# Prediction of Plant MicroRNAs

## Dissertation

der Fakultät für Informations- und Kognitionswissenschaften
der Eberhard-Karls-Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
**Dipl.-Inform. Tobias Dezulian**
aus Böblingen

**Tübingen**
**2006**

| | |
|---|---|
| Tag der mündlichen Qualifikation: | 20.12.2006 |
| Dekan: | Prof. Dr. Michael Diehl |
| 1. Berichterstatter: | Prof. Dr. Daniel H. Huson |
| 2. Berichterstatter: | Prof. Dr. Detlef Weigel<br>(Max-Planck-Institut<br>für Entwicklungsbiologie) |

## Erklärung

Hiermit erkläre ich, daß ich diese Schrift selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und daß alle Stellen, die im Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben der Quellen kenntlich gemacht sind. Eine detaillierte Abgrenzung meiner eigenen Leistungen von den Beiträgen meiner Kooperationspartner und von Implementierungsleistungen, die im Rahmen von mir betreuter Studien- und Diplomarbeiten erbracht worden sind, habe ich explizit in Anhang B vorgenommen.

Tübingen, Oktober 2006 *Tobias Dezulian*

# Zusammenfassung

In allen Lebewesen wird der in Genen kodierte Bauplan von der Zellmaschinerie mit Hilfe von RNA in Proteine übersetzt. Erst Anfang dieses Jahrtausends wurde entdeckt, dass bei diesem Prozess in höheren Lebewesen wie Tieren und Pflanzen einer bislang unbekannten Klasse von RNA Molekülen eine entscheidende regulatorische Rolle zukommt.

Gegenstand dieser Arbeit sind die sogenannten microRNA Gene, welche keine Proteine kodieren, sondern bereits auf RNA Ebene ihre Wirkung auf andere Transkripte entfalten. In diesem Rahmen wurden mehrere bioinformatische Werkzeuge konzipiert. Eines davon erlaubt es, neue microRNA Gene durch den Vergleich zweier Pflanzengenome aufzuspüren. Dies wurde auf die Genome von *Arabidopsis thaliana* und Pappel, sowie von Reis und Hirse angewandt. Die resultierenden Gen-Kandidaten wurden dann in enger Zusammenarbeit von molekularbiologischen Kooperationspartnern in vitro und in vivo näher untersucht. Eine weitere, universell einsetzbare Software ermöglicht dem Benutzer die Visualisierung und interaktive Erforschung zweier beliebiger RNA Sequenzmengen und ist insbesondere nützlich bei der Anwendung auf microRNAs. Die Idee für eine andere Anwendung entstand durch die Verfügbarkeit der Ergebnisse einer neuen Sequenziertechnologie für die Signaturen kleiner RNAs. Diese dienten als Informationsquelle für eine Methode, um auf deren Basis neue microRNAs in *Arabidopsis* vorherzusagen, die dann auch erfolgreich von Kooperationspartnern biologisch verifiziert werden konnten. Ferner wurde ein simples Programm entwickelt, welches ausgehend von einer bekannten microRNA dessen Homologe in nahezu beliebigen Sequenzdaten mit hoher Sensitivität und Spezifizität detektieren kann. Dieses wurde genutzt, um ausgehend von den 286 beim Sanger Institut registrierten pflanzlichen microRNAs weitere 200 Homologe in verschiedenen Pflanzen zu identifizieren. Diese vergrösserte Datenbasis diente schliesslich als Ausgangsbasis für eine Überblicksarbeit über Konservierung und Divergenz von microRNAs in Pflanzen.

# Abstract

The blueprint stored in the genes of each living creature is translated into proteins via RNA. But it was not until the beginning of this millennium that a class of small RNAs was discovered that performs crucial regulatory roles during this process in higher organisms such as animals and plants.

The focus of this thesis are the so–called "microRNAs"—gene products which do not code for proteins but instead have a regulatory impact on other transcripts at the RNA level. In this context, several computational biology tools were designed and implemented. One of them permits the identification of new microRNA genes by comparing two plant genomes. It was applied to the genomes of *Arabidopsis* and poplar, as well as to rice and sorghum. The resulting gene candidates were then analyzed experimentally by a collaboration of molecular biologists. Another, universally applicable software allows the user to visualize and interactively explore two arbitrary RNA sequence sets and is especially useful in the context of microRNAs. The idea for a further application was spawned by the availability of the small RNA signatures resulting from a new sequencing technology. These signatures served as the basis on which we could predict new microRNAs—some of which could afterwards be validated by experimental collaboration. In addition, we developed a simple program that can—given a known microRNA—identify its homologs in almost arbitrary sequence data with high sensitivity and specificity. When this tool was supplied with the 286 microRNAs registered at the Sanger institute, an additional 200 homologs could be found across different plants. Eventually, this enlarged data set served as the basis for a survey article on conservation and divergence of microRNAs in plants.

# Acknowledgements

Tremendous thanks go to my advisor Daniel Huson who felt adventurous enough to give "the guy from the bank" (his words) the chance to conduct this PhD in his department. I am very grateful for his professional advice, encouragement and support. He generously provided an ideal working environment in every respect. In addition, he introduced me to an amazing crowd of Kiwis and provided me with the opportunity to work with Mike Steel in the area of phylogenetic reconstruction—a fascinating field which I initially worked on during my PhD (but is not part of this thesis). In addition to thanking him for assisting my professional growth, I am indebted to Mike for more than those great times out in the New Zealand mountains.

Detlef Weigel deserves my deepest gratitude for his insightful ideas, numerous discussions and his co-supervision of this thesis. I feel honored to be part of this fruitful microRNA collaboration. In particular, I would also like to thank Javier Palatnik, Rebecca Schwab, Felipe Felippes, Heike Wollmann and Benjamin Czech, who work in Detlef's laboratory at the Max-Planck-Institute for Developmental Biology, for infecting me with an enthusiasm for plant biology, for giving me a glimpse into lab work and for productively sharing their ideas with me.

I thank current and former staff members of the department of "Algorithms in Bioinformatics", namely Alexander Auch, Olaf Delgado Friedrichs, Marine Gaudefroy-Bergmann, Stefan "Robbie" Henz, Tobias Klöpper, Christian Rausch and Daniel Richter for a fun, friendly and productive atmosphere. I am especially happy to have shared an office with Christian Rausch—an astounding source of good humor, encouragement and inspiration.

Furthermore, I appreciate countless scientific and non-scientific discussions over coffee with (among many others): Markus Gruber, Nina Lehmann, Stephan Steigele, Janko Dietzsch, Kay Nieselt, Jens Gramm, Markus Eiglsperger, Christian Klug, Michael Schröder, Martin Schaefer, Hannes Planatscher and Martin Siebenhaller.

Last, but certainly not least, I would like to thank the person who —unintentionally—sparked my interest in molecular biology and has been most important for me during these years: Stefanie Röhm.

In accordance with the standard scientific protocol, I will use the personal pronoun “we” to indicate the reader and the writer or (as explained in Appendix B) my scientific collaborators and myself.

# Contents

# Chapter 1

# Introduction

## 1.1 Background

Complex organisms such as plants and animals rely on complex regulatory circuitry for development, growth and adequate response to external conditions. Micro RNAs (microRNAs) are key players in these regulatory networks. They were—astoundingly—overlooked until less than five years ago, when it was discovered that a phenomenon that was taken as a peculiarity of the worm *Caenorhabditis elegans* is in fact part of a mechanism found in many multicellular organisms, including all plants and animals.

MicroRNAs are short RNAs that are excised from longer stem-loop shaped RNA transcripts that are endogenously expressed and do not code for proteins. They thus form a class of non-protein-coding (noncoding) genes which, on the whole, have recently received much attention. In a nutshell, microRNAs act by providing sequence-based target specificity to a ribonucleoprotein complex that homes in on messenger transcripts and other non-protein-coding transcripts for either destruction by cleavage or for translational repression. Therefore, microRNA genes are grouped in families of homologs characterized by producing (almost) identical microRNAs and thus targeting identical or closely related transcripts.

When the work for this thesis was started in the spring of 2004, a wave of tremendous interest in noncoding RNA research had just begun to surge. In contrast to a prevalent viewpoint that RNA was primarily either employed as messenger RNA (mRNA), as part of the ribosome, or otherwise directly involved in protein synthesis, recent large scale efforts aiming to analyze the transcriptional output (transcriptome) of plants [Yamada *et al.*, 2003; Stolc *et al.*, 2005; MacIntosh *et al.*, 2001] and animals [Cawley *et al.*, 2004; Bertone *et al.*, 2004; Glusman *et al.*, 2006] have led to the baffling conclusion that a very large fraction of the transcriptome is not associated with the ribosome and does not code for proteins. Moreover, for quite a few of these transcripts, it could be shown that they are not the result of "transcriptional noise", as

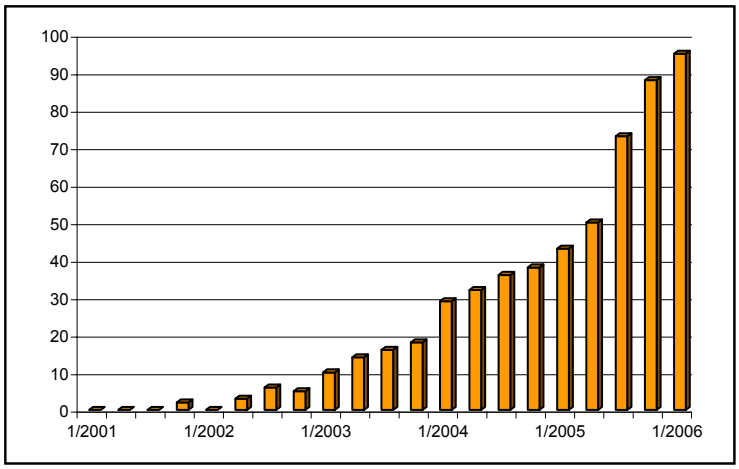



**Figure 1.1: MicroRNA articles at PubMed.** The graph depicts the number of articles having any of the terms '*microRNA*', '*micro RNA*' or '*miRNA*' in their title or abstract for each quarter between Jan 2001 and May 2006.

initially suspected, but that they convey crucial functions [Hüttenhofer & Vogel, 2006; Mattick, 2005] and thus may only be the tip of an iceberg of functional noncoding RNAs. Simultaneously, small RNAs such as microRNAs and short interfering RNAs (siRNAs) that had been discovered via a different route have been found to act as important players in the regulatory circuitry of plants and animals. Together, these intriguing discoveries have triggered the stunning suspicion that a hidden layer of RNA regulation of unknown extent may exist, and consequently led to an era of very intense research on noncoding RNA [Hüttenhofer *et al.*, 2005; Eddy, 2002] at the beginning of the $21^{st}$ century.

Now, a few years later, one has to acknowledge that noncoding RNAs (ncRNA) perform key regulatory functions that have so far been seen exclusively in the hands of proteins and their interactions; ncRNA have been shown to perform roles as diverse as developmental timing of protein expression, mRNA turnover and chromosome architecture, and may also regulate transcription and alternative splicing. The extent and effects of this hidden regulatory layer are only now beginning to emerge [Mattick, 2003].

The work of this thesis was started when microRNA research had already picked up a lot of momentum, and new insight regarding their biogenesis, genomics, modes of action, and regulatory roles was increasing rapidly, as shown in Figure 1.1.

In this fast-moving environment, we performed the computational biology work on which we report in this thesis, concentrating on the following five areas: first, as detailed in Chapter 2, we developed a tool named "microHARVESTER" for predicting microRNA homologs with great sensitivity

and specificity, and made it available for public use via a web interface. Second, using previously published microRNAs plus an additional large set of new homologs that we could contribute to the Sanger microRNA registry, we were able to analyze conservation and divergence of plant microRNAs on an unprecedented database, confirming previous results and reporting new relationships, as elaborated in Chapter 3. Thirdly, we conceived of and implemented a method to interactively explore and visualize relationships between (micro) RNAs, as discussed in Chapter 4. This tool, named CrossLink, is helpful in a wide variety of contexts, and we offer it in a web and in a standalone version. Fourth, we designed an approach for identifying new microRNA genes on the basis of comparative whole-genome predictions. Prospects for this seemed favorable, since a first draft assembly of the poplar genome was about to become available and the coding parts of the sorghum genome were just being assembled. This endeavor led to the implementation and results as outlined in Chapter 5. Finally, we identified new microRNAs in *Arabidopsis* by devising a method that exploits a new database of Massively Parallel Signature Sequencing (MPSS) that was published in 2005 [Lu *et al.*, 2005]. This project is described in Chapter 6. Note that these projects were pursued simultaneously to some extent. Ordering these chapters by the project kickoff date of the corresponding projects would yield the series: 5, 2, 3, 6, 4.

Much of the current insight into microRNAs was gained by the results of studies that were published while this thesis was in progress. In conjunction with the competition exerted by a large number of scientists who turned towards microRNA research at that time, one of the main challenges of this work was speed. Also, besides delivering results quickly, it was crucial to select research topics that were adequate in a temporal sense. Particularly, the work described in Chapters 3, 5 and 6 was crucially dependent upon resources that were just becoming available at the start of these projects.

## 1.2 Plant MicroRNAs

Each cell in a multicellular organism is—in general—equipped with an identical copy of the species' genome. Therefore, sophisticated orchestration of the gene expression in each cell is necessary to ensure that a cell-specific protein profile is expressed at an adequate rate and at specific times.

This regulation occurs at many different levels and includes a wide variety of processes that influence gene expression, ranging from chemical modifications of DNA, structural modifications of chromatin, activation and suppression by transcription factors, and processes that influence messenger RNA stability, to the regulation of the transcription and translation machinery.

Plant microRNAs regulate the expression of target genes posttranscriptionally, predominantly by guiding target transcripts to cleavage or—to a lesser extent—translational repression.

Note that microRNAs are chemically and functionally similar to small interfering RNAs (siRNAs) which mediate the related phenomena of RNA interference (RNAi), post-transcriptional gene silencing (PTGS) and transcriptional gene silencing (TGS) [Jones-Rhoades *et al.*, 2006]. Like microRNAs, siRNAs are processed by the Dicer RNaseIII family of enzymes, but unlike microRNAs, which derive from transcripts that fold into a stem-loop structure, siRNAs are processed from bimolecular duplexes or double stranded precursors with much longer stems [Jones-Rhoades *et al.*, 2006]. Both microRNAs and siRNAs are incorporated into silencing complexes that contain Argonaute proteins to which they confer the sequence-specificity that guides their repression of target genes. Together, microRNAs and siRNAs are the most prominent representatives of small RNAs.

MicroRNAs have been identified exclusively in plants, animals and their viruses. In contrast, RNAi, for which siRNAs are key components, seems to be a very basal mechanism that also operates in unicellular eukaryotes and possibly also in prokaryotes [Makarova *et al.*, 2006]. Since both classes of small RNA share much of their processing machinery, it can be speculated that siRNAs evolved earlier than microRNAs.

### Plant microRNAs vs. animal microRNAs

Plant microRNAs and animal microRNAs employ biogenesis and effector machinery components that have a common origin, such as the Dicer/-DICER-LIKE enzymes and the RISC complex (see below). There are, however, important differences that suggest that plant and animal microRNAs evolved independently in both clades.

Firstly, the biogenesis of animal and plant microRNA, although similar, proceeds differently (see below). Secondly, animal microRNAs are much more uniform in size and structure than plant microRNAs. While the stem-loop of the former is uniformly 60–70 nucleotides in length, the foldback of the latter can range anywhere from 50 to 450 nucleotides.

Thirdly, animal microRNAs seem to select their target much less specifically than their plant counterparts. While for the former, it seems sufficient that a so-called "seed" region comprising a few nucleotides at the 5' end of the microRNA binds with perfect complementarity to the target site and some of the remaining bases can compensate the imperfect binding of others, targeting in plants is much more specific [Mallory *et al.*, 2004] and presumably requires a higher degree of complementarity. This might explain why a single animal microRNA seems to be able to target hundreds of transcripts whereas a plant microRNA typically only targets a handful. Consequently, this looser specificity of animal microRNAs may exert a stronger force in

shaping the organism's transcriptome as mutational drift causes genes to escape or get caught under the control of particular microRNAs.

Fourth, the preferred mode of action of animal microRNAs is translational repression. Here, often several microRNAs bind to a single transcript and conjointly lead to repression—an interplay reminiscent of the way in which combinations of transcription factors cause activation or repression of transcription. In plants, a target transcript typically only displays one target site that is matched with near-perfect complementarity and predominantly leads to cleavage rather than translational repression—although counterexamples are known for both clades.

Fifth, the target sites of animal microRNAs are preferentially located within untranslated regions (UTRs) of target transcripts. In plants, target sites tend to be situated in the coding regions of the target transcripts. This may originate from different evolutionary scenarios for the microRNAs in the two clades.

Finally, there is no significant sequence similarity between any plant and animal microRNA that is indicative of a common origin.

Note that in plants, the functions of the Dicer enzyme, of which only one homolog is found in animals, seems to have been distributed across four homologs. Here, only one of them, DICER-LIKE1, is directly involved in microRNA biogenesis.

### Biogenesis

MicroRNA genes are transcribed by RNA polymerase II [Xie *et al.*, 2005], yielding a primary transcript (sometimes called pri-microRNA) that can be over one kb in length and may undergo canonical splicing, polyadenylation [Kurihara & Watanabe, 2004] and capping [Xie *et al.*, 2005]. The transcript region that contains the microRNA then acquires the typical stem-loop structure that is characteristic of microRNA precursors due to hydrogen bonds that form between the two arms of the precursor that make up the stem (confer Figure 1.2(a)).

MicroRNA transcripts are processed differently in animals and in plants. In animals, the mature microRNA is excised from the pri-microRNA in a stepwise manner by two different RNaseIII-type endonucleases: first, a nuclear-localized enzyme named Drosha cuts the microRNA stem-loop from the pri-microRNA. Then, after the stem-loop has been exported to the cytoplasm, an enzyme called Dicer makes the second set of cuts that separates the loop region from the relevant RNA duplex that consists of the microRNA and its opposite stem segment, which is referred to as the microRNA$\star$. Note that these last cuts are shifted by two nucleotides, resulting in $3'$ overhangs on either side of the duplex. In plants, the mature microRNA is also processed from the pri-microRNA in two steps, but here, a single enzyme, the Dicer homolog DICER-LIKE1 (DCL1), is responsible for both sets of cuts that liberate the microRNA/microRNA$\star$ duplex from the pri-microRNA.

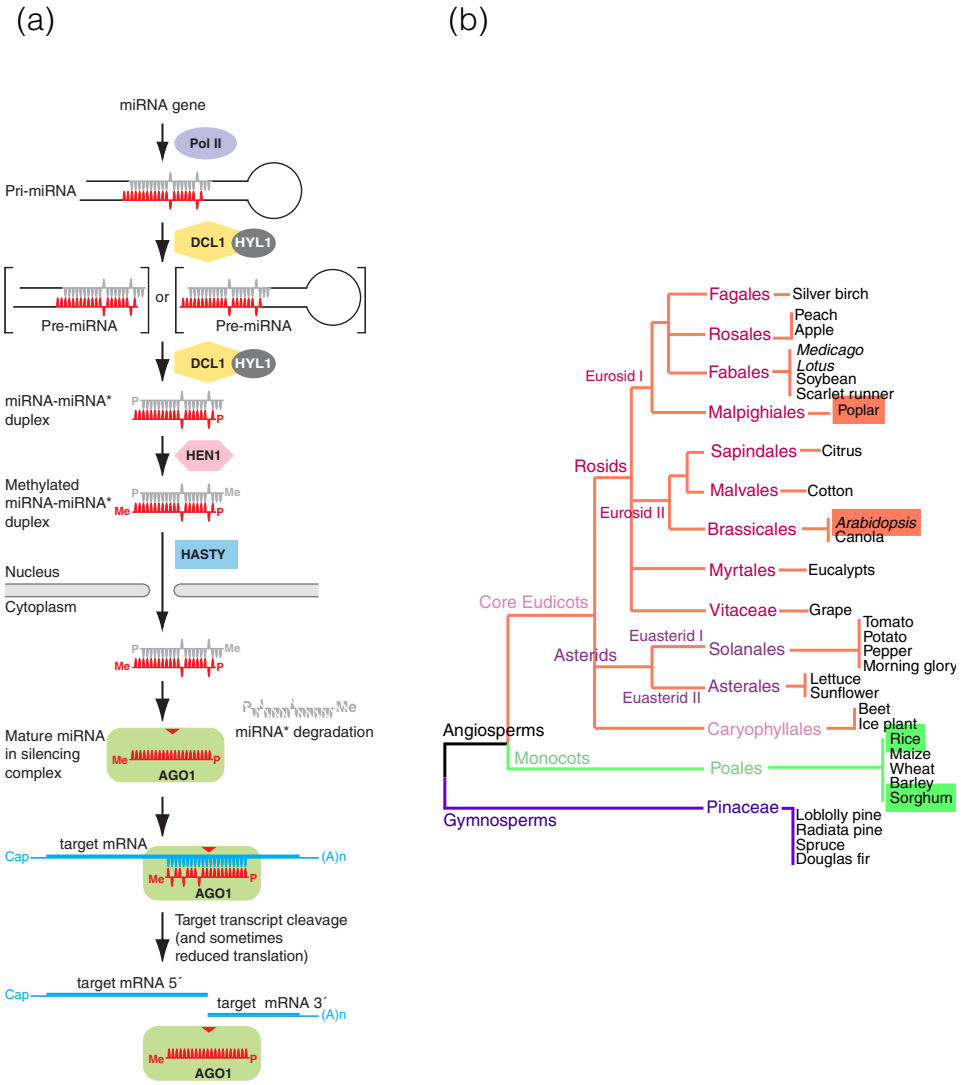



**Figure 1.2: (a)** Biogenesis of microRNAs in plants. See text for details. This figure is adapted from Jones-Rhoades *et al.* [2006] and reprinted with kind permission from the Annual Review of Plant Biology (©2006 by Annual Reviews, `www.annualreviews.org`) **(b)** Angiosperm phylogeny. The eudicotyledon, monocotyledon and gymnosperm clades are colored in red, green and blue, respectively. The two dicot species and the two monocot species that have been used for comparative microRNA prediction in Chapter 5 are highlighted. Note that the most recent common ancestor of dicots and monocots is dated as existing approximately 150 million years ago [Allen *et al.*, 2004]. This figure is adapted from Brunner *et al.* [2004] and reprinted with kind permission from Elsevier (©2004 by Elsevier, `www.elsevier.org`).

The precise mechanism that allows DCL1 to recognize the correct cleavage positions is largely unknown. The secondary structure (rather than the primary sequence) appears to play a dominant role in this decision, as substitutions in the primary sequence combined with compensatory changes in the opposing arm of the stem that sustain the pattern of hydrogen bonds along the stem leave the cleavage positions largely unaltered [Parizotto *et al.*, 2004].

This recognition is thought to involve the double-stranded-RNA-binding domain of DCL1 in collaboration with the HYPONASTIC LEAVES1 (HYL1) gene product. During maturation, the 3′-terminal nucleotides of microRNAs are methylated on their 2′ hydroxyl groups by the methyltransferase domain of the HUA ENHANCER1 (HEN1). Only then, after DCL1-mediated cleavage and HEN1-mediated methylation, are most microRNA/microRNA⋆ duplexes exported to the cytoplasm with the help of the nucleocytoplasmic transporter HASTY (HST).

In the cytoplasm, the microRNA strand of the duplex is loaded into the RNA-induced silencing complex (RISC), to which it confers sequence specificity. In plants, RISC includes AGO1 as its principal component, an Argonaute protein that contains a PAZ small RNA-binding domain and a Piwi RNase H-like domain.

The selection between the microRNA and the microRNA⋆ strand for RISC loading seems to be based upon energetic asymmetry in the bond strength of both ends of the duplex.

### Evolution

Many known plant microRNAs evolved early in the history of land plants [Axtell & Bartel, 2005]. Some microRNA families (e.g. miR160 and miR390) even date back to before the development of vascular systems and their sequence is shared nearly unaltered by eudicots and mosses. The extreme conservation of the mature 21 nucleotide microRNA sequence can be explained by the fact that a microRNA exerts influence on several target transcripts—all of which would have to co-evolve with their target site if regulation remained unchanged.

The remaining sequence of microRNA genes, however, is naturally much less conserved and its evolution is primarily restricted by its need for expression, the adoption of the foldback structure by its transcript, and its recognition by the biogenesis machinery. In Chapter 3, we look at this differential conservation along a microRNA gene in more detail, and in Chapter 5, we describe how to take advantage of this conservation to predict new microRNA genes in *Arabidopsis* and poplar, representatives of the dicot clade, and sorghum and rice, representatives of the monocot clade—see Figure 1.2(b).

As Allen *et al.* [2004] suggest, new microRNAs may evolve from the duplication of a target gene. In this scenario, such a locus could be capable

of adopting a stem-loop structure, if expressed, and could come under the control of RNA interference, spawning siRNAs at the duplication locus. If sequences at the duplication locus could mutate while maintaining the stem-loop structure then they could possibly adapt to the microRNA biogenesis apparatus and evolve into a microRNA gene with specificity for the founder gene and related family members. In Chapter 6, we propose an alternative model for microRNA evolution in addition to the model of Allen *et al.* [2004].

### Function

At first glance, it seems puzzling and uneconomical that an organism would invest resources in generating messenger transcripts on one hand, and, on the other, in an apparatus that simultaneously inactivates the former. Nevertheless, biology is full of examples where evolution does not necessarily optimize efficiency, and the added opportunity for control over gene expression could by itself justify the emergence of the microRNA-based regulatory system [Bartel & Chen, 2004].

In addition, certain scenarios might require that a cell alters its messenger RNA profile quickly. Here, microRNAs can provide rapid removal of unwanted regulatory mRNAs. At the mid-blastula transition of early animal embryogenesis, for example, microRNAs can coordinately destroy a large number of maternal mRNAs [Giraldez *et al.*, 2006; Weigel & Izaurralde, 2006]. In plants, microRNAs are implicated in developmental timing [Palatnik *et al.*, 2003], differentiation [Rhoades *et al.*, 2002] and response to stress conditions such as sulfate starvation [Jones-Rhoades & Bartel, 2004], phosphate starvation [Fujii *et al.*, 2005] and other environmental stresses [Sunkar & Zhu, 2004].

Many plant microRNAs target regulatory proteins such as transcription factors, suggesting that microRNAs are master regulators [Palatnik *et al.*, 2003; Jones-Rhoades *et al.*, 2006].

For metazoans, Bartel et al. propose the "micromanager model" [Bartel & Chen, 2004] as an explanation for the widespread, often subtle and customized influence of microRNAs on gene expression. Using the analogy of a dimmer switch that allows for control of adequate light, they suggest that microRNAs provide additional regulatory options on two levels: first, as mentioned, to accommodate for expression profile changes during cellular differentiation, development and as a response to stress conditions. Second, on an evolutionary level, where an organism would profit from the additional degree of fine-tuning that microRNAs provide.

# Chapter 2

# Prediction of MicroRNA Homologs

## 2.1 Motivation

MicroRNA genes that yield (almost) identical microRNAs are grouped into families because they recognize a common set of target transcripts, based on the signal they confer with their mature microRNA sequence. Another, largely independent signal is supplied by each microRNA gene's (foldback) structure which is decisive for recognition by the RNAse III enzyme DICER–LIKE1 during microRNA biogenesis.

Simultaneously taking advantage of both signals, we have designed and implemented a bioinformatic approach for the identification of microRNA homologs, given a microRNA and an appropriately formatted sequence database. This implementation, coined "microHARVESTER", serves a twofold purpose: first, given a validated microRNA as query, the microHARVESTER program can automatically identify its homologs with good sensitivity and specificity. Second, given a putative microRNA candidate as query, its homolog set, particularly the number of predicted homologs and their phylogenetic distribution, is indicative of the likelihood that this candidate is indeed a microRNA. In order to make use of both usage scenarios from an external application, the microHARVESTER program interacts with a relational database in which all results are stored.

## 2.2 Results

We have designed and implemented the microHARVESTER approach as detailed below and provide it to the community [Dezulian *et al.*, 2006a] both as an open–source standalone program and an online resource. In the latter case, it features an HTML interface that is dynamically generated by a web framework [Biegert *et al.*, 2006] and allows job tracking.

### 2.2.1 Identification Procedure

Given a known microRNA (microRNA precursor sequence plus mature microRNA sequence) as input for our search, we can use the precursor as query for a sequence similarity search against a set of sequences (e.g. reads from a plant genome or a set of EST sequences) to generate a set of candidate homologs. Since the (mature) microRNA sequence is very well conserved across large evolutionary distances [Axtell & Bartel, 2005], using BLAST [Altschul *et al.*, 1997] with the very large E–value cutoff of 10 and a minimal word size of 7, one can generate a hit for almost all microRNA homologs at the price of many false positives.

In the first filter step, we discard the sequences of the candidate set for which aligned segments do not span most of the mature segment of the query. In a second filter step, we apply a modified Smith–Waterman pairwise alignment algorithm [Smith & Waterman, 1981] to determine the mature sequence in the candidate precursor precisely from the optimal alignment of the query mature sequence against the corresponding segment of the BLAST hit. We discard a candidate if the length of the mature sequences differs by $> 2$ nucleotides. In a third filter step, we predict the minimal free energy structure of the candidate sequence using RNAfold [Hofacker *et al.*, 1994] and determine its putative microRNA$\star$ sequence. We discard a candidate if more than six nucleotides of its microRNA$\star$ are not predicted to form bonds with its mature microRNA (keeping in mind the 2 nucleotide offset between microRNA and microRNA$\star$) and pass on all remaining candidates.

### 2.2.2 Sensitivity and Specificity

In order to assess the sensitivity and specificity of this approach, we applied the microHARVESTER with parameters as published in [Dezulian *et al.*, 2006a] to the fully sequenced dicot *Arabidopsis thaliana* (Ath) genome using a set of query sequences from the monocot *Zea mays* (Zma). For each of the currently available (microRNA registry release 7.0) 18 microRNA families shared by Ath and Zma, we selected one Zma microRNA gene at random. Using this query set, the microHARVESTER identified 67 of the 75 Ath microRNA genes of these families—at least one in each family—at the price of five false positives. Analyzing why 8 of the Ath genes were missed revealed that 4 of them are members of the miR166 family which have 5 non-interacting base-pairs within the microRNA/microRNA$\star$ segment and were therefore discarded.

MicroHARVESTER is able to identify plant microRNA homologs with good sensitivity and specificity in any set of sequences, for a given query microRNA. For the identification procedure, the origin of the sequence database to be searched is irrelevant.

