Computational prediction of miRNA targets in plant genomes

Manuel Borges Dias dos Reis

Thesis to obtain the Master of Science Degree in Information Systems and Computer Engineering

Supervisors: Prof. Ana Teresa Correia de Freitas
Dr. Nuno Miguel Dias Mendes

Examination Committee

Chairperson: Prof. João António Madeiras Pereira
Supervisor: Prof. Ana Teresa Correia de Freitas
Member of the committee: Dr. Susana de Almeida Mendes Vinga Martins

May 2015
Acknowledgments

I would like to thank my friends and family, for their encouragement and support throughout this journey.

I would like to thank Professor Ana Teresa Freitas, for taking me under her wing and guiding me throughout this thesis. To Doctor Nuno Mendes, for his valuable advice and patience, for taking the time to guide me and teach me the Biology concepts needed for writing this thesis and also for sharing his knowledge not only in the Biology field, but also in Computer Sciences.

This work was supported by Fundação para a Ciência e a Tecnologia (UID/CEC/50021/2013), Project DataStorm - Large-Scale Data Management in Cloud Environments project (EXCL/EEI-ESS/0257/2012) and the project CAMP - Computational Analysis of MicroRNAs in Plants (PTDC/EIA-EIA/122534/2010).
Abstract

MicroRNAs (miRNAs) are posttranscriptional regulators which act by specifically binding to sites in their target messenger RNAs (mRNAs). They are present in nearly all eukaryotes, in particular in plants, where they play an important role in developmental and stress response processes by targeting mRNAs for cleavage or translational repression. Recent developments in high-throughput sequencing technologies have enabled the identification of a large number of new miRNAs whose function is unknown. A crucial step for proper functional annotation of miRNAs is the systematic identification of their targets. Most existing tools for target prediction were designed for animal miRNAs and, despite their similarities, the features of miRNA:target pairs in plants differ significantly from those found in animals. The need for miRNA target prediction tools specifically designed for plants has led to the development of a small number of widely used programs with a focus on sensitivity. But given that the determinants of an effective miRNA:target interaction remain elusive, existing methods produce an excess of predictions which are likely false positives. Here we present PINETREE – an efficient and customizable target prediction method. We use PINETREE to identify 133 of 167 validated targets in Arabidopsis thaliana and we compare its performance against two other tools: TAPIR and psRNATarget. We show that PINETREE attains comparable sensitivity while reducing the overall number of predictions. Additionally, we present a novel approach to miRNA functional annotation based on the identification of anti-targets – transcripts which are much less likely to contain functional targets, for a given miRNA, than what would be expected by chance. Our approach is suitable for genome-wide miRNA target searches and is much faster than current methods. Search parameters are easy to interpret and can be adjusted to match the characteristics of new plant genomes or miRNA:target interaction models. Moreover, the anti-target detection feature can be used to contribute to a better understanding of the regulatory context of each miRNA.

Keywords

plant microRNA, algorithm for target prediction, complementarity matching, target-site accessibility, anti-targets
Resumo

Os microRNAs (miRNAs) são reguladores pós-transcricionais, que actuam ligando-se em locais específicos do seu RNA mensageiro (mRNA) alvo. Os miRNAs estão presentes em quase todos os eucariotas, particularmente em plantas onde desempenham um papel importante nos processos de desenvolvimento e resposta ao stress, actuando nos miRNAs alvo por mecanismos de clivagem ou repressão da tradução. Desenvolvimentos recentes nas técnicas de sequenciamento em larga escala permitiram a identificação de um grande número de novos miRNAs cuja função é desconhecida.

Um passo crucial para a anotação funcional adequada de miRNAs é a identificação sistémática dos seus alvos. A maioria das ferramentas existentes para a previsão de alvos foi desenhada para miRNAs de animais, e apesar de existirem semelhanças, as características de pares miRNA:alvo em plantas diferem bastante dos pares encontrados em animais. A necessidade de ferramentas para a previsão de alvos de miRNAs que seja especificamente desenhada para plantas levou ao desenvolvimento de um pequeno número de programas com foco na sensibilidade. No entanto, os determinantes de uma interacção miRNA:alvo eficiente permanecem indefinidos, e assim os métodos existentes produzem um excesso de previsões que são provavelmente falsos positivos.

No contexto desta tese apresentamos a ferramenta Pinetree – um método de previsão de alvos eficiente e personalizável. Utilizámos esta ferramenta para identificar 133 de 167 alvos validados em Arabidopsis thaliana e comparámos o seu desempenho com outras duas ferramentas: TAPIR e psRNATarget. Mostramos que este método atinge uma sensibilidade comparável, reduzindo o número total de previsões. Apresentamos também uma nova abordagem para a anotação funcional de miRNAs baseada na identificação de anti-alvos – transcritos que têm menor probabilidade de conter alvos funcionais, para um dado miRNA, do que seria esperado por acaso. A nossa abordagem é apropriada para a procura de alvos de miRNAs à escala do genoma, e é muito mais rápida que os métodos actuais. Os parâmetros de procura são fáceis de interpretar e podem ser ajustados para corresponderem às características de novas espécies de plantas ou modelos de interacções miRNA:alvo. Além disso, a detecção de anti-alvos pode ser usada para perceber melhor o contexto regulatório de cada miRNA.

Palavras Chave

microRNA em plantas, algoritmo para previsão de alvos, procura por complementaridade, acessibilidade do alvo, anti-alvos
## Contents

1 Introduction
   1.1 Introduction ................................................. 1
   1.2 Problem .................................................. 2
   1.3 Objectives and Approach ............................... 3
   1.4 Contributions .......................................... 6
   1.5 Thesis outline ......................................... 6

2 Related work ................................................. 7
   2.1 MicroRNAs .................................................. 8
      2.1.1 MicroRNA Biogenesis ............................. 8
      2.1.2 Function ........................................... 9
   2.2 MiRNA target prediction ............................ 10
      2.2.1 Experimental Methods ......................... 10
      2.2.2 Computational Methods ....................... 12

3 The Pinetree pipeline ....................................... 19
   3.1 Principles and Pinetree back-end implementation ............. 20
      3.1.1 Complementarity ................................ 20
      3.1.2 Accessibility .................................... 22
      3.1.3 Anti-targets .................................... 23
      3.1.4 Implementation details ....................... 25
   3.2 Pinetree front-end implementation .................... 26
      3.2.1 Web-interface ................................ 26
      3.2.2 File output ..................................... 28

4 Results and discussion ..................................... 30
   4.1 Dataset description .................................. 31
   4.2 Performance ........................................... 31
   4.3 Discussion ............................................. 34
List of Figures

1.1 Framework pipeline for target prediction ........................................ 4
1.2 Framework pipeline for anti-target search ..................................... 5

2.1 Major biogenesis pathways of plant miRNAs, adapted from Ding et al. [2012] 9
2.2 High throughput sequence, adapted from Ding et al. [2012] ............ 11
2.3 Example of a Smith-Waterman execution ...................................... 15
2.4 Target site accessibility, adapted from Ding et al. [2012] ............... 15
2.5 Target site multiplicity, adapted from Ding et al. [2012] ............... 17

3.1 Precision / recall curve for different expectation values .................. 21
3.2 Precision and recall curves for different cutoffs of the alignment score .... 21
3.3 Precision and recall curves for different cutoff values of the accessibility score . 22
3.4 Evolutionary score calculation ......................................................... 23
3.5 Creation of statistical models from transcripts .............................. 24
3.6 Front-page PINETREE website ....................................................... 27
3.7 PINETREE server scheme .............................................................. 27
3.8 Output for target prediction execution mode ............................... 28
3.9 Output for anti-target search execution mode .............................. 29

4.1 Venn diagrams showing the comparison of target predictions ............ 32
4.2 Plot showing the distribution of affinities ....................................... 33

A.1 Script for preprocessing miRNA and validated targets datasets ........... 49
A.2 Script for pre-processing the results from tools ........................... 49
List of Tables

4.1 Datasets description .................................................. 31
4.2 Comparison of plant miRNA target prediction tools performance .......................... 32
A.1 Plant miRNA target prediction tools, adapted from Srivastava et al. [2014] ........ 48
List of Tables
Glossary

**AGO** ARGONAUTE protein

**cDNA** complementary DNA

**DNA** deoxyribonucleic acid

**DCLI** Dicer-like protein

**FASTA** DNA sequence alignment software package

**FASTA format** text-based format for representing nucleotide sequences, in which nucleotides are represented using single-letter codes

**GMUCT** genome-wide mapping of uncapped transcripts

**GUI** graphical user interface

**mRNA** messenger RNA

**miRISC** miRNA induced silencing complex

**miRNA** microRNA

**ncRNA** non-coding RNA

**PARE** parallel analysis of RNA ends

**PCR** polymerase chain reaction

**PDO** PHP Data Objects

**POL II** RNA polymerase II enzyme

**pre-miRNA** precursor miRNA

**pri-miRNA** primary miRNA

**qRT-PCR** quantitative real-time PCR

**RACE** rapid amplification of cDNA ends
Glossary

**RISC** RNA-induced silencing complex

**RNA** ribonucleic acid

**sRNA** small RNA

**ta-siRNA** trans-acting siRNA

**UTR** untranslated region
1

Introduction

Contents

1.1 Introduction .................................................. 2
1.2 Problem ...................................................... 2
1.3 Objectives and Approach ................................. 3
1.4 Contributions ................................................ 6
1.5 Thesis outline ............................................... 6
1. Introduction

1.1 Introduction

MicroRNAs (miRNAs) are part of a family of small RNAs (sRNAs) found in both animals and plants, which are distinguished from other small RNAs by their biogenesis, despite sharing some effector mechanisms (Ghildiyal and Zamore 2009).

These small regulatory RNAs (20-24nt) were initially studied about two decades ago, with the discovery that *lin-4* and *let-7*, which control developmental timing in *C.elegans*, did not code for protein but were rather RNA transcripts which negatively regulated gene expression posttranscriptionally (Lee et al., 1993).

MiRNAs bind to their target messenger RNA (mRNA) at particular sites and in plants they primarily silence targets through cleavage-dependent RNA degradation, but also by translational repression pathways (Brodersen et al., 2008). The perfect, or near-perfect (Rhoades et al., 2002) pairing of miRNA and its target site elicits endonucleolytic cleavage of the mRNA by AGO (ARGONAUTE) proteins.

These small molecules have been known to play a part in developmental timing regulation (Reinhart et al., 2000), apoptosis (Brennecke et al., 2003) and cell proliferation (Lecellier et al., 2005). More recently, studies have shown that some of them even act as potential tumor supressors (Johnson et al., 2005), oncogenes (He et al., 2005), and might turn out to be important drug targets (Maziere and Enright, 2007).

The identification of a large number of miRNAs existing across different species has increased the interest in unraveling their regulatory roles. It has been shown that one miRNA may regulate more than one target gene and vice-versa (Enright et al., 2004), and this way they are likely to being master switches in many biological pathways (Enright et al., 2004; Lim et al., 2005), so understanding this network of regulatory control is dependent on the *bona fide* identification of miRNA targets.

