreme values in these features, I included a 90% - 95% confidence interval for each feature distribution mean. I derived four identification criteria: (1) a GC content of 30% - 67%, (2) an entropy of $\geq 1.76$, (3) only allowing simple sequence repeats of 1-4 bp with a maximum copy number of 10, 6, 5, and 4 respectively, with $\geq 80\%$ identity between copies, and (4) a minimum free energy (MFEs) normalized by precursor length of $\leq -0.3$ kcal/mol-bp.

6.3.3 Prediction pipeline of miRNAs and their targets

The prediction of miRNAs and target genes included a number of consecutive steps (Figure 6.2). I initially employed the findMiRNA program (Adai et al., 2005) to search for short sequence segments (18-25 bp) within non-coding transcripts and intronic sequences that showed near perfect matches to the sequences in the TIGR M. truncatula Gene Index (GI). For GIs, transcribed strands only were used for searching where transcriptional orientation was known; both strands were used otherwise. For introns, both sequence strands were used for this search. It can be argued that miRNAs should be located on the transcribed strands only. However, miRNAs that have their own transcription units can also be located on the opposite strands. The strand information has been stored with all pre-
dicted miRNAs and can be used to filter out unwanted candidates at a later stage. This initial search step generated a set of ∼13,000 and ∼500,000 matched miRNA::target pairs found in ncRNA and intronic sequences, respectively. This preliminary set presumably contained a large number of false positive hits. These sequences are subsequently referred to as potential “exonic miRNAs” and “intrinsic miRNAs”. A preliminary set of ∼14,000 exonic and ∼70,000 intronic miRNA::target pairs was also generated for *L. japonicus*. I then applied three filters to reduce the number of false positives from this pool, namely a sequence filter, a miRNA hairpin structure filter, and a target binding filter. The criteria used in these steps were derived both from the analysis of known plant miRNAs (see Section 6.3.2) and from previous research (Bonnet et al., 2004a; Grad et al., 2003; Wang et al., 2004).

The sequence quality filter eliminates sequence repeats, low complexity sequences, vector contaminations, short tandem repeats, and tRNAs. It discarded presumptive sequences that had overlaps with repetitive regions as detected by Dust ([ftp://www.ncbi.nlm.nih.gov/pub/tatusove/dust](ftp://www.ncbi.nlm.nih.gov/pub/tatusove/dust)) and Repeat Masker (Smit et al., 2004). The potential miRNA and target sequences also had to have a GC content of 30%–67% and an entropy of ≥ 1.76 to exclude low complexity sequences. Hairpins that showed similarity to vector contaminations as provided by the UniVec database ([http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html](http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html)) (E-value ≥ 1e-2) were discarded, as were the stem-loops containing tRNAs or tandem repeats. tRNAscan-SE (version 1.21) (Lowe and Eddy, 1997) and Tandem Repeat Finder (version 3.21) (Benson, 1999) were used to predict tRNAs and tandem repeats, respectively. In *M. truncatula* 4,585 exonic and ∼200,000 intronic and in *L. japonicus* 11,333 exonic and 56,553 intronic miRNA and target pairs passed these constraints.

The pri-miRNA hairpin structure filter is based on three rules. First, in accor-
6.3. METHODS

Figure 6.2: Flowchart of the computational pipeline of miRNA and target prediction

The source sequences (yellow boxes), the preliminary scan of miRNA and targets (grey box), three elimination filters and a merging procedure (purple boxes) are shown. Each step is described in detail in the text.
dance with published plant miRNA annotation guidelines (Jones-Rhoades et al., 2006), miRNA::miRNA* duplex matches are restricted to ≤ 7 mismatches, ≤ 3 continuous mismatches, and ≤ 2 gaps in the 25 nucleotides centred on the miRNAs and miRNA*s. Second, the MFEs normalized by miRNA precursor length has to be below 0.3 kcal/mol·bp, which was derived from the aforementioned analysis of the known miRNAs. Finally, the loop length of the miRNA hairpin precursors has to be greater than 15 bp given that the minimum known miRNA precursor length in the RNA registry is about 55 bp.

The target binding filter considers the fact that known plant miRNAs have nearly perfect matches to their targets. The filter removed miRNA::target pairs where their alignments showed gaps, or more than three mismatches, or where the MFEs of the miRNA::target duplex were more than -15 kcal/mol. To avoid that sequence regions homologous to the miRNA precursor regions are mistaken for miRNA targets, I also required that no sequence similarity exists outside the miRNA::target regions. Consequently, I aligned not only the 25 nucleotides centred on the miRNA::target pairs, but also the adjacent left 25, and the adjacent right 25 nucleotide regions. The miRNA::target pairs with a maximum sequence similarity of more than 64% in one of the adjacent regions were discarded. After applying this filter, 863 exonic and 36,726 intronic miRNA::target pairs were retained in M. truncatula and 939 exonic and 7,193 intronic miRNA::target pairs were retained in L. japonicus

As the miRNA length cannot be precisely defined, the prediction pipeline determined three regions: (i) miRNA::target duplex binding regions ranging from 18–25 nucleotides as predicted by the FindMiRNA program, which are further used to show the miRNA::target alignments in the database web-interface (described in Chapter 8), (ii) 25 nucleotides centred on the miRNA::miRNA* duplex used in the miRNA fold-back filter, and (iii) the central 21 nucleotides (as 90% of
known plant miRNAs are 21 nucleotides long) to report the predicted miRNAs, to analyze miRNA conservation, and to cluster miRNAs.

Two additional merging steps were added into the pipeline to further minimize the overprediction that may have been caused by overlapping miRNAs or overlapping precursors.

First, the miRNA merging step handles situations in which two or more predicted miRNAs overlap in the same precursor. The aim of this step is to not only minimize superfluous miRNAs but to also keep most information resulting from the prediction. Instead of simply choosing only one representative as other studies have done (Bonnet et al., 2004a; Wang et al., 2004), overlapping miRNAs that had the same orientation and that were located in the same source sequence were merged if they meet either of two criteria: (1) if more than one predicted miRNA overlapped in both the miRNA and target sequence regions, then only the miRNA with the stronger miRNA::target pairing was kept, (2) if only the miRNA regions overlapped but not their target sites, then miRNAs were merged only if the merged miRNAs could be contained within 21 nucleotides or if their 25 nucleotide regions were identical. This merging procedure resulted in a final set of 850 exonic (445 miRNAs::748 target genes) and 35,809 intronic (16,730 miRNAs::16,525 targets) miRNA::target pairs in *M. truncatula* and a set of 936 exonic (620 miRNAs::887 targets) and 7,157 intronic (4,168 miRNAs::4,533 targets) miRNA::target pairs in *L. japonicus*.

The second step merged overlapping precursors. Given that the alternative precursor structures or overlapping miRNAs extending to outside 25 nucleotide regions may have produced overlapping precursors, I determined unique precursor regions by clustering overlapping precursors irregardless of target sites. I then assigned a unique identifier to each region. This resulted in 115 exonic and 3,216
intrinsic unique precursor regions in *M. truncatula* as well as 256 exonic and 1,052 intrinsic unique precursor regions in *L. japonicus*.

### 6.3.4 Assessment of miRNA prediction

In the previous section, I employed multiple filters to minimize the number of miRNA false positive predictions. Take the prediction of *M. truncatula* miRNAs for example. The sequence quality filter reduced miRNA numbers by about 60% through the elimination of miRNAs and targets that overlap repetitive and low complexity elements. Another 32% of the miRNAs were eliminated by taking into account miRNA precursor structures as well as miRNA::target duplex bindings. A further 3% of the candidate miRNAs were merged to counter overprediction.

To statistically estimate the number of falsely predicted miRNAs, I used the *Medicago* ncRNA set to test the signal-to-noise ratio. I simulated the background noise by randomly shuffling the sequences of the Medicago ncRNA set, keeping nucleotide compositions unchanged. A background miRNA distribution (Figure 6.3) was obtained by a bootstrap sampling of miRNA numbers generated from randomly shuffled sequences. The predicted miRNA frequency from actual sequences was 4.8 times the mean of the background.

This assessment process is similar to several recent publications. There are several additional issues that could be considered in a more extensive estimation of the signal-to-noise ratio in future studies following from this thesis: The signal-to-noise ratio estimated here may be affected by any internal homology unrelated to miRNAs, and the extra local structures not due to miRNAs may not be present in the randomly shuffled sequences.

Conversely, as I aimed to avoid over-prediction, I may have missed some true miRNAs. Particularly, the criteria for miRNA::target matches, MFE of precursor
Figure 6.3: Bootstrap frequency distribution for the miRNAs predicted from shuffled sequences
The graph shows the number of miRNAs predicted from real sequences (red arrow) in relation to the background distribution of miRNAs predicted from shuffled sequences (yellow histogram) obtained by bootstrap sampling. Note that the number of predicted miRNAs (445) is much greater than expected by chance.
structures, GC content, entropy, and tandem repeats were stringent, and a small number of experimentally verified miRNA would not have met these criteria. In addition, as the analysis required GIs to map to genomic sequences, miRNAs located in genomic regions not yet sequenced could not be detected.

6.4 Results and Discussion

6.4.1 Data statistics

I developed a computational pipeline for predicting miRNAs from the expressed sequences (i.e. ESTs and introns) (see Methods). The pipeline predicted miRNAs and their corresponding targets by searching for those short and near perfect matches between non-coding EST transcripts or introns and the TIGR Gene Index. The candidates were examined by multiple purification filters to meet miRNA sequence quality, stem-loop structure, and energy requirements (see methods). Applying this pipeline to the model legumes *Medicago truncatula* and *Lotus japonicus* has resulted in the discovery of thousands of potential miRNA candidates and their corresponding targets.

