In silico Vorhersage von MicroRNAs in konservierten Regionen

Thomas Blank

14. Juni 2007

Diplomarbeit zur Erlangung des akademischen Grades Diplomingenieur (FH) der Studienrichtung Bioinformatik

Betreuer:

Prof. Dr. Bernhard Haubold
FH Weihenstephan
Fakultät Biotechnologie und Bioinformatik
85350 Freising

Fachhochschule Weihenstephan
University of Applied Sciences

Dr. Matthias Scherf
Genomatix Software GmbH
Bayerstr. 85a
80335 München

Genomatix
understanding gene regulation

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass die vorliegende Arbeit von mir selbst und ohne fremde Hilfe verfasst und noch nicht anderweitig für Prüfungszwecke vorgelegt wurde. Es wurden keine anderen als die angegebenen Quellen oder Hilfsmittel benutzt. Wörtliche und sinngemäße Zitate sind als solche gekennzeichnet.

Freising, den .....

.....
Unterschrift

Danksagung

Bedanken möchte ich mich bei Herrn Prof. Dr. Bernhard Haubold und Herrn Dr. Matthias Scherf für die sehr gute Betreuung meiner Diplomarbeit. Weiterer Dank gilt Andreas Klingenhoff und Dr. Korbinian Grote für wissenschaftliche Ratschläge und das Korrekturlesen dieser Arbeit. Besonderer Dank gebührt meiner Familie für die Unterstützung während des Studiums.

Inhaltsverzeichnis

1 Zusammenfassung

MicroRNAs sind eine Klasse von kleinen RNA-Molekülen (etwa 21 Nukleotide lang), die Gene posttranskriptionell regulieren. Ähnlich wie bei small interfering RNAs, wird durch den Mechanismus der RNA-Interferenz die Translation der mRNA des Target-Gens in ein Protein verhindert. MicroRNAs kommen sowohl in Pflanzen als auch in Tieren vor. Man nimmt an, dass etwa 30% der Gene von Säugetieren durch MicroRNAs reguliert werden. Die Wichtigkeit der RNA-Interferenz als neues Prinzip der Genregulation wird durch die Verleihung des Nobelpreises für Physiologie oder Medizin an Fire und Mello, die Entdecker dieses Prinzips, deutlich. So erhofft man sich vor allem im Bereich der Medizin neue Diagnose- und Behandlungsmöglichkeiten durch die Anwendung der RNA-Interferenz.

In dieser Arbeit wurde ein in silico Ansatz zum Identifizieren noch nicht bekannter MicroRNAs etwickelt. Dabei wurde sich auf die Suche nach phylogenetisch konservierten pre-MicroRNAs konzentriert. Pre-MicroRNAs sind Vorläufersequenzen von MicroRNAs und haben, im Gegensatz zu den reifen MicroRNAs, einige markante Merkmale anhand derer man sie identifizieren kann. Die wichtigsten Merkmale sind, neben der Konservierung, die Länge, der Konservierungsgrad, die Sekundärstruktur und die zugehörige minimale freie Energie. Mit dem Ansatz dieser Arbeit wurden in den Genomen von Mensch, Maus, Hund, Rind und Zebrafisch insgesamt etwa 13.000 neue MicroRNA-Kandidaten entdeckt. Anschließend wurden Hinweise auf die Expression dieser Sequenzen mittels einer Korrelation mit Expressed Sequence Tags und Gen-Loci ermittelt. Die vorhergesagten MicroRNAs werden von der Firma Genomatix Software GmbH für die eigene Genomannotation verwendet.

2 Einleitung

MicroRNAs sind kleine RNA-Moleküle, die etwa 21 Nukleotide lang sind und eine wichtige Rolle bei der posttranskriptionellen Genregulation spielen. Man hielt sie anfänglich für einen interessanten Sonderfall der posttranskriptionellen Genregulation in Caenorhabditis elegans. Victor Ambros, Rosalind Lee und Rhonda Feinbaum hatten 1993 herausgefunden, dass ein bestimmtes Gen, das bei der Entwicklung von C. elegans während des Larvenstadiums eine wichtige Rolle spielt, nicht in ein Protein übersetzt wird. Stattdessen produziert dieses Gen (lin-4) zwei kleine RNAs [22]. Eine RNA ist etwa 21 Nukleotide lang (lin-4s), die andere etwa 61 (lin-4l). Es wurde angenommen, dass lin-4s aus lin-4l hervorgeht und lin-4l eine stem-loop-Struktur einnimmt, eine besondere Form der Sekundärstruktur[1], dargestellt in Abbildung 2.1. Außerdem wurde festgestellt, dass die lin-4s RNA komplementär zu mehreren Bereichen der UTR (untranslated region) der lin-14 mRNA ist. In einer vorangegangenen Arbeit wurde vermutet, dass diese Region eine Vermittlerrolle bei der Repression von lin-14 durch das Genprodukt von lin-4 innehat [55]. Aufgrund dieser Erkenntnisse wurde ein Modell entwickelt, in welchem die lin-4 RNAs an die lin-14 3' UTR binden und dadurch die Translation der lin-14 mRNA verhindern [22, 56]. Heute wird die kürzere lin-4 RNA (lin-4s) als erstes Mitglied einer

[1] Liegt eine RNA als Einzelstrang vor, so kann sie eine sogenannte Sekundärstruktur einnehmen. Dabei bilden komplementäre Abschnitte der RNA Wasserstoffbrücken bzw. intramolekulare Doppelstränge aus. Nicht komplementäre Abschnitte bleiben einzelsträngig.