In the context of this thesis, we developed Pinetree (Plant microRNA Target Explorer), a new pipeline that incorporates upgraded versions, in terms of computational performance, of current approaches for target prediction methods in plants, along with a new approach for the miRNA functional annotation. This approach is based on the hypothesis that a transcript evolves towards eliciting or avoiding the action of miRNA in a selective way, introducing a new perspective for miRNA functional annotation that is not biased by current target prediction techniques.

1.2 Problem

Identifying miRNA targets is a major challenge and it is a fundamental step for their functional annotation. Most of the target prediction tools were developed for animal miRNAs which are significantly different from plant miRNAs in the target recognition process (Li et al., 2010). There are two main differences that make animal target prediction tools unsuitable for plants, leaving a
1.3 Objectives and Approach

gap in the computational target prediction for plants. In first place, an animal microRNA typically regulates its target gene's expression at the translational level, as opposed to the plant microRNA that regulates its target by a cleavage mechanism. Furthermore, an animal microRNA elicits the recognition of the 3'UTR region of the target, whereas a plant miRNA doesn’t have a general preference in terms of position.

In second place, the complementarity of a miRNA with its target differs from animals to plants. The animal miRNA mature sequence usually requires a flexible complementarity in the first eight nucleotides of the miRNA with is mRNA target, whilst a plant miRNA requires a near perfect alignment between the miRNA mature sequence and its mRNA target. Both of these differences highlight the need for the development of a separate plant miRNA target prediction tool, one that fits the specific characteristics of the plant target recognition process.

However, this task is even made more difficult due to the many different mechanisms by which miRNAs regulate their target genes, such as target mRNA destabilization, translational repression and even activation of gene expression [Huntzinger and Izaurralde (2011)]. Additionally, the determinants for an effective miRNA:target association remain unknown – the interaction between the target mRNA and the miRNA occurs via incomplete and interrupted base pairing, and the rules that govern such interactions are incompletely characterized.

Moreover, very few mRNAs have been experimentally shown to be subject to miRNA regulation, limiting the number of examples from which the basic rules governing miRNA:target associations can be deduced. This lack of information is particularly serious since, even for mRNAs known to be regulated, the actual target site is usually not experimentally identified but rather inferred from its partial complementarity to the miRNA [Burgler and Macdonald (2005)].

1.3 Objectives and Approach

This Master thesis aims to contribute to the development of bioinformatic resources and tools for miRNAs target prediction in plants. Model organisms like the *A.thaliana* and the *O.sativa* species have been used as examples.

The approach that was followed and that will be detailed in Chapter 3 includes:

- The development of a tool that takes into account the specificities of the binding of plant miRNAs to their targets and incorporates the recent discoveries in plant miRNA target recognition;

- The investigation and implementation into a new framework of alternative hybridisation models;

The framework that was developed has two execution modes; one to predict putative miRNA targets in plant species, and one to identify potential anti-targets using statistical models. Fig-
1. Introduction

Figure 1.1 provides an overview of the target prediction execution mode. Each module was developed separately, making it easy to change and introduce new functionalities. Figure 1.2 provides an overview of anti-target search execution mode.

This work was supported by Fundação para a Ciência e a Tecnologia (UID/CEC/50021/2013), Project DataStorm - Large-Scale Data Management in Cloud Environments project (EXCL/EEI-ESS/0257/2012) and the project CAMP - Computational Analysis of MicroRNAs in Plants (PTDC/EIA-EIA/122534/2010).
Figure 1.2: Overview of the pipeline for anti-target search.
1. Introduction

1.4 Contributions

The main contribution of this thesis comprises a computation pipeline for the target and anti-target prediction for plant species, based on transcriptomic data. This work also led to a contribution with an article (Reis and Mendes [in review 2015]) to an international journal entitled "PINETREE: a plant miRNA target explorer", which was submitted for publication in Bioinformatics. I attended and orally presented my work (Reis et al., 2014a) at the Bioinformatics Open Days 2014, in Braga, Portugal and attended and orally presented my work (Reis et al., 2014b) at the XII Symposium on Bioinformatics, in Seville, Spain. I also presented my work in a INESC-ID seminar.

1.5 Thesis outline

This thesis is organized in five chapters:

Chapter 2 - Related work: Provides a set of notions that we considered important to fully understand the contents of this thesis. We discuss both the experimental and computational techniques to predict miRNA targets in plant, with a special focus on the computational techniques. A survey on the latest discoveries and state-of-the-art miRNA target prediction methods in plants are available.

Chapter 3 - The PINETREE pipeline: Explains the several modules that comprise our tool, and how each module works in detail.

Chapter 4 - Results and discussion: Presents all the obtained results and discusses them based on our own expectations and by comparing them with the results of similar studies.

Chapter 5 - Conclusions: This chapter presents a summary of major achievements of our approach for target prediction of miRNAs in A.thaliana and O.sativa. The future work and perspectives are presented, highlighting the potential improvements of the pipeline.

Appendix A: Compiles additional information that was important during the pipeline development.

Appendix B: This chapter contains the full documentation of PINETREE, including the install and usage guide
Related work

Contents

2.1 MicroRNAs ......................................................... 8
2.2 MiRNA target prediction ................................. 10
2. Related work

2.1 MicroRNAs

MicroRNAs (miRNAs) are endogenous small (20-24 nucleotides long) RNAs that can play important regulatory roles in animals and plants by targeting mRNAs for cleavage or translational repression (Rhoades et al., 2002). Although the first microRNAs were identified in 1993 (Lee et al., 1993), they escaped notice until the last decade. MiRNAs comprise one of the most abundant classes of gene regulatory molecules in multicellular organisms and likely influence the output of many protein-coding genes (Bartel, 2004).

The multitude of small non-coding RNAs (ncRNAs) found in eukaryotes introduces an additional complexity in our understanding of the mechanism and proper classification of miRNAs. Like metazoan miRNAs, plant miRNAs (1) are endogenously expressed ∼22 nt RNAs potentially processed from one arm of foldback precursors; (2) are generally conserved in evolution; and (3) can originate both from their own genes or introns (Bartel, 2004).

2.1.1 MicroRNA Biogenesis

Plant miRNAs are initially transcribed into primary miRNA transcripts (pri-miRNAs) by RNA polymerase II (POL II) (Xie et al., 2005) (see Figure 2.1), which are a few hundred to a few thousand base pairs long.

Subsequently, the pri-miRNA is processed originating a precursor sequence (pre-miRNAs) with a length ranging from 60 to a few hundred nucleotides, occasionally folding into multi-loop structures. The maturation of the miRNA from the pri-miRNA is a stepwise process that involves a Dicer-like protein – DCL1. The mature miRNA (20-24 nucleotides-long) is located in the pre-miRNA and is processed through at least the two RNAse III-mediated steps referred above (Bartel, 2004; Kim, 2005).

The DCL1 processes the pre-miRNA to the mature miRNA, producing a duplex of 20-24 nt small RNAs with 2 nt overhangs at the 3’ end of each strand. One of the strands is selectively incorporated into a ribonucleoprotein complex named RISC (RNA induced silencing complex) (Schwarz et al., 2003), where it guides the cleavage or translational repression of its target mRNAs by hybridizing with the target, thus being the miRNA the effector of gene regulation (Bartel, 2004).

There are two precursor-processing pathways that have been identified for plant miRNA genes. The primary pathway involves stem-to-loop processing in which the sequence and structure beyond the miRNA:miRNA* site are necessary and used by the cleavage pathway components to excise the mature sequences (Cuperus et al., 2010; Song et al., 2010; Werner et al., 2010).

The second pathway involves loop-to-stem processing in which only the structure between the miRNA and miRNA* is necessary for the cleavage pathway components to excise the mature sequences (Addo-Quaye et al., 2009b; Bologna et al., 2009). Most plant pri-miRNA hairpins produce a single miRNA/miRNA* duplex, but some loci consistently produce multiple duplexes (Zhang...
Figure 2.1: Major biogenesis pathways of plant miRNAs, adapted from Ding et al. (2012).

MiRNA biogenesis in plants is completed within the nucleus, and several accessory factors also contribute to the efficiency and fidelity of miRNA/ miRNA* excision in plants (Axtell et al., 2011; Voigt, 2009).

2.1.2 Function

The mechanism by which miRNAs regulate target gene expression has been a controversial subject, as there is evidence for target mRNA destabilization, translational repression and even activation of gene expression (Huntzinger and Izaurralde, 2011).

As animal and plant miRNA5s greatly differ in their biogenesis, these differences are also reflected in their requirements for target recognition. In plants, miRNAs can silence targets through RNA degradation but also well as translational repression pathways (Brodersen et al., 2008).

The perfect, or near perfect pairing of miRNA and its target site elicits endonucleolytic cleavage of the mRNA by AGO (Figure 2.1). It is more frequent to see this kind of mechanism in plants instead of animals. There are also cases in plants in which miRNAs cause reduced levels of protein, but not mRNA, suggesting that translational repression is directed by miRNA-induced silencing complex (miRISC) (Llave et al., 2002).

Overall, miRNAs typically repress gene expression, and it remains to be seen whether positive regulation of targets extends beyond the limited cases that have been uncovered so far. Methods for discerning these different mechanisms of target regulation will be discussed in the next section.
2. Related work

2.2 MiRNA target prediction

Matching miRNAs to specific target sequences is a challenging process, and several approaches have been adopted to identify functional interactions. We can divide these approaches in two kinds: experimental methods and computational methods.

The factors involved in an efficient miRNA:mRNA target association are not fully known, and as new experimental data comes to hand, the knowledge about these associations is expanded (Elefant et al., 2011).

Even though computational methods do not usually incorporate all the knowledge obtained from these experimental data, when comparing to experimental methods, they can output potential targets more easily and efficiently, which can facilitate the downstream investigations (Ding et al., 2012).

In this section, we will discuss the principles considered to exist when using experimental and computational methods, along with some popular available methods/tools based on the discussed principles, conveniently organized in Table A.1 in Appendix A in order to facilitate description.

2.2.1 Experimental Methods

Target-specific validation

From the several approaches designed to identify functional interactions, target-specific methods were introduced first. Target-specific experimental validation with commonly used techniques such as quantitative real-time PCR (qRT-PCR) (VanGuilder et al., 2008), western blot (Towbin et al., 1979) and 5’ rapid amplification of cDNA ends (5’RACE) (Anon, 2005), are used to validate miRNA:mRNA pairs.

In order for the miRNA to regulate the expression of its target, both should be co-expressed. This co-expression is typically demonstrated by performing northern blot analysis or qRT-PCR using total RNA isolated from a specific cell type and probes or using primers specific for a given miRNA and mRNA target (Kuhn et al., 2008).

Even though most miRNA targets appear to be regulated at the mRNA and protein level, such regulation is only visible at the protein level for some targets (Baek et al., 2008). For a target of a given miRNA to be considered functional, the modulation of miRNA concentration should correspond to a predictable change in the amount of protein encoded by the target mRNA.

