Genomisch kodierte microRNAs und Viroid-induzierte kleine RNAs in Viridiplantae

I n a u g u r a l - D i s s e r t a t i o n
zur
Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf
vorgelegt von

Jan-Hendrik Teune
aus Düsseldorf

April, 2008

Aus dem Institut für Physikalische Biologie
der Heinrich-Heine Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: apl. Prof. Dr. G. Steger

Korreferent: Univ.-Prof. Dr. W. Martin

Tag der mündlichen Prüfung: 26.05.2008

Danksagung

Zuallererst möchte ich Herrn apl. Prof. Dr. Gerhard Steger für seine Unterstützung danken, die maßgeblich zum Gelingen dieser Arbeit beigetragen hat. Desweiteren möchte ich mich für die gewährte wissenschaftliche Freiheit bedanken, die ebenfalls zu einem positiven Abschluß dieser Arbeit beigetragen hat.

Herrn Prof. Dr. William Martin möchte ich danken, dass er sich so spontan bereit erklärt hat, diese Arbeit als Korreferent zu beurteilen.

Herrn Prof. Dr. Detlev Riesner und Herrn Prof. Dr. Dieter Willbold danke ich für die Möglichkeit, diese Arbeit im Institut für Physikalische Biologie anzufertigen.

Auch möchte ich mich bei Herrn Dr. Sascha Laubinger vom MPI in Tübingen bedanken, dass er mir an richtiger Stelle eine große Hilfe gewesen war.

Ein besonderer Dank geht an Herrn Dr. Michael „Kiwi“ Schmitz, der mir einen nicht zu kleinen Einblick in die Welt des Pinguins ermöglichte und auch bei fachlichen Fragen immer mit Rat und Tat zur Seite stand.

Weiterhin gilt mein Dank Herrn Dr. Oliver Bannach und Herrn Dr. Axel Schmitz, die nie mit guten Tips und ausführlichen Diskussionen geizten. Auch für die Bereitschaft, diese Arbeit zu korrigieren, möchte ich gerade Herrn Dr. Bannach danken.

Der Rechnergruppe möchte ich für eine stets unterhaltsame und erquickende Arbeitsatmosphäre danken, zudem möchte ich Frau Dr. cand. rer nat Nathalie Diermann danken, die ebenfalls stets für eine ausgezeichnete Atmosphäre verantwortlich war.

Allen anderen Mitgliedern des Instituts und besonders viele der mittlerweile Auswärtigen danke ich für das ausgezeichnete Arbeitsklima, das mir immer in bester Erinnerung bleiben wird.

Weiterhin möchte ich meiner Familie danken, die mir zu jederzeit eine große Hilfe waren und ohne die diese Arbeit erst garnicht möglich gewesen wäre.

Mein besonderer Dank gilt Andrea Marzoll für die moralische Unterstützung, liebevolle Freizeitgestaltung und ihrer Liebe, die besonders in der Endphase eine große Stütze gewesen ist.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1

Einleitung

Im Jahr 1958 stellte Crick das zentrale Dogma der Biologie auf, welches die Desoxyribonukleinsäuren (DNA) als zentralen Informationsträger in den Mittelpunkt rückte. Dieses Dogma besagt, dass die genetische Information durch Ribonukleinsäuren (RNA) in Protein übersetzt wird und die Nukleinsäuren die Aufgabe als Informationsspeicher („Desoxyribonukleinsäuren“, DNA) und -überträger („Ribonukleinsäuren“, RNA) übernehmen, während die Proteine die chemischen Prozesse in der Zelle bewerkstelligen. Seinerzeit waren drei Klassen von RNAs bekannt, die *messenger*-RNAs (mRNA), *transfer*-RNAs (tRNA) und die *ribosomal*-RNAs (rRNA), welche ausnahmslos an der Proteinbiosynthese beteiligt sind. Dieses Dogma wurde strittig, als in den 80er Jahren eine neue Klasse von nicht-Protein-kodierenden RNAs („non-coding RNAs“, ncRNA) mit katalytischen Eigenschaften entdeckt wurde (Guerrier-Takada *et al.*, 1983; Kruger *et al.*, 1982), welche Ribozyme genannt wurden. Seitdem wurden fortlaufend weitere Klassen von ncRNAs entdeckt, welche in der Zelle verschiedenste Funktionen übernehmen. Diese Funktionsvielfalt lässt auf ein großes RNA-gestütztes Netzwerk divergenter Funktionen und Regulationsmechanismen schließen.

Fire *et al.* beschrieben 1998, dass in lebende Zellen injizierte RNA mit endogenen mRNAs wechselwirkt und somit die Degradation der korrespondierenden Ziel-mRNA einleitet. Für diese Experimente erhielten Fire & Mello 2006 den Nobelpreis. Seither wurden weitere Klassen von RNA-vermittelten Regulationsmechanismen aufgedeckt, zu denen unter anderem die Regulation durch microRNAs und *small interfering RNAs* (siRNAs) gehören, auf die in den Abschnitten 1.3.1 und 1.3.1 genauer eingegangen wird. Zu den biochemischen Funktionen und Regulationsmechanismen der kleinen RNAs gehören neben der Entwicklung und Differenzierung von Zellgruppen z. B. auch ein Abwehrmechanismus gegen eindringende Viren (Saumet & Lecellier, 2006). Für das bessere Verständnis der Funktionsweise und den Umfang der RNA-vermittelten Regulationsnetzwerke ist es daher nötig, weitere Stoffwechselwege oder Prozesse in den Zellen zu identifizieren, die sich durch eine Beteiligung von kleinen RNAs als Mediator der Regulation auszeichnen. Zur Bewältigung dieser Aufgabe sind bereits einige microRNA-Vorhersageprogramme publi-

ziert worden; MiRScan (Lim *et al.*, 2003), RNAmicro (Hertel & Stadler, 2006) und microHARVESTER (Dezulian *et al.*, 2006) seien stellvertretend genannt für drei verschiedene Methoden zur microRNA, wobei in Abschnitt 1.5 nur auf MiRScan sowie microHARVESTER eingegangen wird, da diese beiden Ansätze speziell für pflanzliche microRNAs entwickelt wurden.

1.1 Ribonukleinsäuren

RNA und DNA sind vom Aufbau her recht ähnlich und unterscheiden sich vornehmlich dadurch, dass RNA meist einzelsträngig und DNA meist doppelsträngig vorliegt. Auf weitere Unterscheidungsmerkmale zwischen DNA und RNA wird in Abschnitt 1.2 genauer eingegangen. Prinzipiell lassen sich RNAs in zwei große Gruppen unterteilen, die Protein-kodierenden Boten-RNAs oder mRNAs und die nicht-Protein-kodierenden RNAs oder ncRNAs. Dabei besitzen die ncRNAs eine deutlich größere Funktionsvielfalt als die Protein-kodierenden RNAs, auf den in Abschnitt 1.3 anhand einiger Beispiele näher eingegangen wird.

1.2 Primär-, Sekundär- und Tertiärstruktur von RNA

Die RNA ist ein Polymer und besteht aus den vier Nukleobasen Adenin (A), Guanin (G), Cytosin (C) und Uracil (U), welche kovalent über den Phosphatrest am C5' der Ribose miteinander verknüpft sind (siehe Abbildung 1.1). Die RNA unterscheidet sich chemisch von der DNA in zweierlei Hinsicht. In der RNA wird anstelle des Nukleotids Thymin (T) das Nukleotid Uracil (U) eingebaut. Das zweite wichtige Unterscheidungsmerkmal betrifft die beteiligte Zuckerkomponente des Nukleotids. In der RNA ist die Nukleobase kovalent mit einer Ribose verknüpft, während die DNA an gleicher Stelle eine 2'-Desoxyribose besitzt. Die Sequenz ist polar aufgebaut und wird nach der Bindung des Phosphat-Rests am C5' der Ribose als 5'-Ende bzw. dem C3' der Ribose als 3'-Ende bezeichnet. Die Abfolge der Nukleotide in 5'-3'-Richtung gelesen wird auch als Primärstruktur bezeichnet.

Die RNA liegt im Gegensatz zur DNA meist einzelsträngig vor und kann daher durch intramolekulare Wechselwirkungen komplexe Strukturen einnehmen. Grundlage hierfür ist die Möglichkeit der Basenpaarausbildung zwischen nicht-benachbarten Nukleotiden sowie der Stapelwechselwirkungen („*Stacking*“, Dipol-induzierte-Dipol-Wechselwirkungen). Die Ausbildung intramolekularer Wasserstoffbrückenbindungen besitzt dabei nur einen geringen Einfluss auf die Gesamtstabilität einer RNA-Struktur. Den größten energetischen Anteil an der Gesamtstabilität einer RNA-Struktur liefern die Stapelwechselwirkungen, welche sich zwischen den planaren Basen der Nukleotide ausbilden. Dabei induzieren die delokalisierten π-Elektronen der planaren, aromatischen Basen ein Dipolmoment in der benachbarten Base und bewirken somit die Stapelwechselwirkung.
Die einfachste Struktur, die eine RNA einnehmen kann, ist eine sogenannte Hairpin-Struktur, die durch Rückfaltung auf sich selbst gebildet wird und aus einer Helix und

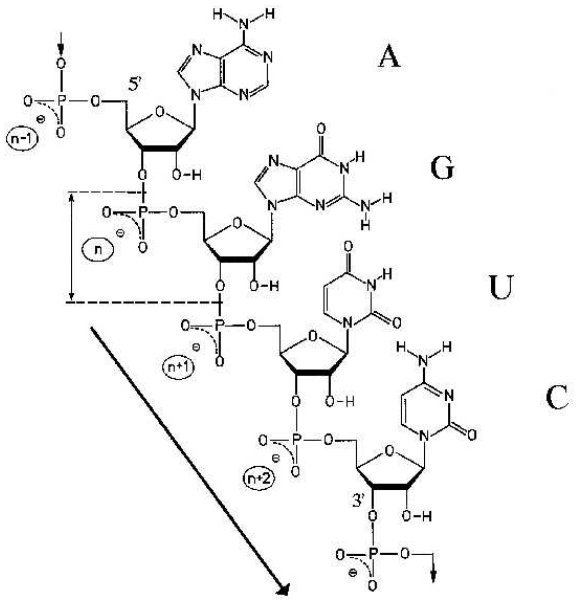

Abbildung 1.1: Primärstruktur der RNA. RNA besteht aus den vier Nukleotiden Adenin (A), Guanin (G), Cytosin (C) und Uracil (U), welche wiederum aus der namensgebenden Base, einem Ribose-Molekül, sowie einem Phosphat-Rest aufgebaut sind. Die Sequenz ist polar aufgebaut und wird nach der Bindung des Phosphat-Rests am C5' der Ribose als 5'-Ende bzw. dem C3' der Ribose als 3'-Ende bezeichnet. Die Sequenz wird der Konvention nach in 5'-3' Richtung gelesen. Das vorliegende Beispiel bezeichnet demnach die Sequenz AGUC. Nach Steger (2003)

einem Loop besteht. Dabei werden die Basenpaare A : U und G : C als Watson-Crick- oder kanonische Basenpaare bezeichnet. Beispiele für Hairpin-Loops und andere Loop-Typen sind in Abbildung 1.2A gezeigt. Helices steuern aufgrund der ausgebildeten Basenpaare und vornehmlich über die Stapelwechselwirkungen den günstigsten Beitrag zur thermodynamischen Gesamtstabilität bei. Die Loops üben hingegen immer einen destabilisierenden Einfluss auf die Gesamtstabilität aus. Die sogenannten extrastabilen Tetraloops, bei denen die Loop-Nukleotide zusätzliche Stapelwechselwirkungen mit der angrenzenden Helix und weitere nichtkanonische Basenpaare ausbilden, zählen zu den am wenigsten destabilisierenden Loops (siehe Abbildung 1.2B).

In der Bioinformatik wird die Sekundärstruktur als eine Liste von Basenpaaren beschrieben, die folgende Bedingungen erfüllen muss: Eine Base kann maximal eine Basenpaarung eingehen und Basenpaare dürfen sich in einem Graphen nicht überkreuzen, d. h. zwei Basenpaare (i, j) und (k, l) müssen die Bedingungen für die Bildung eines Basenpaares $i < j < k < l$ oder $i < k < l < j$ erfüllen.

Auf Basis der Sekundärstruktur bildet sich die Tertiärstruktur einer RNA aus. Unter der Tertiärstruktur einer RNA versteht man die Wechselwirkung einzelner Sekundärstrukturelemente untereinander. Die tertiären Wechselwirkungen lassen sich in drei Gruppen kategorisieren. Zu den Helix-Helix-Interaktionen gehören die koaxialen Stapelwechselwirkungen, bei denen zwei unterschiedliche Helices weitere Stapelwechselwirkungen untereinander ausbilden, wie z. B. der D- und T-Arm der tRNA (siehe Abbildung 1.3). Ein weiteres wichtiges Tertiärstrukturelement umfasst die Interaktion zwischen Helices und ungepaarten Elementen, zu denen unter anderem sogenannte Tripel-Basenpaare gezählt werden. Bei diesen bildet ein Basenpaar weitere Wasserstoffbrückenbindungen mit einem ungepaarten Nukleotid aus. Eine formelle Beschreibung eines Tripel-Basenpaares erfordert das Vorliegen zweier Basenpaare (i, j) und (k, l), wobei entweder $i = k$ oder $j = l$ gelten muss. Die letzte hier beschriebene Gruppe umfasst die Interaktion zweier ungepaarter Bereiche. Ein Beispiel für ein solches Tertiärstrukturelement ist die Bildung eines Pseudoknotens, bei denen die Loop-Nukleotide einer Hairpin-Struktur mit einem

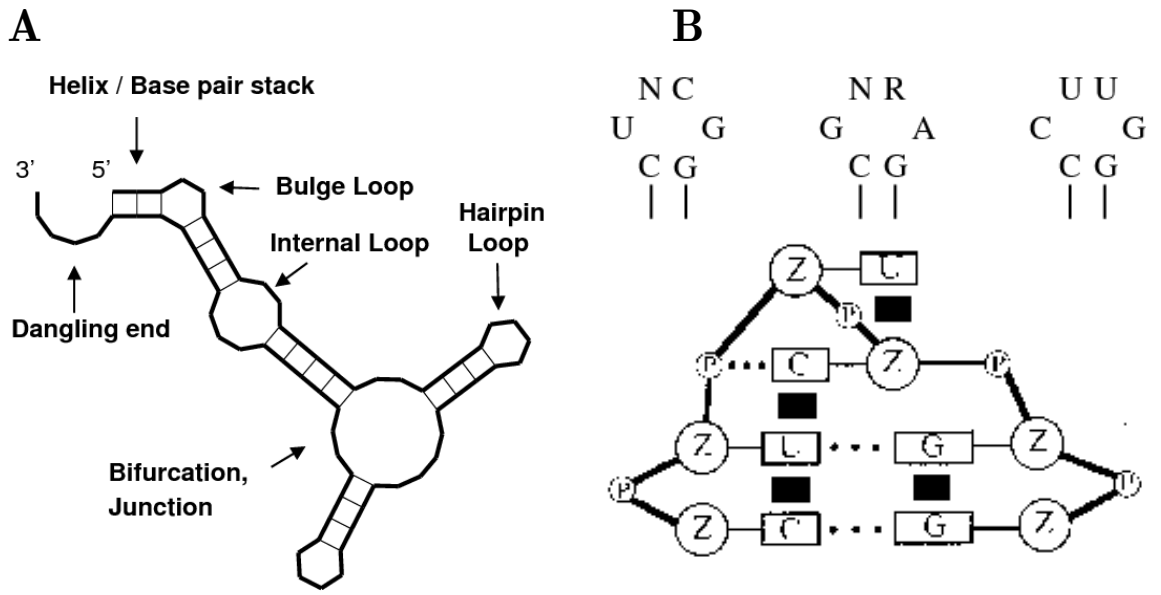

Abbildung 1.2: Sekundärstrukturelemente einer RNA. Darstellung **A** zeigt verschiedene Looptypen sowie einige weitere verschiedene Sekundärstrukturelemente. Die Loopbildung besitzt normalerweise einen destabilisierenden Einfluss auf die Gesamtstabilität. Ausnahmen bilden dabei die sogenannten extrastabilen Tetraloops, von denen in Abbildung **B** drei Konsensussequenzen dargestellt sind. Im unteren Teil von Abbildung **B** sind die Besonderheiten des extrastabilen UNCG-Tetraloops dargestellt. Die Loop-Nukleotide 1 (U) und 4 (G) bilden in dem gezeigten Beispiel ein zusätzliches Basenpaar aus und werden zusätzlich auf die angrenzende Helix gestapelt, was einen positiven Einfluss auf die Gesamtstabilität mit sich bringt. Das Nukleotid 3 (C) stapelt ebenfalls auf das angrenzende Basenpaar und bildet mit einem Phosphat-Rest weitere Wasserstoffbrücken aus. Das Nukleotid 2 (U) zeigt als einziges nach „außen“, stapelt aber ebenfalls auf die Ribose des Nukleotids 3. Nach Steger (2003)

freien Ende weitere Basenpaare ausbilden. Die formelle Beschreibung eines Pseudoknotens benötigt wieder das Vorliegen zweier Basenpaare (i, j) und (k, l), welche die Bedingung $i < k < j < l$ erfüllen.

1.3 Nicht-kodierende RNAs

Traditionell wurden RNAs als Mediator zwischen der auf der DNA kodierten Information und der zellulären Translationsmaschinerie angesehen. Ausnahmen bildeten die bekannten rRNAs als Bestandteil der Ribosomen und die tRNAs als Träger der Aminosäuren zu den Ribosomen. Als nicht-kodierende RNAs (ncRNA) werden hingegen die RNAs verstanden, die unmittelbar nach der Transkription ihre Funktion ausüben. Im Jahr 1982 wurde zum ersten Mal von einer RNA in *Tetrahymena thermophila* berichtet (Kruger *et al.*, 1982), welche in der 26S-rRNA-kodierenden Region lag und für sich Spalt-Aktivität ohne Beteiligung von Protein-Komponenten zeigte. Diese RNA wurde später in die Gruppe I-Intron eingeordnet und stellte das erste entdeckte Ribozym dar. Ribozyme zeichnen sich dabei über ihre autokatalytische Aktivität in Abwesenheit von Proteinen aus. In den folgen-

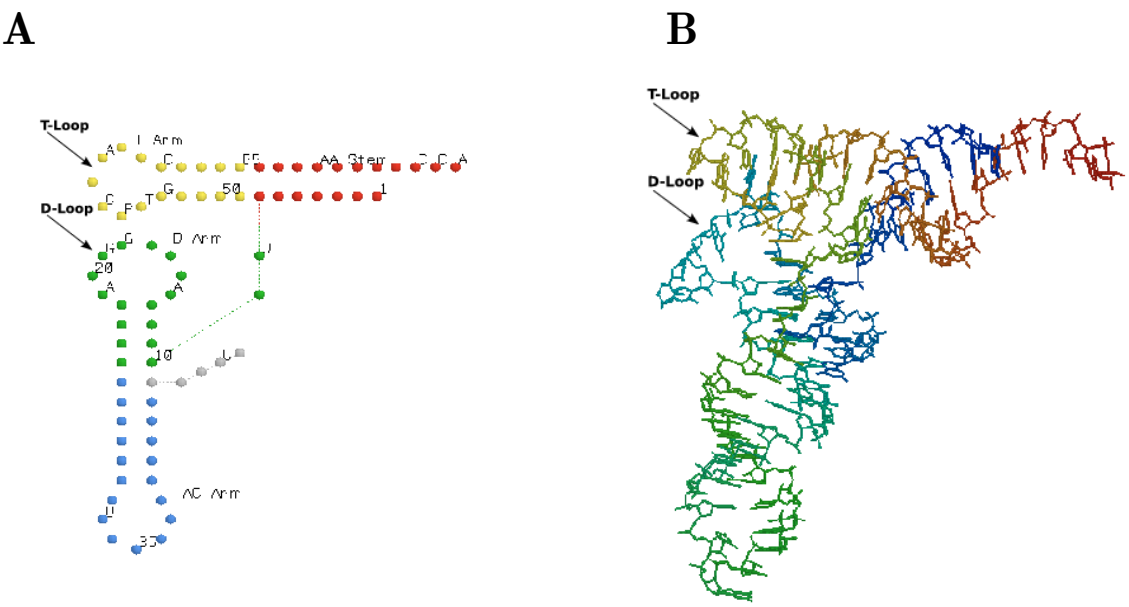

Abbildung 1.3: Tertiärstruktur am Beispiel einer tRNA. Darstellung **A** zeigt die Sekundärstruktur einer tRNA in L-Form. Hierbei sind die Bereiche, welche in der 3D-Darstellung in räumlicher Nähe zueinander stehen, ebenfalls in räumlicher Nähe dargestellt. Anhand dieser Darstellung lassen sich tertiäre Wechselwirkungen verdeutlichen. Darstellung **B** zeigt die Tertiärstruktur einer tRNA. Der D- und T-Loop steht in räumlicher Nähe zueinander und kann somit tertiäre Wechselwirkungen ausbilden.

---

den Jahren wurden fortlaufend weitere Klassen von ncRNAs entdeckt, darunter auch die von Fire *et al.* entdeckten kleinen RNAs („small RNAs“, sRNA). Eine andere Klasse von ncRNAs sind RNase P-RNAs, welche an der Prozessierung der tRNA-Vorläufer (pre-tRNA) beteiligt sind und für deren Maturierung durch Abspaltung eines 5’-seitigen *leader*-Bereichs unverzichtbar in der Zelle sind (Frank & Pace, 1998). Eine weitere Gruppe von ncRNAs sind die *small nucleolar* RNAs (snoRNA). Diese RNAs sind im Nukleus lokalisiert und sind an der Prozessierung und Modifikation einer ganzen Reihe von weiteren ncRNAs, darunter rRNAs und *small nuclear* RNAs (snRNA), beteiligt (Reichow *et al.*, 2007). Die snoRNAs sind zusammen mit einigen Proteinen in sogenannten *small nucleolar ribonucleoprotein particles* (snoRNP) assoziiert und zum einen für die Methylierung des 2’O der Ribose (C/D snoRNA-Klasse) und zum anderen für die Pseudouridylierung (H/ACA snoRNA-Klasse) entsprechender Ziel-RNAs zuständig. Die C/D snoRNA-Klasse bildet dabei über ein konserviertes Sequenzmotiv Basenpaarungen zu einer spezifischen Region in der entsprechenden Ziel-RNA aus, die sich in unmittelbarer Nähe zu der Stelle in der Ziel-RNA befindet, die methyliert werden soll. Die snoRNPs bilden über die snoRNA einen intermolekularen RNA-Duplex aus, von dem ausgehend die Methylierung stattfindet (Kiss-László *et al.*, 1996). Die typische Sekundärstruktur der H/ACA snoRNAs besteht aus zwei Hairpins, welche über eine ungepaarte Region miteinander verbunden sind. Die H/ACA snoRNAs besitzen zwei konservierte Sequenzmotive, wovon das eine in der Region zwischen den beiden Hairpins (H-Motiv, Konsensussequenz: ANANNA) und das andere am 3’-Ende lokalisiert ist (ACA-Motiv, Konsensussequenz: ACA) (Ofengand

et al., 2001). Die beiden konservierten Sequenzmotive bilden einen Duplex mit der zu modifizierenden RNA aus, welcher das zu isomerisierende Uridin direkt flankiert (Ganot *et al.*, 1997).

Dies sind exemplarische Beispiele für den Funktionsumfang der ncRNAs, welche jedoch nur einen kleinen Ausschnitt aus der Funktionsvielfalt unter Beteiligung von ncRNAs darstellen. Durch die Identifikation weiterer neuer Klassen von ncRNAs ist das Verständnis um die Funktionsvielfalt von RNAs in regulatorische aber auch funktionale Prozesse deutlich gestiegen. Der große Umfang an zentralen Aufgaben, der von RNAs in der Zelle übernommen wird und deren Konservierung in allen Lebewesen impliziert, dass deren Funktionalität schon früh in der Geschichte des Lebens entstanden ist und seitdem nicht mehr substantiell verändert wurde (Michalak, 2006).

1.3.1 Small RNAs (sRNAs)

Die sRNAs sind eine spezielle Klasse von ncRNAs. Sie kommen in Pflanzen, Tieren, Pilzen, manchen einzelligen Organismen und einigen Viren gleichermaßen vor. Man differenziert die Klasse der kleinen RNAs in *short interfering RNAs* (siRNA), *trans-acting siRNAs* (ta-siRNA, *natural antisense transcript-derived siRNAs* (nat-siRNA, *repeat-associated siRNAs* (ra-siRNA) und *microRNAs*. In ihrer biologisch aktiven Form sind sie ca. 19–28 Nukleotide lang und für die Regulation der Expression ihrer Zielgene auf unterschiedlicher Ebene von Bedeutung. Zum einen binden sie über Ausbildung von Basenpaaren an die entsprechende Ziel-mRNA und bewirken je nach Grad der Komplementarität entweder die Spaltung der mRNA oder die Inhibition der Translation (*posttranscriptional gene silencing*, PTGS), wie es in tierischen Zellen gezeigt werden konnte (Hutvágner & Zamore, 2002; Zeng *et al.*, 2002). Des Weiteren können sRNAs an epigenetischen Modifikationen im genetischen Kontext beteiligt sein, was als *transcriptional gene silencing* (TGS) bezeichnet wird. Der Oberbegriff für dieses Phänomen ist die sogenannte RNA-Interferenz oder RNAi (Fire *et al.*, 1998). Biologisch und chemisch unterscheiden sich diese RNAs nicht oder nur kaum voneinander. Sie werden allein über ihren Ursprung und ihre Biogenese klassifiziert, auf die im Folgenden genauer eingegangen wird.

MicroRNAs

Die microRNAs sind kleine, genomisch kodierte RNAs, welche die Regulation korrespondierender Gensequenzen in Pflanzen, Tieren, Pilzen und einzelligen Algen vermitteln. Eine grob vereinfachte Darstellung der microRNA-Prozessierung und -Funktionsweise ist in Abbildung 1.5 zu sehen. Im biologisch aktiven Zustand sind sie ca. 19–24 Nukleotide lang, und werden von der RNA-Polymerase II in langen Einheiten (mehrere 100 bis mehrere 1.000 Nukleotide) transkribiert (*primary*-microRNAs, pri-microRNA). Diese pri-microRNA-Transkripte besitzen die für eine Transkription durch die RNA-Polymerase II typische 3'-seitige Polyadenylierung (Aukerman & Sakai, 2003; Kurihara & Watanabe, 2004) sowie eine 5'-Kappe (Xie *et al.*, 2005). Zudem besitzen einige, jedoch nicht alle microRNA-Gene eine für die RNA-Polymerase II typische TATA-Box (Xie *et al.*, 2005). Zudem konnte

an zwei microRNA-kodierenden Bereichen aus *Oryza sativa* gezeigt werden, dass innerhalb des vermuteten microRNA-Vorläufers (*precursor*-microRNA, pre-microRNA) eine mögliche Exon-*Junction* liegt. Dies legt den Schluss nahe, dass das Splicing eine grundlegende Voraussetzung für die Funktionsfähigkeit einiger microRNAs darstellt (Sunkar *et al.*, 2005). Die primären microRNA-Transkripte werden anschließend von einem *Dicer-like protein 1* (DCL1) und *hyponastic leaves 1* (HYL1) prozessiert. Das DCL1 besitzt zwei RNaseIII-Domänen, welche spezifisch doppelsträngige RNAs erkennen und das Ausschneiden der funktionalen Stem-Loop-Struktur (pre-microRNA) aus dem pri-microRNA-Transkript bewerkstelligen. An diesem Prozess ist das Protein HYL1 beteiligt, von dem bekannt ist, dass es zwei RNA-Bindedomänen (dsRBD) besitzt und spezifisch doppelsträngige RNAs erkennt und bindet (Song *et al.*, 2007). Die pre-microRNA-Strukturen bilden charakteristische, unvollständig gepaarte Stem-Loop-Sekundärstrukturen aus (siehe Abbildung 1.4). In tierischen Zellen sind sie ca. 70 Nukleotide lang, während ihre Größe in pflanzlichen Zellen stärker divergiert. Hier variiert sie zwischen 70 und 690 Nukleotiden bei einer durchschnittlichen Länge von ungefähr 170 Nukleotiden. Des Weiteren besitzen die pre-microRNA-Strukturen ein zwei Nukleotide umfassendes, 3'-seitig überhängendes Ende, welches für die weitere Prozessierung der pre-microRNA eine wichtige Funktion einnimmt (Sekundärstruktur in Abbildung 1.4 dargestellt). Die weitere Prozessierung der Stem-Loop-Strukturen übernehmen ebenfalls DCL1 und HYL1, welche auch an der Prozessierung der pri-microRNA-Sequenz beteiligt sind. Im Anschluss daran wird der microRNA/microRNA*-Duplex ausgeschnitten, der die reife microRNA-Sequenz beinhaltet. Der microRNA*-Strang bezeichnet dabei den zur reifen microRNA komplementären Strang. Der microRNA/microRNA*-Duplex wird anschließend von einer Methyltransferase (HEN1) am 2'O des jeweiligen 3'-Endes methyliert. Die Methylierung der 3'-Enden des microRNA/microRNA*-Duplex schützt diesen vor einer Degradation durch Exonukleasen (Yu *et al.*, 2005). In tierischen Zellen wird der microRNA/microRNA*-Duplex durch das Kerntransporter-Protein Exportin5 aus dem Nukleus in das Zytoplasma exportiert. In Pflanzen geschieht dies durch das Protein HASTY, welches eine signifikante Homologie zum tierischen Exportin5 aufweist (Park *et al.*, 2005). An dem Export des microRNA/microRNA*-Duplex ist weiterhin eine RAN-GTPase beteiligt, die essentiell für die Translokation von RNAs und Proteinen durch den nuklearen Poren-Komplex ist (Kim, 2004; Moore, 1998; Moy & Silver, 2002).

Das Zytoplasma ist der eigentliche Funktionsort der reifen microRNA-Sequenz. Die reife microRNA-Sequenz wird dort in den *RNA-induced silencing complex* (RIS-Komplex, RISC) integriert. Aus dem microRNA/microRNA*-Duplex wird dafür die Sequenz als reife microRNA-Sequenz gewählt, welche das thermodynamisch weniger stabile 5'-Ende besitzt (Khvorova *et al.*, 2003; Schwarz *et al.*, 2003). Der korrespondierende microRNA*-Strang wird sofort degradiert und steht somit nicht mehr als funktionelle microRNA zur Verfügung. Der RIS-Komplex führt die RNA-vermittelte Genregulation in Pflanzen sowie Tieren durch. Dabei fungiert der RIS-Komplex als eine Endonuklease, der die Ziel-mRNA zwischen den Nukleotiden zehn und elf der gepaarten microRNA schneidet. Der Hauptbestandteil jedes RIS-Komplexes ist ein Mitglied der Argonauten-Familie (AGO1), das eine zentrale PAZ-Domäne (benannt nach den beteiligten Proteinen PIWI (*P-element induced*

Abbildung 1.4: Sekundär-Struktur des *ath-mir156a*-microRNA-Vorläufers. Die Darstellung zeigt die Sekundärstruktur des in Pflanzen hoch konservierten *ath-mir156a*-Vorläufers. In rot ist die reife microRNA-Sequenz hervorgehoben, welche in *Arabidopsis thaliana* in die Regulation von sogenannten *Squamosa-Promotor Binding Proteins* involviert ist.

wimpy testis), Argonaut und Zwille) sowie eine C-terminale PIWI-Domäne besitzen. Die PIWI-Domäne bindet die microRNA an ihrem 5'-Ende, wohingegen die PAZ-Domäne das 3'-Ende einzelsträngiger RNAs bindet (Lingel *et al.*, 2004; Ma *et al.*, 2005; Parker *et al.*, 2004; Song *et al.*, 2003; Yan *et al.*, 2003). Weitere Studien zeigten, dass die PIWI-Domäne einer RNase H ähnelt, die RNA-Stränge eines RNA-DNA-Hybrids schneidet (Nowotny *et al.*, 2005). Strukturelle und biochemische Studien zeigten, dass AGO1 die Komponente mit Endonuklease-Aktivität im RIS-Komplex darstellt (Liu *et al.*, 2004; Song *et al.*, 2004). Nach bisherigem Kenntnisstand aus Experimenten an tierischen Zellen hängt es vom Grad der Komplementarität ab, ob eine Spaltung oder eine Inhibition der Translation der korrespondierenden Ziel-mRNA stattfindet (Hutvágner & Zamore, 2002; Zeng *et al.*, 2002). Bei perfekter Komplementarität wird die Ziel-mRNA geschnitten. Bei geringerer Komplementarität findet dagegen eine, über einen nicht identifizierten Mechanismus, Inhibition der Translation statt.

Small interfering RNA (siRNA)

Zur Klasse von sRNAs gehören die sogenannten *small interfering RNAs* (siRNA). Die siRNAs sind zuerst in Pflanzen beobachtetworden und dort in viele Regulations- und Kontrollprozesse involviert. Hierzu gehören ein RNA-vermitteltes Immunsystem gegen RNA-Viren (Almeida & Allshire, 2005), die epigenetische Modifizierung bestimmter Regionen im Genom (Grewal & Moazed, 2003), die Kontrolle von Transposons und Transgenen (Almeida & Allshire, 2005) sowie die transkriptionelle (Furner *et al.*, 1998) und posttranskriptionelle Regulation von mRNAs (Que & Jorgensen, 1998).

Der Unterschied zwischen siRNAs und den microRNAs liegt in ihrem Ursprung und ihrer Biogenese. Die microRNAs sind, wie in Abschnitt 1.3.1 beschrieben, genomisch kodierte Hairpin-Strukturen, die von einer RNA-Polymerase II transkribiert werden. Der Ursprung der siRNAs ist hingegen deutlich divergenter. Zum einen liegt der Ursprung in langen doppelsträngigen RNAs und zum anderen in langen, perfekt gepaarten einzelsträngigen RNAs. Sie entstammen endogenen genomischen Bereichen, können jedoch auch von exogener Natur sein. Endogene siRNAs entstammen aus Transposons, Retrotransposons, repetitiven Elementen im Genom sowie anormalen mRNAs. Exogene siRNAs besitzen ihren Ursprung in viralen oder transgenen RNAs. Eine gemeinsame Eigenschaft einiger dieser siRNAs ist das Vorliegen einer einzelsträngigen RNA, welche von einer RNA-abhängigen RNA-Polymerase (RDR-Polymerasen, RDRP) in einen RNA-Doppelstrang

überführt wird. Dabei sind unterschiedliche RDR-Polymerasen für RNA-Einzelstränge unterschiedlichen Ursprungs zuständig.

Ein Beispiel für den Ursprung von siRNAs sind die sogenannten *trans-acting siRNAs* (ta-siRNAs) (Peragine *et al.*, 2004; Vazquez *et al.*, 2004). Die ta-siRNAs haben ihren Ursprung in den sogenannten TAS-Genen, welche ebenfalls von der RNA-Polymerase II transkribiert werden. Diese ta-siRNAs werden im Anschluss von einem RIS-Komplex, welcher mit einer microRNA beladen ist, geschnitten. Eines der Teilstücke des ta-siRNA-Vorläufers wird von einer RNA-abhängigen-RNA-Polymerase (RDR6) in lange doppelsträngige RNAs überführt. Dieser lange RNA-Doppelstrang wird ebenfalls wieder von einem Mitglied der *Dicer-like proteins* (DCL4) in Kooperation mit einem RNA-Doppelstrang-bindenden Protein (DRB4) in sogenannte ta-siRNA-Duplices geschnitten. Dies führt zu zahlreichen neuen siRNA-Duplices, welche in die Regulation anderer Zielgene involviert sind. Diesen Vorgang der Produktion neuer siRNA-Sequenzen durch eine microRNA-Sequenz wird auch als Transitivität bezeichnet. Auch hier werden die entstandenen siRNA-Duplices von einer Methyltransferase (HEN1) am 2'-O der 3'-Enden methyliert und somit vor einer Degradation durch eine Exonuklease geschützt. Aus diesen siRNA-Duplices wird auch wieder der Strang mit der reifen siRNA in einen RIS-Komplex integriert. In RIS-Komplexen können unterschiedliche Mitglieder der Argonauten-Familie (AGO) integriert sein, welche die weitere Funktionalität des Komplexes bestimmen. Im Fall der ta-siRNAs können unterschiedliche Argonauten-Proteine beteiligt sein. Ist AGO1 in den RIS-Komplex inkorporiert, so findet je nach Grad der Komplementarität eine Degradation oder Inhibition der Translation statt (zur Übersicht Jones-Rhoades *et al.*, 2006).

Weiterhin gibt es in Pflanzen neben der posttranskriptionellen Regulation korrespondierender Ziel-mRNAs zusätzliche Regulationsmechanismen. Hierzu gehören unter anderem die systemische Verbreitung der kleinen RNAs in der Pflanze sowie die RNA-vermittelte Methylierung und die Regulation auf transkriptioneller Ebene. Für die Regulation auf transkriptioneller sowie posttranskriptioneller Ebene bzw. der systemischen Verbreitung des *Silencing*-Signals sind unterschiedliche siRNAs zuständig, welche sich durch ihre Länge unterscheiden (Hamilton *et al.*, 2002). Diese distinkten Größenklassen sind zum einen die siRNAs der Länge von 19–22 Nukleotiden sowie die siRNAs der Länge von 24–26 Nukleotiden (Hamilton *et al.*, 2002). Die Klasse kürzerer siRNAs scheint für die siRNA-vermittelte Degradation der Ziel-mRNAs verantwortlich zu sein, während die Klasse mit längeren siRNAs für die epigenetische Modifikation genomischer Regionen sowie für die systemische Inhibition der Zielgene zuständig zu sein scheint. Die Größe der kleinen RNAs wird durch das jeweils verantwortliche Dicer-Protein bestimmt. DCL1 ist vermutlich für die Produktion der 21 Nukleotide langen siRNAs sowie der microRNAs zuständig (Bonnet *et al.*, 2006). Das *Dicer-like protein 3* (DCL3) scheint dagegen für die Produktion von kleinen RNAs zuständig zu sein, die für die Heterochromatinanordnung benötigt werden. Eine vereinfachte Darstellung der Prozessierung von siRNAs ist in Abbildung 1.6 veranschaulicht.

1.4 Viroide

Die Viroide stellen eine eigene Klasse von nicht-Protein-kodierenden RNAs dar, werden jedoch aufgrund ihres exogenen Ursprungs formell nicht zu den ncRNAs gezählt. Viroide wurden zum ersten Mal in den 70er Jahren von Diener (1971) bei dem Versuch, den Erreger der Kartoffel-Spindel-Knollen-Sucht zu identifizieren, beschrieben. Da die Infektiösität von Extrakten erkrankter Pflanzen durch eine Behandlung mit Phenol und Proteasen nicht eliminiert werden konnte, postulierte Diener die Hypothese, dass es sich bei dem infektiösen Agens um eine „nackte“ RNA mit niedrigem Molekulargewicht handeln müsse. In nachfolgenden Arbeiten wurde die Sequenz der Viroide beschrieben, welche eine Länge von 246–401 Nukleotiden besitzen. Somit ist ihr Genom ca. zehnmal kleiner als das kleinste Genom eines RNA-Virus. Das Viroid-Genom ist zirkulär kovalent geschlossen und besitzt einen hohen Grad an Selbstkomplementarität, was zur Folge hat, dass der RNA-Zirkel in nativem Zustand eine stäbchenförmige Struktur einnimmt, die in Abbildung 1.7 am Beispiel des *Potato Spindle Tuber Viroid* (*PSTVd*) dargestellt ist. Dabei wechseln sich basengepaarte und ungepaarte Bereiche in der Sekundärstruktur des *PSTVd* ab. Weiterhin ist die Tatsache bemerkenswert, dass Viroide für keine Proteine kodieren, sodass jegliche Funktion wie Replikation und Transport von Wirtsfaktoren vermittelt werden muss (Gross *et al.*, 1978). Derzeit sind ca. 40 Viroid-Spezies mit verschiedenen Sequenzvarianten beschrieben. Diese lassen sich in zwei Familien einteilen: die *Avsunviroidae* und *Pospiviroidae*. Der namensgebende Vertreter der *Avsunviroidae* ist das ***Av****ocado* ***Sun*** *Blotch Viroid* (*AVSVd*), der namensgebende Vertreter der *Pospiviroidae* ist das *PSTVd*.

1.4.1 Organisation des Viroids am Beispiel des *PSTVd*

Mittlerweile sind viele Vertreter der *Pospiviroidae* bekannt, was einen Sequenzvergleich ermöglichte. Es zeigten sich Bereiche mit deutlich unterschiedlichen Sequenzhomologien, was eine Einteilung des *PSTVd* in verschiedene distinkte Domänen erlaubte. Die zentrale konservierte Region (CCR) ist durch ihren hohen GC-Gehalt thermodynamisch sehr stabil und in die Prozessierung von oligomeren Replikationsintermediaten zu Monomeren und deren Ligation zu Zirkeln involviert. An die CCR schließt sich rechtsseitig die weniger konservierte variable Region (VR) und linksseitig die pathogenitätsmodulierende Region (PM oder *Virulence-Modulating*, VM) an. Nukleotidaustausche in dieser Region sind für stammspezifische Symptomatiken verantwortlich. Diese stammspezifischen Symptomatiken reichen von relativ milden Symptomen bis hin zu nahezu letalen Symptomen, wobei dieser Unterschied schon durch einen einzigen Nukleotidaustausch bedingt sein kann. Begrenzt wird das *PSTVd* durch zwei terminale Loops, *Terminal Left* (TL) und *Terminal Right* (TR). Das Fehlen offener Enden verleiht den Viroiden zudem eine gewisse Resistenz gegenüber RNasen. Verzweigungen oder tertiäre Strukturelemente wurden für das *PSTVd* ausgeschlossen, jedoch sind einige der ungepaarten Bereiche durch Ausbildung nicht-kanonischer Basenpaare in der Lage, komplexe Strukturmotive zu bilden, welche aus anderen nicht-kodierenden RNAs bekannt sind. Ein Beispiel für ein solches Motiv ist der zentral gelegene Loop E, welcher eine große Ähnlichkeit mit dem namens-

gebenden Loop E der 5S rRNA oder dem Sarcin/Ricin-Loop aus der 23S rRNA aufweist. Über solche lokalen Strukturelemente scheint das Viroid mit seiner molekularen Umwelt zu kommunizieren. Somit beinhaltet die Sekundärstruktur einen Teil der genetischen Information, während die Rolle der Primärstruktur des Viroids noch nicht geklärt werden konnte.

1.4.2 Replikationszyklus

Die Replikation des *Potato Spindle Tuber Viroid* (*PSTVd*) ist im Zellkern lokalisiert und wird von einer wirtseigenen DNA-abhängigen RNA-Polymerase II durchgeführt. Die Replikation des *PSTVd* folgt dabei einem asymmetrischen Replikationszyklus (Branch *et al.*, 1981), bei dem die RNA-Polymerase II zunächst den RNA-Zirkel mehrfach umläuft und somit in einen multimeren Gegenstrang transkribiert. Dieser nach Definition als (−)-Strang bezeichnete Strang dient wiederum als Matrize für die Synthese eines multimeren (+)-Strangs, der zunächst in die monomeren Einheiten geschnitten und schließlich wieder zu zirkulären Molekülen ligiert wird. Welche Enzyme an der Spaltung und Ligation beteiligt sind, ist derzeit noch nicht geklärt.

Die Replikation der Vertreter der *Avsunviroidae* (*Avsunviroidae*) unterscheidet sich grundlegend vom Replikationszyklus der *Pospiviroidae*. Im Replikationszyklus der *Avsunviroidae* wurden neben den monomeren reifen Viroiden zusätzlich monomere (−)-strängige zirkuläre Replikationsintermediate beobachtet. Somit folgt die Replikation der *Avsunviroidae* einem symmetrischen Replikationszyklus (Daròs *et al.*, 2006). Des Weiteren bilden die Vertreter der *Avsunviroidae* eine Ribozymstruktur aus, welche eine autokatalytische Prozessierung bewirkt; somit wird für deren Spaltung in die monomeren Einheiten kein weiterer Wirtsfaktor benötigt. Ein weiterer wichtiger Unterschied in der Replikation dieser beiden großen Familien von Viroiden ist die Lokalisation derselbigen: *Avsunviroidae* sind in den Chloroplasten lokalisiert und rekrutieren zur Replikation eine plastidäre RNA-Polymerase, während *Pospiviroidae* im Nukleus repliziert werden.

1.4.3 Transport und Pathogenität

Die Infektion einer Pflanze mit Viroiden geschieht durch Verletzungen der Zelloberfläche mit anschließendem Eindringen der Viroid-Moleküle. Die Infektion erfolgt in der Regel systemisch, was dazu führt, dass nahezu alle Pflanzenteile von Viroiden befallen werden. Dabei erfolgt der Transport in der Pflanze keineswegs passiv durch Diffusionsprozesse, vielmehr sind spezifische, durch den Wirt vermittelte Faktoren an diesem Prozess beteiligt (Ding *et al.*, 2005).

Für die Replikation des Viroids muss dieses zunächst in den Zellkern gelangen. Zu diesem Zweck wird das Viroid mit hoher Wahrscheinlichkeit in einem Ribonukleoprotein-Komplex (RNP-Komplex) integriert und spezifisch in den Kern transportiert. Als Mediator dieses Transports wird derzeit das Viroid-bindende Protein VirP1 diskutiert (de Alba *et al.*, 2003). VirP1 besitzt ein Kern-Lokalisations-Signal und bindet das Viroid spezifisch an

einem asymmetrischen Loop in der terminalen rechten Region (TR-Region, siehe Abbildung 1.7), dem sogenannten RY-Motiv. Wegen der charakteristischen Basenabfolge, abwechselnd Purin-Pyrimidin (RY nach der IUPAC-Nomenklatur), wird dieses Bindemotiv als RY-Motiv bezeichnet.

Der Langstrecken-Transport in der Pflanze erfolgt über das Phloem. Dabei folgt der Transport des Viroids dem Strom der Photoassimilate von den photosynthetisch aktiven Organen hin zu den Trieben der Blattspitze (Palukaitis, 1987). Es handelt sich jedoch auch bei dem Langstrecken-Transport nicht um einen passiven Transport-Mechanismus, vielmehr ist auch dieser Transport von Wirtsfaktoren vermittelt (Owens *et al.*, 2001; Stark-Lorenzen *et al.*, 1997).

Der Pathogenitätsmechanismus der Viroide ist derzeit noch weitestgehend unverstanden. Es ist nicht geklärt, weshalb unterschiedliche *PSTVd*-Varianten unterschiedlich stark ausgeprägte Symptome ausbilden und schon einzelne Nukleotidaustausche solche drastischen Unterschiede bewirken können. Anfänglich wurde die pathogene Wirkung der Viroide auf die Wechselwirkung der Viroide mit essentiellen Wirtsfaktoren zurückgeführt. Dabei wurden unterschiedlichste Ansätze diskutiert, zu denen unter anderem Beeinflussung des mRNA-Splicings oder eine gestörte rRNA-Reifung durch Bildung von Basenpaaren gehörten (Diener, 2001). Derzeit werden andere Ansätze zur Erklärung der Pathogenität diskutiert, wonach die Pathogenität auf die Bildung von kleinen Viroid-spezifischen RNAs zurückzuführen ist (Itaya *et al.*, 2001; Papaefthimiou *et al.*, 2001), die von den in Abschnitt 1.3.1 beschriebenen siRNAs und miRNAs nur durch ihre Sequenz unterscheidbar sind. Die in infizierten Pflanzen beobachteten kleinen Viroid-spezifischen RNAs könnten dabei als eine Reaktion der Pflanze auf die eindringenden RNA-Stränge verstanden werden (Tabler & Tsagris, 2004). Weiterhin fällt auf, dass die stäbchenförmige Sekundärstruktur eine auffallende Ähnlichkeit zu genomisch kodierten microRNA-Vorläufer-Strukturen aufweist, was ein Indiz dafür sein könnte, dass *PSTVd* von der Pflanze fälschlicherweise als microRNA-Vorläufer-Struktur erkannt und prozessiert wird. Zudem spricht die Lokalisation des *PSTVd* ebenfalls für solch einen induzierten, pathogenen Effekt, da bekannt ist, dass microRNA-Vorläufer-Strukturen im Nukleus der Zelle prozessiert werden. Somit könnte über bioinformatische Methoden eine qualitative Aussage über *PSTVd* und seine Ähnlichkeit zu wirtseigenen microRNAs ein weiteres Indiz darstellen, dass der Pathogenitätsmechanismus über eine fehlgeleitete Regulation essentieller Wirtsfaktoren vermittelt wird (siehe Abschnitt 1.7). Dieser Mechanismus zur Pathogenität der Viroide könnte zudem auch die unterschiedliche Ausprägung der Symptome erklären, da kleine RNAs durch Hybridisierung an entsprechende Ziel-mRNAs ihre destruktive Wirkung entfalten. Einzelne Nukleotidaustausche könnten somit unterschiedlich stark ausgeprägte Hybridisierungen bedingen, was zu einer unterschiedlich stark ausgeprägten Regulation der Ziel-mRNAs führen würde.