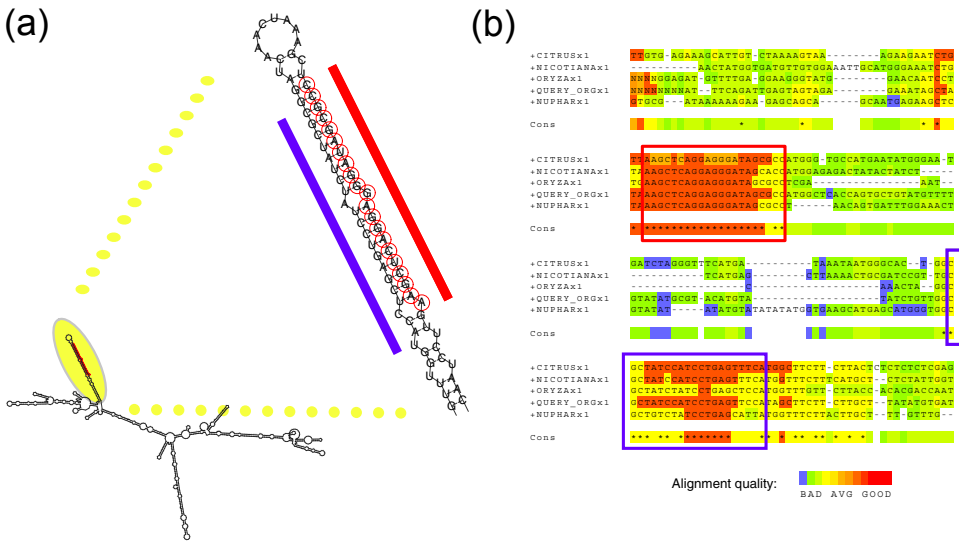



**Figure 2.1:** **(a)** As part of the resulting overview document, the predicted folding structure of each putative microRNA homolog is displayed (bottom left). The microRNA and the microRNA⋆ segments are colored in red and blue, respectively, as shown in the detailed view on the top right. **(b)** A multialignment of the relevant segment of selected homolog candidates is displayed. The alignment quality for each position was calculated using [Notredame & Abergel, 2003] and is indicated by shading, ranging from blue (poor alignment quality) to red (excellent).

However, using an EST database as the sequence pool offers the additional assurance that the predicted microRNA homologs are actually expressed [Zhang *et al.*, 2005].

### 2.2.3 Additional Procedures

Besides generating homolog candidates and filtering them, the microHARVESTER also determines any similarity of the query precursor with sequences in any of the the following databases using BLAST:

- All plant microRNAs of the Sanger microRNA registry.
- All *Arabidopsis* repetitive sequences as provided by TIGR.
- All rice repetitive sequences as provided by TIGR.

This is particularly useful in the second usage scenario, when microRNA candidates are used as the input for the microHARVESTER to determine whether any relationship to previously published precursors or repetitive sequences exists.

After homolog candidates have been determined, each of them is classified phylogenetically. Taxonomic information is derived in two ways, depending on the sequence database used. When an EST database provided by the NCBI is used, taxonomic information is derived using special retrieval options of the FASTACMD program that is part of the NCBI BLAST package in conjunction with the NCBI taxonomy database. If, on the other hand, a researcher uses his/her own sequence database, a special tag is provided that can be included in each sequence's header line to provide taxonomic information to the microHARVESTER. The database we provide contains all genomic sequences available at this time, tagged in this way, of the following plants: *Arabidopsis*, poplar, medick, lotus, rice, sorghum and maize. When results are tracked in the SQL database, this information is used to classify each resulting homolog candidate taxonomically. This is valuable since, in the second scenario, because of the strong phylogenetic conservation of most published microRNAs the likelihood that a given candidate is indeed a microRNA increases with the number of species in which a (putative) homolog can be detected. Additionally, we classify each query microRNA on the basis of whether its predicted homologs are spread across the plant kingdom (monocots and dicots), occur in either only monocots or only dicots, or are restricted to a single species—in decreasing order of interest.

### 2.2.4 Output Preprocessing

After the identification procedure and the homolog classification procedure we prepare a summary document in PDF format that captures all relevant information of the program run, formatted in such a way as to enable easy manual inspection of the putative homologs. In addition, a FASTA file is generated which contains the query and all predicted homolog sequences.

The summary document is partitioned in four sections:

1. An overview section that captures the program parameter settings, the results of the similarity searches against the Sanger registry and the repeat databases, and the number of predicted homologs in each genus and species.
2. A multiple alignment of the relevant section of a number of representative homologs.
3. A half–page detail description of each predicted homolog, including the homolog's predicted folding structure and predicted mature microRNA sequence.
4. Two tables, one listing all headers of the homolog candidates and the other listing all their accession numbers for easy cut–and–paste into other documents.

### Multiple Alignment

We construct the multiple sequence alignment of homologs using the T–Coffee software [Notredame *et al.*, 2000] to align a region that includes the microRNA, the microRNA⋆ and the "loop" sequence located in between the microRNA and the microRNA⋆. The reliability of each position of this multiple alignment is visualized using a color scheme, cf. Figure 2.1(b). In Figure 2.1(a), the corresponding mature microRNA region and microRNA⋆ regions of the hairpin are shown in a detail view of an EST–derived pri–microRNA transcript. When only few homologs are predicted we align all of them. If many homologs from different species have been predicted, we select at most ten representative homologs from as wide a taxonomic distribution as possible and use them for the alignment to capture maximal diversity.

### Candidate Details

Each predicted homolog is described on half a page in the overview document. This description consists of a picture of the predicted 2D folding structure of the 500 nucleotide segment containing the microRNA, with the microRNA and the microRNA⋆ marked in color. This greatly helps in deciding whether the hairpin containing the microRNA is part of a larger foldback section or not. Furthermore, the number and distribution of bulges, length of the hairpin and the size of the loop region can be examined and compared to those of the input query microRNA. A written description provides the following information: taxonomic information, accession number, sequence and position index of both the microRNA and the microRNA⋆, the E–value of the BLAST search, the strand of the microRNA (Watson or Crick), and the number of mismatches between the microRNA and the microRNA⋆.

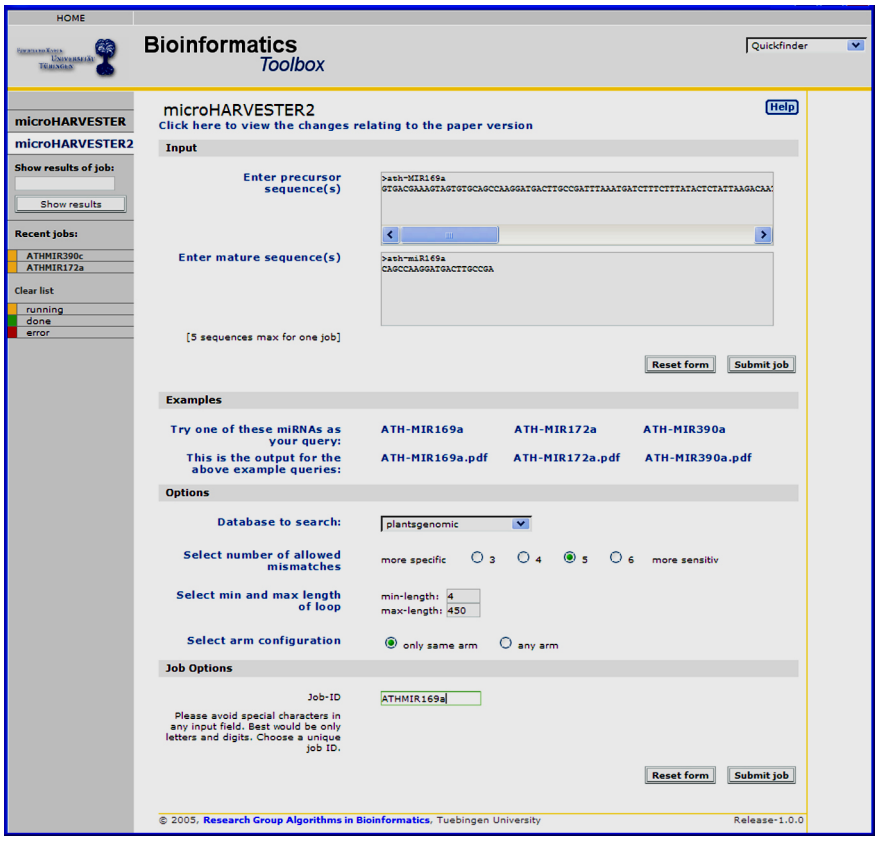



**Figure 2.2:** The graphical interface of the web service version of the microHARVESTER. In the *Input* section, a number of query microRNAs can be entered, supplying the mature sequence and the precursor sequence for each microRNA gene separately. In the *Examples* section, three validated microRNAs are provided for automatic insertion into the input section, along with the corresponding output in PDF format, which will result from applying the microHARVESTER to these queries. In the *Options* section, the database to be searched plus other options (see text) can be selected. Each run of the microHARVESTER is assigned a name (*Job Options* section) by which the result can be accessed later on using the task bar located on the left of the GUI. This task bar also provides color–coded status information for each recent job.

### 2.2.5 Web Interface

We have fitted the microHARVESTER with a web front–end that takes a batch of up to 5 microRNA sequences as input plus all relevant parameters as shown in Figure 2.2. A sophisticated, database–based job tracking facility allows monitoring the progress of several jobs running simultaneously. Also, by providing a fixed URL for each submitted job, a user can submit a job one day, then shut down his/her computer and pick up the results the next using this URL. Job results are provided in the form of the PDF summary document as explained above. A detailed discussion of the web interface's architecture is described in [Biegert *et al.*, 2006].

Two versions of the microHARVESTER are provided on our web site: *microHARVESTER* is the version of our approach detailed in our publication [Dezulian *et al.*, 2006a]. The implementation *microHARVESTER2* is an improved version which allows additional parameters to be specified, in particular, the greatest number of mismatches allowed between the microRNA and microRNA$\star$ can be chosen within a given range, thus balancing sensitivity and specificity. In addition, the minimally and the maximally allowed length of the loop segment between the microRNA and microRNA$\star$ can be restricted; a choice of databases is offered, and a user may choose not to discard predicted homologs on the opposite arm ($5'$ or $3'$) to that of the query precursor—a useful option for quality assessment of the prediction procedure, since homologous microRNAs always originate from the same arm of their respective precursor.

## 2.3 Methods

MicroHARVESTER is available as a web–service at `www-ab.informatik.uni-tuebingen.de/software/microHARVESTER`. Source code for the microHARVESTER is also available from the authors upon request. In order to run this standalone version on a standard Linux operating system, the following free software is also needed: Java 1.5, NCBI BLAST, RNAfold, T–Coffee and a standard LaTeX installation. Results can optionally be stored in a MySQL database.

For efficiency reasons, the sequence set which is to be used as a search database should consist of sequence fragments each being no longer than 1000 nucleotides each. This is due to the fact that microHARVESTER uses the program FASTACMD from the NCBI BLAST family of tools to efficiently retrieve candidate sequences directly from the BLAST database, and FASTACMD does not allow retrieval of sequence fragments. A tool, CHOPPER, that performs appropriate preprocessing of long sequences is provided in the microHARVESTER software package.

Also, taxonomic information needs to be provided when compiling one's own library (databases derived from the NCBI FTP server provide their own

taxonomic information as described in section 2.2.3). The repeat databases used are the following:

- *Arabidopsis thaliana*: file `TIGR_Arabidopsis_Repeats.v2` retrieved from `ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats/`
- *Oryza sativa*: file `TIGR_Oryza_Repeats.v3.1` retrieved from `ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats/`

The following genomic plant sequences have been used to compile the "plants-genomic" database selectable from the web interface:

- *Medicago truncatula*: file `all_bac_ends.Z` retrieved from `http://www.tigr.org/tdb/e2k1/mta1/` (downloaded on July $14^{th}$, 2005)
- *Sorghum bicolor*: `SAMIs v2.0 Contigs w/ Singletons` retrieved from `http://www.plantgenomics.iastate.edu/maize/` (downloaded on July $14^{th}$, 2005)
- *Zea mays*: file `ISU_MAGIs_3.1w_sing.fas.zip` retrieved from `http://www.plantgenomics.iastate.edu/maize/` (downloaded on July $14^{th}$, 2005)
- *Populus trichocarpa*: `Assembly v1.0` (file `poplar.unmasked.fasta.gz`) retrieved from `http://genome.jgi-psf.org/Poptr1/Poptr1.download.html` (downloaded on July $14^{th}$, 2005)
- *Oryza sativa*: TIGR Version 3.0 retrieved from `ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_3.0/all_chrs/all.con` (downloaded on July $18^{th}$, 2005)
- *Arabidopsis thaliana*: file `ATH1_chr_all.5con.gz` retrieved from `ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES/` (downloaded on July $18^{th}$, 2005)
- *Lotus japonicus*: files `BAC_end` and `Tac_end` retrieved from `ftp://ftp.kazusa.or.jp/pub/lotus/Endseq/BAC_end` (downloaded on July $18^{th}$, 2005)

In addition, the NCBI "EST–others" database which includes all EST sequences hosted by the NCBI except human and mouse sequences (in the version downloaded on the $27^{th}$ of July 2005) is provided for searching in the web service version of the microHARVESTER.

## 2.4 Discussion

Successful approaches for plant microRNA homolog identification have previously been described [Maher *et al.*, 2004; Adai *et al.*, 2005]. However, microHARVESTER is the first such tool that is available through a web interface. It complements a very recently published animal microRNA homolog identification approach [Wang *et al.*, 2005].

We have used a predecessor of the microHARVESTER to identify a large set of additional microRNA homologs, on the basis of which a survey of sequence– and structure–based properties could be conducted (cf. chapter 3). Also, we have used the microHARVESTER on the NCBI EST database, taking all published microRNAs as query set and identifying hundreds of homologs distributed across a wide taxonomic range. Other studies, similar in nature, have done the same and published an analysis of their results [Zhang *et al.*, 2005, 2006] making further research in this direction less attractive for us. The microHARVESTER, though simple in its approach, has also been of invaluable help in filtering intermediate candidate sets of our comparative prediction approach (cf. chapter 5) by evaluating the taxonomic distribution of putative homologs of new candidate microRNAs. For this, it has been essential that the results are deposited in an SQL database —enabling loose coupling of the microHARVESTER with other software modules.

# Chapter 3

# Conservation and Divergence of MicroRNA Families in Plants

## 3.1 Motivation

Plant microRNAs have been identified using one of three primary strategies. The first relies on the direct cloning of small RNAs. Several labs have prepared small RNA libraries from *Arabidopsis* and from rice, including different tissues and conditions [Llave *et al.*, 2002; Reinhart *et al.*, 2002; Sunkar & Zhu, 2004]. A second strategy is based on computational procedures which take advantage of the extensive conservation of microRNAs during evolution. Small RNAs that are conserved between *Arabidopsis* and rice with surrounding sequences that are able to form fold-back structures have allowed the computational identification of several new microRNAs and the postulation of many others [Jones-Rhoades & Bartel, 2004; Wang *et al.*, 2004]. A third approach has been the identification of microRNAs through forward genetics, an approach that had led to the first identification of small RNAs in plants [Palatnik *et al.*, 2003]. Although there are a few cases where a microRNA family is unique to *Arabidopsis* or rice, the majority of validated plant microRNA families is largely conserved across the plant kingdom [Axtell & Bartel, 2005].

Taking advantage of the extensive conservation of the mature part of microRNA precursors, we have used a predecessor version of the microHARVESTER tool (cf. Chapter 2) on large sequence databases to generate a multitude of microRNA homolog candidates for each family across several plant species. After extensive manual inspection, we have thus been able to identify and contribute a large set of additional microRNA homologs to the Sanger microRNA registry [Griffiths-Jones, 2004; Griffiths-Jones *et al.*, 2006]—as listed in Table C.1. Based on this enlarged set of plant microRNA

precursors, we have performed a number of analyses regarding structural features and sequence-level characteristics.

## 3.2 Results

The current version of the Sanger microRNA registry [Griffiths-Jones, 2004; Griffiths-Jones *et al.*, 2006] (Release 5.1) contains 286 plant microRNAs: 112 from *Arabidopsis*, 134 from rice and 40 from maize. We used a manually curated homology search (cf. Methods section) on the recently available genomic sequences of sorghum, maize, medick and poplar to identify 200 new microRNA genes, and we used the same strategy an a large database of expressed sequences tags to identify 37 additional microRNA transcripts. Overall, this increased the number of available microRNA genes by roughly 83%. The family distribution of this enlarged set of microRNAs is depicted in Figure 3.1(a). This enlarged set of sequences will be the basis for all following analyses.

For some microRNA families we did not detect any additional homologs. This may most likely be because of one of two reasons. One possibility is that this family has only recently evolved in *Arabidopsis*, as was shown for the families miR161 and miR163 [Allen *et al.*, 2004]. An alternative explanation is that the original *Arabidopsis* query sequences do not constitute *bona fide* microRNAs, since DICER-LIKE1 dependent biogenesis has not been determined for all small RNAs contained in the microRNA registry.

### 3.2.1 Family-Specific Conservation

First, we decided to examine the pairwise sequence similarity of microRNA genes on a per-family basis. For this, we conducted pairwise BLAST comparisons between all precursors belonging to a family. To standardize this comparison, we only took the microRNA/loop/microRNA$\star$ segment of each precursor into account and normalized all scores to prevent any possible bias caused by differing family size. As depicted in Figure 3.1(b), the variance of similarity scores varies greatly across families. The pairwise scores for family miR399, for example, are all very similar, despite the large number of family members, while, in contrast, the scores for family miR408 are spread across a wide range. The family-specific variance in similarity might reflect a combination of structural constraints and/or be related to the evolutionary time since the first speciation/duplication event that the founding member of this family was involved in.

### 3.2.2 Clade-Specific Divergence

MicroRNA families vary in size due to the number of species they have been identified in and because of different numbers of paralogs in each species.

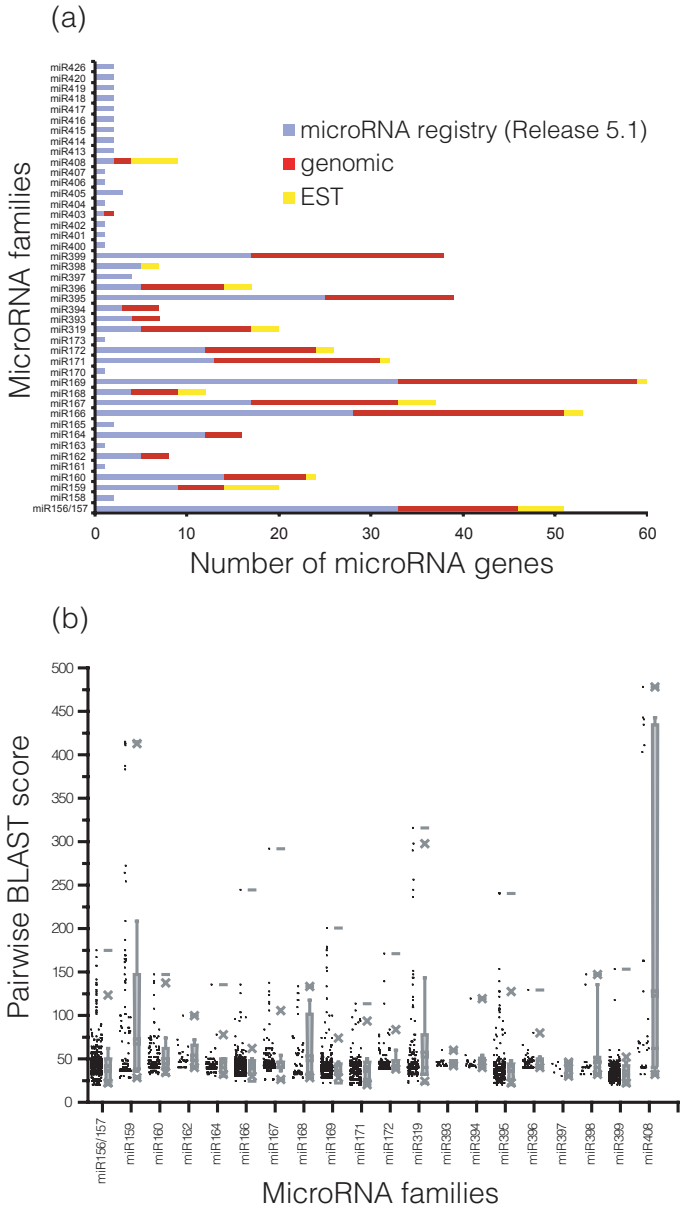



**Figure 3.1: (a)** Additional microRNA genes per family. Blue indicates genes contained in the RFAM microRNA registry (Release 5.1) and used as starting queries; red indicates newly identified genes from genomic sequences and yellow indicates newly identified genes obtained from expressed sequence (EST) databases. **(b)** Pairwise BLAST score of stem-loop precursor sequences (microRNA + loop + microRNA$\star$) for members of each family. Each data point is plotted in black. Statistical symbols are drawn in gray, maximal and minimal values by horizontal marks, first and $99^{th}$ percentile by crosses, and mean values by a square. A gray box covers the range from 25% to 75% with whiskers extending from 10% to 90%. This figure was adapted from our manuscript [Dezulian *et al.*, 2005].

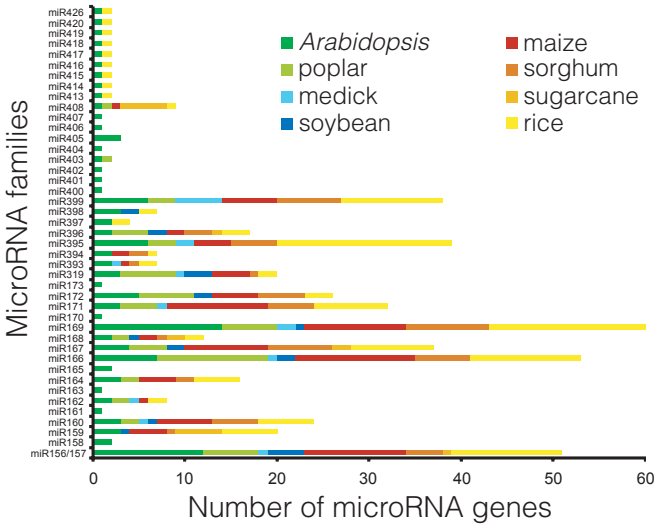



**Figure 3.2:** Histogram of microRNA genes. Bar segments are colored according to species: dark green (*Arabidopsis thaliana*), light green (poplar), light blue (medick), dark blue (soybean), dark red (maize), light red (sorghum), orange (sugarcane), yellow (rice). Figure adapted from our manuscript [Dezulian *et al.*, 2005].

Some families, like miR162, contain only two paralogs in a single species (*Arabidopsis*), while 14 members of family miR169 have been found in total across 7 species. Figure 3.2 depicts the family-specific distribution of microRNA genes across species.

In general, different microRNA families regulate different target genes. Since it seems likely that the importance of these target genes varies in a clade-specific manner this might have implications on the evolutionary pressure a family is exposed to and hence have implications for the number of paralogs a particular family contains in a specific clade (e.g. because of subfunctionalization).

To explore these possibilities, we tallied the number of family members separately for the sequenced (or almost fully sequenced) dicot species *Arabidopsis* and poplar and for the sequenced (or almost fully sequenced) monocot species rice, sorghum and maize. Figure 3.3(a) depicts the resulting plot of the clade-specific number of family members against each other. None of the families behaves as a crude outlier. On the contrary: the number of microRNA genes in monocots and dicots is roughly the same for each family, which is interesting in itself since the genome size of monocot species is often much larger than that of dicots.

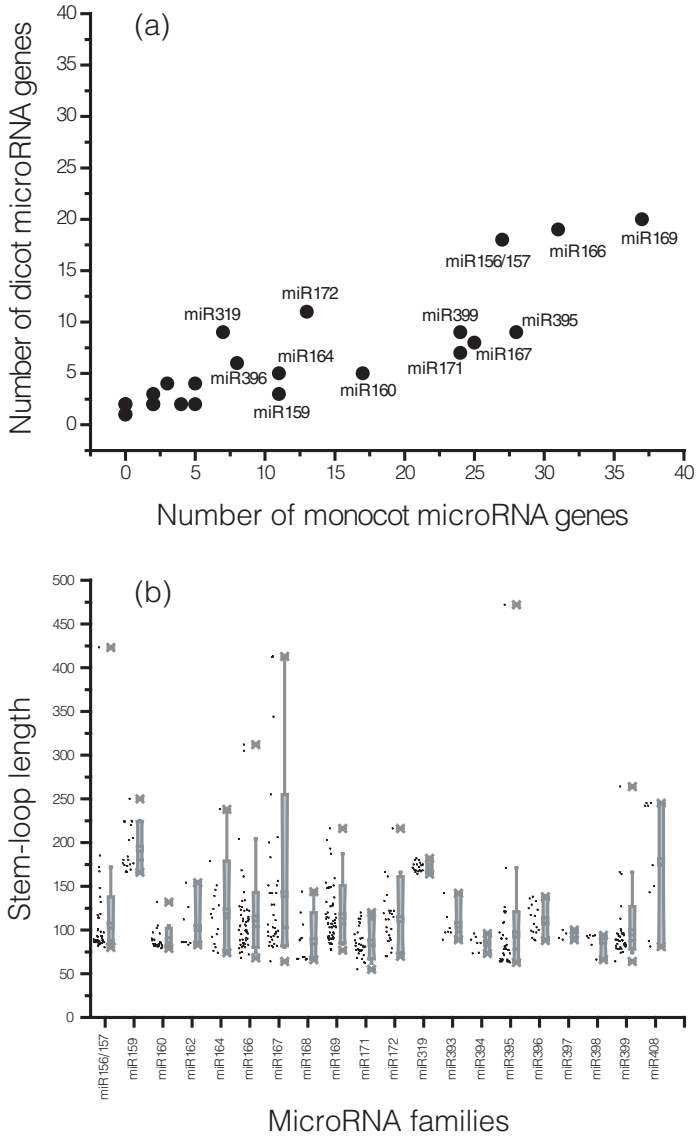



**Figure 3.3:** Conservation of microRNAs. **(a)** Number of monocot family members (rice, maize and sorghum) plotted against the number of dicot family members (*Arabidopsis thaliana* and poplar). **(b)** Stem-loop length variation for different microRNA families. Statistical symbols as in Figure 3.1. This figure was adapted from our manuscript [Dezulian *et al.*, 2005].

### 3.2.3 Structural Variation

Plant microRNAs are much more homogeneous in length than animal microRNAs. The length of the hairpin precursor of plant microRNAs can be anywhere between 50 and 500 nucleotides, whereas in animals it is usually around 70 nucleotides [Bartel, 2004]. As a first step to characterize plant microRNA precursors, we plotted the length distribution for each family in Figure 3.3(b). We found that several microRNA families, such as miR164 or miR408, have variable foldback sizes while others, such as miR319, showed little variation. This difference may be attributable to different structural constraints imposed by the microRNA processing machinery.

Next, we analyzed the structural conservation of plant microRNA precursors in detail. We compared all microRNAs of each family regardless of the originating species. For this analysis, we used only those *enlargeable* microRNA precursors for which the complete segment, starting from 50 nucleotides upstream and 50 nucleotides downstream of the microRNA/-loop/microRNA$\star$ hairpin, was available and excluded sequences which were shorter due to incomplete sequencing (especially pertaining to EST-derived sequences). We used the multiple alignment program T-Coffee [Notredame *et al.*, 2000] (cf. Methods section) for both alignment and visualization of the conservation at each position. T-Coffee makes use of the CORE algorithm [Notredame & Abergel, 2003] to indicate the conservation quality of the resulting alignment at each position using a color code as depicted in Figure 3.4.

As expected, we found a maximum of two nucleotides deviation from the consensus sequence in each microRNA family—resulting in a block of excellent conservation. The microRNA$\star$, being structurally constrained to tightly pair with the microRNA (with the exception of a few possible bulges) is second best in conservation and forms a second block of mostly good conservation. Compensatory mutations caused by wobble pairing and mutations at bulge positions are tolerated structurally and thus contribute to the degraded conservation quality with respect to the microRNA segment. These two blocks of conservation were clearly identifiable for each microRNA family. The segments upstream and downstream of the microRNA/loop/-microRNA$\star$ segment were much more poorly conserved in all families.

A very interesting result was obtained when we examined the conservation of the loop segment for each family. In most cases, such as the miR160 (cf. Figure 3.4) and miR164 families, there was essentially no conservation of this sequences across different species. Astonishing exceptions to this divergence pattern are the families miR159, miR319 and miR394. In these families, we found that the loop region of the foldback displays an unusual amount of conservation. Especially in the families miR159 and miR319, there were two other distinct blocks that gave signals similar in quality to that of the microRNA and microRNA$\star$.

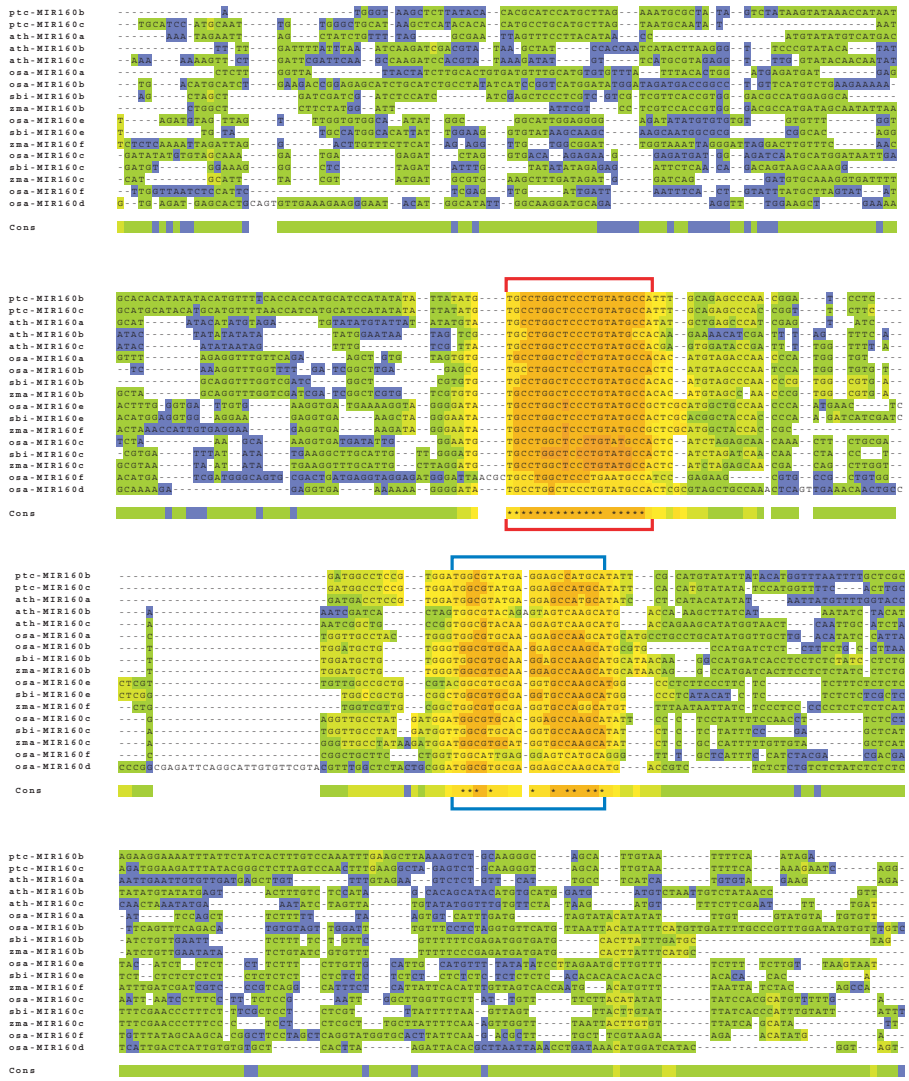



**Figure 3.4:** Alignment of all genes of the microRNA family miR160. The alignment software used, T-Coffee [Notredame *et al.*, 2000], provides an algorithm and a coloring scheme to indicate the degree of conservation: red/yellow/green/blue symbolizes excellent/good/average/poor conservation, respectively. The mature microRNA is marked by a red rectangle and the sequence segment pairing to the microRNA is marked by a blue rectangle. The alignment has not been curated manually. This figure was adapted from our manuscript [Dezulian *et al.*, 2005].