A typical approach to validate the functional importance of a miRNA:target association is a transient over-expression of a given miRNA mimic in a cell type known to express the putative target protein and subsequent western analysis using a specific antibody against that protein. One limitation of monitoring protein concentration is that it may not be selective enough to distinguish between members of the same miRNA family with similar sequences.

Generally, the downstream effects of differential miRNAs can be observed at the protein level.
2.2 MiRNA target prediction

by western blot and at the mRNA level by qRT-PCR, although these measures will not distinguish between direct and secondary miRNA targets and it is hard to determine whether the target mRNA is regulated by one single miRNA predominantly or several miRNAs simultaneously.

In the specific situation where a miRNA target is directly cleaved, 5' RNA ligase mediated-RACE (5' RLM-RACE) may be used to evaluate such targeting. 5'-RACE is a PCR-based technique, whereby an RNA adapter is ligated to the free 5' phosphate of an uncapped mRNA produced from, among other nucleolytic activities, AGO2-directed mRNA cleavage.

The ligation product can be reversely transcribed using a forward primer directed against the linker and a gene specific reverse primer which is subsequently PCR amplified, cloned and identified by sequencing. 5'-RACE has been employed extensively to validate products of RISC-mediated cleavage in plants (Llave et al., 2002).

High-throughput

In order to map RNA-RNA cleavage sites in a more comprehensible way, parallel analysis of RNA ends (PARE) was introduced (also known as Degradome-seq or genome-wide mapping of uncapped transcripts (GMUCT)).

Degradome libraries are built by ligation of polyA-enriched RNA samples to a custom RNA adaptor containing a 3' Mmel (restriction enzyme) site, followed by reverse transcription, second-strand synthesis, Mmel digestion, ligation of a 3' dsDNA adaptor, gel purification and PCR amplification, as shown in Figure 2.2.

Widespread mRNA cleavage events regulated by miRNAs have been identified in Arabidop-
2. Related work

sis (Addo-Quaye et al., 2008; German et al., 2008, 2009), rice (Li et al., 2010; Wu et al., 2009), grapevine (Pantaleo et al., 2010) and limited cleavages in mammals (Bracken et al., 2011; Karginov et al., 2010; Shin et al., 2010), which shows the capabilities of PARE and high-throughput sequencing.

The use of PARE identifies a large subset of miRNA targets that are subject to direct cleavage thus being most suited for plants, since extensive base-pairing between miRNAs and mRNAs leads to direct RISC-mediated cleavage, which does not appear to be a major mechanism of miRNA activity in mammals (Eckardt).

Degradome data can be used to find evidence of cleaved sRNA targets without resorting to computational predictions. The current task with this process is to develop effective and efficient pipelines to make better use of these data. From this task, three pipelines have been proposed to process Degradome data: CleaveLand (Addo-Quaye et al., 2009a), Seq-Tar (Zheng et al., 2012) and PARE-snip (Folkes et al., 2012).

However, there are some issues related with the processing of Degradome data. The computational cost to do so is high, when it comes to computation time and hardware cost. Also, current pipelines always generate much more candidate pairs than conventional computational methods, leading to the belief that not all of them are real predictions. Furthermore, considering that Degradome-seq uses a principle similar to 5'-RACE, novel experimental methods are required to evaluate candidates identified by these pipelines. Finally, when using Degradome-seq technology we can only identify cleavable targets, which means that non-cleavable targets will be missed.

2.2.2 Computational Methods

Complementarity - Principles

When predicting targets in plants, complementarity is considered a key principle, since the stability of a miRNA:mRNA association is determined by the complementarity between the miRNA and the target site (Mallory et al., 2004; Schwab et al., 2005).

A major conclusion for metazoan miRNA target prediction is that requiring conserved Watson-Crick pairing to the 5' region of the miRNA, centered in nucleotides 2-7 (which is called seed), reduces the occurrence of false-positive prediction (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2003, 2005). The discovery that perfect seed pairing substantially improves prediction reliability implied that it was also important for miRNA target recognition, and this principle was later adopted for target prediction in plants.

However, tools based in complementarity have yet to find a scoring scheme that is not based in excessively stringent criteria. Generally, the target prediction tools based on complementarity expect a perfect match, or a single G:U wobble between the seed region of the miRNA and target site of the mRNA, but exceptions to this criterion have been reported (Jones-Rhoades and Bartel...
2.2 MiRNA target prediction

[2004], leading to the conclusion that a stringent seed region matching rule may miss potential target sequences for some miRNAs.

On the other hand, some systematic mutagenesis studies emphasize the importance of seed regions for target recognition in *Arabidopsis* [Mallory et al. 2004, Schwab et al. 2005]. Additionally, the experimental characterization of a number of different canonical seed types of varying lengths or with a specific initial base [Bartel 2009], rather than a single class of seed sequences, makes it hard to develop a consistent computational method based in complementarity alone.

The specificity of long seeds has been verified in a study, but in general the functional target sites are formed by less specific seeds with a length of 6-nt [Ellwanger et al. 2011]. An alternative binding mode by which miR-124 can regulate its target with a G-bulge site (position 5-6) has been identified, which cannot be explained by canonical seed matches [Chi et al. 2012].

The degree of complementarity in the central region (position 9-11) is often presented as a decisive characteristic for determining whether slicing or translational repression and mRNA decay follows from RISC recruitment [Brodersen and Voinnet 2009]. Also, a study in humans showed that functional miRNA target sites that only pair up with this central region lead to translational inhibition or mRNA decay [Shin et al. 2010], and this may also apply to plants. Many computational methods already consider this region, when scoring a miRNA:mRNA association [Dai and Zhao 2011, Ding et al. 2011a, Moxon et al. 2008, Xie and Zhang 2010]. As showed in Table A.1 (Appendix A), psRNATarget [Dai and Zhao 2011] and imiRTP [Ding et al. 2011a] predict the functional type of miRNA based on the complementary at the central region of the miRNA:target pair.

The differences between these computational methods are the position of the central region, and whether a mismatch or a wobble pair is allowed in this region and, if it is allowed, whether the functional type of this miRNA:mRNA pair is also predicted. In addition to seed and central regions, it is thought that miRNA nucleotides 16 and 19 have an important role for target recognition in *Arabidopsis* [Palatnik et al. 2007].

In another study, it is stated that perfect complementarity between the 3’ end of miR173 and the 5’ end of AT2G39675 (TAS1c) is fundamental, by systematically mutating the miR173 target site. Mismatches at the 3’ end of miR173 abolish trans-acting siRNA (ta-siRNA) formation, while mismatches at the 5’ end had less effect [Zhang et al. 2012]. Unlike the seed and central region, the 3’ backbone is not fully considered by current computational methods.

All of these exceptions, as well as the fact that not all factors for an efficient association are known, show the complexity of the mechanisms behind target recognition, hence the requirement for improved approaches that do not solely apply simple matching rules between miRNA and target mRNA. Many potential targets may be missed because of stringent-pair types, where less stringent-pair may increase the number of false positives.
2. Related work

Complementarity - Algorithms

Typical approaches take a given nucleotide sequence and search a corresponding sequence database by using local sequence alignment to find matches of similar database sequences. FASTA and ssearch are two widely used tools for this purpose.

The FASTA program follows a largely heuristic method which contributes to the high speed of its execution. It initially observes the pattern of word hits, word-to-word matches of a given length, and marks potential matches before performing a more time-consuming optimized search using a Smith-Waterman implementation. The size taken for a word, given by the parameter $ktup$, controls the sensitivity and speed of the program. Increasing the $ktup$ value decreases number of background hits that are found. From the word hits that are returned the program looks for segments that contain a cluster of nearby hits. It then investigates these segments for a possible match.

The ssearch performs a rigorous Smith-Waterman alignment between a DNA sequence to another DNA sequence or a DNA library which is much slower than the previous algorithm, but performs a more comprehensive search.

Algorithm 1 presents the pseudo-code for a Smith-Waterman implementation, where $a$ and $b$ are the strings to align, $m = \text{length}(a)$, $n = \text{length}(b)$, $s(a, b)$ is a similarity function on the alphabet, $H(i, j)$ is the maximum similarity score between $a_i$ and $b_j$, $g$ is the gap penalty and $T(i, j)$ is the backtrace of the maximizing pair between $a_i$ and $b_j$.

Algorithm 1 Smith-Waterman alignment pseudocode

**Initialize:** $H(i, 0) \leftarrow 0, 0 \leq i \leq m$  $H(0, j) \leftarrow 0, 0 \leq j \leq n$

1: for $1 \leq i \leq m$ do
2:     for $1 \leq j \leq n$ do
3:         Insertion $\leftarrow H(i, j - 1) + g$
4:         Deletion $\leftarrow H(i - 1, j - 1) + g$
5:         Match/Mismatch $\leftarrow H(i - 1, j - 1) + s(a_i, b_j)$
6:         $H(i, j) \leftarrow \max(0, \text{Insertion}, \text{Deletion}, \text{Match/Mismatch})$
7:         $T(i, j) \leftarrow \text{maximizing pair } (i', j')$
8:     ($i, j$) $\leftarrow (0, 0)$
9: while ($i, j$) $\neq (0, 0)$ do
10:     if $T(i, j) = (i - 1, j - 1)$ then
11:         print $a_{i-1}, b_{j-1}$
12:     else if $T(i, j) = (i - 1, j)$ then
13:         print $a_{i-1}, -$  
14:     else
15:     print $-, b_{j-1}$
16:     ($i, j$) $\leftarrow T(i, j)$

Figure 2.3 depicts an example of the algorithm execution, where the strings to be aligned are $a = ACACACTA$, $b = AGCACACA$ and the scoring schema used is $s(a, b)_{a=b} = 2$, $s(a, b)_{a\neq b} = -1$ and $g = -1$. 

14
2.2 MiRNA target prediction

(a) Scoring matrix

\[
H = \begin{pmatrix}
- & A & C & A & C & T & A \\
- & 0 & 0 & 0 & 0 & 0 & 0 \\
A & 0 & 2 & 1 & 2 & 1 & 0 & 2 \\
G & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\
C & 0 & 3 & 2 & 3 & 2 & 3 & 2 & 1 \\
A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 3 & 4 \\
C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 5 \\
A & 0 & 3 & 3 & 6 & 6 & 9 & 8 & 7 & 8 \\
C & 0 & 2 & 4 & 5 & 8 & 8 & 11 & 10 & 9 \\
A & 0 & 4 & 3 & 6 & 7 & 10 & 10 & 10 & 10 & 10 \\
\end{pmatrix}
\]

(b) Traceback matrix

\[
T = \begin{pmatrix}
- & A & C & A & C & T & A \\
- & 0 & 0 & 0 & 0 & 0 & 0 \\
A & 0 & \_ & \_ & \_ & \_ & \_ \\
G & 0 & \_ & \_ & \_ & \_ & \_ \\
C & 0 & \_ & \_ & \_ & \_ & \_ \\
A & 0 & \_ & \_ & \_ & \_ & \_ \\
C & 0 & \_ & \_ & \_ & \_ & \_ \\
A & 0 & \_ & \_ & \_ & \_ & \_ \\
C & 0 & \_ & \_ & \_ & \_ & \_ \\
A & 0 & \_ & \_ & \_ & \_ & \_ \\
\end{pmatrix}
\]

Figure 2.3: Example of a Smith-Waterman execution

Thermodynamics and target site accessibility

An effective miRNA:target association begins with the hybridization reaction on an open structure at the target site (see figure 2.4), making thermodynamic stability, or thermostability an important factor to consider when predicting targets. The thermostability comprises the free energy of a miRNA:mRNA duplex, since hybridization between two complementary RNA sequences occurs in a more favorable state, meaning that the lower the free energy between the miRNA:target pair, the more energy is required to disrupt this association.