1.5 MicroRNA-Vorhersagemethoden

Die Hypothese der fehlgeleiteten Regulation essentieller Wirtsfaktoren über die Prozessierung der Viroide über den microRNA-Stoffwechselweg könnte durch microRNA-Vorhersagemethoden bestätigt oder abgelehnt werden. Derzeit existieren bereits einige microRNA-Vorhersagemethoden, welche durch unterschiedliche Vorgehensweisen gegebene Sequenzen als microRNA bzw. nicht-microRNA klassifizieren. Zwei dieser Ansätze werden im Folgenden vorgestellt: microHARVESTER (Dezulian *et al.*, 2006) erstellt Homologie-basiert Vorhersagen über das Vorliegen einer microRNA; findMiRNA (Adai *et al.*, 2005) erstellt basierend auf vorhandenen intergenischen Sequenzen und vorhandenen cDNA-Sequenzen eine Vorhersage über das Vorliegen von neuen microRNAs in den untersuchten intergenischen Sequenzen.

1.5.1 microHARVESTER

microHARVESTER (Dezulian *et al.*, 2006) ist ein Beispiel für einen Homologie-basierten Ansatz zur Identifikation von möglichen microRNA-Kandidaten. Dieser Ansatz geht davon aus, dass einige microRNA-Gene ein typisches Konservierungsmuster aufweisen. In microRNA-Familien sind die reifen microRNAs sehr stark konserviert, da ihre Sequenz essentiell für die Interaktion mit der zu regulierenden mRNA ist. Dagegen ist der microRNA*-Strang weniger konserviert, unterliegt jedoch einigen Restriktionen, da er mit der reifen microRNA Basenpaare ausbilden muss. Die restliche Sequenz der pre-microRNA unterliegt nur wenigen weiteren Beschränkungen und weist daher meist nur eine sehr geringe Konservierung auf. Der Bereich um den microRNA/microRNA*-Duplex muss die Möglichkeit besitzen, in eine stabile Hairpin-Struktur zu falten, was die einzige Beschränkung für diesen Bereich darstellt (Zhang *et al.*, 2006).

microHARVESTER beginnt seine Klassifizierung mit einer BLAST-Ähnlichkeitssuche zur Generierung einer Menge von Kandidatensequenzen, auf denen basierend die Identifikation der microRNA-Homologen stattfindet. Zur weiteren Unterscheidung von microRNA-Homologen und zufällig ähnlichen Sequenzen wurden einige Filter-Schritte eingebaut, welche charakteristische strukturelle Eigenschaften von pflanzlichen pre-microRNA-Strukturen ausnutzen. Die initiale BLAST-Suche findet zum einen mit der reifen microRNA-Sequenz und zum anderen mit der Sequenz der pre-microRNA statt. Da eine BLAST-Suche sehr viele falsch-positive Treffer produziert, werden in einem ersten Filter-Schritt all diejenigen Sequenzen verworfen, bei denen die alignierte Sequenz nicht den größten Teil der reifen microRNA-Sequenz abdeckt. In einem weiteren Filter-Schritt wird nun mit Hilfe eines modifizierten optimalen, paarweisen Smith-Waterman-Algorithmus (Smith & Waterman, 1981) die Lokalisation der reifen microRNA durch ein Alignment der Ausgangssequenz und den korrespondierenden BLAST-Treffern bestimmt. Unterscheidet sich die Länge der reifen microRNA-Sequenzen um mehr als zwei Nukleotide, so wird der entsprechende Kandidat verworfen. Der letzte Filter-Schritt umfasst die thermodynamische Faltung der Treffer-Sequenzen anhand von RNAfold und

daran anschließend die Lokalisation der microRNA*-Sequenz. Eine Sequenz wird verworfen, wenn bei mehr als sechs Nukleotiden keine Basenpaarung vorhergesagt werden kann. Die verbleibenden Sequenzen werden nun als mögliche, zur Ausgangssequenz homologe, Kandidaten angesehen.

Ein solcher Homologie-basierter Ansatz zur Identifikation von microRNA-Homologen ist durch die Menge an bekannten microRNA-Sequenzen begrenzt und eignet sich demnach auch nicht zur Identifikation von neuen, nicht sehr stark konservierten microRNA-Sequenzen.

1.5.2 findMiRNA

Ein weiterer Ansatz zur Identifikation von möglichen microRNA-Sequenzen wurde von Adai *et al.* (2005) beschrieben. findMiRNA ist ein sogenannter *single-genomic*-Ansatz und benötigt für die Vorhersage von möglichen microRNA-Sequenzen nur einen genomischen Sequenzdatensatz inklusive der Informationen über das Transkriptom und intergenische Bereiche. Die Vorgehensweise des Programms verfolgt prinzipiell den umgekehrten Ansatz zu bisherigen Programmen für die Vorhersage von microRNA-Sequenzen. Im ersten Schritt werden alle überlappenden Heptamere (Sequenzen der Länge 7) mit einigen Einschränkungen indiziert. Die Einschränkungen umfassen die Nukleotid-Komposition, wobei jedes Heptamer mindestens zwei Guanin- oder Cytosin-Nukleotide beinhalten muss. Heptamere mit sich wiederholenden Regionen der Länge zwei oder vier werden dagegen verworfen. Jedes betrachtete mRNA-Transkript wird in überlappende Heptamere aufgeteilt und mit den indizierten Heptameren der intergenischen Regionen bezüglich ihrer Komplementarität verglichen. Für jeden erfolgreichen Vergleich werden die umgebenden Sequenzen beider Heptamere verlängert und ohne die Einführung von Gaps aligniert. Die entstandenen Alignments werden in einer Fenstergröße von 18–25 Nukleotiden durchlaufen und nach dem folgenden Schema bewertet:

- kanonische Watson-Crick-Basenpaare werden mit einem Score von Zwei bewertet
- G:U-Wobble-Basenpaare werden mit einem Score von Eins bewertet
- Mismatche werden mit Null bewertet

Die errechneten Scores werden auf die Länge des alignierten Bereichs normiert und dabei der Bereich als microRNA vorhergesagt, der den normierten Score maximiert. Übersteigt der errechnete Score einen definierten Schwellenwert (normierter Score von $1,55$ bei mindestens 35 Punkten nach obigem Schema), so wird die die mögliche microRNA-Sequenz umgebende Region auf ihre Fähigkeit zur Ausbildung einer Hairpin-Sekundärstruktur untersucht. Dabei wird die mögliche microRNA-Sequenz in einer 400 Nukleotide umfassenden Region zentriert und der umliegende Bereich über einen dynamischen Programmieransatz weiter analysiert.

Der Ansatz der dynamischen Programmierung für überlappende Treffer (Durbin, 1998) aligniert die mögliche microRNA-Sequenz an die umgebende, 400 Nukleotide umfassende

Region durch eine modifizierte Version der Score-Matrix, welche für das Alignment der komplementären Bereiche verwendet wurde. Die Modifikation liegt einzig in der Vergabe eines negativen Scores für Mismatche; hier wird ein Score von -1 vergeben. Die vorhergesagte microRNA-Sequenz wird bewertet und nicht verworfen, wenn ihr normierter überlappender Score (Alignment-Score/microRNA-Länge) größer Eins ist. Die so prozessierten überlappenden Heptamere der Transkripte werden absteigend nach ihrem kombinierten Score aus beiden algorithmischen Ansätzen sortiert und ihre pre-microRNA-Sequenz als die Region definiert, welche von dem microRNA/microRNA*-Duplex flankiert wird.

1.6 Statistische Auswertungsmethoden

Da jede Vorhersage mit Fehlern behaftet ist, werden statistische Kenngrößen benötigt, welche als Maß für die Güte der Vorhersagemethode dienen. Hierzu wurden in der Statistik die Begriffe Sensitivität, Spezifität, Segreganz, Relevanz, Korrektklassifikationsrate sowie Falschklassifikationsrate definiert. Zur Definition der Begriffe werden die in Abbildung 1.8 dargestellten und erläuterten Abkürzungen TP, FP, TN und FN benötigt.

1.6.1 Statistische Kenngrößen

Der Begriff Sensitivität (Formel 1.1) wird auch als Richtig-Positiv-Rate oder Empfindlichkeit eines Klassifikators bezeichnet und beschreibt die Wahrscheinlichkeit, ein tatsächlich positives Objekt auch als positiv vorherzusagen. Die Sensitivität gibt den Teil der als positiv vorhergesagten Objekte im Verhältnis zur Gesamtheit der positiven Objekte an.

$$\text{Sensitivität} = \frac{TP}{TP + FN} \tag{1.1}$$

Die Spezifität (Formel 1.2) wird auch als Richtig-Negativ-Rate bezeichnet und beschreibt die Wahrscheinlichkeit, ein tatsächlich negatives Objekt als Negativ vorherzusagen. Die Spezifität gibt somit das Verhältnis der als negativ vorhergesagten Objekte zur Gesamtmenge der negativen Objekte an.

$$\text{Spezifität} = \frac{TN}{TN + FP} \tag{1.2}$$

Die Falsch-Positiv-Rate (Formel 1.3) ist ebenfalls ein Ausdruck für die Spezifität eines Klassifikators und beschreibt den Anteil, dass ein tatsächlich positives Objekt als negativ klassifiziert wurde.

$$\begin{aligned} \text{FP-Rate} &= 1 - \text{Spezifität} \\ &= \frac{FP}{FP + TN} \end{aligned} \tag{1.3}$$

Der Begriff der Relevanz (Formel 1.4) wird in der Statistik auch als Wirksamkeit des Klassifikators beschrieben. Die Relevanz ist die Wahrscheinlichkeit, dass ein positiv-klassifiziertes Objekt tatsächlich positiv ist. Die Relevanz gibt das Verhältnis der korrekterweise als positiv vorhergesagten Objekte an der Gesamtheit der als positiv vorhergesagten Objekte an.

$$\text{Relevanz} = \frac{TP}{TP + FP} \tag{1.4}$$

Die Segreganz oder Trennbarkeit eines Klassifikators (Formel 1.5) beschreibt die Wahrscheinlichkeit, dass ein negativ-klassifiziertes Objekt tatsächlich zur Gesamtmenge der negativen Objekte gehört. Die Segreganz gibt also den Anteil der korrekterweise als negativ vorhergesagten Objekte zur Gesamtmenge der als negativ vorhergesagten Objekte an.

$$\text{Segreganz} = \frac{TN}{TN + FN} \tag{1.5}$$

Sind durch eine Vorhersage Objekte in eine Klasse eingeteilt worden, so ist diese Einteilung zumeist mit Fehlern behaftet. Man spricht dann von einer sogenannten Falschklassifikation. Die Falschklassifikationsrate (Formel 1.7) beschreibt den Anteil an falschvorhergesagten Objekten an der Gesamtheit der zu klassifizierenden Objekte. Die Korrektklassifikationsrate (Formel 1.6) beschreibt analog dazu den prozentualen Anteil der korrekt-klassifizierten Objekte an der Gesamtmenge.

$$\text{Korrektklassifikationsrate} = \frac{TP + TN}{TP + FP + FN + TN} \tag{1.6}$$

$$\text{Falschklassifikationsrate} = \frac{FP + FN}{TP + FP + FN + TN} \tag{1.7}$$

1.6.2 *Receiver Operating Characteristics* (ROC)

Eine ROC-Kurve ist eine Methode zur Messung der Effizienz von sogenannten Klassifikatoren und wurde im Zweiten Weltkrieg zur Analyse von Radarbildern angewendet.

Später wurde diese Methode zur Analyse und Visualisierung des Verhaltens von Systemen in medizinischen Bereichen wie z. B. der Krankheitsdiagnostik übernommen. Hier dient die ROC-Analyse häufig als Interpretations- und Evaluationshilfe in der medizinischen Diagnostik. In den späten 80er Jahren wurde von Spackman (1989) die Methode zum ersten Mal zu Auswertungs- und Vergleichszwecken von Algorithmen beschrieben und dient heute als Standard-Methode zur Evaluierung und Visualisierung von Klassifizierungsalgorithmen. Ein Beispiel für eine solche ROC-Kurve ist in Abbildung 1.9 zu sehen.

Wie schon in Abschnitt 1.6 erwähnt, ist jeder Versuch einer Klassifizierung mit Fehlern behaftet. Der Klassifizierung liegt ein sogenanntes Klassifizierungsmodell zugrunde, welches für die Objekte der Testmenge eine Klassifikation bzw. Zuordnung zu einer bestimmten Klasse trifft. Diese Zuordnung steht in Abhängigkeit eines Schwellenwerts, der darüber entscheidet, in welche Klasse die Zuordnung stattfindet. Um die Effizienz eines Klassifikators in Abhängigkeit eines zu wählenden Schwellenwertes beurteilen zu können, werden die in Abbildung 1.8 und Abschnitt 1.6.1 eingeführten statistischen Begriffe benötigt. Ein Klassifikator, der Objekte aus einer Testmenge in nur zwei diskrete Klassen unterteilt, wird auch als binärer Klassifikator bezeichnet. Dieser unterteilt, wie der Name schon sagt, gegebene Objekte in zwei Klassen. Wird davon ausgegangen, dass die tatsächliche Einteilung bekannt ist, lassen sich den klassifizierten Objekten die Messgrößen aus der Tabelle 1.8 zuordnen.

Die für die ROC-Analyse benötigten Kenngrößen können aus diesen Messgrößen hergeleitet werden: die Richtig-Positiv-Rate oder Sensitivität (Formel 1.1, Y-Achse) und die Falsch-Positiv-Rate (Formel 1.3, X-Achse). Diese Werte, die ein Maß für die Effizienz des Klassifikators darstellen, werden gegeneinander im Koordinatensystem aufgetragen. Da die Werte für Richtig-Positiv-Rate oder Sensitivität und Falsch-Positiv-Rate immer Werte zwischen 0 und 1 einnehmen, markieren diese Werte die Grenzen des Raumes, in dem die Punkte aufgetragen werden. Dieser durch die Messgrößen definierte Raum wird auch als ROC-Raum bezeichnet.

Um Aussagen über die Effizienz eines Klassifikators beurteilen zu können, werden an dieser Stelle einige markante Punkte des ROC-Raums erläutert. Es existieren, wie in Abbildung 1.9 dargestellt, fünf Punkte bzw. Bereiche des ROC-Raumes, welche unterschiedlichste Aussagekraft besitzen. Liegt ein Wertepaar auf der Diagonalen zwischen den Punkten $(0, 0)$ und $(1, 1)$, so entspricht der analysierte Klassifikator dem bloßen Raten des Ergebnisses, da in diesem Bereich mit gleicher Wahrscheinlichkeit richtig-positive und falsch-positive vorhergesagt werden. Befindet sich ein Wertepaar in der Nähe des Punktes $(0, 1)$ (siehe Abbildung 1.9 Punkt B), ist dies kennzeichnend für einen guten Klassifikator, welcher nahezu nur richtig-positive Aussagen macht und im Gegenzug annähernd keine falsch-positiven Klassifizierungen trifft. Allgemein ausgedrückt bedeutet dies, dass ein Klassifikator umso besser ist, je geringer die Distanz zwischen dem Punkt $(0, 1)$ und seinem tatsächlichen Wertepaar ist. Der Punkt A an Position $(0, 0)$ in der Abbildung 1.9 bedeutet, dass keinerlei falsch-positive Vorhersagen getroffen werden, allerdings auch keinerlei richtig-positive Vorhersagen. Dies ist vergleichbar mit dem bloßen Ablehnen eines jeden gegebenen Objekts. Der Punkt C an der Position $(1, 1)$ ist der umgekehrte Fall zum

Punkt A an der Position $(0, 0)$. Dieser beschreibt das Verhalten, dass alle richtig-positiven Objekte als positiv, im Gegenzug jedoch auch alle richtig-negativen als positiv vorhergesagt werden. Anders ausgedrückt bedeutet dies, dass jedes Objekt ohne Ausnahme als positiv klassifiziert wird. Des Weiteren besitzt der ROC-Raum noch zwei charakteristische Regionen, welche besonders hervorzuheben sind. Die Fläche um den Buchstaben D (in Abbildung 1.9 rot dargestellt) wird als konservativ bezeichnet und steht für geringe Falsch-Positiv-Raten bei gleichzeitig geringen Richtig-Positiv-Raten; der Bereich um den Buchstaben E (in Abbildung 1.9 rot dargestellt) wird als liberal bezeichnet und beschreibt ein Diskriminierungsverhalten mit hohen Richtig-Positiv-Raten, die jedoch mit hohen Falsch-Positiv-Raten einhergehen.

Für die Berechnung eines Schwellenwerts, der beide Klassen optimal diskriminiert, wird die gesamte Menge an möglichen Schwellenwerten getestet und anhand derer die Falsch-Positiv- und Richtig-Positiv-Rate berechnet und gegeneinander im ROC-Raum aufgetragen. Der optimale Schwellenwert zeichnet sich durch eine möglichst hohe Richtig-Positiv-Rate bei gleichzeitig möglichst niedriger Falsch-Positiv-Rate aus. Der optimale Schwellenwert eines Klassifikators wird hauptsächlich auf zweierlei Art und Weise berechnet. Zum einen ist es möglich, die Distanz zwischen den aufgetragenen Punkten und dem Punkt der bestmöglichen Diskriminierung zu berechnen und den Schwellenwert zu wählen, für den diese Distanz minimal wird. Eine zweite Möglichkeit ist der sogenannte Youden-Index (Perkins & Schisterman, 2006) (1.8). Der Youden-Index markiert den optimalen Schwellenwert, welcher die Summe aus Sensitivität und Spezifität maximiert. Der Youden-Index (J) ist definiert als:

$$J = \max\{\text{Sensitivität} + \text{Spezifität} - 1\} \tag{1.8}$$

Neben der maximierten Summe aus Sensitivität und Spezifität hat das Wertepaar des Youden-Index die Eigenschaft, dass die vertikale Distanz zwischen der ROC-Kurve und der Diagonalen des bloßen Ratens maximal ist (siehe Abbildung 1.9).

1.6.3 Fläche unter der Kurve (*Area under the Curve*, AUC)

Wie im vorigen Abschnitt bereits beschrieben ist die ROC-Kurve eine zwei-dimensionale Darstellung der Effizienz eines Klassifikators. Um mehrere Vorhersagemethoden bezüglich ihrer Effizienz vergleichen zu können ist es nötig, diese Effizienz in einem einzigen skalaren Wert auszudrücken. Eine gebräuchliche Methode ist die Berechnung der Fläche unterhalb der Kurve, auch als *Area under the Curve* (AUC) bezeichnet. Da die Fläche unterhalb der Kurve Element des ROC-Raumes ist, liegen die Werte für die AUC immer zwischen 0 und 1.

Anstelle der Flächenberechnung durch Integration wird meist die Fläche durch das Anlegen von geometrischen Figuren an die Kurve angenähert. Durch Rechtecke bzw. Trapeze wird mit unterschiedlicher Genauigkeit die Fläche unter Kurve abgeschätzt. Die Fläche des von zwei benachbarten Datenpunkten aufgespannten Trapezes wird über die folgenden

Formel berechnet:

$$F_{\text{Trapez}} = \frac{(x2 - x1) \cdot (y1 + y2)}{2} \tag{1.9}$$

Die Fläche wird für alle benachbarten Datenpunkte ausgerechnet und aufsummiert. Bei entsprechend hoher Anzahl Datenpunkte erhält man somit eine relativ genaue Berechnung der Fläche unter der ROC-Kurve.

Die AUC wird meist im Zusammenhang mit einem 95%igem Konfidenzintervall angegeben, da die AUC eines bestimmte Klassifikators kein absoluter Wert ist, sondern eine statistische Größe, die einem Fehler unterliegt. Dies bedeutet, dass der wahre Wert der AUC innerhalb des 95%igen Konfidenzintervalls liegt, jedoch eine Chance von 5% vorliegt, dass diese Annahme falsch ist. Liegt demnach die untere Grenze der AUC bei einem 95%igem Konfidenzintervall oberhalb von 0,5, so ist der Klassifikator statistisch signifikant besser, als eine Vorhersage basierend auf dem bloßen Raten des Ergebnisses.

Die AUC ist somit ein Maß der Effizienz eines Klassifikators und wird als Durchschnittswert der Sensitivität für alle Werte der Spezifität interpretiert. Die AUC kann nur Werte zwischen 0 und 1 annehmen, da die X- und Y-Achse nur jeweils Werte zwischen 0 und 1 annehmen können. Grundsätzlich lässt sich sagen, dass AUC-Werte nahe 1 eine sehr gute Effizienz des vorliegenden Klassifikators beschreiben. Praktisch betrachtet ist die untere Grenze der AUC eines Klassifikators eine Wert von 0,5, da dies dem bloßen Raten entspricht.

1.7 Ziel dieser Arbeit

Viroide stellen eine eigene Klasse von pflanzlichen Krankheitserregern dar und entfalten ihre Funktionalität über die Rekrutierung von wirtseigenen Faktoren. Sie werden aufgrund der fehlenden Protein-kodierenden Regionen als nicht-kodierende RNAs bezeichnet und besitzen aufgrund der charakteristischen Sekundärstruktur eine nicht zu vernachlässigende Ähnlichkeit zu bekannten microRNA-Vorläufer-Strukturen. Zudem wurden in infizierten Pflanzen bereits kleine Viroid-spezifische RNAs mit einer Länge von vornehmlich 21–22 Nukleotiden identifiziert. Die microRNA-ähnliche Sekundärstruktur der Viroide sowie das Vorliegen von kleinen Viroid-spezifischen RNAs, welche eine für reife microRNA-Sequenzen charakteristische Länge besitzen, lässt die Hypothese zu, dass die Viroid-vermittelte Pathogenität in der Pflanze auf einen microRNA-ähnlichen Mechanismus zurückzuführen ist. Des Weiteren scheint es möglich zu sein, dass das Viroid von der wirtseigenen RNA-Interferenz-Maschinerie aufgrund der charakteristischen Sekundärstruktur und der Lokalisation im Zellkern fälschlicherweise als microRNA-Vorläufer erkannt und prozessiert wird. Von *PSTVd* ist bekannt, dass es in *Arabidopsis thaliana* repliziert und prozessiert wird, einzig der Transport der Viroide in alle Pflanzenteile scheint nicht vermittelt zu werden (Daròs & Flores, 2004; Matousek *et al.*, 2004). In der Literatur gibt es derzeit keine Hinweise auf eine Akkumulation kleiner Viroid-spezifischer RNAs in

Arabidopsis thaliana, jedoch könnte es aufgrund der Konservierung des Regulationsmechanismus durch kleine RNAs auch dort zu einer Prozessierung der Viroide und Akkumulation kleiner Viroid-spezifischer RNAs kommen. Dies muss jedoch nicht zwangsläufig mit einer Ausprägung von Symptomen einhergehen, da mögliche Zielsequenzen eventuell zu divergent sind und demnach keine Bindestellen besitzen.

Zelluläre microRNAs bedürfen charakteristischer Sekundärstrukturen und eventuell auch Primärstrukturelemente für ihre Prozessierung in der Zelle. Sollte die Viroid-vermittelte Pathogenität über den microRNA-vermittelten Regulationsmechanismus funktionieren, so müsste das Viroid Sequenz- und/oder Strukturcharakteristika aufweisen, die denen der wirtseigenen microRNAs ähnlich sind. Derzeit gibt es einige microRNA-Vorhersageprogramme, welche jedoch für diesen Ansatz nicht erfüllbare Voraussetzungen mit sich bringen. Dies wird besonders am Beispiel von microHARVESTER deutlich, das einzig über eine Homologie-Suche neue microRNA-Kandidaten ermittelt. Weitere Ansätze zur *de novo*-Identifikation von microRNA-Kandidaten benötigen zu spezifische Informationen über den zugrundeliegenden Organismus wie z. B. Kenntnis über das Transkriptom und intergenischer Bereiche (siehe Abschnitt 1.5.2). *Arabidopsis thaliana* als zugrundeliegenden Organismus zu wählen begründet sich zumindest in zweierlei Hinsicht: die Fülle an verfügbaren Sequenzinformationen und experimentellen Daten sowie die Tatsache, dass zumindest ein Teil der Funktionalität des *PSTVd* durch Rekrutierung von *Arabidopsis thaliana*-spezifischen Faktoren vermittelt wird.

Da an dieser Stelle nicht davon auszugehen ist, dass die Viroide eine Homologie zu bekannten microRNA-Familien aufweisen und derzeit auch keinerlei detaillierte Informationen über das Transkriptom natürlicher Wirtspflanzen vorliegen, sollte im Rahmen dieser Arbeit ein neuer Ansatz zur *de novo*-Identifikation von microRNA-kodierenden Bereichen allein basierend auf Informationen über Sequenz und Struktur von bekannten microRNAs entwickelt werden.

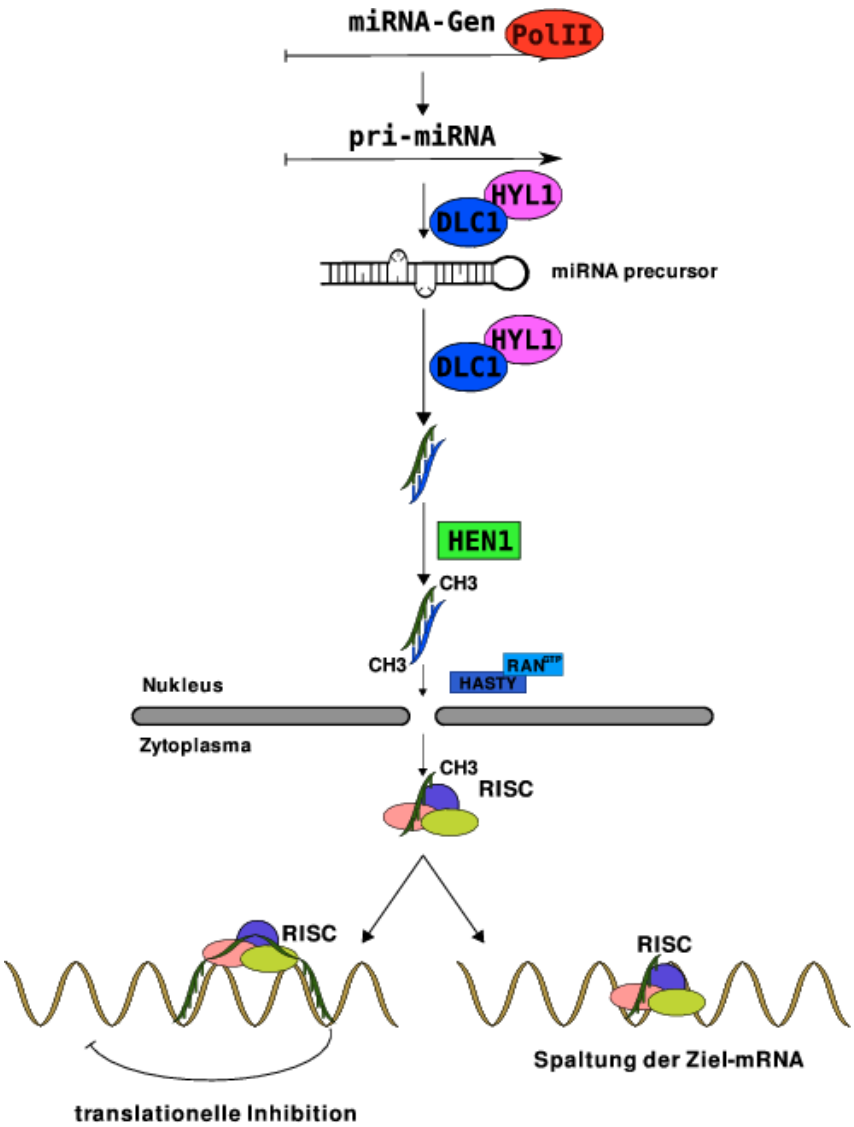

Abbildung 1.5: Schematische Darstellung der microRNA-Biogenese in Pflanzen. Die microRNA-kodierende Region wird von der RNA-Polymerase II transkribiert. Daraus resultieren die sogenannten pri-microRNA, welche mehrere 100 bis 1.000 Nukleotide lang sein kann. Das *Dicer-like protein 1* (DCL1) und *hyponastic leaves 1* (HYL1) prozessieren die pri-microRNAs und schneiden den microRNA-Vorläufer heraus. DCL1 übernimmt in Zusammenarbeit mit HYL1 zusätzlich die Prozessierung des microRNA-Vorläufers und schneidet den microRNA/microRNA*-Duplex heraus, der die reife microRNA enthält. Anschließend findet eine 2'O-Methylierung der 3'-Enden durch HEN1, einer Methyl-Transferase, statt. Der Export in das Zytoplasma wird durch das Exportin5-Ortholog HASTY vermittelt. Von Exportin5 ist bekannt, dass es in tierischen Zellen für den Kernexport kleiner RNAs verantwortlich ist (Park *et al.*, 2005). Im Zytoplasma wird die reife microRNA in den *RNA-induced silencing complex* (RIS-Komplex) eingelagert. Der RIS-Komplex beinhaltet unter anderem ein Mitglied der Argonauten-Familie (AGO1), welches essentiell für die microRNA-vermittelte Regulation ist. AGO1 vermittelt anschließend je nach Komplementaritätsgrad die Inhibition der Translation oder die Spaltung der korrespondierenden Ziel-mRNA (Hutvágner & Zamore, 2002; Zeng *et al.*, 2002).

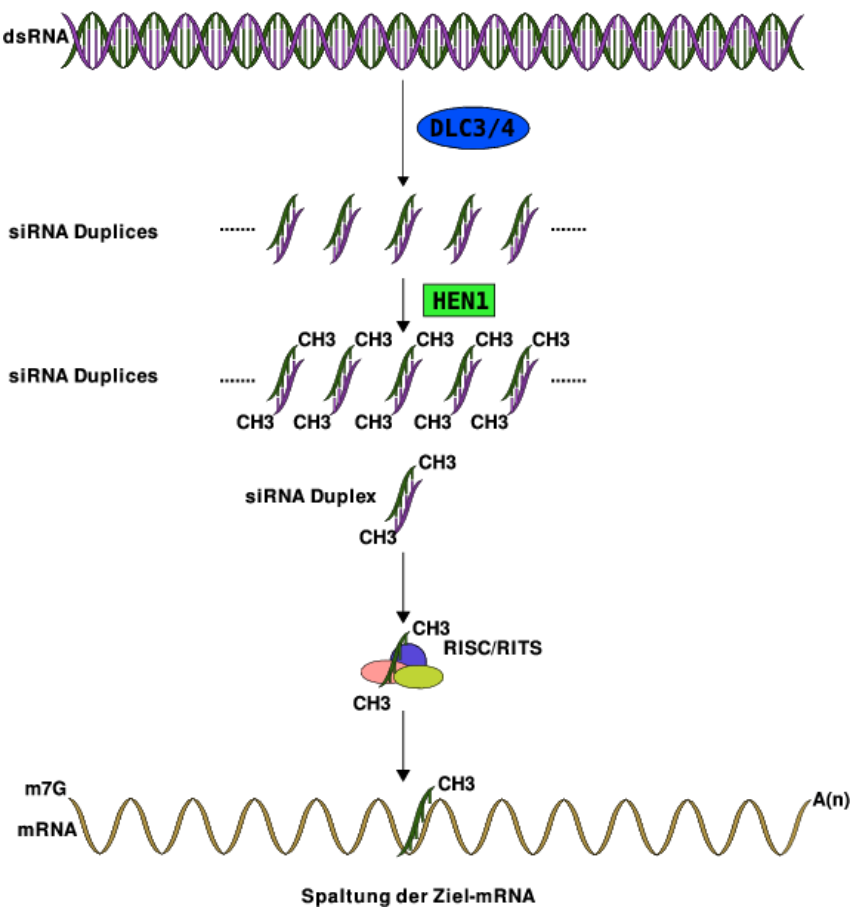

Abbildung 1.6: *Small-interfering-RNA*-Biogenese. Ursprung der siRNAs ist ein RNA-Doppelstrang, der unterschiedlichen Ursprungs sein kann. Der lange RNA-Doppelstrang wird von einem Mitglied der *Dicer-like proteins* (DCL), je nach Ursprung DCL3 oder DCL4, in kurze siRNA-Duplices geschnitten, welche anschließend von HEN1, einer Methyl-Transferase, am 2'-O des 3'-Endes methyliert werden. Diese siRNA-Duplices werden je nach Ursprung und Funktion in unterschiedliche *Silencing*-Komplexe integriert, welche die Funktion vermitteln. Zum einen gibt es den *RNA-induced silencing complex* (RISC), welcher die Inhibition der Translation vermittelt, und zum anderen den *RNA-induced transcriptional silencing*-Komplex (RITS), welcher die Chromatin-Modifikationen vermittelt. Unterschiedliche Mitglieder der Argonauten-Familie sind in die RIS- und RITS-Komplexe involviert, von ihnen hängt die jeweilige vermittelte Funktion ab (Übersicht in Jones-Rhoades *et al.*).

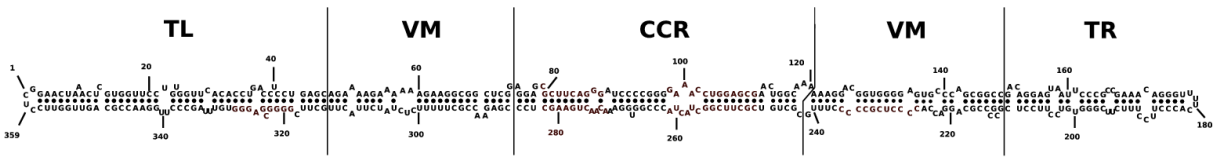

Abbildung 1.7: Sekundärstruktur des *PSTVd*. Die Darstellung zeigt die Sekundärstruktur des *PSTVd*. Die fünf durch Sequenzhomologien ermittelten funktionell relevanten Domänen (Keese & Symons, 1985) sind durch TL und TR (linker und rechter Terminus), VM (Virulenz-modulierende Region), CCR (zentral konservierte Region) und VR (variable Region) gekennzeichnet. Basenpaarungen sind durch Punkte dargestellt.

	Positive Klasse	Negative Klasse
als Positiv vorhergesagt	Richtig Positiv (=TP)	Falsch Positiv (=FP)
als Negativ vorhergesagt	Falsch Negativ (=FN)	Richtig Negativ (=TN)

Abbildung 1.8: Wahrheits-Matrix. Die sogenannte Wahrheitsmatrix wird auch als Vierfelder-Tabelle bezeichnet und veranschaulicht die Begriffe Richtig-Positiv, Falsch-Positiv, Falsch-Negativ und Richtig-Negativ. Die in dieser Arbeit verwendeten Abkürzungen sind in rot dargestellt.

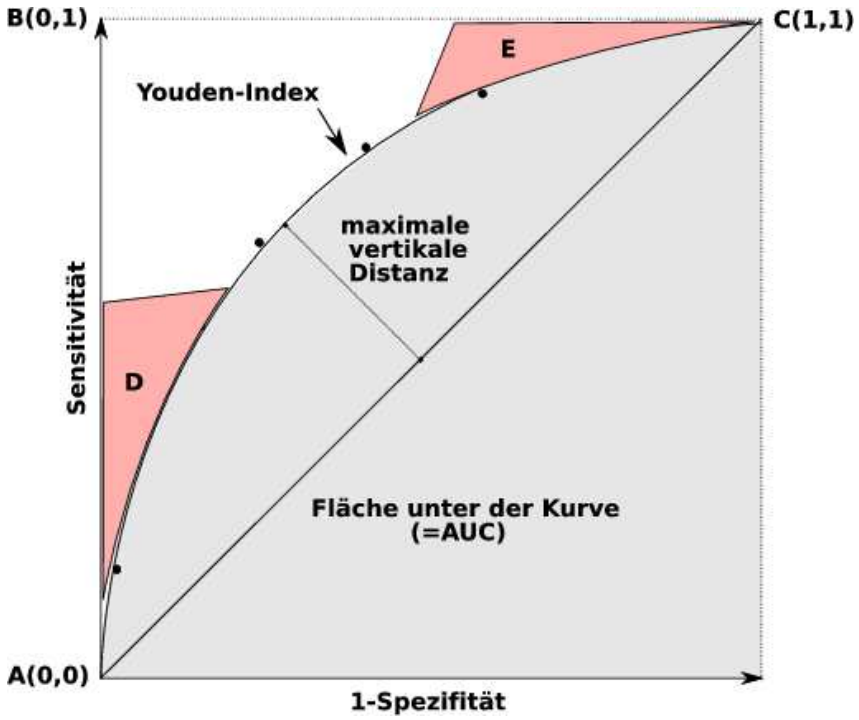

Abbildung 1.9: Beispielhafte ROC-Kurve. Beispielhafte Darstellung einer ROC-Kurve mit fünf charakteristischen Punkten: **A** bedeutet keine falsch-positive Klassifizierung, allerdings auch keine richtig-positive; der Punkt **B** beschreibt eine perfekte Klassifizierung; der Punkt **C** bedeutet nur positive Klassifizierungen, allerdings auch viele falsch-positive; die Fläche um den Buchstaben **D** (rot) wird als konservativ bezeichnet und steht für geringe Falsch-Positiv-Raten, allerdings auch geringe Richtig-Positiv-Raten; die Fläche um den Buchstaben **E** (rot) wird als liberal bezeichnet und bedeutet, dass nahezu alle Objekte korrekt klassifiziert werden, was aber mit hohen Falsch-Positiv-Raten einhergeht. Der Youden-Index stellt beispielhaft das Wertepaar dar, an dem die Summe aus Sensitivität und Spezifität maximal wird. Weiterhin stellt dieser Punkt die maximal mögliche Distanz zwischen der Diagonalen des bloßen Ratens sowie der ROC-Kurve dar.

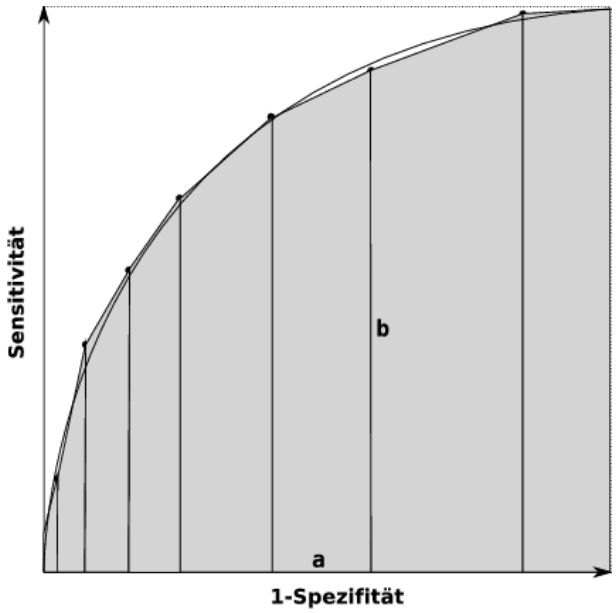

Abbildung 1.10: Abschätzung der *Area under the Curve*. Eine Möglichkeit zur Abschätzung der Fläche unter der Kurve (*Area under the Curve*, AUC) ist in dieser Abbildung gezeigt. Zwei benachbarte Punkte im ROC-Raum definieren ein Trapez im ROC-Raum, dessen Fläche berechnet werden kann. Wird für alle Punkte-Paare im ROC-Raum ein solches Trapez erstellt, dessen Fläche berechnet und die Flächeninhalte aller Trapeze aufsummiert, so ergibt sich daraus eine Abschätzung für die Fläche unter der Kurve.

2

Material und Methoden

2.1 Entwicklungsumgebung

2.1.1 Hardware

Die Entwicklungsarbeit wurde auf einem Dualcoreprozessorsystem mit zwei Intel-Pentium D 2,80 GHz Prozessoren sowie 1 GB Arbeitsspeicher durchgeführt.
Die Rechnungen fanden zudem auf einem Dualprozessorsystem mit zwei Intel Xeon CPU 3,0 GHz sowie 4 GB Arbeitsspeicher statt.

Die Kommunikation dieser Computer fand in einem geswitchten 100 MB Ethernet-Netzwerk statt.

2.1.2 Betriebssystem

Als Betriebssystem diente Ubuntu[1] mit einem GNU/Linux-Betriebssystemkern. Der Betriebssystemkern lag in Version 2.6.22-14 vor.

2.2 Programme

2.2.1 Perl

Perl[2] ist eine Skriptsprache und eignet sich in der Bioinformatik besonders zum Verarbeiten von Zeichenketten und ist daher sehr geeignet für den Umgang mit großen geno-

[1] `http://www.ubuntu.com/`

[2] `http://www.perl.com/`

mischen Sequenzdaten. Perl wurde in Version 5.8.8 verwendet. Zudem wurde Perl für die Entwicklung des Programms Yet Another miRNA Predictor (YamP) verwendet.

2.2.2 BioPerl

BioPerl[3] (Version 1.4) wurde zum Erstellen von Skripten benutzt und ist eine Zusammenstellung von Perl-Modulen, welche die Entwicklung von Skripten zu bioinformatischen Fragestellungen erleichtert. Es beinhaltet keine kompletten Programme, bietet jedoch wiederverwendbare Module zu häufig auftretenden bioinformatischen Problemen, wie z. B. Sequenz-Manipulation, Zugang zu biologischen Datenbanken, Schnittstellen zu diversen molekularbiologischen Programmen sowie die Möglichkeit, deren Ausgaben zu analysieren.

2.2.3 HyPa

Für die Identifikation neuer möglicher microRNAs war es nötig, positiv klassifizierte Sequenzen aus intergenischen und intronischen Sequenzbereichen im Genom von *Arabidopsis thaliana* lokalisieren zu können. Zu diesem Zweck wurde das Programm HyPa (Gräf, 2004; Strothmann, 2005) verwendet, welches auf der Musterbeschreibungssprache HyPaL (Gräf, 2004; Gräf *et al.*, 2001) beruht und anhand erstellter Muster genomische Sequenzdaten effizient durchsuchen kann. Das Programm HyPa wurde in der Version von 2006-09-01 verwendet.

2.2.4 RNAfold

Das Programm RNAfold, entwickelt von Hofacker *et al.* (1994), wurde zu Auswertungszwecken und der grafischen Darstellung von möglichen microRNA-Kandidaten verwendet. RNAfold berechnet die thermodynamisch optimale Struktur einer gegebenen Nukleinsäuresequenz über die Verwendung thermodynamischer Parameter, die aus experimentell ermittelten Daten abgeleitet wurden. Verwendete wurde die Option „-T“, welche die Temperatur spezifiziert, bei der die thermodynamisch optimale Sekundärstruktur einer gegebenen RNA-Sequenz berechnet werden soll. RNAfold lag in Version 1.6 vor und ist Teil des Vienna-RNA-Paketes (Hofacker, 2003).

2.2.5 RNALfold

Das Programm RNALfold, entwickelt von Hofacker *et al.* (2004), wurde für die lokale Faltung von großen genomischen Sequenzbereichen verwendet um lokal stabile Strukturen aufzudecken, welche Basis für die anschließende Klassifizierung waren. RNALfold berechnet lokal thermodynamisch optimale Strukturen in langen Sequenzen unter

[3] `http://bioperl.org/`

Berücksichtigung einer maximalen Spannweite für die Bildung eines Basenpaares. Die verwendeten Optionen waren die „-T“-Option zur Angabe der Temperatur, bei der die Faltung berechnet werden soll, sowie die „-L“-Option, welche die maximale Spannweite erlaubter Basenpaare in der Sequenz vorgibt. RNALfold lag in Version 1.6 vor und ist Teil des Vienna-RNA-Paketes.

2.2.6 RNAshapes

Das Programm RNAshapes, entwickelt von Giegerich *et al.* (2004), wurde zur Berechnung von optimalen und suboptimalen Sekundärstrukturen verwendet. Da thermodynamisch optimale Strukturen nicht den nativen Strukturen entsprechen müssen, ist es nötig auch suboptimale Strukturen zu betrachten. Viele Programme bewerkstelligen die suboptimale Faltung, produzieren jedoch große Mengen an Ausgabe und zeichnen sich durch eine lange Laufzeit aus. RNAshapes ist ebenfalls in der Lage, optimale und suboptimale Sekundärstrukturen gegebener Sequenzen zu berechnen. RNAshapes gibt jedoch nur die Sekundärstrukturen aus, die sich in ihrer Sekundärstruktur auf einem gegebenen Abstraktionsniveau von anderen Sekundärstrukturen unterscheiden. Die niedrigste Abstraktionsstufe enthält fast alle Informationen einer Sekundärstruktur, hält jedoch nur die Länge gepaarter Bereiche und einzelsträngiger Regionen zurück. Eine mittlere Abstraktionsstufe enthält Informationen über vorliegende gepaarte Bereiche, die Position von Bulge-Loops und interner Loops, hält jedoch Informationen über andere einzelsträngige Bereiche zurück. Die höchste Abstraktionsstufe enthält ausschließlich Informationen über gepaarte Bereiche, jedoch nicht über deren Länge sowie aller einzelsträngiger Bereiche. Als Abstraktionsniveau wurde das mittlere Niveau über die Option „-t“ gewählt, da die Kenntnis der Länge nicht-gepaarter Bereiche innerhalb einer Sekundärstruktur für diese Arbeit nicht nötig war. RNAshapes lag in Version 2.0 vor.

2.2.7 miRU

Die Vorhersage von neuen möglichen microRNA-Vorläufer-Sequenzen erfordert zudem die Suche nach möglichen Zielgenen. Der miRU-Webserver (Zhang, 2005) sucht in den mRNA-Sequenzen von *Arabidopsis thaliana* nach komplementären Sequenzen und weist ihnen eine Bewertung zu. Dabei werden Standard-Watson-Crick-Basenpaare mit 0 und G : U-Wobble-Basenpaare mit $0,5$ gewichtet. Die Einführung von Insertionen bzw. Deletionen wird mit $2,0$ gewichtet und jeder andere Mismatch mit $1,0$. Zudem wird jeder Mismatch in der komplementären Sequenz zwischen den Positionen 2 und 7 zusätzlich mit $0,5$ bestraft, da diese sogenannte *Seed*-Region für die Funktion essentiell ist und nur wenige Mismatche zulässt. Bleibt der Gesamtscore einer komplementären Sequenz unterhalb eines zu wählenden Schwellenwerts, so wird diese als mögliche Hybridisierungsstelle ausgegeben.

2.2.8 miRanda

Die in Abschnitt 2.2.7 vorgestellte Methode zur Vorhersage von Zielgenen basiert allein auf Sequenzkomplementarität und bezieht keinerlei thermodynamische Berechnungen mit ein. Das Programm miRanda (Enright *et al.*, 2003) bietet eine solche Erweiterung an. Zunächst wird mit Hilfe eines dynamischen Programmier-Ansatzes eine, zu einer vorgegebenen kleinen RNA, komplementäre Stelle in gegebenen genomischen Sequenzen gesucht. Die Erweiterung stellt nun die Berechnung der minimalen freien Energie (*Minimum Free Energy*, *mfe*) für die vorher ermittelten komplementären Bindestellen dar. Diese Berechnung stellt eine weitere Verifikation für die Bindung einer gegebenen kleinen RNA an die vorliegenden genomischen Sequenzen dar. Das Programm miRanda wurde mit den Standardeinstellungen und der zusätzlichen Option zur zufälligen Neusortierung der Positionen in der ermittelten komplementären Bindestelle und der damit verbundenen Z-Score-Berechnung ausgeführt. Das Programm miRanda wurde in Version 1.9 verwendet.