For *M. truncatula*, it generated a set of 445 putative miRNAs encoded in 88 ncRNA transcripts and 748 corresponding target genes in the model legume *M. truncatula*. It further predicted 16,730 intronic miRNAs and their 16,525 target genes. For *L. japonicus*, it generated a set of 620 putative miRNAs encoded in 234 ncRNA transcripts and 887 corresponding target genes as well as 4,168 intronic miRNAs and their 4,533 target genes (Table 6.1). These candidate miRNAs and their targets have three main features: (i) they are free from repetitive and low complexity elements, (ii) they meet the two-dimensional structure and energy requirements described previously, and (iii) they have near perfect miRNA::target
Table 6.1: Summary of miRNA and target prediction

<table>
<thead>
<tr>
<th></th>
<th><em>M. truncatula</em></th>
<th><em>M. truncatula</em></th>
<th><em>L. japonicus</em></th>
<th><em>L. japonicus</em></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GIs</td>
<td>Introns</td>
<td>GIs</td>
<td>Introns</td>
</tr>
<tr>
<td>Num of miRNAs</td>
<td>445</td>
<td>16730</td>
<td>620</td>
<td>4168</td>
</tr>
<tr>
<td>Num of target genes</td>
<td>748</td>
<td>16525</td>
<td>887</td>
<td>4533</td>
</tr>
<tr>
<td>Num of unique</td>
<td>115</td>
<td>3216</td>
<td>256</td>
<td>1052</td>
</tr>
<tr>
<td>precursor regions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

duplex binding. Further, the *M. truncatula* candidate miRNAs are contained in 115 exonic and 3,216 intronic unique precursor regions, and the *L. japonicus* candidate miRNAs are contained in 256 exonic and 1,052 intronic unique precursor regions.

Although I reduced the number of false positives using the aforementioned filtering steps, the data can be further refined to choose a subset to garner more confidence. First, the lower MFEs of miRNA::target duplex would provide a more stable hybridization between miRNAs and target sites. Figure 6.4 displays the proportion of candidate miRNA and targets obtained using different miRNA::target duplex MFE cutoffs. For example, when the MFE cutoff is set to be -25 kcal/mol or less, the number of *M. truncatula* putative miRNA::targets is 8643:7308. Further confidence can be gained by selecting candidates targeting more than one target, as the chance of false positives predicting with more than one target gene is slim. Accordingly, 8,877 (52%) *M. truncatula* and 1,838 (38%) *L. japonicus* miRNA candidates target more than one gene and, using this filter, the signal-to-noise ratio is increased (16:1). Additional confidence can be also obtained if the predicted miRNAs are conserved in other species (see Chapter 7) (Bonnet et al., 2004a; Jones-Rhoades and Bartel, 2004). Prediction of conserved miRNAs is likely to be more reliable with a signal-to-noise ratio of 25:1. How-
ever, conservation is possibly too stringent as a filter, given that at first it would exclude the less-conserved miRNAs, and second the sequence data from other related legumes are currently far from complete. Predicted miRNAs and their targets were recorded in the MIRATdb database (see Chapter 8) if they met the criteria detailed in the aforementioned purification procedures. Additional filters including the miRNA::target duplex MFES and the restriction to conserved miRNAs can be applied at the user’s discretion via the public database interface.

6.4.2 Exonic and intronic miRNAs

Exonic miRNAs were detected from mRNA-like ncRNA transcripts represented in the TIGR GI database which assembles primary EST sequences to tentative consensus sequences. The number of composite ESTs for a given consensus (GI) can be used as a measure of gene expression. The 88 *M. truncatula* and 234 *L. japonicus* ncRNA harbouring miRNAs have an average of 1.7 ESTs/GI and 1.92 ESTs/GI, respectively. This finding is consistent with the predicted ncRNAs (see Chapter 4).

It has been suggested that ncRNAs derived from introns play important regulatory roles in gene expression (Mattick, 1994) and many animal miRNAs are located in introns (Millar and Waterhouse, 2005). It is therefore possible that plant miRNAs also reside in introns. I have revealed a number of putative miRNAs encoded in introns, 24% in *M. truncatula* and 19% in *L. japonicus* which were evolutionarily conserved in other plant species. Some of these intronic miRNAs appear a perfect match to homologous miRNAs in more than one species, indicating that they are authentic. Of the predicted 16,730 *M. truncatula* intronic miRNAs, 56% and 44% of the miRNAs are in sense and anti-sense orientation to the host genes, respectively. In *L. japonicus*, 58% of the miRNAs are in sense and
6.4. RESULTS AND DISCUSSION

Figure 6.4: Statistics of using different miRNA::target duplex MFE thresholds obtained in *M. truncatula*

The proportions of candidate miRNAs, targets, and miRNA::target pairs by using different miRNA::target duplex MFE cutoffs are given.
42% are in anti-sense orientation to the host genes. The host gene orientation is
determined by the presence of splice consensus of ESTs mapping to the genome.
It is reasonable to assume that intronic miRNAs in the antisense orientation to
their host genes are processed from their own transcription units (Weber, 2005)
as is the case for exonic miRNAs. It is not clear whether sense miRNA precursors
also have their own transcription unit, or whether they are transcribed together
with their encompassing genes.

6.4.3 Advantages and disadvantages of using expressed
sequences as source sequences

In this study, expressed sequences (i.e. non-protein-coding ESTs and introns)
were used as source sequences to predict miRNAs. Advantages of this approach
are twofold. First, it gives some confidence that the identified miRNAs are ex-
pressed as their host genes are expressed sequence tags, which provides an al-
ternative solution to the difficulties associated with large-scale experimentally
validating expression of miRNAs. Second, the expression information of ESTs
harbouring the predicted miRNAs, such as library expressions, tissue type, and
microarray data, provide a readily available resource in order to assist miRNA
experimental design and functional analysis.

The disadvantage associated with this approach is that miRNA coverage de-
creases because the prediction depends upon available EST sequences. Because
we required both potential miRNAs and targets to be predicted from ESTs and
each analyzed miRNA to have a corresponding target site, miRNAs were not
identified if host ESTs carrying either miRNAs or target sites are not available.
Chapter 7

Characterization of the predicted miRNAs and targets

7.1 Aim

In Chapter 6, I predicted putative miRNAs and target genes for two model legumes *M. truncatula* and *L. japonicus*. In this Chapter, I further characterize the predicted miRNAs and targets by searching for miRNA conservation in other species, clustering highly similar miRNAs, and functionally classifying potential miRNA target genes.

7.2 Results and Discussion

7.2.1 Taxonomic conservation of miRNAs

To assess the conservation of miRNAs across plant species, I searched *L. japonicus* genomic sequences for homologues to *M. truncatula* miRNAs, and *M. truncatula* genomic sequences for homologues to *L. japonicus* miRNAs. *A. thaliana* genomic sequences were used to search for homologues to both species. Using precExact
I determined all sites which are similar to the 21 nucleotide miRNA sequence (≤ 3 mismatches) and have a matching miRNA* at a distance of 15–400 bp. In addition, the homologous miRNAs must reside on the same arm of the precursor loop as shown in the known homologous miRNAs. The matched homologous miRNA sequences were also required to pass the aforementioned sequence quality and miRNA fold-back structure filters. I found 2,647 of the 17,175 *M. truncatula* predicted miRNAs to be conserved; 2,511 in *L. japonicus* and 188 in *A. thaliana* with only 52 conserved in both plants. For the 4,788 predicted *L. japonicus* miRNAs, 303 are conserved; 262 in *M. truncatula* and 50 in *A. thaliana* with only 9 conserved in both plants.

To see whether the putative *M. truncatula* miRNAs are also conserved in other plants, I extended the search to ESTs of both TIGR consensus sequences and sequences in dbEST ([http://www.ncbi.nlm.nih.gov/dbEST](http://www.ncbi.nlm.nih.gov/dbEST)). I retrieved all assembled GI transcripts of 34 plant species from the TIGR Gene Indexes. For ESTs in dbEST, I first blasted the predicted miRNA sequences against NCBI dbEST and retrieved all plant EST sequences that contained a region with at least 85% sequence identity. All these EST sequences were then used to search for homologous miRNAs by repeating the procedure used to locate homologous miRNAs in genomic sequences. For *M. truncatula* miRNAs, I found 2,179 of the predicted miRNAs matching to 65 plant species. Of these, 1,192 miRNAs match to other legumes, 1,597 to dicotyledons, 808 to monocotyledons, 167 to gymnosperms, 52 to green algae, and 15 to mosses, whereby one miRNA could match to many different species. For *L. japonicus* miRNAs, I found 677 of the predicted miRNAs matching to 52 plant species. Of these, 309 miRNAs match to other legumes, 495 to dicotyledons, 227 to monocotyledons, 49 to gymnosperms, 21 to green algae, and 1 to mosses.

In summary, a total of 4,075 (24%) *M. truncatula* and 853 (18%) *L. japoni-
icu" miRNAs were found conserved in plants and are hereafter referred to as “conserved miRNAs”. No matches were found for the remaining 13,100 *M. truncatula* and 3,935 *L. japonicus* miRNAs in the available sequence data and are subsequently referred to as “non-conserved miRNAs” (Table 7.1).