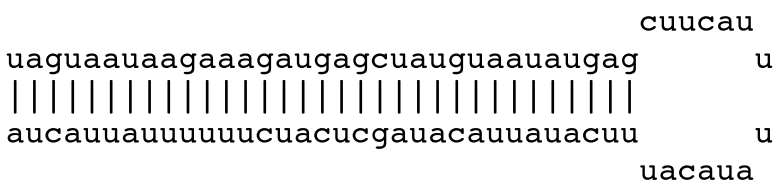


Abbildung 2.1: Darstellung einer idealen stem-loop-Struktur

Klasse von kurzen, genregulierenden RNAs, den MicroRNAs, angesehen.

In den letzten Jahren haben zahlreiche Studien gezeigt, dass die Genregulation durch MicroRNAs keineswegs eine Ausnahme, sondern eher die Regel darstellt, in Tieren als auch in Pflanzen. So spielen MicroRNAs beim Entwicklungsablauf bei Würmern, beim Zelltod und Fettmetabolismus bei Fliegen, bei der Reifung von Blutzellen bei Säugetieren und bei der Blattentwicklung von Pflanzen eine wichtige Rolle [2, 16]. Außerdem konnten kürzlich Korrelationen zwischen MicroRNA-Expression und Krankheiten beim Menschen gezeigt werden. Bestimmte MicroRNAs weisen in Tumoren eine veränderte Expression auf. Es könnte also ein direkter Zusammenhang zwischen MicroRNA-Expression und Krebserkrankungen beim Menschen bestehen [8]. Beispielsweise reguliert die MicroRNA let-7 das Onkogen[2] RAS, und let-7 ist bei Lungenkrebs häufig herunterreguliert [46]. All diese Erkenntnisse machen die Bedeutung von MicroRNAs für die Medizin deutlich. Sobald man den Mechanismus der Genregulierung durch MicroRNAs ausreichend gut verstanden hat, können neue Ansätze zur Diagnostik und Therapie von vielen Krankheiten entwickelt werden.

2.1 Lebenszyklus einer MicroRNA

Primärtranskripte von MicroRNA-Genen (pri-MicroRNAs) sind für gewöhnlich mehrere Kilobasen lang und enthalten eine stem-loop-Struktur. Es wird angenommen, dass die Transkription von MicroRNA-Genen durch die RNA-Polymerase II gesteuert wird. Hierfür gibt es verschiedene Hinweise. Primärtranskripte von MicroRNAs besitzen eine sogenannte m^7G-CAP-Struktur (diese Struktur kommt am 5' Ende von eukariontischen mRNAs vor) und einen Poly(A)-Schwanz. Dies sind typische Merkmale einer RNA-Polymerase II Transkription [53, 7]. Die Transkription von MicroRNAs reagiert empfindlich auf eine bestimmte Konzentration von α-amanitin, bei der RNA-Polymerase II blockiert wird, nicht aber RNA-Polymerase I oder RNA-Polymerase III [53], ein weiteres Indiz für eine RNA-Polymerase II Transkription. Durch eine Chromatin-Immunopräzipitationsanalyse konnte außerdem ein physischer Zusammenhang zwischen den Promotoren einiger MicroRNAs und RNA-Polymerase II nachgewiesen werden [53]. Anschließend an die Transkription, wird die stem-loop-Struktur durch das Protein Drosha (Typ III RNase) von der pri-MicroRNA abgetrennt [52]. Es entsteht

[2]Onkogene sind Gene, die an der Entstehung von Krebs beteiligt sind.

ein sogenannter MicroRNA-precursor (pre-MicroRNA), wie in Abbildung 2.2 dargestellt. Pre-MicroRNAs sind etwa 65 bis 115 Nukleotide lang. Nach der Abtrennung durch Drosha wird die pre-MicroRNA vom Zellkern ins Zytoplasma transportiert. Dieser Vorgang wird durch den Transportfaktor Exportin-5 reguliert [27, 58, 4].

Im Zytoplasma angekommen, werden die pre-MicroRNAs durch Dicer (Typ III RNase) weiter verarbeitet [3, 32, 38, 45, 18]. Es entstehen etwa 21 Nukleotide lange, doppelsträngige RNAs. Diese doppelsträngigen RNAs bleiben in der Regel nicht lange bestehen. Üblicherweise wird ein Strang abgebaut, der andere bleibt als reife MicroRNA übrig. Die Selektion eines Stranges als reife MicroRNA wird durch die relative thermodynamische Stabilität der beiden Enden der doppelsträngigen RNA bewerkstelligt. Der Strang, dessen 5'-Ende auf der thermodynamisch instabileren Seite ist, bleibt meist bestehen [37, 15]. Diese, nun reife, MicroRNA wird vom sogenannten miRISC-Komplex (microRNA-containing RNA-induced silencing complex) aufgenommen und führt diesen zur Target-mRNA. Nur sechs oder sieben Nukleotide von den etwa 21 Nukleotiden einer MicroRNA sind hauptverantwortlich für die Bindungsspezifität zwischen dem miRISC-Komplex und der Target-mRNA [11, 33, 6, 19]. Diese sechs (bzw. sieben) Nukleotide liegen auf den Positionen zwei bis sieben (bzw. acht) vom 5'-Ende der MicroRNA her gesehen und werden als sogenannte „seed sequence“ bezeichnet. Vom ersten Nukleotid auf der 5'-Seite der MicroRNA hingegen wird angenommen, dass es nicht mit dem Target bindet [14, 41, 30].