2.2.9 RNAup

RNAup (Hofacker *et al.*, 1994; Mückstein *et al.*, 2006) wurde für die Berechnung von RNA-RNA-Interaktionen zur Lokalisation und Verifikation von möglichen Bindestellen herangezogen. Die RNA-RNA-Interaktion wird dabei in zwei Schritte aufgeteilt. Zunächst wird die Wahrscheinlichkeit berechnet, das die Bindestelle in der Ziel-RNA ungepaart vorliegt. Als nächstes wird die Energie der Bindung der beiden RNA-Sequenzen berechnet. Dies geschieht unter der Annahme, dass die Bindestelle ungepaart ist. Als Option wurde die „-Xf“-Option verwendet, welche vorgibt, wie Sequenzen in einer Datei z verarbeiten sind. RNAup wurde in Version 1.6 verwendet und ist wie RNAfold und RNALfold Teil des Vienna-RNA-Paketes.

2.2.10 *Receiver-Operator-Characteristics*-Analyse (ROC-Analyse)

Die Bewertung der Effizienz eines Klassifikators wurde mit Hilfe einer ROC-Analyse durchgeführt. Hierzu wurde das Perl-Paket Statistics::ROC[4] in Version 0.04 verwendet. Das Perl-Paket enthält Bibliotheken zur Analyse gegebener statistischer Messgrößen. Des Weiteren ist die Möglichkeit der grafischen Darstellung inklusive des Abspeicherns der ermittelten Koordinaten gegeben, welche für die Berechnung der Fläche unter der Kurve (AUC) benötigt wurden. Eine genaue Beschreibung der ROC-Analyse wird in Abschnitt 1.6.2 gegeben.

2.3 Sequenzdaten

Die Entwicklung des Programms YamP (Yet Another miRNA Predictor), auf das in Kapitel 3 näher eingegangen wird, benötigte zu Test- und Trainingszwecken Sequenz-

[4] `http://search.cpan.org/~hakestler/Statistics-ROC-0.04/lib/Statistics/ROC.pm`

datensätze (siehe Abschnitt 3.1). Hierzu wurden die folgenden Sequenzdatenbanken verwendet:

2.3.1 NCBI

Die Sequenzdatenbank des *National Center for Biotechnology Information*[5] (NCBI) diente als Bezugsquelle für genomische Sequenzdaten von *Arabidopsis thaliana*, welche für die genaue Lokalisation neuer möglicher microRNA-Sequenzen in den genomischen Sequenzdaten benötigt wurden. Die genomischen Sequenzdaten inklusive ihrer Annotationen lagen in ihrer Aktualisierung vom 20.April 2007 vor.

2.3.2 *The microRNA-Registry* (miRBase)

Die miRBase[6] (Griffiths-Jones, 2004; Griffiths-Jones *et al.*, 2006) ist eine Sammlung bekannter und teilweise verifizierter microRNA-Sequenzen. Zusätzlich gespeicherte Informationen zu den Sequenzen der microRNA-Vorläufer umfassen die reife microRNA-Sequenz, den Bereich im Genom sowie teilweise Informationen über die korrespondierenden regulierten Zielgene. Zu Trainings- und Testzwecken wurden die reifen microRNAs und Vorläufer-microRNAs aus *Arabidopsis thaliana* von der miRBase bezogen. Es lagen zwei Versionen der miRBase vor. Der erste Datensatz lag in Version 7.0 vor und beinhaltete 117 microRNA-Sequenzen. Der zweite Datensatz lag in Version 10.0 vor und beinhaltete neben den 117 microRNA-Sequenzen der Version 7.0 67 weitere microRNA-Sequenzen. Beide Datensätze dienten in den Abschnitten 3.1 und 3.3 als Trainings- und Testdatensatz.

2.3.3 Rfam

Für das Training und Testen des im Rahmen dieser Arbeit entwickelten Programms YamP wurden richtig-negative Sequenzen benötigt. Zu diesem Zweck wurden nicht-kodierende RNAs („non-coding RNAs“, ncRNAs) aus der Rfam-Datenbank[7] (*RNA families database of alignments and Covariance Models*) (Griffiths-Jones *et al.*, 2005) bezogen. Die Rfam-Datenbank beinhaltet eine große Sammlung multipler Sequenzalignments diverser ncRNA-Familien, darunter auch einige microRNA-Vertreter. Im Rahmen dieser Arbeit wurde der Rfam-Alignmentdatensatz von den microRNA-Vertretern bereinigt und die verbleibenden Sequenzen durch das Entfernen der Gaps dealigniert. Für die Zusammenstellung des richtig-negativen Sequenzdatensatzes wurde weiterhin nur ein zufälliger Vertreter einer ncRNA-Familie in den Trainingsdatensatz aufgenommen. Der Rfam-Alignmentdatensatz lag in Version 7.0 vor und beinhaltete zu diesem Zeitpunkt 455 verschiedene ncRNA-Familien.

[5] `http://www.ncbi.nlm.nih.gov/`

[6] `http://microrna.sanger.ac.uk/sequences/index.shtml`

[7] `http://rfam.janelia.org`

2.3.4 TAIR

The Arabidopsis Information Resource (TAIR; (Garcia-Hernandez *et al.*, 2002)) diente als Sequenzdatenbank für die Suche nach möglichen neuen microRNA-Sequenzen sowie zur Identifikation von Genen, welche durch die neuen microRNAs reguliert werden. Weiterhin dienten die kodierenden Sequenzen („coding sequences“, CDS) als Basis für eine zufällige Auswahl von Hairpin-Strukturen, welche in den Trainingsdatensätzen als richtig-negative Sequenzen eingesetzt wurden (siehe Abschnitt 3.1.2). Hierzu wurden von TAIR[8] intergenische, intronische sowie kodierende Sequenzen bezogen. Die verwendeten Sequenzdaten lagen in der Version 7 vor und waren auf den 25.04.2007 (CDS), den 26.02.2007 (intergenische Sequenzen) sowie den 27.02.2007 (intronische Sequenzen) datiert.

2.3.5 *The Gene Index Project*

Eine weitere Sequenzdatenbank stellt das *Gene Index Project*[9] dar. Das *Gene Index Project* bietet eine Sammlung von Genen für eine ganze Fülle von Organismen an und versucht Informationen über die funktionellen Rollen der Gene und ihrer Produkte zu sammeln. Vom *Gene Index Project* wurden die aktuellen Gen-Informationen von *Arabidopsis thaliana* verwendet, die auf den 16.06.2006 datiert waren.

2.4 Expressionsdatenbanken

Die Vorhersage von möglichen microRNA-Sequenzen bedurfte einer zusätzlichen Validierung über öffentliche Expressionsdatenbanken, welche Informationen über das Vorhandensein von kleinen RNAs in den entsprechenden Sequenzbereichen zur Verfügung stellen. Hierzu wurden zwei Expressionsdatenbanken verwendet, die unterschiedliche Informationen enthalten.

Die erste in dieser Arbeit verwendete Datenbank ist die ASRP[10]-Datenbank (Gustafson *et al.*, 2005). Die in dieser Expressionsdatenbank eingetragenen und dem Genom zugeordneten kleinen RNAs sind in Pyrosequencing-Experimenten (Margulies *et al.*, 2005) sequenziert worden. Die sequenzierten kleinen RNAs stammen dabei unterschiedlichen Entwicklungsstadien von *Arabidopsis thaliana* (Col-0 Ökotyp) wie z. B. dem Keimling, dem Blütenstand oder dem Blattgewebe. Viele der eingetragenen kleinen RNAs entstammen zusätzlich aus Mutanten von *Arabidopsis thaliana*, welche einen Defekt in einer der Komponenten des sRNA-Stoffwechsels aufwiesen. Hierzu gehören unter anderem alle vier bekannten *Dicer-like proteins* (*dcl1-7*, *dcl2-1*, *dcl3-1* und *dcl4-2*) sowie drei bekannte RNA-abhängiger RNA-Polymerasen (RDRP) (*rdr1-1*, *rdr2-1* und *rdr6-15*). Derzeit umfasst die ASRP-Expressionsdatenbank 218.585 einzigartige kleine RNAs (Backman *et al.*, 2008)

[8] `http://www.arabidopsis.org`

[9] `http://compbio.dfci.harvard.edu/tgi/`

[10] `http://asrp.cgrb.oregonstate.edu/`

Die zweite im Rahmen dieser Arbeit verwendete Expressionsdatenbank ist die *Arabidopsis MPSS plus*[11]-Datenbank. Die *Massive Parallel Signature Sequencing*-Methode (MPSS-Methode; Brenner *et al.* (2000)) identifiziert nahezu jedes in der betrachteten Zelle vorkommende Transkript. Hierzu werden 17bp- bzw. 20bp-Signaturen von jedem vorliegenden Transkript erstellt und dem Genom zugeordnet. Bei dieser Methode wird davon ausgegangen, dass 17bp bzw. 20bp ausreichen um ein Transkript eindeutig zu identifizieren. Weiterhin setzt diese Methode voraus, dass bereits Informationen über das Transkriptom bzw. des Genoms vorliegen, um Aussagen über die Expressionsrate von beobachteten Transkripten zu treffen. In einem modifizierten Ansatz wurden die in *Arabidopsis thaliana* exprimierten kleinen RNAs sequenziert, in einer Datenbank hinterlegt und der Öffentlichkeit zu Auswertungszwecken zur Verfügung gestellt.

[11] `http://mpss.udel.edu/at/`

3

Ergebnisse

Das Ziel dieser Arbeit war die Identifikation möglicher microRNAs in genomischen Sequenzdaten von *Arabidopsis thaliana* sowie die möglicher Bereiche im Genom des *Potato Spindle Tuber Viroid* (*PSTVd*), die das Potenzial besitzen, Ursprung kleiner Viroid-spezifischer RNAs zu sein. Zu diesem Zweck wurde das Programm Yet Another miRNA Predictor (YamP) entwickelt. Derzeit sind 184 mehr oder weniger konservierte microRNA-Sequenzen bekannt und es liegt die Vermutung nahe, dass weitere existieren müssen (Lindow & Krogh, 2005). Schätzungen, die auf der Komplementarität von intergenischen zu Protein-kodierenden Bereichen beruhen, ergaben, dass weitere mehrere hundert mögliche microRNA-Sequenzen im Genom von *Arabidopsis thaliana* kodiert sein müssen. Besonders die nicht-konservierten microRNA-Sequenzen sind schwer zu lokalisieren, da keinerlei Homologie-Untersuchung von bereits bekannten microRNA-Sequenzen aus anderen Organismen durchgeführt werden kann. Daher ist es von großem Interesse weitere, nicht konservierte oder auch konservierte microRNA-Sequenzen in *Arabidopsis thaliana* zu identifizieren. Das Programm YamP erstellt auf Basis bekannter microRNA-Sequenzen und -Strukturen ein statistisches Modell und klassifiziert darauf basierend mögliche Kandidatensequenzen. Vor die eigentliche Klassifizierung wurden zudem einige Filter-Schritte geschaltet, die besonders im genomischen Einsatz den Suchraum reduzieren und die Menge der falsch-positiven Sequenzen minimieren sollten.

Das Programm YamP lässt sich in zwei voneinander abhängige Einheiten aufteilen:
Das Training umfasst dabei die Aufstellung eines statistischen Modells und beruht auf einem richtig-positiven microRNA-Sequenzdatensatz und einem richtig-negativen nicht-microRNA-Sequenzdatensatz mit unterschiedlicher Komposition. Das Training bildet die erste Einheit des hier entwickelten Programms (siehe Abschnitt 3.1).

Die zweite Einheit des Programms YamP beschäftigt sich mit dem Klassifizierungsprozess und somit der Vorhersage neuer möglicher microRNA-Kandidaten, der auf dem in der ersten Einheit erstellten statistischen Modell beruht. Die Klassifizierungseinheit bildet somit den Kern des hier entwickelten Programms, auf das in Abschnitt 3.2 im Detail eingegangen wird.

3.1 Training

Die eigentliche Ausführung des Programms YamP für die Klassifizierung gegebener Sequenzen benötigt zunächst eine Trainingsphase, in der ein statistisches Modell von richtig-positiven microRNA- und richtig-negativen nicht-microRNA-Sequenzen erstellt wird, dass auf Sequenz- und Struktureigenschaften beruht. Das in der Trainingsphase erstellte statistische Modell stellt die Grundlage für die Klassifizierung dar (Abschnitt 3.2). Im Folgenden werden zunächst die zugrundeliegenden Sequenzdatensätze, die Trainingsphase und schließlich die statistische Erhebung des Modells im Detail erläutert bevor in Abschnitt 3.2 detailliert auf den Klassifizierungsprozess eingegangen wird. Das Training ist dabei einmal initial für einen spezifischen Organismus und die zugrundeliegenden Trainingsdatensätze durchzuführen bevor sie dem folgenden Klassifizierungsprozess zur Verfügung stehen.

3.1.1 Richtig-positive Sequenzen

Der Trainingsdatensatz der richtig-positiven Sequenzen umfasste alle in der miRBase annotierten und teilweise verifizierten microRNA-Sequenzen. Dabei stehen dem Training zwei verschiedene richtig-positive Datensätze zur Verfügung. Dabei handelte es sich bei dem ersten richtig-positiven Datensatz um die microRNA-Sequenzen der miRBase in Version 7.0. Der zweite richtig-positive Trainingsdatensatz umfasste neben der miRBase in Version 7.0 mit 117 microRNA-Sequenzen weitere 67 microRNA-Sequenzen und beinhaltete insgesamt 184 teilweise verifizierte und annotierte microRNA-Sequenzen (miRBase in Version 10.0).

3.1.2 Richtig-negative Sequenzen

Der richtig-negative Trainingsdatensatz war eine Zusammenstellung von unterschiedlichen richtig-negativen Sequenzen. Die Grundmenge der richtig-negativen Sequenzen bestand aus den folgenden Sequenzklassen:

- 710 mRNA-Sequenzen
- 631 tRNA-Sequenzen
- 602 5S-rRNA-Sequenzen
- 63 5,8S-rRNA-Sequenzen
- 454 RNA-Familien aus der Rfam-Datenbank
- 10.000 pseudo-Hairpin-Sequenzen aus *Arabidopsis thaliana*
- 8.494 pseudo-Hairpin-Sequenzen aus *Homo sapiens* (Ng & Mishra, 2007).

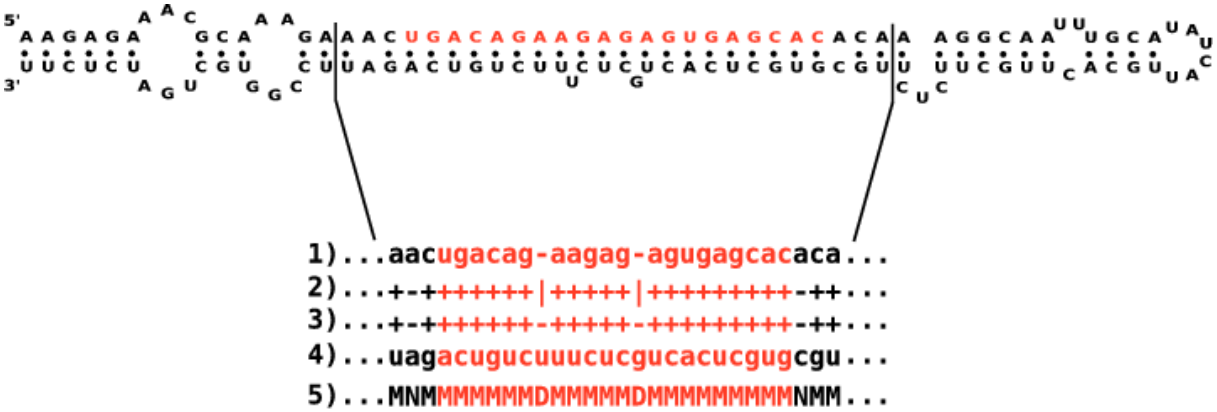

Abbildung 3.1: Umformatierte Struktur der *ath-mir156a*-Vorläufer-microRNA. Die Abbildung zeigt im obigen Teil die Sekundärstruktur des *ath-mir156a*-Vorläufers. In rot ist die reife microRNA-Sequenz dargestellt. Die Sekundärstruktur wird in ein Alignment-ähnliches Format überführt, welches im unteren Teil des Bildes ausschnittsweise zu sehen ist. Die erste Zeile beschreibt dabei die Vorwärts-Sequenz der Hairpin-Struktur, die vierte Zeile die Rückwärts-Sequenz der Hairpin-Struktur. In rot ist wieder der Bereich der reifen microRNA-Sequenz dargestellt. Die Zeilen 2 und 3 beschreiben den strukturellen Zustand, in dem sich die gegenüberstehenden Nukleotide befinden. Ein „+“ bedeutet ein Basenpaar der beiden Nukleotide, ein „-“ in beiden Zeilen bedeutet ein interner Loop und ein „|“ bedeutet einen internen asymmetrischen Loop oder einen Bulge-Loop. Die Zeile 5 beschreibt den strukturellen Zustand, in dem sich die gegenüberliegenden Nukleotide befinden. Ein **M** beschreibt eine Basenpaarung, ein **N** beschreibt die Nicht-Paarung, ein **D** die Deletion eines Nukleotids im Vorwärts-Strang und ein **I** eine Insertion im Vorwärts-Strang.

Die pseudo-Hairpins stellen Sequenzen dar, welche über die Fähigkeit zur Ausbildung einer Hairpin-Struktur aus den *complementary DNA*-Sequenzen (cDNA) ausgewählt und extrahiert wurden. Der richtig-negative Trainingsdatensatz stellt im weiteren Verlauf dieser Arbeit eine unterschiedliche Zusammenstellung der aufgelisteten Sequenzgruppen dar. Diese wurden dabei so ausgewählt, dass das Vorliegen von falsch-negativen Sequenzen mit Sicherheit ausgeschlossen werden konnte.

3.1.3 Umformatierung der Sekundärstrukturen

Die Trainingsphase beginnt mit der Faltung der richtig-negativen und richtig-positiven Sequenzdatensätze durch RNAShapes. RNAShapes wird verwendet, um eine Strukturverteilung um die *Minimum Free Energy*-Struktur (*mfe*-Struktur) zu berechnen. Die so berechneten Strukturen, maximal drei Strukturen pro Sequenz, werden anschließend in ein Alignment-ähnliches Format überführt (siehe Abbildung 3.1) (Nam *et al.*, 2005). Charakteristisch für die umformatierten Sekundärstrukturen ist, dass Vorwärts- und Rückwärts-Strang durch die Einführung von Gaps, bedingt durch asymmetrische Loops und Bulge-Loops, dieselbe Länge besitzen. Die umformatierte Struktur im unteren Bereich der Abbildung 3.1 zeigt den Bereich um die reife microRNA-Sequenz im *ath-mir156a*-Vorläufer von *Arabidopsis thaliana*.

Weist die Sekundärstruktur einer Sequenz mehr als einen distinkten Hairpin auf, so werden die von RNAshapes berechneten Sekundärstrukturen aufgeteilt, jeder Hairpin extrahiert und gesondert in der Trainingsphase betrachtet. Da die reifen microRNA-Sequenzen stets in Hairpin-Regionen auftreten, wurden alle terminalen Loops sowie überhängenden Enden der Hairpin-Strukturen entfernt, so dass nur die tatsächlichen Hairpin-Regionen in das Training und die Klassifizierung eingingen (Zhang *et al.*, 2006).

3.1.4 Trainingsverfahren

Dem eigentlichen Training des Programms liegen die in Abschnitt 3.1.3 eingeführten Alignment-ähnlichen Strukturen zugrunde und werden im Folgenden als umformatierte Strukturen bezeichnet. Diese werden linear von vorne nach hinten durchlaufen, wobei sich gegenüberliegende Nukleotidpaare in Abhängigkeit ihres jeweiligen Zustands analysiert werden. Der Begriff Zustand besitzt im Folgenden eine duale Bedeutung. Zum einen beschreibt er den strukturellen Zustand und zum anderen die Umgebung, in der sich das betrachtete Nukleotidpaar befindet. Der strukturelle Zustand beschreibt, wie in Abbildung 3.1 erwähnt, das Vorliegen eines Basenpaares, eines nicht gepaarten Nukleotidpaares bzw. die Einführung einer Insertion oder Deletion. Der Umgebungszustand definiert, ob sich das betrachtete Nukleotidpaar innerhalb oder außerhalb des Bereiches einer reifen microRNA-Sequenz befindet. Eine Übersicht der im Programm YamP implementierten Zustandsmöglichkeiten ist in Tabelle 3.1 gegeben.

Die beiden oberen Zustandsnamen, *state-is* und *state-not* in der Tabelle 3.1, beinhalten sowohl den strukturellen Zustand als auch den umgebungsabhängigen Zustand. Dabei sei erwähnt, dass nur der Bereich der reifen microRNA-Sequenz als *state-is* und der Rest der microRNA-Vorläufer-Struktur als *state-not* betrachtet wird. Richtig-negative Sequenzen werden komplett als *state-not* betrachtet. Die umgebungsabhängigen Zustände werden weiterhin noch in Abhängigkeit ihres strukturellen Zustands des aktuell betrachteten Nukleotidpaares unterteilt. Die fünfte Zeile in Abbildung 3.1 zeigt dabei drei der vier möglichen strukturellen Zustände. Dabei bedeutet **M** (Match) das Vorhandensein eines Basenpaares, **N** (Mismatch) zwei sich gegenüberstehende, nicht gepaarte Nukleotide, **D** (Deletion) die Deletion eines Nukleotids im Vorwärts-Strang (Zeile 1) und **I** die Insertion eines Nukleotids im Rückwärts-Strang (Zeile 4). Somit liegen für die beiden oberen Zustandsnamen insgesamt acht unterschiedliche Zustandskombinationen vor.

Die unteren vier Zustandsnamen, *di-nuk-is*, *di-nuk-not*, *di-nuk-plus* und *di-nuk-min* aus der Tabelle 3.1 beschreiben hingegen nur den umgebungsabhängigen Zustand. Der Zustand *di-nuk-is* bedeutet dabei, dass das betrachtete Nukleotidpaar innerhalb einer reifen microRNA-Sequenz liegt. Der Zustand *di-nuk-not* beschreibt hingegen, dass das betrachtete Nukleotidpaar außerhalb einer reifen microRNA-Sequenz liegt. Die beiden weiteren Zustände beschreiben dabei den Übergang zwischen den beiden eben erwähnten Zuständen. Dabei gilt das 5'-Nukleotidpaar einer reifen microRNA-Sequenz als *di-nuk-plus*- und das 3'-Nukleotidpaar als *di-nuk-min*-Zustand.

Tabelle 3.1: Strukturelle und umgebungsabhängige Zustände Die vorliegende Tabelle zeigt die im Programm YamP verwendeten strukturellen und umgebungsabhängigen Zustände, welche als Basis für die Erhebung des statistischen Modells aus Abschnitt 3.1 dienten.

Zustandsname	Beschreibung
state-is	gegenüberstehendes Nukleotidpaar in Abhängigkeit des strukturellen Zustandes und innerhalb einer reifen microRNA-Sequenz
state-not	gegenüberstehendes Nukleotidpaar in Abhängigkeit des strukturellen Zustandes und außerhalb einer reifen microRNA-Sequenz
di-nuk-is	benachbartes Nukleotidpaar befindet sich im Bereich einer reifen microRNA
di-nuk-not	benachbartes Nukleotidpaar befindet sich nicht im Bereich einer reifen microRNA
di-nuk-plus	benachbartes Nukleotidpaar beschreibt den Übergang vom Zustand *ist-nicht-mirna* zu *ist-mirna*
di-nuk-min	benachbartes Nukleotidpaar beschreibt den Übergang vom Zustand *ist-mirna* zu *ist-nicht-mirna*

Eine weitere Prämisse für die Einteilung der Strukturen in die jeweiligen Zustände ist die Länge der betrachteten Hairpin-Struktur. Dabei muss die umformatierte Struktur eine Mindestlänge von 19 Nukleotiden besitzen, was der minimalen Länge einer reifen microRNA-Sequenz entspricht.

Es werden zunächst die Frequenzen der Nukleotidpaare in Abhängigkeit ihres umgebungsabhängigen und strukturellen Zustands ermittelt. Es werden für jeden umgebungsabhängigen und strukturellen Zustand die auftretenden Nukleotidpaare gezählt und durch die Gesamtzahl der auftretenden strukturellen Zustände dividiert. Dies bedeutet, dass sich die Frequenzen eines jeden strukturellen Zustands zu 1 aufsummieren. Bei den nur umgebungsabhängigen Zuständen wird die Frequenz durch Division durch die Gesamtzahl an Nukleotidpaaren für den jeweiligen umgebungsabhängigen Zustand ermittelt.

Die so ermittelten Frequenzen in Abhängigkeit des strukturellen und umgebungsabhängigen Zustands werden in einer Datei gespeichert und stehen der Klassifizierung im Abschnitt 3.2 zur Verfügung.

3.2 Klassifizierung

Die in Abschnitt 3.1 geschilderte Trainingsphase und das daraus resultierende statistische Modell der microRNA- und nicht-microRNA-Sequenzen steht im Folgenden der Klassifizierung gegebener Sequenzen zur Verfügung. Die eigentliche Klassifizierung neuer möglicher microRNA-Sequenzen ist Inhalt des Abschnitts 3.2.3. Dabei werden die grundlegende Funktionsweise, der eigentliche Klassifizierungsprozess, sowie der Einfluss des statistischen Modells auf die Klassifizierung im Detail erläutert.

3.2.1 Vorbereitende Maßnahmen

Das Programm YamP (Yet Another miRNA Predictor) beginnt zunächst mit dem Einlesen der Sequenzen und der Entscheidung, ob es sich um einen genomischen Ansatz handelt oder einzelne Sequenzen klassifiziert werden sollen. Dabei kann der genomische Ansatz das Einlesen eines gesamten Genoms oder Chromosoms bedeuten, aber auch das Einlesen und die anschließende Klassifizierung eines langen Sequenzstücks. Beides wird jedoch im Folgenden als genomischer Ansatz bezeichnet. Um eine möglichst gute thermodynamische Strukturvorhersage zu erreichen, wurde die Grenze zur Entscheidung für einen genomischen Einsatz auf 400 Nukleotide festgelegt. Von Sequenzen mit weniger als 400 Nukleotiden wird mit Hilfe von RNAshapes eine Kollektion von Strukturen um einen vom Benutzer definierten Energiebetrag um die *Minimum Free Energy* (*mfe*) erstellt und für die folgende Klassifizierung verwendet. Dagegen werden von Sequenzen mit einer Länge größer als 400 Nukleotide und kleiner als 1.000 Nukleotide durch RNALfold direkt die thermodynamisch optimalen Strukturen mit einer Basenpaar-Spannweite von maximal 400 Nukleotiden erstellt, aus der Gesamtsequenz extrahiert und für die weitere Klassifizierung vorgesehen. Sequenzen mit einer Länge von mehr als 1.000 Nukleotiden werden in um 500 Nukleotide überlappende 1.000 Nukleotide umfassende Sequenzstücke geschnitten, um eine vollständige Abdeckung möglicher vorliegender microRNA-Sequenzen erreichen zu können. Von diesen Sequenzstücken werden ebenfalls wieder mit Hilfe von RNALfold die thermodynamisch optimalen Strukturen berechnet und deren Sequenz aus der Gesamtsequenz extrahiert. Dieser Schritt dient einzig der Extraktion lokal stabiler Strukturelemente in langen Sequenzen. Die Sekundärstrukturen der extrahierten Sequenzen werden dabei wieder verworfen.

Noch vor der thermodynamischen Faltung der zu klassifizierenden Sequenzen durch RNAshapes findet ein erster Filter-Prozess statt, um möglichst viele falsch-positive Sequenzen aus der Grundmenge zu eliminieren. Dieser Filter untersucht die gegebenen Sequenzen auf ihre Nukleotid-Zusammensetzung (Abschnitt 3.2.2) und verwirft alle Sequenzen, die diese Filter-Bedingungen nicht erfüllen. Eine Übersicht über den Ablauf von YamP ist in Abbildung 3.2 gezeigt.

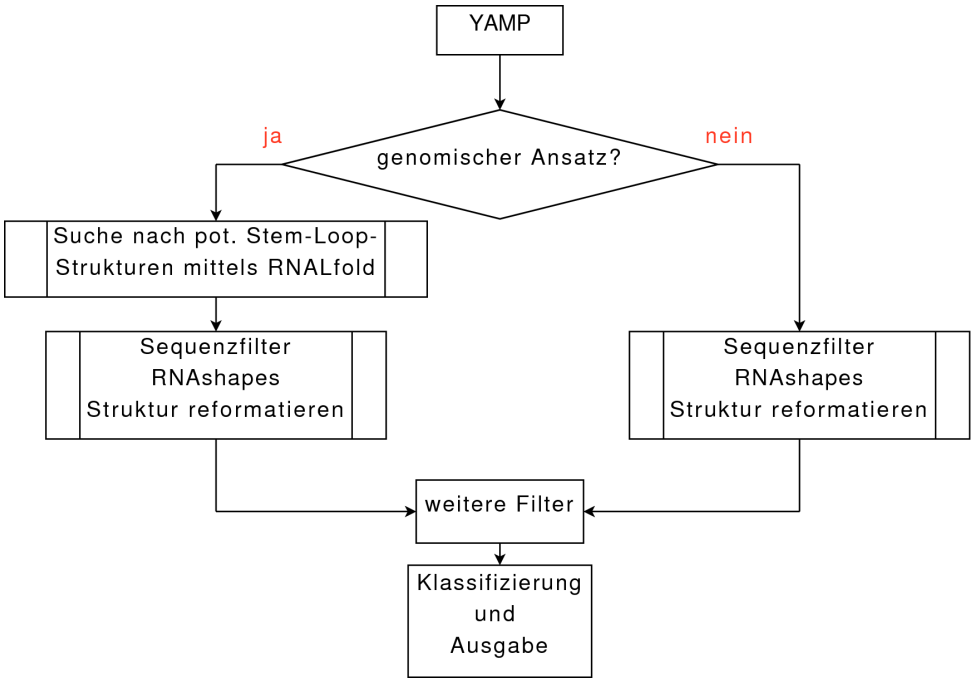

Abbildung 3.2: Generelle Vorgehensweise von YamP. Zunächst wird die Entscheidung getroffen, ob es sich um einen genomischen Ansatz oder eine Einzel-Sequenz-Klassifizierung handelt. Davon hängt die weitere Behandlung der Eingabe-Sequenzen ab. Einzel-Sequenzen werden direkt mit RNAshapes gefaltet während längere Sequenzen aus dem genomischen Ansatz zunächst durch RNALfold auf mögliche Hairpin-Strukturen untersucht werden. Nach Extraktion der Hairpin-Strukturen werden beide Ansätze wieder vereinigt, die Filter auf die Sequenzen und Strukturen angewendet und zuletzt klassifiziert.

3.2.2 Filter-Methoden

Um schon vor der eigentlichen Klassifizierung den Suchraum bzw. die Anzahl falsch-positiver Sequenzen zu verringern, wurden in YamP einige Filter-Methoden implementiert. Diese Filter-Methoden umfassen zum einen die Sequenzkomposition und zum anderen strukturbasierte Filter-Methoden. Da aus der Literatur bekannt ist (Bonnet *et al.*, 2006), dass die Sequenzinformation nur einen geringen Einfluss auf die Funktionalität einer microRNA-Sequenz hat, basiert der größte Teil der Filter-Methoden auf strukturellen intrinsischen Eigenschaften.

Nukleotid-Filter

Die Eingabesequenzen werden noch vor der thermodynamischen Faltung durch RNAshapes bezüglich ihrer Nukleotid-Sequenz untersucht. Sequenzen, in denen wenigstens ein Nukleotid deutlich unterrepräsentiert ist, werden verworfen. Eine Analyse der bekannten microRNA-Sequenzen, erste Testläufe über das Genom von *Arabidopsis thaliana* und Literaturangaben (Zhang *et al.*, 2006) ergaben, dass der Anteil eines jeden Nukleotids mindestens 10% der Gesamtsequenz ausmachen muss. Auf eine statistische Auswertung bezüglich der Sensitivität und Spezifität dieser Filter-Methode wurde an dieser Stelle

verzichtet, da dieser Filter hauptsächlich dazu dient, repetitive Sequenzen, in denen einzelne Nukleotide häufig deutlich unterrepräsentiert sein können, herauszufiltern.

Energie-Filter

Der Energie-Filter folgt im Anschluss an die Faltung der Sequenzen durch RNAshapes und untersucht die Sekundärstrukturen bezüglich ihrer errechneten thermodynamischen Stabilität. Der verwendete Energie-Filter basiert auf den Untersuchungen von Zhang *et al.* (2006), die Indizien darlegten, dass microRNA-Sekundärstrukturen sich bezüglich ihrer *mfe* signifikant von anderen RNAs unterscheiden. Der Ansatz aus den geschilderten Untersuchungen wurde adaptiert und in einer modifizierten Form in das Programm integriert. Die Modifikation bezog sich dabei auf die Normierung der *mfe*. Während Zhang *et al.* die *mfe* auf den GC-Gehalt und die Länge normierten, fand in der Implementierung des Energie-Filters in YamP die Normierung über die Länge statt. Die durch RNAshapes berechnete *mfe* in kcal/mol wird durch die Länge der Sequenz dividiert (3.1). Besitzt die betrachtete Struktur eine normalisierte Energie, die größer als ein definierter Schwellenwert (*Threshold*, $T_{Energie-Filter}$) ist, so wird die Struktur verworfen.

$$\frac{\Delta G}{\text{length}(Hairpin)} \leq T_{Energie-Filter} \tag{3.1}$$

Hairpin-Filter

Der Hairpin-Filter schließt sich der Umformatierung der Punkt-Klammer-Notation an. Es werden alle Strukturen bzw. Teilstrukturen, die nicht mindestens eine definierte Länge aufweisen, verworfen. Dieser Filter begründet sich in der Tatsache, dass die reife microRNA über ihre gesamte Länge in einem Hairpin vorliegen muss und sich nicht über Verzweigungs- oder Hairpin-Loops erstreckt (Zhang *et al.*, 2006). Über die Kenntnis der Mindestlänge einer microRNA und ihrer Lokalisation in der Hairpin-Struktur folgt die Bedingung, dass microRNA-Vorläufer für ihre Funktionalität eine Mindestlänge besitzen müssen.

Konsekutive Basenpaare

Biologisch relevante microRNA-Vorläufer unterscheiden sich, wie bereits erwähnt, statistisch signifikant von anderen konservierten Hairpins bezüglich ihrer *mfe* und Gesamtgröße. Weiterhin tendieren diese statistisch signifikant zu weniger Loops; wenn diese auftreten, so meist deutlich verkleinert (Ritchie *et al.*, 2007). Größere Loops würden der dreidimensionalen Struktur eines Hairpins eine irreguläre bzw. asymmetrische Form verleihen und somit die Funktionalität beeinträchtigen bzw. die Prozessierung durch *Dicer-like protein* (DCL) behindern.

In diesem Filter-Ansatz wird die maximale Anzahl konsekutiver Basenpaare in allen umformatierten Strukturen ermittelt. Da interne Loops oder Bulge-Loops seltener auftre-

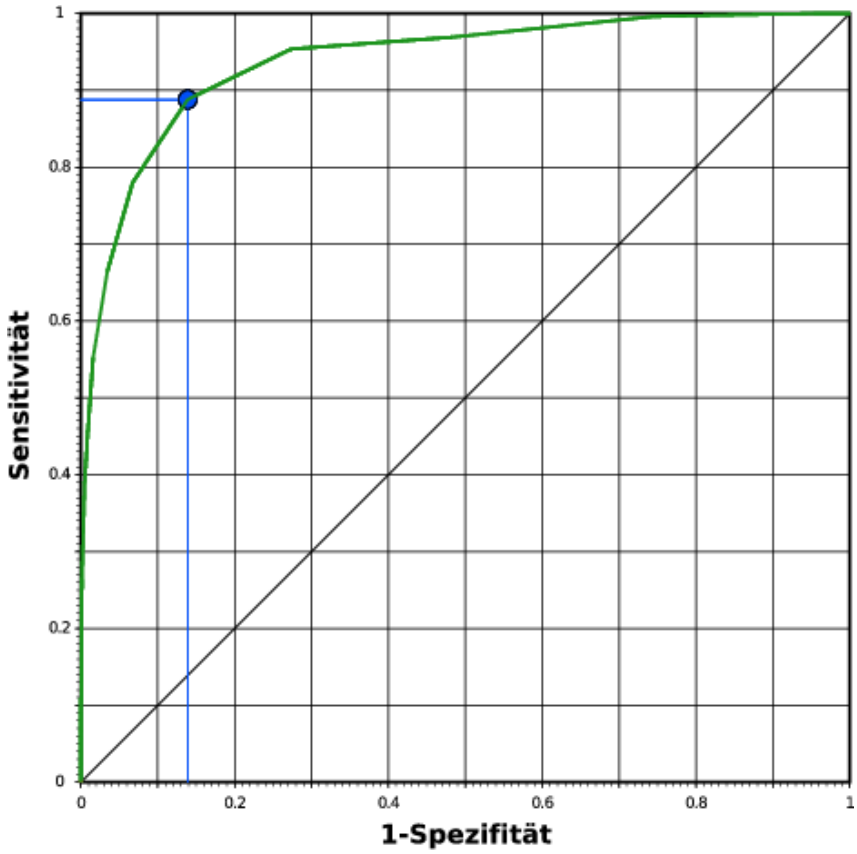

Abbildung 3.3: Effizienz des Filters für die Anzahl konsekutiver Basenpaare. Das Diagramm zeigt eine *Receiver-Operator-Characteristics*-Kurve (ROC-Kurve) Die Y-Achse beschreibt die Sensitivität (1.1), die X-Achse die Falsch-positiv-Rate (1.3). Die Auftragung stellt die Abhängigkeit der Sensitivität und der Falsch-positiv-Rate eines Klassifikators in Abhängigkeit des zugehörigen Schwellenwertes dar (grüne Kurve). Der blaue Punkt beschreibt das Wertepaar aus Sensitivität und Falsch-positiv-Rate, der den zugrundeliegenden Datensatz basierend auf dem untersuchten Klassifikator optimal diskriminiert.

---

ten, ist die Gesamtzahl an aufeinanderfolgenden Basenpaaren gegenüber anderen nichtkodierenden RNAs deutlich erhöht. In Abbildung 3.3 ist die Effizienz dieses Filters über eine sogenannte *Receiver-Operator-Characteristics*-Kurve (ROC-Kurve) bei einem Konfidenzintervall von 95% abgebildet. Die Abbildung 3.3 zeigt, dass der untersuchte Filter die zugrundeliegenden Sequenzen sehr gut diskriminieren kann. Dies ist an dem Verlauf der Kurve und der damit verbundenen großen Fläche unterhalb der Kurve (*Area under the Curve*, AUC) erkennbar, welche im vorliegenden Fall einen Wert von ca. $0,91$ betrug. In die Bewertung flossen zum einen alle derzeit bekannten 184 microRNA-Sequenzen sowie über 8.494 aus *cDNA*s extrahierte Hairpin-Strukturen und die komplette Rfam-Datenbank ein.

Sequenz-Hairpin-Korrelation

Dieser Filter zielt auf das Verhältnis zwischen der Gesamt-Sequenzlänge und der Anzahl Hairpins einer Struktur ab. Sequenzen mit nur wenig verzweigten Strukturen, bei denen

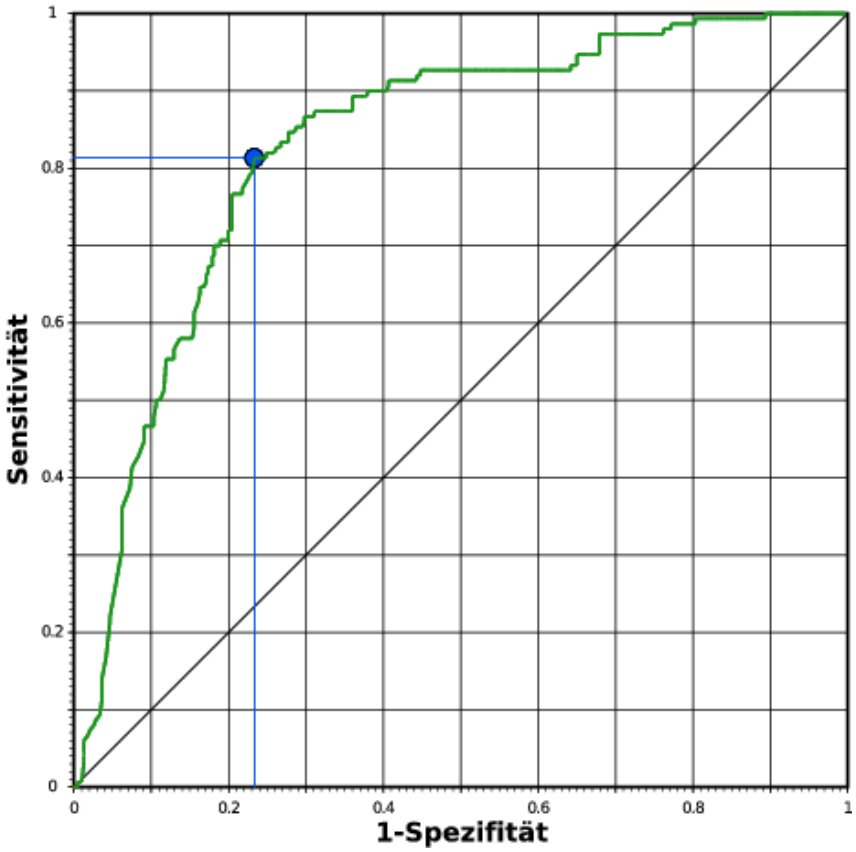

Abbildung 3.4: Effizienz des Sequenz-Hairpin-Korrelations-Filters. Das Diagramm zeigt eine *Receiver-Operator-Characteristics*-Kurve (ROC-Kurve). Die Y-Achse beschreibt die Sensitivität (1.1) und die X-Achse die Falsch-positiv-Rate (1.3). Die Auftragung stellt die Abhängigkeit der Sensitivität und der Falsch-positiv-Rate eines Klassifikators in Abhängigkeit des zugehörigen Schwellenwertes dar (grüne Kurve). Der blaue Punkt beschreibt das Wertepaar aus Sensitivität und Falsch-positiv-Rate, der den zugrundeliegenden Datensatz basierend auf dem untersuchten Klassifikator optimal diskriminiert. Die Erhebung der Daten für diesen Filter basiert auf den microRNA-Sequenzen aus der miRBase in Version 10.0 und nicht-microRNA-Sequenzen aus der Rfam-Datenbank.

---

nahezu die gesamte Sequenz in einem Hairpin vorliegt, erzielen hier die besten Werte. Der günstigste Wert, den eine Sequenz dabei einnehmen kann, konvergiert gegen 2 mit zunehmender Sequenzlänge. Dies gilt für den Fall, dass bis auf die terminalen Loop-Nukleotide alle Nukleotide in einem einzigen Hairpin in der Gesamtstruktur involviert sind. Liegt der errechnete Wert über einem definierten Schwellenwert ($T_{Sequenz-Hairpin-Korrelation}$) (3.2), so wird die betrachtete Struktur verworfen.

Die Bewertung der Effizienz dieses Filters ist in Abbildung 3.4 dargestellt. Dieser ROC-Analyse lagen nur die Rfam-Sequenzen zugrunde, da hier die Effizienz des Sequenz-Hairpin-Korrelations-Filters am deutlichsten hervorgehoben werden konnte.

$$\frac{\text{length}(Sequenz)}{\text{length}(Hairpin)} \geq T_{Sequenz-Hairpin-Korrelation} \tag{3.2}$$

Die Abschätzung der AUC ergab dabei eine Fläche von ca. $0,69$.

***Window-Slide*-Filter**

Die letzte Filter-Methode zielt ebenfalls wieder auf die Anzahl der Basenpaare bei gleichzeitiger Minimierung von Loop-Anzahl und -Größe ab. Bei dieser Methode werden in einem Fenster von jeweils 20 Nukleotiden die Anzahl der Basenpaare in der umformatierten Struktur ermittelt und über die Fenstergröße normiert. Die m größten Werte jeder Struktur werden ebenfalls summiert und der Mittelwert gebildet (siehe Formel 3.3). Dabei wurde der Wert für m wurde so gewählt, dass die entsprechenden Werte für die richtig-positiven microRNA-Sequenzen maximiert und die Werte für die richtig-negativen Sequenzen minimiert wurden. Liegt dieser Wert unter einem definierten Schwellenwert ($T_{Window-Slide-Filter}$), so wird diese Struktur verworfen.

Die Bewertung der Effizienz über eine ROC-Analyse ist in Abbildung 3.5 dargestellt. In die Berechnungen flossen hierzu die Sequenzen aus der RFAM-Datenbank, 10.000 aus cDNAs extrahierte Hairpin-Strukturen, sowie die derzeit bekannten 185 microRNA-Vorläufer-Sequenzen, ein. Dabei zeigt sich, dass der *Window-Slide*-Filter die zugrundeliegenden Sequenzen mit sehr guter Effizienz diskriminieren konnte, was durch den Verlauf der Kurve in Abbildung 3.5 deutlich wird und durch die Berechnung der Fläche unter der Kurve (*Area under the Curve*, AUC) bestätigt werden konnte. Die Abschätzung der AUC betrug für den *Window-Slide*-Filter ca. $0,97$.

$$\frac{\sum_{i=1}^{\text{length(seq)}-\text{window size}} \text{sort desc } \frac{\text{number of basepairs}}{\text{window size}}}{m} \geq T_{Window-Slide-Filter} \tag{3.3}$$

3.2.3 Klassifizierungsmethode

Die Klassifizierung der Eingabe-Sequenzen in mögliche microRNA- oder nicht-microRNA Sequenzen beinhaltet zwei von einander abhängige Prozesse. Zum einen ist die generelle Klassifizierung der Eingabe-Sequenzen zu nennen und zum anderen, falls eine positive Klassifizierung stattgefunden hat, die Lokalisation der reifen microRNA-Sequenz in der Hairpin-Struktur. Um diese beiden voneinander abhängigen Prozesse durchführen zu können, wurde zunächst ein Gitter mit vier verschiedenen Zuständen definiert. Eine grafische Veranschaulichung eines derartigen Gitters ist in Abbildung 3.6 dargestellt. Dieses Gitter umfasst vier Zustände:

- *ist-miRNA*
- *ist-miRNA* $\rightarrow$ *ist-nicht-miRNA*
- *ist-nicht-miRNA* $\rightarrow$ *ist-miRNA*
- *ist-nicht-miRNA*

Das Gitter hat somit die Dimensionen $n \times 4$ mit n als Länge der umformatierten Struktur. Eine Einschränkung in dem definierten Gitter stellen die möglichen Transitionen dar. So

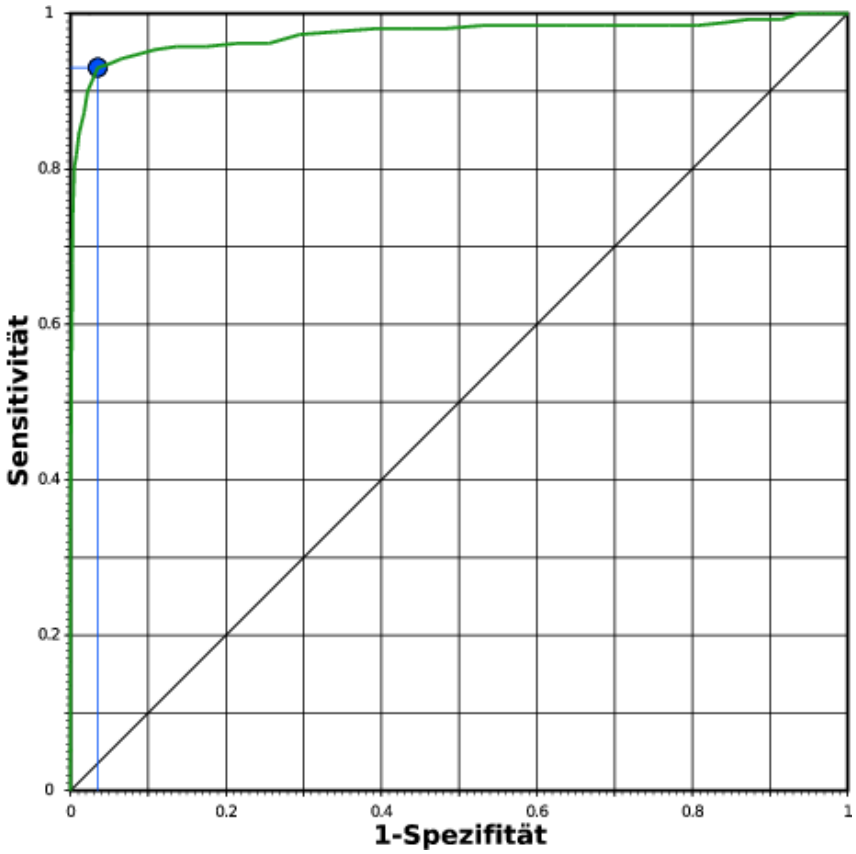

Abbildung 3.5: Effizienz des *Window-Slide*-Filters. Das Diagramm zeigt eine *Receiver-Operator-Characteristics*-Kurve (ROC-Kurve). Die Y-Achse beschreibt die Sensitivität (1.1) und die X-Achse die Falsch-positiv-Rate (1.3). Die Auftragung stellt die Abhängigkeit der Sensitivität und der Falsch-positiv-Rate eines Klassifikators in Abhängigkeit des zugehörigen Schwellenwertes dar (grüne Kurve). Der blaue Punkt beschreibt das Wertepaar aus Sensitivität und Falsch-positiv-Rate, der den zugrundeliegenden Datensatz basierend auf dem untersuchten Klassifikator optimal diskriminierte.