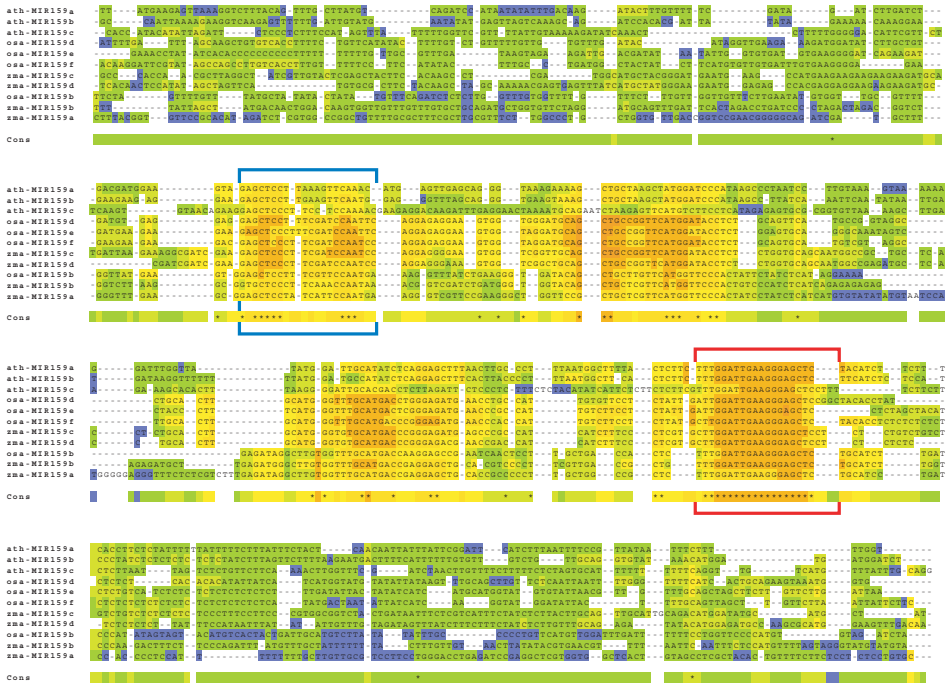


**Figure 3.5:** Alignment of all genes of the microRNA family miR159. The mature microRNA is marked by a red rectangle and the sequence segment pairing to the microRNA is marked by a blue rectangle. Note the additional conserved blocks of about 20 nucleotides length, about 15 nt downstream of the microRNA⋆ and 15 nt upstream of the annotated microRNA. The alignment has not been curated manually. This figure was adapted from our manuscript [Dezulian *et al.*, 2005].

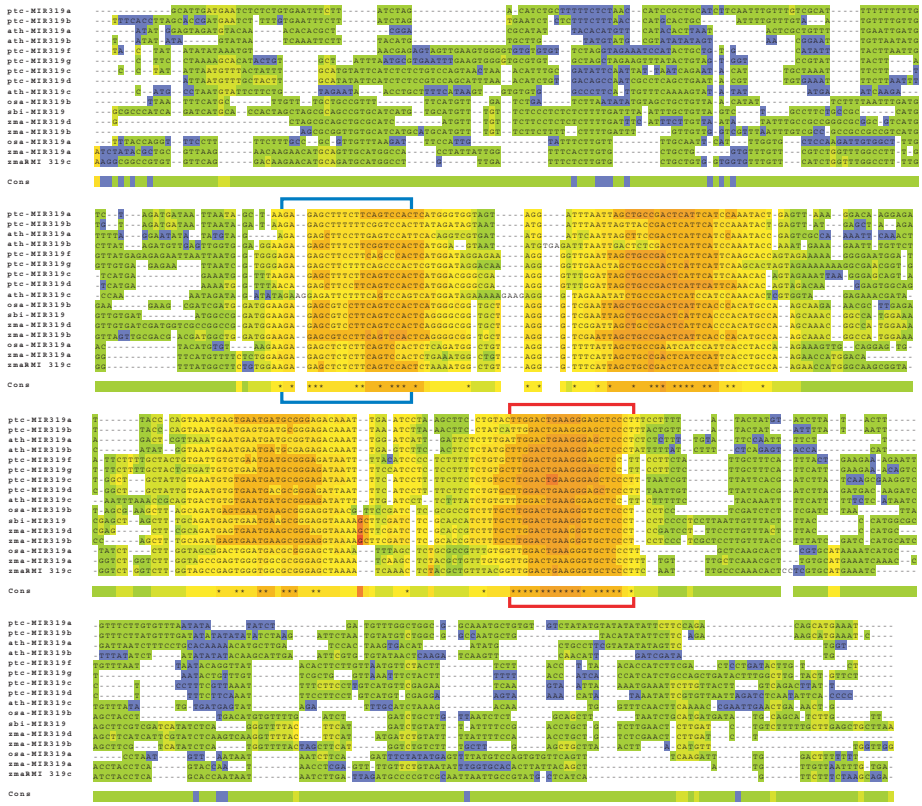


**Figure 3.6:** Alignment of all genes of the microRNA family miR319. The mature microRNA is marked by a red rectangle and the sequence segment pairing to the microRNA is marked by a blue rectangle. Note the additional conserved blocks of about 20 nucleotides length, about 15 nt downstream of the microRNA⋆ and 15 nt upstream of the annotated microRNA. The alignment has not been curated manually. This figure was adapted from our manuscript [Dezulian *et al.*, 2005].

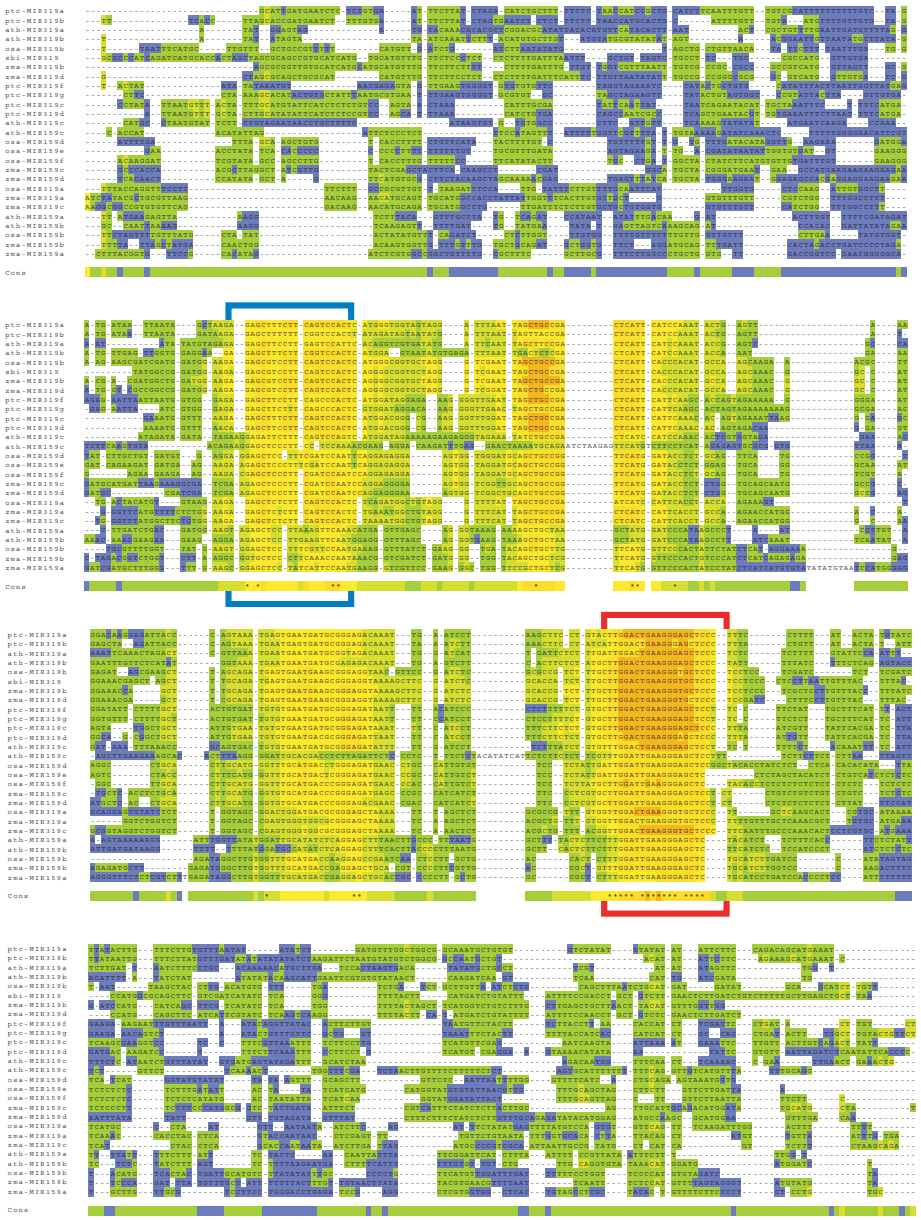


**Figure 3.7:** Alignment of all members of both family miR159 and miR319 together. Note that the conserved blocks of both the microRNA and the microRNA⋆ of each family align well together. Also, the additional blocks observed in each of these two families (cf. Figure 3.6 and Figure 3.6) are conserved in this united alignment. The alignment has not been curated manually. This figure was adapted from our manuscript [Dezulian *et al.*, 2005].

Figure 3.5 shows conservation in the miR159 family and Figure 3.6 shows conservation in the miR319 family.

When clustering and graphing all microRNA precursors by their sequence similarities we had previously noted that the miR159 and miR319 families blend together exceptionally well (cf. Chapter 4, data not shown). Thus in the case of the miR159 and miR319 families, this shared exceptional pattern of conservation was not astounding. Also, the mature microRNAs of these two families are very similar in sequence, though they appear to have largely non–overlapping target sets [Palatnik *et al.*, 2003; Achard *et al.*, 2004; Millar & Gubler, 2005].

We decided to use the precursor sequence without the microRNA and microRNA$\star$ segments to avoid any bias due to these sequences and performed a BLAST search against our microRNA precursor database. When we used the sequence of the miR319 precursor as query, we found that the best match is a miR159 and vice versa. Consequently, we performed an alignment of all enlargeable miR159 members together with all enlargeable miR319 members, which yielded the result shown in Figure 3.7. Four conserved blocks are clearly visible, resulting from the microRNA blocks of both families which align together, the microRNA$\star$ blocks of both families which align together, plus the two additional blocks of conservation that each family exhibits on its own aligning together as well. Taking into account the close similarity of the microRNA sequence of miR159 and miR319 and their shared extraordinary conservation pattern in the additional blocks, we hypothesized that these two families might share a common ancestry.

How can the additional conserved blocks in families miR159, miR319 and miR394 be explained? One possibility is that these segments code for a second microRNA. However, a search for putative target mRNAs using the tool WMD [Schwab *et al.*, 2006] yielded no plausible results. A more likely possibility might be that these segments are required for adequate microRNA biogenesis and to provide structural cues.

### 3.2.4 Position-Specific Nucleotide Preferences

Previous studies have reported a position-specific nucleotide bias in the mature microRNA sequence, with uracil being the most common base at the extreme $5'$ end. We were interested in the position-specific nucleotide bias in our enlarged precursor set and constructed position-weight-matrices (PWMs) for all microRNA sequences of length 21 for which 50 nucleotides were available upstream and downstream. Likewise, we constructed PWMs for their (microRNA$\star$) counterparts.

We first analyzed 307 microRNA precursors of this type (out of the 523 overall) and graphed their nucleotide bias per position (Figure 3.8(a)), confirming the preference for uracil at the $5'$ position. To rule out the possibility that this result is influenced by the differential size of microRNA

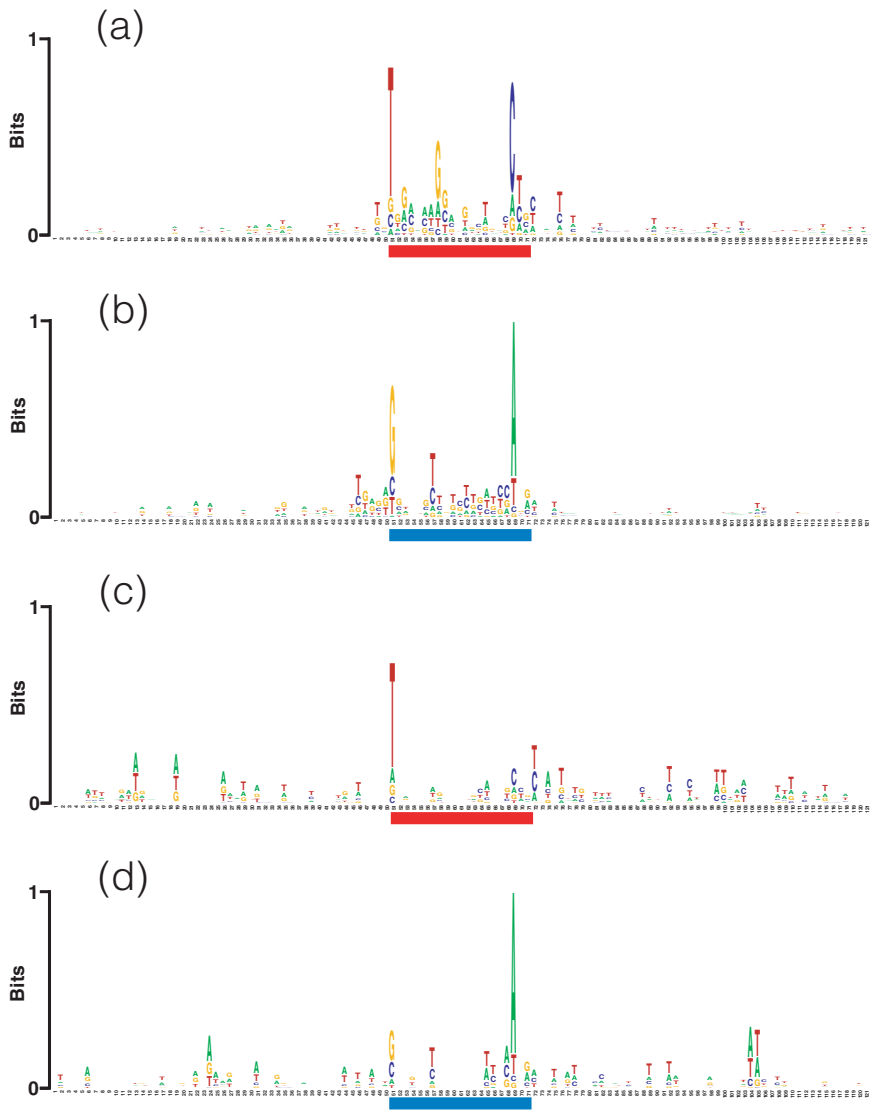



**Figure 3.8:** Position weight matrices (PWMs) of microRNA genes are displayed using the software WebLogo [Crooks *et al.*, 2004]. The y-axis indicates the total number of informative bits for each position. Within each column, the fraction of the height covered by each nucleotide is equal to its proportion of occurrences at the corresponding position. **(a)** and **(c)** display sequence logos of the mature microRNA plus $50$ nt $5$' and $3$' across all families and across one randomly chosen representative for each family, respectively. **(b)** and **(d)** display sequence logos of the microRNA$\star$ plus $50$ nt $5$' and $3$' across all families and across one randomly chosen representative for each family, respectively. The microRNA$\star$ is underlined with a blue bar. Only microRNAs of length $21$ for which $50$ nt upstream and downstream were available have been included in this analysis. This figure was adapted from our manuscript [Dezulian *et al.*, 2005].

families, we repeated this procedure after having selected only one such representative from each family at random and constructing a PWM for this set of 35 representatives (Figure 3.8(c)). Note that the guanine and cytosine preferences at positions 8 and 19, respectively, which are visible in Figure 3.8(a), are mostly due to the bias introduced by the large families miR166 and miR169, and disappear when we use per-family representatives (Figure 3.8(c)).

We found weak sequence signatures that have not been reported before, such as a pyrimidine preference at the first position downstream of the mature microRNA and a thymine preference at the fifth position downstream of the mature microRNA, although these signals were much weaker than the previously described uracil at position 1.

Figures 3.8(b) and (d) show analogous PWMs for enlargeable microRNA$\star$ sequences of length 21, using all 187 available sequences and the 28 sequences derived by selecting one representative each per family at random.

### 3.2.5 Bond-Specific Strand Selection

Finally, we analyzed the secondary structure of the microRNA and microRNA$\star$ to detect any difference in position-specific bond strength preference. We scored the strength of the bond at each position of the microRNA and of the microRNA$\star$, scoring each from its $5'$ end to compensate for variations in microRNA length and the effect of asymmetric bulges. We used the following *ad hoc* scoring scheme: GC pairs were assigned a score of 3; AU, 2; GU, 1; unpaired nucleotides (bulges) scored zero at that position. We found that the $5'$-most position of the microRNA scores an average of 1.6, while the $5'$-most position of the microRNA$\star$ receives an average score of roughly 2.4. This indicates that the first nucleotide of the microRNA is more likely to be unpaired than that of the microRNA$\star$ — a finding that is consistent with previous reports in animals [Khvorova *et al.*, 2003] which claimed that the protein complex in charge of loading RISC discriminates the microRNA from the microRNA$\star$ on the basis of differential $5'$ end stability.

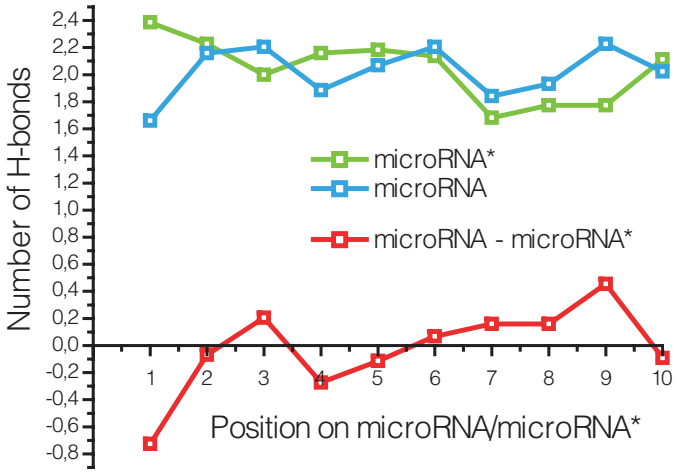



**Figure 3.9:** The average score (as described in the text) for the hydrogen bonds at the first 10 positions of the microRNA and the microRNA⋆, respectively, is displayed in blue and green. In red, the difference between the score for microRNA and the microRNA⋆ is plotted along the sequence. This figure was adapted from our manuscript [Dezulian *et al.*, 2005].

## 3.3 Methods

For the microRNA homology search, we used a predecessor of the microHARVESTER approach (cf. chapter 2) followed by substantial manual curation. Essentially, we used NCBI BLAST [Altschul *et al.*, 1997] with the large E-value cutoff of 10, using all 286 plant microRNA precursors (112 from *Arabidopsis*, 134 from rice and 40 from maize) currently hosted by the microRNA registry [Griffiths-Jones, 2004; Griffiths-Jones *et al.*, 2006] (Release 5.1, March 2005) as queries to generate a set of possible candidates. All matches were automatically folded at $19°C$ using RNAfold from the Vienna RNA package [Hofacker *et al.*, 1994] and checked for the adoption of a stem-loop-stem structure. Furthermore, we applied a second structural filter by requiring that the folding pattern of each homolog candidate is roughly similar in structure in the microRNA/microRNA⋆ region to its query—to this end, we required that at least the same number of interacting bases between the microRNA and the microRNA⋆ are found in the homolog candidate. All candidates that passed these test were manually inspected for homology with the query.

We used genomic sequence databases of sorghum, maize, medick and poplar. Of these four species, only poplar has been completely sequenced. It has been estimated that about two-thirds of the genic fraction of the sorghum and maize genomes have been obtained [Bedell *et al.*, 2005; Fu *et al.*, 2005].

We obtained the sorghum (*Sorghum bicolor*) genome [Bedell *et al.*, 2005], which has recently been sequenced by methyl filtration, from the MAGI website of the Iowa State University on $10^{th}$ of March 2005: (SAMI Version 2.0 Contigs w/ Singletons) at `http://magi.plantgenomics.iastate.edu/downloadall_s.html`. We obtained a re-assembly of the maize (*Zea mays*) genome [Emrich *et al.*, 2004; Fu *et al.*, 2005] from the MAGI website of the Iowa State University on $10^{th}$ of March 2005 (MAGI Version 3.1 Contigs w/ Non-repetitive Singletons) [20] at `http://magi.plantgenomics.iastate.edu/downloadall_s.html`. We obtained the genomic sequences of medick (*Medicago truncatula*) on $10^{th}$ of March 2005 from the NCBI Genome Survey Sequence database [Benson *et al.*, 2003] at `ftp://ftp.ncbi.nih.gov/blast/db/` (Files 'gss.0X.tar.gz'). A first assembly of the yet unpublished poplar (*Populus trichocarpa*) genome was downloaded from the DoE Joint Genome Institute and Poplar Genome Consortium web page at `http://genome.jgi-psf.org/Poptr1/Poptr1.download.html` on the $10^{th}$ of March 2005 (Version 1.0, preliminary draft). Furthermore, for searching for homology in EST sequences, we used the NCBI EST database [Boguski *et al.*, 1993], downloaded on the $10^{th}$ of March 2005.

All microRNA homologs identified in this approach have been deposited with the microRNA registry.

For the pairwise BLAST comparisons, we normalized the scores by setting a virtual (effective) database size equal to the length of the current NCBI NR/NT database ($13,371,533,914$ nucleotides). We find this useful since the BLAST E-value and score are directly related to the effective database length and omitting this would introduce a bias in similarity caused by differences in family size.

Post-processing of the search results was done using customized Java software and centered on a MySQL database. Overview documents for manual inspection were generated automatically for each candidate microRNA homolog using the Latex typesetting system.

## 3.4 Conclusion

Employing a similarity search of genomic and EST sequences with subsequent structural verification, we have been able to increase the number of plant microRNAs by 83% to 523 microRNA genes. Our analysis of this enlarged set has led to the following conclusions:

- In contrast to animals, plant microRNA precursors were already known to be more variable in length. While we can confirm that there is size variation both across and between families, we also found that not all families are equally variable. In some families, all members are uniform in size. This phenomenon might reflect evolutionary trajectories and/or differential functional constraints.

- The number of microRNA family members is roughly similar in monocots and in dicots across all families.
- With the exception of the previously reported uracil at the 5′-most position of the microRNA sequence, we find no obvious sequence bias. Consistent with the strand selection model for incorporating microRNAs into RISC, we observe differential bond strength between the 5′ end of the microRNA and 5′ end of the microRNA⋆.
- It seems that there are two classes of microRNA precursors with different structural properties. The most abundant class includes precursors that have only two strongly conserved regions or blocks comprising the microRNA and microRNA⋆. The foldbacks of these precursors contain a short stem consisting mainly of the microRNA/microRNA⋆ duplex. A second and less frequent class, which includes the microRNA families miR159, miR319 and miR394, display four conserved sequence blocks. This is reflected in the secondary structure of these precursors, which typically contain two adjacent, strongly paired stem segments. This possibly reflects a processing mechanism that requires two consecutive steps by DICER-LIKE enzymes, in a similar way to the progressive action of DICER in siRNA production.
- Finally and most importantly, for microRNA families miR159 and miR319, the close similarity of their mature microRNA sequence, the extensive similarity of their secondary structure and the existence of four conserved blocks that align well together provide convincing evidence for the hypothesis that they share a common evolutionary history, despite non-overlapping target sets.

# Chapter 4

# Visualization and Exploration of Sequence Relationships between (micro) RNAs

## 4.1 Overview

CrossLink is a versatile tool for the exploration of relationships between RNA sequences. After a parametrization phase, CrossLink delegates the determination of sequence relationships to established tools (BLAST, Vmatch and RNAhybrid) and then constructs a network. Each node in this network represents a sequence and each link represents a match or a set of matches. Match attributes are reflected by graphical attributes of the links and corresponding alignments are displayed when clicked on. The distributions of match attributes such as E–value, match length and proportion of identical nucleotides are displayed as histograms. Sequence sets can be highlighted and the visibility of designated matches can be suppressed by real-time adjustable thresholds for attribute combinations. Powerful network layout operations (such as spring-embedding algorithms) and navigation capabilities complete the exploration features of this tool. CrossLink can be especially useful in a microRNA context since Vmatch and RNAhybrid are suitable tools for determining the antisense and hybridization relationships, which are decisive for the interaction between microRNAs and their targets. CrossLink is available both online and as a standalone version at `http://www-ab.informatik.uni-tuebingen.de/software`.

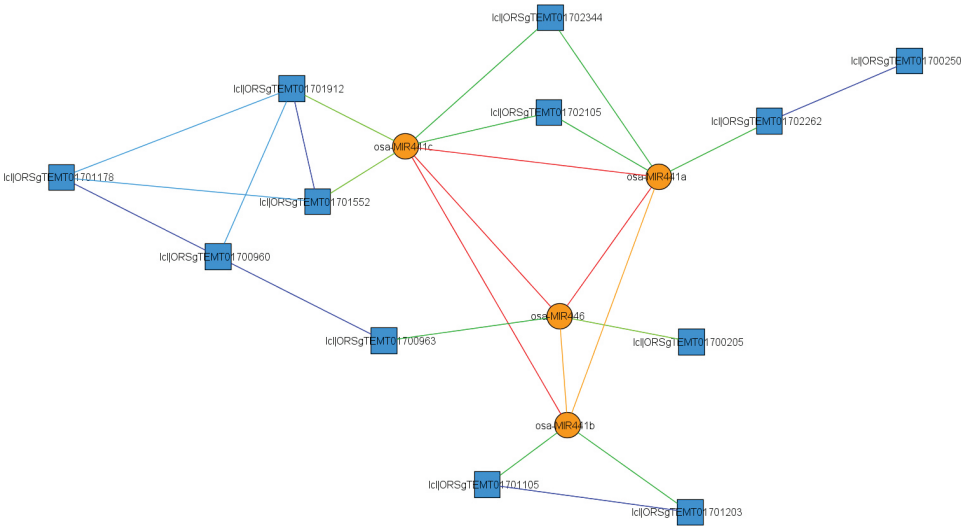



**Figure 4.1:** Network of sequence sets $A$ (red nodes) and $B$ (blue nodes) with corresponding matches of set $M_{AA}, M_{AB}$ and $M_{BB}$ represented by links in shades of red, green and blue, respectively.

## 4.2 Motivation

Many times during the prediction work, the need arose to compare different sets of microRNAs with each other or to compare a set of microRNAs with a set of putative target mRNAs, transposable elements or other types of RNA, for instance, when comparing the different resulting candidate sets of a prediction method which would result from varying parameters. The need also arose when determining the number of microRNAs (or homologs) in this type of result set that were already contained in the RFAM registry or contained in the resulting candidate set of another approach (e.g. [Adai *et al.*, 2005; Lindow & Krogh, 2005]) or library (e.g. [Gustafson *et al.*, 2005]). Furthermore, some sort of clustering by sequence similarity would be helpful to delineate microRNA families and examine their relationships.

As an extensive literature search brought no results, we realized that not only does no such tool exist for the comparison of microRNA sets, but that, astonishingly, a suitable exploration tool does not exist for the more general case of RNA for visualizing sequence sets and their relationships in an intuitive fashion. We thus decided to design an interactive exploration tool named "CrossLink" that would make use of the graph visualization analogy to visualize RNA sequences and their relationships in the form of a network (i.e. graph) in which sequences are represented by nodes and matches are represented by edges.

## 4.3 Results

Explicitly visualizing sequences and their relationships as a network provides concise and intuitive exploration possibilities. In this respect, CrossLink nicely complements the software CLANS [Frickey & Lupas, 2004] which uses a network to visualize sequence similarity between amino acid sequences. CrossLink delegates the determination of sequence relationships to the established tools BLAST [Altschul *et al.*, 1997], Vmatch [Kurtz *et al.*, 2001] and RNAhybrid [Rehmsmeier *et al.*, 2004]. Users versed with these tools will appreciate that (almost) all tool specific parameters may be set from within CrossLink. Furthermore, CrossLink allows relationships determined by distinct tools to be visualized within the same network.

Both BLAST and Vmatch can detect local sequence similarity in both sense and antisense directions and are suitable for a wide range of scenarios. BLAST is a standard tool using a fast seed–and–extend strategy. Vmatch employs a suffix array–based approach that permits constraints on the match length and on the number of mismatched bases within a match.

RNAhybrid is a specialized tool that can predict potential binding sites of microRNAs in large target RNAs using an extension of the classical RNA secondary structure prediction algorithm [Zuker & Stiegler, 1981]. In general, RNAhybrid finds the active sites that are most favorable to hybridizing a small RNA sequence in a large RNA sequence.

Although CrossLink can be put to use in many scenarios amenable to the above tools, it can be especially useful in a microRNA context: microRNAs interact with target transcripts by complementary base–pairing and can be classified into families on the basis of sequence similarity relationships that can be detected by using Vmatch/RNAhybrid and BLAST, respectively (cf. Examples below). Balancing flexibility and complexity, CrossLink allows the user to independently specify three different kinds of relationship searches, each with its own strategy (BLAST, Vmatch and RNAhybrid) and a set of parameters. To this end, CrossLink's input consists of two sets of RNA, $A$ and $B$, each provided in the FASTA [Pearson & Lipman, 1988] format.

The first kind of similarity search, $S_{AA}$, is performed between all sequences of set $A$, yielding the set of matches $M_{AA}$. Likewise, similarity searches $S_{AB}$ and $S_{BB}$ are performed to yield the set of matches $M_{AB}$ (between all sequences of set $A$ and all sequences of set $B$) and the set of matches $M_{BB}$ (between all sequences of set $B$), respectively. For clarity, a color scheme is associated with each kind of match: reddish colors frame the parameter input controls for $S_{AA}$ as well as the match representations of $M_{AA}$ in the network, corresponding alignment windows and histograms. Similarly, $S_{AB}$ and $M_{AB}$ are associated with greenish colors and $S_{BB}$ and $M_{BB}$ are associated with bluish colors (Figure 4.1).