Sobald der miRISC-Komplex an eine Target-mRNA gebunden hat, gibt es zwei Möglichkeiten zur Verhinderung der Translation. Welche der Möglichkeiten stattfindet, hängt davon ab, wie gut der Rest der MicroRNA an das Target bindet. Dies ist in Abbildung 2.3 dargestellt. Binden die meisten Nukleotide der MicroRNA an die Target-mRNA, so bildet sich in diesem Bereich die Form einer α-Helix aus. Diese Struktur wiederum wird von einem Argonaute-Protein erkannt. Dadurch wird der Abbau der Target-mRNA eingeleitet [9, 28]. Findet die Bindung zwischen MicroRNA und Target-mRNA nur partiell statt, so entsteht keine α-Helix und die mRNA wird nicht direkt abgebaut. Stattdessen wird die Translation der mRNA in ein Protein blockiert [29, 10]. Im weiteren Verlauf wird der gesamte Komplex von miRISC und mRNA in einen bestimmten Bereich des Zytoplasmas transportiert, den sogenannten P-Body. Dort angekommen, wird die mRNA schließlich abgebaut [25, 31]. Bei beiden Fällen der translationalen Repression von mRNAs, wird nach dem Abbau der mRNAs der miRISC-Komplex wieder frei. Dadurch kann ein miRISC-Komplex bzw. eine MicroRNA mehrere potentielle Transla-