I traced homologues of 24% of the *M. truncatula* and 18% of the *L. japonicus* miRNAs in more than 60 other plant species in angiosperms, gymnosperms, green algae, and mosses. This suggests that some of the miRNAs have an ancient origin and have remained highly conserved throughout evolution. This has also been suggested by other authors (Axtell and Bartel, 2005). I may have missed homologues due to the conservative homology criteria. I found nearly 10 times as many miRNAs conserved in other legumes as in *Arabidopsis*, indicating that many miRNAs are rapidly evolving. Conversely, I found 185 *M. truncatula* and 130 *L. japonicus* miRNAs with homologues in *Arabidopsis* but not in other legume species. This could reflect a differential loss or more rapid sequence divergence in other legume lineages, or more likely, missing legume sequence data.
7.2.2 Clustering miRNAs

All predicted 21 nucleotide miRNAs and 751 known plant miRNAs were clustered by sequence similarity using CLANS (Frickey and Lupas, 2004). The clustering of sequences was based on alignment scores calculated for all sequence pairs. Due to the large number of sequences involved and the high degree of conservation expected between related miRNAs, alignment scores were calculated by scoring the number of k-mer identities for all sequence pairs. Prior to import into CLANS the pairwise scores were converted to “attraction” values by normalizing pairwise scores to the range 0:1. Clustering of the data was performed at attraction values better or equal to 0.7. Using these parameters, I could correctly reconstruct the known miRNA families. Using this clustering approach, 75% and 59% of the *M. truncatula* and the *L. japonicus* candidate miRNAs could be assigned to 3,651 and 973 groups, respectively, each containing two or more sequences (Figure 7.1). miRNA clustering results are summarized in Table 7.1. Multiple sequence alignments for each clan were performed using emma (Rice et al., 2000) and are used to graphically display alignments through the database web interface (see Chapter 8) by Jalview (Clamp et al., 2004).

Four of the predicted miRNAs showed high sequence similarity to one or more known plant miRNAs, that were in turn categorized into miR166, miR165, and miR398 families. It is important to note the possibility that other known miRNAs were not recovered because the source set of non-coding transcripts did not contain precursors that were similar to known miRNA hairpins. As most known plant miRNAs were predicted from intergenic regions, I set out to see whether the non-coding transcripts used as the source sequences in this study have sequence similarity to known miRNA hairpin sequences miRBase, that my prediction may have missed. A sequence similarity search revealed no further hits.
aside from the miR166, miR165 miR398 families. The miRNA analysis will be extended as more genomic sequences are sequenced and more ESTs are classified into non-coding RNA transcripts.

7.2.3 Classifying miRNA target function

I classified the putative miRNA target gene functions using Gene Ontology (GO) (The Gene Ontology Consortium, 2000). Both “biological process” and “molecular function” categories were assigned to miRNA target genes. Consistent with other reports (reviewed in (Zhang et al., 2006)), the results indicate that plant miRNAs are potentially important for a wide variety of functions and many of the target genes are involved in gene regulation or response to environmental change. Gene ontology analysis of target gene functions suggests that in *M. truncatula*, about 18% of target genes are involved in stimulus response activities, 13% in transporter activities, 10% in transcriptional factor activities, and 9% in both developmental and stress response activities. In *L. japonicus*, the abundant functional categories associated with target genes include 12.3% of target genes that are involved in stimulus response activities, 11.6% in regulation of biological process, 10.5% in transporter activities, and 9.5% in transcriptional regulator activities. Table 7.2 and Table 7.3 give the proportions of both the putative target genes and the entire set of *M. truncatula* and *L. japonicus* genes that fall into each of these two categories.

To further assess whether miRNAs preferentially target certain functions, I searched for statistically over-represented GO terms associated with target genes using GeneMerge (version 1.2) (Castillo-David and Hartl, 2003) with a threshold of the corrected p-value $\leq 0.1$. I identified 230 and 54 GO terms in the “biological process” category and 165 and 33 GO terms in the “molecular function” category.
Figure 7.1: Clustering of 21mer putative miRNAs
The graph illustrates the clustering of *M. truncatula* and *L. japonicus* putative miRNAs using CLANS. Colored dots draw groups clustered at attraction values better or equal to 0.7, each containing two or more sequences. Different colors represent the clusters containing different numbers of sequences. For instance, purple represents the clusters containing seven sequences and red represents each cluster containing equal or more than 15 sequences. The window on the top right corner gives the cumulative distribution plot of attraction values better or equal to 0.7.
7.2. RESULTS AND DISCUSSION

Table 7.2: Functional analysis of predicted miRNA target genes versus all genes in *Medicago truncatula*

<table>
<thead>
<tr>
<th>GO accession</th>
<th>GO term</th>
<th>miRNA target genes</th>
<th>All TIGR genes</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0005554</td>
<td>molecular function unknown</td>
<td>5037/13137</td>
<td>38.3%</td>
<td>11067/34362</td>
</tr>
<tr>
<td>GO:0003824</td>
<td>catalytic activity</td>
<td>7134/13137</td>
<td>54.3%</td>
<td>15889/34362</td>
</tr>
<tr>
<td>GO:0005488</td>
<td>binding</td>
<td>5688/13137</td>
<td>43.3%</td>
<td>12501/34362</td>
</tr>
<tr>
<td>GO:0005215</td>
<td>transporter activity</td>
<td>1793/13137</td>
<td>13.6%</td>
<td>3529/34362</td>
</tr>
<tr>
<td>GO:0030528</td>
<td>transcription regulator activity</td>
<td>1539/13137</td>
<td>11.7%</td>
<td>3077/34362</td>
</tr>
<tr>
<td>GO:0045182</td>
<td>translation regulator activity</td>
<td>260/13137</td>
<td>2.0%</td>
<td>491/34362</td>
</tr>
<tr>
<td>GO:0030234</td>
<td>enzyme regulator activity</td>
<td>220/13137</td>
<td>1.7%</td>
<td>451/34362</td>
</tr>
<tr>
<td>GO:0049871</td>
<td>signal transducer activity</td>
<td>265/13137</td>
<td>2.0%</td>
<td>563/34362</td>
</tr>
<tr>
<td>GO:0016209</td>
<td>antioxidant activity</td>
<td>166/13137</td>
<td>1.3%</td>
<td>347/34362</td>
</tr>
<tr>
<td>GO:0005198</td>
<td>structural molecule activity</td>
<td>446/13137</td>
<td>3.4%</td>
<td>1100/34362</td>
</tr>
<tr>
<td>GO:0030188</td>
<td>chaperone regulator activity</td>
<td>4/13137</td>
<td>0.0%</td>
<td>6/34362</td>
</tr>
<tr>
<td>GO:0045735</td>
<td>nutrient reservoir activity</td>
<td>64/13137</td>
<td>0.5%</td>
<td>132/34362</td>
</tr>
<tr>
<td>GO:0003774</td>
<td>motor activity</td>
<td>96/13137</td>
<td>0.7%</td>
<td>208/34362</td>
</tr>
<tr>
<td>GO:0000004</td>
<td>biological process unknown</td>
<td>5760/11647</td>
<td>49.5%</td>
<td>12937/29458</td>
</tr>
<tr>
<td>GO:0007582</td>
<td>physiological process</td>
<td>7919/11647</td>
<td>68.0%</td>
<td>17993/29458</td>
</tr>
<tr>
<td>GO:0009987</td>
<td>cellular process</td>
<td>7382/11647</td>
<td>63.4%</td>
<td>16619/29458</td>
</tr>
<tr>
<td>GO:0050896</td>
<td>response to stimulus</td>
<td>2116/11647</td>
<td>18.2%</td>
<td>4233/29458</td>
</tr>
<tr>
<td>GO:0050789</td>
<td>regulation of biological process</td>
<td>1672/11647</td>
<td>14.4%</td>
<td>3452/29458</td>
</tr>
<tr>
<td>GO:0007257</td>
<td>development</td>
<td>1067/11647</td>
<td>9.2%</td>
<td>2147/29458</td>
</tr>
<tr>
<td>GO:0040007</td>
<td>growth</td>
<td>452/11647</td>
<td>3.9%</td>
<td>820/29458</td>
</tr>
<tr>
<td>GO:0000003</td>
<td>reproduction</td>
<td>190/11647</td>
<td>1.6%</td>
<td>364/29458</td>
</tr>
<tr>
<td>GO:0044419</td>
<td>interaction between organisms</td>
<td>20/11647</td>
<td>0.2%</td>
<td>31/29458</td>
</tr>
<tr>
<td>GO:0043473</td>
<td>pigmentation</td>
<td>6/11647</td>
<td>0.1%</td>
<td>14/29458</td>
</tr>
<tr>
<td>GO:0016032</td>
<td>viral life cycle</td>
<td>32/11647</td>
<td>0.3%</td>
<td>68/29458</td>
</tr>
</tbody>
</table>

associated with *M. truncatula* and *L. japonicus* target genes respectively, which can be accessed in the “target details” table of the database via the web-interface.