Within each color scheme, shades indicate the orientation of each match: a dark shade is associated with matches in sense orientation and a light

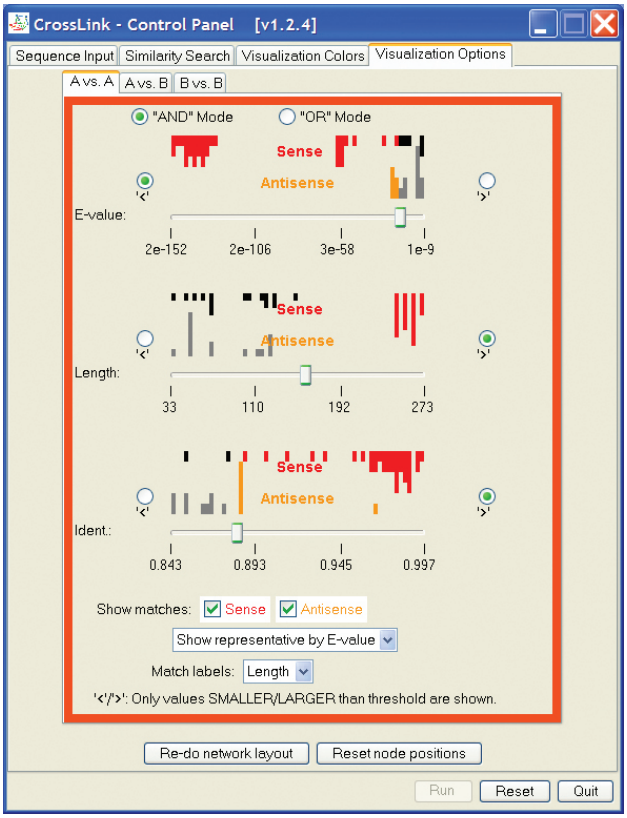



**Figure 4.2:** The visualization options panel for matches in set $M_{AA}$ displaying the histograms associated with each match attribute.

shade is associated with matches in antisense orientation. In addition to orientation, each match has the following attributes: E–value, length and the proportion of identical nucleotides within the alignment when the match was determined by using BLAST or Vmatch; minimal free energy (MFE), length and the proportion of paired nucleotides within the alignment when the match was determined by using RNAhybrid.

For each match set, a visualization option panel (Figure 4.2) is provided that uses a histogram for each match attribute to display the corresponding value distribution. Sense matches and antisense matches are tallied separately in each histogram. Note that the E–value and MFE histograms run on a logarithmic scale and the length and identity/paired proportion histograms run on a linear scale. Serving a twofold purpose, the visualization option panel also allows manipulation of the network: a threshold may be set for each attribute and a specified combination of thresholds then determines which matches will be considered for analysis and represented as links in the network and which will be suppressed.

This feature allows the user to rapidly focus on matches with interesting characteristics. A threshold is set by adjusting a slider for each attribute and selecting a combination mode. Two combination modes are available: in conjunction mode (logical "AND") only matches that pass all thresholds will be displayed. In disjunction mode (logical "OR") only matches that pass at least one of the thresholds will be displayed. Whether the threshold acts as a cutoff for smaller or higher values of an attribute is specified by radio buttons located on the left and right of each attribute histogram. In addition, all sense and/or antisense matches may be suppressed for a given match set.

Exploration can be focused further on an arbitrary selection of sequences by removing all remaining sequences (along with their relationships) from the exploration session using the menu bar (▶ *View* ▶ *Remove all unselected nodes*). All histograms are accordingly recalculated on the basis of the remaining relationships. An exploration session involves three phases that occur in order: first, during a parametrization phase, the two input files are chosen and a strategy is selected for each of the three relationship searches (BLAST, Vmatch or RNAhybrid) and the corresponding parameters are specified. Next, in the search phase, CrossLink uploads all necessary information to the server and the search is performed remotely. Upon completion the results are passed back. During the final exploration phase the resulting network is visualized and relationships can be explored. A reset button permits the user to jump back to the parametrization phase with the current parameters.

Any two sequences can give rise to several distinct local sequence similarities. Representing each match by its own link may clutter up the network visualization when many sequence pairs each yield a multitude of local matches.

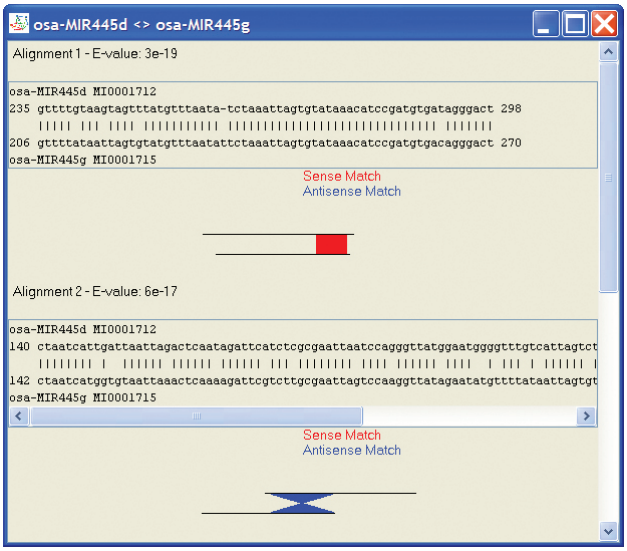



**Figure 4.3:** Alignment window showing two separate matches between one pair of sequences.

Therefore, each match set can independently be displayed in either "single match representative mode" or "multiple match representative mode". In "single match representative mode", each link between two network nodes represents a single match between the corresponding sequences. In the case of several matches between this pair of sequences, each is represented by its own link running alongside the other links between the two nodes. In "multiple match representative mode" a link between a pair of sequences represents all corresponding matches. One can select whether the representative of this match set should be the one with the smallest E–value/MFE, greatest length or highest identity/paired proportion—as this may be relevant for the mentioned attribute histograms.

Clicking on a node or link of the network spawns a separate window displaying detail information about the corresponding sequence or match(es) (Figure 4.3). Note that the alignment is displayed in text form exactly as it was produced by the originating tool. Clicking on a subset of selected nodes generates a separate window displaying the corresponding sequences in the FASTA format. This enables sequence subsets to be exported for further scrutiny using other tools.

By default, sequences of set $A$ and set $B$ are displayed as red and blue nodes, respectively, in the network. Arbitrary colors may be assigned to subsets of sequences using the following mechanism: a color can be associated with a text pattern and each sequence which contains the text pattern exactly as a substring in its FASTA header will be colored accordingly.

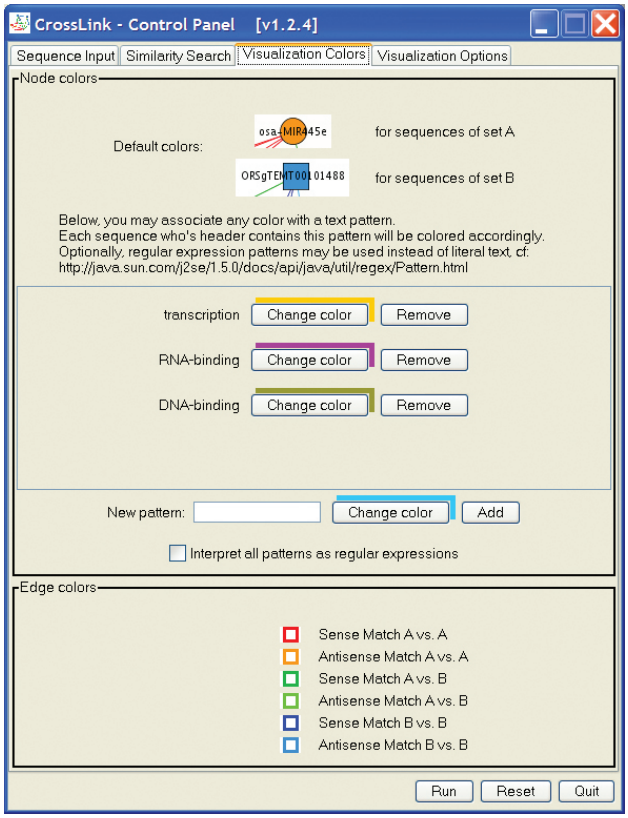



**Figure 4.4:** The visualization color panel, showing custom pattern–color associations in the center.

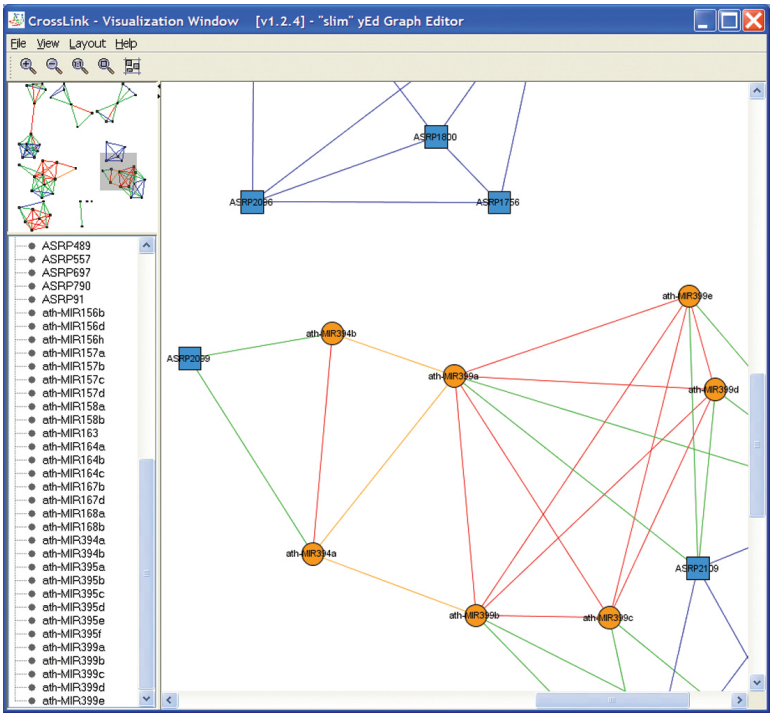



**Figure 4.5:** The visualization window, with an overview area on the top left and a sequence selection panel on the lower left.

Optionally, the pattern may contain a regular expression that is matched accordingly. Any number of these pattern-color associations may be specified (Figure 4.4). A sequence associated with several colors will appear multicolored.

To facilitate repeated exploration runs, the current parameter set can be named and saved as a configuration template. Any subsequent exploration task can be based on a configuration template either "as is" or after modification. Each configuration template contains the following parameters: each of the three sequence similarity search strategies including all parameters, the custom pattern–color associations and the two sequence input file names (as associated default file names). Note that, for consistency, selecting a different configuration template does not change the currently stated input file names. However, the default file names associated with the current template can be chosen explicitly.

A visualization window offers fast and powerful navigation of the network shown in the main view area (Figure 4.5). An overview area displays the currently visible clipping as a gray rectangle, which can be dragged, focussing the main view area accordingly. The mouse wheel permits rapid zooming. Network nodes can be selected and moved. Double–clicking on

a sequence in the sequence selection pane (Figure 4.5, lower left) centers the view onto this sequence. Dragging the mouse cursor over a sequence displays its FASTA header.

Several algorithms are available for network layout. The default layout algorithm is a Fruchterman–Reingold [Fruchterman & Reingold, 1991] spring–embedding, similar to the one used in the BioLayout [Enright & Ouzounis, 2001] library, where each link acts as a spring pulling at the sequences it is attached to. A "Reset node positions" Button undoes all node movements performed since the last application of a layout algorithm. CrossLink's visualization component is based on the yFiles [Wiese *et al.*, 2001] graph library which provides the spring–embedding implementation.

## 4.4 Examples

CrossLink provides three sample configuration templates along with the corresponding sequence files. To try out CrossLink, one merely has to select one of the samples and press the "Run" button. The following sample scenarios are provided:

- Example 1:
  Sequence set $A$ consists of all rice microRNAs of families 440–446 available from miRBase [Griffiths-Jones *et al.*, 2006]. Sequence set $B$ contains a subset of repetitive rice sequences downloaded from the TIGR Rice Genome Annotation Database. It is immediately visible that, for example, the rice microRNA family 445 exhibits very close sequence similarity to a family of repetitive rice sequences. Initially displaying a multitude of links in a tangle, this example demonstrates the power of the interactive histograms to focus on relevant relationships.

- Example 2:
  Sequence set $A$ consists of all *Arabidopsis* microRNA precursors available at miRBase. Sequence set $B$ contains all ($\sim$2000) sequences contained in the *Arabidopsis* Small RNA Project database [Gustafson *et al.*, 2005] to date. Setting these two sets in relationship with each other allows one to assess which microRNA families have been sequenced by the ASRP project. This example also demonstrates CrossLink's ability to handle large sets of sequences and also shows the power of the spring-embedding algorithm in clustering microRNAs into families.

- Example 3:
  Sequence set $A$ consists of the *Drosophila* microRNAs dme-miR-3, dme-miR-4 and dme-miR-5. Sequence set $B$ contains all corresponding targets which have been predicted (with an E–value $< 1$) in a study by Rehmsmeier et al. [Rehmsmeier *et al.*, 2004], plus some

randomly picked sequences from the same study that have not been predicted as potential targets of these microRNAs. This example demonstrates the use of RNAhybrid, revealing that one sequence (accession no. CG15125) is simultaneously targeted by two different microRNAs. Furthermore, the capability of custom pattern–color associations is shown as each predicted target set of the Rehmsmeier et al. [Rehmsmeier *et al.*, 2004] study is associated with its own color (yellow, magenta and cyan for the targets of dme-miR-3, dme-miR-4 and dme-miR-5, respectively) and the non–targets are shown in blue.

## 4.5 Methods

CrossLink is available both online and as a downloadable local version. Both versions require an installed Java Runtime Environment (JRE1.4.2 or later). To prevent overload of our server, the online version restricts the size of the two input files to 1 MB. The local version requires locally installed NCBI BLAST, Vmatch and RNAhybrid tools and a TCSH command line. We distribute the client software both as a Java Web Start client and as a Java Applet—both secured with a code signing certificate of our department issued by Thawte. Client/server communication is HTTP–based, secured by a firewall and guarded against SQL–injection attacks. A PHP script handled by an Apache server receives the similarity search request along with all parameters and sequences, and passes it on to a Java server program. All client/server communication is logged in a MySQL database and can be monitored remotely via HTML interface supplied using PhpMyAdmin. The CrossLink website (at `http://www-ab.informatik.uni-tuebingen.de/software`) provides a user manual including a quick start guide, instruction on how to set up Java Web Start on different platforms, plus detailed descriptions of the sample input data that CrossLink supplies.

A crucial component of CrossLink is the yFiles graph library and the yEd graph editor of the yWorks company (`www.yworks.com`). We gratefully acknowledge the permission to use these great libraries.

## 4.6 Discussion

CrossLink enables quick, intuitive and interactive exploration of arbitrary RNA sequence similarities. In addition, it provides features especially suited to the exploration of microRNA sequences (e.g. target prediction). Because of its universality, we have used CrossLink in most of the projects mentioned in this thesis and followed this procedure: Firstly, CrossLink is run with very unspecific relationship search setting (e.g. an E–value of 10) on all available sequences of the study in question. Then, one experiments with the view parameters to focus on relevant relationships. Re-layouting the graph reveals

clusters formed under these modified constraints. Focussing further, one can exclude all sequences which did not show interesting relationships for the question at hand from the study. Finally, one can export selected sequences for further scrutiny using specialized tools for the relevant question (e.g. for multialignment or refined target prediction).

# Chapter 5

# Comparative Prediction of Plant MicroRNAs

## 5.1 Introduction

Most plant microRNAs have been identified either by sequencing small RNAs or by computational approaches. The latter take advantage of the extensive conservation of most known microRNAs across the plant kingdom [Floyd & Bowman, 2004; Axtell & Bartel, 2005], coupled with a series of filters that distinguish microRNA genes from other conserved sequences on the basis of structural and sequence characteristics. We have developed and implemented a whole-genome comparative approach, termed "microSECTOR" for the *de novo* identification of microRNAs, given the genomic sequences of a pair of plant species.

In this chapter, we briefly sketch the design of this approach and then report on the resulting microRNA candidates of two analyses, each of which is based on the application of the microSECTOR to a unique pair of plant genomes.

Our endeavor is similar in spirit to the one presented by Matthew Jones-Rhoades and David Bartel [Jones-Rhoades & Bartel, 2004] but unique in three decisive aspects, which we will briefly list here and will discuss in more detail in Section 5.2.

Firstly, at the time of this project, of all plants, only the genomes of *Arabidopsis* and rice had been fully sequenced and assembled, so therefore published comparative approaches had used these two plants for comparison [Jones-Rhoades & Bartel, 2004; Bonnet *et al.*, 2004]. Also, a few studies had included the sequenced portion of the maize genome [Maher *et al.*, 2004]. We applied our microSECTOR approach to two pairs of genomes, both of which had not previously been used for comparative microRNA prediction:

- *Arabidopsis* and poplar, and
- rice and sorghum.

Since *Arabidopsis* and poplar are dicotyledons, and rice and sorghum are monocotyledons, one could hope to identify clade-specific microRNAs for these very important clades. Especially for the dicot clade, this would be the first analysis able to identify clade-specific microRNAs.

Secondly, the microSECTOR approach incorporates a very recently developed tool, RNA**L**fold [Hofacker *et al.*, 2004], which is able to predict locally stable secondary structures with unprecedented efficiency and is especially suitable for a genome-wide survey in higher eukaryotes. To our knowledge, no other microRNA study had made use of this tool before.

Thirdly, the last step of the microSECTOR approach incorporates the use of a modified version of the microHARVESTER tool (cf. Chapter 2). Applying it to the resulting candidate set, using genomic sequences of a wide variety of plants and the NCBI EST database as background information, allows further filtering of the resulting candidate set before it is subjected to manual scrutiny.

The microSECTOR approach has been implemented in two separate versions: the first implementation (microSECTOR1) was programmed by Christian Klug in the course of his diploma thesis [Klug, 2005] and applied to the *Arabidopsis* and poplar genomes. A second code-independent and slightly different implementation (microSECTOR2) was programmed by Christian Mayer in the course of his student project [Mayer, 2005] and applied to the sorghum and rice genomes (as well as to the *Arabidopsis* and poplar genomes for comparison).

In the next section, we provide background information regarding each of the unique aspects of our endeavor mentioned above. Furthermore, we sketch the preparatory work that led to the microSECTOR approach and our analyses. Then, in Section 5.3, we outline the microSECTOR approach. For details, we refer to the Master's thesis of Christian Klug [Klug, 2005] and the student project of Christian Mayer [Mayer, 2005] as we will not repeat these here. The main part of this chapter, Section 5.4, is devoted to the two analyses conducted with the help of the microSECTOR, one identifying new microRNAs using the two dicot genomes and the other using the two monocot genomes.

## 5.2 Background

### The Poplar Genome

Following *Arabidopsis* and rice, the poplar species black cottonwood(*Populus trichocarpa*) was the third plant whose genome was fully sequenced. As poplar is a dicot like *Arabidopsis*, the most important model plant for molecular biology, its sequence harbors great potential for comparative genomics. Consequently, when a first assembly of the poplar genome was made available on the internet `http://genome.jgi-psf.org/poplar0/` by the Joint

Genome Institute in June 2004, we immediately started to devise a comparative approach for the identification of microRNAs using the *Arabidopsis* and poplar genomes. This was complicated by the fact that absolutely no annotation or gene prediction for any of the sequences in this first assembly was available—this situation only changed more than two years later when, in September 2006, the poplar genome was officially published [Tuskan *et al.*, 2006]. By that time, of course, our analysis was long completed. Furthermore, a very large proportion of the poplar genome sequence was still fairly unassembled in the June 2004 assembly version and not even assigned to a chromosome.

### RNA Secondary Structure Prediction

As microRNAs exhibit a characteristic hairpin-like folding structure, any prediction approach incorporates an RNA secondary structure prediction step. Secondary structure prediction is a non-trivial process and typical free energy minimization techniques, such as those employed by the programs mfold ([Zuker & Stiegler, 1981; Zuker *et al.*, 1999]) and RNAfold ([Hofacker *et al.*, 1994]) make use of dynamic programming and exhibit a computational complexity that is cubic in the length of the input sequence for which the structure is to be determined.

Another important aspect of RNA folding is that the structure that a specific segment of an RNA molecule adopts is very dependent on the neighboring (5' and 3') sequence segments. This results from the fact that the mapping from RNA sequence space to RNA secondary structure space is not continuous (in a mathematical sense)—implying that two almost identical sequences may yield completely different structures (cf. [Voss, 2004; Shen *et al.*, 1999]).

These two problems of RNA structure prediction, computational complexity and strong context-dependance, are aggravated when facing whole eukaryotic genomes. In a preparatory project [Bitsch, 2004], we decided to devise our own approach for the predicting regions that were capable of forming stable hairpin-like structures. After trying different approaches, we ultimately used a seed-and-extend approach inside a sliding window (that progressed along the genome) that searched for almost perfectly matching microRNA/microRNA$\star$ pairs. Then we used dynamic programming to implement the Nussinov algorithm [Nussinov & Jacobson, 1980] and checked whether the surrounding region would form a foldback.

This preparatory project ended fruitlessly in terms of newly discovered microRNAs, due to our lack of experience regarding crucial microRNA-specific features and tools, but it paved the way for the two follow-up projects on whose results we report below. One helpful outcome, that was a prerequisite for these projects, was the implementation of our own secondary structure prediction web server termed "microFOLD" (available

at `http://www-ab.informatik.uni-tuebingen.de/toolbox/index.php?view=rnafold`) that makes use, internally, of a modified version of RNAfold [Hofacker *et al.*, 1994]. Given a microRNA precursor sequence and a putative mature sequence, it quickly produces a picture of the predicted structure in which the mature segment is highlighted within the precursor.

Another helpful outcome was the decision to install our own dedicated sequence similarity web server that internally uses the BLAST and Vmatch tools to detect similarity to a set of specific sequence databases. These included databases containing established microRNA precursors, expressed sequence tags and genomic sequences of a variety of plants including all available genome survey sequences hosted at GenBank [Benson *et al.*, 2003].

Most importantly, we became aware of a brand-new tool, RNA**L**fold [Hofacker *et al.*, 2004], which is able to predict locally stable structures in very large sequences with unprecedented efficiency. Its algorithm is based on a sliding window that is moved along the input sequence. For each position of this window, it essentially uses the established algorithm of RNAfold to predict the structure. The trick is that it incrementally updates its dynamic programming matrix as the sliding window is moved, thus preserving a cubic complexity in the length of the input sequence for a fixed window size. As no plant microRNA precursor is larger than 500 nucleotides, and a window of this size is acceptable, this tool is ideally suited for predicting their stable hairpins along a genome.

The final step of the microSECTOR approach involves the use of the microHARVESTER tool (see Chapter 2). Applying the microHARVESTER to the candidate set against the NCBI EST database provides additional filtering, as the number of detected homologs and their taxonomic distribution are indicative of the likelihood that a particular candidate is indeed a microRNA precursor.

After the first project (on poplar and *Arabidopsis*) was completed, the coding portion of the sorghum (*Sorghum bicolor*) genome was sequenced by methyl filtration [Bedell *et al.*, 2005] and a draft of a first assembly [Emrich *et al.*, 2004; Fu *et al.*, 2005] was available. This prompted us to begin the second project (comparing sorghum and rice), which essentially used the same approach as described in the next section, but made use of a different, improved implementation.

## 5.3 Outline of the MicroSECTOR Approach

The microSECTOR approach takes two plant genomes as input, each as a FASTA file, plus a set of parameters. As output, it generates a set of microRNA candidates, for each of which it produces an overview document that summarizes valuable information for manual scrutiny. In addition, information about each stage of the program and information about each

candidate is stored in a result database. The different stages of the approach are:

- Structure prediction,
- Length and energy filter,
- Cover filter,
- Conservation filter,
- Sequence complexity filter,
- Repeat filter,
- Exon filter,
- Hairpin filter,
- Arm filter and microRNA$\star$ filter,
- Check for known microRNA,
- PSSM evaluation,
- Target type evaluation,
- Target conservation evaluation, and
- Structure evaluation.

We will discuss each stage in turn, using the poplar and *Arabidopsis* project as our example—for details, we refer to [Klug, 2005] and [Mayer, 2005]. Note that we used a reference set of published microRNAs to determine our filter parameters empirically, balancing sensitivity and specificity.

During the structure prediction stage, each of the two input genomes is subjected to structure prediction by RNALfold. This stage takes several weeks of computation on a dedicated server. Its output is a set of folds, each of which is a segment of the genome with its minimal free energy structure prediction. As any given nucleotide may participate in several minimal free energy structures, depending on the position of the sliding window, RNALfold generated more than eleven million folds for the *Arabidopsis* genome, many of which overlapped identical genomic loci.

### Filter Stages

As microRNA hairpins are structurally very stable, the following length and energy filter excluded all folds from further analysis for which the combined length and energy values did not satisfy an empirically devised criterion.

The cover filter removed all folds which covered any genome segment with an excessive overlap. This was useful, as we had determined that the folds containing our reference set of established microRNAs contained significantly less overlap than an average fold.

So far, each stage had been applied to each of the input genomes separately. The conservation filter takes the fold set from each genome and uses a suffix array-based procedure to determine which two folds, each from a different genome, shared a segment 21 nucleotides long with a maximal edit distance (also called "Levenshtein distance") of one. The rationale behind this is that the mature segment of a microRNA precursor is almost perfectly conserved, particularly between closely related genomes such as two dicots. We excluded all folds that did not satisfy this criterion and used the fold pairs (i.e. putative precursor pairs—one from each genome) determined in this step as the subjects of the following stages.

The sequence complexity filter discarded all putative precursor pairs for which the sequence of one of the partners was not of sufficient complexity as determined by an *ad hoc* procedure relying on nucleotide distribution. By discarding a large fraction of repetitive sequences, this filter was able to reject roughly 36% of the pairs available at this stage.

As we assumed that microRNAs are not located within an exon, the exon filter rejected all candidate pairs where one of the partners was significantly similar to a database of full-length cDNA.

The goal of the hairpin filter was to reject all precursor pairs for which the predicted structure did not form the stable hairpin that is characteristic of a microRNA. First, this filter used RNAfold to calculate the structure's partition function [Hofacker *et al.*, 1994] for each putative precursor which yields a matrix of base-pairing probabilities. We then integrated the matrix cells diagonally in order to determine a value which had to pass an empirically determined threshold in order for a pair to pass this filter.

All microRNA members of a family are located in the same arm (3' or 5') of their precursor. The arm filter therefore excluded all pairs in which the predicted microRNA (as determined by the conservation filter) was not located in the same arm. The microRNA$\star$ of established precursors is constrained to form bonds with its microRNA counterpart in only a limited number of ways within the predicted precursor structure. The microRNA$\star$ filter excludes all pairs that do not conform to an acceptable range of patterns from further analysis.

Next, all pairs which overlapped identical genomic segments were grouped together. In the *Arabidopsis* and poplar project, after all filters had been

applied, only 2187 distinct loci in *Arabidopsis* were left. We checked the relationship of each group to known microRNAs using sequence similarity as our criterion. This revealed that more than 70 of the 111 established *Arabidopsis* microRNA precursors (representing 23 out of 43 microRNA families) had passed all filters using our approach.

#### Evaluation Stages

The final stages evaluated each precursor group with the purpose of ultimately ranking all groups so that the most promising precursors would be subjected to manual scrutiny first. Each of the four evaluation stages yielded a numerical value which was ultimately integrated (using an empirical formula based on our reference set) into one number that determined the rank of this group.

We used a position-specific scoring matrix (PSSM, which is also called position-weight matrix) compiled from all known microRNAs and used it to evaluate each group's putative mature sequence. This procedure awarded a higher score to, for example, a thymine at the first position of the putative microRNA.

For the target type evaluation, we used targeting rules established by Rebecca Schwab (personal communication; later published in [Schwab *et al.*, 2005]) to find putative targets of each microRNA and then used the Gene Ontology database [Berardini *et al.*, 2004] to map each target to an associated molecular function. A candidate was assigned a higher score in this evaluation if its targets were homogeneously assigned to a narrow and non-ubiquitous range of associated functions such as transcriptional regulation.

We also evaluated whether the predicted targets in both input genomes were homologous—based on the rationale that homologous microRNAs would participate in regulatory pathways that have also been inherited from the most recent common ancestor.

Finally, we evaluated the difference in bond strength between the 5' end of the predicted microRNA and the 5' end of the corresponding microRNA$\star$—as microRNAs have been shown to exhibit a strand bias [Khvorova *et al.*, 2003].

## 5.4 Results

### 5.4.1 The Dicot Project: *Arabidopsis* and Poplar

The application of microSECTOR1 to *Arabidopsis* and poplar resulted in roughly 2000 *Arabidopsis* candidates and 3000 poplar candidates. We subsequently used the microHARVESTER tool, as outlined in Chapter 2, to filter these sets against the NCBI EST database. Using an EST database

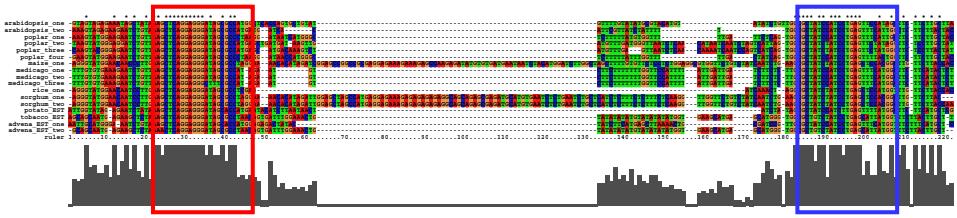


**Figure 5.1:** Alignment of miR390 homologs constructed using the ClustalW alignment software. This conservation profile, displayed in the lower part of the graph, is typical of authentic microRNAs: the mature microRNA segment (marked with a red box), is most conserved; second best for conservation is the microRNA⋆ segment (marked with a blue box) and the loop in between, and the segments upstream and downstream are least conserved. Sequences used for the alignment include two *Arabidopsis* paralogs and four poplar paralogs as well as genomic and/or EST sequences from maize, medick, potato, sorghum, tobacco, rice and spatterdock (*Nuphur advena*).

provides some evidence that the resulting refined set of microRNA candidates is actually expressed *in vivo*. We found that two sequence classes frequently passed through all filters, including the microHARVESTER step, although they are not microRNA-related: sequences coding for ribosomal RNA and common vector sequences. Obviously, ribosomal RNA is very conserved across the plant kingdom and thus easily meets the conservation criteria which we used as a filter. In addition, rRNA segments also adopt a stable hairpin structure with low free energy, which explains why some segments passed the filters applied by the microSECTOR approach. Vector sequences, on the other hand, are artificial sequences used for the amplification of DNA during the sequencing effort. Usually, vector sequences are removed after a DNA fragment is sequenced and before the genome is assembled. Nevertheless, in the *Arabidopsis* and poplar genomes, our method found putative microRNA candidates which, upon scrutiny, turned out to be remnants of the sequencing effort.

In the following sections, we look at a selection of interesting sequences contained in our result set.

### 5.4.2 A First Candidate: miR390

As we evaluated the resulting candidate set of microSECTOR1, we compared each candidate with the set of previously published microRNAs, as disseminated by the microRNA registry [Griffiths-Jones, 2004] at the Sanger institute. One sequence in particular stuck out when we applied the microHARVESTER to each candidate to determine potential homologs: we found its homologs in more than a dozen plant species. Furthermore, we also identified homologs in EST databases, which is an additional indication that this sequence is expressed and therefore is quite likely to be functional. In

addition, these homologs aligned well together and the conservation profile as displayed in Figure 5.1 was very similar to that of many other known microRNAs: the mature microRNA segment (left side of the profile), is most conserved; the second most conserved is the microRNA$\star$ segment (right side of the profile) and the loop in between and the segments upstream and downstream are least conserved.

Unfortunately, as we later discovered, our enthusiasm was premature: this sequence had previously been found by another group [Gustafson *et al.*, 2005; Adai *et al.*, 2005] and had not yet found its way into the microRNA registry. We could find some consolation in the confirmation this gave to our identification procedure.