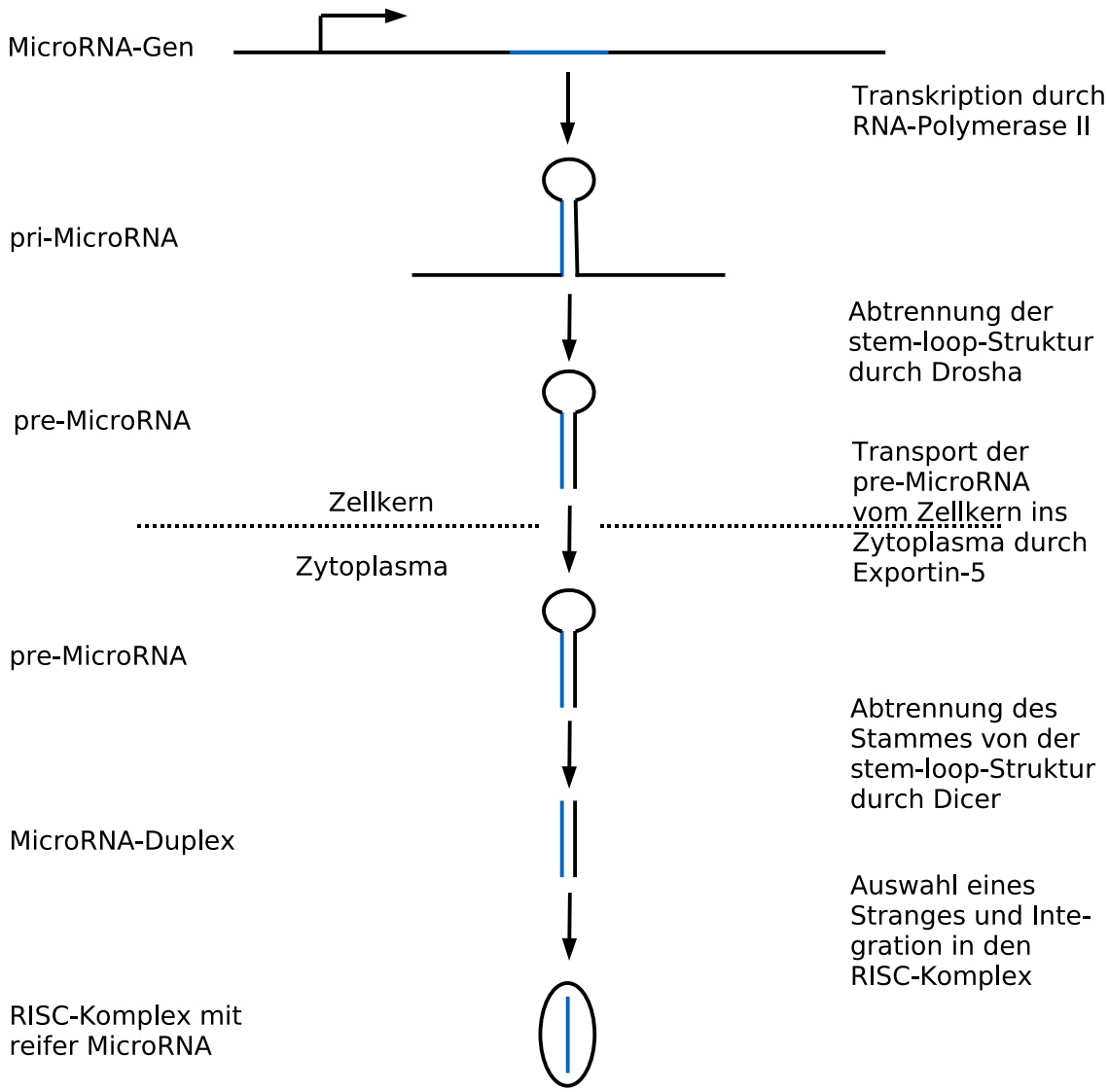


Abbildung 2.2: Modell des MicroRNA-Lebenszyklus. MicroRNA-Gene werden durch RNA-Polymerase II transkribiert, wodurch ein mehrere Kilobasen langes Primärtranskript entsteht (pri-MicroRNA). Durch Drosha (Typ III RNase) wird die stem-loop-Struktur (pre-MicroRNA) vom Primärtranskript abgetrennt. Der Transportfaktor Exportin-5 erkennt die pre-MicroRNA und transportiert diese vom Zellkern ins Zytoplasma. Im Zytoplasma angelangt, wird die pre-MicroRNA von Dicer (Typ III RNase) erkannt. Dicer trennt den Stamm der pre-MicroRNA ab und es entsteht ein etwa 21 Nukleotide langer MicroRNA-Duplex. Der MicroRNA-Duplex wird aufgetrennt und ein Strang wird als reife MicroRNA in den RISC-Komplex integriert.

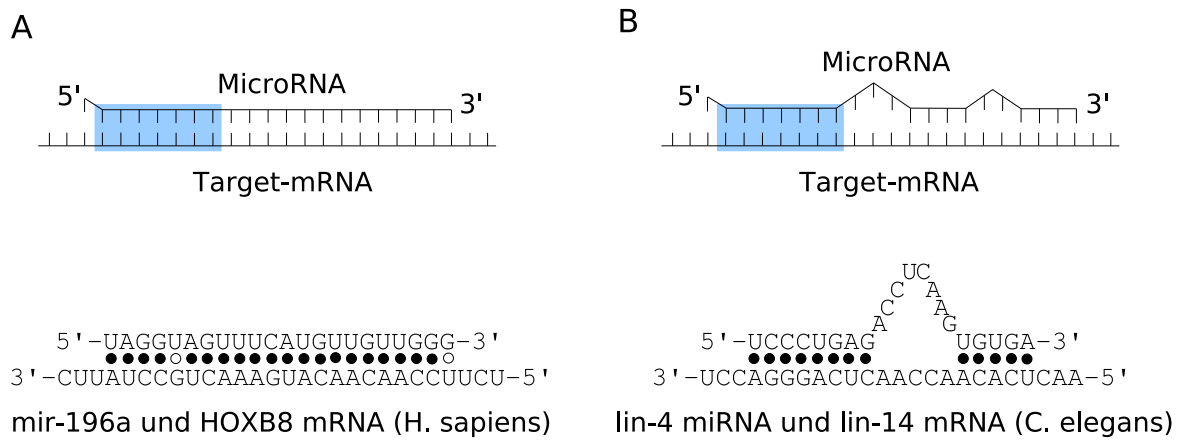


Abbildung 2.3: Bindungsarten zwischen MicroRNA und Target-mRNA.

tionsvorgänge verhindern.

2.2 Definition einer MicroRNA

Damit eine kurze RNA der Klasse der MicroRNAs zugeordnet werden kann, sollten laut Ambros [48] folgende Kriterien erfüllt sein:

- Vorhersage einer potentiellen stem-loop-Struktur, wobei ein Strang des Stammes die reife MicroRNA enthält. In Tieren ist diese stem-loop-Struktur etwa 80 Nukleotide lang, in Pflanzen dagegen variiert die Länge stark. Der Stamm der stem-loop-Struktur sollte keine großen Schleifen aufweisen.
- Phylogenetische Konservierung der reifen MicroRNA und der zugehörigen stem-loop-Sequenz in nah verwandten Organismen. Dieses Kriterium trifft häufig zu, aber es gibt auch einige MicroRNAs, die nicht phylogenetisch konserviert sind.
- Erhöhung der pre-MicroRNA-Konzentration in Organismen mit reduzierter Dicer-Funktion.
- Nachweis der reifen MicroRNA durch eine Northern Blot-Analyse, oder mit Hilfe einer RNA-Bibliothek.

2.3 Möglichkeiten der MicroRNA-Identifizierung

Alle unten aufgeführten Möglichkeiten, MicroRNAs zu identifizieren, basieren auf den Definitionen von Abschnitt 2.2 und können in zwei Gruppen eingeteilt werden. Bei experimentellen Methoden wird zuerst nach kleinen, exprimierten RNAs gesucht, die anschließend mit bioinformatischen Methoden auf bestimmte strukturelle Eigenschaften, die für MicroRNAs typisch sind, getestet werden. Der zweite Ansatz besteht darin, zuerst mit bioinformatischen Methoden Genomsequenzen auf die entsprechenden strukturellen Eigenschaften hin zu untersuchen und dann die so gefundenen MicroRNA-Kandidaten durch Expressionsanalysen zu bestätigen.