As stated, a miRNA:target interaction occurs if the secondary structure at the target site is opened, and target site accessibility has been proven to be an important determinant for a successful target binding process. A target site is considered accessible for binding to a miRNA if it can be opened and not interact with other sites within the mRNA (Hofacker et al., 2007; Kertesz et al., 2007). The physicochemistry of miRNA-mRNA associations can be used for miRNA target recognition. The thermostability of this associations is not generally a very good predictor, whereas the miRNA target site accessibility has been proven to be one of the effective structural features for the prediction of target site functionality (Hausser et al., 2009; Li and Zhang, 2005).

Figure 2.4: Target site accessibility, adapted from Ding et al. (2012)

Thus, the secondary structure is very important for predicting miRNA targets, as an effective
2. Related work

target site on mRNA is expected to have less mRNA secondary structure blockage, which affects miRNA access to the target site (Bergauer et al., 2009; Bonnet et al., 2004; Ding et al., 2010; Hauser et al., 2009; Li and Zhang, 2005).

The first step of an effective miRNA-mRNA interaction is the hybridization reaction on an opening structure at the target site. Upon binding, miRISC (miRNA induced silencing complex) can disrupt the secondary structure at the site to elongate hybridization (Kertesz et al., 2007; Long et al., 2007). A detailed investigation of accessibility of target sites was performed recently, where they found that the efficacy of miRNA strongly depends on the accessibility of both the 5’ and 3’ end of their binding sites (Kiryu et al., 2011).

UEA sRNA (Moxon et al., 2008), WMD3 (Ossowski et al., 2008), TAPIR and p-TAREF first calculate the hybridization energy of miRNA:target duplex, then compare it with the optimal hybridization energy, and finally use the percentage as a filter to choose potential targets. Another two methods, psRNATarget and imiRTP, calculate the accessibility by RNAup program in Vienna RNA package (Lorenz et al., 2011).

Despite being a central principle for target recognition, calculating accessibility can be computationally expensive, depending on the size of the mRNA sequence. Also, the thermodynamic models used in RNA secondary structure prediction algorithms are not very accurate (Eddy, 2004) for sequences with a length beyond a few hundred nucleotides.

Evolutionary conservation of target sites

The conservation of binding site is another commonly applied criterion to filter miRNA target candidates. miRNA families are well-conserved among related species, bound to have the same seed site and have targets that are conserved within related species (Ding et al., 2011b).

In some target prediction algorithms, the existence of conserved potential miRNA targets in orthologous sequences in closely related species have been used to re-enforce the prediction, in a way that reduces the rate of false positive targets (Ha et al., 2008). Several recent methods, like Targetfinder (Fahlgren et al., 2007), TAPIR (Bonnet et al., 2010), p-TAREF (Jha and Shankar, 2011), psRNATarget and imiRTP, still consider this principle as part of a scoring schema.

However, different definitions of conservation are used by the various prediction programs (Rajewsky, 2006). In general, target sites are considered conserved if they appear at orthologous locations in multiple genomes, meaning that they have to appear exactly at the same position in the alignment of the 3’UTR sequences. In other cases, a target site is regarded conserved when the seed region of a miRNA falls within overlapping alignment positions (Krek et al., 2005).

Target sites are also sometimes considered conserved if they can be found in the sequences regardless of being in aligned positions (Hammell et al., 2008). Additionally, we can consider that the target is missing or has mutated in one of the sequences of the multiple organisms that are considered, and we regard these target sites as poorly conserved.
Applying a conservation filter may decrease the false positive rate, but will only be effective for conserved miRNAs. When species-specific miRNAs are of interest, it is important to be able to identify non-conserved targets.

**Multiplicity of target sites**

Even though the importance of multiple target sites was generally underestimated in reported plant miRNA target prediction tools, increasing reports on translation inhibition of the miRNA targets in *Arabidopsis* have underlined the importance of the effect of such multiple target sites.

There have been reports of dual target sites on AtTAS3 which is the target gene of miR390, where AtTAS3 is the phasing precursor of ta-siRNA TAS3 and the biogenesis of the ta-siRNA shares similar target recognition mechanism with miRNA [Axtell et al., 2006]. For this case (see figure 2.5), two target sites represent two cleavage events, which constitute a stronger signal for degradation of target mRNA [Dai et al., 2011].

![Figure 2.5: Target site multiplicity, adapted from Ding et al. (2012)](image)

**mRNA degradation or translational inhibition**

The discovery of a widespread translational inhibitory component that is genetically separable from endonucleolytic cleavage was a major update on plant miRNA-guided silencing. It was concluded that translational repression can occur due to central mismatches in a miRNA:target duplex, since it prevents slicing, while central matches in miRNA:target duplexes tend to cleave the
target mRNA and exclude translational repression, apart from a few mismatches in other regions of the duplex (Brodersen et al., 2008).

Also, the translational target sites seem to be found in the coding region rather than in the 3’UTR region as in animals. These differences make target prediction tools designed for animals unsuitable for plant miRNA target prediction. This suggests that many ‘non-functional’ miRNAs, especially non-conserved ‘orphan’ miRNAs, need to be re-identified since experimental methods such as 5'-RACE PCR method would fail to detect the change of target genes at protein level (Brodersen et al., 2008).

Correlation expression for miRNA and mRNA targets

The expression level of target mRNAs or proteins is generally negatively correlated with the expression level of its corresponding miRNAs. The use of transcriptome data, particularly microarray and RNA-seq based expression data, should significantly reduce false positive predictions (Garber et al., 2011; Huang et al., 2007). Targets that are regulated at the RNA level can be identified by profiling mRNAs on microarrays. This technology is easily accessible but does not identify targets regulated only at protein level (Pais et al., 2010).

Messenger RNA expression data can be used to detect direct miRNA-mRNA target interactions (Lim et al., 2005) by looking for transcripts that appear to be downregulated after overexpression of a tissue-specific miRNA. Provided that data sets that profile the expression of miRNAs as well as mRNAs across many tissues are available, one is expected to find miRNA target relationships in two separate expression data sets profiling both miRNAs and mRNAs.
3

The PINETREE pipeline

Contents

3.1 Principles and PINETREE back-end implementation ............ 20
3.2 PINETREE front-end implementation .......................... 26
3. The Pinetree pipeline

3.1 Principles and Pinetree back-end implementation

Pinetree uses a pipeline-based architecture, fully implemented in the C programming language, which facilitates the integration of the FASTA application and RNAup program present in ViennaPackage.

The pipeline combines two different approaches: a search by complementarity, using the FASTA [Pearson and Lipman 1988] local alignment program (Section 3.1.1), followed by an accessibility analysis, using the RNAup [Mückstein et al. 2006] program, that calculates locally stable secondary structures (Section 3.1.2).

A third module, developed separately, introduces a novel approach, giving a score to each miRNA:mRNA pair based on their sequence characteristics with the aim of detecting anti-targets, as described in Section 3.1.3. The pipeline gives users the flexibility to set various parameters relevant for each processing step. A full list of the many command-line options is given in the user guide, present in Appendix B.

In the following section we will explain in detail how each module of our tool works.

3.1.1 Complementarity

As explained in 2.2.2, complementarity is an important principle to consider when predicting targets in plants, since the stability of a miRNA:mRNA association is largely determined by the complementarity between the miRNA and the target site [Mallory et al. 2004].

Pinetree assesses complementarity between the target transcript and miRNA using the FASTA program (version 36.3.7a) [Pearson and Lipman 1988]. In order to improve the precision of the predictions while retrieving at least as many targets as those predicted by other tools, we calculated the results for multiple expectation values (see Figure 3.1) for *A. thaliana*, setting the cutoff to 30. The remaining parameters were set to their default values.

In our pipeline, the FASTA program outputs multiple alignments, which are then analyzed to search for potential targets. To each of these alignments Pinetree assigns a Smith-Waterman-type score, applying a scoring scheme which has shown good results in evaluating the likelihood of functional miRNA:target pairs [Allen et al. 2005]. The scoring scheme distinguishes different portions of the hybridization site, including the “seed” region, which is set to occur at positions 2 – 12.

Our score relies on the local alignment determined by FASTA but then extends the alignment to the full length of the miRNA. A perfect alignment has score 0, all other possibilities have positive values according to the following sum of penalties:

\[
S = S_{\text{mis}} + \frac{S_{\text{wob}}}{2} + S_{\text{gap}} + 2 \left( S_{\text{mis}}^* + \frac{S_{\text{wob}}^*}{2} + S_{\text{gap}}^* \right)
\]

(3.1)

where \(S_{\text{mis}}\), \(S_{\text{wob}}\), and \(S_{\text{gap}}\) represent the number of mismatches, G:U wobbles and gaps outside
3.1 Principles and Pinetree back-end implementation

Figure 3.1: Precision / recall curve for different expectation values in the *A. thaliana* dataset.

The seed region, respectively, and $S_{\text{mis}}^*$, $S_{\text{wob}}^*$, and $S_{\text{gap}}^*$, the same counts in the seed region.

The alignment score is then used to filter the most promising target candidates by introducing another cutoff (here, set to 3.5), which maximizes the weighted harmonic mean between recall and precision, giving recall twice as much weight (see Fig. 3.2).

Figure 3.2: Precision and recall curves for different cutoffs of the alignment score in the *A. thaliana* dataset.

When mismatches occur in the central region of the miRNA:mRNA duplex, at positions 9 – 11, the cleavage-dependent regulatory process is disabled, but the miRNA effector mechanism might still be functional, inhibiting translation instead (Brodersen et al., 2008). Pinetree is able to identify and report such potential translational inhibition events.
3. The Pinetree pipeline

3.1.2 Accessibility

Pinetree evaluates target site accessibility using RNAup (Mückstein et al., 2006), which is part of the Vienna Package (Lorenz et al., 2011). RNAup calculates the thermodynamics of RNA-RNA interactions. In particular, it can calculate the opening energies associated with exposing a stretch of a potentially structured portion of an RNA molecule.

To serve our purposes, the tool was modified to only output relevant information. Here we are interested only in calculating the opening energies of a stretch of \( n \) consecutive nucleotides around the candidate target site, which directly translate to our accessibility score. In our case, \( n \) corresponds to the size of the miRNA:mRNA alignment determined previously, plus an upstream and downstream region. These flanking regions account for the bulk of the RNA-induced silencing complex (RISC), which benefits from encountering an unstructured region for a more favourable hybridisation. We follow the recommendations in the literature, stating that the size of these regions should correspond to 17-nt upstream and 13-nt downstream (Kertesz et al., 2007).