### 5.4.3 The PUZZLING Candidate

We also found a most interesting sequence, which we coined "PUZZLING", in the *Arabidopsis* genome. It is located in positions 14344616–14344716 (TAIR version 6.0) of chromosome 4, which is just a few nucleotides downstream of the predicted 5' UTR of gene $AT4G29100.1$, a gene annotated as "ethylene-responsive family protein" with the additional annotation "contains similarity to ethylene-inducible ER33 protein (Lycopersicon esculentum) `gi|5669656|gb|AAD46413`". Judging from the gene's ORF, it is a basic helix-loop-helix (bHLH) domain containing transcription factor (Javier Palatnik, personal communication; [Morgenstern & Atchley, 1999]). The PUZZLING sequence is most probably located in the 3′UTR of this gene, although the 3′UTR is annotated in the TAIR database [Rhee *et al.*, 2003] to end a few nucleotides upstream.

A homolog of PUZZLING which is likewise capable of adopting a hairpin fold upon expression is also found in the 3′UTR of gene $AT2G20100.1$ (Javier Palatnik, personal communication). These two *Arabidopsis* genes are recent paralogs and are probably redundant (Detlef Weigel, personal communication). Their expression patterns are similar, with $AT4G29100.1$ expressed more highly. The highest expression levels are in stems, roots, and hypocotyls, which makes it possible that it can be used as a vascular marker (Detlef Weigel, personal communication). The sequence adopts a hairpin-like structure in the region of 80–100 nucleotides, a segment approximately 25 nucleotide long that is almost perfectly conserved across the plant kingdom. Figure 5.3 displays exemplary predicted folds of four homologs of different species. One can make the following observations, which apply to most validated microRNAs, in the PUZZLING sequence:

- the length of the crucial region, which would be the mature microRNA segment in an authentic microRNA, is conserved almost perfectly;

- the length of the loop region varies across species;

- the putative precursor adopts a hairpin structure outside of the crucial region as well as allowing compensatory mutations; and
- the segment predicted to pair with the crucial region in the foldback is less conserved than the crucial region itself.

Using the microHARVESTER, we could detect homologs in 15 different species within the NCBI EST database, as listed in Table 5.1. Interestingly, only two of these species are monocotyledons; the other 13 are dicotyledons. Rebecca Schwab from Detlef Weigel's laboratory performed a range of experiments, aiming to determine whether this sequence is a microRNA. Although PUZZLING seems to exhibit several typical microRNA-specific features, its processing into a small RNA could not be shown.

As part of her experiments, Rebecca cloned the PUZZLING sequence into a binary vector. Then, using agroinfiltration, *Nicotiana benthamiana* leaves were transiently transfected by *Agrobacteria tumefaciens* cultures harboring the generated vectors. Performing small RNA Northern blots, using radiolabeled oligos of a sequence antisense to the predicted microRNA, yielded a negative result for the processing of the PUZZLING sequence. In addition, the public database of small RNA MPSS signals [Nakano *et al.*, 2006] in *Arabidopsis* contains no small RNA sequences that come from this locus.

If the PUZZLING sequence is not a microRNA, what else could it be? Its remarkable ubiquity across the plant kingdom, the extent and pattern of conservation, and its predicted foldback structure make it highly unlikely that this is a chance phenomenon. In addition, its location in the 3′UTR of a bHLH transcription factor may support the conjecture that this sequence is functional and subject to evolutionary pressure. Since the bHLH domain is able to unwind RNA duplexes, an adventurous thought would be that the protein could be involved in unwinding its own 3′UTR and would thus regulate itself; in addition, the gene is very probably differentially spliced and the foldback in the 3′UTR could play some role here, too (Detlef Weigel, personal communication). Unfortunately, we can only pose these questions in the context of this work and not present their answers yet. Further experiments will be needed to shed light on the purpose and mechanism of the PUZZLING sequence.

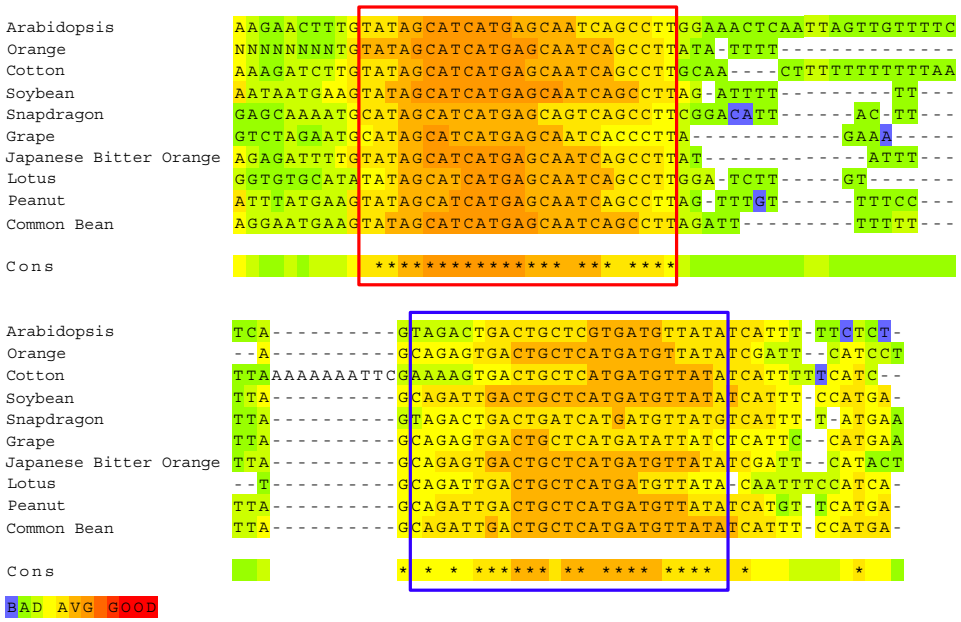



**Figure 5.2:** Alignment of EST sequences, which are homologs of PUZZLING, from different species. The two conserved blocks are marked in red and blue. NCBI accession numbers and scientific species names are as follows (ordered as in the alignment): gi|42528730|emb|BX834882.1| (*Arabidopsis thaliana*), gi|34524853|gb|CF509669.1| (*Citrus sinensis*), gi|48805921|gb|CO107235.1| (*Gossypium raimondii*), gi|23053843|gb|BU577597.1| (*Glycine max*), gi|51058961|emb|AJ789999.1| (*Antirrhinum majus*), gi|33409609|gb|CF215236.1| (*Vitis vinifera*), gi|57876300|gb|CX641471.1| (*Poncirus trifoliata*), gi|45637515|dbj|BP080854.1| (*Lotus japonicus*), gi|30420018|gb|CD038180.1| (*Arachis hypogaea*), gi|59937050|gb|CB542381.1| (*Phaseolus vulgaris*).

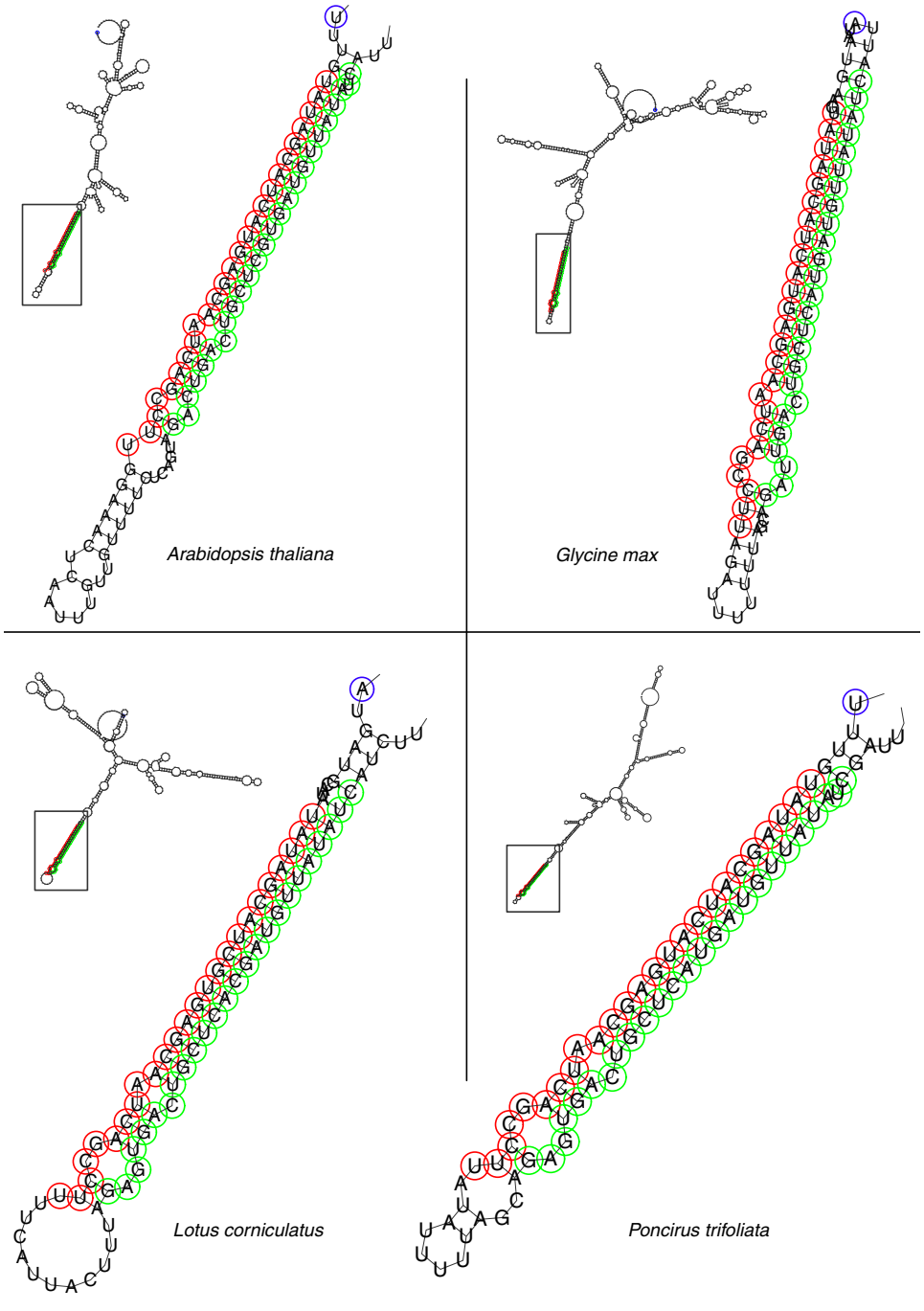



**Figure 5.3:** Predicted folds of homologs of PUZZLING as discussed in Section 5.4.3. Accession numbers for these sequences are as follows: *Arabidopsis thaliana* (gi|501997|gb|T20556.1|), *Glycine max* (gi|31464669|gb|CD406682.1|), *Lotus corniculatus var. japonicus* (gi|45411427|dbj|BP034267.1|), *Poncirus trifoliata* (gi|57876300|gb|CX641471.1|). Note that the bulges at both ends of the predicted mature segments are also conserved. The 5' end of each RNA sequence, its most conserved segment and the corresponding binding segment are marked in blue, red and green, respectively.

| Species | NCBI EST Accession |
|---|---|
| *Arabidopsis thaliana* | gi\|501997\|gb\|T20556.1\| |
| *Antirrhinum majus* | gi\|51058961\|emb\|AJ789999.1\| |
| *Arachis hypogaea* | gi\|30420018\|gb\|CD038180.1\| |
| *Citrus paradisi x Poncirus trifoliata* | gi\|57926697\|gb\|CX667982.1\| |
| *Citrus sinensis* | gi\|34524853\|gb\|CF509669.1\| |
| *Glycine max* | gi\|31464669\|gb\|CD406682.1\| |
| *Hordeum vulgare* | gi\|59945723\|gb\|DN159829.1\| |
| *Lactuca sativa* | gi\|22235299\|gb\|BQ849830.1\| |
| *Lotus corniculatus var. japonicus* | gi\|45411427\|dbj\|BP034267.1\| |
| *Malus x domestica* | gi\|51237894\|gb\|CO898104.1\| |
| *Medicago truncatula* | gi\|9682265\|emb\|AL382514.1\| |
| *Phaseolus vulgaris* | gi\|59937050\|gb\|CB542381.1\| |
| *Poncirus trifoliata* | gi\|57876300\|gb\|CX641471.1\| |
| *Sorghum bicolor* | gi\|57807075\|gb\|CX608355.1\| |
| *Vitis vinifera* | gi\|33409609\|gb\|CF215236.1\| |

**Table 5.1:** This table lists one sample homolog per species for the PUZZLING sequence as discussed in Section 5.4.3. Each of these sequences is contained in the NCBI EST database and thus expressed. Note that only two species (*Hordeum vulgare* and *Sorghum bicolor*) are monocotyledons, while all others are dicotyledons.

### 5.4.4 The RESISTANT Candidate

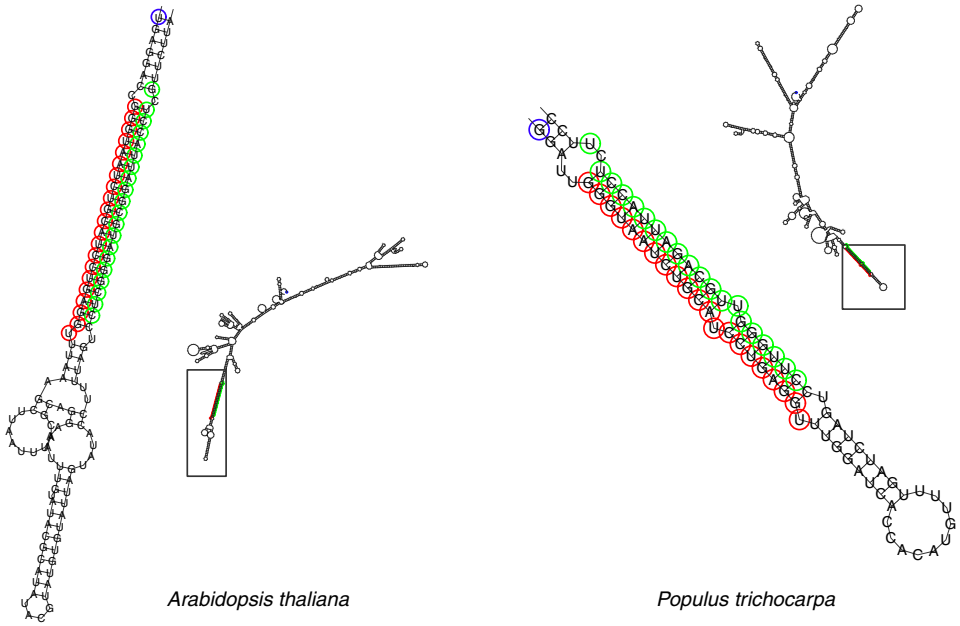
Another interesting result is a microRNA precursor candidate, which we named "RESISTANT". It is located on chromosome 3 of *Arabidopsis*, covering positions 2854307–2854440 (watson strand) in TAIR version 6 and has the following sequence: GAGGACC GGGTAATCTGCATCCTGAGGT TTAAAGCTTAATTTACGCAGGAAATTTGTATACGCATATACGTATGTGTATTAGTATACCTTTTAGTC CTCGGGATGCGGATTACCTCG TTCTTACTTACAATACA (putative microRNA and microRNA⋆ are highlighted in red and green, respectively).

The predicted secondary structure of RESISTANT, as shown in Figure 5.4(a), resembles that of established microRNA precursors. The RESISTANT precursor is located on the watson strand of the genome, between gene AT3G09280.1 and gene AT3G09290.1, and is upstream of both genes, since AT3G09280.1 is transcribed from the crick strand of the genome and AT3G09289.1 is transcribed from the watson strand. Gene AT3G09280.1, which is roughly 4000 nucleotides upstream of RESISTANT, is scantily annotated as being an "expressed protein" (expression supported by MPSS), and gene AT3G09290.1, which is roughly 1800 nucleotides downstream of RESISTANT, codes for a zinc finger (C2H2 type) family protein. We could determine one other sequence homologous to RESISTANT in *Arabidopsis* and four homologs in poplar—but could not find any homologs of this sequence in any other plant genome.

Benjamin Czech, Rebecca Schwab, Heike Wollmann and Felipe Felippes from the Weigel lab performed a number of experiments involving the RESISTANT precursor, in an attempt to find out whether it is indeed a microRNA precursor. First, they amplified the genomic segment containing the RESISTANT precursor by PCR and cloned it into a binary vector. Then, using agroinfiltration, *Nicotiana benthamiana* leaves were transiently transfected by *Agrobacteria tumefaciens* cultures harboring the generated vectors. Performing small RNA blots using radiolabeled oligos of sequence ACCTCAGGATGCAGATTACCC, which is antisense to the predicted microRNA, they could detect evidence for processing of the mature sequence as predicted, both in *Nicotiana benthamiana* leaves (see Figure 5.4(b)) and also in a follow-up experiment, in which *Arabidopsis* plants were transformed to contain this segment under the control of a strong constitutive (CaMV 35S) promoter.

The public database of small RNA MPSS signals [Nakano *et al.*, 2006], however, does not register any small RNA for the 17mer signature GGGTAATCTGCATCCTG —which one might expect for an at least moderately expressed microRNA. We could predict one target gene in *Arabidopsis* for this putative microRNA: *AT*1*G*07010, using the WMD tool of Schwab *et al.* [2006]. This target is a calcineurin-like phosphoesterase family protein. However, the target site is not conserved in its poplar ortholog and therefore further scrutiny—possibly involving DICER-LIKE1 mutants—is necessary.

(a)




(b)

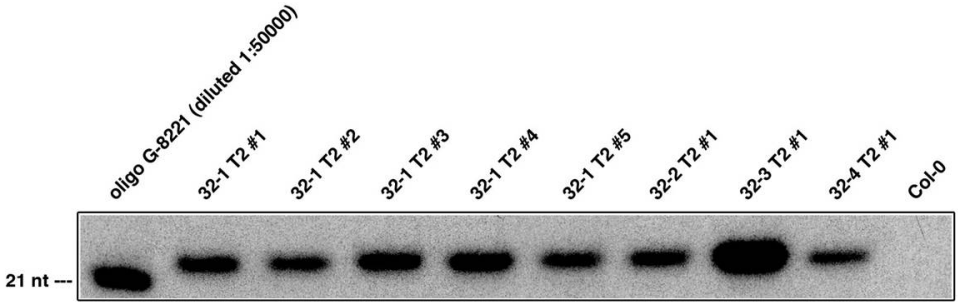



**Figure 5.4: (a)** Predicted folds of the *Arabidopsis* and poplar homologs of the RESISTANT sequence discussed in Section 5.4.4. Accession numbers for these sequences are as follows: *Arabidopsis thaliana* (gi|501997|gb|T20556.1|), *Populus trichocarpa* (gi|45411427|dbj|BP034267.1|), *Poncirus trifoliata* (gi|57876300|gb|CX641471.1|). The 5' end of each sequence, its microRNA segment and the corresponding microRNA$\star$ are marked in blue, red and green, respectively. **(b)** This blot shows that small RNAs which correspond to the predicted microRNA of the RESISTANT precursor are processed in the leaves of transgenic *Arabidopsis* plants under a constitutive promoter. The far left lane is for control and the far right lane used wild-type RNA. All other lanes show RNA harvested from the leaves of stable RESISTANT-transfected *Arabidopsis* lines.

### 5.4.5 The Monocot Project: Sorghum and Rice

The application of microSECTOR2 to the sorghum and rice genomes yielded roughly 10 000 microRNA precursor candidates in the sorghum genome and 70 000 candidates in the rice genome [Mayer, 2005]. One reason that the number of rice candidates was several times larger than that of sorghum candidates might be that the genomic rice sequence we used as input was 480 MB in size and the genomic sorghum sequence was only 150 MB in size. In addition, we noted that in many cases a particular candidate in sorghum was associated with several similar sequences in rice. This might indicate that many of these sequences were associated with repetitive sequences.

We applied the microHARVESTER to the sorghum candidates in conjunction with the NCBI EST database as described in Section 5.4.1. A quarter of the candidates, 2729 precursors, passed this step as roughly 75% of the candidates were rejected. We ranked these candidates by the number of species for which each candidate had predicted homologs in and manually scrutinized the top 700 candidates.

There were many candidates which exhibited the typical features of microRNAs: they were predicted to form microRNA-like hairpins in several species (at least in rice and sorghum) with near perfect conservation of the mature segment and the typical conservation pattern of microRNA precursors; they are expressed, as they occur in EST sequences—in many cases, even the pattern of bulges within the stem is conserved across species. To meet the crucial criterion, the experimental validation that a small RNA is processed from a putative precursor sequence *in vivo*, we compiled a small set of the 13 most promising sequences (See Table 5.2 and Figures C.2,C.3,C.4 and C.5).

Our collaborator, Ramanjulu Sunkar from the laboratory of Jian-Kang Zhu at the University of California, Riverside, then performed a number of wet lab experiments. First, he extracted and concentrated RNA from rice plants. Then, he fractionated the small RNA content of the RNA and performed Northern blots using radioactively labeled oligonucleotides with the complementary sequences to our microRNA candidates. If a Northern blot displayed a small RNA signal for a particular candidate sequence, he additionally performed Northern blots for this sequence using *Arabidopsis* and maize RNA. The results of these experiments for the 13 candidate sequences is displayed in Table 5.2. As shown, four of these candidate sequences yield a significant signal for the predicted small RNA sequence.

Prompted by the very large number of predicted microRNA candidates in rice, we wanted find out whether any of our 13 selected candidates had any relationship to a known repetitive sequence. Therefore, we used BLAST to find similarity matches in the "*Oryza* Repeat DB" hosted at The Institute for Genomic Research (TIGR) at `http://www.tigr.org/tdb/e2k1/osa1/blastsearch.shtml`.

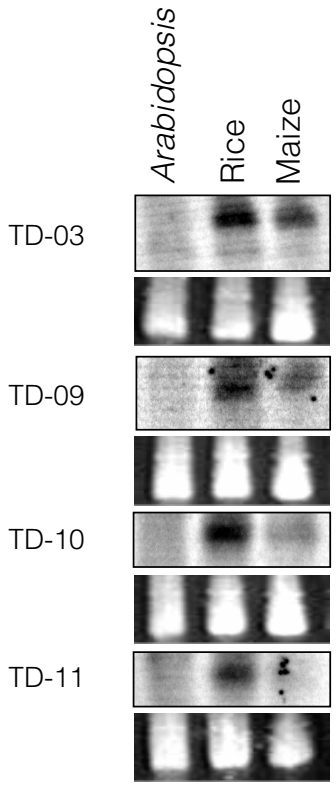



**Figure 5.5:** RNA gel blots of total RNA isolated from different tissues were probed with labeled oligonucleotides (as explained in the text) for selected candidate sequences TD-03, TD-09, TD-10 and TD-11. The tRNA and 5S rRNA bands were visualized by ethidium bromide staining of polyacrylamide gels and serve as loading controls.

We discovered that one of our candidates, TD-3, which yielded a positive blot (see Table 5.2) was significantly related to miniature inverted-repeat transposable elements (MITEs) [Feschotte *et al.*, 2002; Jiang *et al.*, 2004]. To our knowledge, no connection between MITEs and microRNAs has been reported in the literature so far.

Since there are currently no DICER-LIKE mutants available for rice, we are unable at present to experimentally establish whether these microRNA candidates are authentic microRNAs or not. These sequences do, however, display the typical pattern of conservation across the monocot clade: a microRNA-typical hairpin folding of their putative precursors. Further experiments exploring this question are ongoing and beyond the goal of this work.

(a)

| Candidate | Sequence | Ath | Rice | Maize |
|---|---|---|---|---|
| TD-01 | TTATAAGTCACTTTGACTTTTTT | n.a. | − | n.a. |
| TD-02 | TTCCCGAGCTCCTCGTCGTTGCGG | n.a. | − | n.a. |
| TD-03 | TTATAATTTGGAACGGAGGGAGTA | − | + | + |
| TD-04 | TTGATGTGCATACACCGCATG | n.a. | − | n.a. |
| TD-05 | TCCGTTTTACAATATAAGTCATT | n.a. | − | n.a. |
| TD-06 | TTATAAGTTGCTTTGACTTTT | n.a. | − | n.a. |
| TD-07 | TGTCCATAGCCACCATAGT | n.a. | − | n.a. |
| TD-08 | TGAAGTGTTTGGGGGAACT | n.a. | − | n.a. |
| TD-09 | TTAAAAAGGAACGGAGGGAG | − | + | + |
| TD-10 | TTATGGGACGGAGGGAGTA | − | + | (weak) |
| TD-11 | TTTGGTGGAGCAATGGGTGTAT | − | + | + |
| TD-12 | TACTCCCTCTGTCCCAAAATA | n.a. | − | n.a. |
| TD-13 | TGAATAAGACGAGTGATCAAA | n.a. | − | n.a. |

(b)

| Candidate | Possible originating loci | Annotation |
|---|---|---|
| TD-03 | Chrom. 04, pos. 19260001(W) | intergenic |
| TD-09 | Chrom. 09, pos. 11460472(C) | intergenic |
| | Chrom. 09, pos. 11430472(C) | intergenic |
| TD-10 | Chrom. 12, pos. 27011219(C) | intergenic |
| | Chrom. 12, pos. 26918904(W) | intergenic |
| | Chrom. 12, pos. 24457858(C) | intergenic |
| | Chrom. 12, pos. 22910561(C) | intergenic |
| | Chrom. 12, pos. 22594864(C) | intergenic |
| TD-11 | Chrom. 05, pos. 20468374(C) | inside of OS05G34650 |

**Table 5.2: (a)** The results of validation experiments for each of the 13 microRNA candidate sequences are shown in this table. The plant descriptor columns either states "+" or "−" to show whether this particular experiment has been performed with a positive or negative result ( i.e. small RNAs could be detected or not), or "n.a." if this experiment was not performed since the corresponding rice experiment was negative. **(b)** Possible originating loci are given for each small RNA that we could validate. Strand (Watson/Crick) is indicated in parentheses. All possible originating loci, except the single one for TD-11, are located in intergenic regions. TD-11 originated from a locus from which the gene OS05G34650 is transcribed in the same direction. This gene, OS05G34650, is solemnly annotated as "expressed gene". All positions have been determined using the *Rice Functional Genomic Browser* available at the the Salk Institute Genomic Analysis Laboratory (SIGnAL) website (`http://signal.salk.edu/`) pertaining to TIGR V4 pseudomolecules.

## 5.5 Discussion

In this chapter, we have first briefly sketched an approach for the comparative prediction of microRNAs that uses a pair of plant genomes as input. Next, we have reported on the results of the application of this approach to two pairs of genomes: *Arabidopsis* and poplar, and sorghum and rice. In both projects, many established microRNAs were contained in the resulting candidate set—a fact indicative of the effectiveness of this approach.

In the *Arabidopsis* and poplar project, we report on three promising candidates: one turned out to be an established microRNA (which was added to the microRNA Registry while our work was in progress). A second sequence is puzzling us, as it shows features typical of a microRNA, although our experiments determined that it is not processed into a small RNA. Nevertheless, the discovered sequence seems very interesting to us because of its widespread distribution across the plant kingdom, its perfectly conserved 24 nucleotide segment, and its conspicuous secondary structure. Our third microRNA candidate exhibits many features characteristic of microRNAs, and experimental lab work showed that transgenic plants hosting this candidate equipped with a constitutive promoter do produce small RNAs as predicted. Nevertheless, more lab work is needed (possibly involving DICER-LIKE1 mutants) to establish firmly that this candidate is indeed a new microRNA.

In the sorghum and poplar project, we report on 13 promising candidates. For four of these, we could validate processing into small RNAs in rice and maize wild-type plants—but not in *Arabidopsis*. At least one of our candidates, TD-03, shows similarity to a miniature inverted-repeat transposable element (MITE). This poses interesting questions, since no relationship between MITEs and microRNAs have been known so far. Further scrutiny is required and is ongoing.

# Chapter 6

# Prediction Based on MPSS Expression Data

## 6.1 Motivation

Massively parallel signature sequencing (MPSS) [Brenner *et al.*, 2000] is a novel sequencing technology that combines simultaneous cloning of millions of DNA fragments with non–gel–based signature sequencing and usually yields hundreds of thousands of sequence tags that are 16–20 nucleotides long. Among other applications, MPSS has previously been used for a whole–genome transcriptional analysis [Meyers *et al.*, 2004] of *Arabidopsis thaliana*. Very recently, Green and colleagues have adapted the MPSS technology to small RNA molecules and sequenced more than 2 million small RNAs from an *Arabidopsis* inflorescence and two seedling libraries [Lu *et al.*, 2005], which they made available publicly [Nakano *et al.*, 2006]. Each of the obtained signatures is 17 nucleotides long so most of these only match a few locations in the *Arabidopsis* genome and some can be unambiguously mapped. In total, 100 452 non–redundant signatures were contained in the 2 million reads.

Matching these signatures to all published microRNAs, Lu *et al.* [Lu *et al.*, 2005] found that signatures exactly matching 73 microRNAs accounted for $\approx 40\%$ of the total abundance of genome–matched signatures from the inflorescence library, and 72 known microRNAs accounted for $\approx 62\%$ of the seedling library derived signatures.

We decided to use this dataset of small RNA MPSS signatures to predict new microRNAs.

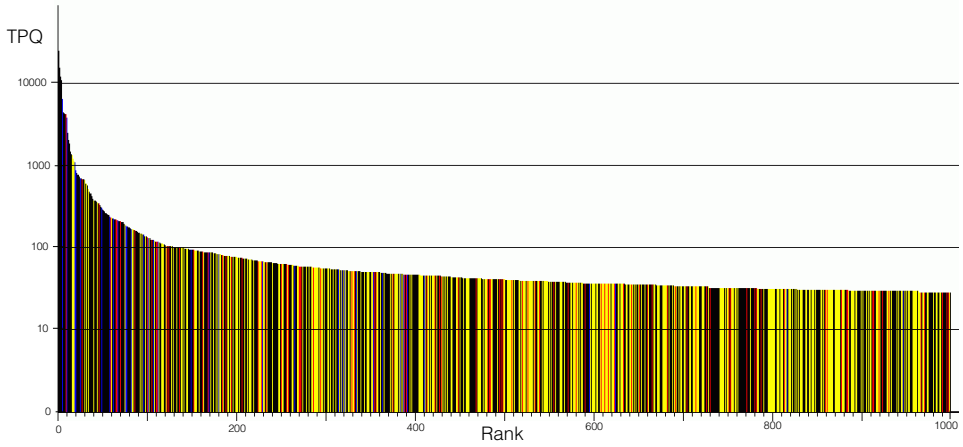



**Figure 6.1:** Histogram showing abundance of MPSS signatures vs. rank. X–axis: rank of MPSS signature by decreasing abundance. Y–axis: abundance measured in TPQ. Each bar is colored according to the classification of the corresponding signature which is determined by its similarity to published mature microRNA sequences, microRNA precursors and repetitive sequences (see text): black: class "known microRNA"; red: class "known precursor"; blue: class "repetitive sequence"; yellow: class "unknown sequence".