2.3.1 Identifizierung mittels RNA-Bibliotheken

Der bevorzugte Ansatz bei der de novo Identifizierung von MicroRNAs ist die Sequenzierung von kurzen cDNAs. Eine RNA-Probe wird mittels eines denaturierenden Polyacrylamid-Gels aufgetrennt. Die Fraktion mit einer Länge von 20-25 Nukleotiden wird extrahiert. An diese RNAs werden 5'- und 3'-Adapter ligiert, anschließend wird eine RT-PCR durchgeführt. Die Fragmente werden in Vektoren eingebracht und eine cDNA-Bibliothek wird erstellt. Dann werden individuelle Klone sequenziert und mit Hilfe der Bioinformatik wird der Ursprung der Sequenzen im Genom festgestellt. Der größte Anteil (gewöhnlich weit über 50%) der klonierten Sequenzen stammt von Fragmenten von größeren RNAs, tRNAs oder rRNAs. Die klonierten Sequenzen müssen also noch anhand der Kriterien von Abschnitt 2.2 klassifiziert werden, um entscheiden zu können, ob es sich um eine MicroRNA handelt oder nicht. Es gibt allerdings auch Limitierungen für diese Methode. Es ist beispielsweise schwierig, MicroRNAs zu finden, die eine relativ geringe Expression aufweisen, oder nur in seltenen Zelltypen exprimiert werden. Manche MicroRNAs sind außerdem wegen ihrer Sequenzzusammensetzung, oder aufgrund von posttranskriptionellen Modifikationen nur schwer zu klonieren [26, 49, 57].

2.3.2 Identifizierung durch forward genetics

Forward genetics bezeichnet einen Ansatz zur Identifizierung von Genen, ausgehend vom Phänotyp. Durch zufällige Mutagenese werden bei einer Gruppe von Individuen Mutanten erzeugt. Weist eine dieser Mutanten eine interessante Eigenschaft auf, so wird versucht, die Mutation im Genom der Mutante zu lokalisieren. Findet man die Mutati-

on, kann man gleichzeitig auch die Position des Gens, das von der Mutation betroffen ist, feststellen. Durch diese Methode wurden die ersten MicroRNAs, z. B. wie oben beschrieben lin-4, gefunden. Allerdings ist der Anteil der so gefundenen MicroRNAs relativ gering. Einer der Gründe hierfür ist die geringe Größe der MicroRNAs und ihre Toleranz gegenüber Mutationen, die nicht in der sogenannten „seed sequence“ liegen (die seed sequence spielt die Hauptrolle bei der Bindung der MicroRNA an ihr Target). Dadurch ist es sehr schwierig, ein MicroRNA-Gen durch Mutagenese zu erkennen. Selbst wenn eine MicroRNA durch eine Mutation ausgeschaltet wurde, kommt es vor, dass sie nicht gefunden wird, da sich die Bemühungen, die Mutation zu lokalisieren, meist auf proteinkodierende Regionen konzentrieren.

2.3.3 In silico Identifizierung

Nachdem genügend MicroRNAs durch die zwei vorgenannten Methoden gefunden wurden, um allgemeine Eigenschaften abzuleiten, konnten verschiedene in silico Methoden zur MicroRNA-Identifizierung entwickelt werden. Die ersten Methoden und Programme basierten hauptsächlich auf dem Kriterium der phylogenetischen Konservierung. Das Programm MirScan [43] beispielsweise identifiziert konservierte stem-loop-Strukturen und klassifiziert diese anhand ihrer Ähnlichkeit mit stem-loop-Strukturen von experimentell verifizierten MicroRNAs. Mit diesem Ansatz wurden 35 neue MicroRNA-Kandidaten in C. elegans [43] und 107 im Menschen [24] gefunden. Ein weiteres Programm, das sich auf die Konservierung der MicroRNAs stützt, ist miRSeeker [20]. Mit miRSeeker wurden 48 MicroRNA-Kandidaten in D. melanogaster gefunden [20].

Einen etwas anderen, auch auf Konservierung basierenden Ansatz, haben Xie u. a. gewählt. Anstatt der Konservierung von stem-loop-Strukturen, wurde hier die Konservierung von potentiellen Target-Sequenzen ausgenutzt. Xie u. a. analysierten konservierte Sequenzmotive in der 3'-UTR von Genen. Sie fanden heraus, dass viele Sequenzmotive mit Komplementen der seed-Sequenzen von MicroRNAs zusammen passen. Durch die Verwendung von Sequenzmotiven, die nicht auf bekannte MicroRNAs passen, konnten 129 neue MicroRNA-Kandidaten im Menschen gefunden werden [51]. Andere Methoden versuchen homologe Sequenzen zu bekannten MicroRNAs zu finden, indem Sequenzalignments und Strukturvergleiche angestellt werden [23, 42, 50].

Die thermodynamische Stabilität der Sekundärstruktur von pre-MicroRNAs ist ein weiteres, häufig verwendetes Kriterium, um MicroRNAs von anderen stem-loop-

Strukturen zu unterscheiden. Bonnet u. a. haben gezeigt, dass pre-MicroRNAs, im Gegensatz zu tRNAs und rRNAs, eine signifikant niedrigere freie Faltungsenergie als Zufallssequenzen haben [5]. Eine Software, die diese Tatsache ausnutzt, ist RNAz [54]. RNAz kombiniert thermodynamische Stabilität und die Konservierung der Sekundärstruktur, um RNAs zu finden, die nicht für ein Protein kodieren. In der Arbeit von Hsu u. a. [44] wurde RNAz erfolgreich zum Finden neuer MicroRNA-Kandidaten eingesetzt.

Methoden die auf dem Kriterium der phylogenetischen Konservierung aufbauen, haben den Nachteil, dass sie nur konservierte MicroRNAs detektieren. Um dieses Problem zu umgehen, wurden Methoden entwickelt, die sich komplett auf strukturelle Eigenschaften von MicroRNAs stützen [39, 34, 36]. Jede dieser Methoden verwendet Klassifikatoren, die bestimmen, wie ähnlich ein MicroRNA-Kanditat zu bekannten MicroRNAs ist. Verwendete Merkmale sind die freie Faltungsenergie, die durchschnittliche Größe von internen Schleifen, die Nukleotidzusammensetzung und andere. Viele nicht konservierte MicroRNAs wurden bereits mit diesen Methoden in Viren [47] und im Menschen [39] identifiziert.