---

ist nicht jeder Zustand von jedem beliebigen Vorgängerzustand aus zu erreichen, vielmehr kann jeder Zustand nur von zwei bestimmten Zuständen erreicht werden. Von jedem Zustand aus gibt es demnach nur genau zwei Möglichkeiten, den Zustand j an Position i zu verlassen. Die Tabelle 3.2 stellt eine Übersicht der möglichen Übergänge zwischen den einzelnen Zuständen dar, die in Abbildung 3.6 grafisch veranschaulicht sind.

Im Folgenden bezeichnet n die Länge der umformatierten Struktur und $m = 4$ die Anzahl der möglichen Zustände im Gitter. $M_{n,m}$ bezeichnet die der Klassifizierung zugrundeliegenden Matrix oder Gitters. Die Initialisierung der Matrix erfolgt dabei wie folgt (Formel 3.4):

$$M_{0,m} = 0 \tag{3.4}$$

Tabelle 3.2: Mögliche Zustandsübergänge. Die Tabelle stellt in der linken Spalte alle möglichen Zustände des in Abschnitt 3.2.3 vorgestellten Gitters dar. Die rechte Spalte listet die aktuellen Zustände auf und die Transitionen zu den möglichen Nachfolgezuständen. Die vier Zustände sind: *ist-miRNA*; *ist-miRNA → ist-nicht-miRNA*; *ist-nicht-miRNA → ist-miRNA*; *ist-nicht-miRNA*. Ausgehend von jedem Zustand können nur genau zwei Nachfolgezustände erreicht werden.

aktueller Zustand	mögliche Nachfolge-Zustände
ist-miRNA	*ist-miRNA* **und** *ist-miRNA → ist-nicht-miRNA*
ist-miRNA → ist-nicht-miRNA	*ist-nicht-miRNA → ist-miRNA* **und** *ist-nicht-miRNA*
ist-nicht-miRNA → ist-miRNA	*ist-miRNA → ist-nicht-miRNA* **und** *ist-miRNA*
ist-nicht-miRNA	*ist-nicht-miRNA* **und** *ist-nicht-miRNA → ist-miRNA*

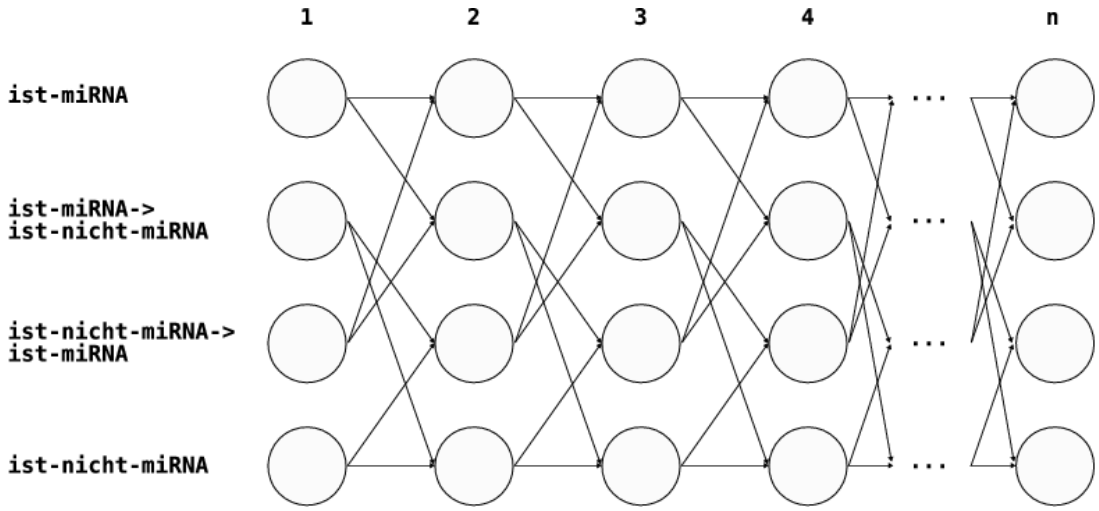

Abbildung 3.6: Aufbau der Matrix. Die Darstellung zeigt die implementierte Matrix. Diese besitzt die Dimensionen $n \times 4$ mit n für die Länge der umformatierten Struktur und die 4 Zustände, in denen sich die jeweilig betrachtete Position befinden kann. Die vier möglichen Zustände sind in der Abbildung links aufgeführt, welche an jeder Position $1 \dots n$ durch einen Kreis symbolisiert sind. Die möglichen Transitionen sind beschränkt, was bedeutet, dass nicht jeder Zustand in jeden beliebigen anderen Zustand überführt werden kann und durch die Pfeile verdeutlicht wird.

Die Felder der Matrix mit der Dimension $n \times 4$ werden wie folgt initialisiert:

$$M_{i,0} = \log(state - is) + \log(p(di - nuk - is)) \tag{3.5}$$
$$M_{i,1} = \log(p(di - nuk - is)) + \log(p(di - nuk - plus)) \tag{3.6}$$
$$M_{i,2} = \log(p(di - nuk - not)) + \log(p(di - nuk - min)) \tag{3.7}$$
$$M_{i,3} = \log(state - not) + \log(p(di - nuk - not)) \tag{3.8}$$

Die Ausdrücke $p(di-nuk-is)$, $p(di-nuk-not)$, $p(di-nuk-plus)$ und $p(di-nuk-min)$ bedeuten, dass die Frequenzen für das aktuell betrachtete, benachbarte Nukleotidpaar aus der Vorwärts- oder Rückwärts-Sequenz der umformatierten Struktur gewählt wurde, welches den positiveren Einfluss auf eine positive Klassifizierung besitzt.

In einem dynamischen Programmieransatz werden nun die Felder der Matrix M wie folgt berechnet:

$$M_{i,0} = M_{i,0} + \max(M_{i,0}, M_{i,2}) \tag{3.9}$$
$$M_{i,1} = M_{i,1} + \max(M_{i,0}, M_{i,2}) \tag{3.10}$$
$$M_{i,2} = M_{i,2} + \max(M_{i,1}, M_{i,3}) \tag{3.11}$$
$$M_{i,3} = M_{i,3} + \max(M_{i,1}, M_{i,3}) \tag{3.12}$$

Für die weitere Klassifizierung in die positive oder negative Klasse wird die vorliegende Matrix in eine zweite Matrix N überführt, welche in ihren Dimensionen reduziert ist und nur noch die Zustände *ist-miRNA* und *ist-nicht-miRNA* repräsentiert. Dabei werden die Positionen $N_{i,j}$ mit $0 \leq i \leq n$ und $0 \leq j \leq 1$ wie folgt berechnet.

Zunächst wird der Mittelwert der gesamten betrachteten Spalte i ermittelt:

$$\bar{M_i} = \frac{(M_{i,0} + M_{i,1} + M_{i,2} + M_{i,3})}{4} \tag{3.13}$$

Die Zustände *ist-miRNA* und *ist-nicht-miRNA* in der Matrix N werden wie folgt errechnet:

$$N_{i,0} = -1 \cdot (\bar{M}_{\text{Spalte}\,i} - M_{i,0}) \tag{3.14}$$
$$N_{i,1} = -1 \cdot (\bar{M}_{\text{Spalte}\,i} - M_{i,3}) \tag{3.15}$$

Dabei bezeichnet $j = 0$ den Zustand *ist-nicht-miRNA* und $j = 1$ der neuen Matrix N den Zustand *ist-miRNA*.

Die eigentliche Entscheidung über das Vorliegen eines microRNA-Vorläufers ist ein zweistufiger Entscheidungsprozess. Im ersten Entscheidungsprozess wird die Summe der beiden Zustandsspalten gebildet:

$$\text{ist-miRNA} = \sum_{i=1}^{n} N_{i,0} \tag{3.16}$$
$$\text{ist-nicht-miRNA} = \sum_{i=1}^{n} N_{i,1} \tag{3.17}$$

Über diese beiden Spaltensummen erfolgt der erste Schritt der Klassifizierung. Der Quotient aus (3.18) wird dabei für jeden möglichen Kandidaten für die abschließende Klassifizierung vermerkt.

$$\text{Pred1} = \left(\frac{\text{ist-nicht-miRNA}}{\text{ist-miRNA}}\right)^2 \tag{3.18}$$

Der zweite Entscheidungsprozess durchläuft die umformatierte Struktur ähnlich dem *Window-Slide*-Filter in einer Fenstergröße von 20 Positionen. Es wird für jede Position

in diesem Fenster die Differenz zwischen den Zuständen *ist-miRNA* und *ist-nicht-miRNA* gebildet und summiert (3.19) und in einer Liste gespeichert.

$$max_{i,i+20} = \sum_{i=1}^{n-20} \sum_{k=i}^{i+20} N_{k,1} - N_{k,0} \tag{3.19}$$

Dabei beschreibt die erste Summe das Durchlaufen der kompletten umformatierten Struktur und die zweite Summe das Durchlaufen der Fenstergröße beginnend bei jedem einzelnen Nukleotid. Es werden somit pro umformatierter Struktur der Länge n insgesamt $n-20$ Fenster der umformatierten Struktur untersucht. Dieser Wert wird inklusive der Indices seines Auftretens vermerkt und absteigend sortiert. Die neun oder weniger höchsten Werte werden ebenfalls aufsummiert wobei der Mittelwert *Pred2* den zweiten Wert für die Klassifizierung darstellt. Die einzelnen aufsummierten Fensterwerte werden ebenfalls vermerkt und dienen der anschließenden Lokalisation der möglichen reifen microRNA-Sequenz.

In der abschließenden Klassifizierung wird zunächst der Wert *Pred1* (3.18) mit einem empirisch ermittelten Schwellenwert verglichen. Der zweite abschließende Klassifizierungsschritt involviert und multipliziert beide Klassifizierungswerte aus den Formeln 3.18 und 3.19:

$$\text{Pred3} = \text{Pred1} \cdot \text{Pred2} \tag{3.20}$$

Dabei wird auch hier der errechnete Wert *Pred3* mit einem empirisch ermittelten Schwellenwert verglichen. Erfüllen beide Klassifizierungswerte *Pred1* und *Pred3* die Schwellenwert-Bedingungen, so wird die vorliegende Sequenz als microRNA-Vorläufer klassifiziert.

Der in (3.20) definierte Klassifizierungswert wurde ebenfalls mit der ROC-Analyse auf seine Effizienz untersucht. Dabei wurden die microRNA-Sequenzen aus der miRBase 10.0 zusammen mit 455 richtig-negativen Sequenzen aus der Rfam-Datenbank und 8.000 pseudo-Hairpins aus *Arabidopsis thaliana* (siehe Abschnitt 3.3.1) klassifiziert und die Ergebnisse bezüglich ihrer Sensitivität und Spezifität bei einem Konfidenzintervall von 95% aufgetragen. Die Effizienz des Klassifikators ist in Abbildung 3.7 zu sehen. Die Berechnung der AUC für den finalen Klassifikator ergab einen Wert von ca. $0,95$ und zeigt somit eine sehr gute Klassifizierungseffizienz bezogen auf die Eingabesequenzen.

Das Durchlaufen der umformatierten Struktur und die anschließende Suche nach dem Maximum der Werte für die einzelnen Fenstergrößen (siehe Formel 3.19) wurde neben der Ermittlung des finalen Klassifizierungswertes zusätzlich zur Lokalisation der reifen microRNA-Sequenz verwendet. Von der in Formel 3.19 gespeicherten Liste wurden die sechs höchsten Werte grafisch an der umformatierten Struktur angelegt und zeigten, wie in Abbildung 3.8 dargestellt, den Bereich der reifen microRNA-Sequenz innerhalb der umformatierten Struktur.

In einer Testumgebung mit dem richtig-negativen Trainingsdatensatz 1 (Abschnitt 3.3.1) und der miRBase in beiden Versionen wurde die Lokalisation der reifen microRNA-Sequenzen in den umformatierten Strukturen durch das Programm YamP getestet. Die

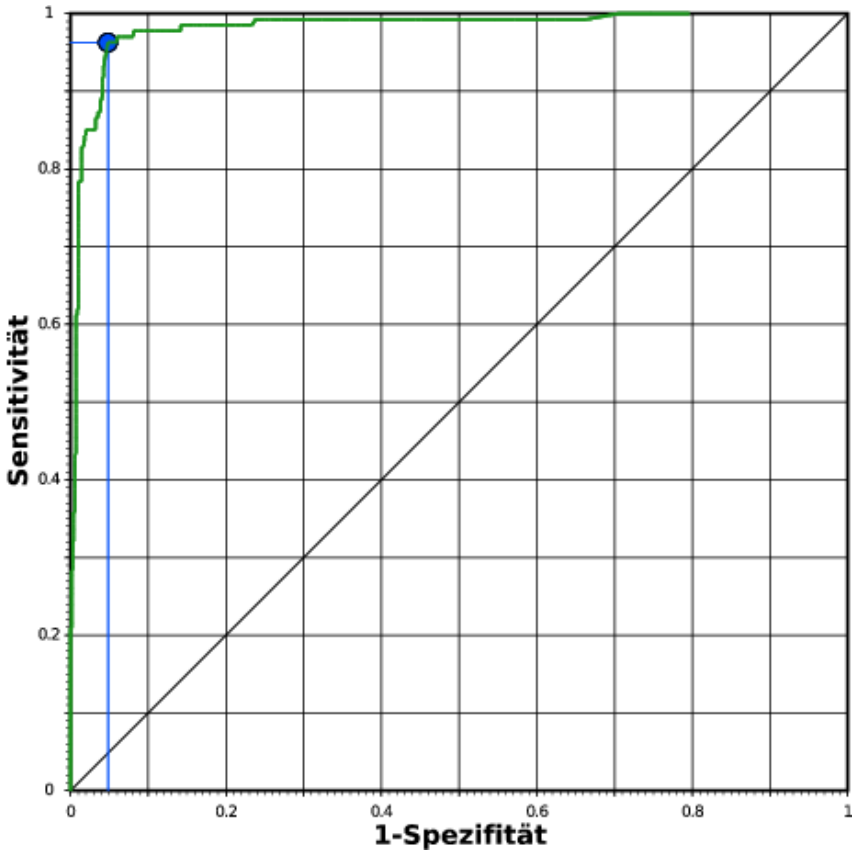

Abbildung 3.7: Darstellung der Effizienz des finalen Klassifizierungsschritts von YamP. Das Diagramm zeigt eine *Receiver-Operator-Characteristics*-Kurve (ROC-Kurve). Die Y-Achse beschreibt die Sensitivität (1.1) und die X-Achse die Falsch-positiv-Rate (1.3). Die Auftragung stellt die Abhängigkeit der Sensitivität und der Falsch-positiv-Rate eines Klassifikators in Abhängigkeit des zugehörigen Schwellenwertes dar (grüne Kurve). Der blaue Punkt beschreibt das Wertepaar aus Sensitivität und Falsch-positiv-Rate, der den zugrundeliegenden Datensatz basierend auf dem untersuchten Klassifikator optimal diskriminiert. Die Erhebung der Daten basierte auf den microRNA-Sequenzen der miRBase in Version 10.0, den nicht-microRNA-Sequenzen der Rfam-Datenbank sowie 8.000 pseudo-Hairpins aus *Arabidopsis thaliana*. Als Klassifikator wurde die Formel 3.20 gewählt und in dem Diagramm aufgetragen.

Ausgangsbedingungen für diese Testumgebung waren neben den bereits erwähnten Trainingsdatensätzen die voreingestellten Werte für die Filter- und den Klassifizierungsschritt (siehe Tabelle 3.4). Weiterhin galt eine microRNA als lokalisiert, wenn mindestens 50% der reifen microRNA-Sequenz (entnommen aus der miRBase) mit der vorhergesagten Region übereinstimmten. Lag eine Übereinstimmung von mindestens 25% und maximal 50% vor, so wurde die reife microRNA-Sequenz als partiell-identifiziert deklariert. Die Tabelle 3.3 zeigt, dass ca. 63% richtig identifiziert wurden und mindestens weitere ca. 8% noch partiell-identifiziert werden konnten. Aus der Literatur ist zu entnehmen, dass nicht alle in der miRBase annotierten microRNAs allgemein als tatsächliche microRNAs angesehen werden (Xie *et al.*, 2005). Daher ist davon auszugehen, dass die Werte in der Tabelle nach oben zu korrigieren sind. Der hier vorgestellte Ansatz zur Identifikation reifer microRNA-Sequenzen in ihren microRNA-Vorläufern wird bei der Suche nach neuen möglichen Kandidaten keine Rolle spielen (siehe Kapitel 4), da für *Arabidopsis thaliana*

Abbildung 3.8: Lokalisation der reifen microRNA-Sequenz im *ath-mir156a*-Vorläufer. Die Darstellung zeigt die umformatierte Struktur des *ath-mir156*-Vorläufers. In rot ist die annotierte reife microRNA-Sequenz hervorgehoben. Zur Lokalisation der reifen microRNA-Sequenz wurde die Formel 3.19 verwendet. Unterhalb der umformatierten Struktur ist an den Positionen der Maxima aus der Formel 3.19 für die betrachtete Fenstergröße eine Sequenz von `x` angelegt worden, welche als Indikator für die reife microRNA-Sequenz dienten.

Tabelle 3.3: Lokalisation der reifen microRNA-Sequenzen. Die Tabelle zeigt den Anteil richtig bzw. partiell-identifizierter reifer microRNA-Sequenzen bezogen auf die Gesamtheit der bekannten microRNAs. Die Werte sind mit Vorsicht zu genießen, da nicht alle in der miRBase annotierten microRNAs allgemein als microRNAs angesehen werden (Bsp: *ath-mir405abd* (Xie *et al.*, 2005), die ihren Ursprung in Transposons haben). Daher sind diese Werte eher nach oben zu korrigieren.

Identifizierte reife microRNAs	63,04%
inklusive partiell-identifizierter reifer microRNAs	71,37%

---

ausreichend experimentelle Daten zur Verfügung (Expressionsdatenbanken MPSS und ASRP), um die möglichen reifen microRNA-Sequenzen in ihren Vorläufer-Strukturen zu identifizieren. Dieser Ansatz ist im Rahmen dieser Arbeit nur zur Lokalisation möglicher kleiner Viroid-spezifischer RNAs in Abschnitt 4.4 eingesetzt worden und diente nicht der genauen Lokalisation reifer microRNA-Sequenzen.

3.3 Validierung

Um das in Abschnitt 3.2 dargestellte und im Rahmen dieser Arbeit entwickelte Programm YamP zur Identifikation von möglichen microRNA-Sequenzen auf biologische Sequenzen anwenden zu können, wurde das Programm zusätzlich in Bezug auf seine Diskriminierungs-Effizienz und -Genauigkeit untersucht. Die Untersuchungen umfassten eine Anwendung auf die Trainingsdatensätze, das Bootstrapping-Verfahren sowie eine Anwendung auf weitere, dem Programm unbekannte, microRNA- und nicht-microRNA-Sequenzen.

3.3.1 Anwendung auf Trainingsdaten

Im ersten Validierungsschritt bestand der Trainingsdatensatz aus 117 bekannten und teilweise verifizierten microRNA-Sequenzen sowie unterschiedlicher nicht-microRNA-Trainingsdatensätze. Die Menge der nicht-microRNA-Sequenzen umfasste verschiedene Klassen von RNAs. Dazu gehörten tRNA-Sequenzen, 5S-rRNA-Sequenzen, 5.8S-rRNA-Sequenzen, mRNA-Sequenzen, und pseudo-Hairpins aus cDNA-Sequenzen von *Arabidopsis thaliana*, sowie pseudo-Hairpins aus *Homo sapiens* (Ng & Mishra, 2007) und der Rfam-Datenbank. Pseudo-Hairpins beschreiben, wie schon im vorigen Abschnitt 3.1 erwähnt, aus cDNAs extrahierte Sequenzen mit der Eigenschaft potentiell eine Hairpin-Struktur ausbilden zu können. Die getesteten Sequenz-Zusammenstellungen sind im Folgenden dargestellt:

- Datensatz 1: 631 tRNA-Sequenzen, 63 5.8S rRNA-Sequenzen und 602 5S rRNA-Sequenzen
- Datensatz 2: 631 tRNA-Sequenzen, 63 5.8S rRNA-Sequenzen, 602 5S rRNA-Sequenzen und 710 mRNA-Sequenzen
- Datensatz 3: 710 mRNA-Sequenzen
- Datensatz 4: 710 mRNA-Sequenzen und 454 ncRNA-Familien aus der Rfam-Datenbank mit je einem Vertreter
- Datensatz 5: 454 ncRNA-Familien aus der Rfam-Datenbank mit je einem Vertreter
- 10.000 pseudo-Hairpins aus *Arabidopsis thaliana*
- 8.494 pseudo-Hairpins aus *Homo sapiens*

Die Tabelle 3.4 zeigt die verwendeten, empirisch ermittelten, Schwellenwerte für die in Abschnitt 3.2.2 beschriebenen Filter-Methoden, die beiden Vorhersage-Schwellenwerte *Pred1* (3.18) und *Pred3* (3.20) sowie die minimale Hairpin-Länge, welche ein Hairpin besitzen muss, um in die Klassifizierung von YamP einzugehen.

Die statistische Auswertung bezüglich der Sensitivität, Spezifität, Korrekt-Klassifikationsrate und Falsch-Klassifikationsrate umfasste bei jedem Testlauf die Menge der microRNA-Sequenzen (117 bzw. 184 Sequenzen), die pseudo-Hairpins aus *Arabidopsis thaliana* sowie die Rfam-Sequenzen aus den Trainingsdaten. Durch die Faltung der zu testenden Sequenzen durch RNAshapes um einen definierten Energiebetrag ergaben sich neben der *mfe*-Struktur eine oder mehrere thermodynamisch suboptimale Strukturen, welche ebenfalls in die Klassifizierung eingingen. Wurden zwei einer Sequenz zugehörigen Strukturen widersprüchlich klassifiziert, so wurde die Sequenz als microRNA-Sequenz klassifiziert angesehen.

Die Tabelle 3.5 zeigt die statistische Auswertung der microRNA-Sequenzen sowie der nicht-microRNA-Sequenzen.

Tabelle 3.4: Verwendete Schwellenwerte für die Testdatensätze. Die Tabelle zeigt die verwendeten Schwellenwerte für die Anwendung auf die Trainingsdaten (siehe Abschnitt 3.3.1).

Pred1 (3.18)	0,9
Pred3 (3.20)	4,37
Konsekutive Basenpaare	8
Korrelation Sequenz-Hairpin	6
Window-Slide-Filter	0,65
normalisierte Energie	−0,2
minimale Hairpin-Länge	32
Energiebereich für RNAshapes	0,1

Tabelle 3.5: Auswertungen aus der Klassifizierung der Trainingsdatensätze. Die Tabelle zeigt die Resultate aus der Anwendung auf die Trainingsdaten. Exemplarisch dargestellt sind die Resultate mit den richtig-negativen Sequenzdatensätzen 1 und 2 sowie den microRNA-Sequenzdatensätzen aus der miRBase in Version 7.0 und 10.0.

Datensatz 1 und 117 microRNA-Sequenzen aus der miRBase 7.0	
Sensitivität	87,61%
Spezifität	98,88%
Korrekt-Klassifikationsrate	97,70%
Falsch-Klassifikationsrate	2,29%
Datensatz 2 und 117 microRNA-Sequenzen aus der miRBase 7.0	
Sensitivität	77,08%
Spezifität	99,84%
Korrekt-Klassifikationsrate	99,44%
Falsch-Klassifikationsrate	0,56%
Datensatz 1 und 184 microRNA-Sequenzen aus der miRBase 10.0	
Sensitivität	85,87%
Spezifität	96,38%
Korrekt-Klassifikationsrate	96,15%
Falsch-Klassifikationsrate	3,84%
Datensatz 2 und 184 microRNA-Sequenzen aus der miRBase 10.0	
Sensitivität	79,89%
Spezifität	99,63%
Korrekt-Klassifikationsrate	99,21%
Falsch-Klassifikationsrate	0,78%

Die gezeigten statistischen Auswertungen für die Klassifizierungseffizienz des hier entwickelten Programms YamP zeigen Ergebnisse, welche bezüglich der Sensitivität und Spezifität im Vergleich zu den anderen Ansätzen am besten sind. Die anderen Ergebnisse zeigten teilweise deutlich schlechtere Werte in ihrer Effizienz zur Diskriminierung der Test-Datensätze (Daten nicht gezeigt). Interessanterweise zeigte auch der Trainingsdaten-

satz 7 mit den humanen pseudo-Hairpin-Strukturen eine sehr gute Effizienz bezüglich der Sensitivität und Spezifität, welche vergleichbar mit den gezeigten Ergebnissen ist.

3.3.2 Bootstrap-Ergebnisse

Das sogenannte Bootstrapping-Verfahren beschreibt in der Bioinformatik eine Methode zur Testung der Aussagekraft eines zugrundeliegenden Modells. Am häufigsten findet dieses Verfahren in der Phylogenie und der Baumrekonstruktion Anwendung. Dabei wird durch Umgestalten der Eingangsdaten eine Abschätzung des statistischen Fehlers einer Hypothese ermöglicht. Im Folgenden bezeichnet das Bootstrapping immer das sukzessive Auslassen einer definierten Menge an richtig-positiven microRNA-Sequenzen und das anschließende Training, dass auf der reduzierten Menge richtig-positiver Sequenzen und den in Abschnitt 3.3.1 richtig-negativen Trainingsdaten beruht. Das sukzessive Auslassen von richtig-positiven Sequenzen umfasste dabei eine Menge von jeweils zehn Sequenzen. Diese zehn Sequenzen wurden anhand ihrer Reihenfolge des Auftretens in der Sequenzdatei ausgewählt, wobei diese alphanumerisch sortiert waren.

Das Bootstrapping wurde für alle in Abschnitt 3.3.1 erwähnten richtig-negativen Trainingsdatensätze durchgeführt. Die richtig-positiven Trainingsdatensätze umfassten zum einen die miRBase in Version 7.0 (117 microRNA-Sequenzen) und zum anderen in Version 10.0 (184 microRNA-Sequenzen).

In den Abbildungen 3.9, 3.10, 3.11 und 3.12 sind die Ergebnisse des Bootstrap-Verfahrens dargestellt. Die Ausgangsbedingungen, welche in Tabelle 3.5 die Grundlage für das Training und die Klassifizierung in Abschnitt 3.3.1 waren, sind in den erwähnten Abbildungen dargestellt. Diese wurden so gewählt, dass die Robustheit des Programms YamP gezeigt werden kann.

Die Bootstrap-Ergebnisse zeigen, dass trotz Auslassens von jeweils zehn richtig-positiven microRNA-Sequenzen die Effizienz nur in einem Fall stärkeren Schwankungen unterliegt. Diese Schwankungen betreffen lediglich die Sensitivität, jedoch nicht die Spezifität des Programms YamP. Dabei schwanken die Werte für die Sensitivität um ca. 11% zwischen $\approx 67\%$ und $\approx 78\%$ (siehe Abbildung 3.11). Die drei weiteren Bootstrap-Ergebnisse zeigen hingegen ein robustes Diskriminierungsverhalten bei Auslassen von jeweils zehn richtig-positiven microRNA-Sequenzen.

3.3.3 Weitere Testdatensätze

In den Abschnitten 3.3 und 3.3.2 wurde die prinzipielle Funktionalität des entwickelten Programms bereits gezeigt. Bei diesen Ansätzen erfolgte die Diskriminierung der zu testenden Sequenzen allein auf den Trainingsdaten. Um zeigen zu können, dass das Programm YamP ebenfalls in der Lage ist, unbekannte Sequenzen korrekt klassifizieren zu können, wurden weitere richtig-negative Sequenzdatensätze erstellt und klassifiziert. Als Trainingsdatensatz diente hierbei, sofern nicht anders erwähnt, jeweils der Datensatz 2 als richtig-negativer Datensatz und die miRBase in Version 10.0 (184 microRNA-Sequenzen)

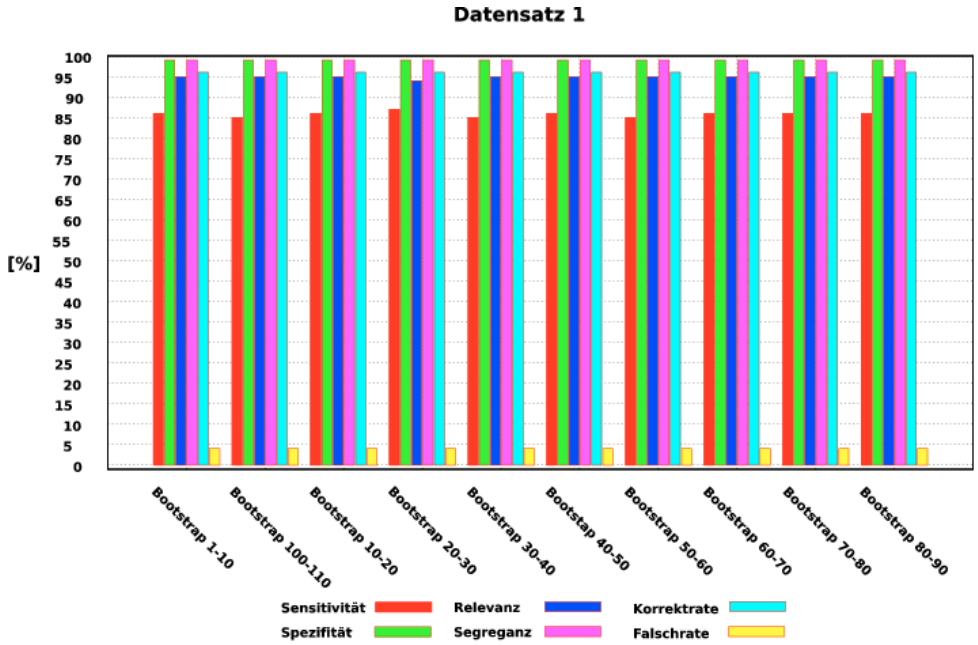

Abbildung 3.9: Bootstrap-Diagramm basierend auf dem Datensatz 1 und der miRBase 7.0. Die Erhebung der statistischen Daten des dargestellten Diagramms fand durch Auslassen von jeweils zehn microRNA-Sequenzen aus dem richtig-positiven Sequenzdatensatz (miRBase 7.0), Training mit den verbleibenden microRNA-Sequenzen und dem richtig-negativen Sequenzdatensatz 1 statt. Die anschließende Klassifizierung fand über die Gesamtmenge der richtig-positiven sowie richtig-negativen Sequenzdaten statt. Die X-Achse stellt die unterschiedlichen richtig-positiven Sequenzdatensätze mit den jeweilig ausgelassenen Sequenzen dar. Die Y-Achse stellt die Sensitivität, Spezifität, Relevanz, Segreganz, Falschrate sowie Korrektrate in % dar.

---

als richtig-positiver Sequenzdatensatz. Desweiteren sollte gezeigt werden, dass YamP zudem in der Lage ist, unbekannte microRNA-Sequenzen korrekt klassifizieren zu können. Zu diesem Zweck wurde als richtig-positiver Trainingsdatensatz die miRBase in Version 7.0 und als Testdatensatz die miRBase in Version 10.0 verwendet.

Als erster neu erstellter richtig-negativer Sequenzdatensatz wurden Sequenzen mit zufälliger Neusortierung der Basenreihenfolge richtig-positiver microRNA-Sequenzen verwendet. Die zufällige Neusortierung der Basenreihenfolge geschah mit dem Programm *shuffle* aus dem BioSQUID-Programmpaket (Eddy, 2008). Es wurden insgesamt drei verschiedene Ansätze zur zufälligen Neusortierung gewählt, welche im Folgenden aufgeführt sind:

- Die Beibehaltung der Mono-Nukleotid-Frequenz
- Die Beibehaltung der Mono-Nukleotid- und Di-Nukleotid-Frequenz
- Die Beibehaltung der Mono-Nukleotid-Frequenz in Fenstern der Größe 20

Diese drei unterschiedlichen Methoden zur Neusortierung wurden für jede einzelne microRNA-Sequenz jeweils fünfmal durchgeführt, was einen nicht-microRNA-Sequenzdatensatz mit insgesamt 2.760 Sequenzen ergab, die als zu klassifizierende

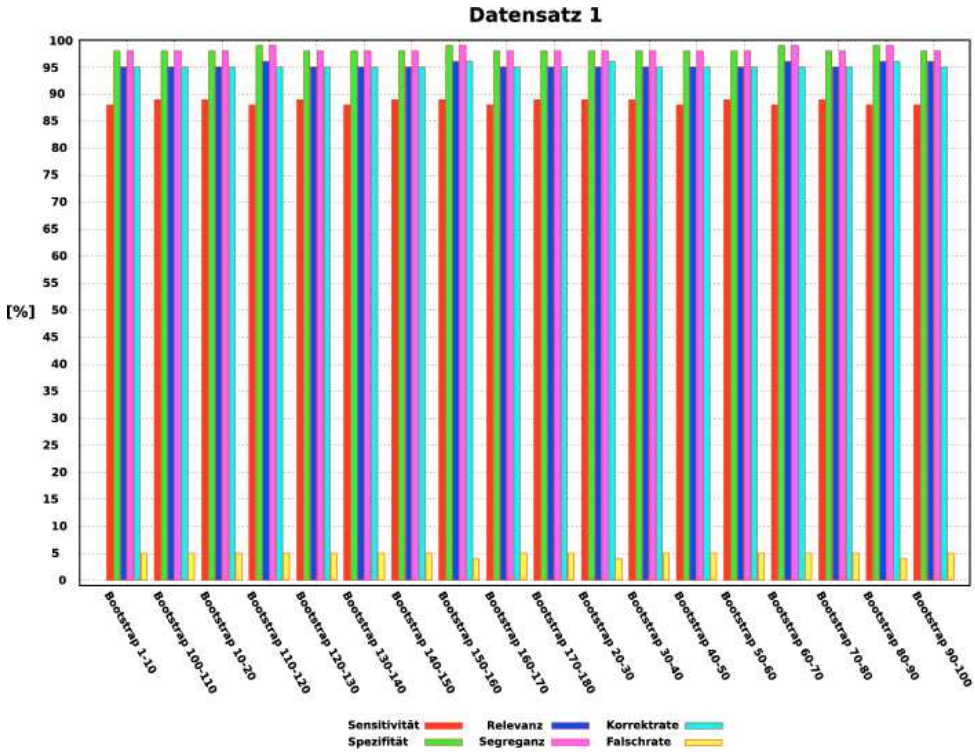

Abbildung 3.10: Bootstrap-Diagramm basierend auf dem Datensatz 1 und der miRBase 10.0. Die Erhebung der statistischen Daten des dargestellten Diagramms fand durch Auslassen von jeweils zehn microRNA-Sequenzen aus dem richtig-positiven Sequenzdatensatz (miRBase 10.0), Training mit den verbleibenden microRNA-Sequenzen und dem richtig-negativen Sequenzdatensatz 1 statt. Die anschließende Klassifizierung fand über die Gesamtmenge der richtig-positiven sowie richtig-negativen Sequenzdaten statt. Die X-Achse stellt die unterschiedlichen richtig-positiven Sequenzdatensätze mit den jeweilig ausgelassenen Sequenzen dar. Die Y-Achse stellt die Sensitivität, Spezifität, Relevanz, Segreganz, Falschrate sowie Korrektrate in % dar.

Tabelle 3.6: Effizienz des Programms YamP basierend auf der zufällig neusortierten Nukleotid-Reihenfolge der richtig-positiven microRNA-Sequenzen.

Spezifität	99,53%
Korrekt-Klassifikationsrate	99,53%
Falsch-Klassifikationsrate	0,47%

Sequenzen verwendet wurden. Die Tabelle 3.6 zeigt die Effizienz von YamP bezüglich der neusortierten, richtig-negativen Sequenzen. Es ist zu deutlich erkennen, dass die Klassifizierung auch in diesem Fall mit hoher Zuverlässigkeit funktioniert.

Als weiterer Testdatensatz diente die RNA-Familie der sogenannten *small nucleolar RNA* (snoRNAs). Diese bilden ebenso wie die microRNA-Sequenzen eine stabile Struktur mit zwei Hairpins aus und sind daher aufgrund auf ihrer Struktur prädestiniert für falsch-positive Klassifizierungen. Als richtig-negative Testmenge wurden 139 snoRNA-

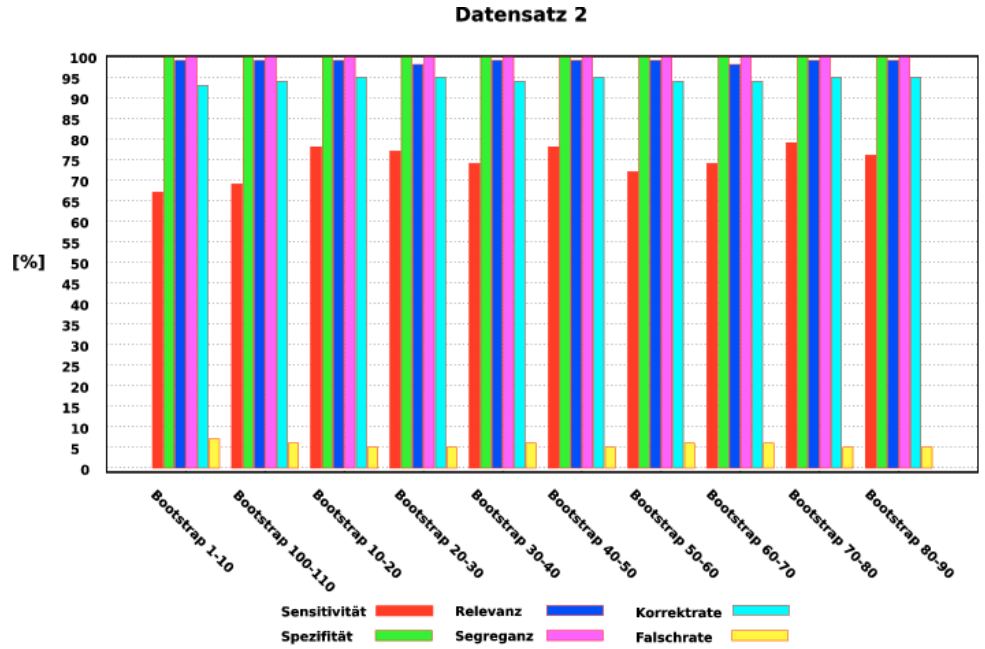

Abbildung 3.11: Bootstrap-Diagramm basierend auf dem Datensatz 2 und der miRBase 7.0. Die Erhebung der statistischen Daten des dargestellten Diagramms fand durch Auslassen von jeweils zehn microRNA-Sequenzen aus dem richtig-positiven Sequenzdatensatz (miRBase 7.0), Training mit den verbleibenden microRNA-Sequenzen und dem richtig-negativen Sequenzdatensatz 2 statt. Die anschließende Klassifizierung fand über die Gesamtmenge der richtig-positiven sowie richtig-negativen Sequenzdaten statt. Die X-Achse stellt die unterschiedlichen richtig-positiven Sequenzdatensätze mit den jeweilig ausgelassenen Sequenzen dar. Die Y-Achse stellt die Sensitivität, Spezifität, Relevanz, Segreganz, Falschrate sowie Korrektrate in % dar.

Sequenzen[1] aus *Arabidopsis thaliana* gewählt. Jede dieser 139 nicht-microRNA-Sequenzen wurde bereits in den Filter-Schritten korrekt als nicht-microRNA-Sequenz klassifiziert und verworfen.

Als nächster richtig-negativer Testdatensatz wurden Sequenzen aus den in der TAIR-Sequenzdatenbank enthaltenen cDNAs extrahiert, welche potentiell in der Lage sind eine Hairpin-Struktur zu bilden. Zu diesem Zweck wurden sämtliche in der TAIR-Sequenzdatenbank enthaltenen cDNAs mit Hilfe von RNALfold und einer maximalen Spannweite erlaubter Basenpaare von 400 Nukleotiden thermodynamisch gefaltet und solche Sequenzen mit einer Mindestlänge von 50 Nukleotiden extrahiert, welche die Eigenschaft zur Ausbildung einer Hairpin-Struktur besitzen. Weitere energetische Beschränkungen lagen bei der Extraktion der Sequenzen nicht vor. Für die anschließende Klassifizierung der richtig-negativen Sequenzen wurden insgesamt zehn Durchläufe vorgenommen, bei denen jeweils 10.000 Sequenzen zufällig aus der Gesamtmenge von 560.193 Sequenzen ausgewählt und klassifiziert wurden. Die Ergebnisse der zehn Klassifizierungsdurchläufe sind in Tabelle 3.7 dargestellt. Auch hier zeigte sich, dass das Programm YamP

[1] `http://lowelab.ucsc.edu/snoRNAdb/`

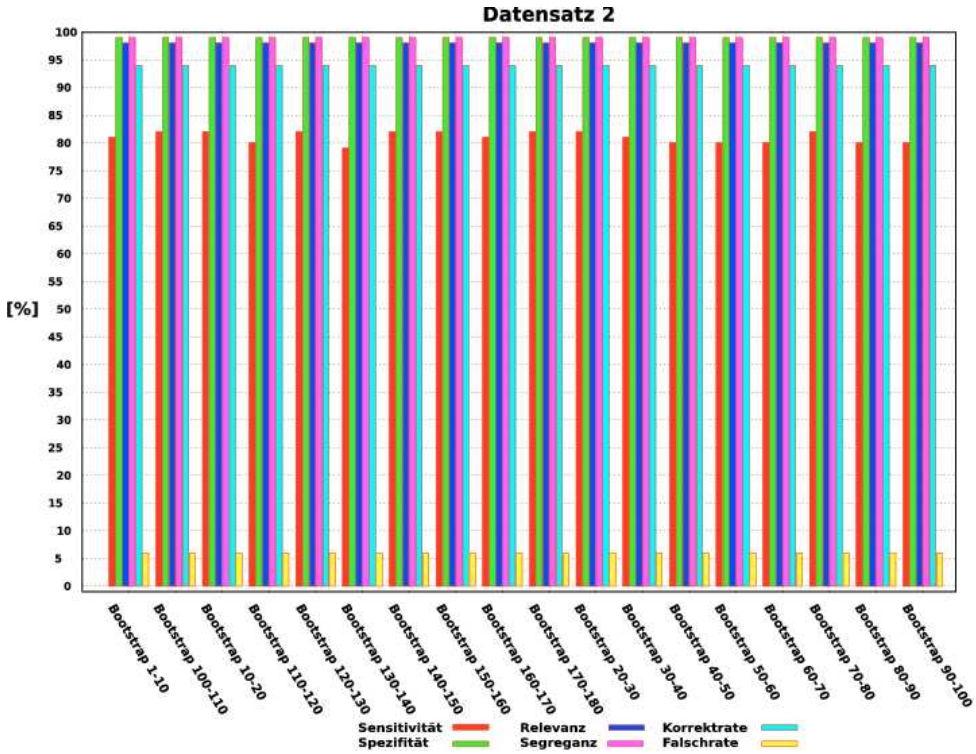

Abbildung 3.12: Bootstrap-Diagramm basierend auf dem Datensatz 2 und der miRBase 10.0. Die Erhebung der statistischen Daten des dargestellten Diagramms fand durch Auslassen von jeweils zehn microRNA-Sequenzen aus dem richtig-positiven Sequenzdatensatz (miRBase 10.0), Training mit den verbleibenden microRNA-Sequenzen und dem richtig-negativen Sequenzdatensatz 2 statt. Die anschließende Klassifizierung fand über die Gesamtmenge der richtig-positiven sowie richtig-negativen Sequenzdaten statt. Die X-Achse stellt die unterschiedlichen richtig-positiven Sequenzdatensätze mit den jeweilig ausgelassenen Sequenzen dar. Die Y-Achse stellt die Sensitivität, Spezifität, Relevanz, Segreganz, Falschrate sowie Korrektrate in % dar.

---

die gegebenen nicht-microRNA Sequenzen mit hoher Zuverlässigkeit korrekt klassifizieren konnte.

Als letzten richtig-negativen Testdatensatz zur Validierung von YamP wurden repetitive Elemente aus dem Genom von *Arabidopsis thaliana* gewählt. Repetitive Elemente umfassen z. B. sogenannte *Tandem-Repeat*-Sequenzen und *Interspersed Elements*, welche bei der simulierten thermodynamischen Faltung durch RNAfold, RNALfold und RNAshapes dazu neigen, thermodynamische stabile Hairpin-Strukturen auszubilden. Als richtig-negativer Sequenzdatensatz wurden daher die in der RepeatMasker-Bibliothek (Smit & Green, 2004) aufgelisteten repetitiven Sequenzen aus *Arabidopsis thaliana* gewählt. Die Ergebnisse der statistischen Auswertung sind in Tabelle 3.8 dargestellt. Es zeigte sich, dass YamP immer noch mit relativ hoher Zuverlässigkeit die vorliegenden richtig-negativen Sequenzen korrekt klassifizieren kann. Der Vergleich zu den bisherigen Werten bezüglich der Sensitivität und Spezifität zeigt jedoch einen nicht zu vernachlässigenden Abfall diesbezüglich. Daher scheint eine korrekte Klassifizierung repetitiver Regionen durch YamP in einem späteren genomischen Ansatz problematisch. Ein

Tabelle 3.7: Validierung durch zufällige Auswahl von pseudo-Hairpins. Die Tabelle zeigt die Klassifizierungseffizienz von YamP bei Anwendung auf eine zufällige Auswahl von pseudo-Hairpins aus den cDNA-Sequenzen von *Arabidopsis thaliana*. Pseudo-Hairpins beschreiben dabei Sequenzen, die nach Berechnung der Sekundärstruktur durch RNALfold eine Mindestlänge von 50 Nukleotiden und eine maximale Spannweite erlaubter Basenpaare von 400 Nukleotiden besitzen. Thermodynamische Beschränkungen lagen nicht vor. Die Erhebung der Sensitivität beruhte auf dem Trainingsdatensatz 2 und den microRNA-Sequenzen aus der miRBase in Version 7.0. Der Einsatz der miRBase in Version 10.0 änderte diese Ergebnisse nicht.

Testdatensatz	FP	Sensitivität	Sensitivität (bezogen auf die Gesamtmenge)
1	28	99,72%	98,57%
2	22	99,78%	98,36%
3	28	99,72%	98,56%
4	20	99,8%	98,97%
5	25	99,75%	98,74%
6	27	99,73%	98,58%
7	27	99,73%	98,57%
8	15	99,85%	99,2%
9	27	99,73%	98,56%
10	17	99,83%	99,13%
Mittelwert	23,6	99,76%	98,72%

Tabelle 3.8: Validierung durch Klassifizierung repetitiver Elemente aus dem Genom von *Arabidopsis thaliana*. Die Tabelle zeigt die Klassifizierungseffizienz von YamP bei Anwendung auf die RepeatMasker-Bibliothek. Die RepeatMasker-Bibliothek ist Zusammenstellung repetitiver genomischer Elemente aus *Arabidopsis thaliana*. Als Trainingsdatensätze lagen die microRNA-Sequenzen der miRBase in Version 7.0 sowie der richtig-negative Trainingsdatensatz 2 vor.