## 6.2 Results

### 6.2.1 Analysis of the MPSS Tag Set

The abundance of MPSS tags is measured in "transcripts per quarter million" (TPQ). In our tag set, the most abundant tag was detected with 91 365 TPQ in one library and the least abundant with only 1 TPQ. In the following, we used the maximal abundance of each tag in any of the three libraries (one inflorescence and two seedling libraries) as the relevant abundance.

Since we suspected that the abundance of an arbitrary tag correlates with the probability of it being derived from a microRNA, we decided to first order all MPSS tags by their abundance and then assign each to one of four classes, based on similarity searches against each of these sequences:

- Class "known microRNA": all plant microRNA precursor sequences hosted at the Sanger registry.
- Class "known precursor": all plant microRNA mature sequences hosted at the Sanger registry.
- Class "repetitive sequence": all TIGR *Arabidopsis* repeat sequences.

For this classification, we used BLAST to determine the E–value of each tag against each of these sequence sets, using a cutoff of 0.001. In case

there was no significant similarity with any of these databases, we assigned the tag to the class "unknown sequence". In case a tag was associated with several of the above databases, we classified it to a class according to its first occurrence in the following order: "known microRNA", "known precursor", "repetitive sequence".

Figure 6.1 shows a bar chart of the 1000 most abundant signatures, ranked by abundance. As can be seen, the most abundant signatures were microRNA–associated: the nine tags with highest rank were derived (in order of decreasing abundance) from members of the microRNA families miR167, miR169, miR170, miR166, miR390, miR157, miR160 miR161 and miR168. Of these, only the tag on rank six was assigned to the class "precursor" and was possibly a degradation product of a microRNA precursor of family miR157—all other tags were identical or very similar (e.g. shifted by 1 nt) to the corresponding mature microRNAs. On rank 11, we found the highest ranked tag that was derived from a repetitive sequence (denoted "ARSgRGRR00000002" in the TIGR database of repetitive sequences) which was therefore assigned to class "repetitive". Note that only 20 tags had an abundance greater than 1000 TPQ and only 140 tags had an abundance greater than 100 TPQ. Of these 140 tags, 47 were derived from known mature microRNAs.

Next, we wanted to discover where MPSS tags mapped in relation to all published microRNA precursors. We found that, in most cases, several MPSS tags mapped onto each precursor, especially in the case of high-abundance microRNA families such as e.g. miR166 and miR169. One can observe the pattern that the MPSS tag with highest abundance maps exactly with the first 17 nucleotides of the mature microRNA and other tags map shifted one or two nucleotides up and downstream. Additionally, tags map with the microRNA$\star$ sequence in many cases—at a significantly lower abundance than with the microRNA—and often a similar spread of the tags can be observed on the microRNA$\star$ as on the microRNA.

### 6.2.2 Prediction of New MicroRNA Candidates

Since so many of the high–abundance MPSS signatures seemed microRNA–associated, we speculated that a significant fraction of the high–abundance signatures which are neither associated with known microRNAs nor with repetitive sequences might be derived from as yet unidentified microRNAs.

Therefore, we decided to start with the MPSS tag set and filter this down to a small candidate set for experimental verification. As outlined on the flowchart in Figure 6.2, we began with all 100 452 MPSS sequence tags from the small RNA database and removed all signatures with an abundance of less than 15 TPQ, resulting in 7582 tags. Then we tried to map each of these tags onto the genome and counted the number of possible originating loci. We removed a sequence if we could not map it onto the genome or if we

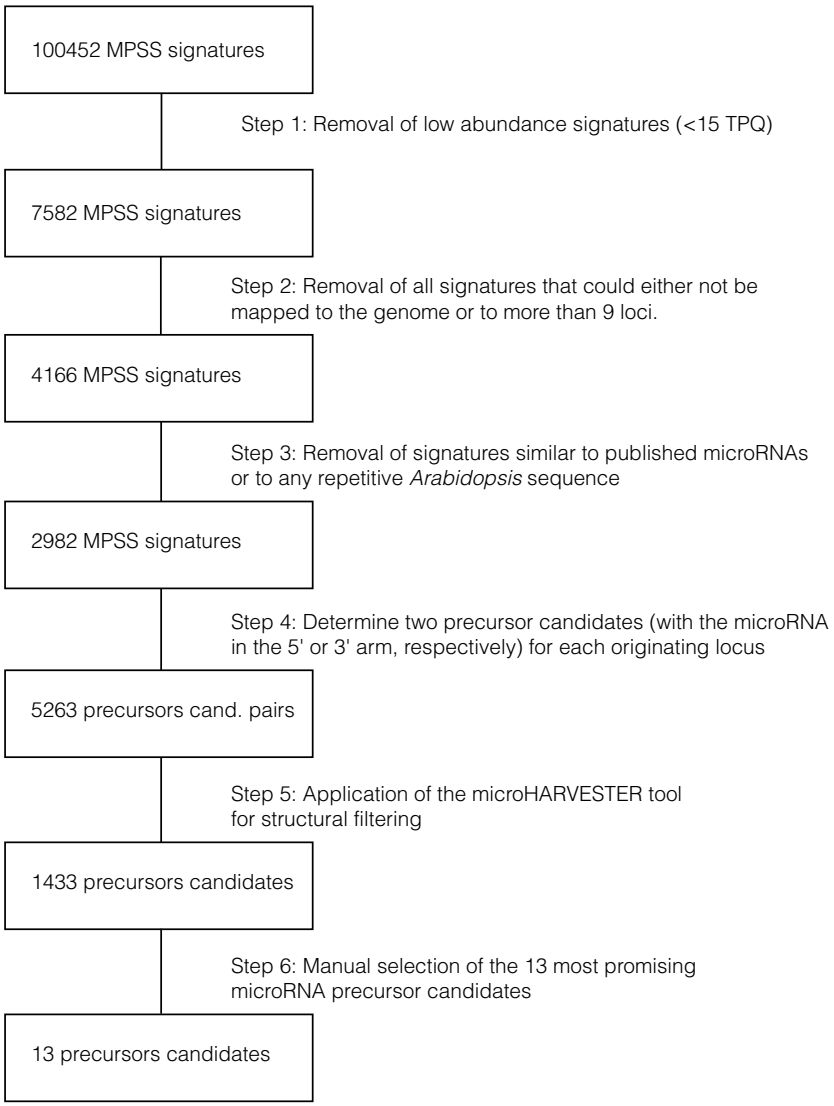



**Figure 6.2:** Flowchart of our prediction approach based on the MPSS sequence data. Essentially we begin with MPSS sequences, which we filter as detailed in Section 6.2.2. Then, we focus on pairs of precursor candidates as the result of filtering at Step $4$ and finally on single precursor candidates from Step 5 onwards.

found more than nine possible originating loci—4166 sequences passed this step. Next, we determined for each sequence its similarity to any published microRNA precursor or repetitive sequence using BLAST. We removed all sequences which were similar to a sequence in any of these databases with a cutoff E–value of 0.1; 2982 sequences passed this test.

For each possible originating locus of each remaining sequence, we extracted two preliminary microRNA precursor candidates from the genome: one, in case the mature microRNA (derived from the tag) would be located in the $5'$ arm of the precursor and the other, in case it would be located in the $3'$ arm of the precursor. For this, we extended the putative microRNA–matching locus 20 nucleotides to one side of the microRNA and 650 nucleotides on the other side. This procedure resulted in 5263 microRNA precursor candidate pairs.

In the next step, we used each microRNA precursor candidate together with the 21 nucleotide putative mature microRNA segment as input to the microHARVESTER2 server as detailed in chapter 2, using default settings except that we allowed up to six mismatches between mature and microRNA$\star$ segments and thus increased sensitivity at the price of additional false positives. This procedure effectively imposed the structural constraints observed in published microRNAs onto our candidates. A total of 1433 precursor candidates passed the test applied using the microHARVESTER.

After this, PDF overview documents showing the putative RNA folding structure were generated for each of the precursor candidates. We manually inspected each document and selected 13 precursor candidates for further analysis, which we labeled consecutively from mpss01 to mpss13. Our primary selection criteria were: strength of expression (TPQ), a preference for a thymine at the first position, as few originating loci in the genome as possible and the foldback quality of the predicted RNA folding structure.

### 6.2.3 Experimental Validation of MicroRNA Candidates

To determine whether these candidates could actually generate small RNAs, Felipe Felippes from the Weigel lab performed the following experiments: The genomic sequences containing the precursors for 11 of our 13 candidates (cf. Table 6.1(a); the remaining two candidates were resistant to PCR amplification) were amplified by PCR and cloned in binary vectors under the control of the Cauliflower Mosaic Virus promoter CaMV 35S, a strong and constitutive promoter in plants. Next, *Nicotiana benthamiana* leaves were transfected by *Agrobacterium tumefaciens* cultures harboring the generated binary vectors using agroinfiltration and, four days later, the leaves were harvested and their RNA extracted. Performing small RNA blots using radiolabeled oligos, Felipe Felippes could detect evidence for expression and processing for the following 5 (out of 11) candidates: mpss01, mpss02, mpss05, mpss07 and mpss11—as detailed in Table 6.1(b).

(a)

| Candidate | MPSS tag | TPQ | #loci | cloned | processed |
|---|---|---|---|---|---|
| mpss01 | TTGGTTACCCATATGGC | 106 | 1 | yes | yes |
| mpss02 | TCATGGTCAGATCCGTC | 97 | 1 | yes | yes |
| mpss03 | GGTGAACGACCTGTGTC | 50 | 9 | yes | no |
| mpss04 | TTCACTACCGAACGATT | 75 | 2 | no | n.a. |
| mpss05 | TGGCCTTGTCATCTCAA | 67 | 1 | yes | yes |
| mpss06 | TGGTCGTGATCTACTGG | 62 | 1 | yes | no |
| mpss07 | TCGGCTCAGGACCATTG | 82 | 1 | yes | yes |
| mpss08 | TACCAACCTTTCATCGT | 168 | 1 | yes | no* |
| mpss09 | TTGGCTTCTACCGCAAG | 154 | 1 | yes | no |
| mpss10 | TTGACGGAATTGTGGCG | 120 | 1 | yes | no |
| mpss11 | TGCGGGAAGCATTTGCA | 589 | 1 | yes | yes |
| mpss12 | CTTCATCGCAATGGCTA | 58 | 1 | yes | no |
| mpss13 | TCAACTCCAGGATTGGA | 114 | 1 | no | n.a. |

(b)

| Candidate | microRNA | originating locus |
|---|---|---|
| mpss01 | TTGGTTACCCATATGGCCATC | Chrom. 1, intergenic, between AT1G60070.1 and AT1G60075.1 |
| mpss02 | TCATGGTCAGATCCGTCATCC | Chrom. 1, intergenic, between AT1G61215.1 and AT1G61230.1 |
| mpss05 | TGGCCTTGTCATCTCAACCGT | Chrom. 1, within an intron of AT1G4410.0, an Amino acid permease |
| mpss07 | TCGGCTCAGGACCATTGCGGT | Chrom. 1, intergenic, between AT1G67480.1 and AT1G67490.1 |
| mpss11 | TGCGGGAAGCATTTGCACATG | Chrom. 5, intergenic, between AT5G03550.1 and AT5G03555.1 |

**Table 6.1: (a)** shows the following attributes for each of our microRNA candidates: MPSS tag sequence; *TPQ*: the abundance in transcripts per quarter million molecules; *loci*: the number of loci this tag can be mapped to in the *Arabidopsis* genome; *cloned*: whether we could clone one of these loci; *processed*: whether the construct was shown to be processed in *Arabidopsis*, using a viral promoter as detailed in the text. (*): We could not provide processing evidence for candidate mpss08. However, an identical sequence is contained in the ASRP database (see text). **(b)** shows the putative microRNA sequence (with some additional precursor nucleotides in gray) and the genomic segment of the originating locus for each candidate that could be shown to be processed.

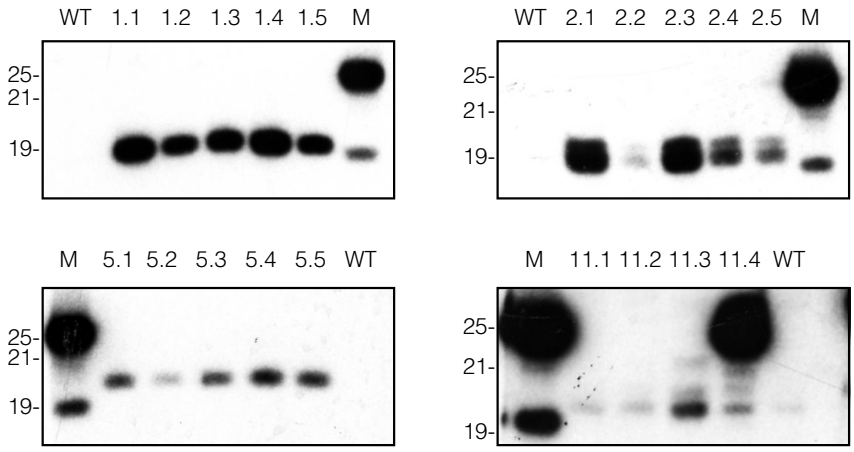



**Figure 6.3:** Blots of sequences mpss01, mpss02, mpss05 and mpss11 after transformation into *Arabidopsis* plants using agrobacteria.

Next, we wanted to know whether these candidates could also be processed in *Arabidopsis* and which phenotypes would result. Consequently, Felipe Felippes generated *Arabidopsis* transgenic plants for each of the 5 constructs. Processing of all 5 constructs could be validated in the transgenic plants—corresponding blot results for mpss01, mpss02, mpss05 and mpss11 are shown in Figure 6.3. Unfortunately, no conspicuous phenotypes ensued from this experiment.

### 6.2.4 Prediction and Validation of MicroRNA Targets

We used the target search facility of the Web MicroRNA Designer (WMD) web site [Schwab *et al.*, 2006] to determine potential target genes for each of those five microRNA candidates, for which we could validate processing to small RNAs. Since the predicted target of mpss11 is weakly expressed in wild type plants and the WMD tool predicted no protein-coding target for mpss07, we focussed on the WMD-predicted targets of the remaining microRNA candidates as shown in Table 6.2. As indicated, Felipe Felippes could verify experimentally in a plant that for each of these precursor constructs, at least one corresponding target is cleaved. Furthermore, he could show by RT–PCR that the target RNA in each case is underregulated in these transfected plants with respect to the wild type.

| Candidate | Target accession / Protein family | Cleavage |
|---|---|---|
| mpss01 | AT3G19890 / F–box | yes |
| | AT3G17490 / F–box | no |
| mpss02 | AT3G43610 / Tubulin | no |
| | AT1G60130 / Jacalin lectin | yes |
| | AT1G57570 / Jacalin lectin | no |
| | AT3G63400 / Peptidyl–prolyl cis–trans isomerase | no |
| | AT5G38550 / Jacalin lectin | yes |
| | AT1G62750 / Elongation factor Tu | no |
| | AT2G37340.2 / Splicing factor RSZ33 | no |
| | AT1G19570 / Dehydroascorbate reductase | no |
| mpss05 | AT1G43130 / Expressed Protein | yes |

**Table 6.2:** This table lists each predicted target of the three processed candidates with promising protein–coding targets along with the results of the *in vivo* experiments that established whether these targets are cleaved.

### 6.2.5 Evolution of MicroRNA Genes

In a landmark paper, Edwards Allen, James Carrington and colleagues proposed the inverted duplication hypothesis for the evolution of genes encoding microRNAs [Allen *et al.*, 2004]:

> Loci capable of forming a transcript that adopts an extended foldback structure can arise by inverted duplication events. If the originating sequence is a protein–coding gene, then the originating gene and closely related family members could be brought under negative regulation by RNAi through short interfering RNAs (siRNAs) spawned at the duplication locus. If sequences at the duplication locus diverge under constraints to maintain a foldback structure and adapt to the miRNA biogenesis apparatus, then the new locus might evolve into a miRNA gene with specificity for one or more targets related to the founder gene.

This hypothesis predicts that the foldback arms of microRNA genes are initially very similar to their originating loci which—together with related family members—become the target of the microRNA. It can be expected that this similarity will be most prominent for recently evolved microRNA genes and will decrease in the course of evolution due to mutations in both the originating locus and the microRNA gene. This similarity might become undetectable, i.e. decrease to the similarity expected by chance for this pair of sequences, if the duplication event has occurred in the distant past.

Allen *et al.* wanted to analyze, which of the published microRNA genes similarity to a target gene could be determined for. For this, they ran FASTA

[Pearson & Lipman, 1988] searches using each microRNA as a query against the set of all protein–coding *Arabidopsis* sequences. They found a significant similarity between a microRNA gene arm and its protein–coding target gene only for microRNAs of the families miR161 and miR163 [Allen *et al.*, 2004]. To evaluate the significance of this similarity, they repeated this search 1000 times using shuffled foldback arm sequences as their query.

### An Alternative Route to MicroRNA Evolution

Folding the *Arabidopsis* genome *in silico* generates hundreds of thousands of foldbacks that resemble those of microRNA precursors. By chance, some of these might be captured by transcriptional regulatory sequences and subsequently be expressed and processed by the microRNA biogenesis machinery, yielding microRNAs that either target no other transcript or a transcript unrelated to the foldback except at the $\approx 21$ nucleotide binding site. If, on one hand, the expression of this nascent microRNA would not confer a selective advantage to the host plant it would eventually cease to be expressed due to acquired mutations. If, on the other hand, its expression would by chance regulate a group of genes in a beneficial way it would be stabilized, fixated and fine–tuned [Bartel & Chen, 2004] through co–evolution with its target(s). We propose that this could be an alternative path to microRNA evolution that would explain why Allen *et al.* were only able to detect significant similarity between two published microRNAs and their protein–coding target genes.

### Similarity of MicroRNA Candidates to Their Target Genes

To find evidence for either of the microRNA gene evolution scenarios mentioned above, we devised a procedure to evaluate the significance of the similarity between the microRNA gene and the corresponding target gene. For each microRNA precursor/target pair, we performed the following:

- We aligned the microRNA to the reverse–complemented target gene. This divides the target gene into three segments, which we will call “5prime”, “microBinder” and “3prime”. The microBinder is the segment which binds to the microRNA; 5prime and 3prime are the segments upsteam and downstream of this segment.

- We determined the predicted folding structure of the precursor and thus determined the sequence that folds opposite the microRNA. We will call this segment “star” throughout this section, although technically the microRNA$\star$ segment is offset by two nucleotides with respect to the microRNA. Furthermore, we divided the microRNA precursor in the middle of the “loop” of the foldback into the microRNA–containing arm and the star–containing arm.

- We aligned this star with the (uncomplemented) target gene. This divides the target into three segments, which we will call "5prime", "starBinder" and "3prime".
- We scored the microRNA–containing arm against the target gene using a global alignment algorithm (see below).
- We obtained a new sequence from the microRNA arm by independently permuting the bases in the 5prime and 3prime segments of the original microRNA arm. Then we scored this sequence against the reverse complemented target gene using a global alignment procedure. We repeated this procedure 10 000 times and noted the resulting score of each run.
- We scored the star–containing arm against the target gene using a global alignment algorithm (see below).
- We obtained a new sequence from the star arm by permuting the bases in the 5prime and 3prime segments of the original star arm. Then we scored this sequence against the target gene using a global alignment procedure. We repeated this procedure 10 000 times and noted the resulting score of each run.

This segmentation procedure ensures that the microRNA part of the precursor is aligned to the same position within the permuted and the un–permuted target arm. Furthermore, the nucleotide composition of each arm segment remains unchanged with respect to the original arm. As can be expected, the scores of the 10 000 permuted sequences are spread out in a bell curve. For example, Figure 6.4 depicts a histogram of the 10 000 scores obtained from permuting the micro arm of candidate mpss01 and aligning it with the reverse complement of its target gene AT3G19890.1. In this case, the average score of the 10 000 permuted micro arms was 130.0 and the score of the un–permuted micro arm was 137.5.

In order to evaluate the significance of the un–permuted score, we ranked all scores obtained from permuting the micro arms in increasing order and noted the rank of the un–permuted score. Likewise, we permuted and ranked the star arms. The results of all candidate/target pairs for which the target could be experimentally validated (cf. Table 6.2) are shown in Table 6.3.

The significance of the similarity evaluated by this procedure suggests that candidate mpss02 is indeed related to its targets outside of the binding segment and thus most likely is a product of the inverted duplication hypothesis scenario. For the other two candidates, mpss01 and mpss05, the evidence for either scenario is ambiguous: In each case, one of the arms scores among the top 10% scores of the permutation test which would argue for the inverted duplication hypothesis scenario. On the other hand, we checked for homologs of these candidates in the NCBI EST database,

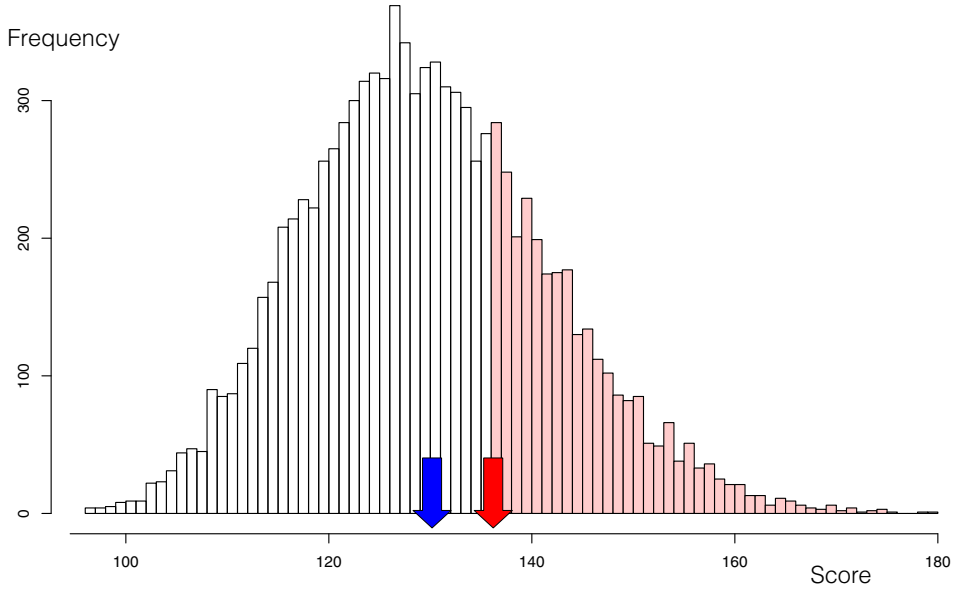



**Figure 6.4:** This histogram displays the distribution of scores across the $10\,000$ permutations (as detailed in the text) for the micro arm of mpss01 aligned against the AT3G19890 target gene. The score of the un–permuted micro arm ($137.5$) is marked by the red arrow; the average score of all permutations ($130.0$) is marked by the blue arrow. Bars of permuted scores at least as high as that of the un–permuted micro arm are shaded in light red.

| microRNA | arm | target gene | score | avg. score | rank |
|---|---|---|---|---|---|
| mpss01 | micro arm | AT3G19890.1 | 137.5 | 130.0 | 7447 |
| mpss01 | star arm | AT3G19890.1 | 126.5 | 113.7 | 9097 |
| mpss02 | micro arm | AT1G60130.1 | 138.0 | 108.5 | 9930 |
| mpss02 | star arm | AT1G60130.1 | 161.0 | 124.0 | 9965 |
| mpss02 | micro arm | AT5G38550.1 | 144.5 | 112.9 | 9916 |
| mpss02 | star arm | AT5G38550.1 | 154.5 | 109.9 | 9996 |
| mpss05 | micro arm | AT1G43130.1 | 137.5 | 121.8 | 9027 |
| mpss05 | star arm | AT1G43130.1 | 116.5 | 118.3 | 4673 |

**Table 6.3:** This table shows the results of the global alignments and the corresponding permutation test. Columns denote the following: *microRNA* denotes the microRNA candidate name. *arm* denotes whether the microRNA–containing arm or the star–containing arm have been used for this test. The *score* column contains the score of the un–permuted microRNA arm (or star arm) when globally aligned to the target gene. The *avg. score* column contains the average of the 10 000 scores obtained from the permutation test. The *rank* column contains the rank of the original arm score with respect to the permuted arm scores, ordered ascendingly.

and all genomic plant databases included with the microHARVESTER (cf. Chapter 2)—with negative results. We also manually checked for homologs in the poplar genome using BLAST analysis—again with a negative result.

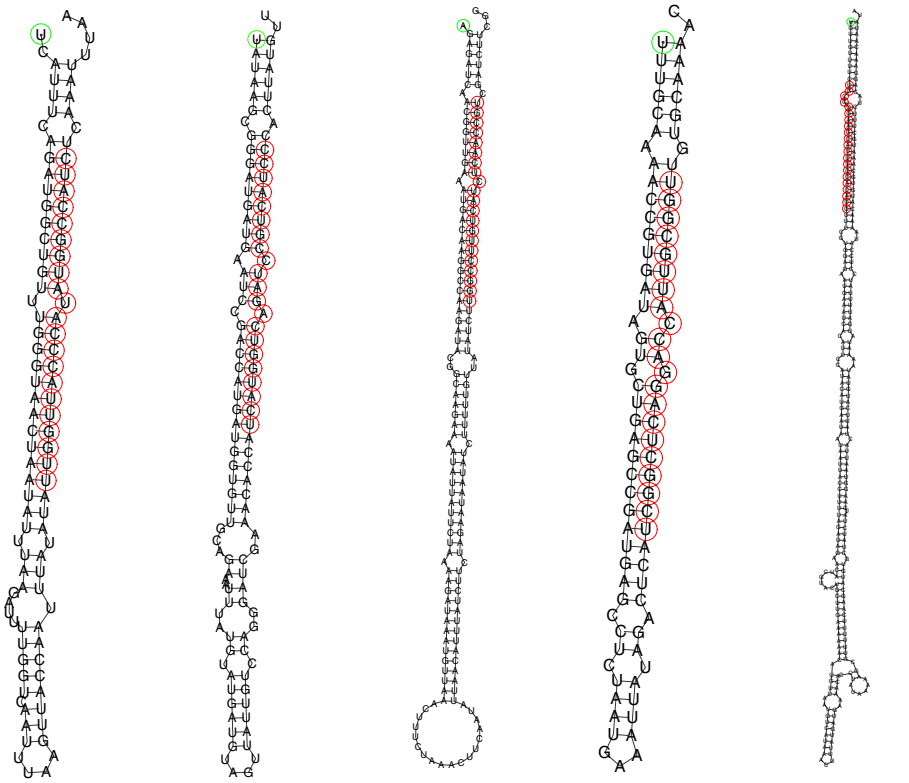
From this we can conclude that mpss01 and mpss05 have evolved recently and are younger than the most recent common ancestor of *Arabidopsis* and poplar—which lends support for the alternative route for microRNA evolution in plants.

### 6.2.6 Relationships of Our Candidates to Other Sequences

In addition to comparing our 13 microRNA candidates to repetitive sequences and published microRNAs we wanted to see how they related to two other sets of microRNA–related sequences: The 1953 small RNAs contained in the *Arabidopsis* Small RNA Project database (ASRP) [Gustafson *et al.*, 2005] and the 592 microRNAs predicted *in silico* by Lindow and Krogh [Lindow & Krogh, 2005].

#### ASRP Sequences

The ASRP database contains 1953 small RNAs from both in–house cloning projects of the Carrington Lab and sequences deposited in the microRNA registry. All contained sequences have thus been expressed and processed into small RNAs and, therefore, many are likely candidates for being derived from or associated with microRNA. For the comparison with the ASRP sequences, we used the online tool CrossLink (cf. Chapter 4). We used our



(a) Mpss01 (b) Mpss02 (c) Mpss05 (d) Mpss07 (e) Mpss11

**Figure 6.5:** Predicted folds of the five microRNA precursors for which processing into small RNAs could be confirmed as detailed in the text. The putative microRNA sequence is highlighted in red and the 5′ end of each sequence is marked by a green circle.

candidate microRNAs as set A and the ASRP sequences as set B, parameterizing CrossLink to use Vmatch for the A/B similarity search, enforcing perfect matches of minimal length 16 nucleotides between them. This yielded the following results:

- **mpss07/mpss08:**
  One 21 nucleotide long ASRP sequence (ASRP1983) is identical to our predicted candidate mpss08, providing evidence for its expression and processing—although we were not able to detect its processing using Northern blots (see above). Another ASRP sequence (ASRP2025), which is 24 nucleotides long, matches perfectly but shifted with respect to mpss08 to the opposing strand of the precursor as predicted by us. As shown in Figure C.1, we predicted that mpss07 and mpss08 would derive from the same long hairpin. The ASRP database does not

contain a small RNA corresponding to mpss07, for which we could provide evidence for expression and processing. Figure C.1 shows the positioning of mpss07, mpss08, ASRP1983 and ASRP2025 within the same foldback.

- **mpss09:**
  The 22 nucleotide small RNA sequence named ASRP1896 from the ASRP database is identical in its last 17 nucleotides to the first 17 nucleotides of our candidate mpss09. In addition, the sequences ASRP-1742, ASRP1842, ASRP1922, ASRP1923, ASRP2079 and ASRP2088 are identical in at least 21 nucleotides to a part of the precursor of mpss09.

- **mpss11:**
  The 21 nucleotide small RNA sequence named ASRP1729 from the ASRP database is identical to the 19 nucleotides of mpss11, and the two remaining nucleotides are identical to the two bases downstream of mpss11 within its precursor. One can see in the Northern blot for mpss11 that the major band is at 19 nucleotides (cf. Figure 6.3) but there are additional bands at 20 and 21 nucleotides' length. This confirms our experimental results that this predicted precursor yields a small RNA at the predicted position, possibly varying in length between 19 and 21 nucleotides.

- **mpss12:**
  The 24 nucleotide small RNA sequence named ASRP1110 from the ASRP database is identical in its reverse complement to the last 21 nucleotides of our candidate mpss12. Regarding the precursor of mpss12, the reverse complements of ASRP607, ASRP895, ASRP862 and ASRP-1331 are identical to a position on either the micro arm or the star arm of the mpss12 precursor.

### Sequences Predicted by Lindow and Krogh

Lindow and Krogh [Lindow & Krogh, 2005] recently used a new *in silico* approach to predict 592 microRNAs. Their procedure did not incorporate comparative information and instead exclusively relied on the *Arabidopsis* genome sequence. In essence, this approach started out with the assumption that each microRNA has one or more protein–coding target genes. Consequently, a large set of candidate microRNA-originating loci was generated, which was then refined step by step when structural constraints of known microRNA precursors were enforced. Unfortunately, no experimental validation was performed following their predictions.

We used the CrossLink software to compare our set of 13 microRNA candidates to the set of 592 candidates predicted by Lindow and Krogh. In the following, we will use this nomenclature for their sequences: "$L\#12345$" denotes the sequence which they name "*Locus–id: 12345*" on their web–frontend that allows browsing of their results. We found similarities for the following four sequences:

- **mpss01:**
  The Lindow/Krogh predicted precursor sequence $L\#91580$ is identical to that of our candidate mpss01. The predicted microRNA sequences are shifted by one nucleotide. Both Lindow/Krogh and we predicted the same target gene.