Eine weitere Möglichkeit, neue MicroRNAs zu finden, besteht darin, in der Nähe von bereits bekannten MicroRNAs zu suchen, da viele MicroRNAs geclustert auftreten [21].

MicroRNAs, die durch in silico Methoden gefunden werden, müssen anschließend noch validiert werden. Dies kann durch die Demonstration der Expression der reifen MicroRNA erreicht werden.

2.4 Über diese Arbeit

Wie aus Abschnitt 2.3.3 ersichtlich, gibt es keinen allgemeingültigen in silico Ansatz, um neue MicroRNA-Kandidaten zu identifizieren, sondern viele unterschiedliche, jeweils angepasst an unterschiedliche Situationen. Auch der in dieser Arbeit verfolgte Ansatz, neue MicroRNAs vorherzusagen, ist ein solcher Spezialfall. Wie bei einigen bereits erwähnten Methoden, soll nach phylogenetisch konservierten MicroRNAs gesucht werden. Dies schließt alle nicht konservierten MicroRNAs von vorneherein aus. Um weitere Kriterien bei der Suche nach MicroRNAs einsetzen zu können, wurden verschiedene Eigenschaften bereits bekannter, konservierter MicroRNAs ermittelt. Es wurde geprüft, in wievielen von den fünf verwendeten Organismen (Mensch, Maus, Hund, Rind, Zebrafisch) diese MicroRNAs konserviert sind und inwiefern sie mit bekannten Gen-Loci

korrelieren. Weiterhin wurde untersucht, ob nur die MicroRNAs konserviert sind, oder ob sie in längeren konservierten Regionen liegen. Schließlich wurde die Sekundärstruktur und die zugehörige minimale freie Energie der konservierten pre-MicroRNAs ermittelt. Mit all diesen Informationen wurden dann die konservierten Regionen der fünf Organismen analysiert. Am Ende steht dann eine Auswahl von konservierten Regionen, die MicroRNA-Kandidaten enthalten.

Der in dieser Arbeit verwendete Ansatz, MicroRNAs vorherzusagen, hat teilweise Gemeinsamkeiten mit den in Abschnitt 2.3.3 dargestellten Methoden. So wurden beispielsweise die phylogenetische Konservierung und die Sekundärstruktur auch in den Programmen MirScan [43] und miRSeeker [20] verwendet. Im Gegensatz zu dieser Arbeit wurde dort allerdings mit den Genomen von C. elegans und C. briggsae beziehungsweise D. melanogaster und D. pseudoobscura gearbeitet. Außerdem wurden bei MirScan und miRSeeker von vornherein MicroRNA-Kandidaten ausgeschlossen, die in Exons liegen. Mittlerweile gibt es aber eine Reihe von experimentell verifizierten MicroRNAs, die in Exons liegen. Eine weitere Methode, die Parallelen zu dieser Arbeit aufweist, ist in [44] beschrieben. Auch hier wurden die Genome verschiedener Vertebraten nach neuen MicroRNAs durchsucht, allerdings nicht die Genome von Rind und Zebrafisch. Eine Vorauswahl der konservierten Regionen anhand der Länge, vor dem Schritt der Sekundärstrukturanalyse, wurde ebenfalls bei keiner der drei oben erwähnten Methoden vorgenommen. In dieser Arbeit wurde dadurch eine Zeitersparnis bei der Sekundärstrukturanalyse erreicht. Auch ein Abgleich der gefundenen MicroRNA-Kandidaten mit Expressed Sequence Tags (ESTs) wurde bei den drei oben genannten Arbeiten nicht durchgeführt. Die Lage eines MicroRNA-Kandidaten innerhalb eines ESTs ist ein deutlicher Hinweis auf die Expression dieses MicroRNA-Kandidaten.

3 Material und Methoden

3.1 Konservierte Regionen

Im Zuge dieser Arbeit wurden paarweise konservierte Regionen zwischen den genomischen Sequenzen von Mensch, Maus, Rind, Hund und Zebrafisch verwendet. Der Vorwärtsstrang eines jeden Genoms wurde jeweils mit Vorwärtsstrang und Rückwärtsstrang aller anderen Genome verglichen. Konservierte Regionen wurden als Sequenzen von mindestens 25 Basenpaaren Länge definiert, die eine Ähnlichkeit von mindestens 80% zwischen zwei Organismen bezüglich ihrer Nukleotidanordnung aufweisen. Gaps/Inserts wurden nicht zugelassen. Die Entscheidung, 80% anstatt der sonst üblichen 100% Sequenzähnlichkeit zu fordern, beruht auf der Tatsache, dass die stem-loop-Sequenzen der MicroRNAs, trotz vereinzelter Punktmutationen, ihre Funktion beibehalten können. Der Datensatz der konservierten Regionen war bereits in Form von Flat-Files bei Genomatix vorhanden. Aufgrund des verwendeten Alignment-Algorithmus können konservierte Regionen, die sich gegenseitig überlappen, nicht ausgeschlossen werden. Für diese Arbeit wurden die sich überlappenden Regionen mittels Perl-Skripten zusammengefasst. Zur besseren Verwendbarkeit für die verschiedenen Analysen wurden die konservierten Regionen anschließend in ein relationales Datenbanksystem eingepflegt. Hierfür wurde MySQL in der Version 5.1 verwendet. In Tabelle 3.1 sind die Anzahlen der zwischen den Organismen konservierten Regionen aufgelistet.