The spectrum of the identified putative target gene functions resembles those of previous studies (Bonnet et al., 2004a), reviewed in (Jones-Rhoades et al., 2006; Zhang et al., 2006) and includes auxin-response transcription factors involved in plant signal transduction, HD-Zip transcription factors involved in meristem development, AP2 domain transcription factor involved in floral development, NAC domain transcription factor involved in embryo, floral, and root development, DCL-like protein and argonaute protein involved in miRNA biogenesis, and F-box proteins. Table 7.4 lists the numbers of the putative miRNAs and predicted target genes that are associated with the known miRNA target functions.
### Table 7.3: Functional analysis of predicted miRNA target genes versus all genes in *Lotus japonicus*

<table>
<thead>
<tr>
<th>GO accession</th>
<th>GO term</th>
<th>miRNA target genes</th>
<th>All TIGR genes</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Fraction</td>
<td>Frequency</td>
<td>Fraction</td>
</tr>
<tr>
<td>GO:0005554</td>
<td>molecular function unknown</td>
<td>920/3118</td>
<td>29.5%</td>
<td>4330/16721</td>
</tr>
<tr>
<td>GO:003824</td>
<td>catalytic activity</td>
<td>1608/3118</td>
<td>51.6%</td>
<td>7483/16721</td>
</tr>
<tr>
<td>GO:0030528</td>
<td>transcription regulator activity</td>
<td>297/3118</td>
<td>9.5%</td>
<td>1169/16721</td>
</tr>
<tr>
<td>GO:0005488</td>
<td>binding</td>
<td>1125/3118</td>
<td>36.1%</td>
<td>5295/16721</td>
</tr>
<tr>
<td>GO:0005215</td>
<td>transporter activity</td>
<td>327/3118</td>
<td>10.5%</td>
<td>1381/16721</td>
</tr>
<tr>
<td>GO:0030234</td>
<td>enzyme regulator activity</td>
<td>46/3118</td>
<td>1.5%</td>
<td>179/16721</td>
</tr>
<tr>
<td>GO:0030188</td>
<td>chaperone regulator activity</td>
<td>2/3118</td>
<td>0.1%</td>
<td>4/16721</td>
</tr>
<tr>
<td>GO:004871</td>
<td>signal transducer activity</td>
<td>44/3118</td>
<td>1.4%</td>
<td>195/16721</td>
</tr>
<tr>
<td>GO:0045182</td>
<td>translation regulator activity</td>
<td>50/3118</td>
<td>1.6%</td>
<td>218/16721</td>
</tr>
<tr>
<td>GO:0045735</td>
<td>nutrient reservoir activity</td>
<td>3/3118</td>
<td>0.1%</td>
<td>26/16721</td>
</tr>
<tr>
<td>GO:003374</td>
<td>motor activity</td>
<td>20/3118</td>
<td>0.6%</td>
<td>93/16721</td>
</tr>
<tr>
<td>GO:0016209</td>
<td>antioxidant activity</td>
<td>34/3118</td>
<td>1.1%</td>
<td>149/16721</td>
</tr>
<tr>
<td>GO:0005198</td>
<td>structural molecule activity</td>
<td>99/3118</td>
<td>3.1%</td>
<td>504/16721</td>
</tr>
</tbody>
</table>

### Table 7.4: Known miRNA target function class associated with the putative miRNAs and targets

<table>
<thead>
<tr>
<th>Target Gene Class</th>
<th><em>M. truncatula</em></th>
<th><em>L. japonicus</em></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of miRNAs</td>
<td>Target gene fraction</td>
</tr>
<tr>
<td>Zinc finger</td>
<td>517</td>
<td>251/553</td>
</tr>
<tr>
<td>F-box</td>
<td>345</td>
<td>171/373</td>
</tr>
<tr>
<td>MYB transcription factor</td>
<td>173</td>
<td>81/162</td>
</tr>
<tr>
<td>bHLH transcription factor</td>
<td>148</td>
<td>61/99</td>
</tr>
<tr>
<td>E3 ubiquitin-protein ligase</td>
<td>107</td>
<td>43/75</td>
</tr>
<tr>
<td>NAC domain-containing protein</td>
<td>67</td>
<td>29/81</td>
</tr>
<tr>
<td>Ubiquitin-conjugating enzyme</td>
<td>47</td>
<td>28/55</td>
</tr>
<tr>
<td>Auxin response factor</td>
<td>53</td>
<td>28/46</td>
</tr>
<tr>
<td>HD-Zip transcriptional factor</td>
<td>97</td>
<td>26/49</td>
</tr>
<tr>
<td>PRR-repeat protein</td>
<td>46</td>
<td>24/56</td>
</tr>
<tr>
<td>Sulfate transporter</td>
<td>52</td>
<td>17/33</td>
</tr>
<tr>
<td>Amine oxidase</td>
<td>26</td>
<td>16/27</td>
</tr>
<tr>
<td>APETALA2-like transcriptional factor</td>
<td>31</td>
<td>15/34</td>
</tr>
<tr>
<td>MADS</td>
<td>19</td>
<td>13/24</td>
</tr>
<tr>
<td>Thioredoxin-like protein</td>
<td>24</td>
<td>12/24</td>
</tr>
<tr>
<td>Scarecrow-like</td>
<td>16</td>
<td>9/19</td>
</tr>
<tr>
<td>Argonaute</td>
<td>19</td>
<td>8/16</td>
</tr>
<tr>
<td>Squamosa promoter-binding protein</td>
<td>13</td>
<td>5/17</td>
</tr>
<tr>
<td>HARP</td>
<td>6</td>
<td>4/9</td>
</tr>
<tr>
<td>laccase</td>
<td>12</td>
<td>4/7</td>
</tr>
<tr>
<td>Plantacyanin</td>
<td>4</td>
<td>3/9</td>
</tr>
<tr>
<td>CCAAT-box binding transcription factor</td>
<td>7</td>
<td>3/4</td>
</tr>
<tr>
<td>Dicer-like protein</td>
<td>2</td>
<td>2/7</td>
</tr>
<tr>
<td>TCL transcription factor</td>
<td>3</td>
<td>2/5</td>
</tr>
<tr>
<td>3-phosphoglycerate kinase</td>
<td>3</td>
<td>1/1</td>
</tr>
<tr>
<td>SAMT</td>
<td>1</td>
<td>1/1</td>
</tr>
</tbody>
</table>

* Target gene fraction and percentage are calculated by dividing the number of target genes by total genes in the given functional class.
Chapter 8

MIRATdb – legume putative miRNA and target database

8.1 Aim

In Chapter 6 and Chapter 7, I reported on the predicted thousand of miRNAs and their corresponding targets for two model legumes. I further characterized predicted miRNAs and targets by searching for miRNA conservation across other plant species, clustering highly similar miRNAs, and classifying functions of target genes. To make such information available, to provide easy access to the data, and to support future miRNA research, I constructed the MIRATdb (MiRNA And Target gene Database) for *M. truncatula* and *L. japonicus*. The database provides a user-friendly web interface and serves as an analysis tool that allows users to retrieve and analyze data. It provides experimentalists with many *in silico* identified miRNA and target candidates as well as associated information to initialize biochemical design and analysis of miRNAs. The database is therefore proffered in hopes that it will assist experimentalists in the design of miRNA biochemical experiments.
8.2 Implementation

The implementation of MIRATdb involves three traditional tiers: a frontend interface for both displaying and interactively manipulating the data, a relational database backend, and a middle dynamic web application for accessing databases to generate dynamic content. The relational database backend consists of two databases, Medicago-MIRATdb and Lotus-MIRATdb, for storing the predicted miRNAs and target genes in *M. truncatula* and *L. japonicus*, respectively.

Data were generated by a pipeline of Perl scripts that run a 20 node Linux Cluster and were stored in a MySQL relational database. The database schema is described in greater detail in Appendix A. The web interface, which was implemented in PHP and Perl scripts and runs on an Apache web-server, queries the data from the database and displays results in a graphic form. An example of programming code for data retrieval from the database is given in Chapter 9.

8.3 Database interface

Various queries, tools, and graphical displays were incorporated into the database interface to provide easy access to the data and to support future miRNA research. The interface is divided into four sections: miRNA information, target information, query facility, and analysis tools.

The miRNA page provides a list of putative miRNAs. Info for each miRNA is accessed by clicking the hyperlink on the “details” button, which brings up a “miRNA details” table (Figure 8.1), providing information about each miRNA entry including the predicted miRNA sequence, the alignment of miRNA::miRNA*, the precursor sequence, as well as the EST libraries and the number of ESTs per GI. The table further includes multiple alignments of, and the hyperlinks to,
similar miRNAs, the number of target sites, as well as a hyperlink to the cor-
responding miRNA::target duplex binding energy and alignment. The target page
provides entries of potential miRNA target genes. The “target details” table
(Figure 8.2) for each entry shows the target GI features, the best three Blast hits
against the non-redundant protein databases, GO classification of target genes,
and an alignment hyperlink to their corresponding miRNAs. External links to
the TIGR Gene Index, GenBank, and UniProt databases are also provided.

Users can search for miRNAs or target genes through a search form. Genes
can be searched by miRNA or GI identifiers, or an advanced combination search of
GI tissue category, miRNA conservation to other species, miRNA::target duplex
MFE, and keywords in target gene annotation. The search results page lists gene
entries that are linked to the tables described above. Users can choose to view
results online or download search results as a comma separated csv file or fasta
sequence file. The miRNAs, target sites, and miRNA precursors are available for
download as flat files.

In addition to the data query functions, two on-line analytical tools are avail-
able to allow users to search for miRNAs or target sites on their submitted se-
quen ces. A BLAST tool is integrated into the database to facilitate users finding
miRNAs in their query sequences of interest from within the putative miRNAs
in the database. Users can define varying BLAST parameters including E-values,
word size, query strand, match reward, mismatch penalty, and the number of
alignments to display. The Blast result page is linked to the “miRNA detail
table” for matched miRNAs.

I further developed an online target prediction interface that allows users to
locate potential target sites in the query sequences of interest. The program
RNAhybrid (Krüger and Rehmsmeier, 2006) is incorporated into the database to
predict miRNA target sites matched to putative miRNAs. The interface offers
users the option of miRNA::target binding minimum free energy, p-value of binding significance, and sequence strand to perform searches. The resultant page displays the miRNA::target binding alignments and target positions as well as matched miRNAs that are, in turn, linked to the “miRNA detail table”.