- **mpss02:**
  The Lindow/Krogh predicted precursor sequence $L\#92546$ is identical to that of our candidate mpss02. The predicted microRNA sequences are shifted by three nucleotides. Lindow/Krogh predict 2 targets of the jacalin lectin family of proteins: $AT1G60110$ and $AT5G38550$. We predicted eight targets and could experimentally show cleavage products for two of them as displayed in Table 6.2. In summary, one of the targets predicted by Lindow and Krogh, $AT5G38550$, was identified independently by both approaches and validated by us while other target predicted by Lindow and Krogh, $AT1G60110$, is a close homolog to the second target we predicted and validated, $AT1G60130$.

- **mpss07:**
  The Lindow/Krogh predicted precursor sequence $L\#97832$ is identical to that of our candidate mpss07. The predicted microRNA sequences, however, are unrelated and stem from different parts of the precursor. We did not find any target gene for our microRNA. Lindow/Krogh predict AT5G46540, an ABC transporter family protein to be targeted by their microRNA candidate with sequence `ACACCGTTTTGCACAACCGC` which is most likely not a correct target due to mismatches at the crucial positions 9 and 11, counting from the $5'$ end of the microRNA.

- **mpss11:**
  Lindow and Krogh predict a precursor sequence ($L\#243447$) on the other strand of our predicted precursor mpss11. In addition to thus being reverse complementary, their predicted microRNA is shifted approximately five nucleotides in its position within the foldback.

## 6.3 Methods

We downloaded the MPSS tag data from `http://mpss.udel.edu/at/public_data/small/smallRNA_17_summary.txt` and loaded it into a MySQL database (`http://www.mysql.com`). We used BLAST [Altschul *et al.*, 1997] to match MPSS tags onto all published microRNA mature and precursor sequences which we had downloaded from the Sanger microRNA registry [Griffiths-Jones, 2004; Griffiths-Jones *et al.*, 2006] in Release 8.0 downloaded on $2^{nd}$ of May 2006. We downloaded repetitive *Arabidopsis* sequences (file `TIGR_Arabidopsis_Repeats.v2`) from TIGR at `ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats/` We used the suffix array based Vmatch program [Kurtz *et al.*, 2001] to map MPSS tags onto the *Arabidopsis* genome.

In a first attempt to quantify similarities between our microRNA candidates and their putative targets, we followed the procedure described in [Allen *et al.*, 2004] and performed FASTA searches of the candidate arms against the genome—obtaining identical results to Allen and colleagues. For the refined method to evaluate these similarities, we implemented our own program that made use of the NEEDLE and SHUFFLESEQ programs as a subroutine. NEEDLE and SHUFFLESEQ are part of the Emboss package [Rice *et al.*, 2000]. Note that NEEDLE performs a Needleman–Wunsch [Needleman & Wunsch, 1970] global alignment based on the smaller of the two sequences but does not penalize unaligned excess sequence in the longer one. Consequently, when aligning the small microRNA sequence against the much longer target sequence the alignment is global with respect to the microRNA but local with respect to the target sequence.

The R package for statistical computing [R Development Core Team, 2005] was used to rank scores and generate histograms. We used the microHARVESTER [Dezulian *et al.*, 2006a] as mentioned in the text. We used the CrossLink [Dezulian *et al.*, 2006b] software to determine relationships of our candidates to other sequence sets as detailed in the text. We downloaded the microRNA precursor sequences predicted by Lindow and Krogh [Lindow & Krogh, 2005] from their accompanying webpage `http://www.binf.ku.dk/users/morten/mimatcher/arabidopsis/` and the ASRP sequences [Gustafson *et al.*, 2005] from `http://asrp.cgrb.oregonstate.edu`.

## 6.4 Conclusion

We designed a procedure to predict new microRNAs based on a large set of recently published MPSS sequence tags of three *Arabidopsis thaliana* small RNA libraries [Lu *et al.*, 2005]. Starting with the MPSS tags, we refined our sequence set and after several steps, we selected 13 loci as possible microRNA precursor candidates. Using transgenic overexpressors, we were able to show processing of five of these precursors *in vivo*. We chose to determine and further analyze the putative targets of three of these precursors and subsequently were able to show *in vivo* that in overexpressor plants, targets for each of these precursors were cleaved. In addition, the target RNA was downregulated in the transgenic plants as expected.

We propose an alternate route to microRNA evolution which complements the inverted duplication hypothesis put forward by Allen, Carrington and colleagues [Allen *et al.*, 2004]. We provide evidence that one of our three new microRNA genes has evolved according to the inverted duplication hypothesis while the other two new microRNA genes do not show significant similarity to their experimentally verified target genes although they seem to have recently evolved—a situation well in agreement with the alternative evolutionary route that we suggest.

# Chapter 7

# Discussion

A young and exciting field of research like that of microRNAs does not stand still—much to the contrary. Thus, as in other fast-moving areas, this work was heavily influenced "just-in-time" by the research of other groups—for better or worse.

On one hand, we could immediately integrate new insights of others into our own research and thus profit from them. On the other hand, we found ourselves in a highly competitive setting with other groups where the rule "winner takes all" would apply.

Awareness of this setting prevented us, for example, from performing a detailed analysis on a large set of microRNA homologs which we had derived from EST databases—a decision that was justified only a few weeks later when a publication entitled "Identification and characterization of new plant microRNAs using EST analysis" [Zhang *et al.*, 2005] reported on such an analysis. Competition also led to emotional roller coasters, when, in one instance, the new microRNA family that we had just found comparatively between *Arabidopsis* and poplar with our microSECTOR tool, turned out to have been discovered by others shortly before and had not quite made its entry into the public section of the Sanger microRNA registry (under the name "miR390"). As a third example, shortly after we had made our survey manuscript entitled "Conservation and divergence of microRNA families in plants" publicly available in the "Deposited Research" section of the *Genome Biology* journal, an article with the almost identical title ("Conservation and divergence of plant microRNA genes") and very similar content was published in *The Plant Journal* [Zhang *et al.*, 2006].

MicroRNAs are typical objects of research intrinsic to the field of (molecular) biology. We have thus been particularly happy (and lucky) that we have been able to contribute to this field from a bioinformatic perspective: by providing software tools and algorithms that help to answer biological questions and by performing analyses which require computers because of either the amount of data involved or the complexity of the algorithms. Success in this endeavor has only been possible because of our close collaboration with biologists—a context which allows scientific problems to be tackled synergistically from both angles.

During the course of this thesis, we have consequently been able to provide several things to the community: first, a program for identifying microRNA homologs in large databases (microHARVESTER; cf. Chapter 2); secondly, a tool for the exploration and visualization of (micro) RNA sequences (CrossLink; cf. Chapter 4) and thirdly, an approach for the comparative prediction of microRNAs from two plant genomes (microSECTOR; cf. Chapter 5). Applying these tools to the genome pairs *Arabidopsis* and poplar, and sorghum and rice, we derived several promising microRNA candidates plus several other interesting genomic loci. In addition, we have made an analysis on the conservation and divergence of microRNAs in plants available and contributed numerous microRNA homologs to the Sanger microRNA registry in this context. Finally, we have had the idea of filtering an MPSS sequence database for microRNA candidates and developed appropriate methods—a project that ultimately led to the discovery of several new microRNAs and aided in substantiating the proposition of an alternative evolution scenario for microRNAs.

Looking into the future, the challenge of "microRNA prediction" may soon become obsolete in its present form because of the advent of new DNA sequencing technologies (such as pyrosequencing, nanopore sequencing and 454 sequencing) that hold the potential to sequence the transcriptome of a eukaryote with unprecedented speed and accuracy [Bonetta, 2006]. Reassuringly, the usefulness of several products of this thesis will be unaffected by this change, although some will—sadly—become obsolete. On the biology side, a small revolution has also just happened with the discovery of yet another small RNA molecule, the Piwi-interacting RNA (piRNA) that corresponds to regions of the (human) genome that were previously thought not to be transcribed [Carthew, 2006].

Naturally, nature still holds plenty of challenges in store.

# Appendix A

# Publications

## A.1 Published Manuscripts

1. Olaf Delgado Friedrichs, Tobias Dezulian and Daniel H. Huson. **A meta-viewer for biomolecular data.** *GI Jahrestagung* (2003), volume 1, pages 375–380.

   > The development of powerful visualization tools is a major challenge in bioinformatics. Although many good special purpose viewers exist, there is a need for configurable meta-viewers that provide enough flexibility to support many different types of data and visualizations. Here we present CGViz, a new software tool that fulfills many of the requirements placed on such a configurable meta-viewer.

2. Tobias Dezulian and Mike Steel. **Phylogenetic closure operations and homoplasy-free evolution.** In *Classification, Clustering, and Data Mining Applications* (Proceedings of the meeting of the International Federation of Classification Societies (IFCS) 2004). (eds D. Banks, L. House, F.R. McMorris, P. Arabie and W. Gaul), pages 395–416. Springer-Verlag, Berlin.

   > Phylogenetic closure operations—on partial splits, and quartet trees—turn out to be both mathematically interesting, and computationally useful. Although these operations were defined two decade ago, until recently little had been established concerning their properties. Here we present some further new results and links between these closure operations, and show how they can be applied in phylogeny reconstruction and enumeration. Using the operations we study how effectively one may be able to reconstruct phylogenies from evolved multi-state characters that take values in a large state space (such as may arise with certain genomic data).

3. Daniel H. Huson, Tobias Dezulian, Tobias Klöpper, and Mike A. Steel. **Phylogenetic Super-Networks from Partial Trees.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2004), volume 1, pages 151–158.

   In practice, one is often faced with incomplete phylogenetic data, such as a collection of partial trees or partial splits. This paper poses the problem of inferring a phylogenetic super-network from such data and provides an efficient algorithm for doing so, called the Z-closure method. Additionally, the questions of assigning lengths to the edges of the network and how to restrict the dimensionality of the network are addressed. Applications to a set of five published partial gene trees relating different fungal species and to six published partial gene trees relating different grasses illustrate the usefulness of the method and an experimental study confirms its potential. The method is implemented as a plug-in for the program SplitsTree4.

4. Tobias Dezulian, Michael Remmert, Javier F. Palatnik, Detlef Weigel and Daniel H. Huson. **Identification of plant microRNA homologs.** *Bioinformatics* (2006), volume 22, pages 359–360.

   MicroRNAs (miRNAs) are a recently discovered class of non-coding RNAs that regulate gene and protein expression in plants and animals. MiRNAs have so far been identified mostly by specific cloning of small RNA molecules, complemented by computational methods. We present a computational identification approach that is able to identify candidate miRNA homologs in any set of sequences, given a query miRNA. The approach is based on a sequence similarity search step followed by a set of structural filters.

5. Tobias Dezulian, Martin Schaefer, Roland Wiese, Detlef Weigel and Daniel H. Huson. **CrossLink: visualization and exploration of sequence relationships between (micro) RNAs.** *Nucleic Acids Research* (2006), volume 34, pages W400–W404.

   CrossLink is a versatile tool for the exploration of relationships between RNA sequences. After a parametrization phase, CrossLink delegates the determination of sequence relationships to established tools (BLAST, Vmatch and RNAhybrid) and then constructs a network. Each node in this network represents a sequence and each link represents a match or a set of matches. Match attributes are

reflected by graphical attributes of the links and corresponding alignments are displayed on a mouse-click. The distributions of match attributes such as E-value, match length and proportion of identical nucleotides are displayed as histograms. Sequence sets can be highlighted and visibility of designated matches can be suppressed by real-time adjustable thresholds for attribute combinations. Powerful network layout operations (such as spring-embedding algorithms) and navigation capabilities complete the exploration features of this tool. CrossLink can be especially useful in a microRNA context since Vmatch and RNAhybrid are suitable tools for determining the antisense and hybridization relationships, which are decisive for the interaction between microRNAs and their targets. CrossLink is available both online and as a standalone version at `http://www-ab.informatik.uni-tuebingen.de/software`.

## A.2 Submitted Manuscripts

6. Tobias Dezulian, Javier F. Palatnik, Daniel H. Huson and Detlef Weigel. **Conservation and divergence of microRNA families in plants.** Submitted for publication in *Genome Biology* and deposited at `http://genomebiology.com/2005/6/11/P13`.

Background: MicroRNAs (miRNAs) are 20 to 24 nucleotides short RNAs involved in posttranscriptional regulation in plants and animals. MiRNAs are processed from larger precursors with extensive secondary structure. In plants, a total of 286 miRNA genes in *Arabidopsis*, rice and maize had been identified by March 2005, clustered in 43 families.
Results: Here, we report the bioinformatic identification of 200 members of the 43 miRNA families in the genomes of maize, sorghum, medick and poplar. Furthermore, we report evidence for expression of 37 miRNA precursors that are present in EST collections of soybean and sugarcane. We have used the enlarged data set to systematically analyze several parameters of the plant precursors including stem length, conservation of the precursors and variation in the secondary structure of the miRNA along the precursor.
Conclusion: Based on this 83% increase in available miRNA

precursor sequences, we present an improved view of phylogenetic distribution, positional nucleotide preference, structural features and conservation of miRNA genes. Our results suggest that there are two different classes of plant miRNA precursors. The most abundant class includes precursors that have only two strongly conserved regions, corresponding to the mature miRNA and its complementary sequence. A less frequent class, which includes the miRNA families miR159/319 and miR394, displays two additional conserved sequence blocks. These precursors have larger stems with more extensive secondary structure.

7. Javier F. Palatnik, Heike Wollmann, Carla Schommer, Rebecca Schwab, Jerôme Boisbouvier, Edwards Allen, Ramiro Rodriguez, Tobias Dezulian, Daniel H. Huson, James C. Carrington and Detlef Weigel. **Differential Targeting of *MYB* and *TCP* Transcription Factor Genes by two related microRNAs in *Arabidopsis*.** Submitted for publication in *Nature Structural & Molecular Biology.*

   The miR159 and miR319 families of plant microRNAs (miRNAs) are closely related in sequence, yet seem to affect distinct sets of transcription factor genes *in vivo*. MiR159 regulates several *MYB* mRNAs, while miR319 predominantly targets *TCP* mRNAs. We demonstrate that miR319 can regulate both *MYB* and *TCP* mRNAs, but *MYB* targeting by miR319 plays at most a minor role in plants because of low endogenous miR319 levels. In contrast, computational predictions suggest that miR159 targets only *MYB* genes, and mutational and overexpression studies confirmed that its sequence prevents miR159 from affecting *TCP* mRNAs. This finding is supported by NMR spectroscopy, which shows that miR159 differentially interacts with potential *MYB* and *TCP* target sites. Finally, we identify nucleotide positions relevant for miRNA activity with mutants recovered from a suppressor screen. Together, our findings reveal that sequence and expression differences contribute to differential *in vivo* effects of miR159 and miR319.

8. Felipe Fenselau de Felippes*, Tobias Dezulian*, Michael Schröder, Daniel H. Huson and Detlef Weigel. **Evidence of chance evolution of functional microRNAs in plants.** Submitted for publication in *Current Biology*.
   [*: these authors contributed equally.]

   Plant miRNAs are produced from precursors that contain self-complementary foldbacks of variable lengths. *In silico* folding of the *Arabidopsis* genome shows that it has the potential to form hundreds of thousands of such foldbacks. A small number of known, low abundance MPSS signatures comes from regions that have a structure typical for miRNA genes. The low abundance suggests, however, either weak transcription or inefficient processing of these potential miRNA genes. Overexpression shows that five out of 13 tested foldbacks, when placed behind a constitutive promoter, can robustly give rise to small RNAs in the typical miRNA size range. Three of these miRNAs were predicted to be able to target *Arabidopsis* mRNAs for cleavage, and appropriate cleavage products are found in plants overexpressing these miRNAs. Because neither the foldbacks nor the target sites are conserved in the poplar genome, these findings suggest a new route for miRNA evolution in plants. By chance, the evolving genome routinely generates structures that can give rise to miRNAs once they are captured by transcriptional regulatory sequences. Subsequent stabilization through co-evolution with potential targets may lead to fixation of a small number of these.

# Appendix B

# Contribution

The ideas, work and results stated in this thesis have been the outcome of several years of work during my PhD. In addition to many fruitful discussions with my PhD supervisor, Daniel Huson, and colleagues at the Wilhelm-Schickard-Institute for Computer Science (WSI), I was able to draw much inspiration from collaborators at the Max-Planck-Institute, especially Detlef Weigel, Rebecca Schwab, Javier Palatnik and Felipe Felippes.

Being a scientific employee at the WSI at this time, I conceived of and supervised several diploma projects and numerous student projects. Naturally, I chose to supervise project topics located in my prime area of interest. For several of my ideas, I thus "outsourced" the implementation of a (sub-)task as a student or diploma project and then supervised the latter. Here, I want to separate the contributions of others from my work clearly and in detail.

**Chapter 2. Prediction of MicroRNA Homologs**

Javier Palatnik once mentioned that Ed Allen had picked up microRNA homologs using a simple BLAST search. This sparked my interest and I got the idea that it might be useful to look into this more closely, possibly leading to a tool that could detect microRNAs automatically in a database. After some experimentation, I implemented a first "quick-and-dirty" version of the microHARVESTER. In order to make this tool easily accessible for others, I offered the task of implementing a web interface for the microHARVESTER as a student project. Michael Remmert, who was already involved in the construction of the Bioinformatics Toolkit [Biegert *et al.*, 2006], took the job and provided the HTML frontend under my supervision. At the same time, I improved my core libraries that handled the search, filtering and PDF generation, and built upon this a second independent microHARVESTER version which I used (among other things) for analyses in Chapters 3 and 6. Originally, a second goal of the student's project had been the re-implementation of the microHARVESTER (using my core libraries) for better connectivity to the web frontend and also for

possibly improving the effectiveness of the approach. Since I found that my second version of microHARVESTER far exceeded this re–implementation in terms of effectiveness, I combined Michael's web interface stubs with my implementation for our publication [Dezulian *et al.*, 2006a].
The contributions to this publication were as follows: I wrote the manuscript, selected the journal and interacted with the editor and reviewers. Daniel Huson contributed useful comments. Detlef Weigel and Javier Palatnik substantially revised the biological part of the introduction.

**Chapter 3. Conservation and Divergence of MicroRNA Families in Plants**
With the microHARVESTER software and extensive manual inspection, I identified many microRNA homologs in a multitude of databases and contributed these to the Sanger microRNA registry. Then Detlef Weigel, Javier Palatnik and I conceived of a survey analysis of all plant microRNAs on the basis of this enlarged set of all plant microRNAs. I performed all analyses. Javier Palatnik, Detlef Weigel and I wrote the manuscript which is available on the *Genome Biology* deposited research server at `http://genomebiology.com/2005/6/11/P13`.

**Chapter 4. Visualization and Exploration of Sequence Relationships between (micro) RNAs**
Working with microRNAs, I frequently came across the need to explore relationships between different sets of microRNAs. I realized that no tool for this exploration task existed, not even for RNAs in general. Becoming aware that this would be a valuable contribution to the community, I decided to build an intuitive, versatile tool and to offer it as a web service. Two people helped me build CrossLink: Matthias Zschunke and Martin Schaefer. Building on library classes of mine, Matthias Zschunke helped circumvent restrictions imposed by the WSI firewall. Martin Schaefer took (as a student project under my supervision) the job of implementing large parts of the GUI built on top of my GUI framework, and he wrote the code that interacted with BLAST and Vmatch and dealt with the resulting matches following my ideas. Implementation-wise, I conceived of the application architecture, implemented the framework for the GUI and implemented everything else on the client and the server side. I handled all issues that were not implementation-related, including the cooperation with Roland Wiese from yWorks, which kindly provided us with their yFiles graph library. I wrote the manuscript ([Dezulian *et al.*, 2006b]), selected the journal and interacted with the editor and reviewers.

Detlef Weigel spawned the idea of taking the yet unpublished poplar genome which was at that time available in a first assembly (with restrictions) from the DOE Joint Genome Institute (JGI), and conceive of a new approach to predict new microRNAs comparatively using *Arabidopsis thaliana* as the second organism. Initially, I handed this task to Jó Bitsch as a student project and merely took the role of a consultant. When his project was finished, we had both gained much experience but although Jó had done an excellent job, this endeavor did not result in the discovery of any new microRNA. I realized that this was a far bigger challenge than originally expected.

I devoted all my energy into this problem for several months and was able to conceive of a detailed approach, identifying the crucial filters that one would need to apply and pinpointing the specialized tools that would be needed in each step (e.g. it was crucial to use the brand-new tool RNALfold [Hofacker *et al.*, 2004] (contained in the beta-release of the Vienna RNA package), which (while still taking weeks) was able to RNA-fold a whole plant genome). I handed the implementation of this approach to Christian Klug as a diploma thesis which I supervised. During his work, we met on an almost daily basis and I—being far senior in this topic—contributed the conceptual ideas while all implementation work was proficiently handled by Christian. During the last days of Christian's thesis, we both manually inspected the resulting set of 592 microRNA candidates and I applied the first version of my microHARVESTER software for further refinement, which led to the discovery of 4 candidates which looked promising enough to justify serious wet lab work. One of these candidates turned out to be miR390, which had just been published but had not yet found its way into the microRNA registry with which we compared our results. The other candidates were scrutinized in the wet lab section of Detlef Weigel's lab by Rebecca Schwab, Ben Czech and Felipe Felippes—while I handled the bioinformatic side of things.

In the following months, I further refined this comparative approach and—on Javier Palatnik's hint that the coding parts of the sorghum genome had been sequenced by methyl filtration—decided to comparatively identify microRNAs between rice and sorghum. Christian Mayer implemented this refined approach under my supervision as his student project. I scrutinized the resulting microRNA candidates using a modified version of the microHARVESTER software and—with the help of Detlef Weigel—set up a collaboration with Ramanjulu Sunkar, who worked in the lab of Jian-Kang Zhu at the University of California, Riverside (UCR). Ramanjulu Sunkar performed all wet lab work on the resulting set of promising candidates.

**Chapter 6. Prediction Based on MPSS Expression Data**

Reading [Lu *et al.*, 2005] I realized that the resulting public database of MPSS signals would be an invaluable resource for the identification of new microRNAs. I mapped the MPSS tags onto the *Arabidopsis thaliana* genome, excised the surrounding sequences, applied several filtering steps and finally came up with a set of 13 candidates that would be worth further scrutiny in the wet lab. Michael Schröder had contributed to this initial bioinformatic work by helping me to manually inspect about half of the approximately 1000 PDF documents that my software generated as a final step. Detlef Weigel agreed that these sequences would be worth further study and Felipe Felippes performed all wet lab work, while I conceived of and performed all further bioinformatic analyses. Detlef Weigel had the idea of focusing on the evolutionary implications of these non-conserved candidates. The three of us jointly compiled the manuscript ([de Felippes *et al.*, 2006]).

# Appendix C

# Supplementary Material

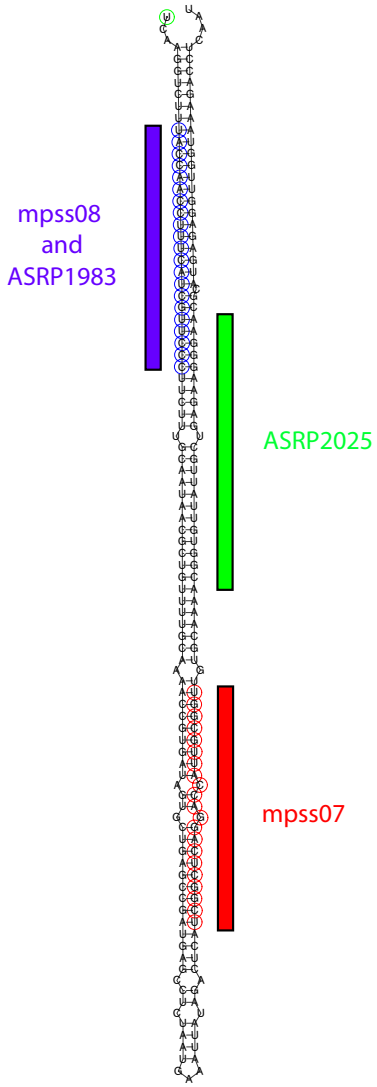



**Figure C.1:** Chapter 6: Candidates mpss07 and mpss08 of are located in the same segment on chromosome 1 that can form a foldback. The 21 nucleotide sequence ASRP1983 of the ASRP database is identical to mpss08, thus providing evidence for its expression and processing—although we could not provide this evidence experimentally. The 24 nucleotide sequence ASRP2025 (ACGGTGTTATTGCTGAGAAGGGAA) is located overlapping mpss08 on the opposite arm. The 5' end of this RNA sequence is marked in green.

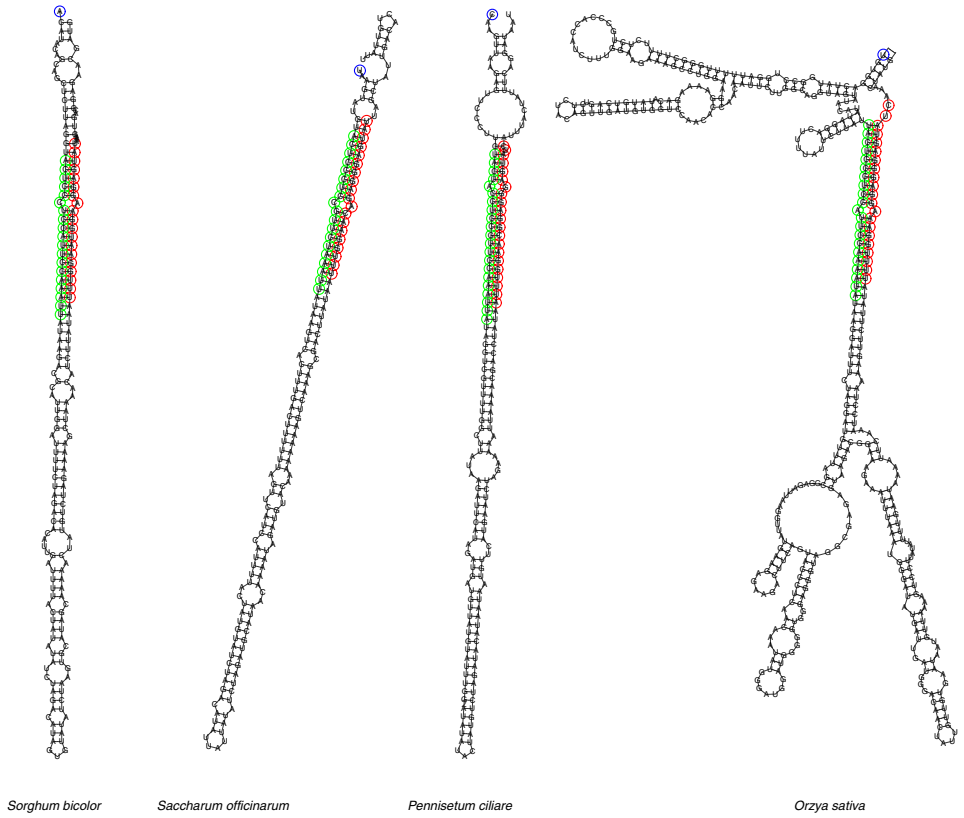



**Figure C.2:** Section 5.4.5 homologs of candidate TD-03. Each predicted RNA structure has been generated using RNAfold at a temperature of 20°C. Some sequences have been truncated to fit into this figure better. Each sequence is contained in the NCBI EST database and therefore expressed. The predicted mature sequence for each species is given in square brackets and its accession number in round brackets: *Sorghum bicolor* (gi|45955802|gb|CN128693.1|) [TTTGGAATGGAAGGAGTAT], *Saccharum officinarum* (gi|35035962|gb|CA141705.1|) [TTTGGAACAGAGGGAGTAT], *Pennisetum ciliare* (gi|27531974|gb|BM084065.1|) [TTTGGAACGGAGGGAGTAT], *Oryza sativa* (gi|29625944|gb|CB630955.1|) [TTTGGAAAGGAGGGAGTAT]. The 5' end of each RNA sequence, its predicted microRNA and the corresponding microRNA$\star$ are marked in blue, red and green, respectively.

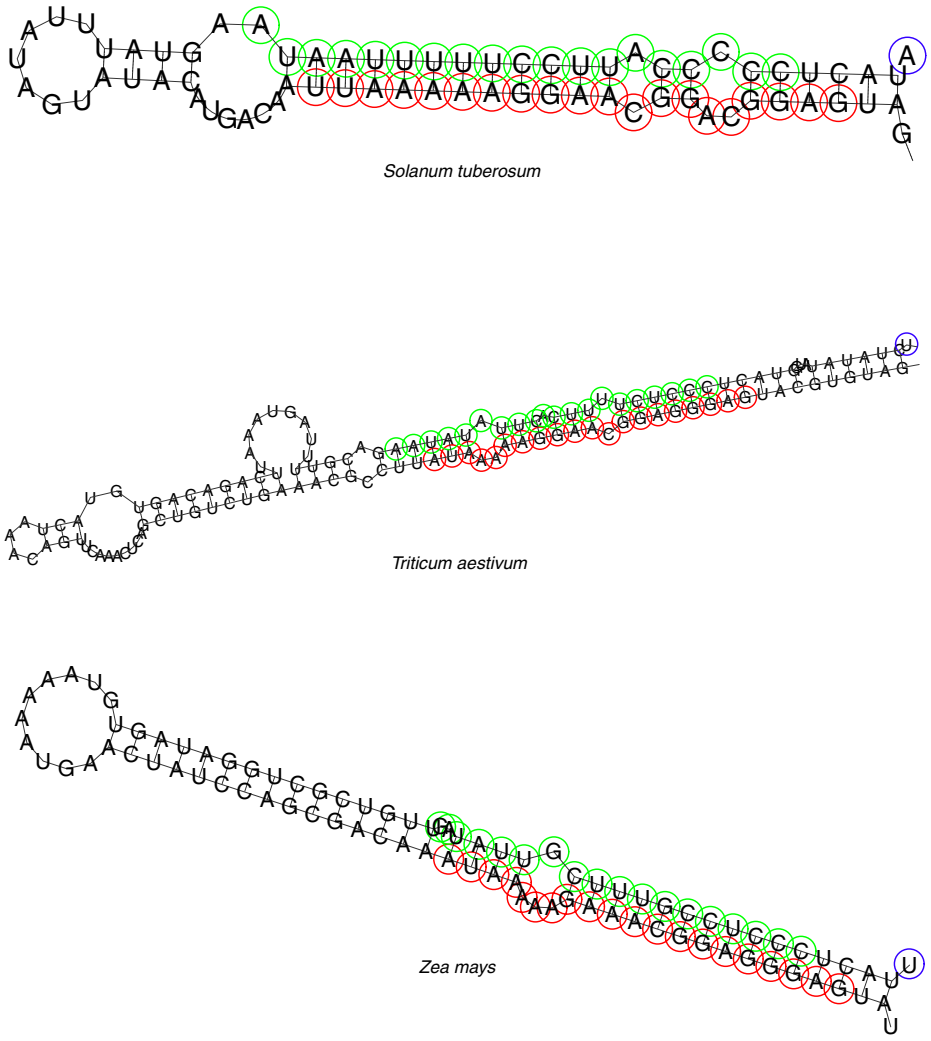



**Figure C.3:** Section 5.4.5 homologs of candidate TD-09. Each predicted RNA structure has been generated using RNAfold at a temperature of 20°C. Some sequences have been truncated to fit into this figure better. Each sequence is contained in the NCBI EST database and therefore expressed. The predicted mature sequence for each species is given in square brackets and its accession number in round brackets: *Solanum tuberosum* (gi|21372066|gb|BQ513197.1|) [TTAAAAAGGAACGGACGGAG], *Triticum aestivum* (gi|20078423|dbj|BJ253849.1|) [ATAAAAAGGAACGGAGGGAG], *Zea mays* (gi|18648869|gb|BM497688.1|) [ATAAAAAGAAACGGAGGGAG]. The 5′ end of each RNA sequence, its predicted microRNA and the corresponding microRNA⋆ are marked in blue, red and green, respectively.