Tabelle 3.1: Anzahlen der zwischen den Organismen konservierten Regionen.

	Mensch	Maus	Rind	Hund	Zebrafisch
Mensch (NCBI Build 36)	-	1.489.981	4.236.997	8.154.936	141.245
Maus (NCBI Build 36)	-	-	571.424	1.261.759	153.100
Rind (NCBI Build 2)	-	-	-	4.780.079	87.692
Hund (NCBI Build 2)	-	-	-	-	163.182
Zebrafisch (NCBI Build 1)	-	-	-	-	-

3.2 MicroRNAs

Ein Ziel dieser Arbeit ist es, Eigenschaften von MicroRNAs zu finden, die für die Identifizierung neuer MicroRNAs nützlich sind. Um solche Eigenschaften zu finden, benötigt man einen Datensatz bereits bekannter MicroRNAs. Dieser Datensatz sollte möglichst groß sein, da die gefundenen Eigenschaften dann als repräsentativ für alle MicroRNAs angenommen werden können. Die momentan größte und frei zugängliche MicroRNA-Datenbank ist miRBase vom Sanger-Institut [1, 13, 12]. Diese Datenbank enthält für eine Vielzahl von Organismen die Sequenzen reifer MicroRNAs, vorhergesagte stem-loop-Sequenzen und die jeweiligen MicroRNA-Targets. Jede MicroRNA, die dort aufgenommen wird, erhält einen eindeutigen Zugriffsschlüssel (z. B. MIMAT0000001) und einen Namen (z. B. cel-let-7). Informationen, wie genomische Position, die Sequenz selbst, Art der Identifizierung und Verifizierung und weitere Annotationsdaten sind in der miRBase ebenfalls enthalten. Für diese Arbeit wurden die vorhergesagten stem-loop-Sequenzen von Mensch, Maus, Hund, Rind und Zebrafisch verwendet. Bevor mit diesen Sequenzen gearbeitet wurde, mussten sie noch zwei Verarbeitungsschritte durchlaufen. Zur Ermittlung der konservierten Regionen wurden die jeweils neuesten Genomversionen vom NCBI (National Center for Biotechnology Information) verwendet. Da sich die MicroRNAs in der miRBase teilweise auf andere (ältere) Genomversionen beziehen, mussten die Positionen der MicroRNAs in Bezug auf die aktuellen Genomversionen vom NCBI neu bestimmt werden. Dies geschah durch Alignments der pre-MicroRNA-Sequenzen mit den aktuellen Genomversionen. Dadurch wurde gewährleistet, dass die Annotation von konservierten Regionen und MicroRNAs übereinstimmt. Außerdem kann es vorkommen, dass eine MicroRNA, die durch ein Experiment gefunden wurde, beim Bestimmen der genomischen Position auf mehrere Positionen gleich gut passt. Ein Beispiel hierfür ist mmu-mir-467a, eine MicroRNA, die im Hodengewebe von Mäusen vorkommt [59]. Die Sequenz von mmu-mir-467a passt exakt auf mehrere Stellen des Mäusegenoms, kann also nicht eindeutig einer Position zugeordnet werden. Es ist nicht bekannt, ob alle oder nur ein Teil dieser genomischen Positionen der Ursprung dieser MicroRNA sind. Um die Analysen, die mit den MicroRNA-Daten von Sanger durchgeführt werden sollen, nicht zu verfälschen, wurden nur MicroRNAs verwendet, die exakt einer Stelle im Genom zugeordnet werden können. Die Anzahl der MicroRNAs, die in der miRBase enthalten sind und wieviele davon für diese Arbeit verwendet wurden, ist in Tabelle 3.2 aufgelistet. Der bereinigte Datensatz wurde anschließend, analog zu den konservierten Regionen, in

Tabelle 3.2: Gegenüberstellung der Anzahlen der MicroRNAs in der miRBase und der MicroRNAs, die eindeutig einer Stelle im Genom zugeordnet werden konnten. Stand: August 2006.

	Mensch	Maus	Rind	Hund	Zebrafisch	Gesamt
MicroRNAs in miRBase	462	358	98	6	337	1.261
MicroRNAs mit eindeutigen Positionen	454	344	73	6	152	1.029

einer MySQL-Datenbank abgespeichert.