The database will be freely accessible at http://bioinfoserver.rsbs.anu.edu.au/utils/Medicago-MIRATdb/. (Note that the database currently needs password access)
### Detail Info for Putative miRNA n379

<table>
<thead>
<tr>
<th>Putative miRNA (mature form)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>miRNA ID</strong></td>
<td>n379</td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>TCLI06017 (164-194)</td>
</tr>
<tr>
<td><strong>Sequence</strong></td>
<td>cgTCATGTGCTATTGATCGATc</td>
</tr>
<tr>
<td><em><em>miR and miR</em> Alignment</em>* (Centered on 25 nt)</td>
<td>CTAGAGCTAGATGCCTCTAGCTGC</td>
</tr>
<tr>
<td><strong>miRNA Precursor</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Positions (In Source)</strong></td>
<td>165-201</td>
</tr>
<tr>
<td><strong>Precursor DeltaG (normalized by length)</strong></td>
<td>-6.32 (kcal/mol-bp)</td>
</tr>
<tr>
<td><strong>Sequence</strong></td>
<td>CGGTAAGCTTTCTAGATCGATCTGAAATACAAATCCGATTTGAT</td>
</tr>
</tbody>
</table>

### EST info

<table>
<thead>
<tr>
<th>GI Accession</th>
<th>TCLI06017</th>
</tr>
</thead>
<tbody>
<tr>
<td>EST Libraries (TIGR Cast)</td>
<td>T1727 T1841</td>
</tr>
<tr>
<td>Tissue Category</td>
<td>Root</td>
</tr>
<tr>
<td>Number of ESTs</td>
<td>2</td>
</tr>
</tbody>
</table>

### miRNA CLAN

<table>
<thead>
<tr>
<th>CLAN ID</th>
<th>55</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment View</td>
<td>DnarJaekes</td>
</tr>
</tbody>
</table>

**Similar Genes in CLAN 55**

<table>
<thead>
<tr>
<th>Genes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>110773</td>
<td>116774</td>
</tr>
<tr>
<td>110775</td>
<td>116776</td>
</tr>
<tr>
<td>116777</td>
<td>116778</td>
</tr>
<tr>
<td>116779</td>
<td>116780</td>
</tr>
<tr>
<td>116781</td>
<td>116782</td>
</tr>
<tr>
<td>116783</td>
<td>116784</td>
</tr>
<tr>
<td>116785</td>
<td>116786</td>
</tr>
</tbody>
</table>

**Conserved in other Plant Species**

<table>
<thead>
<tr>
<th>In Legumes</th>
<th>Trifolium pratense</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Non-Legume Plants</td>
<td></td>
</tr>
</tbody>
</table>

**miRNA Targets**

<table>
<thead>
<tr>
<th>Target number</th>
<th>Align</th>
</tr>
</thead>
</table>

---

Figure 8.1: Screenshot of the miRNA details table. This screenshot presents the details of putative miRNA n379. The inserted window gives the multiple alignments of all miRNAs in clan 55 (see text).
**Figure 8.2:** Screenshot of the target details table  
This screenshot presents the details of putative target gene TC103458. The alignments of miRNAs (i5106 and i7965) to the this target gene are given (see text).
Chapter 9

Implementations of programming source code

Data collection, generation, and analysis reported throughout the project and described in Chapter 3 to Chapter 5 were primarily implemented through four types of programming codes. First, Perl and Bioperl scripts were used to engage pipelines of non-coding RNA prediction, miRNA and target prediction, and RT-PCR primer design. Second, PHP, Perl, and SQL were used to develop web-based application of MIRATdb. Third, R scripts were used to analyze statistical tests and plots. Finally, Perl scripts distributed the codes to run softwares in parallel on the Linux cluster nodes. To show all codes would exceed the space avoided for this thesis, thus each type of programming and one example of a coinciding code fragment is presented and explains in more detail below.

9.1 Perl scripts

I developed about 150 Perl and Bioperl scripts to conduct data analysis. An example is provided below. The program “findInvertedRepeats.pl” takes in-
put sequences in FASTA-format file and iterates each sequence by the Bioperl “Bio::SeqIO” module to find inverted repeats using the Smith-Waterman local alignment.

I explored a scoring matrix for aligning RNA sequences with parameters of a G-C score of 3, an A-T score of 2, and a G-U score of 1, a mismatch of 4, a gap-opening of 5, and a gap-extension of 4. The matrix was used to facilitate the local alignments for finding inverted repeats in RNA sequences. The output files are parsed through a parser and stored into a MySQL database for further analysis. The code is shaded in grey and a brief description is given in red text starting with comment sign “ #”.

Listing 9.1: Perl Code

```perl
#!/usr/bin/perl

# this program is to find inverted repeats using dynamic programming algorithm (Smith-Waterman). The program take a sequence as input and reverse it and use program "mLocalS" (local alignment in Waterman package) to align the program can find "AT GC GU" pairs
# matrix: G-C 3 AT 2 GU 1 mismatch 4 gap-opening 5 gap-extension 4
# input: sequences or reverse complement of the input seqs
# output: inverted repeats scoring database and frequency database

# import Perl and Bioperl module
use Bio::Seq;
use Bio::SeqIO;
use Bio::SearchIO;
use Getopt::Long;
use Carp;
use DBI;
use strict;

my ($db_invert, $infile, $revcomp, $dbusr, $dbpw) = '';
# Program input options
&GetOptions("-db_invert=s"=>\$db_invert,
           ":-i=s"=>\$infile

# T reverse complement of the input seq, F otherwise
```
9.1. PERL SCRIPTS

```perl
unless ($db_invert && $revcomp && $infile && $dbpw) {
    print "prompt> perl -Mdb_invert -i
    input_file -rc -u username -p password
";
    exit;
}
my $dsn = "DBI:mysql:Jean_Medicago:node0";
my $dbh;

#connect to MySQL db
$dbh = DBI->connect($dsn, $dbusr, $dbpw)
or die "Cannot connect to $dsn: $errstr\n";
&find_inverted_repeats;

sub find_inverted_repeats {
    my $in_seq = Bio::SeqIO->new(-file => $infile,
      -format => "fasta");
    while ( my $seq_obj = $in_seq->next_seq() ) {
        my $id = $seq_obj->id;
        my $seq_len = $seq_obj->length;
        my $fw_str;
        $fw_str = $seq_obj->seq if ($revcomp eq "F");
        $fw_str = $seq_obj->revcomp->seq if ($revcomp eq "T");
        my $rc_str = &revSeq($fw_str);
        open FW, ">fw_tmp" or croak "Error opening fw_tmp\n";
        open RC, ">rc_tmp" or croak "Error opening rc_tmp\n";
        print FW "$id\n$fw_str\n";
        print RC "$id\n$rc_str\n";
        &run_mlocalS("fw_tmp", "rc_tmp");
        &parse_results($seq_len);
    }
}

# run Smith–Waterman local alignments mlocalS
sub run_mlocalS {
    my ($fw_tmp, $rc_tmp) = @_; 
    my $matrix = "/opt/seqaln-2.0/seqaln/matrix/InvertedRepeats";
    system("/opt/seqaln-2.0/seqaln/bin/mlocalS"
        "-rc=s" => $revcomp,
        "-u=s" => $dbusr,
        "-p=s" => $dbpw);
    }
```

sub parse_results {
    my ($seq_len) = @_; 
    open(IN_tmp, "<out_tmp") || die "cannot open out_tmp for reading";
    my ($score, $id1, $id2, $id1_start, $id1_end, $id2_start, $id2_end, $id1_len, $id2_len, $loop_len, $gt_count)=0;
    my ($fw_align, $rc_align) = "";
    my $first = 0;
    my $align_flag = 1;
    while (<IN_tmp>) {
        my $line = ";
        if ($line =~ /Score: \((d+) at \((\(w+)\))\)((d+)..\((d+)..(d+)\))/) {
            $first ++;
            if ( ($fw_align && $rc_align ) { 
                $fw_align = $fw_align . $1;
                $rc_align = $rc_align . $1;
            } elsif ( $align_flag == 2 ) {
                $fw_align = $fw_align . $1;
                $rc_align = $rc_align . $1;
                $align_flag = 1;
            }
            $score = $1;
            ($id1, $id1_start, $id1_end, $id2, $id2_start, $id2_end) = ($2, $3, $4, $5, ($seq_len-$7+1), ($seq_len-$6+1));
        } elsif ($line =~ /\d+\s+([A-Z\-]{}1,})\s+\d+/) {
            if ( $align_flag == 1 ) {
                $fw_align = $fw_align . $1;
                $align_flag = 2;
            } elsif ( $align_flag == 2 ) {
                $rc_align = $rc_align . $1;
                $align_flag = 1;
        }}
if ($fw_align && $rc_align) {
    if ($loop_len > 0) {
        $gt_count = &count_GI($fw_align, $rc_align);
        &insert_into_db($id1, $id1_start, $id1_end, $id2_start, $id2_end, $score, $loop_len, $fw_align, $rc_align, $gt_count);
    }
}
close(IN_tmp);