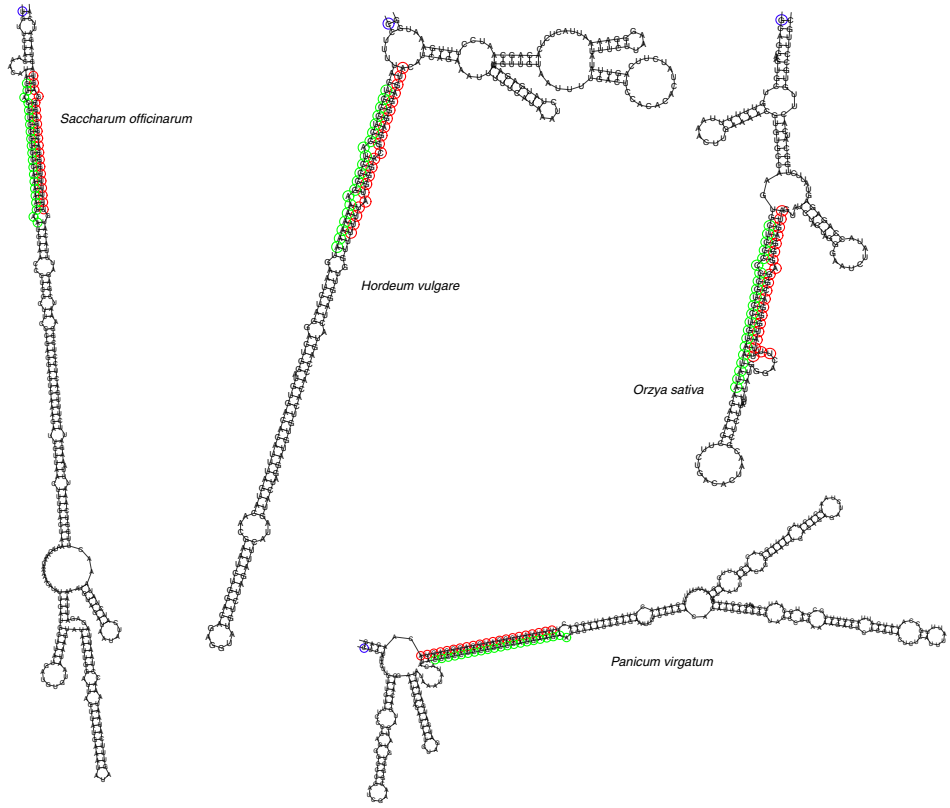



**Figure C.4:** Section 5.4.5 homologs of candidate TD-10. Each predicted RNA structure has been generated using RNAfold at a temperature of 20°C. Some sequences have been truncated to fit into this figure better. Each sequence is contained in the NCBI EST database and therefore expressed. The predicted mature sequence for each species is given in square brackets and its accession number in round brackets: *Saccharum officinarum* (gi|35278556|gb|CA222833.1|) [TTTTTTGGGACGGAGGGAGTA], *Hordeum vulgare* (gi|57827937|gb|CX629150.1|) [TTTTATGGGACGGAGGGAGTA], *Oryza sativa* (gi|33676391|gb|CF304630.1|) [TTTTATGGGACGGAGGGAGTA], *Panicum virgatum* (gi|59864031|gb|DN143174.1|) [TTTTATGGGATGGAGGGAGTA]. The 5' end of each RNA sequence, its predicted microRNA and the corresponding microRNA⋆ are marked in blue, red and green, respectively.

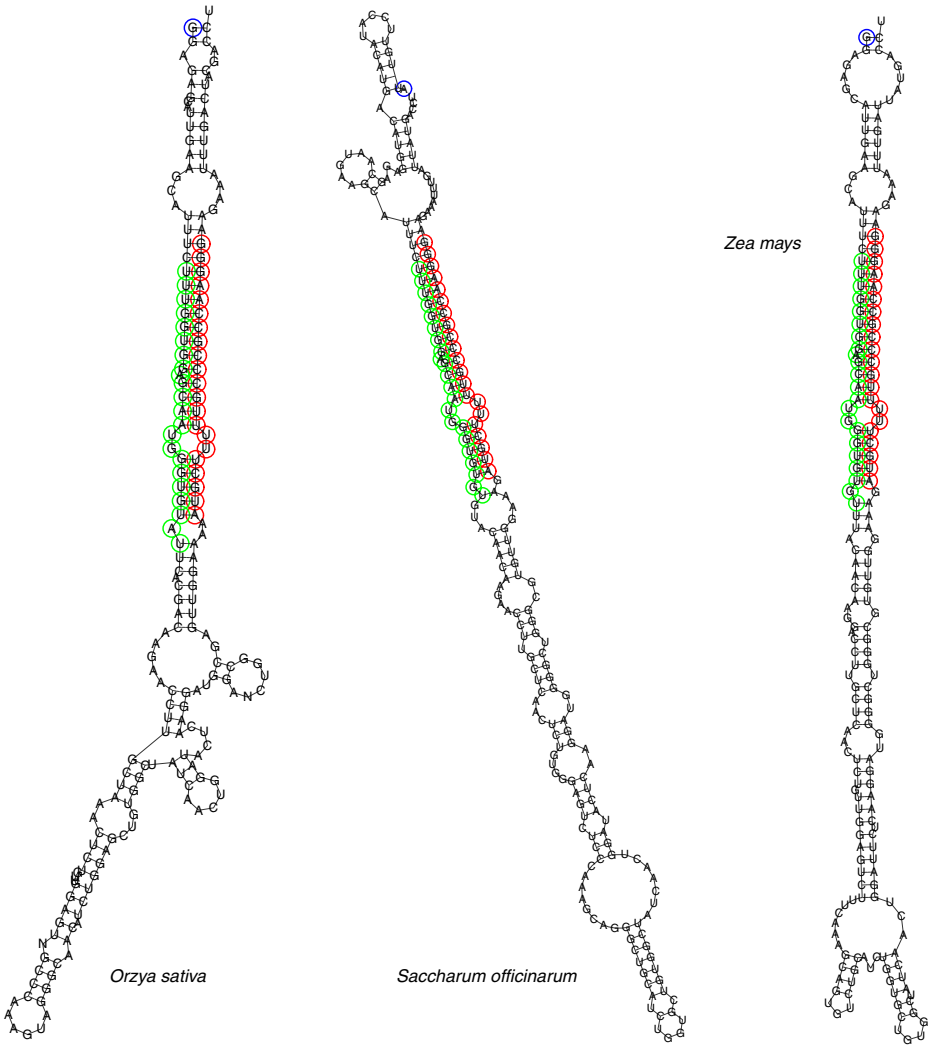



**Figure C.5:** Section 5.4.5 homologs of candidate TD-11. Each predicted RNA structure has been generated using RNAfold at a temperature of 20°C. Some sequences have been truncated to fit into this figure better. Each sequence is contained in the NCBI EST database and therefore expressed. The predicted mature sequence for each species is given in square brackets and its accession number in round brackets: *Oryza sativa* (gi|2442748|dbj|C74519.1|) [ATGCTTTTTGCCCGCCAAGGG], *Saccharum officinarum* (gi|36057997|gb|CA294783.1|) [ATGCTTTTTGCCCGCCAAGGG], *Zea mays* (gi|18450494|gb|BM428772.1|) [ATGCTTTTTGCCCGCCAAGGG]. The 5' end of each RNA sequence, its predicted microRNA and the corresponding microRNA$\star$ are marked in blue, red and green, respectively.

| | | | | |
|---|---|---|---|---|
| gma-MIR156b | ptc-MIR164e | ptc-MIR399f | sbi-MIR169c | zma-MIR167f |
| gma-MIR156c | ptc-MIR166a | ptc-MIR399g | sbi-MIR169d | zma-MIR167g |
| gma-MIR156d | ptc-MIR166b | ptc-MIR403a | sbi-MIR169e | zma-MIR167h |
| gma-MIR156e | ptc-MIR166c | ptc-MIR408 | sbi-MIR169f | zma-MIR167i |
| gma-MIR159 | ptc-MIR166d | sof-MIR156 | sbi-MIR169g | zma-MIR168a |
| gma-MIR160 | ptc-MIR166e | sof-MIR159a | sbi-MIR169h | zma-MIR168b |
| gma-MIR166a | ptc-MIR166f | sof-MIR159b | sbi-MIR169i | zma-MIR169c |
| gma-MIR166b | ptc-MIR166g | sof-MIR159c | sbi-MIR171a | zma-MIR169d |
| gma-MIR167a | ptc-MIR166h | sof-MIR159d | sbi-MIR171b | zma-MIR169e |
| gma-MIR167b | ptc-MIR166j | sof-MIR159e | sbi-MIR171c | zma-MIR169f |
| gma-MIR168 | ptc-MIR166n | sof-MIR167a | sbi-MIR171d | zma-MIR169g |
| gma-MIR169 | ptc-MIR166o | sof-MIR167b | sbi-MIR171e | zma-MIR169h |
| gma-MIR172a | ptc-MIR167b | sof-MIR168a | sbi-miR172a | zma-MIR169i |
| gma-MIR172b | ptc-MIR167d | sof-MIR168b | sbi-miR172b | zma-MIR169j |
| gma-MIR319a | ptc-MIR167i | sof-MIR396 | sbi-miR172c | zma-MIR169k |
| gma-MIR319b | ptc-MIR167j | sof-MIR408a | sbi-MIR172d | zma-MIR171c |
| gma-MIR319c | ptc-MIR168a | sof-MIR408b | sbi-MIR172e | zma-MIR171d |
| gma-MIR396a | ptc-MIR168b | sof-MIR408c | sbi-MIR319 | zma-MIR171e |
| gma-MIR396b | ptc-MIR169a | sof-MIR408d | sbi-MIR393 | zma-MIR171f |
| gma-MIR398a | ptc-MIR169g | sof-MIR408e | sbi-MIR394a | zma-MIR171g |
| gma-MIR398b | ptc-MIR169i | sbi-miR156a | sbi-MIR394b | zma-MIR171h |
| mtr-MIR156 | ptc-MIR169k | sbi-miR156b | sbi-MIR395a | zma-MIR171i |
| mtr-MIR160 | ptc-MIR169l | sbi-miR156c | sbi-MIR395b | zma-MIR171j |
| mtr-MIR162 | ptc-MIR169m | sbi-MIR156d | sbi-MIR395c | zma-MIR171k |
| mtr-MIR166 | ptc-MIR171a | sbi-MIR159 | sbi-MIR395d | zma-MIR172e |
| mtr-MIR169a | ptc-MIR171b | sbi-miR160a | sbi-MIR395e | zma-MIR319a |
| mtr-MIR169b | ptc-MIR171f | sbi-miR160b | sbi-MIR396a | zma-MIR319b |
| mtr-MIR171 | ptc-MIR171i | sbi-miR160c | sbi-MIR396b | zma-MIR319c |
| mtr-MIR319 | ptc-MIR172b | sbi-miR160d | sbi-MIR396c | zma-MIR319d |
| mtr-MIR393 | ptc-MIR172d | sbi-miR160e | sbi-MIR399a | zma-MIR393 |
| mtr-MIR395a | ptc-MIR172e | sbi-miR164 | sbi-MIR399b | zma-MIR394a |
| mtr-MIR395b | ptc-MIR172f | sbi-MIR164b | sbi-MIR399c | zma-MIR394b |
| mtr-MIR399a | ptc-MIR172g | sbi-miR166a | sbi-MIR399d | zma-MIR395a |
| mtr-MIR399b | ptc-MIR172h | sbi-miR166b | sbi-MIR399e | zma-MIR395b |
| mtr-MIR399c | ptc-MIR319a | sbi-miR166c | sbi-MIR399f | zma-MIR395c |
| mtr-MIR399d | ptc-MIR319b | sbi-miR166d | sbi-MIR399g | zma-MIR395d |
| mtr-MIR399e | ptc-MIR319c | sbi-MIR166e | zma-MIR156j | zma-MIR396a |
| ptc-MIR156b | ptc-MIR319d | sbi-MIR166f | zma-MIR156k | zma-MIR396b |
| ptc-MIR156d | ptc-MIR319f | sbi-miR167a | zma-MIR159a | zma-MIR399a |
| ptc-MIR156g | ptc-MIR319g | sbi-miR167b | zma-MIR159b | zma-MIR399b |
| ptc-MIR156h | ptc-MIR395c | sbi-MIR167c | zma-MIR159c | zma-MIR399c |
| ptc-MIR156i | ptc-MIR395g | sbi-MIR167d | zma-MIR159d | zma-MIR399d |
| ptc-MIR156j | ptc-MIR395i | sbi-MIR167e | zma-MIR160f | zma-MIR399e |
| ptc-MIR160b | ptc-MIR396a | sbi-MIR167f | zma-MIR166j | zma-MIR399f |
| ptc-MIR160c | ptc-MIR396b | sbi-MIR167g | zma-MIR166k | zma-MIR408 |
| ptc-MIR162a | ptc-MIR396d | sbi-MIR168 | zma-MIR166l | |
| ptc-MIR162b | ptc-MIR396e | sbi-miR169a | zma-MIR166m | |
| ptc-MIR164b | ptc-MIR399a | sbi-miR169b | zma-MIR167e | |

**Table C.1:** Identification numbers of the microRNA homologs which we have been able to contribute to the Sanger microRNA registry. For details, refer to Chapter 3 and to our manuscript [Dezulian *et al.*, 2005].

# Bibliography


Achard, P., Herr, A., Baulcombe, D.C. & Harberd, N.P. (2004). Modulation of floral development by a gibberellin-regulated microRNA. *Development*, **131**, 3357–3365.

Adai, A., Johnson, C., Mlotshwa, S., Archer-Evans, S., Manocha, V., Vance, V. & Sundaresan, V. (2005). Computational prediction of miRNAs in *Arabidopsis thaliana*. *Genome Res*, **15**, 78–91.

Allen, E., Xie, Z., Gustafson, A.M., Sung, G.H., Spatafora, J.W. & Carrington, J.C. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. *Nat Genet*, **36**, 1282–1290.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–3402.

Axtell, M.J. & Bartel, D.P. (2005). Antiquity of microRNAs and their targets in land plants. *Plant Cell*, **17**, 1658–1673.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Bartel, D.P. & Chen, C.Z. (2004). Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nat Rev Genet*, **5**, 396–400.

Bedell, J.A., Budiman, M.A., Nunberg, A., Citek, R.W., Robbins, D., Jones, J., Flick, E., Rholfing, T., Fries, J., Bradford, K., McMenamy, J., Smith, M., Holeman, H., Roe, B.A., Wiley, G., Korf, I.F., Rabinowicz, P.D., Lakey, N., McCombie, W.R., Jeddeloh, J.A. & Martienssen, R.A. (2005). Sorghum genome sequencing by methylation filtration. *PLoS Biol*, **3**, e13.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. (2003). GenBank. *Nucleic Acids Res*, **31**, 23–27.

Berardini, T.Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L.A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D. & Rhee, S.Y. (2004). Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiol*, **135**, 745–755.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M. & Snyder, M. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.

Biegert, A., Mayer, C., Remmert, M., Söding, J. & Lupas, A.N. (2006). The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Res*, **34**, W335–W339.

Bitsch, J.A. (2004). MiRNA detection in plant genomes. Student project report, Wilhelm-Schickard-Institute for Informatics, Tübingen University.

Boguski, M.S., Lowe, T.M. & Tolstoshev, C.M. (1993). dbEST–database for "expressed sequence tags". *Nat Genet*, **4**, 332–333.

Bonetta, L. (2006). Genome sequencing in the fast lane. *Nature Methods*, **3**, 141–147.

Bonnet, E., Wuyts, J., Rouzé, P. & de Peer, Y.V. (2004). Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. *Proc Natl Acad Sci U S A*, **101**, 11511–11516.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J. & Corcoran, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol*, **18**, 630–634.

Brunner, A.M., Busov, V.B. & Strauss, S.H. (2004). Poplar genome sequence: functional genomics in an ecologically dominant plant species. *Trends Plant Sci*, **9**, 49–56.

Carthew, R.W. (2006). Molecular biology. A new RNA dimension to genome control. *Science*, **313**, 305–306.

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K. & Gingeras, T.R. (2004). Unbiased mapping

of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.

Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. (2004). WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188–1190.

de Felippes, F.F., Dezulian, T., Schröder, M., Huson, D.H. & Weigel, D. (2006). Evidence of chance evolution of functional microRNAs in plants. Submitted for publication to *Current Biology*.

Dezulian, T., Palatnik, J.F., Huson, D.H. & Weigel, D. (2005). Conservation and divergence of microRNA families in plants. Deposited research (not peer-reviewed) with *Genome Biology* (`http://genomebiology.com/2005/6/11/P13`).

Dezulian, T., Remmert, M., Palatnik, J.F., Weigel, D. & Huson, D.H. (2006a). Identification of plant microRNA homologs. *Bioinformatics*, **22**, 359–360.

Dezulian, T., Schaefer, M., Wiese, R., Weigel, D. & Huson, D.H. (2006b). CrossLink: visualization and exploration of sequence relationships between (micro) RNAs. *Nucleic Acids Res*, **34**, W400–W404.

Eddy, S.R. (2002). Computational genomics of noncoding RNA genes. *Cell*, **109**, 137–140.

Emrich, S.J., Aluru, S., Fu, Y., Wen, T.J., Narayanan, M., Guo, L., Ashlock, D.A. & Schnable, P.S. (2004). A strategy for assembling the maize (*Zea mays L.*) genome. *Bioinformatics*, **20**, 140–147.

Enright, A.J. & Ouzounis, C.A. (2001). BioLayout—an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, **17**, 853–854.

Feschotte, C., Jiang, N. & Wessler, S.R. (2002). Plant transposable elements: where genetics meets genomics. *Nat Rev Genet*, **3**, 329–341.

Floyd, S.K. & Bowman, J.L. (2004). Gene regulation: ancient microRNA target sequences in plants. *Nature*, **428**, 485–486.

Frickey, T. & Lupas, A. (2004). CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**, 3702–3704.

Fruchterman, T.M.J. & Reingold, E.M. (1991). Graph drawing by force-directed placement. *Software—Practice and Experience*, **21**, 1129–1164.

Fu, Y., Emrich, S.J., Guo, L., Wen, T.J., Ashlock, D.A., Aluru, S. & Schnable, P.S. (2005). Quality assessment of maize assembled genomic islands

(MAGIs) and large-scale experimental verification of predicted genes. *Proc Natl Acad Sci U S A*, **102**, 12282–12287.

Fujii, H., Chiou, T.J., Lin, S.I., Aung, K. & Zhu, J.K. (2005). A miRNA involved in phosphate-starvation response in *Arabidopsis*. *Curr Biol*, **15**, 2038–2043.

Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Dongen, S.V., Inoue, K., Enright, A.J. & Schier, A.F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science*, **312**, 75–79.

Glusman, G., Qin, S., El-Gewely, M.R., Siegel, A.F., Roach, J.C., Hood, L. & Smit, A.F.A. (2006). A Third Approach to Gene Prediction Suggests Thousands of Additional Human Transcribed Regions. *PLoS Comput Biol*, **2**, e18.

Griffiths-Jones, S. (2004). The microRNA Registry. *Nucleic Acids Res*, **32**, D109–D111.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. & Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, **34**, D140–D144.

Gustafson, A.M., Allen, E., Givan, S., Smith, D., Carrington, J.C. & Kasschau, K.D. (2005). ASRP: the *Arabidopsis* Small RNA Project Database. *Nucleic Acids Res*, **33**, D637–D640.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M. & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.

Hofacker, I.L., Priwitzer, B. & Stadler, P.F. (2004). Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, **20**, 186–190.

Hüttenhofer, A. & Vogel, J. (2006). Experimental approaches to identify non-coding RNAs. *Nucleic Acids Res*, **34**, 635–646.

Hüttenhofer, A., Schattner, P. & Polacek, N. (2005). Non-coding RNAs: hope or hype? *Trends Genet*, **21**, 289–297.

Jiang, N., Feschotte, C., Zhang, X. & Wessler, S.R. (2004). Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Curr Opin Plant Biol*, **7**, 115–119.

Jones-Rhoades, M.W. & Bartel, D.P. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell*, **14**, 787–799.

Jones-Rhoades, M.W., Bartel, D.P. & Bartel, B. (2006). MicroRNAS and Their Regulatory Roles in Plants. *Annu Rev Plant Biol*, **57**, 19–53.

Khvorova, A., Reynolds, A. & Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell*, **115**, 209–216.

Klug, C. (2005). Computational prediction of plant non-coding RNA using a comparative approach. Master's thesis, Wilhelm-Schickard-Institute for Informatics, Tübingen University.

Kurihara, Y. & Watanabe, Y. (2004). *Arabidopsis* micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc Natl Acad Sci U S A*, **101**, 12753–12758.

Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J. & Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res*, **29**, 4633–4642.

Lindow, M. & Krogh, A. (2005). Computational evidence for hundreds of non-conserved plant microRNAs. *BMC Genomics*, **6**, 119.

Llave, C., Kasschau, K.D., Rector, M.A. & Carrington, J.C. (2002). Endogenous and silencing-associated small RNAs in plants. *Plant Cell*, **14**, 1605–1619.

Lu, C., Tej, S.S., Luo, S., Haudenschild, C.D., Meyers, B.C. & Green, P.J. (2005). Elucidation of the small RNA component of the transcriptome. *Science*, **309**, 1567–1569.

MacIntosh, G.C., Wilkerson, C. & Green, P.J. (2001). Identification and analysis of *Arabidopsis* expressed sequence tags characteristic of non-coding RNAs. *Plant Physiol*, **127**, 765–776.

Maher, C., Timmermans, M., Stein, L. & Ware, D. (2004). Identifying MicroRNAs in Plant Genomes. In *Computational Systems Bioinformatics* (ed. IEEE), (ed. F. Titsworth), pp. 718–723. IEEE, Stanford, CA.

Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I. & Koonin, E.V. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct*, **1**, 7.

Mallory, A.C., Reinhart, B.J., Jones-Rhoades, M.W., Tang, G., Zamore, P.D., Barton, M.K. & Bartel, D.P. (2004). MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region. *EMBO J*, **23**, 3356–3364.

Mattick, J.S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, **25**, 930–939.

Mattick, J.S. (2005). The functional genomics of noncoding RNA. *Science*, **309**, 1527–1528.

Mayer, C. (2005). Comparative miRNA prediction in *Oryza sativa* and *Sorghum bicolor*. Student project report, Wilhelm-Schickard-Institute for Informatics, Tübingen University.

Meyers, B.C., Tej, S.S., Vu, T.H., Haudenschild, C.D., Agrawal, V., Edberg, S.B., Ghazal, H. & Decola, S. (2004). The use of MPSS for whole-genome transcriptional analysis in *Arabidopsis*. *Genome Res*, **14**, 1641–1653.

Millar, A.A. & Gubler, F. (2005). The *Arabidopsis* GAMYB-like genes, MYB33 and MYB65, are microRNA-regulated genes that redundantly facilitate anther development. *Plant Cell*, **17**, 705–721.

Morgenstern, B. & Atchley, W.R. (1999). Evolution of bHLH transcription factors: modular evolution by domain shuffling? *Mol Biol Evol*, **16**, 1654–1663.

Nakano, M., Nobuta, K., Vemaraju, K., Tej, S.S., Skogen, J.W. & Meyers, B.C. (2006). Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Res*, **34**, D731–D735.

Needleman, S.B. & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48**, 443–453.

Notredame, C. & Abergel, C. (2003). Using multiple alignment methods to assess the quality of genomic data analysis. In *Bioinformatics and Genomes : Current Perspectives*. Horizon Scientific Press, Wymondham, UK, 30–55.

Notredame, C., Higgins, D.G. & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**, 205–217.

Nussinov, R. & Jacobson, A.B. (1980). Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci U S A*, **77**, 6309–6313.

Palatnik, J.F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J.C. & Weigel, D. (2003). Control of leaf morphogenesis by microRNAs. *Nature*, **425**, 257–263.

Parizotto, E.A., Dunoyer, P., Rahm, N., Himber, C. & Voinnet, O. (2004). In vivo investigation of the transcription, processing, endonucleolytic activity, and functional relevance of the spatial distribution of a plant miRNA. *Genes Dev*, **18**, 2237–2242.

Pearson, W.R. & Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, **85**, 2444–2448.

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3900051-07-0.

Rehmsmeier, M., Steffen, P., Höchsmann, M. & Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B. & Bartel, D.P. (2002). MicroRNAs in plants. *Genes Dev*, **16**, 1616–1626.

Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L.A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D.C., Wu, Y., Xu, I., Yoo, D., Yoon, J. & Zhang, P. (2003). The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res*, **31**, 224–228.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B. & Bartel, D.P. (2002). Prediction of plant microRNA targets. *Cell*, **110**, 513–520.

Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, **16**, 276–277.

Schwab, R., Palatnik, J.F., Riester, M., Schommer, C., Schmid, M. & Weigel, D. (2005). Specific Effects of MicroRNAs on the Plant Transcriptome. *Dev Cell*, **8**, 517–527.

Schwab, R., Ossowski, S., Riester, M., Warthmann, N. & Weigel, D. (2006). Highly Specific Gene Silencing by Artificial MicroRNAs in *Arabidopsis*. *Plant Cell*, **18**, 1121–1133.

Shen, L.X., Basilion, J.P. & Stanton, V.P. (1999). Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc Natl Acad Sci U S A*, **96**, 7871–7876.

Smith, T.F. & Waterman, M.S. (1981). Identification of common molecular subsequences. *J Mol Biol*, **147**, 195–197.

Stolc, V., Samanta, M.P., Tongprasit, W., Sethi, H., Liang, S., Nelson, D.C., Hegeman, A., Nelson, C., Rancour, D., Bednarek, S., Ulrich, E.L., Zhao, Q., Wrobel, R.L., Newman, C.S., Fox, B.G., Phillips, G.N., Markley, J.L. & Sussman, M.R. (2005). Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. *Proc Natl Acad Sci U S A*, **102**, 4453–4458.

Sunkar, R. & Zhu, J.K. (2004). Novel and stress-regulated microRNAs and other small RNAs from *Arabidopsis*. *Plant Cell*, **16**, 2001–2019.

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R., Bhalerao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.L., Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Déjardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjärvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leplé, J.C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouzé, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., de Peer, Y.V. & Rokhsar, D. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

Voss, B. (2004). Advanced tools for RNA secondary structure analysis. PhD thesis, Center for Biotechnology, Bielefeld University.

Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X. & Li, Y. (2005). MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, **21**, 3610–3614.

Wang, X.J., Reyes, J.L., Chua, N.H. & Gaasterland, T. (2004). Prediction and identification of *Arabidopsis thaliana* microRNAs and their mRNA targets. *Genome Biol*, **5**, R65.

Weigel, D. & Izaurralde, E. (2006). A tiny helper lightens the maternal load. *Cell*, **124**, 1117–1118.

Wiese, R., Eiglsperger, M. & Kaufmann, M. (2001). yfiles—visualization and automatic layout of graphs. *LNCS, Proceedings of the 9th International Symposium on Graph Drawing*, **2265**, 453–454.

Xie, Z., Allen, E., Fahlgren, N., Calamar, A., Givan, S.A. & Carrington, J.C. (2005). Expression of *Arabidopsis* MIRNA genes. *Plant Physiol*, **138**, 2145–2154.

Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C., Nguyen, M., Pham, P., Cheuk, R., Karlin-Newmann, G., Liu, S.X., Lam, B., Sakano, H., Wu, T., Yu, G., Miranda, M., Quach, H.L., Tripp, M., Chang, C.H., Lee, J.M., Toriumi, M., Chan, M.M.H., Tang, C.C., Onodera, C.S., Deng, J.M., Akiyama, K., Ansari, Y., Arakawa, T., Banh, J., Banno, F., Bowser, L., Brooks, S., Carninci, P., Chao, Q., Choy, N., Enju, A., Goldsmith, A.D., Gurjal, M., Hansen, N.F., Hayashizaki, Y., Johnson-Hopson, C., Hsuan, V.W., Iida, K., Karnes, M., Khan, S., Koesema, E., Ishida, J., Jiang, P.X., Jones, T., Kawai, J., Kamiya, A., Meyers, C., Nakajima, M., Narusaka, M., Seki, M., Sakurai, T., Satou, M., Tamse, R., Vaysberg, M., Wallender, E.K., Wong, C., Yamamura, Y., Yuan, S., Shinozaki, K., Davis, R.W., Theologis, A. & Ecker, J.R. (2003). Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science*, **302**, 842–846.

Zhang, B., Pan, X., Cannon, C.H., Cobb, G.P. & Anderson, T.A. (2006). Conservation and divergence of plant microRNA genes. *Plant J*, **46**, 243–259.

Zhang, B.H., Pan, X.P., Wang, Q.L., Cobb, G.P. & Anderson, T.A. (2005). Identification and characterization of new plant microRNAs using EST analysis. *Cell Res*, **15**, 336–360.

Zuker, M. & Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, **9**, 133–148.

Zuker, M., Mathews, D.H. & Turner, D.H. (1999). Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. In *RNA Biochemistry and Biotechnology*, J. Barciszewski & B.F.C. Clark, eds., NATO ASI Series, Kluwer Academic Publishers.

## Lebens- und Bildungsweg

| | |
|---|---|
| **Name:** | Tobias Dezulian |
| **Geburtsdatum und -ort:** | 6.3.1972 in Böblingen |

| | |
|---|---|
| **1978 - 1982** | Besuch der Grundschule in Böblingen |
| **1982 - 1985** | Besuch des Max-Planck-Gymnasiums in Böblingen |
| **1985 - 1986** | Besuch der Junior High School, Wappingers Falls, New York, USA |
| **1986 - 1987** | Besuch der High School, Grapevine, Texas, USA |
| **1987 - 1991** | Besuch des Max-Planck-Gymnasiums in Böblingen |
| **06/1991** | Abitur (Note: 1,3) |
| | Leistungskurse: Mathematik und Physik |
| **09/1991 - 11/1992** | Zivildienst im Altenheim St. Vincenz-Haus, Köln |
| **12/1992 - 09/1993** | Auslandsreisen |
| **10/1993 - 02/1999** | Studium der Informatik an der Eberhard-Karls-Universität Tübingen |
| **03/1999 - 09/1999** | Diplomarbeit (Betreuer: Prof. Wolfgang Küchlin) bei DaimlerChrysler, Stuttgart-Möhringen, mit dem Titel *Evaluation of the Enterprise JavaBeans Component Model 1.0* |
| **09/1999** | Diplom in Informatik, Nebenfach Psychologie |
| | (Note: Sehr gut) |
| **04/2000 - 06/2002** | Entwickler für Client/Server Anwendungen bei der Firma Fiducia Karlsruhe/Stuttgart im Bereich Basistechnologie |
| **seit 09/2002** | Promotion an der Fakultät für Informatik, Universität Tübingen, Arbeitsbereich *Algorithmen der Bioinformatik* bei Prof. Daniel H. Huson |