Spezifität	93,33%
Korrekt-Klassifikationsrate	93,33%
Falsch-Klassifikationsrate	6,67%

vorgeschalteter Filter-Schritt zur Identifikation solcher Elemente in den genomischen Sequenzdaten ist somit für einen genomischen Ansatz vonnöten, welcher durch das Programm RepeatMasker realisiert werden kann.

Bisher wurde das entwickelte Programm YamP nur zur Klassifizierung richtig-negativer Sequenzen eingesetzt und daraufhin getestet. Für die *de novo*-Identifikation unbekannter microRNA-Sequenzen musste zusätzlich gezeigt werden ob YamP ebenfalls in der Lage ist, dem Programm unbekannte microRNA-Sequenzen korrekt klassifizieren zu können. Zu diesem Zweck wurde ein letzter Validierungsschritt eingeführt, welcher die Anwendung von YamP auf unbekannte microRNA-Sequenzen beinhaltete. Dazu wurden 67 microRNA-

Sequenzen, die im Verlauf dieser Arbeit zusätzlich in der miRBase annotiert wurden, als microRNA-Datensatz verwendet.

YamP wurde dabei mit den in Abschnitt 3.3 erwähnten Datensätzen 1 und 2 als richtig-negativen Trainingsdatensätzen und der miRBase in Version 7.0 als microRNA-Sequenzdatensatz trainiert und auf die Menge der unbekannten microRNAs angewendet. Wie aus Tabelle 3.5 ersichtlich, handelt es sich bei dem richtig-negativen Datensatz 1 um den Datensatz, der sensitiver microRNA-Sequenzen korrekt klassifizieren kann. Der richtig-negative Datensatz 2 ist hingegen spezifischer bezüglich der Klassifizierung richtig-negativer Sequenzdaten.

Die Klassifizierung durch YamP mit dem statistischen Modell, dass auf dem richtig-negativen Trainingsdatensatz 1 und den microRNA-Sequenzen aus der miRBase 7.0 beruhte, zeigte eine zuverlässige Klassifizierung auch mit bei dem Programm unbekannten microRNA-Sequenzen Es konnten von den 67 microRNA-Sequenzen 55 von YamP korrekt als microRNA-Sequenzen klassifiziert werden, was einer Sensitivität von ca. $82,09\%$ entsprach.

Auch der zweite Durchlauf des Programms YamP mit dem richtig-negativen Trainingsdatensatz 2 und der miRBase 7.0 als richtig-positivem Trainingsdatensatz klassifizierte die dem Programm unbekannten microRNA-Sequenzen noch mit hoher Zuverlässigkeit. Dabei wurden von den 67 microRNA-Sequenzen noch 48 von YamP korrekt als microRNA-Sequenzen klassifiziert, was immer noch einer Sensitivität von ca. $71,64\%$ entsprach.

Beide Durchläufe zeigten, dass YamP nicht nur in der Lage ist, mit hoher Zuverlässigkeit richtig-negative Sequenzen korrekt klassifizieren zu können, sondern zudem unbekannte microRNA-Sequenzen mit hoher Zuverlässigkeit korrekt klassifizieren kann. Die ermittelten Sensitivitäten beruhen auf den unterschiedlichen richtig-negativen Trainingsdatensätzen 1 und 2, sowie dem richtig-positiven Datensatz aus der miRBase 7.0 entsprechen ungefähr denen, die auch schon in Abschnitt 3.5 gezeigt wurden.

Über die beschriebenen, unterschiedlichen Testumgebungen aus Abschnitt 3.3.1 konnte gezeigt werden, dass YamP mit hoher Zuverlässigkeit in der Lage ist, microRNA-Sequenzen von nicht-microRNA-Sequenzen zu unterscheiden. Da die Funktionalität und besonders die Zuverlässigkeit in diesem Abschnitt gezeigt werden konnte, folgt im nächsten Schritt die Anwendung auf genomische Sequenzdaten, worauf in Kapitel 4 eingegangen wird.

4

Anwendung

In Kapitel 3 wurde das in dieser Arbeit entwickelte Programm Yet Another miRNA Predictor (YamP) hinsichtlich seines algorithmischen Hintergrundes erläutert sowie dessen prinzipielle Funktionalität gezeigt. Bei diesen Untersuchungen zeigte sich, dass YamP grundsätzlich in der Lage ist, bekannte und nicht-bekannte microRNA-Vorläufer von nicht-microRNA-Vorläufern mit einer sehr guten Effizienz korrekt zu unterscheiden. Der nächste Schritt sollte die Anwendung auf genomische Sequenzdaten aus dem Modellorganismus *Arabidopsis thaliana* darstellen. Dabei sollten neben den bereits bekannten weitere, bisher unbekannte microRNA-kodierende Bereiche im Genom identifiziert und verifiziert werden, Die Identifikation solcher Bereiche und somit eine positive Klassifizierung durch YamP reicht alleine nicht für eine Verifikation aus. Daher wurden die identifizierten Bereiche mit den bestehenden Expressionsdatenbanken, MPSS und ASRP (siehe Abschnitt 2.4), abgeglichen, um unabhängige, experimentell ermittelte Daten zur Verifikation der ermittelten Bereiche zu erhalten.

Da bekannt ist, dass die Mehrzahl der bekannten microRNA-Sequenzen innerhalb von intronischen und intergenischen Bereichen kodiert ist (Sunkar & Zhu, 2004), wurden diese Sequenzbereiche von *The Arabidopsis Information Resource* (TAIR) zur näheren Analyse durch YamP bezogen. Eine Einschränkung des Suchraums auf die intergenischen und intronischen Sequenzregionen ist zwar aufgrund der Fähigkeiten von YamP nicht zwingend nötig, erleichtert jedoch die Auswertung und verringert die Laufzeit erheblich. Somit konnten bereits im Vorhinein große Bereiche des Genoms und somit mögliche falschpositive Sequenzen ausgeschlossen werden. Diese Sequenzregionen aus dem Genom von *Arabidopsis thaliana* dienten im Folgenden als Suchraum für neue bis dato nicht identifizierte microRNA-Kandidaten.

Neben dem algorithmischen Hintergrund und der prinzipiellen Funktionalität wurden in Kapitel 3 der Einfluss unterschiedlich gewählter Schwellenwerte und Sequenzdatensätze zum Diskriminierungsverhalten des Programms YamP untersucht. Zwei der fünf untersuchten Trainingsdatensätze zeichneten sich bei diesen Untersuchungen besonders aus. Der

Trainingsdatensatz 2 aus dem Abschnitt 3.3.1 erzielte bezüglich der Spezifität die besten Resultate, was jedoch mit Einbußen bei der Sensitivität einherging. Bei dem zweiten Datensatz handelte es sich um den Trainingsdatensatz 1, welcher eine sehr hohe Sensitivität aufwies, welche gleichwohl auf Kosten der Spezifität gewonnen wurde. Die genomweite Suche nach neuen möglichen microRNA-Kandidaten umfasste die Gesamtheit der annotierten intergenischen und intronischen Sequenzbereiche. Diese Sequenzdaten umfassten 180.000 einzelne Sequenzen, welche im Fall der intronischen Sequenzen eine durchschnittliche Länge von 160 Nukleotiden und im Fall der intergenischen Sequenzbereiche eine durchschnittliche Länge von 1.670 Nukleotide besaßen. Da aufgrund dieser Menge an Sequenzdaten eine große Anzahl möglicher Hairpin-Strukturen zu erwarten war, wurde für die genomweite Suche nach neuen möglichen microRNA-Kandidaten das statistische Modell gewählt, welches bezüglich der Spezifität das beste Diskriminierungsverhalten aufwies. Dieses Modell basierte auf dem Trainingsdatensatz 2 sowie der miRBase in Version 10.0.

Die Analyse der in YamP implementierten und in Tabelle 3.4 vorangestellten Filter-Schritte inklusive ihrer Schwellenwerte wurde größtenteils für den genomischen Ansatz übernommen. Zwei Modifikationen sollten aber auch hier die Menge der falsch-positiven Treffer im Vorhinein zusätzlich deutlich einschränken: Die Mindestlänge eines zu klassifizierenden Hairpins in seiner umformatierten Struktur (eingeführt in Abschnitt 3.1.3), welche hier auf 35 Zeichen (vier Nukleotide inklusive des Gap-Symbols „-“) festgesetzt wurde und die Wahl des Schwellenwerts für den in Abschnitt 3.2.2 vorangestellten *Window-Slide*-Filter. Dieser zeigte in den Testläufen die beste Effizienz bezüglich der Diskriminierung microRNA- und nicht-microRNA Sequenzen. In den Testläufen wurde der Schwellenwert für diesen Filter mit $0,65$ noch sehr liberal gewählt. Im genomischen Ansatz wurde er aufgrund der zu erwartenden hohen Anzahl an möglichen, falsch-positiven Hairpin-Strukturen, auf $0,75$ gesetzt und damit konservativer gewählt. Diese beiden Modifikationen ließen erwarten, dass die Menge der falsch-positiven Treffer dadurch deutlich verringert werden kann.

Im Folgenden werden zunächst die allgemeinen Ergebnisse des genomischen Ansatzes vorgestellt und diskutiert, bevor im weiteren Verlauf auf fünf, exemplarisch für die Menge der Ergebnisse ausgewählte Kandidaten genauer eingegangen wird. Bei den hier vorgestellten microRNA-Kandidaten lagen neben der Klassifizierung durch YamP weitere gute Hinweise vor, diese Bereiche zukünftig als microRNA-kodierende Bereiche zu annotieren. Diese beruhen auf experimentell ermittelten und der Öffentlichkeit zur Verfügung gestellten Expressionsdaten kleiner RNA-Sequenzen in *Arabidopsis thaliana* (Expressionsdatenbanken ASRP und MPSS). Um eine möglichst gute Beschreibung der vorliegenden microRNA-Kandidaten zu erhalten, wurde zusätzlich eine Suche nach möglichen Zielgenen der vorgestellten Kandidaten durchgeführt (siehe Abschnitt 4.5).

Ein weiteres Ziel dieser Arbeit stellte die Lokalisation bzw. Klassifizierung der stäbchenförmigen Sekundärstruktur des *PSTVd* dar. Es sollten zudem weitere Indizien gesammelt werden, welche zusätzliche Hinweise auf den Pathogenitätsmechanismus geben sollten. Eine erste Klassifizierung des *PSTVd* auf Basis des statistischen Modells von microRNAs aus *Arabidopsis thaliana* mit anschließender Lokalisation der Berei-

che, welche potentiell Ursprung kleiner Viroid-spezifischer RNAs sein könnten, wird in Abschnitt 4.4 vorgestellt.

4.1 Allgemeines zu den Ergebnissen

Das Programm YamP wurde mit den in Tabelle 3.4 definierten Schwellenwerten und den zwei beschriebenen Modifikationen auf die intergenischen (Abschnitt 4.3) und intronischen Sequenzen (Abschnitt 4.2) angewendet.

Der intergenische Sequenzdatensatz umfasste insgesamt 30.413 Sequenzen verschiedener Länge. Durch RNALfold und RNAshapes wurden aus dieser Grundmenge an Sequenzen 2.827 nicht-redundante Sequenzen extrahiert, welche die Filter-Vorgaben erfüllten und anschließend einer Klassifizierung durch YamP unterzogen wurden. Die Suche nach lokalen, stabilen Strukturelementen durch RNALfold ergab für einen bestimmten Sequenzbereich unter Umständen mehrere stabile Strukturen, da man von der 5'- und 3'-Seite jeweils mehrere Nukleotide entfernen kann und die verbleibende Sequenz immer noch eine thermodynamisch stabile Sekundärstruktur ausbildet. Diese unterschiedlich langen, aber allesamt stabilen Hairpin-Strukturen gingen dabei unabhängig voneinander in die Klassifizierung durch das Programm YamP ein. Eine weitere Quelle für das Auftreten redundanter Sequenzen bzw. Strukturen war die simulierte thermodynamische Faltung durch RNAshapes. RNAshapes faltet die Eingabesequenzen für eine Temperatur von 37 ℃ und selektiert Strukturen mit einem definierten Energiebetrag unter dem thermodynamischen Optimum. Dabei kann eine Sequenz im definierten Energiebereich mehrere stabile Sekundärstrukturen ausbilden, welche sich meist nur durch die Lage einzelner, interner symmetrischer oder asymmetrischer Loops unterscheiden. Auch diese Strukturen wurden dabei unabhängig voneinander durch YamP klassifiziert.

Die folgende Auflistung zeigt die Anzahl der insgesamt in Betracht gezogenen Sequenzen aus dem intergenischen Sequenzdatensatz sowie die Anzahl der insgesamt als microRNA- und nicht-microRNA-Sequenz klassifizierten Sequenzen:

insgesamt klassifizierte Sequenzen	2.827
als microRNA klassifiziert	828
als nicht-microRNA klassifiziert	1.999

Der intronische Sequenzdatensatz umfasste insgesamt 148.558 Sequenzen verschiedener Länge. Aus dieser Grundmenge wurden insgesamt 8.161 nicht-redundante Sequenzen extrahiert, welche auch hier die Filter-Vorgaben erfüllten und von YamP klassifiziert wurden. Die folgende Auflistung zeigt die Anzahl der insgesamt in Betracht gezogenen Sequenzen aus dem intronischen Sequenzdatensatz, sowie die Anzahl der insgesamt als microRNA- und nicht-microRNA-Sequenz klassifizierten Sequenzen:

insgesamt klassifizierte Sequenzen	8.161
als microRNA klassifiziert	649
als nicht-microRNA klassifiziert	7.512

Bei den separat durchgeführten Durchläufen mit dem Programm YamP wurden im Fall der intergenischen Sequenzen ca. 70% der zu klassifizierenden Hairpin-Strukturen als nicht-microRNA-Strukturen abgelehnt und ca. 30% als microRNA-Struktur angenommen. Für den Durchlauf mit den intronischen Sequenzen wurden ca. 92% als nicht-microRNA-Struktur abgelehnt und ca. 8% als microRNA-Struktur angenommen. Insgesamt handelte es sich dabei jedoch immer noch um 1.477 positiv-klassifizierte Strukturen, welche im Folgenden einer näheren Untersuchung unterzogen wurden.

Für das Vorliegen eines tatsächlichen microRNA-Vorläufers in dem Bereich der positiv-klassifizierten Region mussten neben dem bloßen Vorliegen einer Hairpin-Struktur und einer Positiv-Klassifizierung durch YamP weitere Hinweise gesammelt werden. Dazu wurden im ersten Schritt zunächst die genauen genomischen Koordinaten der positiv-klassifizierten Hairpin-Strukturen mit Hilfe von HyPa ermittelt und für jeden möglichen Kandidaten vermerkt. Die so ermittelten genomischen Koordinaten wurden in einem zweiten Schritt mit zwei Expressionsdatenbanken (ASRP und MPSS) abgeglichen und nach Bereichen gesucht, die eine Expression von kleinen RNA-Sequenzen im Bereich der klassifizierten Hairpin-Struktur in der entsprechenden Orientierung aufwiesen. Die entsprechende Orientierung bedeutet, dass der Ursprung der kleinen RNA-Sequenzen tatsächlich im Bereich des Hairpins lokalisiert ist und aus demselben Strang hervorgeht. Aus den bekannten microRNA-kodierenden Regionen aus dem Genom von *Arabidopsis thaliana* geht dabei hervor, dass diese meist, jedoch nicht immer, recht typische Expressionsmuster für kleine RNA-Sequenzen aufweisen, welche im Folgenden anhand zweier Beispiele (siehe Abbildung 4.1) von annotierten und verifizierten microRNA-Vorläufer-Sequenzen veranschaulicht werden. Über diese typischen Expressionsmuster wurden die Kandidaten ausgewählt, die in Abschnitt 4.2 und 4.3 als neue mögliche microRNA-Sequenzen vorgestellt werden, wobei beide in der Abbildung 4.1 dargestellten microRNA-Vorläufer-Regionen ebenfalls von YamP als microRNA-kodierende Bereiche identifiziert worden. Es ist an dieser Stelle zu erwähnen, dass die beiden in Abbildung 4.1 vorgestellten bekannten microRNA-Sequenzen in den intergenischen und intronischen Sequenzdaten vom TAIR noch nicht annotiert und somit noch nicht aus diesen entfernt wurden. Beide microRNAs wurden bei der Suche nach neuen möglichen microRNA-Kandidaten aus den Sequenzdaten als microRNA-Sequenzen klassifiziert, was wiederum zeigt, dass eine prinzipielle Funktionalität durch das Programm YamP gegeben ist.

Ein microRNA-kodierender Bereich im Genom von *Arabidopsis thaliana* zeichnet sich im Regelfall durch ein typisches Expressionsmuster aus, bei dem kleine RNA-Sequenzen mit einer vornehmlichen Länge von 20–21 Nukleotiden einem oder zweier distinkter Bereiche entstammen. In Abbildung 4.1**A** ist der genomische Bereich der verifizierten microRNA *ath-mir158b* dargestellt. Es ist deutlich zu erkennen, dass die Mehrheit der diesem Bereich zugeordneten kleinen RNA-Sequenzen einem Bereich der Hairpin-Struktur zuzuordnen sind, welcher 3'-seitig in der microRNA-kodierenden Region liegt. Von der *ath-mir158b* ist bekannt, dass die reife microRNA-Sequenz in eben diesem Bereich der Hairpin-Struktur kodiert ist. Ein zweiter Fall für einen microRNA-kodierenden Bereich ist in Abbildung 4.1**B** am Beispiel des verifizierten microRNA-Vorläufers *ath-mir842a* veranschaulicht. Dort ist zu erkennen, dass es zwei distinkte Regionen innerhalb des

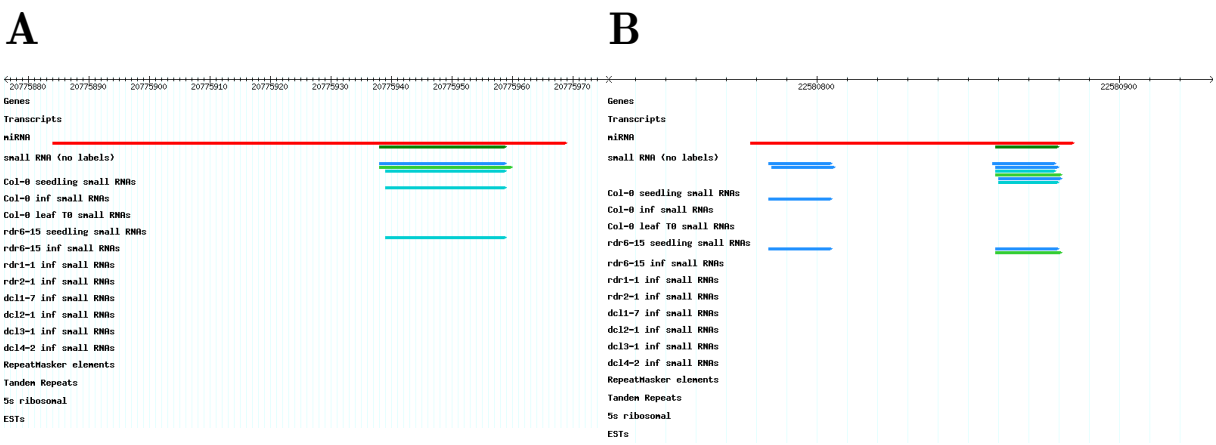

Abbildung 4.1: Typische Expressionsmuster zweier microRNA-Vorläufer. Dargestellt sind zwei für microRNA-kodierende Bereiche im Genom von *Arabidopsis thaliana*. Die roten Pfeile deuten auf die tatsächlichen microRNA-Vorläufer-Sequenzen hin. Die kleineren, hell- und dunkelblauen Pfeile stehen für RNA-Sequenzen der Länge 20 bzw. 21 und die violetten und brauen Pfeile für RNA-Sequenzen der Länge 23 bzw. 25. Abbildung **A** zeigt die genomische Region, welche für den microRNA-Vorläufer *ath-mir158b* kodiert. Es ist deutlich zu erkennen, dass hauptsächlich nur eine Region aus dem Bereich des microRNA-Vorläufers Ursprung kleiner RNA-Sequenzen ist. Abbildung **B** zeigt die genomische Region, welche für den microRNA-Vorläufer *ath-mir842a* kodiert. Daran wird deutlich, dass neben dem Bereich der reifen microRNA-Sequenz noch ein zweiter Bereich Ursprung kleiner RNAs ist, in deren Bereich die korrespondierende microRNA*-Sequenz angesiedelt ist.

Hairpin-kodierenden Bereichs gibt, für den kleine RNA-Sequenzen sequenziert wurden. Von der *ath-mir842a* ist bekannt, dass die reife microRNA-Sequenz ebenfalls 3'-seitig der microRNA-kodierenden Region ist, was ebenfalls aus der Abbildung 4.1**B** hervorgeht. Weiterhin ist zu erkennen, dass 5'-seitig der microRNA-kodierenden Region, und somit im 5'-Strang des Hairpins, ebenfalls ein Bereich Ursprung kleiner RNA-Sequenzen ist. Hierbei handelt es sich um die entsprechende microRNA*-Sequenz, welche in diesem Fall ebenfalls sequenziert und in der Expressionsdatenbank vom ASRP eingetragen wurde.

Neben diesen beiden exemplarisch veranschaulichten Beispielen gibt es jedoch weitere bekannte microRNA-kodierende Bereiche im Genom von *Arabidopsis thaliana*, deren Expressionsmuster sich deutlich von den gezeigten unterscheiden. So sind Beispiele bekannt (hier nicht gezeigt), in denen sequenzierte kleine RNA-Sequenzen, die über den gesamten microRNA-kodierenden Bereich verteilt sind, zugeordnet wurden. Weiterhin ist bekannt, dass reife microRNA-Sequenzen in der Regel 20–21 Nukleotide lang sind, was jedoch nicht allgemeingültig für die annotierten microRNA-kodierenden Bereiche ist. Trotzdem wurden die im Folgenden besprochenen Kandidaten so gewählt, dass sie typischen microRNA-kodierenden Bereichen ähnlich sind.

Das Vorliegen kleiner RNAs in der Expressionsdatenbank von TAIR stellte jedoch nicht die einzige Bedingung dar, die positiv-klassifizierten Regionen im Genom als microRNA-kodierend zu bezeichen. Zu diesem Zweck wurde noch weitere experimentelle Daten herangezogen, die das Vorliegen eines microRNA-kodierenden Bereichs bestätigen sollten.

Weitere experimentelle Daten konnten von der Expressionsdatenbank *Arabidopsis MPSS plus* bezogen werden, die Sequenz-Signaturen kleiner exprimierter RNAs veröffentlichten. Somit konnten neben der Klassifizierung durch YamP unabhängige experimentelle Daten verwendet werden, um die als microRNA-kodierend vorhergesagten Regionen im Genom von *Arabidopsis thaliana* verifizieren zu können. Somit erfüllen die hier vorgestellten Kandidaten drei Kriterien, nach denen sie ausgewählt wurden; die Klassifizierung durch das Programm YamP sowie das Vorliegen exprimierter kleiner RNA-Sequenzen in den entsprechenden genomischen Bereichen aus zwei voneinander unabhängigen Expressionsdatenbanken.

4.2 Intronische Sequenzen

Als Suchraum für neue microRNA-kodierende Bereiche im Genom von *Arabidopsis thaliana* dienten in diesem Ansatz die intronischen Sequenzdaten von der Sequenzdatenbank TAIR. Die vorsortierten Sequenzdaten wurden gewählt, um den Suchraum schon im Voraus einschränken zu können und mögliche Klassifizierungen von bereits bekannten microRNA-Sequenzen umgehen zu können. Bei diesem Ansatz stellte sich jedoch heraus, dass nicht alle bereits bekannten microRNA-Sequenzen aus *Arabidopsis thaliana* in den Sequenzdaten von TAIR annotiert waren. Dies führte dazu, dass bereits bekannte, jedoch nicht annotierte microRNA-Sequenzen ebenfalls klassifiziert wurden. Im Folgenden wird ein Kandidat vorgestellt, der in einem Intron einer mRNA lokalisiert ist.

4.2.1 Kandidat 1

Der erste vorgestellte Kandidat aus den intronischen Sequenzbereichen befindet sich auf Chromosom *I* von *Arabidopsis thaliana*. Die durch RNAfold für 25 ℃ berechnete Sekundärstruktur ist in Abbildung 4.2 zu sehen. Dabei bildet die Sequenz eine für microRNA-Vorläufer typische Hairpin-Struktur aus und besitzt eine minimale freie Energie von $-67,18$ kcal/mol. Dies entspricht einer auf die Länge normierten Energie von ca. $-0,476$ und zeigt somit eine thermodynamische Stabilität, die typisch für microRNA-Vorläufer-Strukturen ist (Zhang *et al.*, 2006). Weitere Eigenschaften der vorhergesagten Sekundärstruktur und Sequenz sind in der folgenden Auflistung zusammengestellt:

Position auf Chromosom *I*	234.152 – 234.013
Länge	141 Nukleotide
mfe (für 25 ℃)	$-67,18$ kcal/mol
GC-Gehalt	38,6%

Der vorgestellte microRNA-Kandidat befindet sich in einem Intron einer mRNA mit der *Accession*-Nummer *AT1G01650* (siehe Abbildung 4.3). Diese Region kodiert nach den vorliegenden Annotationen und einer BLAST-Suche für ein Protein mit 540 Aminosäuren

Abbildung 4.2: Kandidat 1 aus den intronischen Sequenzen. Die Sekundärstruktur-Berechnung für den ersten Kandidaten aus den intronischen Sequenzen ergab durch RNAFOLD für 25 ℃ eine für microRNAs typische Hairpin-Struktur. In rot hervorgehoben sind die in Abbildung 4.3 und Abbildung 4.4 ermittelten sequenzierten kleinen RNA-Sequenzen aus dem Genom von *Arabidopsis thaliana*. Beide befinden sich um zwei Nukleotide verschoben am 5'- und 3'-Ende des Hairpins.

---

und besitzt zwei konservierte Domänen. Die konservierten Domänen besitzen zum einen Ähnlichkeit mit einer Protease-Domäne (N-Terminus) und zum anderen einer Peptidase-Domäne (C-Terminus). Das in der Expressionsdatenbank vom ASRP vorliegende Expressionsmuster (siehe Abbildung 4.3) entspricht dem eines typischen microRNA-kodierenden Bereichs, was als ein gutes Indiz für das Vorliegen einer bisher unbekannten microRNA-kodierenden Region zu deuten ist. In der Abbildung 4.3 ist deutlich zu erkennen, dass zwei distinkte Bereiche Ursprung kleiner RNA-Sequenzen sind, welche vornehmlich dieselbe Orientierung besitzen.

Die weitere Validierung des ersten Kandidaten fand über die Expressionsdatenbank MPSS statt. Dabei wurden die mit Hilfe von HYPA ermittelten genomischen Koordinaten 5'- und 3'-seitig um jeweils 2.000 Nukleotide verlängert, um eine Übersicht über den gesamten genomischen Bereich und die unmittelbare Nachbarschaft zu bekommen (siehe Abbildung 4.4). Die Darstellung zeigt den genomischen Bereich für die von YAMP als microRNA-kodierend vorhergesagten Sequenzbereich. Die schwarzen und schwarz umrandeten Dreiecke zeigen das Vorkommen von Sequenz-Signaturen kleiner RNAs an, wobei die schwarz ausgefüllten Dreiecke für Signaturen stehen, die dem Genom eindeutig zugeordnet werden konnten. Die umrandeten Dreiecke stehen für Sequenzsignaturen, die dem Genom mehr als einmal zugeordnet werden konnten. Die doppelt im Genom vorhandenen kleinen RNAs befinden sich jeweils im reversen Komplement des hier vorliegenden genomischen Bereichs. Dabei sei erwähnt, dass nur die kleinen RNAs duplizierte Sequenzen darstellen. Der restliche Hairpin weist keinerlei interne Duplizität auf.

Der vorliegende genomische Bereich zeigt insgesamt sehr gute Indizien, ihn als microRNA-Vorläufer zu bezeichnen. Eine BLAST-Suche mit der entsprechenden Sequenz über das Genom von *Arabidopsis thaliana* deckte dabei noch ein weiteres Indiz auf, dass es sich hierbei um einen microRNA-Vorläufer handelt. Allen *et al.* beschrieben 2004, dass einige microRNA-kodierende Bereiche aus invertierten Duplikationsereignissen ihrer korrespondierenden Zielgene heraus entstanden sein könnten. Sie belegten dies durch Sequenzvergleiche der microRNAs mit den entsprechenden Zielgenen. Die BLAST-Suche mit der Sequenz des besprochenen Kandidaten zeigte dabei neben der entsprechenden genomischen Region der Sequenz eine signifikante Ähnlichkeit zu einer kodierenden Region auf dem Chromosom *IV*, die für ein Protein kodiert, welches in die Transkriptionsregulation involviert ist und auch in Abschnitt 4.5.1 als mögliches Zielgen vorgestellt wird. Diese Tat-

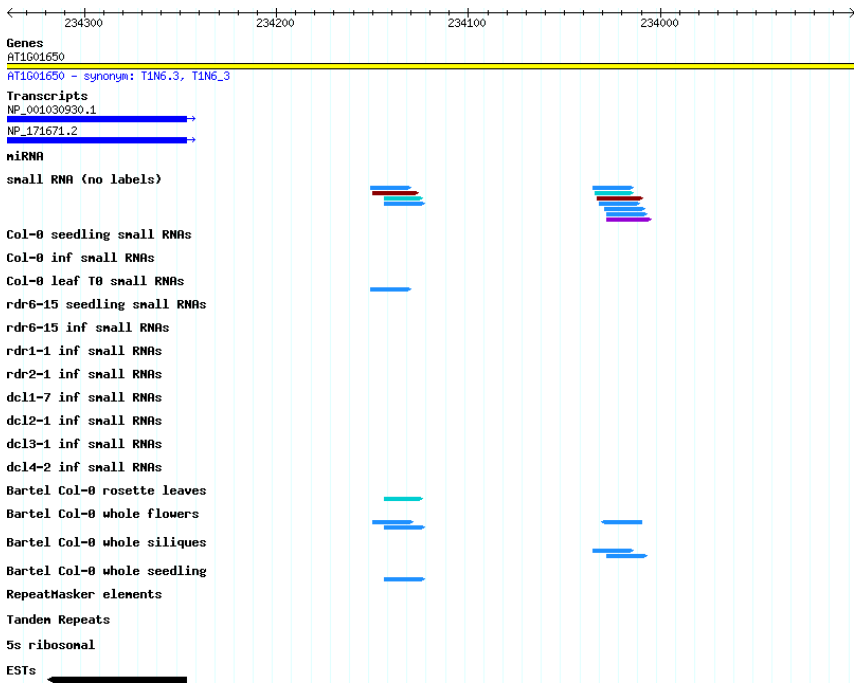

Abbildung 4.3: ASRP-Expressions-Datenbank-Informationen zum ersten Kandidaten. Der erste Kandidat befindet sich auf dem Chromosom *I* von *Arabidopsis thaliana* in einem Intron einer Protein-kodierenden Region. Das für viele microRNAs typische Expressionsmuster mit zwei distinkten Bereichen als Ursprung kleiner RNAs ist deutlich zu erkennen. Weiterhin werden hauptsächlich kleine RNAs mit 20 bzw. 21 Nukleotiden exprimiert, was auf einen tatsächlichen microRNA-kodierenden Bereich hinweist (erkennbar an den hell- und dunkelblauen Pfeilen). Die weiteren sequenzierten kleinen RNAs, erkennbar an den violetten und braunen Pfeilen, stehen für RNAs der Länge 23 bzw. 25 Nukleotide.

---

sache spricht dafür, dass dieser microRNA-Kandidat ebenfalls aus einem von Allen *et al.* beschriebenem Vorgang entstanden ist. Eine detaillierte Analyse der Zielgen-Identifikation für den ersten Kandidaten folgt in Abschnitt 4.5.1.

4.3 Intergenische Sequenzen

Die Grundlage für die Suche nach weiteren microRNA-kodierenden Bereichen im Genom von *Arabidopsis thaliana* stellten die intergenischen Sequenzbereiche dar. Diese wurden wie die intronischen Sequenzen von TAIR bezogen. Im Folgenden werden vier intergenische Regionen vorgestellt, bei denen neben der Klassifizierung durch YAMP weitere gute Kriterien vorliegen, diese künftig als microRNA-kodierend zu bezeichnen. Diese vier Bereiche stellen ebenso nur eine exemplarische Auswahl dar.

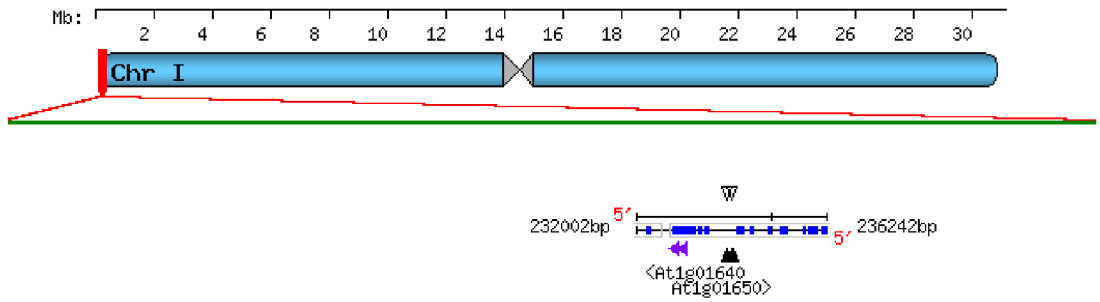

Abbildung 4.4: MPSS-Signaturen aus dem Bereich des ersten Kandidaten. Die Abbildung zeigt die Signaturen für vier kleine RNAs. Die schwarz ausgefüllten Dreiecke stehen dabei für einmalig im Genom vorkommende Signaturen, während die umrandeten Dreiecke für doppelte Signaturen stehen. Es ist zu erwähnen, dass die doppelten Signaturen genau mit den einmalig vorkommenden Signaturen übereinstimmen und auch genau die in der ASRP vermerkten kleinen RNAs darstellen, welche in Abbildung 4.2 in rot gezeigt wurden. Die Exons des Protein-kodierenden Bereichs sind an den blauen Blöcken um den Ort der kleinen RNAs zu erkennen.

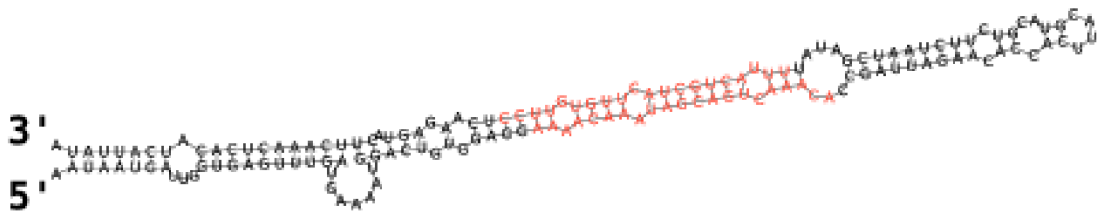

Abbildung 4.5: Kandidat 2 aus intergenischen Sequenzen. Die Darstellung zeigt die Sekundärstruktur des zweiten Kandidaten. Die Lokalisation des möglichen reifen microRNA/microRNA*-Duplex (rot) erfolgte über die ASRP-Expressionsdatenbank. Vermerkt wurden die Sequenzen, welche in der ASRP die zwei distinkten Bereiche darstellten.

4.3.1 Kandidat 2

Die zweite vorgestellte genomische Region ist auf dem Chromosom *II* von *Arabidopsis thaliana* lokalisiert. Die durch RNAFOLD berechnete Sekundärstruktur der entsprechenden Sequenz ist in Abbildung 4.5 dargestellt. Diese bildet einen einzigen stabilen Hairpin aus und besitzt eine minimale freie Energie (*mfe*) von $-72,71\,\text{kcal/mol}$ für 25 ℃. Normiert auf die Länge ergibt dies einen Wert von $-0,482$ und entspricht somit einer für microRNA-Vorläufer typischen thermodynamischen Stabilität (Zhang *et al.*, 2006). Weitere Eigenschaften der vorhergesagten Sekundärstruktur und Sequenz sind in der folgenden Auflistung zusammengestellt:

Position auf dem Chromosom	8.439.729 – 8.439.879
Länge	150 Nukleotide
mfe (für 25 ℃)	$-72,71\,\text{kcal/mol}$
GC-Gehalt	35,8%

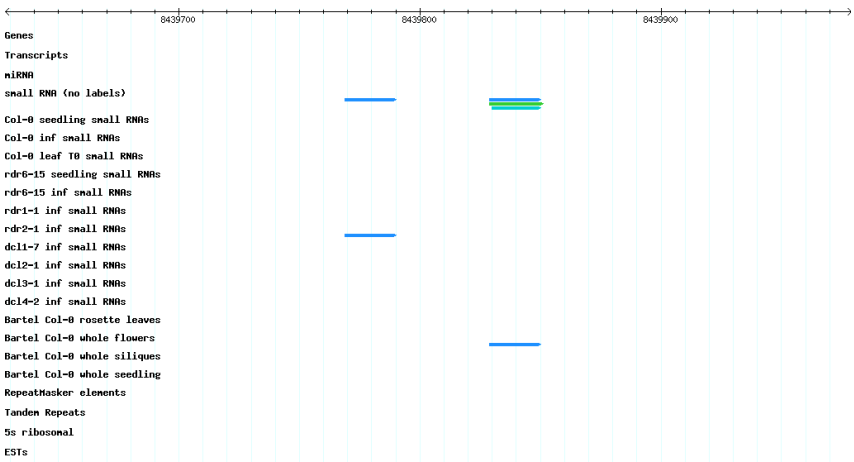

Abbildung 4.6: ASRP-Expressions-Datenbank-Informationen zum zweiten Kandidaten. Die Darstellung zeigt die Expression von kleinen RNA-Sequenzen im Sequenzbereich des zweiten Kandidaten. Erkennbar sind hier zwei distinkte Bereiche als Ursprung für kleine RNAs. Die für diesen Bereich sequenzierten kleinen RNAs besitzen die Länge von 20–22 Nukleotiden, wobei hellblaue Pfeile für RNAs der Länge 20, blaue für RNAs der Länge 21 und grün für RNAs der Länge 22 stehen.

Die Abbildung 4.6 zeigt die sequenzierten kleinen RNAs aus dem Bereich um den zweiten Kandidaten. Auffällig ist, dass, wie für viele microRNA-kodierende Bereiche typisch, zwei distinkte Bereiche Ursprung der kleinen RNAs sind. Die in diesem Bereich exprimierten kleinen RNAs entsprechen der Größenklasse, die 20–22 Nukleotide umfasst und somit der für microRNAs typischen Größenklasse entspricht. Die exprimierten kleinen RNAs sind in der Sekundärstruktur des zweiten microRNA-Kandidaten in Abbildung 4.5 in rot dargestellt.

Die weitere unabhängige Bestätigung der Vorhersage fand über die MPSS-Expressionsdatenbank statt. Der Bereich um den vorhergesagten microRNA-Vorläufer zeigt auch in dieser Expressionsdatenbank die Signatur einer kleinen, nur einmal im Genom vorkommenden Sequenz-Signatur und befindet sich in der Abbildung 4.7 auf dem Vorwärtsstrang des zweiten Chromosoms. Die eingetragene Signatur der kleinen RNA liegt im 3'-Arm der Hairpin-Struktur und entspricht den Expressionsdaten für einen der beiden Bereiche aus der ASRP-Expressionsdatenbank. Betrachtet man die für microRNAs typische Sekundärstruktur gemeinsam mit den experimentellen Daten aus den Expressionsdatenbanken, so liegt nahe, dass die von YAMP vorhergesagte Sequenz für eine microRNA kodiert.

Die weitere Analyse durch eine BLAST-Suche ergab, dass die Sequenz keinerlei signifikante Ähnlichkeit zu anderen Regionen im Genom von *Arabidopsis thaliana* aufweist. Somit lässt dies den Schluss zu, dass dieser microRNA-Kandidat nicht aus einem Duplikationsereignis (wie von Allen *et al.* (2004) beschrieben) entstanden ist. Eine detaillierte Beschreibung der Suche nach möglichen Zielgenen folgt in Kapitel 4.5

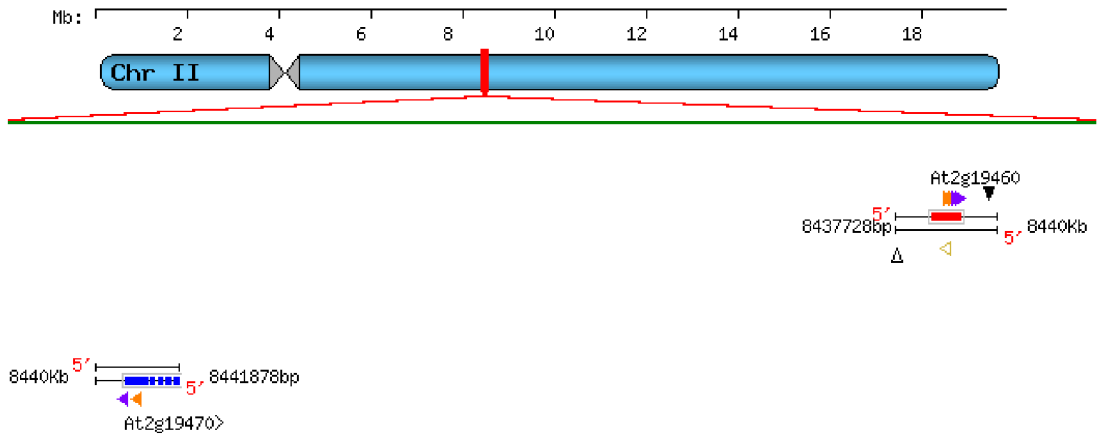

Abbildung 4.7: MPSS-Signaturen aus dem Bereich des zweiten Kandidaten. Die Abbildung zeigt die Signatur für eine kleine RNA aus dem genomischen Bereich für den zweiten Kandidaten. Diese Signatur stimmt mit der kleinen RNA aus der ASRP-Expressionsdatenbank überein, welche in der Sekundärstruktur im 3'-Arm der Hairpin-Struktur liegt. Die roten und blauen Balken stellen die Exons zweier Protein-kodierender Bereiche im Vorwärts- und Rückwärtsstrang der DNA dar, welche ebenfalls durch Signaturen für ebensolche Bereiche (orangene und violette Dreiecke) ausgezeichnet sind.

4.3.2 Kandidat 3

Die dritte vorgestellte genomische Region besitzt ihren Ursprung auf dem Chromosom *III* von *Arabidopsis thaliana* und ist an dieser Stelle von besonderem Interesse. Die Betrachtung der Ergebnisse dieses genomischen Bereichs ergab, dass die zu dem Bereich gehörenden exprimierten kleinen RNAs auch einer zweiten Region auf dem Chromosom *V* zugeordnet werden konnten. Im Folgenden wird hauptsächlich auf den als microRNA-klassifizierten Bereich auf dem Chromosom *III* eingegangen. Die Expressionsdaten für die korrespondierende Region auf dem Chromosom *V* aus beiden Expressionsdatenbanken werden für diesen Kandidaten zusätzlich vorgestellt und diskutiert. Die durch RNAFOLD bei 25 °C berechnete Sekundärstruktur dieses Sequenzbereichs ist in Abbildung 4.8 dargestellt und bildet eine stabile Hairpin-Struktur aus, welche durch eine Loop-reiche Region unterbrochen wird. Die Sekundärstruktur besitzt eine minimale freie Energie (*mfe*) von $-85,93$ kcal/mol, was auf die Länge normiert einen Wert von $-0,544$ ergibt. Somit handelt es sich bei dieser Hairpin-Struktur um eine thermodynamisch sehr stabile Struktur, die typisch für microRNA-Vorläufer ist (Zhang *et al.*, 2006). Einige weitere Eigenschaften der zugehörigen Sekundärstruktur und Sequenz sind in der folgenden Auflistung dargestellt:

Position auf dem Chromosom	2.854.290 – 2.854.448
Länge	158 Nukleotide
mfe (für 25 °C)	$-85,93$ kcal/mol
GC-Gehalt	40,9%

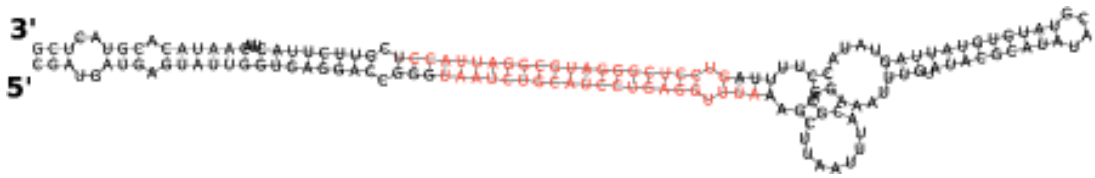

Abbildung 4.8: Kandidat 3 aus intergenischen Sequenzen. Die Darstellung zeigt die durch RNAFOLD bei 25 ℃ berechnete Sekundärstruktur des dritten Kandidaten. Die in der ASRP-Datenbank eingetragenen exprimierten kleinen RNA-Sequenzen sind in rot hervorgehoben (siehe Abbildung 4.9). Die Signaturen aus der MPSS-Expressionsdatenbank korrelieren mit der kleinen RNA auf der 5'-Seite der Hairpin-Struktur (siehe Abbildung 4.10).

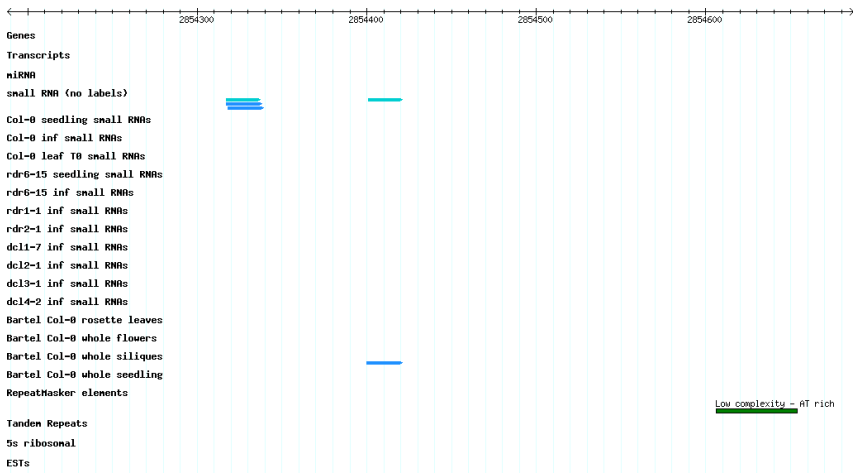

Abbildung 4.9: ASRP-Expressions-Datenbank-Informationen zum dritten Kandidaten. Die Abbildung zeigt die Expressionsdaten aus dem genomischen Kontext des dritten Kandidaten auf dem Chromosom *III*. Erkennbar sind zwei distinkte Bereiche als Ursprung exprimierter kleiner RNA-Sequenzen der Größenklasse von 20–21 Nukleotiden (hellblaue und blaue Pfeile).

Der Abgleich mit der ASRP-Expressionsdatenbank für diesen Kandidaten zeigte das Vorliegen zweier distinkter Bereiche als Ursprung kleiner RNAs (siehe Abbildung 4.9). Diese sind in der Abbildung 4.8 in rot hervorgehoben und korrespondieren mit einem möglichen microRNA/microRNA*-Duplex. Die exprimierten RNAs sind ausnahmslos aus der für microRNAs typischen Größenklasse von 20–21 Nukleotiden, was ein Hinweis für das Vorliegen einer microRNA-kodierende Region ist. Eine detaillierte Analyse und Beschreibung zur Identifikation möglicher Zielgene folgt in Abschnitt 4.5.3.