Using the same rationale as before, with the aim of reducing the number of false positive predictions with a minimal impact on the number of validated targets retrieved, we computed the results, for multiple cutoff values, of adding a filter by the accessibility score (see Fig. 3.3), thereby setting the cutoff to 15.5. This means that target sites requiring 15.5 kcal/mol or less to become unstructured are deemed accessible.

![Figure 3.3: Precision and recall curves for different cutoff values of the accessibility score in the A. thaliana dataset.](image)

Since computing the accessibility for each putative interaction is a computationally expensive procedure, we allow the user to choose to disable this feature by turning off the accessibility module.
3.1 Principles and Pinetree back-end implementation

3.1.3 Anti-targets

Anti-targets are transcripts that have evolved to avoid miRNA-mediated silencing, presumably because such negative regulation would have a deleterious effect. Any attempt at identifying these anti-targets must then find a strategy to detect an unexpected depletion of potential target sites for a particular miRNA.

In our case, this strategy consists of comparing the sequence characteristics of each individual transcript against the background of the whole transcriptome. Thus, for each miRNA, we propose to determine how likely is the occurrence of a perfect target site (reverse-complementary sequence) in a given transcript compared to the same measure taken in the set of all transcripts.

This way we can mitigate the influence of sequence biases unrelated to the miRNA milieu each transcript is exposed to. An anti-target will then be a transcript that is much less likely to harbour a target site for a particular miRNA than what the sequence characteristics of the transcriptome would suggest.

Our assessment is based on the notion of affinity between a miRNA, \( \mu \) and a transcript, \( t \in T \) (\( T \) is the set of all transcripts), given a sequence model for the transcript (\( M_t \)) and the transcriptome (\( M_T \)) (see Figure 3.4).

Figure 3.4: Evolutionary score calculation

The affinity, \( \alpha \), is expressed as a log-odds score:

\[
\alpha(\mu, t) = \log P(\tilde{\mu} \mid M_t) - \log P(\tilde{\mu} \mid M_T)
\]  

(3.2)

\( P(\tilde{\mu} \mid M) \) is the probability that the sequence model \( M \) generates the reverse-complementary sequence of \( \mu \).

Here, sequence models correspond to Markov chains of order \( n \), which are learnt from the corresponding sequences (see Figure 3.5). To account for the zero-frequency problem, pseudo-counts are introduced for unobserved transitions, \( i \to j \), by adding 1 to the frequency of all transitions.
3. The Pinetree pipeline

exiting state \( i \), and to the initial probabilities of all states.

Let \( S = S_1..k \) be a sequence of length \( k \), \( M \) a Markov chain of order \( n \) (with \( 0 < n \leq k \)), \( \pi_0(i) \) the initial probability of state \( i \), and \( \pi(i \rightarrow j) \) the transition probability from \( i \) to \( j \) then, \( P(S \mid M) \) is given by:

\[
P(S \mid M) = \pi_0(S_1..n) \prod_{i=1}^{k-n} \pi(S_{i..i+n-1} \rightarrow S_{i+1..i+n})
\]

(3.3)

The best anti-target candidates correspond to miRNA:mRNA pairs with the lowest affinity.

**Algorithm**

Algorithm 2 shows the basis for the algorithm, where \( \Sigma \) is the alphabet composed by \( \{A, C, T, G\} \), \( S = S_1..k \) is a sequence of length \( k \), \( M \) a Markov chain of order \( x \) (with \( 0 < x \leq k \)), \( \pi_0(i) \) the initial probability of state \( i \), \( \pi(i \rightarrow j) \) the transition probability from \( i \) to \( j \) and \( e \) is the number of errors left to give.

**Algorithm 2** Basis of the evolutionary score calculation

**Initialize:** Offset \( \leftarrow x \), Current \( \leftarrow S_1...,x \), \( P(S \mid M) \leftarrow 0 \)

1: for all \( n \in \{ \text{sequences of size } x \in \Sigma \} \) do
2: if Current \( \neq n \) and \( e \) then
3: Increment \( P(S \mid M) \) by \( \pi_0(n) \ast \text{CalcEscore}(e - 1, \text{Offset} + 1, n) \)
4: else if Current \( = n \) then
5: Increment \( P(S \mid M) \) by \( \pi_0(n) \ast \text{CalcEscore}(e, \text{Offset} + 1, n) \)
6: return \( P(S \mid M) \)

Algorithm 3 shows the recursion for the algorithm.
3.1 Principles and Pinetree back-end implementation

Algorithm 3 Recursion of the evolutionary score calculation

Initialize: \( P(S | M) \leftarrow 0 \)

1: procedure CalcEscore(e, Offset, Prev)
2:  \[ \text{if } \text{Offset} < \text{length}(S) \text{ then} \]
3:     \[ \text{for all } n \in \Sigma \text{ do} \]
4:     \[ \text{Current} \leftarrow (\text{Prev} \ll 2) | n \]
5:     \[ \text{if } S_{\text{Offset}} \neq n \text{ and } e \text{ then} \]
6:     \[ \text{Increment } P(S | M) \text{ by } \pi (\text{Prev} \rightarrow n) * \text{CalcEscore}(e - 1, \text{Offset} + 1, \text{Current}) \]
7:     \[ \text{else if } S_{\text{Offset}} = n \text{ then} \]
8:     \[ \text{Increment } P(S | M) \text{ by } \pi (\text{Prev} \rightarrow n) * \text{CalcEscore}(e, \text{Offset} + 1, \text{Current}) \]
9:     \[ \text{else} \]
10: \[ \text{for all } n \in \Sigma \text{ do} \]
11: \[ \text{if } (S_{\text{Offset}} \neq n \text{ and } e) \text{ or } S_{\text{Offset}} = n \text{ then} \]
12: \[ \text{Increment } P(S | M) \text{ by } \pi (\text{Prev} \rightarrow n) \]
13: \[ \text{return } P(S | M) \]

3.1.4 Implementation details

Pinetree uses a pipeline-based architecture, fully implemented in the C programming language which facilitates the integration with the FASTA (Pearson and Lipman, 1988) tool and the RNAup (Mückstein et al., 2006) program present in the ViennaPackage (Lorenz et al., 2011).

The FASTA program is called using the `popen` command, which allows Pinetree to process its results as they are being produced. We also take advantage of the parallelization mechanisms present in the FASTA program, using the `-T` flag when creating the FASTA process.

Since the RNAup program only accepts one sequence at a time as input, we do not call the program directly, but we link the relevant libraries with Pinetree instead. We parallelize the calls to RNAup using the `fork` system call, which creates a new process. Despite the fact that it is more costly to create new processes, an approach using shared memory space is not possible since RNAup is not reentrant.

For the calculation of the complementarity score, for each score normalization procedure, as well as for the anti-target search, we take advantage of multi-core systems to balance the workload via the OpenMP (open multi-processing) API (v3.1).

The created sub-tasks never communicate with each other because each miRNA:mRNA evaluation does not depend on any other data, and therefore it is trivial to distribute the workload among available cores using OpenMP directives.

Normalization

The model species A. thaliana has served as a basis for the development of the majority of existing tools for miRNA target prediction (Srivastava et al., 2014). This is justified by the fact that there is much more data available about this species compared to other plants, in particular, on validated miRNA targets.

For the same reason, we used an A. thaliana dataset to determine the best cutoffs for each
3. The Pinetree pipeline

module. For these cutoffs to be meaningful in other plant species, whose genome has different characteristics and will thus present a different range of scores (due, inter alia, to GC content disparities or different sequence biases) we introduce a score normalization procedure for the complementarity and accessibility modules. Our normalization consists in replacing the raw score, \( x \), by the corresponding \( Z \)-score, i.e. \( z = (x - \bar{x})/\sigma \).

Data structures

One of the optimizations we made in order to have a smaller memory footprint was to encode our nucleotide alphabet (A,C,T,G) into 2 bits (00, 01, 10, 11), and this way we spend less space as opposed to using 1 byte with the common char representation.

This letter encoding allows Pinetree to perform string concatenations simply by shifting 2 bits to left, and then performing an or operation. We store the string for transactions in a long long int, so we can use a Markov order up to 32 (even though orders bigger than 6 are rarely used).

The Markov chains are represented in the form of bi-dimensional matrices, in which the rows represent the current state (i.e. the subset \( S_{t...,t+x} \) of the sequence \( S \) observed at time \( t \)) and the columns represent the state to transition to. The memory occupied by such a structure is \( O(n|\Sigma|^2) \) where \( n \) is the chosen Markov order.

Testing

The code was compiled and tested with gcc (version 4.8.2), using the flags \(-g\), \(-Wall\), \(-Wextra\), \(-pedantic\) and \(-fopenmp\). For performance purposes the code available to public is compiled using the flags \(-O3\) and \(-fopenmp\).

The code was tested for memory leaks running Valgrind (version 3.1.0) with the flags \(-leak-check=full\) and \(-show-reachable=yes\). The Arabidopsis thaliana dataset was used to run these tests.

3.2 Pinetree front-end implementation

3.2.1 Web-interface

The server front-end integrates a simple user-friendly interface, where the user can make requests for retrieving targets from submitted datasets. The user can also obtain relevant information about the pipeline, using the interface (see Figure 3.6) to download the source code of Pinetree, along with a comprehensive user guide with the available parameters and some examples for proper execution of the program (see Appendix B).

The server is currently in its experimental version, lacking advanced functionality, such as a complete parametrization of the program. The server is implemented in PHP, simplifying the
3.2 Pinetree front-end implementation

Figure 3.6: The front-page for the Pinetree website

Integration with the web pages. The database where the requests are kept is implemented in MySQL, and the communication between the server and the database is done using PHP Data Objects (PDO).

Figure 3.7 provides a complete picture of the process behind the interface.

The interface accepts as input the miRNA and target sequences, submitted by the user. The user is able to submit its sequences in a file with FASTA format, or in a text box, requiring that the inputted sequences follow the same FASTA format. The back-end pipeline is executed with its default parameters, that were based in our literature analysis (see Section 3.1) and in our
3. The Pinetree pipeline

benchmark tests.
Upon a successful submission, the end user will be presented with a temporary, random-generated URL, where it can trace the analysis progress and ultimately, obtain the result file. The page can be bookmarked by the user, and visited later on for the retrieval of the result. All results are kept for 24h in our server. The result file can be downloaded in a CSV format, which facilitates the computational analysis a posteriori.

3.2.2 File output

The output of Pinetree is stored in a text file in either a tabular or human readable format, and presents all predicted pairs with a complementarity score not below the complementarity threshold and with an accessibility score not above the accessibility threshold. The same applies to the anti-targets module, which outputs the miRNA:mRNA interactions with a miRNA:mRNA affinity score not above the affinity threshold.

For the complementarity module, Pinetree will additionally output the miRNA and binding site sequences, the corresponding alignment and its start position in the target. Figures 3.8 and 3.9 depict the output for the two execution modes of Pinetree.