# Results are stored into MySQL database
sub insert_into_db {
    my ($id1, $id1_start, $id1_end, $id2_start, $id2_end, $score, $loop_len, $fw_align, $rc_align, $gt_count) = @_;
    my $statment = qq(INSERT INTO $db_invert('est_name', 'fw_start', 'fw_end', 'rc_start', 'rc_end', 'score', 'loop_len', 'gt_count', 'fw_align', 'rc_align') VALUES("$id1","$id1_start","$id1_end","$id2_start","$id2_end","$score","$loop_len","$gt_count","$fw_align","$rc_align"););
    my $sth = $dbh->prepare($statment);
    $sth->execute;
    $sth->finish;
}
sub count_GI {
    my ($fw_align, $rc_align) = @_;
    my @fw_align = split(/\/, $fw_align);
    my @rc_align = split(/\/, $rc_align);
    my $gt_num = 0;
    for (my $i = 0; $i <= (@fw_align - 1); $i++) {
        if (( $fw_align[$i] eq "G" && $rc_align[$i] eq "T") || ($fw_align[$i] eq "T" && $rc_align[$i] eq "G")) {
            $gt_num++;
        }
    }
    return $gt_num;
}
sub revSeq {
    my ($seq) = @_;
    my @seq = split(/\/, $seq);
    my $revseq = "";
}
Figure 9.1: A example for querying miRNAs through database web interface

```perl
foreach my $rev_letter (reverse @seq) {
    $revseq = $revseq . $rev_letter;
}
return $revseq;
```

### 9.2 Web-based application

The user, for example, wants to search those miRNAs that are derived from ncRNA transcripts, are conserved in legumes, and are expressed in nodules. The options that the user made are illustrated as black arrows in Figure 9.1 below.

A PHP program shown below takes users options and sends it to a Perl script to retrieve the data from the database through a SQL statement. The code
9.2. WEB-BASED APPLICATION

fragments are shown as an example below.

Listing 9.2: PHP Code

```php
<?php

session_start();
if (!isset($_SESSION['usr']) || !isset($_SESSION['pwd'])){  
    header('Location:../..//index.php');
} else {
    require_once '../includes/defination.inc';
    require_once '../includes/db_lotus.inc';
    require_once C_HEADER;
    require_once C_L_TOP;
    print top();
    ................................

    $offset = 0;
    if ($_POST['What'] === "miRNA") {

        if (isset($_POST['SourceId'])) {
            $sourceId = $_POST['SourceId'];
            system("perl_searchMiRNA.pl $hostname $databasename $username $password $offset 2 "$sourceId");
        }
        if (isset($_POST['Source'])) {
            $source = $_POST['Source'];
        }
        if (isset($_POST['Species'])) {
            $species = $_POST['Species'];
        }
        if (isset($_POST['Clone'])) {
            $clone = $_POST['Clone'];
        }
        if (isset($_POST['Tissue'])) {
            $tissue = $_POST['Tissue'];
        }
        if (($source or $species or $clone or $tissue)) {
            system("perl_searchMiRNA.pl $hostname $databasename $username $password $offset 3 "$source "$species "$tissue "$clone");
        }  
    }
```

require_once C_FOOTER;

} # end session

# SQL statement construction
if ( $species ne "all" && $tissue ne "all" ) {
  # all selected
  my $where;
  if ( $source eq "all" && !$clone ) {
    $where .= &tissueString($tissue).' AND'.&speciesString($species);
  } elsif ( $source eq "all" && $clone ) {
    $where .= &cloneString($clone).' AND'.
      &tissueString($tissue).' AND'.
      &speciesString($species);
  } elsif ( $source ne "all" && !$clone ) {
    $where .= &sourceString($source).' AND'.
      &tissueString($tissue).' AND'.
      &speciesString($species);
  } elsif ( $source ne "all" && $clone ) {
    $where .= &sourceString($source).' AND'.
      &cloneString($clone).' AND'.
      &tissueString($tissue).' AND'.
      &speciesString($species);
  }

  $queryString1 = qq(SELECT COUNT(*) FROM
    'Mir25', 'MirConserv_short', 'GiTissue_miRNA'
    Where 'Mir25'. 'Mir25Id'=' MirConserv_short'. 'Mir25Id'
    AND 'Mir25'. 'GiId'='GiTissue_miRNA'. 'GiId'
    AND $where );

  $queryString2 = qq(SELECT 'Mir25'. 'Mir25Id', 'Mir25'. 'GiId', 'Mir25'. 'IntronId', 'Mir25'
    . 'OrSrc', 'Mir25'. 'StartInSrc', 'Mir25'. 'Mir25Seq',
    'Mir25'. 'cntTgt' FROM 'Mir25', 'MirConserv_short',
    'GiTissue_miRNA'
9.3. STATISTICAL COMPUTING USING R

All data statistics conducted in this thesis are implemented in R scripts using R statistical package (http://www.r-project.org). A R script below illustrates a hybrid bootstrap sampling for the longest ORFs of the coding and non-coding RNAs. The resulted graph is in Figure 3.6.

Listing 9.3: R Code

```r
library(bootstrap)

#The function to perform the bootstrap samplings and plot
#histograms of sampling distribution for the longest ORF
boots. ORFs <-function(x, x_r, titles) {
  xmy <- c(x, x_r)
  xmy.size = length(xmy)
  theta <-function(xmy){
    i <- sample(1:xmy.size, xmy.size/2, replace = TRUE)
    xmy.1 <- mean(i)
    j <- sample(1:xmy.size, xmy.size/2, replace = TRUE)
    xmy.2 <- mean(j)
    xmy.1 - xmy.2
  }
  results <- bootstrap(xmy, 1000, theta)
  booted.xmy <- results$thetastar
  Diff.real <- abs(mean(x) - mean(x_r))
  #calculate p-value
  p <- (length(booted.xmy[abs(booted.xmy) > Diff.real])+1)/(bootstrap.size+1)
  #get critical value
  xmy<-sort(booted.xmy, decreasing = TRUE)
  xmy[50]
  #histogram plots of distribution, and data summary
  h<-hist(booted.xmy, br=24,col=”skyblue”)
}
```

```r
xlim = c(min(booted_xmy) - 10, max((Diff_real + 10),
          max(booted_xmy))), main = titles
xlab = paste("longest ORF length difference p-value =", p)

s <- summary(booted_xmy)
sd <- sd(booted_xmy)
var <- var(booted_xmy)
q <- quantile(booted_xmy, probs = c(2.5, 5, 10, 25, 50, 75, 90, 95, 97.5) / 100)

axis(1)
if (Diff_real < max(booted_xmy)) {
  axis(1, round(Diff_real, 1), mgp = c(3, 1, 0), col = "red", lwd = 2)
  arrows(Diff_real, -10, Diff_real, 1, col = "red", lwd = 2)
} else {
  axis(1, round(Diff_real, 1), mgp = c(3, -2, 0), col = "red", lwd = 2)
  arrows(Diff_real, 10, Diff_real, 3, col = "red", lwd = 2)
}

axis(1, round(xmy[50], 1), mgp = c(3, 1, 0), lwd = 2)

return(p, s, sd, var, q)
```

```
# read data into data frames
ncRNA_ORF <- c(read.table("LJ_ncds_longestORF.txt", header = TRUE, sep = " "))
ncRNA_ORF_r <- c(read.table("LJ_ncds_longestORF_r.txt", header = TRUE, sep = " "))
cdRNA_ORF <- c(read.table("LJ_cds_longestORF.txt", header = TRUE, sep = " "))
cdRNA_ORF_r <- c(read.table("LJ_cds_longestORF_r.txt", header = TRUE, sep = " "))

# extract the column of ORF length
ncRNA_ORFlen <- ncRNA_ORF$orf_len
ncRNA_ORFlen_r <- ncRNA_ORF_r$orf_len
cdRNA_ORFlen <- cdRNA_ORF$orf_len
cdRNA_ORFlen_r <- cdRNA_ORF_r$orf_len

par(mfcol = c(1, 2))

boots_ORFs(ncRNA_ORFlen, ncRNA_ORFlen_r, "ncRNA")
boots.ORFs(cdrNA.ORFlen, cdRNA.ORFlen_r, "cdRNA")
```
9.4 Usage of high performance Linux clusters

9.4.1 High performance clusters

High performance clusters are commonly engaged in running parallel programs for time-intensive calculations. A cluster is typically comprised of one head node and several compute nodes (Figure 9.3). In recent years, Linux high performance clusters, with their relatively low cost, are often used in life science research to handle ever-growing data. I have been involved in the purchase decision of RSBS Linux Cluster, and responsible for setting up the cluster, software installation, and system administration.

9.4.2 Specification of RSBS GIG Linux cluster

The Genomic Interaction Group (GIG) Bioinformatics Cluster is comprised of a Dell PowerEdge 2850 head node and 20 Dell PowerEdge SC1425 compute nodes. All nodes are interconnected via 100Mb Ethernet links to a Dell PowerConnect 5324 switch. All the nodes are running Fedora Linux and lamd daemon for MPI services. The head node configuration is:
• CPU - Intel(R) Xeon(TM) Processor 3.2GHz/1M,EM64T,800MHz

• Main Memory - 2GB (4x512), DDR-2 400MHz

• Disk  RAID 5 array based on 300GB Ultra320 (10K RPM, 80-pin) SCSI drives

The compute nodes configuration is

• CPU - Intel(R) Xeon(TM) Processor 3.2GHz/1M, 800MHz

• Disk - 80GB (7200 RPM) SATA

• Main Memory - 2GB (4x512), DDR-2 400MHz ECC

9.4.3 Uses of the linux cluster in this project

In this project, I used the distributed memory approach of parallel programming. In brief, the head nodes divide and distribute jobs to the compute nodes. The compute nodes perform the tasks on their own memory and return the results to the head nodes. The job distribution is scheduled by the SLURM (Simple Linux Utility for Resource Management, http://www.llnl.gov/linux/slurm/). Figure 9.3 demonstrates a simple model of this approach.