Die Vorhersage wurde durch weitere experimentelle Daten aus der MPSS-Expressionsdatenbank bestätigt. Im Bereich des dritten Kandidaten befindet sich eine Sequenz-Signatur, die sich nicht eindeutig dem Genom von *Arabidopsis thaliana* zuordnen lässt (siehe Abbildung 4.10). Weitere kleine RNA-Sequenzen wurden in dem Bereich des microRNA-Kandidaten und in seiner unmittelbaren Nachbarschaft nicht sequenziert. Wie bereits erwähnt, ist die vorliegende Signatur einer kleinen RNA nicht eindeutig im Genom von *Arabidopsis thaliana*, da auf dem Chromosom *V* ein zweiter Bereich existiert, welcher dieselbe kleine RNA-Signatur besitzt (siehe Abbildung 4.12). Dieser

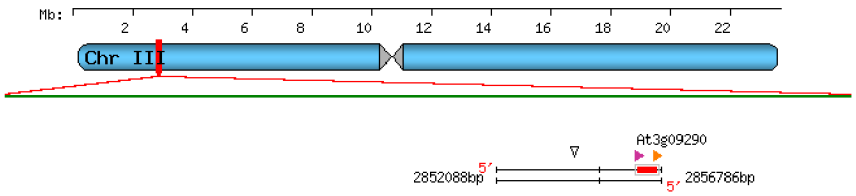

Abbildung 4.10: MPSS-Signaturen aus dem Bereich des dritten Kandidaten. Die Abbildung zeigt die Signatur für eine exprimierte kleine RNA-Sequenz, die zentriert in dem Ausschnitt zu sehen ist. Das schwarz umrandete Dreieck impliziert, dass die dem Genom zugeordnete Signatur nicht eindeutig für diesen genomischen Bereich ist. Diese Signatur kommt in einem zweiten Bereich im Genom von *Arabidopsis thaliana* vor, der in Abbildung 4.12 gezeigt ist.

Bereich ist dort ebenfalls nicht mit repetitiven oder transposablen Elementen assoziiert. Im Folgenden wurde die Sequenz des Bereichs auf dem Chromosom *V* ermittelt und einer weiteren Analyse unterzogen. Dabei erstreckt sich die Sequenzverdoppelung allein auf den microRNA-kodierenden Bereich, während der restliche Bereich nur wenig Ähnlichkeit aufweist. Die Anwendung von YAMP auf diese genomische Region klassifizierte diesen Bereich ebenfalls als möglichen microRNA-Vorläufer. Eine weitere Untersuchung des genomischen Bereichs um den entsprechenden Kandidaten durch den Abgleich mit den beiden Expressionsdatenbanken (MPSS und ASRP) zeigte auch in dieser Region die Expression kleiner RNAs in zwei distinkten Bereichen (siehe Abbildung 4.11). Die Suche nach möglichen Zielgenen der beiden möglichen microRNAs folgt in Abschnitt 4.5.3. Die Tatsache, dass zwei genomische Bereiche Ursprung derselben kleinen RNA-Sequenz sind, lässt den Schluss zu, dass es sich hierbei um eine neue Familie von microRNA-kodierenden Bereichen handeln könnte, welche dieselbe Gruppe von Zielgenen reguliert.

4.3.3 Kandidat 4

Die vierte vorgestellte Region hat ihren Ursprung auf dem Chromosom *IV* von *Arabidopsis thaliana*. In Abbildung 4.13 ist die mittels RNAFOLD für 25 °C berechnete Sekundärstruktur dargestellt Diese bildet eine stabile Hairpin-Struktur mit einer minimalen freien Energie von $-65,90$ kcal/mol aus. Normiert auf die Länge ergibt dies einen Wert von $-0,432$, was somit ebenfalls eine thermodynamische Stabilität aufweist, die der für microRNA-Vorläufer-Strukturen typischen Stabilität entspricht (Zhang *et al.*, 2006). Einige weitere Eigenschaften bezüglich der Sequenz und Sekundärstruktur sind in der folgenden Auflistung dargestellt:

Position auf dem Chromosom	11.963.137 - 11.962.889
Länge	157 Nukleotide
mfe (für 25 °C)	$-65,90$ kcal/mol
GC-Gehalt	28%

Abbildung 4.11: ASRP-Expressionsdaten kleiner RNAs eines Bereichs auf dem Chromosom *V*. Die Abbildung zeigt die Expressionsdaten aus dem genomischen Kontext des dritten Kandidaten auf dem Chromosom *V*. Erkennbar sind zwei distinkte Bereiche als Ursprung exprimierter kleiner RNA-Sequenzen der Größenklasse von 20–21 sowie 25 Nukleotiden (hellblaue, blaue und braune Pfeile).

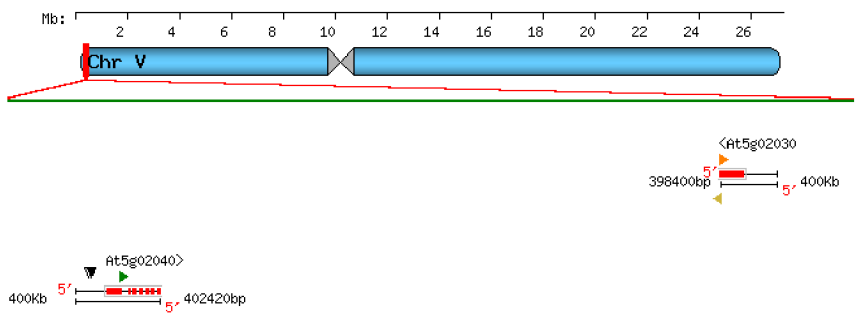

Abbildung 4.12: MPSS-Signaturen kleiner RNAs eines Bereichs auf dem Chromosom *V*. Die Abbildung zeigt einen Ausschnitt aus dem Chromosom *V*. In dieser Region des Genoms befinden sich Signaturen für zwei kleine RNAs, wovon die linke Signatur der Signatur des dritten Kandidaten aus Abbildung 4.10 entspricht.

Die ASRP-Expressionsdatenbank zeigt im Bereich des vierten Kandidaten die Expression von mehreren kleinen RNAs, die vornehmlich zwei distinkten Bereichen aus dieser Region zugeordnet werden können (siehe Abbildung 4.14). Die in diesem Bereich exprimierten kleinen RNAs entsprechen den für microRNAs typischen Größenklassen von 21–22 Nukleotiden, was als gutes Indiz für das Vorliegen einer microRNA-kodierenden Region zu deuten ist. In der näheren Umgebung des möglichen microRNA-Vorläufers befinden sich keine weiteren Bereiche, die Ursprung kleiner RNA-Sequenzen sind.

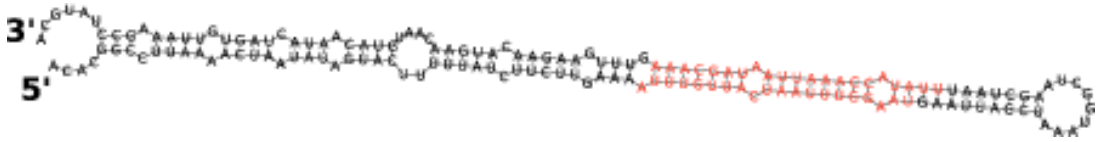

Abbildung 4.13: Kandidat 4 aus intergenischen Sequenzen. Die Darstellung zeigt die durch RNAFOLD bei 25 ℃ berechnete Sekundärstruktur des vierten Kandidaten. Die in rot hervorgehobenen Nukleotide markieren den Ort des möglichen microRNA/microRNA*-Duplex. Die aus der ASRP-Expressionsdatenbank entnommenen Sequenzen für kleine RNAs entsprechen denen auf 5'- und 3'-Seite der Hairpin-Struktur. Die Signatur aus der MPSS-Expressionsdatenbank entspricht der 5'-seitigen Sequenz der Hairpin-Struktur.

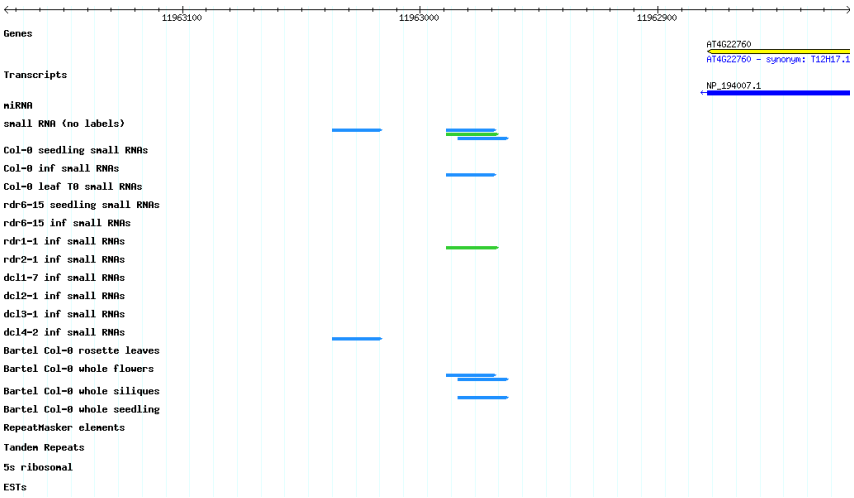

Abbildung 4.14: ASRP-Expressions-Datenbank-Informationen zum vierten Kandidaten. Die Darstellung zeigt die Expression von kleinen RNAs im Bereich des vierten Kandidaten. Es ist zu erkennen, dass auch bei diesem Kandidaten zwei distinkte Bereiche Ursprung kleiner RNA-Sequenzen sind. Die blauen Pfeile stehen für RNAs der Länge 20, die grünen Pfeile für RNAs der Länge 22. 3'-seitig erkennt man den beginnenden offenen Leserahmen einer Protein-kodierenden Region.

Die unabhängige Bestätigung für das Vorliegen exprimierter kleiner RNAs in diesem genomischen Bereich fand ebenfalls über einen Abgleich mit der MPSS-Expressionsdatenbank statt. Die betrachtete genomische Region ist auch hier Ursprung einer Sequenz-Signatur, die dem Genom von *Arabidopsis thaliana* eindeutig zugeordnet werden konnte und sich auf dem Gegenstrang des vierten Chromosoms befindet (siehe Abbildung 4.15). Diese Beobachtung bestätigt den Abgleich mit der ASRP-Expressionsdatenbank und stellt ein weiteres Indiz für das Vorliegen einer microRNA-kodierenden Region dar. Weitere annotierte Bereiche in dieser Region sind eine Protein-kodierende Region (horizontale rote Balken) auf dem Vorwärtsstrang der DNA sowie ein 5'-seitiger *Inverted Repeat*, welcher ebenfalls Signaturen für exprimierte kleine RNAs aufweist. In der ASRP-Expressionsdatenbank ist diese Region einem Transposon zugeordnet.

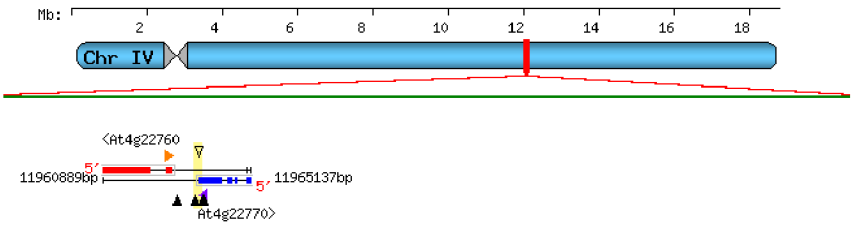

Abbildung 4.15: MPSS-Signaturen aus dem Bereich des vierten Kandidaten. Die Abbildung zeigt die Signatur einer kleinen RNA aus dem Bereich des vierten Kandidaten, erkennbar an dem schwarzen Dreieck auf dem Gegenstrang der DNA. Die zu dieser Signatur passenden Sequenz liegt in der Hairpin-Struktur im 5'-Arm und ist in Abbildung 4.13 in rot hervorgehoben. Des Weiteren sind zwei Protein-kodierende Regionen auf demselben (horizontale blauer Balken) und dem Gegenstrang (horizontale rote Balken) zu erkennen. In unmittelbarer Nachbarschaft ist eine weitere Region ebenfalls Ursprung für kleine RNAs, die jedoch nach den vorliegenden Annotationen einem Transposon zuzuordnen sind (vertikale gelbe Balken).

Eine BLAST-Suche nach ähnlichen Bereichen im Genom von *Arabidopsis thaliana* ergab keine signifikanten Sequenzhomologien, weshalb davon ausgegangen werden kann, dass diese Region nicht aus einem Duplikationsereignis eines möglichen Zielgens hervorgegangen ist. Eine Suche nach möglichen Zielgenen folgt in Abschnitt 4.5.4.

4.3.4 Kandidat 5

Die fünfte und letzte vorgestellte Sequenz hat ihren Ursprung auf dem Chromosom *V* von *Arabidopsis thaliana*. Die durch RNAFOLD für 25 °C berechnete Sekundärstruktur ist in Abbildung 4.16 dargestellt und bildet eine lange Hairpin-Struktur aus, welche an ihrem Kopfende eine Verzweigungsstruktur ausbildet. Da von pflanzlichen microRNA-Strukturen bekannt ist, dass sie nicht zwangsläufig eine Hairpin-Struktur ohne Verzweigungen ausbilden müssen, kann man davon ausgehen, dass allein auf Basis der Sekundärstruktur die Existenz einer microRNA-kodierenden Region nicht abgelehnt werden muss. Die Sekundärstruktur besitzt eine minimale freie Energie von $-159,71$ kcal/mol, was auf die Länge normiert einen Wert von $-0,560$ ergibt und somit einer thermodynamisch sehr stabilen Struktur entspricht (Zhang *et al.*, 2006). Einige weitere Eigenschaften bezüglich der Sekundärstruktur und Sequenz sind in der folgenden Auflistung gezeigt:

Position auf dem Chromosom	21.385.733 – 21.386.017
Länge	285 Nukleotide
mfe (für 25 °C)	$-159,71$ kcal/mol
GC-Gehalt	32,3%

Die genomische Region des fünften vorgestellten Kandidaten zeigt nach den Daten der ASRP-Expressionsdatenbank zwei distinkte Bereiche exprimierter kleiner RNAs (siehe Abbildung 4.17). Diese sind in der Sekundärstruktur in der 5'- und 3'-Seite des Hairpins

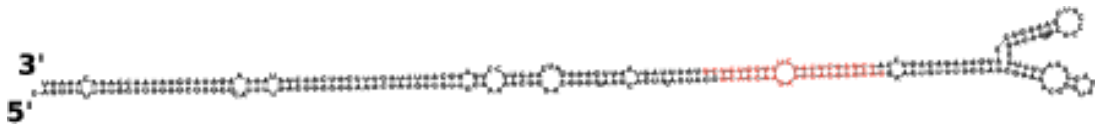

Abbildung 4.16: Kandidat 5 aus intergenischen Sequenzen. Die Darstellung zeigt die durch RNAFOLD für 25 ℃ berechnete Sekundärstruktur des fünften Kandidaten. In rot sind die in der MPSS- und ASRP-Expressionsdatenbank eingetragenen Sequenzen exprimierter kleiner RNAs in dem genomischen Ursprung der gezeigten Hairpin-Struktur vermerkt. Die 3'-seitig in rot vermerkte Sequenz entspricht der in der MPSS-Expressionsdatenbank eingetragenen Signatur (siehe Abbildung 4.18). Die distinkten Bereiche exprimierter RNAs aus der ASRP-Expressionsdatenbank entsprechen hingegen den 5'- und 3'-seitig vermerkten kleinen RNAs in der Sekundärstruktur (siehe Abbildung 4.17).

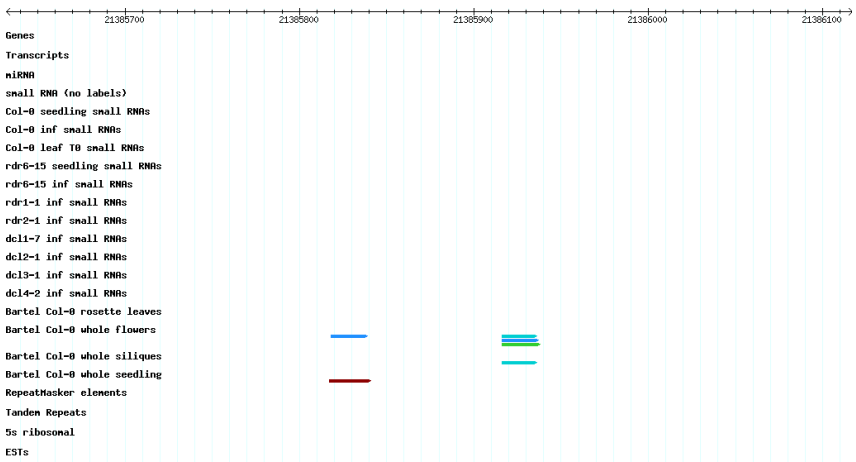

Abbildung 4.17: ASRP-Expressions-Datenbank-Informationen zum fünften Kandidaten. In dieser Abbildung ist die Expression von mehreren kleinen RNAs unterschiedlicher Länge dargestellt. Braune Pfeile stehen dabei für Sequenzen der Länge 25, hellblaue, blaue und grüne Pfeile stehen für Sequenzen der Länge 20, 21 und 22. Die in dieser Region exprimierten RNAs entstammen dabei aus unterschiedlichen Pflanzenteilen von *Arabidopsis thaliana*. Die Mehrheit der RNAs wurde dabei in den Blüten sequenziert (Pfeile in der Spalte *Bartel Col-0 whole flowers*).

---

lokalisiert und stellen einen möglichen microRNA/microRNA*-Duplex dar. Somit liegen auch für diesen Bereich gute Kriterien vor, dass dieser für eine microRNA kodiert. Interessanterweise ist hier zu vermerken, dass die vorliegenden kleinen RNAs nicht aus den RNA-Interferenz-Stoffwechsel-Mutanten sequenziert wurden, sondern fast ausschließlich in den Blüten und Schoten von *Arabidopsis thaliana* weiterer unabhängiger Sequenzieransätze. Der Ort ihrer Expression könnte ein Hinweis auf eine eventuelle Funktion sein, wobei die detaillierte Analyse der möglicherweise regulierten Zielgene in Abschnitt 4.5.5 folgt.

Die unabhängige Bestätigung des vorliegenden Kandidaten über die MPSS-Expressionsdatenbank zeigte ebenfalls die Expression einer Sequenz-Signatur für kleine RNAs im Bereich des fünften Kandidaten, welche dem Genom von *Arabidopsis thaliana* eindeutig zugeordnet werden konnte. Da für den Bereich nur zwei distinkte Bereiche

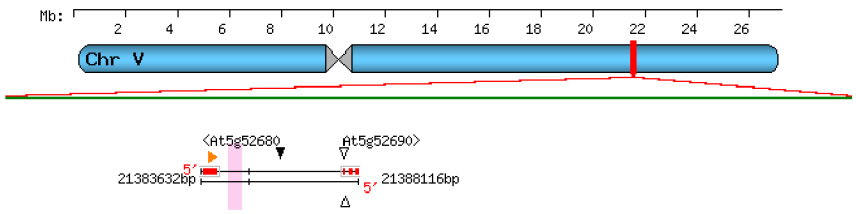

Abbildung 4.18: MPSS-Signaturen aus dem Bereich des fünften Kandidaten. Die Abbildung zeigt die annotierten MPSS-Signaturen aus dem Bereich des fünften Kandidaten. Es ist zu erkennen, dass auf dem Vorwärtsstrang die Signatur einer kleinen RNA ermittelt wurde (erkennbar an dem ausgefüllten schwarzen Dreieck), die dem Genom von *Arabidopsis thaliana* eindeutig zugeordnet werden konnte. Auf der 5'-Seite der Signatur einer kleinen RNA erkennt man die Präsenz eines Retrotransposons an dem vertikalen rosa Balken.

Ursprung kleiner RNAs sind und weiter keine kleinen RNAs exprimiert werden, kann ausgeschlossen werden, dass es sich bei diesem Bereich um ein repetitives Element handelt. Auf dem gleichen Strang befindet sich 5'-seitig des vorgestellten Kandidaten eine Region, die nach den vorliegenden Annotationen einem Retrotransposon entspricht. Auch diese sind oftmals Bereiche, die mit einer hohen Zahl an exprimierten kleinen RNAs einhergehen (ähnlich dem *Inverted Repeat* aus Abbildung 4.15, erkennbar an dem vertikalen gelben Balken).

4.4 Viroid-Vorhersage

Ein Ziel dieser Arbeit stellte die Identifikation von Bereichen aus dem Genom von *Arabidopsis thaliana* dar, die Ursprung kleiner Viroid-spezifischer RNAs sein könnten. Zu diesem Zweck wurde zunächst das Programm YAMP entwickelt.YAMP ist wie in Abschnitt 3.3 gezeigt, in der Lage microRNA-kodierende Sequenzen von nicht-microRNA-kodierenden Sequenzen zu unterscheiden. Um mögliche kleine-RNA-kodierende Bereiche im Genom von *Potato Spindle Tuber Viroid* (*PSTVd*) zu lokalisieren, wurde der *PSTVd*-Stamm AS1 (Matousek *et al.*, 2007) gewählt, welcher die stärksten bekannten Symptome in den Wirtspflanzen induziert. Hierzu wurde die genomische Sequenz des *PSTVd* in vier bzw. acht überlappende Bereiche unterteilt und separat einer Klassifizierung durch YAMP unterzogen. Die vier Regionen umfassten dabei zum einen die obere Hälfte, die untere Hälfte, die linke terminale Region sowie die rechte terminale Region des Viroids (siehe Abbildung 1.7). Die weiteren vier untersuchten Regionen umfassten dieselben Bereiche, wurden jedoch aus dem *PSTVd*-(−)-Strang zusammengesetzt. Als richtig-negativer Trainingsdatensatz wurde der Datensatz 1 aus Abschnitt 3.3.1 gewählt.

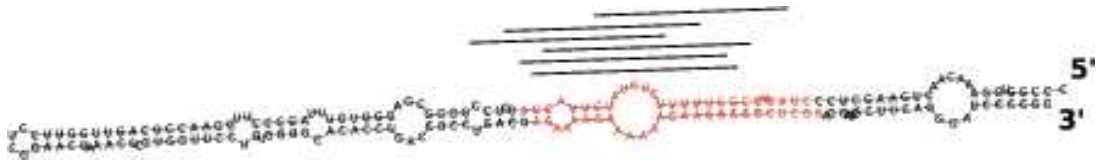

Abbildung 4.19: Terminal-linke Hairpin-Region von *PSTVd*. Die Darstellung zeigt die linke terminale Region des *PSTVd* inklusive der durch YAMP vorhergesagten Region als Ursprung kleiner Viroid-spezifischer RNAs (erkennbar an den Linien oberhalb der Sekundärstruktur). Dabei kann an dieser Stelle keine Aussage über den genauen Strang gemacht werden, da YAMP nur den möglichen RNA-Duplex vorhersagt. Die Sekundärstruktur wurde durch RNAFOLD für eine Temperatur von 25 °C berechnet. In rot ist die VM-Region dargestellt.

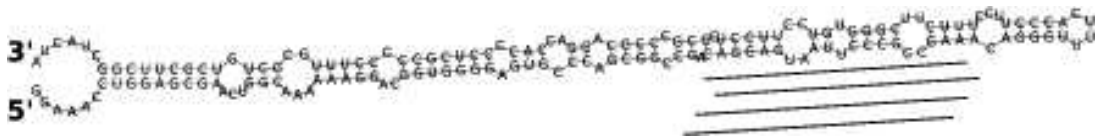

Abbildung 4.20: Terminal-rechte Hairpin-Region von *PSTVd*. Die Darstellung zeigt die rechte terminale Region des *PSTVd* inklusive der durch YAMP vorhergesagten Regionen als Ursprung kleiner RNAs (erkennbar an den Linien unterhalb der Sekundärstruktur). Dabei kann an dieser Stelle keine Aussage über den genauen Strang gemacht werden, da YAMP nur den möglichen RNA-Duplex vorhersagt. Die Sekundärstruktur wurde durch RNAFOLD für eine Temperatur von 25 °C berechnet.

4.4.1 *PSTVd*-(+)-Strang

Die Analyse des *PSTVd*-(+)-Strangs als Ursprung möglicher microRNA-ähnlicher Sequenzen ergab für die obere und untere Hälfte des *PSTVd* keine positiven Klassifizierungen. Dies begründet sich darin, dass diese Regionen für sich betrachtet keine microRNA-Vorläufer-ähnlichen Strukturen ausbilden. Bei der Analyse des Gesamtgenoms bzw. des rechten und linken terminalen Endes von *PSTVd* fielen hingegen zwei Regionen in der stäbchenförmigen Struktur auf, die von YAMP als mögliche microRNA-ähnliche Bereiche identifiziert werden konnten. Die entsprechenden Regionen sind in Abbildung 4.19 und Abbildung 4.20 dargestellt. Die identifizierte Region in der terminalen linken Region (Abbildung 4.19) scheint als microRNA-ähnlich jedoch unwahrscheinlich, da ein großer interner Loop den Bereich unterbricht, was für microRNA-Vorläufer eher untypisch ist.

4.4.2 *PSTVd*-(−)-Strang

Die Suche nach microRNA-Vorläufer-ähnlichen Regionen im *PSTVd*-(−)-Strang ergab ebenfalls einen Bereich, der Ursprung kleiner Viroid-spezifischer RNAs sein könnte. Dieser ist ebenfalls der linken terminalen Region des *PSTVd* zuzuordnen. Interessanterweise handelt es sich bei dieser Region ebenfalls um die Pathogenitäts-modulierende Region (*Virulence-Modulating*-Region, VM-Region). Zudem ist diese Region im *PSTVd*-(−)-Strang nicht von einem großen internen Loop unterbrochen, was die Wahrscheinlichkeit

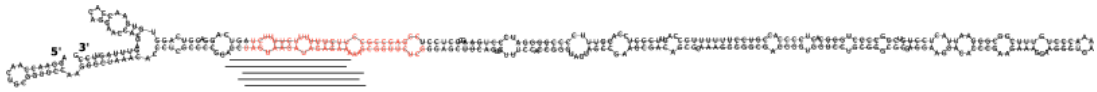

Abbildung 4.21: Sekundärstruktur des *PSTVd*-(−)-Strangs. Die Darstellung zeigt die durch RNAfold für 25 ℃ berechnete Sekundärstruktur des *PSTVd*-(−)-Strangs. Die als microRNA-ähnlich vorhergesagte Region ist, durch Linien unterhalb der Struktur dargestellt, in der VM-Region (rot) lokalisiert. Diese Region entspricht der Region in der terminalen linken Region aus Abbildung 4.19.

erhöht, dass diese Region von der Pflanze fälschlicherweise als microRNA-Vorläufer erkannt wird (siehe Abbildung 4.21). Einführend wurde die Hypothese aufgestellt, dass unterschiedliche Stämme des *PSTVd* durch Nukleotidaustausche in der VM-Region unterschiedlich stark ausgeprägte Symptome hervorrufen und dies auf einen höheren Komplementaritätsgrad zu möglichen Zielgenen zurückzuführen ist. Die Tatsache, dass die VM-Region als Ursprung kleiner RNAs vorhergesagt wurde, untermauert diese Hypothese zusätzlich, da Nukleotidaustausche in eben dieser Region unterschiedlich starke Symptome in den Wirtspflanzen induzieren würden.

4.5 Identifikation der Zielgene

Im vorherigen Abschnitt wurden fünf *de novo*-vorhergesagte microRNA-Kandidaten vorgestellt, bei denen gute Hinweise vorlagen, dass es sich bei ihnen um funktionale microRNAs handelt. Eine weitere Möglichkeit zur Verifikation der vorhergesagten microRNAs umfasst die Vorhersage von möglichen Zielgenen. Jede microRNA greift über Komplementarität zum korrespondierenden Zielgen auf transkriptioneller oder posttranskriptioneller Ebene in die Regulation desselbigen ein und entfaltet über die Komplementarität ihre Funktionalität. Für eine möglichst genaue Beschreibung der vorliegenden Kandidaten wurde neben der Klassifizierung durch YamP und der Validierung durch zwei unabhängige Expressionsdatenbanken (ASRP und MPSS) eine Suche nach möglichen Zielgenen durchgeführt. Da pflanzliche microRNAs über perfekte oder fast-perfekte Komplementarität (Adai *et al.*, 2005) an ihre korrespondierenden Ziel-mRNAs binden, wird die Suche nach den entsprechenden Zielgenen deutlich vereinfacht. Zur Identifikation der entsprechenden Zielgene wurden drei verschiedene, bereits publizierte Programme verwendet. Das erste verwendete Programm war der von Zhang 2005 beschriebene Webservice mit dem dazugehörigen Programm miRU. Es bewertet komplementäre Sequenzen auf Basis eines Bewertungsschemas, welches in Abschnitt 2.2.7 im Detail erläutert wurde. Auf diese Weise konnte eine Grundmenge an möglichen Zielgenen ermittelt werden, die durch das Programm RNAup verifiziert werden sollten. Das Programm RNAup aus dem Vienna-RNA-RNA-Paket berechnet unter Berücksichtigung lokaler Strukturelemente in den langen, Protein-kodierenden mRNA-Sequenzen ein thermodynamisches Bindeprofil einer kleinen RNA-Sequenz und einer langen RNA-Sequenz. Falls durch eine initiale Suche mit miRU keine oder nur unzureichende mögliche Zielgene ermittelt wurden, so wurde das Programm miRanda mit anderen Sequenzdaten als denen, die dem Webservice miRU zur Verfügung stehen, verwendet. miRanda sucht mit Hilfe eines modifizierten Smith-Waterman-Algorithmus für eine gegebene kleine RNA-Sequenz komplementäre Sequenzbereiche in langen RNA-Sequenzen und berechnet die thermodynamische Stabilität der Hybridisierung der kleinen und der langen RNA.

Weiterhin ist bekannt, dass funktionale microRNAs vornehmlich in die Regulation von Entwicklungs- und Wachstums-Prozessen eingreifen, sowie Reaktionen auf veränderte Umwelteinflüsse regulieren (Schauer *et al.*, 2002; Sunkar & Zhu, 2004). Aus diesem Grund wurde das Hauptaugenmerk auf solche möglichen Zielgene gerichtet, welche in die eben genannten Prozesse involviert sein könnten. Die Funktion der entsprechenden Zielgene wurde über die vorliegenden Annotationen ermittelt. Lagen keine eindeutigen Annotationen zu den entsprechenden Zielgenen vor, wurden durch eine BLAST-Suche homologe Protein-kodierende Bereiche in anderen Organismen ermittelt und von denen auf eine Funktion geschlossen.

Zu allen vorgestellten Kandidaten lagen aus den Expressionsdatenbanken die RNA-Sequenzen des möglichen microRNA/microRNA*-Duplex vor. Hier stellt sich die Frage, welche der beiden im Reifungsprozess der microRNAs in den *RNA-induced silencing complex* (RIS-Komplex) integriert wird und die Regulation der Zielgene vermittelt. Khvorova *et al.* (2003) und Schwarz *et al.* (2003) stellten dabei die Asymmetrie-Regel auf, nach

Tabelle 4.1: Mögliche Zielgene des ersten Kandidaten. Die Tabelle zeigt die durch miRU vorhergesagten möglichen Zielgene für den ersten Kandidaten (siehe 4.2.1). Die erste Spalte in der ersten Zeile beschreibt die Orientierung der in der zweiten Spalte gezeigten möglichen microRNA-Sequenz. Die folgenden Zeilen geben die möglichen Zielgene wieder, welche durch miRU ermittelt wurden. Die erste Spalte gibt dabei die *Accession*-Nummer an, unter der das mögliche Zielgen in der Datenbank eingetragen ist, die zweite Spalte die mögliche Bindestelle in der Zielsequenz, die dritte Spalte die Position der Bindestelle und die vierte Spalte den durch miRU berechneten Score. Die besprochenen Kandidaten sind *At4g00760.1* und *At5g15890.1*.

Query (3' - 5')	`ctctactacgcgtttacgcct`	Position der Bindestelle	Score
At4g00760.1	`gagatgatacgcaaatgcgga`	91 - 111	1
At3g19690.1	`aagatgctgcggaaatgtgga`	266 - 286	2.5
At5g15890.1	`gagatgatgagcaaatgtgtg`	542 - 562	3
At1g54280.1	`tagaggatgagcgaatgtgga`	3.539 - 3.559	3
At5g60360.1	`gagatgatcagcgaatgtggt`	1.223 - 1.243	3

der die Sequenz des microRNA/microRNA*-Duplex in den RIS-Komplex integriert wird, welche das thermodynamisch weniger stabile 5'-Ende besitzt. Zu diesem Zweck wurde mit Hilfe des Poland-Algorithmus von Steger (1994) die Sequenz ermittelt, welche nach dem Denaturierungsprofil das weniger stabile Ende besitzt. Konnte auf diese Weise keine eindeutige Aussage über die Stabilität der 5'-Enden getroffen werden, so wurde die Sequenz als reife microRNA-Sequenz gewählt, welche in der MPSS-Expressionsdatenbank eingetragen war. Die Daten des thermodynamischen Denaturierungsprofils durch den Poland-Algorithmus werden im Folgenden nicht gezeigt.

4.5.1 Kandidat 1

Der erste Kandidat aus einem Intron einer Protein-kodierenden Region auf dem Chromosom *I* bildet einen einzigen stabilen Hairpin aus. Die Expressionsdatenbank des ASRP zeigt die Expression kleiner RNAs an zwei distinkten Bereichen im hier diskutierten Hairpin (siehe Abbildung 4.2 und Abbildung 4.3). Die MPSS-Expressionsdatenbank zeigt ebenfalls die Expression kleiner RNAs in diesem Bereich, welche sich zwar auf beiden Strängen befinden, jedoch identisch sind. Ein erstelltes Denaturierungsprofil durch den Poland-Algorithmus erbrachte den Hinweis, dass die 3'-seitige kleine RNA das weniger stabile 5'-Ende besitzt. Eine initiale Identifikation möglicher Zielgene mit dem Webservice miRU ergab für die RNA-Sequenz insgesamt fünf verschiedene mögliche Zielgene, die in Tabelle 4.1 aufgelistet und nach aufsteigendem miRU-Score sortiert sind (niedriger Score bedeutet höherer Grad an Komplementarität).

Von diesen möglichen Zielgenen sind zwei aufgrund ihrer Funktion von besonderem Interesse, da beide an Regulationsprozessen bzw. Entwicklungsprozessen in der Pflanze beteiligt sind. Das erste in der Tabelle 4.1 gezeigte mögliche Zielgen mit der *Accession*-Nummer *At4g00760.1* kodiert für ein Protein, welches nach den vorliegenden Annotationen einem Zwei-Komponenten-Signaltransduktions-Regulationsprotein zuzuordnen ist und regulativ

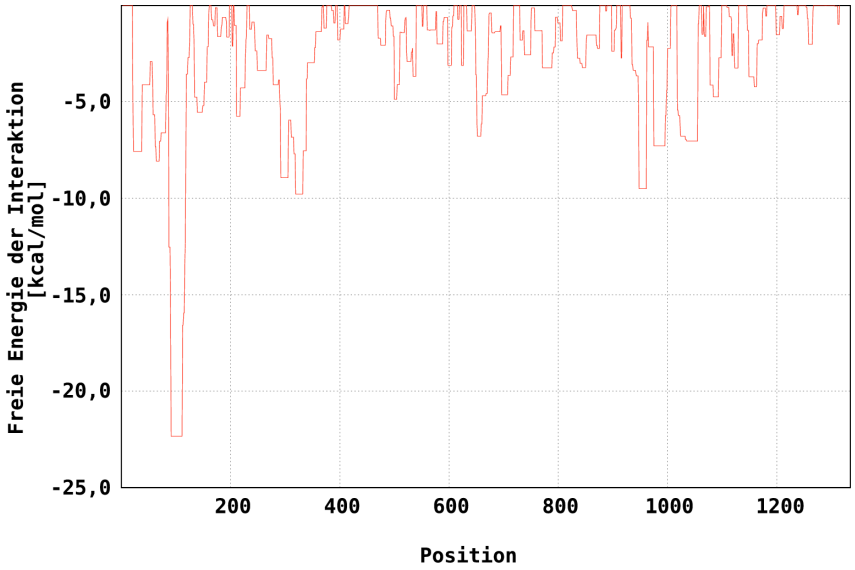

Abbildung 4.22: Minimale freie Energie der Interaktion des ersten Kandidaten. Die Darstellung zeigt die Auftragung der minimalen freien Energie der Hybridisierung des microRNA-Kandidaten und der *coding sequence* (CDS) des Treffers mit minimalem Score (*At4g00760.1*). Die X-Achse beschreibt die Position in der CDS. Die Y-Achse bezeichnet die minimale freie Energie der Hybridisierung der kleinen RNA an die CDS unter Berücksichtigung der benötigten Energie zur Auflösung lokaler Strukturelemente (Einheit in kcal/mol). Die Berechnung wurde durch RNAup ermittelt.

in die Transkription eingreift. Eine BLAST-Suche ergab signifikante Homologien zu anderen Proteinen aus *Arabidopsis thaliana*, welche hauptsächlich in die gleiche funktionale Kategorie einzuordnen sind. Somit liegt für diese microRNA-Sequenz ein interessantes mögliches Zielgen vor. Die Bindestelle befindet sich in der betrachteten mRNA am 5'-Ende in relativer Nähe zum Startkodon und stellt mit nur einem Mismatch eine fast-perfekt komplementäre Bindestelle für die mögliche microRNA-Sequenz dar. Der Mismatch zwischen der möglichen microRNA-Sequenz und der Bindestelle liegt an Position 13 der microRNA-Sequenz und somit 3'-seitig einer möglichen Schnittstelle durch den RIS-Komplex, der üblicherweise zwischen dem zehnten und elften Nukleotid der reifen microRNA schneidet (Elbashir *et al.* (2001), gezeigt in *Drosophila melanogaster*). Die Analyse durch RNAup ergab das in Abbildung 4.22 dargestellte Bindeprofil, welches im Bereich um die von miRU ermittelte Bindestelle (Position 91–111) die minimale freie Energie der Interaktion beider Sequenzen unter Berücksichtigung lokaler Strukturelemente in der mRNA ermittelte.

Das zweite mögliche Zielgen (*At3g19690.1*) ist in *Arabidopsis thaliana* als unbekanntes bzw. hypothetisches Protein annotiert. Dieses Protein ist dennoch als mögliches Zielgen interessant, da durch eine BLAST-Suche eine statistisch signifikante Homologie zu einem Protein aus *Oryza sativa* ermittelt werden konnte, das in die Blattalterung involviert zu sein scheint. Somit wäre auch dieses Zielgen basierend auf seiner Funktion in entwicklungsspezifische Prozesse von großem Interesse.

4.5.2 Kandidat 2

Der zweite Kandidat hat seinen Ursprung in einer intergenischen Region des zweiten Chromosoms (siehe Abschnitt 4.3.1). Die ASRP-Expressionsdatenbank zeigt in dem Bereich der als microRNA-kodierend vorhergesagten Region zwei distinkte Regionen exprimierter RNAs (siehe Abbildung 4.6). Da ein Denaturierungsprofil mit Hilfe des POLAND-Algorithmus keinen Aufschluss über die relative Stabilität der 5'-Enden der beiden exprimierten kleinen RNAs ergab, wurde die Sequenz als reife microRNA-Sequenz gewählt, welche in der MPSS-Expressionsdatenbank durch eine ermittelte Signatur für diesen Bereich eingetragen ist (siehe Abbildung 4.7). Die initiale Identifikation möglicher Zielgene durch MIRU ergab für diese mögliche microRNA insgesamt 17 verschiedene mögliche Zielgene. Die Annotationen zu den möglichen Zielgenen ergaben dabei, dass mindestens sieben der 17 Treffer Protein-Kinasen oder zumindest, über eine BLAST-Suche ermittelt, Homologien zu Protein-Kinasen aufweisen. Hierzu gehört auch der Treffer von MIRU, welcher den minimalsten Score erhielt und somit den besten Kandidaten darstellt (siehe Tabelle 4.2). Eine genauere Analyse der Hybridisierung der möglichen microRNA an die korrespondierende mRNA des ersten Treffers (*At1g53700.1*) ist in Abbildung 4.23 dargestellt. Die von MIRU vorgeschlagene Bindestelle konnte durch RNAUP bestätigt werden, da die Hybridisierung unter Berücksichtigung lokaler Strukturelemente eine, im Vergleich zu anderen Bereichen in der mRNA, relativ hohe thermodynamische Stabilität aufweist, was diese Bindestelle sehr wahrscheinlich macht.

Bei der Sichtung der weiteren Treffer ergab sich ein Hinweis auf die regulatorische Funktion des möglichen microRNA-Kandidaten. Der zweite Treffer (*At2g37680.1*) aus der Tabelle 4.2 kodiert für ein Protein, welches nach den vorliegenden Annotationen eine Komponente der PhytochromA-Signaltransduktionskaskade darstellt. Der dritte Treffer kodiert laut Annotation für ein Protein, welches eine Pektinesterase-Inhibitor-Domäne besitzt. Jedoch zeigte eine BLAST-Suche für dieses mögliche Zielgen eine signifikante Homologie zu einem Photomorphogenese-assoziiertem Protein (*COnstitutive Photomorphogenesis*-Protein, COP1-Protein; Deng *et al.* (1991)). Die Tatsache, dass sich unter den weiteren 15 durch MIRU ermittelten Zielgenen noch weitere Protein-Kinasen bzw. Proteine mit signifikanten Homologien zu anderen Protein-Kinasen befinden, lässt den Schluss zu, dass dieser microRNA-Kandidat möglicherweise die Expression mehrerer Komponenten einer Signaltransduktionskaskade, in diesem Fall die Lichtantwort, reguliert.

4.5.3 Kandidat 3

Der dritte Kandidat hat seinen Ursprung in einer intergenischen Region des dritten Chromosoms von *Arabidopsis thaliana*. Die ASRP-Expressionsdaten zeigen zwei distinkte Bereiche als Ursprung kleiner RNAs im Bereich des vorhergesagten microRNA-Kandidaten (siehe Abbildung 4.9). Diese wurden im Folgenden für die Identifikation möglicher regulierter Zielgenen herangezogen. Als zusätzliche Information lag das Vorliegen einer zweiten Region auf dem Chromosom *V* vor, welche nach den Expressionsdaten der MPSS-

Tabelle 4.2: Mögliche Zielgene des zweiten Kandidaten. Die Tabelle zeigt die durch miRU vorhergesagten möglichen Zielgene für den zweiten Kandidaten 4.3.1. Die erste Spalte in der ersten Zeile beschreibt die Orientierung der in der zweiten Spalte gezeigten möglichen microRNA-Sequenz. Die folgenden fünf Zeilen geben die möglichen Zielgene wieder, welche durch miRU ermittelt wurden. Die erste Spalte gibt dabei die *Accession*-Nummer an, unter der das mögliche Zielgen in der Datenbank eingetragen ist, die zweite Spalte die mögliche Bindestelle in der Zielsequenz, die dritte Spalte die Position der Bindestelle und die vierte Spalte den durch miRU berechneten Score.

Query (3' - 5')	`ccttgtgttcatcgtcatttt`	Position der Bindestelle	Score
At1g53700.1	`tgagcaagagtagcagtaaaa`	1.382 - 1.402	2
At2g37680.1	`tgatcataagtagtagtaaaa`	452 - 472	2
At1g56100.1	`agaacgcgagtagcagcaaaa`	644 - 664	2.5
At3g03720.1	`cgaactcaagtaccagtgaaa`	1.300 - 1.320	2.5
At3g56100.1	`ggaagtcaagtagcagtgaag`	1.417 - 1.437	2.5
At4g23200.1	`ggaacagaagttgcagtgaag`	1.039 - 1.059	2.5
At3g45860.1	`ggagtacaagttgcagtgaag`	1.117 - 1.137	2.5
At2g26620.1	`agaaca-aagtagcagtaaat`	1.034 - 1.053	3
At3g09010.1	`ggcacacaagtagctgtgaaa`	534 - 554	3
At1g18670.1	`ggaagaagagtagtagtaaaa`	161 - 181	3
At1g56140.1	`ggaagagaggtagcagtgaag`	2.224 - 2.244	3
At3g14040.1	`gcaacaaaggtagcagtgaag`	1.237 - 1.257	3
At2g19480.1	`ggaatgtaagcagcagtgaag`	1.267 - 1.287	3
At4g23280.1	`ggagtacaagttgcggtgaag`	1.066 - 1.086	3
At4g23310.1	`ggagtacaagttgcggtgaag`	1.588 - 1.608	3
At5g01310.1	`aggacacaagaggcagtggaa`	2.099 - 2.119	3
At5g58870.1	`agagtactagtagtagtagaa`	300 - 320	3

Datenbank dieselbe kleine RNA exprimiert. Diese kleine RNA wurde im Folgenden für die Suche nach möglichen Zielgenen verwendet, da auch hier keine eindeutige Aussage über die thermodynamische Stabilität der 5'-Enden des möglichen microRNA/microRNA*-Duplex getroffen werden konnte. Die initiale Suche nach Zielgenen erfolgte ebenfalls mit Hilfe von miRU, welches die Sequenzen in Tabelle 4.3 als mögliche Zielgene identifizierte. Besonders auffällig ist der erste Treffer (*At3g27150.1*) mit dem minimalen Score. Die detaillierte Analyse der Bindung der kleinen RNA an die kodierende Region des Treffers *At3g27150.1* ist in Abbildung 4.24 zu sehen. Es ist zu erkennen, dass die Interaktion der kleinen RNA mit der mRNA-Sequenz des Treffers unter Berücksichtigung lokaler Strukturelemente thermodynamisch sehr stabil ist, was auf eine tatsächliche Hybridisierung des Kandidaten an die mRNA schließen lässt. Die Bindestelle ist dabei in relativer Nähe zum 3'-Ende lokalisiert. Die mRNA kodiert im Genom von *Arabidopsis thaliana* für ein Protein, dessen genaue Funktion nicht bekannt ist. Die zugehörigen Annotationen zeigten, dass dieses Protein eine F-Box-Domäne sowie sogenannte *Kelch-Repeats* besitzt. Von F-Box-Domänen ist bekannt, dass ihre Funktion in der Zelle sehr vielfältig sind. F-Box-Domänen vermitteln Protein-Protein-Interaktionen und sind an einer Vielzahl von regula-

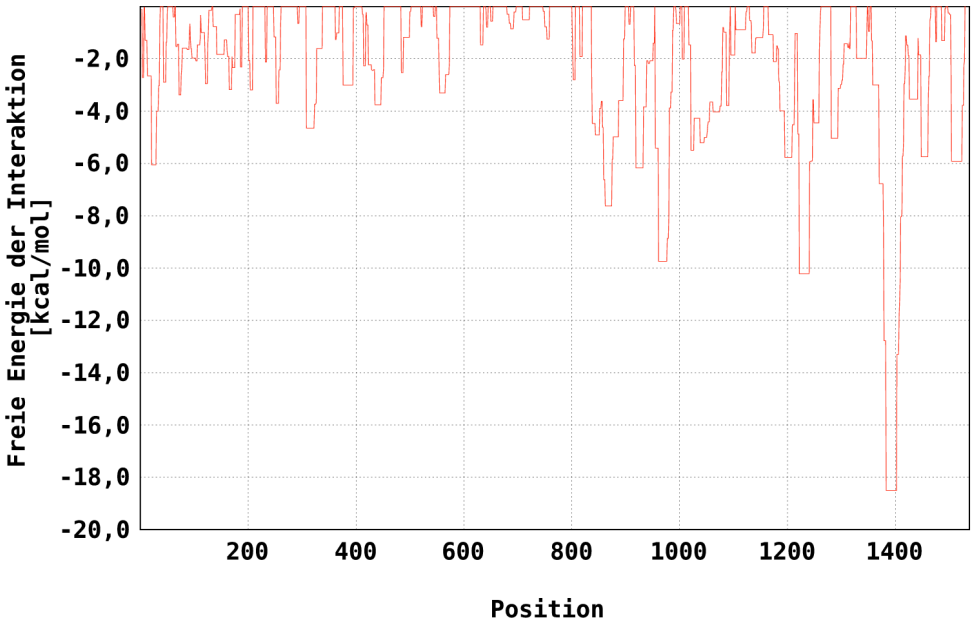

Abbildung 4.23: Minimale freie Energie der Interaktion des zweiten Kandidaten. Die Darstellung zeigt die Auftragung der minimalen freien Energie der Hybridisierung des microRNA-Kandidaten mit der CDS des Treffers mit minimalem Score (*At1g53700.1*). Die X-Achse beschreibt die Position in der *coding sequence* (CDS). Die Y-Achse bezeichnet die minimale freie Energie der Hybridisierung der kleinen RNA an die CDS unter Berücksichtigung der benötigten Energie zur Auflösung lokaler Strukturelemente (Einheit in kcal/mol). Die Berechnung wurde durch RNAup ermittelt.