![Figure 3.8: The several parameters and results for the target prediction mode, along with an output example](image)

The several parameters are presented at the beginning of the file, along with the date The parameters for the score calculation (seed position) and for translational inhibition events (central region position) can be modified through a configuration file read by Pinetree. They can also be
3.2 Pinetree front-end implementation

The several parameters and results for the anti-target search mode, along with an output example:

```bash
# Start time of execution: Sun Apr 26 19:24:14 2015
# End time of execution: Sun Apr 26 19:24:14 2015
# Number of processors: 1
# Transcript file used: ../datasets/athaliana/2.fa
# miRNA file used: ../datasets/athaliana/1.fa
# Annotation file used: ../datasets/athaliana/Athaliana_167_defline.txt
# miRNA:mRNA affinity threshold: 7.6
# Markov order: 3
# Number of errors allowed: 0
# Human readable output: Yes

target id: AT2G28100.1
  target info: UDP-Glycosyltransferase superfamily protein
  miRNA id: ath-mirR5021
  miRNA:mRNA affinity: 1.185649

#

target id: AT3G46730.1
  target info: open reading frame 204
  miRNA id: ath-mirR5021
  miRNA:mRNA affinity: -1.352624
```

Figure 3.9: The several parameters and results for the anti-target search mode, along with an output example

modified using command-line options, which will override the values in the configuration file. If a configuration file is not provided, then Pinetree will run with default values.

When provided with a transcript annotation file, Pinetree will also include this information in the output.
Contents

4.1 Dataset description . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31
4.2 Performance . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31
4.3 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

Results and discussion
4.1 Dataset description

We gathered miRNA and transcript datasets and compiled a set of validated targets for two well-studied species: *Arabidopsis thaliana* and *Oryza sativa*. The miRNAs were obtained from miRBase (release 21) ([Kozomara and Griffiths-Jones, 2010](#)) and the transcript data was obtained from Phytozome (version 10) ([Lamesch et al., 2012](#) [Ouyang et al., 2007](#)).

For *Arabidopsis thaliana*, we obtained the experimented validated targets from literature, but also from a manually curated database, TarBase (version 6.0). The validated targets for *Oryza sativa* were also obtained from TarBase.

Some of the reported targets on TarBase had interactions containing miRNAs that were not present in the current version of miRBase, thus being removed from our analysis. Table 4.1 shows more detailed information about the datasets, including the references for the validated targets obtained.

<table>
<thead>
<tr>
<th>Specie</th>
<th>Transcripts</th>
<th>miRNAs</th>
<th>Validated</th>
<th>Refs.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>35,386</td>
<td>337</td>
<td>167</td>
<td><a href="#">Addo-Quaye et al., 2008</a> <a href="#">Allen et al., 2005</a> <a href="#">Alves-Junior et al., 2009</a> <a href="#">Dai and Zhao, 2011</a> <a href="#">German et al., 2008</a> <a href="#">Vergoulis et al., 2012</a></td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>49,061</td>
<td>273</td>
<td>224</td>
<td><a href="#">Vergoulis et al., 2012</a></td>
</tr>
</tbody>
</table>

Table 4.1: Dataset description for the studied plant species

We also gathered the produced targets of two different target prediction tools, for both species: TAPIR ([Bonnet et al., 2010](#)) and psRNATarget ([Dai and Zhao, 2011](#)). The results obtained from these programs contain information that is not relevant for the execution of PINETREE, requiring a pre-processing step before using these results for comparison.

Since the miRNA datasets, as well as the validated datasets also contain information that is not relevant to the prediction of targets by PINETREE, we also pre-process these datasets.

Additionally, considering the validated targets come from different sources, this information is represented in different ways. Pre-processing this data before using it allows us to establish the standard for the used datasets. Scripts [A.1](#) and [A.2](#) present in Appendix A show how this data is pre-processed.

4.2 Performance

In this section we present the result of using PINETREE to predict miRNA targets in *A. thaliana* and *O. sativa*. We compare the performance of our tool with two other widely used methods for plant miRNA target prediction – TAPIR ([Bonnet et al., 2010](#)) and psRNATarget ([Dai and Zhao, 2011](#)). Additionally, we summarize the results of using our tool to search for anti-targets.
4. Results and discussion

We use the standard definition for recall, precision, and the $F_1$ measure, having $\text{Recall} = \frac{TP}{P}$, $\text{Precision} = \frac{TP}{(TP + FP)}$, and $F_1 = 2\left(\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}\right)^{-1}$. The gold standard for our evaluation (positive set) is the set of validated targets described in section 4.1.

It is important to highlight that the validated set may not be representative of the universe of bona fide targets, since the experimental conditions in which these interactions are observed frequently do not match physiological conditions. In addition, most targets are validated by detecting the occurrence of cleavage events, which only take place when cleavage-dependent silencing is the effector mechanism. This introduces a bias against targets which are silenced through translational repression. As a result, the false positive rates will inevitably be overestimated.

The results for all tools using default parameters for the datasets of *A. thaliana* and *O. sativa* are summarized in Table 4.2. In addition, Venn diagrams showing the predictions of Pinetree without accessibility assessment, TAPIR, psRNATarget and the targets in the validated set are presented in Figure 4.1.

\begin{center}
\begin{tabular}{lcccc}
\hline
 Tool & Total & Recall & Precision & $F_1$ score & Time \\
\hline
 TAPIR / FASTA & 652 & 0.78 & 0.20 & 0.32 & 49m53s \\
 psRNATarget & 1434 & 0.77 & 0.09 & 0.16 & N/A \\
 Pinetree & 647 & 0.80 & 0.21 & 0.33 & 2m28s \\
 Pinetree w/ accessibility & 636 & 0.79 & 0.21 & 0.33 & 4m50s \\
\hline
\end{tabular}
\end{center}

\begin{center}
\begin{tabular}{lcccc}
\hline
 Tool & Total & Recall & Precision & $F_1$ score & Time \\
\hline
 TAPIR / FASTA & 1007 & 0.63 & 0.14 & 0.23 & 125m48s \\
 psRNATarget & 2159 & 0.50 & 0.05 & 0.09 & N/A \\
 Pinetree & 1011 & 0.62 & 0.14 & 0.22 & 6m44s \\
 Pinetree w/ accessibility & 956 & 0.62 & 0.14 & 0.23 & 12m15s \\
\hline
\end{tabular}
\end{center}

Table 4.2: Comparison of plant miRNA target prediction tools for a set of 167 validated targets of *A. thaliana* and 224 validated targets of *O. sativa*, showing the total number of predictions, and prediction performance in terms of precision, recall, $F_1$ measure and running time. Best performances are highlighted.

\begin{center}
\begin{figure}
\begin{minipage}[b]{0.45\textwidth}
\centering
\includegraphics[width=\textwidth]{A_thaliana}
\caption{Venn diagrams showing the comparison of target predictions for *A. thaliana* amongst the evaluated tools and the corresponding set of validated targets.}
\end{minipage} \hfill
\begin{minipage}[b]{0.45\textwidth}
\centering
\includegraphics[width=\textwidth]{O_sativa}
\caption{Venn diagrams showing the comparison of target predictions for *O. sativa* amongst the evaluated tools and the corresponding set of validated targets.}
\end{minipage}
\end{figure}
\end{center}
4.2 Performance

Using default parameters, Pinetree is able to predict 133 out of 167 validated targets for A. thaliana, with a precision and recall rates higher than other tools.

Results show that for the A. thaliana dataset, the accessibility analysis leads to worse results than standard execution – few candidates are eliminated but those that are include validated targets.

On the other hand, an accessibility analysis can further refine the set of putative targets on O. sativa, marginally outperforming the standard execution results. However, the decrease in predicted targets comes at the cost of extra computational time needed to perform such analysis. The time required to perform these calculations increases the running time approximately twofold compared to the standard execution.

When comparing these results to the other evaluated tools we conclude that psRNATarget has the worst results, while Pinetree has a slightly better performance than TAPIR for A.thaliana and comparable results for O. sativa.

It is important to highlight that Pinetree is much faster than the standalone version of TAPIR (a comparison to psRNATarget was not made, since no standalone version is currently available). All the tests were performed on an AMD Opteron(TM) Processor 6276 @ 2.3GHz, with Pinetree using 20 cores of the machine.

![Figure 4.2](image)

Figure 4.2: Plot showing the distribution of affinities in A. thaliana of each transcript against all miRNAs (on the left), and of each miRNA against all transcripts (on the right).

The results for the anti-target search were obtained using Markov chains of order 3. In Figure 4.2 (on the left panel) we observe that a few transcripts have a noticeable avoidance pattern against a number of miRNAs, illustrated by the spikes on the plot elongated towards the left.

A different perspective is given by the panel on the right. We can see that some miRNAs have very similar affinity profiles across the transcriptome. This is due to the fact that they belong
to the same miRNA family and have thus similar sequences. Interestingly, many miRNAs with a greater number of transcripts showing avoidance profiles (spikes on the left), typically exhibit a number of transcripts with greater propensity of harbouring target sites (spikes on the right). With hindsight, this is what might be expected of a master switch: targeting a number of transcripts for silencing, while other messages involved in the same biological process are left intact. Sequences might have evolved so that the former group is much more likely to conserve multiple potential target sites, while the latter is actively avoiding targeting.

4.3 Discussion

The identification of miRNA targets is a major challenge, but it is a fundamental step towards their functional annotation. It stands in the way of a more complete understanding of gene regulatory networks, since miRNAs are known to regulate the expression of many genes, hence influencing the dynamics of several biological pathways.

MiRNA-mediated gene regulation can be operated by different mechanisms, which will either promote target destabilization or translational repression. The diversity of effector mechanisms and the different impacts they have on target mRNA abundance presents challenges both at computational and experimental levels.

Experimentally, it is more difficult to find evidence of regulation via translational repression because it can only be directly observed by measuring protein levels. This influences the design of computational methods insofar as it introduces a bias on the number of examples to generalize from when trying to characterize the rules governing functional miRNA:mRNA interactions. Examples which, in any case, are still very few, even for model plant species, and being produced by a variety of experimental protocols that may or may not provide us with an experimentally-verified binding site for the miRNA. The same can be said of a lack of experimental evidence about non-functional would-be miRNA:mRNA interactions – possibly suggested by computational methods but experimentally invalidated.

Bearing these issues in mind, it is important to develop methods that can accommodate biases while trying to mitigate their effect on the predictions. One important aspect to consider is that, since experimentally-confirmed interactions are currently insufficient, and experimentally-confirmed non-interactions are all but nonexistent, conclusions drawn by their analysis are necessarily provisional. Our best current knowledge has provided us with default parameters for all steps of Pinetree, but experimentalists must be able to tweak the parameters and make sense of the results for their objects of study. In light of this need, we made an effort to make parameters have straightforward effects on the predictions and produce an output which is easy to interpret.

MiRNA:target complementarity is considered – and our results again confirm it – a key factor for predicting functional interactions. However, it fails to account for all validated targets.
Alignment-based complementarity assessment can be seen as a simplified energy model for miRNA:mRNA hybridization, so we used RNAup (Mückstein et al., 2006) as a tool for complementarity assessment to see whether a more accurate energy model could improve on the results. We found that if we relax parameters sufficiently, we can recover all validated targets at the expense of a substantial increase of the false positive rate and a noticeable degradation of the running time (data not shown).