The code shown below divides a large FASTA-formatted sequence file into 20 parts that then sends these parts to run on 20 compute nodes by SLURM scheduler.

Listing 9.4: Cluster Code

```perl
#!/usr/bin/perl -w
use Bio::SeqIO;
use Bio::Seq;
use strict;
use Getopt::Long;
my ($node_num, $infile, $rc, $invdb, $dbusr, $dbpw) = '';
```
Figure 9.3: Distributed memory approach

```perl
&GetOptions("-i=s"=>$infile,
"-rc=s"=>$rc,
"-invdb=s"=>$invdb,
"-np=s"=>$node_num,
"-u=s"=>$dbusr,
"-p=s"=>$dbpw,
);
unless($node_num && $infile && $rc && $invdb && $dbusr && $dbpw) {
    print "prompt>sendToNodes_inv.pl-i_path_to_input_file
    -rc_F-invdb_invdb_-np_number_of_nodes_to_use\n";
    exit;
}
&chunk_file;
&run;
exit;
sub chunk_file {
    my $seq_in = Bio::SeqIO->new ("-file =>"$infile",
        -format => "fasta");
    my $seq_num = 0;
    while (my $seqobj = $seq_in->next_seq()) {
        $seq_num ++;
    }
    my $seqsPerChunk = $seq_num/$node_num;
    if ( ($seq_num % $node_num) != 0 ) {
        $seqsPerChunk = 1 + int($seqsPerChunk);
    }
    print "node_num=$node_num,s_seq_num=$seq_num,
```
```perl
CHAPTER 9. PROGRAMMING CODE

seqsPerChunk=$seqsPerChunk

unless (open (IN, $infile)) {
  die "Could not open $infile";
}
my $line = <IN>;
for (my $i = 1; $i <= $node_num; $i++) {
  mkdir ("./chunk$i");
  chdir("./chunk$i/");
  my $file="./chunk$i.fa";
  unless (open (OUT, ">$file")) {
    die "Could not open chunk$i.fa for writing";
  }
  my $count=1;
  while ((($line) && ($count <= $seqsPerChunk))) {
    if ( ($line =~ />)) {
      $count++;
      print OUT $line;
      $line=<IN>;
      while (($line =~ m/\w+$/) and ($line)) {
        print OUT $line;
        $line=<IN>;
      }
    }
  }
  close (OUT);
  chdir("..");
}

sub run {
  for (my $i=1; $i <= $node_num; $i++) {
    chdir("./chunk$i/");
    system("/opt/slurm/bin/srun -b -N1 perl findInvertedRepeats.pl -i $invdb -i_chunk$i.fa -rc $rc -u $dbusr -p $dbpw &");
    chdir("..");
  }
}
```
Chapter 10

Conclusions and future work

The major contribution of this thesis was to produce first sets of mRNA-like non-coding RNA candidates for the incompletely sequenced model legumes *Medicago truncatula* and *Lotus japonicus*, as a basis for both further experimental study and further computational analysis and characterization. A functional analysis, and in particular, the question of which of the ncRNAs presented encode miRNAs is also presented in this study.

The computational purification pipeline I described produced sets of 673 *M. truncatula* and 1637 *L. japonicus* putative mRNA-like non-coding RNA genes using ESTs and genomic sequences. These sets of candidate ncRNAs differ significantly in several features in comparison to those of mRNAs and also non-transcribed sequences, and I showed several base composition biases previously noted in the literature. Further analysis demonstrated that structure asymmetries exist on transcribed sequences and, especially when used in combination e.g. using SVM classification, the asymmetry features described in this thesis may be a useful starting point for future work on *de novo* ncRNA prediction.

Analysis of transcript expression profiles suggests that these candidate ncRNAs are generally expressed at a low level, and also they have different expres-
sion patterns in several tissue types. Particular, in *M. truncatula*, nodule tissue shows a dramatically increased expression of the candidate ncRNA set relative to mRNA. This suggests an important role for ncRNA in nodule formation and development.

I showed that many of these studied putative ncRNAs are conserved in closely-related legume species, suggesting that many ncRNAs tend to evolve rapidly through evolution and have particular roles in this family of plants. I further showed that candidate ncRNA genes often partner with protein genes in bidirectional transcription units.

I designed PCR primers for the *M. truncatula* ncRNA set to validate their expression (ongoing). Initial experiments with a qRT-PCR assay demonstrated that this putative ncRNA set consists of genuine transcribed sequences.

I further developed a computational pipeline for predicting miRNAs and applied it to putative ncRNA transcripts which I have identified, as well as intronic regions of *Medicago truncatula* and *L. japonicus*. This has led to the discovery of thousands of potential miRNA candidates and their corresponding targets. I investigated miRNA conservation in other plant species, clustered highly similar miRNAs, and provided a functional classification of target genes. To make this information available, I constructed the database, MIRATdb, to provide a flexible access to the large amount of data associated with the prediction of putative miRNAs and target genes for these two model legumes. This database provides the research community with a resource of putative *M. truncatula* miRNAs, their potential target genes, and other supporting information to assist in the selection of candidate miRNAs and target genes for experimental design. A variety of filters available through the database web-interface can be applied to the predicted miRNAs for more stringent settings at the users choice.

Thus, these putative ncRNA and microRNA gene sets provide a resource for
further investigation. Experimentally, we are planning further expression studies to validate and investigate these findings, and in particular the root nodule and other tissue expression differences. In addition, I note that the Medicago affymetrix GeneChip (http://www.affymetrix.com/products/arrays/specific/medicago.affx) contains probe sets for the majority of (i.e. 75%) the candidate non-coding transcripts identified here; so the current study provides useful information that will enable experimental researchers to link their expression data with the identified ncRNA genes on the chip. Computationally, I will further improve the non-coding RNA classifier to discriminate non-coding RNAs from coding RNAs and non-transcribed sequences using a machine learning approach. The characteristics of ncRNAs described in this thesis such as RNA secondary structure and base composition bias should aid future de novo ncRNA classification.

This work constitutes the first systematic ncRNA prediction in legumes and provides a basis for further analysis of ncRNA function. Some of the predicted ncRNAs have been classified as miRNAs. Other classes of ncRNAs will also need to be discovered through both computational and experimental analysis in the future research. As the genome of M. truncatula has not been completely sequenced, a further update of ncRNA analysis and MIRATdb might be needed when most or all the genomic sequences become available.
Appendix A

A.1 miRNA database schema

The MIRATdb schema description:

*Table Mir25*
Table description: Data for the 25mer miRNA and the miRNA*

- **Mir25Id**: VARCHAR(10) (primary key)
- **GiId**: VARCHAR(20) (foreign key)
- **IntronId**: VARCHAR(20) (foreign key)
- **MirClanId**: SMALLINT (foreign key)
- **MirRgnId**: VARCHAR(10) (foreign key)
- **PrecRgnId**: VARCHAR(10) (foreign key)
- **OrSrc**: VARCHAR(10) (Orientation is relative to the source sequence)
- **MirArm**: VARCHAR(10) (5’ or 3’ arm of precursor that miRNA located)
- **StartInSrc**: INT
- **StarStartInSrc**: INT
- **StartInMirRgn**: INT
- **StartInPrecRgn**: INT
- **StarStartInPrecRgn**: INT
- **CntMatch**: SMALLINT (Number of matches between 25mer miRNA and the miRNA*)
- **CntMisMatch**: SMALLINT (Number of mismatches between 25mer miRNA and the miRNA*)
- **CntGap**: SMALLINT (Number of gaps in alignment of 25mer miRNA and the miRNA*)
- **Lcm**: SMALLINT (Longest continues mismatches)
- **AlnStrUpr**: VARCHAR(40) (Upper alignment string (including gap) in 3’→5’ direction)
APPENDIX A. APPENDIX

AlnStrLwr : VARCHAR(40) (Lower alignment string in 5→3 direction)
Mir25Seq : VARCHAR(25) (Sequence in 3’ → 5’ direction)
Star25Seq : VARCHAR(25) (Sequence in 5’ → 3’ direction)
cntTgt : SMALLINT (Number of target genes)

Table Intron
Table description : Data for introns

IntronId : VARCHAR(20) (primary key)
GiId : VARCHAR(20) (foreign key)
cntMir25_asMirTr : SMALLINT (Number of miRNAs)
cntMirRgn : SMALLINT (Number of miRNA regions)
cntMirPrecRgn : SMALLINT (Number of miRNA precursor regions)
StartInGi : INT
Len : INT
Seq : LONGTEXT

Table Gi
Table description : Data for Gene Indices

GiId : VARCHAR(20) (primary key)
CdStatus : VARCHAR(2) (Coding Status)
(Note : possible values are :
NC for non-coding RNA
CD for protein encoding RNA
UN for Undefined RNA
NM for no mapping to BAC)
TrStatus : VARCHAR(1) (Direction of transcription)
(Note : Possible values are :
T for transcribed strand is given
N for non-transcribed strand is given
U for unknown direction of transcription
cntEST : SMALLINT (Number of ESTs)
cntMir25MirTr : SMALLINT (Number of miRNAs contained while considering as miRNA encoded transcript)
cntMirRgn : SMALLINT
cntMirPrecRgn : SMALLINT
cntMir25_asTgtTr : SMALLINT (Number of miRNAs contained while considering as target transcript)
cntTgt_asTgtTr : SMALLINT (Number of target sites contained while considering as target transcript)
Len : SMALLINT
Seq : LONGTEXT
NRAnn : TEXT (Best BLAST hit to SwissProt+trEmbl+NR)
A.1. MIRNA DATABASE SCHEMA