---

torischen Prozessen beteiligt. Zu diesen gehören unter anderem die Poly-Ubiquitinierung, Transkriptions-Elongation, Zentromer-Bindung und die Hemmung der Translation. Proteine mit *Kelch*-Motiven wurden zunächst in *Drosophila melanogaster* identifiziert und nach der zugehörigen Mutante *Kelch* benannt. Die Funktionen von Proteinen mit einem *Kelch*-Motiv sind sehr unterschiedlich und reichen von einer Beteiligung am Zytoskelett bis zu Galactose-Oxidase-Aktivität. Das Vorliegen einer F-Box-Domäne im identifizierten Zielgen und die beteiligten Funktionen von F-Box-Domänen machen dieses Zielgen zu einem interessanten Kandidaten für die Regulation durch eine microRNA.

4.5.4 Kandidat 4

Der vierte microRNA-Kandidat entstammt einer intergenischen Region des vierten Chromosoms von *Arabidopsis thaliana*. Die ASRP- und MPSS-Expressionsdaten zeigten für die intergenische Region das Vorliegen kleiner exprimierter RNAs (siehe Abbildung 4.14 und Abbildung 4.15). Eine initiale Suche nach möglichen Zielgenen durch Verwendung von miRU erbrachte zunächst keine Treffer, bei denen eine perfekte oder fast-perfekte Komplementarität vorlag. Da die vorhandenen Sequenzdaten nicht mehr aktuell waren, wurde eine weitere initiale Suche durch miRanda über den *Arabidopsis thaliana-Gene-Index* in Version 13 durchgeführt. Bei dieser Suche wurde von miRanda ein Kandidat identifiziert (*Accession*-Nummer *AT1G19415*), der eine fast-perfekte Komplementarität zu dem kor-

Tabelle 4.3: Mögliche Zielgene des dritten Kandidaten. Die Tabelle zeigt die durch miRU vorhergesagten möglichen Zielgene für den dritten Kandidaten. Die erste Spalte in der ersten Zeile beschreibt die Orientierung der in der zweiten Spalte gezeigten möglichen microRNA-Sequenz. Die folgenden Zeilen geben die möglichen Zielgene wieder, welche durch miRU ermittelt wurden. Die erste Spalte gibt dabei die *Accession*-Nummer an, unter der das mögliche Zielgen in der Datenbank eingetragen ist, die zweite Spalte die mögliche Bindestelle in der Zielsequenz, die dritte Spalte die Position der Bindestelle und die vierte Spalte den durch miRU berechneten Score.

Query (3' - 5')	`atttggagtcctacgtctaat`	Position der Bindestelle	Score
At3g27150.1	`gaaacctaaggatgcagatta`	290 - 310	1
At1g07010.1	`tgaacttcaggttgcagatta`	1.205 - 1.225	2
At5g52010.1	`taagcctcaggctgcggattg`	889 - 909	2
At2g19590.1	`aaaacctcag-atgcagattg`	356 - 375	2.5
At1g25550.1	`gaaacctaaggctgcagattc`	595 - 615	3
At5g37440.1	`taaacctaaagctgcagattc`	627 - 647	3
At5g45930.1	`gaagcttcagga-gcagatta`	1.042 - 1.061	3
At2g23370.1	`tggaccttaggatgcagatat`	821 - 841	3
At3g47050.1	`ttagcctcaggatatagattt`	1.452 - 1.472	3
At5g45060.1	`cagatcacaggatgtggatta`	1.413 - 1.433	3

respondierenden Zielregion aufwies (siehe Tabelle 4.4). Ein weiteres interessantes Ergebnis stellte das Vorliegen einer zweiten möglichen Zielregion innerhalb dergleichen mRNA dar. Eine Zielregion ist in unmittelbarer Nähe des Startkodons, wohingegen die zweite mögliche Bindestelle mittig der mRNA lokalisiert ist. Die möglichen Zielregionen wurde detaillierter analysiert und beide Bindestellen konnten durch eine Analyse mit RNAup verifiziert werden. Diese weisen zwar im Vergleich zu den anderen bisher vorgestellten Kandidaten eine relativ geringe thermodynamische Stabilität auf (siehe Abbildung 4.25), was jedoch eine Funktion nicht zwingend ausschließen muss. Zudem weist die Bindestelle in relativer Nähe zum Startkodon eine höhere thermodynamische Stabilität auf als die zweite Bindestelle, was durch lokale Strukturelemente bedingt sein könnte.

Die identifizierte mRNA kodiert nach vorliegenden Annotationen für ein *LTR-Retrotransposon-like-Protein*. Dies ist ein interessanter Hinweis auf eine mögliche Funktion des vorliegenden microRNA-Kandidaten. Der microRNA-Kandidat könnte in die Stilllegung von Retrotransposons im Genom eingreifen sein und deren Verbreitung bzw. Umordnung im Genom regulieren. Weiterhin konnte den Annotationen entnommen werden, dass zwei funktionelle Domänen in der Aminosäure-Sequenz auftreten, welche als *DC1* bezeichnet werden. Diese sind in der Lage, Zn^{2+}-Ionen zu binden. Jedoch ergab die BLAST-Suche zusätzlich eine signifikante Ähnlichkeit zu einem Protein, welches – wie ein mögliches Zielgen des zweiten Kandidaten – in die Lichtantwort involviert ist und für ein *UV-B light insensitive*-Protein (ULI3) kodiert. Somit stehen zwei mögliche Regulationsmechanismen im Raum, welche allein über die Bioinformatik nicht geklärt werden können. Es bleibt jedoch festzuhalten, dass der vorliegende Kandidat alle theoretischen

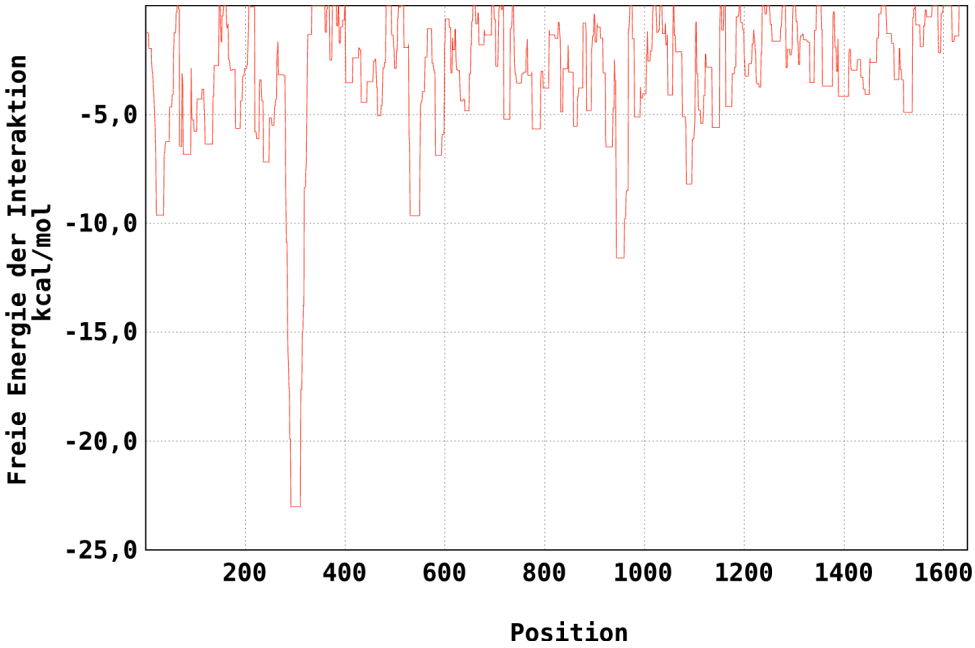

Abbildung 4.24: Minimale freie Energie der Interaktion des dritten Kandidaten. Die Darstellung zeigt die Auftragung der minimalen freien Energie der Hybridisierung der kleinen RNA an die *coding sequence* (CDS) des Treffers mit der *Accession*-Nummer *At3g27150.1* mit minimalem Score. Die X-Achse stellt die Position in der CDS dar. Die Y-Achse beschreibt die minimale freie Energie der Hybridisierung der kleinen RNA an die CDS unter Berücksichtigung der benötigten Energie zur Auflösung lokaler Strukturelemente (Einheit in kcal/mol). Die Berechnung wurde mit Hilfe von RNAup ermittelt.

Voraussetzungen erfüllt, als funktionelle microRNA in der Pflanze zu fungieren und somit ein interessantes Ziel für weitere *in vitro*-Untersuchungen darstellt.

4.5.5 Kandidat 5

Der letzte im Rahmen dieser Arbeit vorgestellte microRNA-Kandidat entstammt einer intergenischen Region des fünften Chromosoms von *Arabidopsis thaliana*, die nach den Expressionsdaten der ASRP- und MPSS-Datenbank kleine RNAs in zwei distinkten Bereichen (siehe Abbildung 4.17 und Abbildung 4.18) exprimiert. Als reife microRNA wurde die Sequenz aus den MPSS-Expressionsdaten gewählt. Die initiale Suche nach möglichen Zielgenen durch miRU ergab die in Tabelle 4.5 dargestellten Treffer, wovon der erste Treffer durch seine fast-perfekte Komplementarität hervorsticht (*Accession*-Nummer *At5g16100*). Er beinhaltet lediglich einen Mismatch im 3'-seitigen Bereich der möglichen microRNA. Eine detailliertere Analyse der Hybridisierung mit Hilfe von RNAup für die mRNA des ersten Treffers mit der möglichen microRNA-Sequenz ergab, dass die von miRU vorhergesagte Bindestelle unter Berücksichtigung lokaler Strukturelemente thermodynamisch sehr stabil ist (siehe Abbildung 4.26). Weiterhin ist der Abbildung zu entnehmen, dass zwei weitere Regionen eine thermodynamische Stabilität aufweisen, welche

Tabelle 4.4: Mögliche Zielgene des vierten Kandidaten. Die Darstellung zeigt zwei durch miRanda ermittelte Bindestellen. Die Suche wurde über den *Arabidopsis thaliana-Gene-Index* in Version 13 durchgeführt. Die Suche ermittelte zwei Bindestellen in einer kodierenden Region mit der *Accession*-Nummer *AT1G19415*. Die Tabelle zeigt in der obersten Spalte die kleine RNA-Sequenz in 3'-5'-Orientierung. Die weiteren Spalten zeigen die *Accession*-Nummer, das Alignment mit dem microRNA-Kandidaten in 3'-5'-Orientierung und der CDS in 5'-3'-Orientierung. Die senkrechten Striche (|) beschreiben kanonische Watson-Crick-Basenpaare, die Doppelpunkte (:) Wobble-Basenpaare. Leere Stellen beschreiben hingegen Mismatche. Die letzte Spalte gibt die Position der möglichen Bindestelle in der CDS an.

Query (3' - 5')	gaaacgataattaaaccatatt	Position der Bindestelle
AT1G19415	3' GAAACGATAATTAAACCATAuu 5' ||||||||| :||||||||| 5' CTTTGCTATCGATTTGGTATgc 3'	27 – 48 −24, 04 kcal/mol
AT1G19415	3' GAAACGATAATTAAACCATAuu 5' ||||||||| :||||||||| 5' CTTTGCTATCGATTTGGTATgc 3'	1.521 – 1.542 −24, 04 kcal/mol

Tabelle 4.5: Mögliche Zielgene des fünften Kandidaten. Die Tabelle zeigt die durch miRU vorhergesagten möglichen Zielgene für den fünften Kandidaten 4.3.4. Die erste Spalte in der ersten Zeile beschreibt die Orientierung der in der zweiten Spalte gezeigten möglichen microRNA-Sequenz. Die folgenden Zeilen geben die möglichen Zielgene wieder, welche durch miRU ermittelt wurden. Die erste Spalte gibt dabei die *Accession*-Nummer an, unter der das mögliche Zielgen in der Datenbank eingetragen ist, die zweite Spalte die mögliche Bindestelle, die dritte Spalte die Position der Bindestelle und die vierte Spalte den durch miRU berechneten Score.

Query (3' - 5')	gctgtggtttctactcaagct	Position der Bindestelle	Score
At5g16100.1	cgactccaaagatgagttcga	252 - 272	1
At3g18140.1	ggacatcaaagatgggtttgg	969 - 989	2
At3g49190.1	caaaaccaaagatgggttcgg	235 - 255	2.5
At3g28510.1	ggaaattgaagatgagttcga	453 - 473	2.5
At1g35940.1	cgagatcaaagattggttcga	2.931 - 2.951	3
At5g55840.1	ctacaccaaaggagagtttga	3.476 - 3.496	3
At2g23890.1	tgacgccaaacttgagtttga	796 - 816	3
At2g28290.1	aggcaccaaagctaagtttga	5.565 - 5.585	3
At2g28290.2	aggcaccaaagctaagtttga	5.565 - 5.585	3
At5g06110.1	tgactctacagatgagtttga	534 - 554	3
At3g51800.1	agttatcagagatgagttcgg	91 - 111	3
At3g51800.2	agttatcagagatgagttcgg	91 - 111	3
At3g62720.1	cgatgccaaggacgagtttgg	1.520 - 1.540	3

der Stabilität des vierten Kandidaten an Position 1.500 entspricht. Dies könnte darauf hindeuten, dass hier ebenfalls mehrere Bindestellen vorliegen. Auf eine genauere Betrachtung der entsprechenden Sequenzen wurde an dieser Stelle verzichtet.

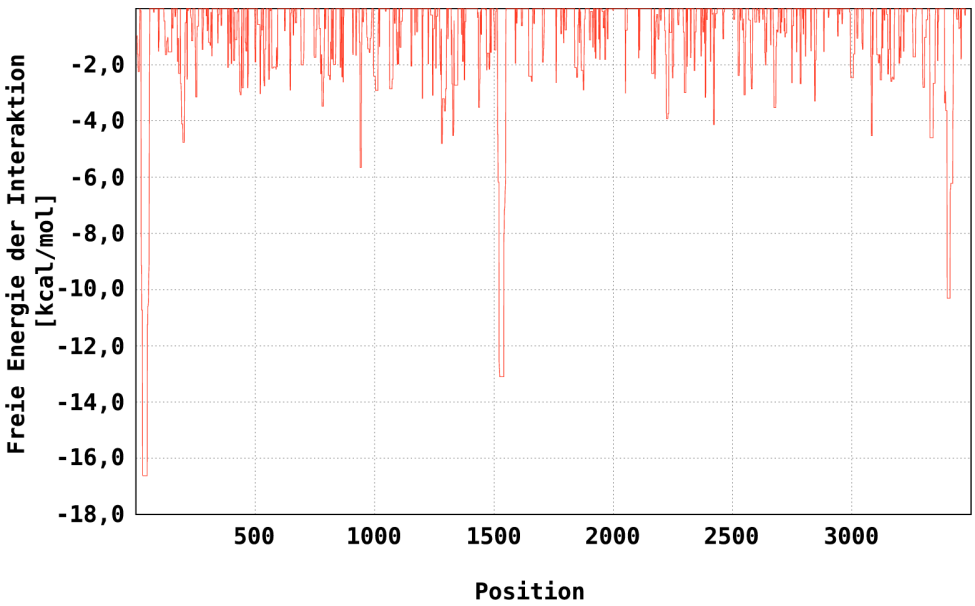

Abbildung 4.25: Minimale freie Energie der Interaktion des vierten Kandidaten. Die Darstellung zeigt die Auftragung der minimalen freien Energie der Hybridisierung der kleinen RNA an die *coding sequence* (CDS) des Treffers *At1g19415.1* mit minimalem Score. Die X-Achse stellt die Position in der CDS dar. Die Y-Achse beschreibt die minimale freie Energie der Hybridisierung der kleinen RNA an die CDS unter Berücksichtigung der benötigten Energie zur Auflösung lokaler Strukturelemente (Einheit in kcal/mol). Die Berechnung wurde durch RNAup ermittelt. Man erkennt beide von miRanda vorhergesagten Bindestellen in der vorliegenden Grafik an den Spitzen um die Position 30 und 1.520 herum.

Das mögliche Zielgen kodiert dabei für ein Protein, welches nach den vorliegenden Annotation ein hypothetisches Protein ist. Eine BLAST-Suche ergab Ähnlichkeiten zu einem Protein aus *Oryza sativa*, welches nach den dort vorliegenden Annotationen eine sogenannte RWP-RK-Domäne besitzt. Diese RWP-RK-Domäne ist ein hoch konserviertes Motiv und wurde initial in Algen und Pflanzen in einem Protein identifiziert, welches in die Stickstoff-kontrollierte Entwicklung involviert ist (Schauser *et al.*, 1999). Die Identifikation des möglichen Zielgens, welches in die Entwicklung involviert zu sein scheint, stellt einen guten Hinweis für das Vorliegen eines realen microRNA-Vorläufers dar. Auch dieser Kandidat weist neben intrinsischen ebenso überzeugende extrinsische Eigenschaften auf, um die vorliegende genomische Region als microRNA-kodierend zu bezeichnen.

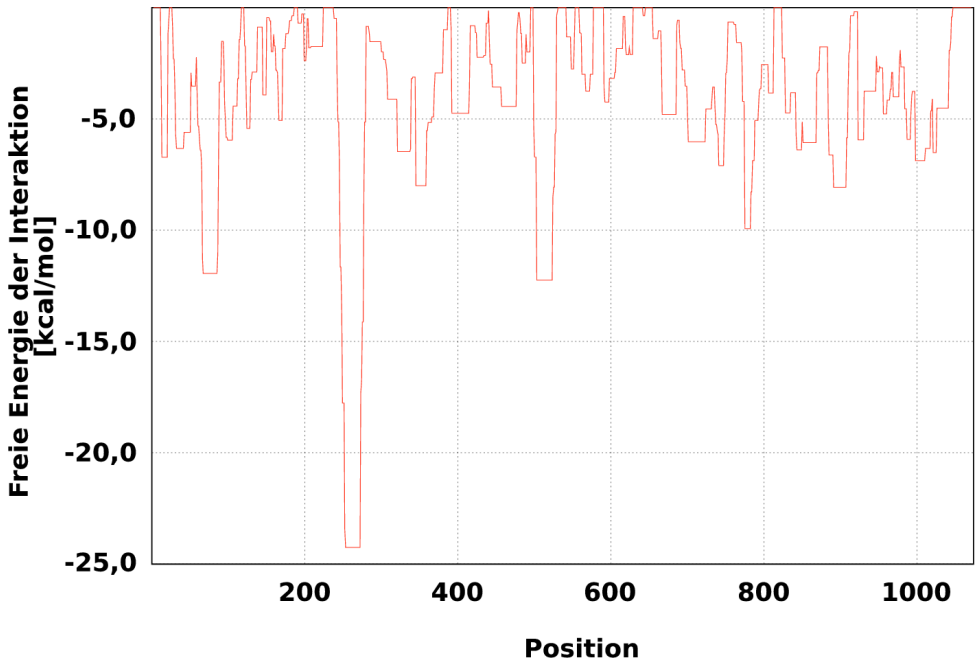

Abbildung 4.26: Minimale freie Energie der Interaktion des fünften Kandidaten. Die Darstellung zeigt die Auftragung der minimalen freien Energie der Hybridisierung der kleinen RNA an die *coding sequence* (CDS) des Treffers *At5g16100.1* mit minimalem Score. Die X-Achse stellt die Position in der CDS dar. Die Y-Achse beschreibt die minimale freie Energie der Hybridisierung der kleinen RNA an die CDS unter Berücksichtigung der benötigten Energie zur Auflösung lokaler Strukturelemente (Einheit in kcal/mol).

5

Diskussion

In diesem Kapitel werden die in den vorangegangenen Abschnitten erzielten Analysen und Ergebnisse behandelt. Zunächst werden dabei das im Rahmen dieser Arbeit entwickelte Programm Yet Another miRNA Predictor (YamP), seine Funktionalität und Grenzen diskutiert (siehe Abschnitt 5.1). Im Speziellen wird dabei vornehmlich auf den Einfluss der zugrundeliegenden Trainingssequenzen (Abschnitt 5.1.1), die Effizienz der implementierten Filter-Methoden (Abschnitt 5.1.2) sowie die abschließende Validierung zur Testung der Funktionalität eingegangen (Abschnitt 5.1.3).

Die Entwicklung eines Ansatzes zur microRNA-Vorhersage und die anschließende Validierung der Leistungsfähigkeit ergaben den logischen Schritt, den entwickelten Ansatz auf genomische Sequenzdaten zur Identifikation neuer microRNA-kodierender Bereiche anzuwenden. Die Diskussion der durch YamP erzielten Ergebnisse findet in Abschnitt 5.2 statt.

Ein Ziel dieser Arbeit bildete die Identifikation möglicher Viroid-spezifischer RNA-exprimierender Bereiche im Genom von *PSTVd*, das in Abschnitt 5.2.3 ausführlich diskutiert wird.

5.1 YamP

Das Programm YamP stellt zusammenfassend betrachtet eine Methode zur microRNA-Vorhersage dar, die microRNA- und nicht-microRNA-Sequenzen mit hoher Effizienz in Bezug auf Sensitivität und Spezifität korrekt klassifizieren kann. Dies konnte insbesondere an den ausführlich durchgeführten Validierungsschritten in Abschnitt 3.3 gezeigt werden. Allerdings zeigte YamP einige Schwächen bei der korrekten Klassifizierung von repetitiven Elementen im Genom von *Arabidopsis thaliana*. Strukturvorhersageprogramme wie RNAfold neigen zur Vorhersage von thermodynamisch stabilen Hairpin-Strukturen für solche Elemente aufgrund ihrer Repetitivität. Repetitive Elemente und insbesondere

Transposons oder Retrotransposons zeichnen sich dadurch aus, dass sie Ursprung einer Vielzahl von kleinen RNAs unterschiedlicher Größenklassen sind (siehe Abbildung 5.1) und demnach ebenfalls durch *Dicer-like proteins* nach ihrer Transkription prozessiert werden. Demnach besitzen diese Elemente einige intrinsische Eigenschaften, die denen der microRNAs sehr ähnlich sind. Diese Regionen aus den intergenischen Bereichen können durch Anwendung zusätzlicher Programme wie z. B. RepeatMasker erkannt und maskiert werden. Diese offensichtliche Schwäche des Programms stellt kein Hindernis dar, neue microRNA-kodierende Bereiche im Genom von *Arabidopsis thaliana* korrekt zu identifizieren.

Potentiell besitzt jede Sequenzregion im Genom von *Arabidopsis thaliana* die Möglichkeit, bei der Berechnung der Sekundärstruktur durch Programme wie z. B. RNAfold mehr oder weniger stabile Sekundärstrukturen auszubilden, deren intrinsische Eigenschaften jedoch stark von denen der microRNAs divergieren. Zu diesem Zweck wurden in YamP einige Filter-Methoden implementiert, welche diese Bereiche schon im Vorhinein ausschließen können. Bei der Anwendung von YamP auf die intergenischen und intronischen Sequenzen konnte bereits im Vorfeld eine deutliche Reduzierung des Suchraums durch diese Filter erreicht werden, was die Anzahl der wahrscheinlich falsch-positiven Treffer deutlich minimierte.

Da es sich bei *Arabidopsis thaliana* nicht um einen natürlichen Wirtsorganismus des *PSTVd* handelt, musste die Funktionalität von YamP die simple Übertragbarkeit auf andere Organismen umfassen. Zwar ist bekannt, dass einige Vertreter der Viroide in *Arabidopsis thaliana* repliziert und auch prozessiert werden (Daròs & Flores, 2004; Matousek *et al.*, 2004), jedoch fehlen noch Hinweise, ob kleine Viroid-spezifische RNAs auch in *Arabidopsis thaliana* akkumulieren. Sollte eine natürliche Wirtspflanze die Viroide und im Speziellen das *PSTVd* fälschlicherweise als microRNA-Vorläufer erkennen, so muss das Viroid Sequenz- und oder Struktur-Charakteristiken aufweisen, welche die Pflanze veranlassen, diesen fälschlicherweise zu prozessieren. Bei dem microRNA-vermittelten Regulationsmechanismus handelt es sich um einen stark konservierten Mechanismus, der wahrscheinlich schon in einem letzten gemeinsamen Vorfahr von Pflanze und Tier vorlag (zur Übersicht Hutvagner & Simard, 2008). Grundsätzlich liegen auch für *Arabidopsis thaliana* einige Hinweise vor, dass das *PSTVd* hier fälschlicherweise als microRNA-Vorläufer erkannt wird. Hierzu gehört seine Lokalisation im Nukleus, dem Ort der Prozessierung der microRNA-Vorläufer, sowie die prinzipielle Fähigkeit, auch in *Arabidopsis thaliana* repliziert und prozessiert zu werden (Daròs & Flores, 2004; Matousek *et al.*, 2004).

Dies und die vorliegenden Ergebnisse, basierend auf *Arabidopsis thaliana*, reichen jedoch nicht aus, um die Hinweise für eine fälschliche Prozessierung des *PSTVd* in seinen Wirtsorganismen zu untermauern. Die Hinweise ließen sich bei Kenntnis von microRNA-Vorläufern aus *Lycopersicon esculentum* oder *Solanum tuberosum* verstärken. Daher muss die Funktionalität von YamP die prinzipielle Anwendbarkeit auf andere pflanzliche Organismen umfassen. Ein erster Versuch, das Programm mit den statistischen Modellen aus *Arabidopsis thaliana* (siehe Abschnitt 3.3.1) und den empirisch ermittelten Schwellenwerten für die implementierten Filter-Methoden auf *Populus trichocarpa* anzuwenden, zeigte dabei bereits eine grundsätzliche Anwendbarkeit mit hoher Effizienz (Daten nicht

gezeigt). Zwar lag die Effizienz noch unterhalb derer, die bei den Validierungsschritten für *Arabidopsis thaliana* erzielt werden konnte, jedoch könnte durch eine genauere Analyse des Diskriminierungsverhaltens bei unterschiedlich gewählten Trainingsdatensätzen sowie einer genaueren Einstellung der Schwellenwerte für die Filter-Methoden insgesamt eine deutlich höhere Effizienz von YamP für weitere Pflanzengenome erzielt werden.

5.1.1 Einfluss der Trainingssequenzen

YamP benötigt für seine Funktionalität ein vorgeschaltetes Training, das auf microRNA- und nicht-microRNA-Sequenzen beruht und anhand dessen ein statistisches Modell erstellt wird. Unterschiedliche Kombinationen der in Abschnitt 3.1.2 verwendeten richtig-negativen Sequenzen zeigten dabei einen unterschiedlichen Einfluss auf die Diskriminierungseffizienz des Programms, wobei aufgrund der Ergebnisse die beiden vorgestellten Trainingsdatensätze bezüglich Sensitivität und Spezifität als die beiden besten Trainingsdatensätze von insgesamt sieben identifiziert werden konnten. Während der Sequenzdatensatz 1 die höchste Sensitivität zeigte, konnte mit dem Sequenzdatensatz 2 eine Kompilation erstellt werden, die bezüglich der Spezifität sehr gute Resultate erzielte. Dies ging jedoch jeweils mit Einbußen in der Spezifität und Sensitivität einher. Beide Trainingsdatensätze weisen eine globale Effizienz auf, die in der Summe betrachtet als sehr gut anzusehen ist, jedoch ist der Anspruch für eine Anwendung auf genomische Sequenzdaten als sehr hoch anzusetzen. Für eine Anwendung auf genomische Sequenzdaten war dabei zu erwarten, dass die Menge der neuen microRNA-kodierenden Bereiche im Vergleich zu den nicht-microRNA-Sequenzen nur eine geringe Anzahl ausmachen würde. Somit würden geringe Unterschiede in der Sensitivität nur in kleinen Unterschieden in der Anzahl falsch-negativer Klassifizierungen resultieren. Betrachtet man jedoch die erwartete große Anzahl von nicht-microRNA-kodierenden Bereichen, so würde ein geringer Unterschied in der Spezifität in einer deutlich höheren Anzahl falsch-positiver Klassifizierungen resultieren. Deshalb wurde der richtig-negative Trainingsdatensatz 2 für die Anwendung auf genomische Sequenzdaten verwendet, da eben dieser die Menge der insgesamt falsch-positiven Klassifizierungen minimierte.

Interessanterweise zeigte auch der Trainingsdatensatz mit den artfremden pseudo-Hairpin-Sequenzen aus cDNA-Sequenzen von *Homo sapiens* (Ng & Mishra, 2007) eine Effizienz, die nahe an die Effizienz der Trainingsdatensätze 1 und 2 mit arteigenen Sequenzen aus *Arabidopsis thaliana* herankommt. Weiterhin wurde ein weiterer Trainingsdatensatz mit arteigenen pseudo-Hairpin-Sequenzen aus *Arabidopsis thaliana* erstellt, der die Effizienz von YamP im Vergleich zu den eben erwähnten Trainingsdatensätzen zwar deutlich herabsenkte, jedoch immer noch eine gute Effizienz aufwies. Somit scheint kein direkter Zusammenhang zwischen der Wahl der richtig-negativen Sequenzen und der Effizienz herstellbar zu sein, jedoch scheint die Wahl von zu microRNAs möglichst divergente Sequenzen die Effizienz von YamP zu steigern.

Da an dieser Stelle bereits eine Effizienz bezüglich des Diskriminierungsverhaltens erreicht wurde, welche für einen genomischen Ansatz ausreichend hoch war, wurde im Folgenden

darauf verzichtet, eine genauere Analyse durchzuführen, welchen Einfluss die Anzahl der richtig-negativen Sequenzen auf das Diskriminierungsverhalten besitzt.

Als Resümee lässt sich festhalten, dass ein Patentrezept zur Erstellung eines guten Trainingsdatensatzes die Auswahl möglichst divergenter Sequenzen hinsichtlich der Organismus-spezifischen microRNAs zu sein scheint. Ein empirischer Beweis dafür steht jedoch noch aus. Die im Rahmen dieser Arbeit erstellten Trainingsdatensätze bildeten eine befriedigende Auswahl an richtig-negativen Sequenzen, was für den genomischen Ansatz möglichst zuverlässige Resultate erwarten ließ.

5.1.2 Effizienz der Filter-Methoden

Es konnte in der Literatur gezeigt werden, dass microRNAs statistisch signifikante Unterschiede bezüglich einiger charakteristischer Eigenschaften im Vergleich zu anderen RNA-Familien wie tRNAs, rRNAs und mRNAs aufweisen (Zhang *et al.*, 2006). Dabei umfasst die Mehrzahl der Eigenschaften die Sequenzkomposition, energetische Charakteristika sowie Eigenschaften, welche die globale Architektur der microRNA-Vorläufer-Strukturen beschreiben (Ritchie *et al.*, 2007).

Diese, der Literatur entnommenen, intrinsischen Eigenschaften der microRNA-Vorläufer-Strukturen und -Sequenzen stellten die Basis für die Implementierung der in Abschnitt 3.2.2 vorgestellten Filter-Methoden dar. Die ausführlich durchgeführten Untersuchungen hinsichtlich der Effizienz der implementierten Filter-Schritte zeigte teilweise sehr gute Resultate, welche sich durch die *Receiver-Operator-Characteristics*-Analyse (ROC-Analyse) und Abschätzung der Fläche unterhalb der Kurve (*Area under the Curve*AUC) besonders hervorheben ließen. Erwartungsgemäß passierte ein Großteil der microRNA-Sequenzen alle Filter-Schritte, was direkt darauf hindeutet, dass die entwickelten Filter-Methoden eine sehr gute Effizienz bei der Vorauswahl der zu klassifizierenden Sequenzen aufweisen. Auch bei der Anwendung von YamP auf die intergenischen und intronischen Sequenzen aus *Arabidopsis thaliana* zeigte sich, dass die vorgeschalteten Filter-Schritte bereits viele der tatsächlichen nicht-microRNA-kodierenden Bereiche aus dem Suchraum entfernten und diese somit garnicht in die finale Klassifizierung eingingen.

Eine der implementierten Filter-Methoden zeichnete sich dabei besonders aus. Der *Window-Slide*-Filter zeigte die beste Effizienz hinsichtlich der Klassifizierung von microRNA- und nicht-microRNA-Sequenzen, was durch eine ROC-Analyse durch Abschätzung der AUC gezeigt werden konnte (siehe Abbildung 3.5). Dieser Filter bezieht zwei charakteristische Eigenschaften der microRNA-Vorläufer-Strukturen mit ein, welche die globale Architektur der microRNA-Vorläufer-Strukturen betreffen. Zum einen wird dabei die Anzahl der Basenpaare maximiert und zum anderen die Anzahl und Größe interner Loops und Bulge-Loops minimiert. Die microRNA-Vorläufer-Strukturen unterscheiden sich hinsichtlich ihrer globalen Architektur statistisch signifikant von anderen RNA-Familien, was ein generelles Charakteristikum und eine wichtige Eigenschaft der microRNAs darzustellen scheint. Eine Struktur mit einer großen Anzahl von Loops und/oder großen Loops, die der Struktur eine asymmetrische Form verleihen, scheint die

korrekte Prozessierung der microRNA-Vorläufer-Strukturen zu beeinträchtigen (Ritchie *et al.*, 2007).

Generell lässt sich sagen, dass die vorliegenden Filter bereits eine sehr gute Effizienz hinsichtlich der Klassifizierung aufweisen. Weitergehende Analysen, experimentelle Daten und Überlegungen zu weiteren charakteristischen Eigenschaften von microRNA-Vorläufer-Strukturen und anderen ncRNA-Familien könnten zusätzliche signifikante Unterschiede aufdecken, die Gegenstand einer Verbesserung von YamP darstellen könnten.

5.1.3 Validierung von YamP

Wie in Abschnitt 5.1.1 diskutiert, besitzt bereits die Wahl des Trainingsdatensatzes einen großen Einfluss auf die Diskriminierungseffizienz von YamP. Die ausführliche Validierung des Programms YamP umfasste zunächst die Anwendung der erstellten statistischen Modelle auf die zugrundeliegenden Trainingsdaten, wobei sich erwartungsgemäß zeigte, dass YamP die Trainingsdaten sehr gut diskriminieren konnte (siehe Tabelle 3.5). Dies allein reicht allerdings für eine zuverlässige Anwendung auf genomische Sequenzdaten von *Arabidopsis thaliana* nicht aus. Ein Programm zur Vorhersage von microRNA-kodierenden Bereichen muss zusätzlich in der Lage sein, unbekannte Sequenzen korrekt zu klassifizieren. Zu diesem Zweck wurden mehrere Tests vorgenommen, welche eben diese Funktionalität sicherstellten.

Die Erstellung des statistischen Modells umfasste die statistische Erhebung umgebungsabhängiger und struktureller Zustände der Nukleotide und Di-Nukleotide in den microRNA- und nicht-microRNA-Sequenzen (siehe Abschnitt 3.2.3). Eine weitere Bestätigung der Funktionalität beruhte auf der Anwendung von Sequenzen, die über eine zufällige Neusortierung der Nukleotidreihenfolge der microRNA-Sequenzen unter spezifischen Bedingungen erstellt wurden. Diese Bedingungen umfassten neben der Beibehaltung der Mono- und Di-Nukleotid-Frequenz auch die Beibehaltung der Mono-Nukleotid-Frequenz in einer Fenstergröße von 20 Positionen, also eben der Größe, welche für reife microRNAs typisch ist (siehe Abschnitt 3.3.3). Somit konnte sichergestellt werden, dass die neu erstellten Sequenzen hinsichtlich der Nukleotidzusammensetzung absolut identisch waren. Die Art der statistischen Erhebung der beiden Zustände hätte das Problem aufwerfen können, dass derartig erstellte Sequenzen für falsch-positive Klassifizierungen prädestiniert sein könnten. Eine Anwendung von YamP auf diese Sequenzen zeigte jedoch genau das gegenteilige Verhalten. Nahezu alle Sequenzen konnten von YamP als nicht-microRNA-Sequenzen korrekt eingeteilt werden, was sich in einer Spezifität von über 99% ausdrückte. Somit konnte an dieser Stelle gezeigt werden, dass YamP auch Sequenzen mit gleicher Mono-Nukleotid- und Di-Nukleotid-Zusammensetzung mit ausreichend hoher Effizienz diskriminieren kann.

Ein weiteres, in der Bioinformatik standardmäßig angewandtes Verfahren zur Validierung von Programmen stellt das sogenannte Bootstrapping-Verfahren dar (siehe Abschnitt 3.3.2). Das Bootstrapping umfasst im betrachteten Fall zunächst das Auslassen einer definierten Anzahl von Sequenzen aus der Grundmenge der zur Verfügung

stehenden microRNA-Sequenzen. Ein anschließendes Training fand mit dem so reduzierten microRNA-Sequenzdatensatz und unterschiedlichen richtig-negativen Trainingsdatensätzen statt. Es folgte die Anwendung auf die Grundmenge der insgesamt zur Verfügung stehenden microRNA-Sequenzen, dessen Ergebnisse in Abschnitt 3.3.2 dargestellt sind. Die zugrundeliegenden Szenarien umfassten als richtig-negative Trainingsdatensätze die bereits erwähnten Datensätze 1 und 2, welche bezüglich ihrer Sensitivität und Spezifität die beste Effizienz der insgesamt verwendeten Trainingsdatensätze zeigten. Die richtig-positiven Trainingsdatensätze mit den Sequenzen der unterschiedlichen Versionen der miRBase stellten die anderen beiden Eckpunkte der analysierten Testläufe dar. Insgesamt wurden also vier unterschiedliche Testläufe ausführlich getestet. Pro Durchlauf wurden dabei jeweils zehn Sequenzen aus der Menge der richtig-positiven Sequenzen ausgelassen.

Das Diagramm in Abbildung 3.9 spiegelt die Ergebnisse des ersten untersuchten Testlaufs wieder. Es ist deutlich zu erkennen, dass die untersuchten statistischen Kenngrößen Sensitivität (1.1), Spezifität (1.2), Relevanz (1.4), Segreganz (1.5), Korrektklassifikations- (1.6) und Falschklassifikationsrate (1.7) unabhängig von den zehn ausgelassenen Sequenzen in konstanten Wertebereichen blieben. Somit konnten mit dem Bootstrapping-Verfahren zwei wichtige und funktionelle Eigenschaften von YamP gezeigt werden. Die erste Eigenschaft stellt die konstant sehr gute Klassifikationsrate unabhängig von den ausgelassenen Sequenzen dar. Dadurch konnte ausgeschlossen werden, dass einzelne microRNA-Sequenzen oder -Gruppen einen höheren Beitrag zur Klassifikationseffizienz beisteuern als andere. Die zweite Eigenschaft lässt sich aus der ersten ableiten und beinhaltet die konstant sehr gute Diskriminierungseffizienz. Hieraus lässt sich ableiten, dass trotz Auslassen einzelner Sequenzen eine korrekte Klassifizierung der ausgelassenen Sequenzen gegeben ist, was für einen genomischen Ansatz und einer *de novo*-Identifikation microRNA-kodierender Bereiche von enormer Bedeutung ist.

Die bisherige Validierung von YamP umfasste allein die Anwendung auf künstlich konstruierte Testdatensätze. Ein genomischer Ansatz stellt jedoch weit größere Herausforderungen an ein Klassifizierungsprogramm dar, da in den untersuchten intergenischen und intronischen Sequenzregionen mit vielen Sequenzen gerechnet werden musste, welche ebenfalls potentiell eine Hairpin-Struktur ähnlich denen der microRNAs ausbilden können. Ein erster Ansatz zur Klassifizierung realer Sequenzen umfasste dabei die in Abschnitt 1.3 eingeführten *small nucleolar* RNAs (snoRNA). Diese entfalten ihre Funktionalität über die Ausbildung von zwei stabilen Hairpin-Strukturen, die denen von microRNA-Vorläufern auf den ersten Blick ähnlich erscheinen. Es zeigte sich jedoch, dass bereits die implementierten Filter-Schritte die Klasse der snoRNAs korrekt klassifizieren konnte, so dass kein Kandidat in die finale Klassifizierung durch YamP einging (siehe Abschnitt 3.3.3).

Eine weitere große Gruppe von genomischen Elementen stellen die repetitiven Elemente, Transposons, Retrotransposons und Satelliten-DNAs dar. Besonders die repetitiven Elemente sind prädestiniert, in thermodynamischen Sekundärstrukturberechnungen mit RNAfold eine sehr stabile Struktur auszubilden. Die Anwendung auf die RepeatMasker-Sequenzbibliothek stellte ein weiteres Szenario dar. Zwar war YamP in der Lage, diese mit einer guten Effizienz zu diskriminieren (siehe Tabelle 3.8), jedoch sank

die Effizienz für einen genomischen Ansatz auf nicht mehr akzeptable Werte herab. Da bekannt ist, dass besonders intergenische Bereiche viele repetitive Bereiche beinhalten, musste der Suchraum bereits im Vorfeld um diese repetitiven Elemente bereinigt werden. Eine Möglichkeit zum Umgehen dieses Problems stellt der Einsatz des Programms RepeatMasker dar, welches für solche Zwecke entwickelt wurde.

Neben der korrekten Klassifizierung von nicht-microRNA-Sequenzen muss zusätzlich die korrekte Klassifizierung unbekannter microRNA-Sequenzen mit guter Effizienz für ein Vorhersageprogramm wie YamP gegeben sein. Bisher wurden entweder künstlich konstruierte oder richtig-negative Sequenzen mit sehr guter Effizienz korrekt klassifiziert. Im Verlauf dieser Arbeit wurde eine neue Version der miRBase veröffentlicht, welche eine interessante Möglichkeit lieferte, die Funktionalität des Programms zu bestätigen. Die neue Version der miRBase enthält 67 neue, teilweise verifizierte microRNA-Vorläufer, welche für die Bestätigung der Funktionalität von YamP eingesetzt wurde. YamP wurde mit den richtig-negativen Trainingsdatensätzen 1 und 2 sowie den microRNA-Sequenzen trainiert und auf die neue Version 10.0 der miRBase angewendet. Dabei zeigte sich, dass bereits der spezifischere Trainingsdatensatz 2 von den insgesamt 67 dem Programm unbekannten microRNA-Sequenzen 48 ($\approx 72\%$) korrekt klassifizieren konnte, der sensitivere Trainingsdatensatz vermochte sogar 55 ($\approx 82\%$) zu identifizieren. Somit konnte eine sehr gute Funktionalität auch bei Anwendung auf unbekannte microRNA-Sequenzen gezeigt werden.

Insgesamt zeigte YamP bei allen Validierungsschritten eine sehr gute Funktionalität, was einen erfolgreichen Einsatz auf genomische Sequenzdaten impliziert. Die Identifikation neuer microRNA-kodierender Bereiche wurde bereits in Kapitel 4 im Detail erläutert.

5.1.4 Vergleich mit anderen Vorhersage-Methoden für pflanzliche microRNAs

Die Funktionalität und Effizienz von YamP konnte, wie in den Abschnitten 3.2 und 3.3 beschrieben, deutlich gezeigt werden. Ein weiterer interessanter Aspekt an dieser Stelle ist der direkte Vergleich mit bereits bestehenden und öffentlich zugänglichen Methoden zur Identifikation neuer microRNA-Kandidaten. Viele der aus der Literatur bekannten microRNA-Vorhersagemethoden beruhen auf einer Homologie-Suche und eignen sich nicht für einen direkten Vergleich mit YamP. Hierzu gehört das in Abschnitt 1.5.1 vorgestellte Programm microHARVESTER von Dezulian *et al.* (2006) sowie ein weiterer Homologie-basierter Ansatz von Wang *et al.* (2004). Ein zusätzlicher Grund für den nicht möglichen Vergleich mit YamP ist die Kenntnis einer möglichen microRNA-Vorläufer-Struktur und -Sequenz sowie ein Hinweis auf den Ort der reifen microRNA. Da es sich bei dem im Rahmen dieser Arbeit entwickelten Ansatz um eine sogenannte *de novo*-Vorhersagemethode handeln sollte, stehen solche Informationen anfänglich nicht zur Verfügung. Prinzipiell wäre microHARVESTER zur Bestätigung einiger möglicher microRNA-kodierender Bereiche denkbar, jedoch nur, wenn es sich bei diesen um Mitglieder einer bereits bekannten microRNA-Familie handelt.

Ein verfügbares Programm, welches nach den Angaben in der Literatur zur *de novo*-Identifikation pflanzlicher microRNA-Sequenzen befähigt ist, stellt das von Adai *et al.* (2005) entwickelte Programm findMiRNA dar. Zunächst wurde findMiRNA auf die Fähigkeit getestet, microRNA-Sequenzen korrekt zu klassifizieren. findMiRNA verlangt als Eingabe-Sequenzen eine Datei mit intergenischen sowie eine Datei mit *messenger*-RNA-Sequenzen (mRNA-Sequenzen). Die microRNA-Sequenzen wurden dabei als intergenische Sequenzen behandelt und umfassten die miRBase in Version 7.0. Als Eingabe-mRNA-Sequenzen wurden die entsprechenden Sequenzen der TAIR-Sequenzdatenbank verwendet. Die Sichtung der Ergebnisse von findMiRNA gestaltete sich dabei sehr kompliziert, da dieses Programm große Mengen an unsortierter Ausgabe produzierte. Jede der 117 microRNAs wurde auch von findMiRNA als solche identifiziert, wobei jedoch auffiel, dass die Eingabemenge von 117 microRNA-Vorläufer-Sequenzen insgesamt 3.088 unterschiedliche mögliche microRNA-Vorläufer produzierte. Auf eine genaue Analyse der korrekten Klassifizierung und Erkennung der tatsächlich korrekten microRNA-Sequenzen wurde an dieser Stelle verzichtet. Erwähnenswert ist an dieser Stelle jedoch die Tatsache, dass viele vorhergesagte Vorläufer-Strukturen eine für microRNAs untypische Sekundärstruktur einnahmen oder eine Hairpin-Struktur ausbildeten, welche eine Länge aufwiesen, die unterhalb der Mindestlänge eines microRNA-Vorläufers zur Ausprägung seiner Funktionalität liegt. In einem zweiten Vergleichsansatz wurde die Fähigkeit zur korrekten Klassifizierung richtig-negativer Sequenzen durch findMiRNA getestet. An dieser Stelle dienten die in Abschnitt 3.3.3 vorgestellten Sequenzen mit zufälliger Neusortierung der Basenreihenfolge unter den in Abschnitt 3.3.3 erwähnten Voraussetzung der Beibehaltung der Mono-Nukleotid- und/oder Di-Nukleotid-Frequenzen. Auch hier wurden sämtliche definitiv richtig-negativen Sequenzen als microRNA-Vorläufer klassifiziert.

YamP besitzt, wie in Abschnitt 3.3.3 beschrieben, einige Schwächen bei der korrekten Klassifizierung repetitiver Elemente, Transposons oder Retrotransposons, welche oftmals als Ursprung kleiner RNAs identifiziert wurden (zur Übersicht Okamoto & Hirochika, 2001). Ein weiterer Vergleich zeigte, dass auch findMiRNA bei der korrekten Klassifizierung solcher Regionen Schwächen aufweist. In einem weiteren Vergleichsansatz wurde die RepeatMasker-Bibliothek aus Abschnitt 3.3.3 als falsch-negativer Sequenzdatensatz gewählt und durch findMiRNA klassifiziert. Dabei zeigte sich auch hier, dass alle Sequenzen der RepeatMasker-Bibliothek als mögliche microRNA-Kandidaten klassifiziert wurden. Aus dem direkten Vergleich beider Programme lassen sich zwei wichtige Schlussfolgerungen ableiten. Zum einen weisen Programme zur *de novo*-Identifikation microRNA-kodierender Bereiche in genomischen Sequenzdaten Schwächen bei der korrekten Klassifizierung repetitiver und transposabler Elemente auf. Dies ist bedingt durch den hohen Grad an Komplementarität dieser Regionen und daraus resultierenden Sekundärstrukturen, welche eine deutliche Ähnlichkeit zu microRNA-Vorläufern aufweisen. Desweiteren lässt sich sagen, dass Versuche zur *de novo*-Identifikation microRNA-kodierender Bereiche über die Suche nach möglichen Zielgenen mit kurzen RNA-Sequenzen und anschließender Herleitung des microRNA-Vorläufers deutlich zu viele falsch-positive Vorhersagen trifft.

Abschließend lässt sich zur gezeigten Funktionalität von YamP zusammenfassen, dass der entwickelte Ansatz zur Vorhersage von microRNAs einen effizienten und vielversprechen-

den Eindruck hinterlässt, was über die gezeigten Validierungsschritte sowie den Vergleich mit einem anderen verfügbaren Programm gezeigt werden konnte.