Recent studies (Hausser et al., 2009) suggest that target accessibility is also an important factor for target prediction, but our results seem to indicate otherwise. It is possible that the accessibility parameters used (e.g. folding temperature, length of the assessed region) are not well-adjusted or may not be uniformly suitable for every putative interaction. We tested both RNAplfold (Bernhart et al., 2006) and RNAhybrid (Mückstein et al., 2006) from the Vienna Package (Lorenz et al., 2011) as surrogate tools to assess accessibility, but could not find any score threshold that significantly improved upon the predictions obtained through complementarity assessment alone, especially for A. thaliana. We thus made the accessibility module optional.

In addition, we sought to compare the complementarity and accessibility scores against corresponding empirical distributions obtained from a set of randomized sequences exhibiting the same dinucleotide frequencies of the original transcripts, hence obtaining $p$-values for each miRNA/mRNA pair. We failed, however, to show that validated targets have statistically significant scores (data not shown).

PINETREE also provides a means to search for anti-targets – transcripts whose sequence show signs of having evolved in such a way that they are depleted from potential target sites of particular, presumably co-expressed, miRNAs. The detection of anti-targets offers a new perspective on miRNA functional annotation, and one which is not limited nor biased by current target prediction techniques, and which may help unravel the regulatory role of miRNAs.

Our results show varying patterns of affinity between miRNAs and each transcript, illustrating that different mRNAs have uneven propensities to harbouring targets for particular miRNAs. When the affinity is strongly negative, we deem the pair an anti-target. There are cases, however, of validated targets with weakly negative affinities. This can be interpreted as a bias against multiple targets for the same miRNA. On the off side, there is no clear way to experimentally determine the validity of these anti-targets, which may explain the disinvestment in developing tools to specifically search for these miRNA/transcript pairs.
Conclusions

Contents

5.1 Conclusion ................................................. 37
5.2 Future work ................................................. 37
5.1 Conclusion

Predicting plant miRNA targets is not as easy a task as was initially thought, showing much of the complexities that were recognized early on for the same endeavor in animals. The determinants for an effective regulation are not completely understood, and thus computational methods are only left with a limited set of constraints to apply, struggling to enhance sensitivity at the expense of risking an excess of false positive predictions.

Hoping to improve upon previous tools we developed PINETREE, and indeed we obtain slightly more accurate results, providing the community with a piece of software that is faster and more flexible than pre-existing programs.

In addition to predicting targets, it can also be used to identify transcripts which may be actively avoiding miRNA-mediated regulation. An insight which is useful in unraveling the regulatory interplay between miRNAs and mRNAs.

PINETREE results from chaining different modules in a processing pipeline. It is simple to customize and can easily be extended with new modules. Parameters such as the size of the seed region, location of the central region, or the order of the Markov chain in the anti-target module, are meaningful and can be modified by the users to better serve their needs.

5.2 Future work

The next step in the development of the proposed tool is to further develop the web interface, one that end users such as biologists can use without requiring a standalone installation or dealing with a command line.

The skeleton for such a web interface can be found at https://aleph.inesc-id.pt/~camp/pinetree, where the source code can be downloaded, and the target-prediction function can be used.

An automated service for anti-targets is also to be constructed, supporting the upload of user datasets for the purpose, and notification after these datasets being processed.

Even though PINETREE shows good recall results for A.thaliana, the results are not as good for O. sativa, and its adaptation to other plant species is yet to be proven. Some other known species such as Vitis vinifera and Medicago truncantula should be evaluated in order to understand and adjust the performance of the pipeline for other plant species.
Bibliography


Eric Bonnet, Jan Wuyts, Pierre Rouzé, and Yves Van de Peer. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. 20(17), 2004.

Eric Bonnet, Ying He, Kenny Billiau, and Yves Van de Peer. TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. Bioinformatics, 26(12):1566–1568, 2010.


Nancy A Eckardt. Investigating translational repression by microRNAs in *Arabidopsis*.


Bibliography


Bibliography


Manuel Reis and Nuno D Mendes. Pinetree: a plant miRNA target explorer. in review 2015.


Manuel Reis, Nuno D Mendes, and Ana T Freitas. Computational prediction of microRNA targets in plant genomes. XII Symposium on Bioinformatics, September 2014b.


Appendix A - Approach
supplementary material
<table>
<thead>
<tr>
<th>Name</th>
<th>Comp¹</th>
<th>Cons²</th>
<th>Acc³</th>
<th>Mul⁴</th>
<th>Func⁵</th>
<th>Code</th>
<th>Link</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patscan</td>
<td>✓</td>
<td></td>
<td>N/A</td>
<td></td>
<td></td>
<td>N/A</td>
<td><a href="http://wmd3.weigelworld.org/">http://wmd3.weigelworld.org/</a></td>
<td>Dsouza et al. (1997)</td>
</tr>
<tr>
<td>miRNAssist</td>
<td>✓</td>
<td>✓</td>
<td>N/A</td>
<td></td>
<td></td>
<td>N/A</td>
<td><a href="http://bioinformatics.psb.ugent.be/webtools/tapir/">http://bioinformatics.psb.ugent.be/webtools/tapir/</a></td>
<td>Xie et al. (2007)</td>
</tr>
<tr>
<td>miRU</td>
<td>✓</td>
<td>✓</td>
<td>N/A</td>
<td></td>
<td></td>
<td>N/A</td>
<td><a href="http://rna-tools.cmp.uea.ac.uk/plant/">http://rna-tools.cmp.uea.ac.uk/plant/</a></td>
<td>Zhang (2005)</td>
</tr>
<tr>
<td>WMD3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>N/A</td>
<td>Bonnet et al. (2010)</td>
<td>Ossowski et al. (2008)</td>
</tr>
<tr>
<td>TAPIR</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>N/A</td>
<td>Xie and Zhang (2010)</td>
<td>Bonnet et al. (2010)</td>
</tr>
<tr>
<td>UEA sRNA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>N/A</td>
<td>Xie and Zhang (2010)</td>
<td>Moxon et al. (2008)</td>
</tr>
<tr>
<td>Target-align</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>N/A</td>
<td>Fahlgren et al. (2007)</td>
<td>Jha and Shankar (2011)</td>
</tr>
<tr>
<td>Targetfinder</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>N/A</td>
<td>Dai and Zhao (2011)</td>
<td>Jha and Shankar (2011)</td>
</tr>
<tr>
<td>p-TAREF</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>N/A</td>
<td>p-TAREF/form1.html</td>
<td>Jha and Shankar (2011)</td>
</tr>
<tr>
<td>psRNATarget</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>N/A</td>
<td>psRNATarget/</td>
<td>Dai and Zhao (2011)</td>
</tr>
<tr>
<td>imiRTP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>N/A</td>
<td><a href="http://admis.fudan.edu.cn/projects/imiRTP.htm">http://admis.fudan.edu.cn/projects/imiRTP.htm</a></td>
<td>Ding et al. (2011a)</td>
</tr>
</tbody>
</table>

Note: ¹ Complementarity; ² Conservation; ³ Accessibility; ⁴ Multiplicity; ⁵ Functionality; ⁶ Reference; ⁷ Currently offline.
#!/bin/sh

cd "$(dirname "$0")"

TARBASE=validated_targets_tarbase.csv

#generate a processed miRNA file
sed ' s / .∗/g' mirna_unprocessed.fa > mirna.fa

sed -re '/ath-///d' ../tarbase_data.csv > temp

cut -f2,4 temp | awk '{ $2=toupper ( $2 ) }1' | sed ' s / /,/g' | sed ' s/mir/miR/g' | awk -F, '{ print $2,$1 }' OFS=, > $TARBASE

rm temp
rm $TARBASE

Figure A.1: Script for preprocessing miRNA and validated targets datasets

#!/bin/sh

cd "$(dirname "$0")"

#processing psRNATarget targets
sed */#/d" psRNATarget/unprocessed_targets.txt | cut -f1,2 | tail -n +2 | sed 's/\t/ ,/g' | awk -F, '{ print $2,$1 }' OFS=, | sort | uniq > psRNATarget/processed_targets.csv

#processing TAPIR targets
sed -e "/#/d" tapir/unprocessed_targets.txt -re '/miRNA /target /!d' -e 's/([^[:space:]]+)/\t/g' | cut -f2 | paste - - -d" | tr " " ',' | awk -F, '{ print $2,$1 }' OFS=, | sort | uniq > tapir/processed_targets.csv

rm temp | uniq > validated_targets.csv

cat psRNATarget/processed_targets.csv tapir/processed_targets.csv validated_targets.csv | sort | uniq > subset.csv

Figure A.2: Script for pre-processing the results from tools
Appendix B - PINETREE documentation
Introduction

This documentation describes version 1.0 of the Pinetree framework. Version 1.0 contains the programs for target-prediction and anti-target detection. This document is divided in 3 sections: (1) A brief overview of the programs in this version of our software; (2) A guide to use the programs, including the available parameters for modification; (3) A guide to install the programs.

1 A brief overview of Pinetree usage

Version 1.0 of Pinetree is divided in two modes of execution: (1) a target prediction mode, where the user can find the miRNA targets for a given plant specie; (2) an anti-target search mode, where the user can search for targets that seem to avoid the action of specific miRNAs.

2 Using Pinetree programs

2.1 Overview

Both programs in the Pinetree package use similar command line options and arguments, differing in the commands for specific functionality. The simplest
command line arguments just include a miRNA file and a transcript file for both execution modes. The arguments can be provided in any order, as long as they follow the corresponding flag. This way, the command:

```
pinetree -m miRNA.fa -t transcript.fa
```

will try to predict targets for the miRNAs in `miRNA.fa` in the transcripts in `transcript.fa`, using the default parameters given in the `constants.h`.