_Table MirTgt_
Table description: miRNA and target-sequence combination. Concerns only the region of the actual sequence match

Mir25Id : VARCHAR(10) (foreign key)
TgtId : VARCHAR(10) (foreign key)
StartInMir25 : INT
Len : SMALLINT
   (Note : As no gap is allowed in miRNA::target matching regions, only on length field is required for both the miRNA and target sites)
CntMatch : SMALLINT
CntMisMatch : SMALLINT
Lcm : SMALLINT
ScoreMatch : SMALLINT
TgtDeltaG : DOUBLE(10,3)

_Table Tgt_
Table description: Data for target sites

GiId : VARCHAR(20) (foreign key)
Orgi : VARCHAR(10) (Orientation relative to source Gi)
   (Note : possible values are :
   FORWARD
   REV_COM)
StartInGi : INT
Len : SMALLINT
TgtGC : DOUBLE(10,2)
TgtEntropy : DOUBLE(10,2)
TgtSeq : LONGTEXT

_Table PrecRgn_
Table description: miRNA precursor region is produced by merging overlapping precursors that are delimited by miRNA to miRNA*

PrecRgnId : VARCHAR(10) (primary key)
GiId : VARCHAR(20) (foreign key)
IntronId : VARCHAR(20) (foreign key)
   (Note : Either GiId or IntronId must be NULL. This indicates whether the miRNA stems from an intron or ncRNA exon.)
OrSrc : VARCHAR(10)
   (Note : Orientation is relative to the source sequence. Note that both Gi and Intron sequences may be in reverse complement orientation. However the Mir25 sequence will always have the same orientation as the PrecRng sequence.)
Possible values: FORWARD or REV_COM

StartInSrc : INT
Len : INT
Seq : LONGTEXT
  (Note: This is a derived field and contains a lot of data duplication. However, it may be worth storing the sequence as the sequence orientation can differ (i.e if OrScr is. REV-COM) relative to the Gi or Intron sequence)

cntMir25 : SMALLINT

Table MirRgn
Table description : Regions of overlapping 25mer miRNAs

MirRgnId : INT (primary key)
GiId : VARCHAR(20) (foreign key)
IntronId : VARCHAR(20) (foreign key)
StartInSrc : INT
Len : SMALLINT
cntMir25 : SMALLINT

Table MirPrec
Table description : The original miRNA precursor sequences (and miRNAs) are defined by the ends of the target matching sequences. The miRNAs where then merged or deleted to produce the Mir25 table. Therefore I do not need the precursor sequences for deleted miRNAs. This table only stores the precursor data that correspond to a Mir25 record (one to one). I ignore that the values are actually calculated from the target match boundaries

Mir25Id : VARCHAR(10) (primary key and foreign key)
PrecRgnId : VARCHAR(10)
StartInSrc : INT
StartInPrecRgn : INT
DeltaG : DOUBLE(10,2)
NormDeltaG : DOUBLE(10,2) (Normalized DeltaG by sequence length)
Len : SMALLINT
Seq : LONGTEXT

Table EstLib
Table description : Summarize TIGR EST libraries and www.medicago.org

EstLibId : VARCHAR(20) (primary key)
Description : VARCHAR(200)
TissueType : VARCHAR(20)
A.1. MIRNA DATABASE SCHEMA

Supplier : VARCHAR(100)
Tissue : VARCHAR(20)
Organ : VARCHAR(20)
Stage : VARCHAR(20)
CellType : VARCHAR(20)
CellLine : VARCHAR(10)
DiseaseState : VARCHAR(10)
Sex : VARCHAR(10)
Subtracted : VARCHAR(10)
TCs : SMALLINT
Singletons : SMALLINT
ESTTotal : SMALLINT

Tissue
Table description : Tissue catalogues to resolve many to many relationship of tissue and EST libraries

TissueId : SMALLINT (primary key)
TissueName : VARCHAR(20)

LibTissue
Table description : Relationships between EstLib and Tissue

LibTissueId : SMALLINT (primary key)
EstLibId : VARCHAR(20) (foreign key)
TissueId : SMALLINT (foreign key)

LibGi
Table description : Relationship between EstLib and Gi

LibGiId : SMALLINT (primary key)
EstLibId : VARCHAR(20) (foreign key)
GiId : VARCHAR(20) (foreign key)

GiEst
Table description : Relationship between Gi and EST

GiId : VARCHAR(20) (foreign key)
ESTId : VARCHAR(20)

MirClan
Table description : Clusters of Mir25 sequences with known miRNA thrown in. The 21nt miRNAs were used to perform clustering
APPENDIX A. APPENDIX

**MirClanId** : VARCHAR(10) (primary key)  
**Mir25Id**. **KnownMirId** : VARCHAR(10) (foreign key)

**KnownMir**  
Table description: Know miRNAs. Only need the ones that cluster together with my miRNAs

**KnownMirId** : VARCHAR(10) (primary key)  
**MirClanId** : VARCHAR(10) (foreign key)  
**MirRegistryId** : VARCHAR(20) (Ids from miRNA Registry)  
**MirFamily** : VARCHAR(10)  
**Len** : SMALLINT  
**Seq** : VARCHAR(30)

**Table HomMir25**  
Table description: Data for a homologues miRNA in a different species

**HomMir25Id** : VARCHAR(10) (primary key)  
**SeqDb** : VARCHAR(5)  
(Note: Possible values are: LJ/MT, AT, DbEST)  
**Accession** : VARCHAR(20)  
**OrDbSeq** : VARCHAR(10) (Orientation relative to SeqDb)  
**HomMirArm** : VARCHAR(20) (5’ or 3’ arm of precursor that miRNA located)  
**StartInDbSeq** : INT  
**StarStartInDbSeq** : INT  
**CntMatch** : SMALLINT (Number of matches between 25mer miRNA and the miRNA*)  
**CntMisMatch** : SMALLINT (Number of mismatches between 25mer miRNA and the miRNA*)  
**CntGap** : SMALLINT (Number of gaps in alignment of 25mer miRNA and the miRNA*)  
**Lcm** : SMALLINT (Longest continues mismatches)  
**AlnStrUpr** : VARCHAR(40) (Upper alignment string (including gap) in 3’ → 5’ direction)  
**AlnStrLwr** : VARCHAR(40) (Lower alignment string in 5’ → 3’ direction)  
**HomMir25Seq** : VARCHAR(30)  
**HomStar25Seq** : VARCHAR(30)

**Table HomMirPrec**  
Table description: This table is like an extension of the HomMir table and could probably be incorporated there. The exception being that the HomMir table refers to the 25mer, whereas this table refers to the 21mer. This is so because the precursor calculations were based on a sequence delimited by the 21mer miRNA and miRNA*
### A.1. MIRNA DATABASE SCHEMA

| HomMirId | VARCHAR(10) (Primary and foreign key) |
| DeltaG   | DOUBLE(10,2)                           |
| NormDeltaG | DOUBLE(10,2)                         |
| HomPrecEntropy | DOUBLE(10,2)                    |
| HomPrecGC  | DOUBLE(10,2)                          |
| HomPrecSeq | LONGTEXT                              |

**Table HomMirs**
Table description: Relationship between homologous miRNAs

| Mir25Id   | VARCHAR(10) (foreign key) |
| HomMir25Id | VARCHAR(10) (foreign key) |
| HomMisMatch | INT(1) (Number of mismatches) |

**Table Taxonomy**
Table description: Taxonomy data from DbEST. It enables me to specify the homologous taxa if a homologous miRNA was found in DbEST

| Accession | VARCHAR(200) (primary key) |
| TaxaId    | SMALLINT                   |
| Species   | VARCHAR(10)                |
| CommonName | VARCHAR(50)                |
| Classification | VARCHAR(100)            |

**Table MirConserv**
Table description: The table is also an extension to the Mir25 table and facilitates the taxonomic group where the miRNAs are conserved.

| Mir25Id   | VARCHAR(10) (primary key and foreign key) |
| InPlants  | BOOL (21nt miRNA is conserved in plant species) |
| InLegumes | BOOL (21nt miRNA is conserved in legume species) |
| InLJ(InMt) | BOOL (21nt miRNA is conserved in *Lotus japonicus* (for *Medicago truncatula* miRNAs) or in *Medicago truncatula* (for *Lotus japonicus* miRNAs) |
| InAT      | BOOL (21nt miRNA is conserved in *Arabidopsis thaliana*) |

**Table GiGO**
Table description: Data for Gene Ontology classification of Gis
GiId : VARCHAR(20) (foreign key)
GOAccession : VARCHAR(20)
GOCatalogue : VARCHAR(20)
(Note: Possible values F for molecular function and P for biological process
catalogue)
GODesc : TEXT

Table GiNR
Table description : Data for Gi blast against SwissProt+trEmbl+NR

GiId : VARCHAR(10) (primary key)
DB : VARCHAR(10)
(Note: Possible values : SProt,TrEmbl, GenBank)
E_value : DOUBLE(10,3)
Score : SMALLINT
Identity : DOUBLE(10,2)
Len : SMALLINT (Number of amino acids
Accession : VARCHAR(20)
Annotation : TEXT

MIRAT schema is also represented in the graphic form in Figure A.1
Figure A.1: MIRATdb schema
References


REFERENCES


REFERENCES


REFERENCES


References


REFERENCES