5.2 Biologische Ergebnisse

Das Programm YamP zeigte in den umfangreichen Testumgebungen eine sehr gute Effizienz bezüglich der korrekten Klassifizierung von microRNA- und nicht-microRNA-Sequenzen. Auch der Vergleich zu anderen verfügbaren Programmen hob die sehr gute Funktionalität deutlich hervor. Der logische Schritt an dieser Stelle war die Ausweitung des Einsatzgebietes auf genomische Sequenzdaten zur *de novo*-Identifikation neuer microRNA-Kandidaten.

5.2.1 Generelle Funktionalität

Die prinzipielle Funktionalität von YamP konnte in Abschnitt 3.3 gezeigt werden. Ob auch der Einsatz von RNALfold als Werkzeug zur Identifikation von microRNA-Vorläufer-Strukturen in genomischen Sequenzen geeignet ist, bedurfte eingehender Analysen. Hierzu wurden die bekannten microRNA-Sequenzen aus der miRBase 7.0 inklusive 500 Nukleotide 5'- und 3'-seitig aus den genomischen Sequenzdaten mittels HyPa extrahiert. Dieses Szenario diente der Verdeutlichung, dass der Ansatz unter Zuhilfenahme von RNALfold prinzipiell in der Lage ist, die korrekten microRNA-Vorläufer-Strukturen aus ihrem genomischen Kontext zu extrahieren. Es zeigte sich, dass von den 117 microRNA-Sequenzen insgesamt 114 microRNA-Sequenzen korrekt extrahiert wurden (Daten nicht gezeigt). Bei den verbleibenden drei Sequenzen (*ath-mir166f*, *ath-mir420* und *ath-mir426*) zeigte sich, dass diese relativ zu ihrem genomischen Kontext nicht in der Lage sind, eine stabile Hairpin-Struktur auszubilden. Zudem konnten Xie *et al.* (2005) zeigen, dass *ath-mir420* einem Transposon entstammt und wahrscheinlich keine microRNA-kodierende Region darstellt. Jones-Rhoades *et al.* (2006) zeigten zudem, dass auch bei *ath-mir426* Zweifel bestehen, diese als microRNA-kodierend zu bezeichnen. Somit ist einzig bei *ath-mir166f* eine falsch-negative Extraktion der Sequenz aufgetreten, was jedoch als vertretbar angesehen wurde.

Zwei der in der miRBase annotierten microRNAs, *ath-mir420* und *ath-mir426*, stellen nicht die Gesamtheit an zweifelhaften Einträgen in der miRBase dar. Zusätzlich bestehen auch bei den Einträgen *ath-mir413* bis *ath-mir419* berechtigte Zweifel an der microRNA-kodierenden Funktion. Interessanterweise wurden diese von YamP, obwohl in der microRNA-Sequenzdatenmenge vertreten, ebenfalls als nicht-microRNA-Sequenz klassifiziert. Gleiche Ergebnisse wurden auch von Lindow *et al.* (2007) erzielt und lieferten somit eine Bestätigung der Negativ-Klassifizierungen der wahrscheinlich fälschlicherweise als microRNA annotierten Einträge.

Die durchgeführte Suche nach möglichen neuen microRNA-Kandidaten ergab neue Hinweise, dass der entwickelte Ansatz zur *de novo*-Identifikation eine sehr gute Funktionalität

aufwies. Es zeigte sich, dass in den intergenischen und intronischen Sequenzdaten der TAIR-Sequenzdatenbank einige verifizierte microRNA-Sequenzen in der aktuellen Version noch nicht als solche annotiert und entfernt wurden. Die Abbildung 4.1 zeigt zwei Beispiele der Identifikation eben solcher microRNA-kodierender Bereiche. Auf eine genaue Auswertung bezüglich dieser Bereiche, die noch nicht in der Sequenzdatenbank des TAIR annotiert sind, wurde an dieser Stelle verzichtet. Somit konnte auch für den in Abschnitt 3.2.1 beschriebenen genomischen Ansatz, und im Speziellen dem Einsatz von RNALfold als Werkzeug zur Extraktion lokal stabiler Strukturelemente, eine sehr gute Funktionalität zugesprochen werden.

5.2.2 MicroRNA-Kandidaten

Die in Kapitel 4 beschriebene *de novo*-Identifikation microRNA-kodierender Bereiche in den intergenischen und intronischen Sequenzbereichen des Genoms von *Arabidopsis thaliana* ergab insgesamt ca. 1.500 positive Klassifizierungen. Ein Abgleich mit der Expressionsdatenbank vom *Arabidopsis Small RNA Project* (ASRP) zeigte dabei, dass viele der als positiv klassifizierten genomischen Bereiche in der Mehrheit zwei Muster der Expression kleiner RNAs aufwiesen, welche sich deutlich von den typischen Expressionsmustern verifizierter microRNA-kodierender Regionen unterscheiden. Verifizierte microRNA-Vorläufer-Regionen weisen, wie in Abbildung 4.1 dargestellt, meist sehr typische Muster auf, die dem reifen microRNA-Duplex entsprechen. Der Großteil der positiv klassifizierten Sequenzen zeigte jedoch entweder keine exprimierten kleinen RNAs oder eine Vielzahl exprimierter kleiner RNAs auf, welche den unterschiedlichsten Größenklassen und beiden Orientierungen zuzuordnen waren. Diese Muster der Expression kleiner RNAs sind häufig bei z. B. repetitiven Elementen oder Transposons im Genom von *Arabidopsis thaliana* zu finden (siehe Abbildung 5.1). Wie schon in Abschnitt 5.1.3 dargestellt, besaß das Programm z. B. bei repetitiven oder transposablen Elementen Schwächen, welche sich jedoch durch einen zusätzlichen Filter-Schritt beheben lassen (in Abschnitt 5.1.3 diskutiert). Viele der positiv-klassifizierten Regionen zeigten hingegen keine Expression kleiner RNAs. Unter der Annahme, dass die beiden Expressionsdatenbanken vom ASRP und MPSS als Standard angesehen werden, könnten diese Bereiche ohne exprimierte kleine RNAs als falsch-positiv klassifizierte Bereiche gedeutet werden. Tagami *et al.* (2007) klonierten und sequenzierten kleine RNAs aus *Arabidopsis thaliana*, die mit dem *Tobacco Mosaic Virus* infiziert waren. Sie konnten zeigen, dass einige microRNAs zu einem höheren Prozentsatz in der Pflanze akkumulierten als in nicht-infizierten Pflanzen. Sie führten dies auf die spezifische Bindung eines viralen replikationsassoziierten Proteins an den microRNA/microRNA*-Duplex zurück. Bei dem Versuch, die sequenzierten microRNAs im Genom zu lokalisieren, stellten sie fest, dass viele der sequenzierten microRNAs nicht in den beiden Expressionsdatenbanken ASRP und MPSS eingetragen waren. Viele der durch YamP ermittelten möglichen microRNAs zeigten in ihrer genomischen Region ebenfalls keinerlei Expression kleiner RNAs. Dies legt den Schluss nahe, dass aufgrund fehlender Expressionsdaten keine endgültige Aussage getroffen werden kann, ob es sich hier um das Vorliegen eines microRNA-kodierenden

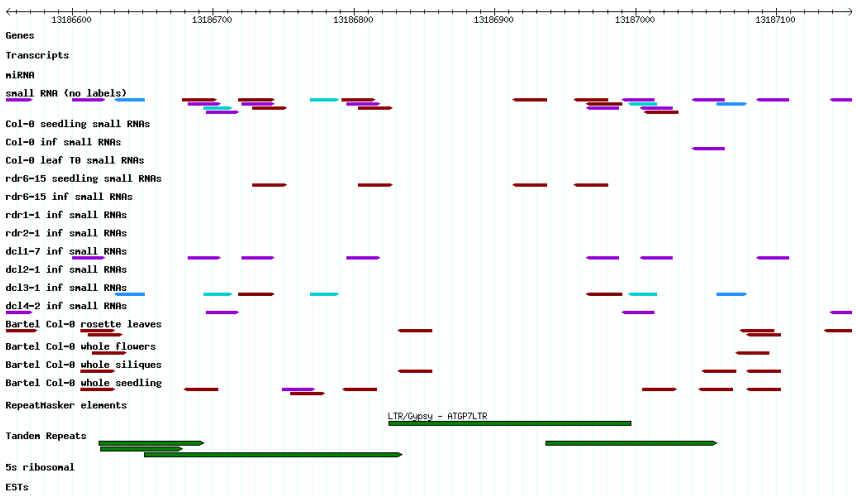

Abbildung 5.1: ASRP-Expressionsdaten aus dem Bereich eines Retrotransposons. Die Darstellung zeigt die genomische Region eines Retrotransposons. Erkennbar ist die Vielzahl unterschiedlicher RNA-Größenklassen sowie die unterschiedliche Orientierung der exprimierten kleinen RNAs.

Bereichs handelt oder nicht. Da die intrinsischen Eigenschaften dieser Regionen denen typischer microRNA-Vorläufer entsprechen und zudem keinerlei repetitive Elemente oder Transposons für diesen Bereich angegeben sind, scheint die Vermutung naheliegend zu sein, dass es sich bei diesen Bereichen um microRNA-kodierende Regionen handeln könnte.

Die ca. 1.500 positiven Klassifizierungen warfen die Frage auf, ob YAMP eine große Anzahl falsch-positiver Klassifizierungen vorgenommen hat oder diese doch als positiv klassifiziert werden könnten. Nimmt man die Vorhersagen in repetitiven Bereichen heraus, bleibt eine große Anzahl an nicht eindeutigen Vorhersagen stehen, welche sich ohne die Bestätigung über experimentelle Daten nicht verifizieren lassen. Jedoch legten Lindow *et al.* (2007) Hinweise über einen deutlich größeren regulatorischen Funktionsumfang durch microRNAs dar. Sie konnten über einen intragenomischen Abgleich komplementäre Regionen im Genom von *Arabidopsis thaliana* aufdecken und führten dies auf einen deutlich größeren Umfang an microRNA-vermittelten Regulationsprozessen zurück. Weiter warfen sie die Frage auf, inwiefern repetitive Elemente unbekannter Art eventuell ebenfalls als regulatorische Elemente fungieren können. Betrachtet man die vorliegenden Ergebnisse unter diesen Gesichtspunkten, scheint die vorliegende Klassifizierungsmethode nicht zu viele positive Klassifizierungen zu treffen.

Die in Abschnitt 3.2 beschriebene und in Abbildung 3.8 beispielhaft dargestellte Lokalisation des microRNA/microRNA*-Duplex konnte auch bei den fünf hier vorgestellten Kandidaten gezeigt werden: Die vorhergesagten Duplices stimmten sehr gut mit denen in der ASRP- und MPSS-Expressionsdatenbank eingetragenen kleinen RNAs überein. Somit konnte zusätzlich an unbekannten und möglichen neuen microRNA-Kandidaten die Funktionalität der Lokalisation gezeigt werden, welche für eine Vorhersage in *PSTVd* erforderlich ist (siehe Abschnitt 4.4).

Neben den erforderlichen intrinsischen Eigenschaften für microRNA-Vorläufer-Sequenzen und -Strukturen konnten auch unabhängige experimentelle Daten zur Verifikation der getroffenen Vorhersagen herangezogen werden. Eine folgende Identifikation möglicher Zielgene der vorgestellten Kandidaten zeigte dabei interessante mögliche Regulationsmechanismen auf. So konnte ein Kandidat identifiziert werden, der mit hoher Wahrscheinlichkeit in die Regulation der Photomorphogenese involviert ist (siehe Abschnitt 4.5.2). Auch für die anderen vorgestellten Kandidaten deuten gute Hinweise darauf hin, dass die ermittelten Zielgene durch diese reguliert werden, da die sequenzierten kleinen RNAs eine annähernd perfekte Komplementarität zu ihren potentiellen Zielgenen aufweisen. So scheinen alle ermittelten Zielgene mit hoher Wahrscheinlichkeit in die Regulation der Signaltransduktion äußerer Einflüsse oder in die Regulation entwicklungsspezifischer Prozesse involviert zu sein.

5.2.3 Kleine Viroid-spezifische RNAs

Eines der Ziele dieser Arbeit stellte die Identifikation möglicher Bereiche im Genom des *Potato Spindle Tuber Viroid* (*PSTVd*) dar, welche das Potenzial besitzen, Ursprung kleiner Viroid-spezifischer RNAs zu sein. Die Prozessierung bzw. Produktion von kleinen Viroid-spezifischen RNAs bedingt, dass das vorliegende stäbchenförmige Viroid von einem Mitglied der *Dicer-like proteins* (DCL) als RNA-Doppelstrang bzw. microRNA-Vorläufer erkannt und fälschlicherweise prozessiert wird. In der Tat konnten Itaya *et al.* (2007) beobachten, dass *PSTVd* ein Substrat für ein noch nicht identifiziertes Mitglied der DCL-Proteine darstellt. Weiterhin konnten sie beobachten, dass das Viroid keiner *Dicer*-vermittelten Degradation unterliegt, was wohl auf die thermodynamisch stabile Sekundärstruktur zurückzuführen ist. Somit kann an dieser Stelle zunächst einmal ausgeschlossen werden, dass die kleinen Viroid-spezifischen RNAs als eine Antwort der pflanzlichen Zelle auf eindringende RNA-Doppelstränge zu werten sind, die auch experimentell nie nachgewiesen werden konnten. Auch scheint bedingt durch die thermodynamisch sehr stabile Sekundärstruktur des *PSTVd* eine Wirts-vermittelte Degradation der eindringenden „nackten“ RNA ausgeschlossen. Schubert *et al.* (2005) zeigten den Einfluss lokaler Strukturelemente auf die Effizienz der *small-interfering*-RNA-vermittelten Degradation der Zielgene (siRNA). Sie konnten zeigen, dass hochstrukturierte und thermodynamisch sehr stabile Elemente in der Zielregion der *messenger*-RNAs (mRNA) die Effizienz der Degradation der Ziel-mRNAs oder Inhibition der Translation dramatisch absenken. Das Viroid-Genom besitzt in seiner nativen Form eine sehr stabile stäbchenförmige Sekundärstruktur, die es wahrscheinlich vor einer Wirts-vermittelten Degradation schützt. Daher scheint es logischer, dass die kleinen Viroid-spezifischen RNAs in RNA-vermittelte Regulationsmechanismen involviert sind und spezifisch oder unspezifisch einzelne Wirtsgene durch eine entsprechende Komplementarität fälschlicherweise regulieren. Denkbar wäre auch das Szenario, dass die pflanzliche Zelle das *PSTVd* von einem Mitglied der DCL-Proteine gespalten wird, die Degradation des *PSTVd* im Zytoplasma jedoch aufgrund der stabilen Sekundärstruktur und der vornehmlichen Lokalisation im Nukleus (Qi & Ding, 2003) nicht stattfinden kann. Itaya *et al.* (2007) zeigten jedoch, dass die kleinen

Viroid-spezifischen RNAs hauptsächlich drei distinkten Bereichen zuzuordnen sind. Dies spricht eher für eine Involvierung in den microRNA-Stoffwechselweg, da bei diesen die reifen microRNAs zum Großteil zwei distinkten Bereichen entstammen.

Somit ist auch im Nachhinein der Versuch einer Lokalisation der kleinen Viroid-spezifischen RNAs über ein Vorhersage-Modell für microRNAs sinnvoll. Die Vorhersage für oder gegen das Vorliegen einer möglichen microRNA-kodierenden Region ergab dabei, dass vornehmlich zwei verschiedene Bereiche von YamP als microRNA-kodierend identifiziert werden konnten. Dabei handelt es sich im (+)-Strang zum einen um eine Region in der Terminal-rechten Region (TR-Region) sowie zum anderen um die Pathogenitäts-modulierende Region (*Virulence-Modulating*-Region, VM-Region). Diese Ergebnisse korrelieren sehr gut mit den experimentellen Daten von Itaya *et al.* (2007), welche kleine Viroid-spezifische RNAs sequenzieren und diese dem Genom eines *Intermediate*-Stamm zuordnete. Sie konnten zeigen, dass die kleinen Viroid-spezifischen RNAs dabei hauptsächlich zwei Regionen in der TR-Region sowie einer Region in der Terminal-linken Region (TL-Region) zugeordnet werden konnten. Einer der Bereiche in der TR-Region wurde auch von YamP als möglicher Ursprung kleiner Viroid-spezifischer RNAs identifiziert, was die Vorhersage von YamP bestätigen würde (siehe Abbildung 4.20). In weiteren *in vitro*-Experimenten konnten Itaya *et al.* (2007) desweiteren kleine Viroid-spezifische RNAs sequenzieren, die der VM-Region zugeordnet wurden. Auch diese Region konnte von YamP im *PSTVd*-(+)- und (−)-Strang als möglicherweise microRNA-kodierend identifiziert werden. Die VM-Region als Ursprung der funktionellen Viroid-spezifischen RNAs scheint die wahrscheinlichste zu sein. Mutationen in eben dieser Region bewirken unterschiedlich stark ausgeprägte Symptome in der Wirtspflanze. Eine Erklärung dafür könnte eine durch Nukleotidaustausche herabgesetzte Komplementarität der kleinen Viroid-spezifischen RNAs zu den Zielgenen und somit einer verminderten Effizienz der Viroid-spezifisch-vermittelten Degradation oder Inhibition der Zielgene sein.

Inwiefern die kleinen Viroid-spezifischen RNAs der TR-Region, welche von Itaya *et al.* (2007) und in dieser Arbeit ermittelt wurden, in die Pathogenität involviert sind, kann an dieser Stelle nicht abschließend geklärt werden. Aus der Literatur ist nicht bekannt, dass Nukleotidaustausche in der TR-Region oder auch in der *Variable*-Region unterschiedlich stark ausgeprägte Symptome induzieren. Somit kann durch YamP ausschließlich ein zusätzlicher Hinweis erbracht werden, dass die VM-Region einen Bereich als Ursprung kleiner Viroid-spezifischer RNAs darstellt, wodurch sich die bekannten Beobachtungen über unterschiedlich stark ausgeprägte Symptome erklären lassen würden.

6

Zusammenfassung

Derzeit stellen die microRNAs die wohl wichtigste Gruppe an regulatorischen Molekülen dar, welche die Regulation der Genexpression auf posttranskriptioneller Ebene durch Basenpaarungen zur Ziel-mRNA vermitteln. Sie nehmen eine wichtige Rolle in der Entwicklung (Lee *et al.*, 1993), Stammzell-Differenzierung (Houbaviy *et al.*, 2003), Signaltransduktion (Guo *et al.*, 2005) und diversen Krankheiten wie z. B. Krebs ein (Lu *et al.*, 2005). Die Klasse der microRNAs wurde als erstes in *C. elegans* von Lee *et al.* (1993) beschrieben. Seither wurden in immer weiteren Organismen neue microRNA-Familien identifiziert. Dabei zeichnete sich ab, dass es sich bei der Regulation durch microRNAs um einen evolutiv stark konservierten Mechanismus handeln muss. Die Mehrheit der microRNAs ist evolutionär stark zwischen entfernt verwandten Spezies wie Mensch und Wurm, Moosen und höheren Pflanzen konserviert.

Viroide sind strukturell, funktionell und evolutiv von den Viren zu unterscheiden. Trotz ihrer nur geringen Größe von 246–401 Nukleotiden und ihrer Unfähigkeit für Proteine zu kodieren, sind sie in der Lage, höhere Pflanzen zu infizieren und unterschiedlichste Krankheitsformen auszulösen (Daròs *et al.*, 2006). Viroid-Infektionen gehen mit der Akkumulation kleiner Viroid-spezifischer RNAs einher, wobei davon ausgegangen wird, dass diese die pathogenen Effekte in den infizierten Pflanzen vermitteln (Matousek *et al.*, 2007).

In dieser Arbeit sollte über einen *de novo*-microRNA-Vorhersageansatz die Hypothese untermauert werden, wonach Viroide und speziell *Potato Spindle Tuber Viroid* (*PSTVd*) ihre Pathogenität über kleine Viroid-spezifische RNAs vermitteln. Existierende Vorhersagemethoden (Adai *et al.*, 2005; Dezulian *et al.*, 2006) zur *de novo*-Identifikation von microRNAs sind für diesen Zweck nicht verwendbar, worin sich die Entwicklung eines neuen Ansatzes begründet. Das in dieser Arbeit entwickelte Programm Yet Another miRNA Predictor (YamP) konnte in einer Reihe von Validierungsschritten seine Funktionalität mit herausragender Effizienz unter Beweis stellen, die sich auch in der durchgeführten, genomischen Suche nach neuen möglichen microRNA-kodierenden Bereichen im Modellorganismus *Arabidopsis thaliana* deutlich zeigen ließ.

Es konnten etliche neue microRNA-kodierende Bereiche identifiziert werden, welche in die diversesten Stoffwechsel-physiologischen Prozesse involviert zu sein scheinen. So konnte z. B. eine mögliche microRNA identifiziert werden, die an der Regulation eines Proteins, das an der Lichtantwort in der Zelle beteiligt sein könnte (COP1-Protein, Deng *et al.*, 1991) und somit in die Photomorphogenese der Pflanze involviert ist. Über die Evolution von microRNAs ist derzeit noch nicht viel bekannt. Es wird jedoch davon ausgegangen, dass manche microRNAs durch Duplikationsereignisse ihrer Zielgene entstanden sind (Allen *et al.*, 2004). Eine der von YamP detektierten, neuen microRNAs scheint in der Evolution von *Arabidopsis thaliana* genau so entstanden zu sein, da der Sequenzbereich, in dem die reife microRNA liegt, eine signifikante Seqeunzähnlichkeit zum Zielgen der microRNA aufweist.

Die Klassifizierung einzelner *PSTVd*-Domänen bzw. des Gesamtgenoms als microRNA-Vorläufer durch YamP konnte neue Indizien erbringen, dass *PSTVd* in der Wirtspflanze fälschlicherweise als microRNA-Vorläufer erkannt und prozessiert wird. Durch die zusätzlich vorgenommene Lokalisation der Bereiche, die nach dem vorliegenden Modell das Potenzial besitzen, einen reifen RNA-Duplex zu beinhalten, ergab ein Resultat, wel-

ches mit den experimentellen Daten von Itaya *et al.* (2007) korreliert. Es konnte ebenfalls die Terminal-rechte sowie die Pathogenitäts-modulierende Region als Ursprung kleiner RNAs identifiziert werden.

Currently, microRNAs are likely the most important molecules that regulate gene expression at the posttranscriptional level by targeting their mRNAs through complementary basepairing interactions and mediate degradation or translational inhibition. They play a crucial role in important biological processes like developmental timing (Lee *et al.*, 1993), stem cell differentiation (Houbaviy *et al.*, 2003), signal transduction (Guo *et al.*, 2005) and diverse disease like cancer (Lu *et al.*, 2005). MicroRNAs are initially discovered in *C. elegans* (Lee *et al.*, 1993). Since their discovery, more microRNAs are identified in further organisms suggesting that it must be an ultra-conserved regulatory mechanism. The majority of microRNAs is evolutionary conserved between distant related organisms like human and worm, mosses and higher eudicot plants.

Viroids are structurally, functionally and evolutionary distinct from viruses. They are capable of infecting higher land plants and cause severe diseases despite their minimal size (256–401 nucleotides) and lack of protein coding capacity (Daròs *et al.*, 2006). Viroid infections are accompanied with an accumulation of small viroid-specific RNAs which implies that they mediate the pathogenic effects in infected plants (Matousek *et al.*, 2007).

In this work a computational microRNA prediction method should corroborate this hypothesis whereupon viroids and especially *Potato Spindle Tuber Viroid* (*PSTVd*) mediates their pathogenicity through small viroid-specific RNAs. Existing methods for computational *de novo* prediction of microRNAs (Adai *et al.*, 2005; Dezulian *et al.*, 2006) are not applicable for this purpose thus asking for the development of a new computational method for predicting new microRNAs. The program Yet Another miRNA Predictor (YamP) developed in this work exhibits a very good functionality which was demonstrated in several validation steps and was supported by a genomic search for microRNAs in the model organism *Arabidopsis thaliana*.

Several, hitherto unknown new microRNA-coding regions were identified in intronic and intergenic sequence data of *Arabidopsis thaliana*; these are possibly involved in diverse biological pathways. One of the new microRNA seems to be involved in the regulation of a protein-coding region that exhibits a significant homology to another protein-coding region which is involved in photomorphogenesis (COP1-Protein, Deng *et al.*, 1991). Only little is known about the evolution of microRNA-coding regions in genomes; Allen *et al.* (2004) claim, however, that some microRNAs are a result of an inverted gene duplication event that formed both arms of the newly evolved microRNA-coding region and that these microRNAs have unique target specificity. One candidate might be a result of such a duplication event because the region surrounding the mature microRNA exhibits a significant sequence similarity to one of the possible target mRNAs.

The classification of single *PSTVd*-domains and the whole *PSTVd*-genome as a precursor-microRNA yield new indications that *PSTVd* could misleadingly be detected and processed by the host as a precursor-microRNA, respectively. The additionally conducted localization of regions with the potential to contain a mature RNA duplex with respect to the corresponding model correlated well with experimental results from Itaya *et al.* (2007). So the terminal right and virulence modulating region could be asserted as these regions, that likely contain the functional mature small RNAs.

7

Literatur

Adai, Alex, Johnson, Cameron, Mlotshwa, Sizolwenkosi, Archer-Evans, Sarah, Manocha, Varun, Vance, Vicki & Sundaresan, Venkatesan (2005). Computational prediction of miRNAs in Arabidopsis thaliana. *Genome Res*, 15(1), 78–91.

Allen, Edwards, Xie, Zhixin, Gustafson, Adam M, Sung, Gi-Ho, Spatafora, Joseph W & Carrington, James C (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. *Nat Genet*, 36(12), 1282–1290.

Almeida, Ricardo & Allshire, Robin C (2005). RNA silencing and genome regulation. *Trends Cell Biol*, 15(5), 251–258.

Aukerman, Milo J & Sakai, Hajime (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *Plant Cell*, 15(11), 2730–2741.

Backman, Tyler W H, Sullivan, Christopher M, Cumbie, Jason S, Miller, Zachary A, Chapman, Elisabeth J, Fahlgren, Noah, Givan, Scott A, Carrington, James C & Kasschau, Kristin D (2008). Update of ASRP: the Arabidopsis Small RNA Project database. *Nucleic Acids Res*, 36(Database issue), D982–D985.

Bonnet, Eric, de Peer, Yves Van & Rouzé, Pierre (2006). The small RNA world of plants. *New Phytol*, 171(3), 451–468.

Branch, Andrea D, Robertson, Hugh D & Dickson, Elizabeth (1981). Longer-than-unit-length viroid minus strands are present in RNA from infected plants. *Proc Natl Acad Sci U S A*, 78(10), 6381–6385.

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(6), 630–634.

Crick, F. H. (1970). Central dogma of molecular biology. *Nature*, 227, 561–563.

Daròs, José-Antonio, Elena, Santiago F & Flores, Ricardo (2006). Viroids: an Ariadne's thread into the RNA labyrinth. *EMBO Rep*, 7(6), 593–598.

Daròs, José-Antonio & Flores, Ricardo (2004). Arabidopsis thaliana has the enzymatic machinery for replicating representative viroid species of the family Pospiviroidae. *Proc Natl Acad Sci U S A*, 101(17), 6792–6797.

de Alba, Angel Emilio Martínez, Sägesser, Rudolf, Tabler, Martin & Tsagris, Mina (2003). A bromodomain-containing protein from tomato specifically binds potato spindle tuber viroid RNA in vitro and in vivo. *J Virol*, 77(17), 9685–9694.

Deng, X. W., Caspar, T. & Quail, P. H. (1991). cop1: a regulatory locus involved in light-controlled development and gene expression in Arabidopsis. *Genes Dev*, 5(7), 1172–1182.

Dezulian, Tobias, Remmert, Michael, Palatnik, Javier F, Weigel, Detlef & Huson, Daniel H (2006). Identification of plant microRNA homologs. *Bioinformatics*, 22(3), 359–360.

Diener, T. O. (1971). Potato spindle tuber "virus". IV. A replicating, low molecular weight RNA. *Virology*, 45(2), 411–428.

Diener, T. O. (2001). The viroid: biological oddity or evolutionary fossil? *Adv Virus Res*, 57, 137–184.

Ding, Biao, Itaya, Asuka & Zhong, Xuehua (2005). Viroid trafficking: a small RNA makes a big move. *Curr Opin Plant Biol*, 8(6), 606–612.

Durbin, R. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Eddy, S. R. (2008). SQUID – C function library for sequence analysis. `http://selab.wustl.edu/cgi-bin/selab.pl?mode=software`.

Elbashir, S. M., Lendeckel, W. & Tuschl, T. (2001). RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev*, 15(2), 188–200.

Enright, Anton J, John, Bino, Gaul, Ulrike, Tuschl, Thomas, Sander, Chris & Marks, Debora S (2003). MicroRNA targets in Drosophila. *Genome Biol*, 5(1), R1.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. & Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669), 806–811.

Frank, D. N. & Pace, N. R. (1998). Ribonuclease P: unity and diversity in a tRNA processing ribozyme. *Annu Rev Biochem*, 67, 153–180.

Furner, I. J., Sheikh, M. A. & Collett, C. E. (1998). Gene silencing and homology-dependent gene silencing in Arabidopsis: genetic modifiers and DNA methylation. *Genetics*, 149(2), 651–662.

Ganot, P., Bortolin, M. L. & Kiss, T. (1997). Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell*, 89(5), 799–809.

Garcia-Hernandez, Margarita, Berardini, Tanya Z, Chen, Guanghong, Crist, Debbie, Doyle, Aisling, Huala, Eva, Knee, Emma, Lambrecht, Mark, Miller, Neil, Mueller, Lukas A, Mundodi, Suparna, Reiser, Leonore, Rhee, Seung Y, Scholl, Randy, Tacklind, Julie, Weems, Dan C, Wu, Yihe, Xu, Iris, Yoo, Daniel, Yoon, Jungwon & Zhang, Peifen (2002). TAIR: a resource for integrated Arabidopsis data. *Funct Integr Genomics*, 2(6), 239–253.

Giegerich, Robert, Voss, Björn & Rehmsmeier, Marc (2004). Abstract shapes of RNA. *Nucleic Acids Res*, 32(16), 4843–4851.

Grewal, Shiv I S & Moazed, Danesh (2003). Heterochromatin and epigenetic control of gene expression. *Science*, 301(5634), 798–802.

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

Griffiths-Jones, Sam, Grocock, Russell J, van Dongen, Stijn, Bateman, Alex & Enright, Anton J (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34(Database issue), D140–D144.

Griffiths-Jones, Sam, Moxon, Simon, Marshall, Mhairi, Khanna, Ajay, Eddy, Sean R & Bateman, Alex (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, 33(Database issue), D121–D124.

Gross, H. J., Domdey, H., Lossow, C., Jank, P., Raba, M., Alberty, H. & Sänger, H. L. (1978). Nucleotide sequence and secondary structure of potato spindle tuber viroid. *Nature*, 273(5659), 203–208.

Gräf, S. (2004). *Strukturbasierte Beschreibung, Suche und Annotation nicht-kodierender RNAs.* PhD thesis, Heinrich-Heine-Universität Düsseldorf.

Gräf, S., Strothmann, D., Kurtz, S. & Steger, G. (2001). HyPaLib: a database of RNAs and RNA structural elements defined by hybrid patterns. *Nucleic Acids Res*, 29(1), 196–198.

Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3 Pt 2), 849–857.

Guo, Hui-Shan, Xie, Qi, Fei, Ji-Feng & Chua, Nam-Hai (2005). MicroRNA directs mRNA cleavage of the transcription factor NAC1 to downregulate auxin signals for arabidopsis lateral root development. *Plant Cell*, 17(5), 1376–1386.

Gustafson, Adam M, Allen, Edwards, Givan, Scott, Smith, Daniel, Carrington, James C & Kasschau, Kristin D (2005). ASRP: the Arabidopsis Small RNA Project Database. *Nucleic Acids Res*, 33(Database issue), D637–D640.

Hamilton, Andrew, Voinnet, Olivier, Chappell, Louise & Baulcombe, David (2002). Two classes of short interfering RNA in RNA silencing. *EMBO J*, 21(17), 4671–4679.

Hertel, Jana & Stadler, Peter F (2006). Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, 22(14), e197–e202.

Hofacker, IL, Fontana, W, Stadler, PF, Bonhoeffer, S, Tacker, M & Schuster, P (1994). Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte für Chemie*, **125**, 167–188.

Hofacker, Ivo L (2003). Vienna RNA secondary structure server. *Nucleic Acids Res*, 31(13), 3429–3431.

Hofacker, I. L., Priwitzer, B. & Stadler, P. F. (2004). Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, 20(2), 186–190.

Houbaviy, Hristo B, Murray, Michael F & Sharp, Phillip A (2003). Embryonic stem cell-specific MicroRNAs. *Dev Cell*, 5(2), 351–358.

Hutvagner, Gyorgy & Simard, Martin J (2008). Argonaute proteins: key players in RNA silencing. *Nat Rev Mol Cell Biol*, 9(1), 22–32.

Hutvágner, György & Zamore, Phillip D (2002). A microRNA in a multiple-turnover RNAi enzyme complex. *Science*, 297(5589), 2056–2060.

Itaya, A., Folimonov, A., Matsuda, Y., Nelson, R. S. & Ding, B. (2001). Potato spindle tuber viroid as inducer of RNA silencing in infected tomato. *Mol Plant Microbe Interact*, 14(11), 1332–1334.

Itaya, Asuka, Zhong, Xuehua, Bundschuh, Ralf, Qi, Yijun, Wang, Ying, Takeda, Ryuta, Harris, Ann R, Molina, Carlos, Nelson, Richard S & Ding, Biao (2007). A structured viroid RNA serves as a substrate for dicer-like cleavage to produce biologically active small RNAs but is resistant to RNA-induced silencing complex-mediated degradation. *J Virol*, 81(6), 2980–2994.

Jones-Rhoades, Matthew W, Bartel, David P & Bartel, Bonnie (2006). MicroRNAS and their regulatory roles in plants. *Annu Rev Plant Biol*, 57, 19–53.

Keese, P. & Symons, R. H. (1985). Domains in viroids: evidence of intermolecular RNA rearrangements and their contribution to viroid evolution. *Proc Natl Acad Sci U S A*, 82(14), 4582–4586.

Khvorova, Anastasia, Reynolds, Angela & Jayasena, Sumedha D (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115(2), 209–216.

Kim, V. Narry (2004). MicroRNA precursors in motion: exportin-5 mediates their nuclear export. *Trends Cell Biol*, 14(4), 156–159.

Kiss-László, Z., Henry, Y., Bachellerie, J. P., Caizergues-Ferrer, M. & Kiss, T. (1996). Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell*, 85(7), 1077–1088.

Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E. & Cech, T. R. (1982). Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, 31(1), 147–157.

Kurihara, Yukio & Watanabe, Yuichiro (2004). Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc Natl Acad Sci U S A*, 101(34), 12753–12758.

Lee, R. C., Feinbaum, R. L. & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5), 843–854.

Lim, Lee P, Lau, Nelson C, Weinstein, Earl G, Abdelhakim, Aliaa, Yekta, Soraya, Rhoades, Matthew W, Burge, Christopher B & Bartel, David P (2003). The microRNAs of Caenorhabditis elegans. *Genes Dev*, 17(8), 991–1008.

Lindow, Morten, Jacobsen, Anders, Nygaard, Sanne, Mang, Yuan & Krogh, Anders (2007). Intragenomic matching reveals a huge potential for miRNA-mediated regulation in plants. *PLoS Comput Biol*, 3(11), e238.

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

Lingel, Andreas, Simon, Bernd, Izaurralde, Elisa & Sattler, Michael (2004). Nucleic acid 3'-end recognition by the Argonaute2 PAZ domain. *Nat Struct Mol Biol*, 11(6), 576–577.

Liu, Jidong, Carmell, Michelle A, Rivas, Fabiola V, Marsden, Carolyn G, Thomson, J. Michael, Song, Ji-Joon, Hammond, Scott M, Joshua-Tor, Leemor & Hannon, Gregory J (2004). Argonaute2 is the catalytic engine of mammalian RNAi. *Science*, 305(5689), 1437–1441.

Lu, Jun, Getz, Gad, Miska, Eric A, Alvarez-Saavedra, Ezequiel, Lamb, Justin, Peck, David, Sweet-Cordero, Alejandro, Ebert, Benjamin L, Mak, Raymond H, Ferrando, Adolfo A, Downing, James R, Jacks, Tyler, Horvitz, H. Robert & Golub, Todd R (2005). MicroRNA expression profiles classify human cancers. *Nature*, 435(7043), 834–838.

Ma, Jin-Biao, Yuan, Yu-Ren, Meister, Gunter, Pei, Yi, Tuschl, Thomas & Patel, Dinshaw J (2005). Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. *Nature*, 434(7033), 666–670.

Margulies, Marcel, Egholm, Michael, Altman, William E, Attiya, Said, Bader, Joel S, Bemben, Lisa A, Berka, Jan, Braverman, Michael S, Chen, Yi-Ju, Chen, Zhoutao, Dewell, Scott B, Du, Lei, Fierro, Joseph M, Gomes, Xavier V, Godwin, Brian C, He, Wen, Helgesen, Scott, Ho, Chun Heen, Ho, Chun He, Irzyk, Gerard P, Jando, Szilveszter C, Alenquer, Maria L I, Jarvie, Thomas P, Jirage, Kshama B, Kim, Jong-Bum, Knight, James R, Lanza, Janna R, Leamon, John H, Lefkowitz, Steven M, Lei, Ming, Li, Jing, Lohman, Kenton L, Lu, Hong, Makhijani, Vinod B, McDade, Keith E, McKenna, Michael P, Myers, Eugene W, Nickerson, Elizabeth, Nobile, John R, Plant, Ramona, Puc, Bernard P, Ronan, Michael T, Roth, George T, Sarkis, Gary J, Simons, Jan Fredrik, Simpson, John W, Srinivasan, Maithreyan, Tartaro, Karrie R, Tomasz, Alexander, Vogt, Kari A, Volkmer, Greg A, Wang, Shally H, Wang, Yong, Weiner, Michael P, Yu, Pengguang, Begley, Richard F & Rothberg, Jonathan M (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), 376–380.

Matousek, Jaroslav, Kozlová, Petra, Orctová, Lidmila, Schmitz, Axel, Pesina, Karel, Bannach, Oliver, Diermann, Natalie, Steger, Gerhard & Riesner, Detlev (2007). Accumulation of viroid-specific small RNAs and increase in nucleolytic activities linked to viroid-caused pathogenesis. *Biol Chem*, 388(1), 1–13.

Matousek, Jaroslav, Orctová, Lidmila, Steger, Gerhard, Skopek, Josef, Moors, Michaela, Dedic, Petr & Riesner, Detlev (2004). Analysis of thermal stress-mediated PSTVd variation and biolistic inoculation of progeny of viroid "thermomutants"to tomato and Brassica species. *Virology*, 323(1), 9–23.

Michalak, P. (2006). RNA world - the dark matter of evolutionary genomics. *J Evol Biol*, 19(6), 1768–1774.

Moore, M. S. (1998). Ran and nuclear transport. *J Biol Chem*, 273(36), 22857–22860.

Moy, Terence I & Silver, Pamela A (2002). Requirements for the nuclear export of the small ribosomal subunit. *J Cell Sci*, 115(Pt 14), 2985–2995.

Mückstein, Ulrike, Tafer, Hakim, Hackermüller, Jörg, Bernhart, Stephan H, Stadler, Peter F & Hofacker, Ivo L (2006). Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10), 1177–1182.

Nam, Jin-Wu, Shin, Ki-Roo, Han, Jinju, Lee, Yoontae, Kim, V. Narry & Zhang, Byoung-Tak (2005). Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res*, 33(11), 3570–3581.

Ng, Kwang Loong Stanley & Mishra, Santosh K (2007). De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, 23(11), 1321–1330.

Nowotny, Marcin, Gaidamakov, Sergei A, Crouch, Robert J & Yang, Wei (2005). Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. *Cell*, 121(7), 1005–1016.

Ofengand, J., Malhotra, A., Remme, J., Gutgsell, N. S., Campo, M. Del, Jean-Charles, S., Peil, L. & Kaya, Y. (2001). Pseudouridines and pseudouridine synthases of the ribosome. *Cold Spring Harb Symp Quant Biol*, 66, 147–159.

Okamoto, H. & Hirochika, H. (2001). Silencing of transposable elements in plants. *Trends Plant Sci*, 6(11), 527–534.

Owens, R. A., Blackburn, M. & Ding, B. (2001). Possible involvement of the phloem lectin in long-distance viroid movement. *Mol Plant Microbe Interact*, 14(7), 905–909.

Palukaitis, P (1987). Potato spindle tube viroid: Investigation of the long-distance, intra-plant transport route. *Virology*, 158(1), 239–241.

Papaefthimiou, I., Hamilton, A., Denti, M., Baulcombe, D., Tsagris, M. & Tabler, M. (2001). Replicating potato spindle tuber viroid RNA is accompanied by short RNA fragments that are characteristic of post-transcriptional gene silencing. *Nucleic Acids Res*, 29(11), 2395–2400.

Park, Mee Yeon, Wu, Gang, Gonzalez-Sulser, Alfredo, Vaucheret, Hervé & Poethig, R. Scott (2005). Nuclear processing and export of microRNAs in Arabidopsis. *Proc Natl Acad Sci U S A*, 102(10), 3691–3696.

Parker, James S, Roe, S. Mark & Barford, David (2004). Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. *EMBO J*, 23(24), 4727–4737.

Peragine, Angela, Yoshikawa, Manabu, Wu, Gang, Albrecht, Heidi L & Poethig, R. Scott (2004). SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. *Genes Dev*, 18(19), 2368–2379.

Perkins, Neil J & Schisterman, Enrique F (2006). The inconsistency of öptimal"cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am J Epidemiol*, 163(7), 670–675.

Qi, Yijun & Ding, Biao (2003). Differential subnuclear localization of RNA strands of opposite polarity derived from an autonomously replicating viroid. *Plant Cell*, 15(11), 2566–2577.

Que, Q. & Jorgensen, R. A. (1998). Homology-based control of gene expression patterns in transgenic petunia flowers. *Dev Genet*, 22(1), 100–109.

Reichow, Steve L, Hamma, Tomoko, Ferré-D'Amaré, Adrian R & Varani, Gabriele (2007). The structure and function of small nucleolar ribonucleoproteins. *Nucleic Acids Res*, 35(5), 1452–1464.

Ritchie, William, Legendre, Matthieu & Gautheret, Daniel (2007). RNA stem-loops: to be or not to be cleaved by RNAse III. *RNA*, 13(4), 457–462.

Saumet, Anne & Lecellier, Charles-Henri (2006). Anti-viral RNA silencing: do we look like plants? *Retrovirology*, 3, 3.

Schauer, Stephen E, Jacobsen, Steven E, Meinke, David W & Ray, Animesh (2002). DICER-LIKE1: blind men and elephants in Arabidopsis development. *Trends Plant Sci*, 7(11), 487–491.

Schauser, L., Roussis, A., Stiller, J. & Stougaard, J. (1999). A plant regulator controlling development of symbiotic root nodules. *Nature*, 402(6758), 191–195.

Schubert, Steffen, Grünweller, Arnold, Erdmann, Volker A & Kurreck, Jens (2005). Local RNA target structure influences siRNA efficacy: systematic analysis of intentionally designed binding regions. *J Mol Biol*, 348(4), 883–893.

Schwarz, Dianne S, Hutvágner, György, Du, Tingting, Xu, Zuoshang, Aronin, Neil & Zamore, Phillip D (2003). Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, 115(2), 199–208.

Smit, A. F. A. Hubley, R. & Green, P. (1996-2004). RepeatMasker Open-3.0. `http://www.repeatmasker.org`.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147(1), 195–197.

Song, Ji-Joon, Liu, Jidong, Tolia, Niraj H, Schneiderman, Jonathan, Smith, Stephanie K, Martienssen, Robert A, Hannon, Gregory J & Joshua-Tor, Leemor (2003). The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. *Nat Struct Biol*, 10(12), 1026–1032.

Song, Ji-Joon, Smith, Stephanie K, Hannon, Gregory J & Joshua-Tor, Leemor (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. *Science*, 305(5689), 1434–1437.

Song, Liang, Han, Meng-Hsuan, Lesicka, Joanna & Fedoroff, Nina (2007). Arabidopsis primary microRNA processing proteins HYL1 and DCL1 define a nuclear body distinct from the Cajal body. *Proc Natl Acad Sci U S A*, 104(13), 5437–5442.

Spackman, K.A. (1989). Signal detection theory: valuable tools for evaluating inductive learning. *Proceedings of the sixth international workshop on Machine learning table of contents*, S. 160–163.

Stark-Lorenzen, P., Guitton, M. C., Werner, R. & Mühlbach, H. P. (1997). Detection and tissue distribution of potato spindle tuber viroid in infected tomato plants by tissue print hybridization. *Arch Virol*, 142(7), 1289–1296.

Steger, G. (1994). Thermal denaturation of double-stranded nucleic acids: prediction of temperatures critical for gradient gel electrophoresis and polymerase chain reaction. *Nucleic Acids Res*, 22(14), 2760–2768.

Steger, Gerhard (2003). *Bioinformatik: Methoden zur Vorhersage von RNA- und Proteinstruktur.* Birkhäuser Verlag, Basel.

Strothmann, D. (2005). *A system for the declarative description and efficient search of hybrid patterns.* PhD thesis, Universität Bielefeld.

Sunkar, Ramanjulu, Girke, Thomas, Jain, Pradeep Kumar & Zhu, Jian-Kang (2005). Cloning and characterization of microRNAs from rice. *Plant Cell*, 17(5), 1397–1411.

Sunkar, Ramanjulu & Zhu, Jian-Kang (2004). Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. *Plant Cell*, 16(8), 2001–2019.

Tabler, Martin & Tsagris, Mina (2004). Viroids: petite RNA pathogens with distinguished talents. *Trends Plant Sci*, 9(7), 339–348.

Tagami, Yuko, Inaba, Naoko, Kutsuna, Natsumaro, Kurihara, Yukio & Watanabe, Yuichiro (2007). Specific enrichment of miRNAs in Arabidopsis thaliana infected with Tobacco mosaic virus. *DNA Res*, 14(5), 227–233.

Vazquez, Franck, Vaucheret, Hervé, Rajagopalan, Ramya, Lepers, Christelle, Gasciolli, Virginie, Mallory, Allison C, Hilbert, Jean-Louis, Bartel, David P & Crété, Patrice (2004). Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. *Mol Cell*, 16(1), 69–79.

Wang, Xiu-Jie, Reyes, José L, Chua, Nam-Hai & Gaasterland, Terry (2004). Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. *Genome Biol*, 5(9), R65.

Xie, Zhixin, Allen, Edwards, Fahlgren, Noah, Calamar, Adam, Givan, Scott A & Carrington, James C (2005). Expression of Arabidopsis MIRNA genes. *Plant Physiol*, 138(4), 2145–2154.

Yan, Kelley S, Yan, Sherry, Farooq, Amjad, Han, Arnold, Zeng, Lei & Zhou, Ming-Ming (2003). Structure and conserved RNA binding of the PAZ domain. *Nature*, 426(6965), 468–474.

Yu, Bin, Yang, Zhiyong, Li, Junjie, Minakhina, Svetlana, Yang, Maocheng, Padgett, Richard W, Steward, Ruth & Chen, Xuemei (2005). Methylation as a crucial step in plant microRNA biogenesis. *Science*, 307(5711), 932–935.

Zeng, Yan, Wagner, Eric J & Cullen, Bryan R (2002). Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. *Mol Cell*, 9(6), 1327–1333.

Zhang, B. H., Pan, X. P., Cox, S. B., Cobb, G. P. & Anderson, T. A. (2006). Evidence that miRNAs are different from other RNAs. *Cell Mol Life Sci*, 63(2), 246–254.

Zhang, Yuanji (2005). miRU: an automated plant miRNA target prediction server. *Nucleic Acids Res*, 33(Web Server issue), W701–W704.

Die hier vorgelegte Dissertation habe ich eigenständig und ohne unerlaubte Hilfe angefertigt. Die Dissertation wurde in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den 2. Juli 2008

(Jan-Hendrik Teune)