The simplest command of the Pinetree programs expects a target and miRNA file, so if you type "pinetree", you will see a short help message:

```
%%%%%%%%%%%%%%%%%%%%%%%%
PINETREE: Version 1.0
%%%%%%%%%%%%%%%%%%%%%%%%

USAGE:
./pinetree -t [transcript file] -m [mirna file] [-options]

Common options:
-A, --annotation [annotation file]
  provides information about the transcripts functions
-b, --accessibility
  turns on accessibility mode
-h, --help
  detailed information about the command-line options
-n, --no_normalization
  turns off normalization mode
-o, --output [file name]
  name of the file where the output os stored (e.g -o pinetree)
-p --processors
  number of processors
-P, --pretty
  enable human readable output
-V, --verbose
  output extended information about the program execution in the header

"pinetree -h" provides a complete listing of the options available for the program and their default values.

Our package includes several test files. To check if everything is working correctly you can run:

```
pinetree -m test/mirna.fa -t test/transcript.fa
```
2.2 File formats

2.2.1 Sequence files

The pinetree programs can read miRNA and transcript files that are in the standard FASTA format. FASTA format files consist of a description line, beginning with a ”>” symbol, followed with one or more lines containing the sequence itself:

```
>ATCG00065.1 \sequence name
ATGCCAACATTTAAACAACCTTTAGAAAAATACAA \first line of the first sequence
CCCTCAAGGACAGGAACATGTACTCGGGGTAT \second line
>ATCG00070.1
ATGCTTAATATATTTAATTTGATCTGTATTTTT \second sequence
```

Several sample test files are included in the Pinetree package that can be found in FASTA files processed by Pinetree should then contain only sequence descriptions, sequences, and end-of-line characters.

2.2.2 Annotation files

The user can provide annotation files to the program (using flag --annotation). These files give additional information about the transcripts included in the transcript dataset, and therefore helping to understand the function of the transcripts involved in the predicted interactions.

The format recognized by Pinetree consists of the sequence name of the transcript, followed by a tab separator and the description of the transcript function:

```
AT1G01010.1 NAC domain containing protein 1
AT1G01020.1 Arv1-like protein
AT1G01020.2 Arv1-like protein
AT1G01030.1 AP2/B3-like transcriptional factor family protein
```

2.3 Running the programs

Pinetree provides a variety of command line options that modify the thresholds (e.g. --c_cut 3.5, --evalue 100), as well as other algorithm and output parameters (e.g. --mirna mirna.fa, --pretty). A complete list of command line options for the different execution modes (see Table 1) is shown in this document.

The several arguments can be passed to the program in any order. Note that the parameters passed over the command line override all the default parameters. If you also want to override the parameters that may be passed in a CONFIG file (--config CONFIG) simply pass the CONFIG file as the first argument of the command line.
Table 1: Command line options for the execution modes of Pinetree

<table>
<thead>
<tr>
<th>Flag</th>
<th>Mode</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-a, --a_cut [value]</td>
<td>1</td>
<td>cutoff value for the accessibility criterion</td>
</tr>
<tr>
<td>-A, --annotation [file]</td>
<td>1,2</td>
<td>path to a file that contains information about the transcripts function. Section 2.2.2 describes the file format in more detail.</td>
</tr>
<tr>
<td>-b, --accessibility</td>
<td>1</td>
<td>turns on accessibility mode. It will also output an accessibility score</td>
</tr>
<tr>
<td>-c, --c_cut [value]</td>
<td>1</td>
<td>cutoff value for the complementarity criterion</td>
</tr>
<tr>
<td>-C, --config [file]</td>
<td>1,2</td>
<td>path to a CONFIG file, that contains the values for the module parameters</td>
</tr>
<tr>
<td>-e, --evalue [value]</td>
<td>1</td>
<td>cutoff value for the e-value parameter of fasta36</td>
</tr>
<tr>
<td>-e, --e_cut [value]</td>
<td>2</td>
<td>cutoff value for the miRNA:mRNA affinity</td>
</tr>
<tr>
<td>-E, --errors[value]</td>
<td>2</td>
<td>sets the number of errors allowed to give when calculating the escore</td>
</tr>
<tr>
<td>-h, --help</td>
<td>1,2</td>
<td>detailed information about the command-line options</td>
</tr>
<tr>
<td>-m, --mirna [file]</td>
<td>1,2</td>
<td>path to a file that contains the miRNA sequences</td>
</tr>
<tr>
<td>-n, --no_normalization</td>
<td>1</td>
<td>turns off normalization mode. Instead of outputting the stand-</td>
</tr>
<tr>
<td>-M, --markov [value]</td>
<td>2</td>
<td>sets the Markov order used in the statistical models</td>
</tr>
<tr>
<td>-o, --output [name]</td>
<td>1,2</td>
<td>name of the file where the output will be stored (e.g. if the name</td>
</tr>
<tr>
<td>-p, --processors [value]</td>
<td>1,2</td>
<td>sets the number of processors for the program</td>
</tr>
<tr>
<td>-P, --pretty</td>
<td>1,2</td>
<td>enables human readable output. The name of the outputted file will have a .txt extension</td>
</tr>
<tr>
<td>-t, --target [file]</td>
<td>1,2</td>
<td>path to a file that contains the transcript sequences</td>
</tr>
<tr>
<td>-v, --version</td>
<td>1,2</td>
<td>prints the current version of the program</td>
</tr>
<tr>
<td>-V, --verbose</td>
<td>1,2</td>
<td>prints extended information about the program execution in the header</td>
</tr>
<tr>
<td>-x, --central_start [value]</td>
<td>1</td>
<td>sets the start position for the central region</td>
</tr>
<tr>
<td>-X, --central_stop [value]</td>
<td>1</td>
<td>sets the end position for the central region</td>
</tr>
<tr>
<td>-z, --seed_start [value]</td>
<td>1</td>
<td>sets the start position for the seed region</td>
</tr>
<tr>
<td>-Z, --seed_stop [value]</td>
<td>1</td>
<td>sets the end position for the seed region</td>
</tr>
</tbody>
</table>

Note: ¹Pinetree; ²Pinetree_evo
2.4 Interpreting the results

PINETREE has two different types of output format: (1) a default output, that is easily parsed, enabling the computational analysis of the outputted data (Figure 1); (2) a "pretty" output, that makes it easier for humans to analyse the output (Figure 2).

![Figure 1: Output for target prediction execution mode](image1)

![Figure 2: Human readable output for target prediction execution mode](image2)

2.4.1 Headers

If the `--verbose` option is used, extended information about the program execution is printed. The headers in the file output are started by the `#` symbol,
followed by a line that contains relevant information about the execution that
gave origin to the file. The first two lines pertain the duration of the program
execution, giving the start and end time of such execution. These lines can also
help to identify and distinguish different executions of the program.

The next line indicates the number of used processors. Again, this informa-
tion can be used for reproducibility purposes, or simply to adjust the number
of processors used later on. Usually, 20 is the maximum number of processors
recommended, as a bigger number won’t speed up the program as it reaches the
limits of parallelization of the fasta36 module.

The following lines contain information relative to the used datasets, namely
the file paths of such files. If an annotation file is not provided, the matching line
will not be outputted.

The next lines represent the parameters used for each module, for repro-
ducibility purposes. The last line indicates the which of several modes available
were used (accessibility, normalization, pretty output). If the --pretty flag
was not on, a line with the name of each column is also outputted, separated
by commas.

2.4.2 Target prediction output

This section clarifies the information present in the output file for the target
prediction program. If the program is not executed in --pretty mode, each
line will contain the name of the miRNA and the transcript for the predicted
interaction.

If the annotation file is provided, a description about the function of the
transcript is also present. If the annotation file is provided, but there is not a
description about the transcript in that file, the description will be filled with
the string "N/A". By default, all calculated scores are normalized afterwards.
The user can turn off the normalization mode using the flag -n.

That line will also contain the complementarity score, if the score of such
interaction is below the chosen threshold, based on the formula:

\[ S = S_{\text{mis}} + \frac{S_{\text{wob}}}{2} + S_{\text{gap}} + 2 \left( S^{*}_{\text{mis}} + \frac{S^{*}_{\text{wob}}}{2} + S^{*}_{\text{gap}} \right) \]  

where \( S_{\text{mis}} \), \( S_{\text{wob}} \), and \( S_{\text{gap}} \) represent the number of mismatches, G:U wob-
bles and gaps outside the seed region, respectively, and \( S^{*}_{\text{mis}} \), \( S^{*}_{\text{wob}} \), and \( S^{*}_{\text{gap}} \),
the same counts in the seed region.

The lower the score, the more complementary the predicted interaction is.
The same applies for the accessibility score if the --accessibility flag was on.
The lower the score, the more accessible the transcript is in the predicted
region.

Depending on the chosen central region – which is adjustable in the parame-
ters – the output will present information about if the transcript may be a target
for a cleavage, or translational inhibition by a given miRNA in the predicted
interaction.
In the same line, there is also graphical information about the alignment between the miRNA and its target. A match between two characters is represented with the symbol ‘|’, a G:U wobble is represented with ‘o’, a gap is represented with ‘-‘ and a mismatch is represented with ‘x’. If the program is executed in \texttt{--pretty} mode, than additional information about the miRNA and target sequences is displayed, stating the alignment between the two sequences.

\subsection{2.4.3 Anti-target detection output}

This section explains the output of the execution of the anti-target program. Again, as the target prediction program, each line will contain the name of the miRNA and the transcript for the predicted interaction.

If the annotation file is provided, a description about the function of the transcript is also present.

For each miRNA, we propose to determine how likely is the occurrence of a perfect target site (reverse-complementary sequence) in a given transcript compared to the same measure taken in the set of all transcripts.

Our assessment is based on the notion of \textit{affinity} between a miRNA, \( \mu \) and a transcript, \( t \in T \) (\( T \) is the set of all transcripts), given a sequence model for the transcript \( (M_t) \) and the transcriptome \( (M_T) \). The affinity, \( \alpha \), also present in the output line, is expressed as a log-odds score:

\[
\alpha(\mu, t) = \log P(\bar{\mu} \mid M_t) - \log P(\bar{\mu} \mid M_T) \quad (2)
\]

\( P(\bar{\mu} \mid M) \) is the probability that the sequence model \( M \) generates the reverse-complementary sequence of \( \mu \).

Here, sequence models correspond to Markov chains of order \( n \), which are learnt from the corresponding sequences. To account for the zero-frequency problem, pseudo-counts are introduced for unobserved transitions, \( i \rightarrow j \), by adding 1 to the frequency of all transitions exiting state \( i \).

Let \( S = S_{1..k} \) be a sequence of length \( k \), \( M \) a Markov chain of order \( n \) (with \( 0 < n \leq k \)), \( \pi_0(i) \) the initial probability of state \( i \), and \( \pi(i \rightarrow j) \) the transition probability from \( i \) to \( j \) then, \( P(S \mid M) \) is given by:

\[
P(S \mid M) = \pi_0(S_{1..n}) \prod_{i=1}^{k-n} \pi(S_{i..i+n-1} \rightarrow S_{i+1..i+n}) \quad (3)
\]

if the score of such interaction is below the chosen threshold. The lower the score, the lower the affinity between the miRNA and the target in the predicted interaction. The top-scoring anti-targets correspond to miRNA:mRNA pairs with the lowest affinity.

\section{3 Installing Pinetree}

The Pinetree install package includes the source files for the two programs, a folder with test files, this document, a README file, a CONFIG file that contains all the modifiable parameters and a script to assist the installing process.
This script will download the latest version of the FASTA program (version 36.3.7a), and generate a `fasta36` executable. This executable is required to assess the complementarity between the set of miRNAs and the set of targets.

This script will also download the latest version of the `ViennaPackage` (version 2.1.8) and install it locally. This package is necessary to calculate the target site accessibility.

After this step, it will also install the Pinetree programs – `pinetree` and `pinetree_evo`.

If you wish to tweak the code feel free to do so. The Pinetree Makefile compile time options are `-O3` and `-fopenmp`, but you can run the command `make all_debug` in order to generate debuggable programs. The programs were tested against a Linux/Unix environment.

If you encounter any bugs, or you wish to see new features/extended functionalities, please contact:

KDBIO/INESC-ID
mreis@kdbio.inesc-id.pt