3.3 Korrelationsanalysen

3.3.1 Bekannte pre-MicroRNAs und konservierte Regionen

Für die Korrelationsanalyse von bekannten pre-MicroRNAs und konservierten Regionen wurden einfache Positionsvergleiche vorgenommen und die Ergebnisse in drei Klassen aufgeteilt. In Abbildung 3.1 sind die verschiedenen Klassen schematisch dargestellt. Klasse 1 enthält pre-MicroRNAs, die komplett in einer konservierten Region enthalten sind. Klasse 2 umfasst alle pre-MicroRNAs, die eine Überlappung um mindestens ein Nukleotid mit einer konservierten Region aufweisen. Klasse 3 beinhaltet pre-MicroRNAs, die einen Abstand von maximal 1000 Nukleotiden zu einer konservierten Region aufweisen. Keine Klasse enthält pre-MicroRNAs einer anderen Klasse. Um zu überprüfen, ob sich das Vorkommen von pre-MicroRNAs in oder in der Nachbarschaft von konservierten Regionen signifikant von zufälligen Sequenzen unterscheidet, wurde eine Simulation durchgeführt. Zur Simulation der Positionsvergleiche wurden für jeden in dieser Arbeit betrachteten Organismus zufällige pre-MicroRNAs erzeugt. Zufällig heißt, dass sie eine zufällig ausgewählte Anfangsposition erhalten und die Länge beliebig aus dem Bereich von 65 bis 115 Nukleotiden ausgewählt wird. Die Anzahl der so erzeugten „Zufalls-pre-MicroRNAs“ pro Organismus entspricht der Anzahl der echten pre-MicroRNAs, die für diese, oben beschriebenen, Positionsvergleiche verwendet wurden. Mit diesen „Zufalls-pre-MicroRNAs“ wurden anschließend ebenfalls Positionsvergleiche mit den konservierten Regionen durchgeführt. Diese Simulation wurde insgesamt 100 mal wiederholt. Anschließend wurden Erwartungswerte und Varianzen für die verschiedenen Klassen berechnet. Um die Unterschiede zwischen echten und „Zufalls-pre-MicroRNAs“, bezüglich ihrer Lage zu konservierten Regionen auf Signifikanz zu testen, wurde schließlich die

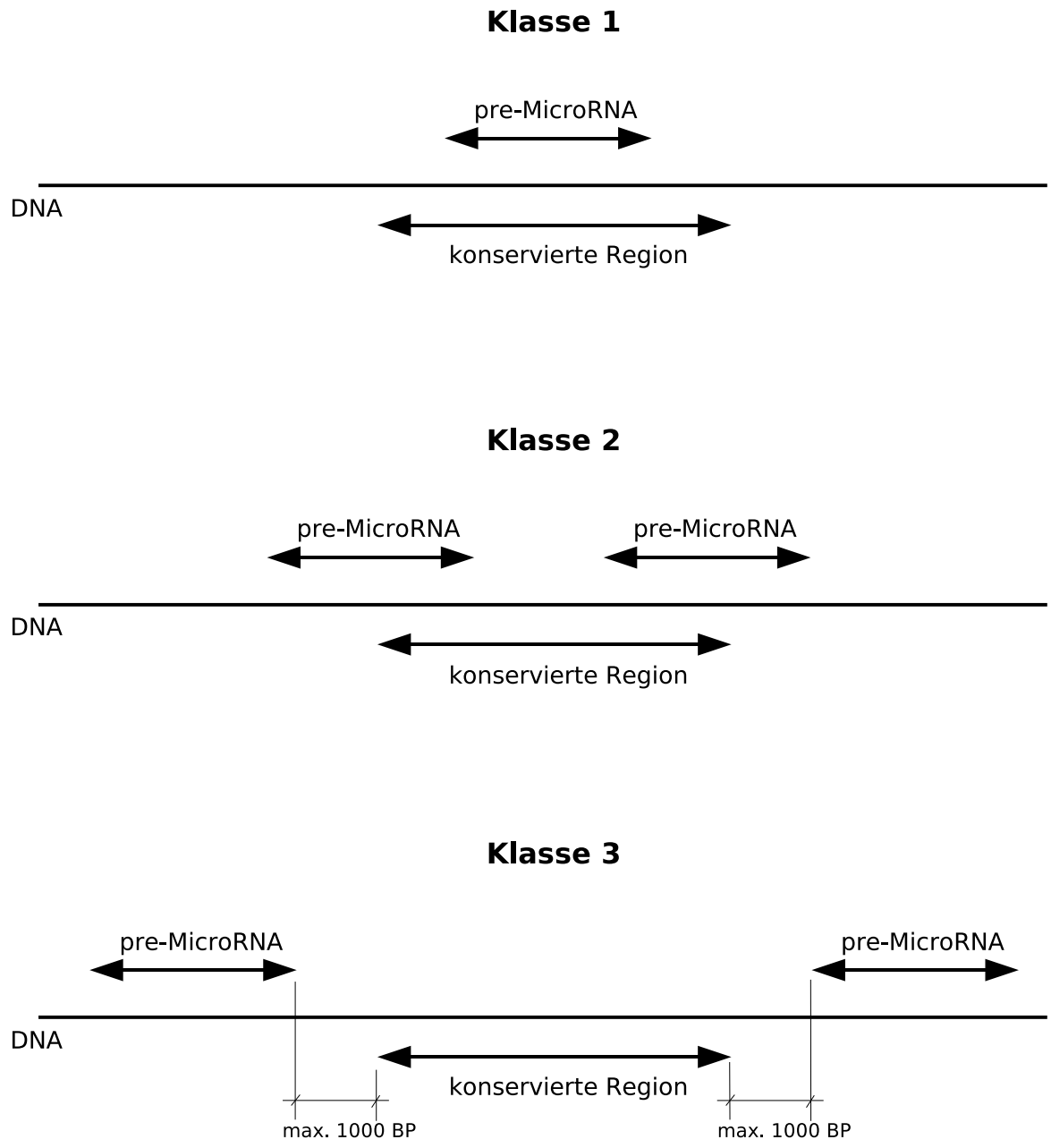


Abbildung 3.1: Schematische Darstellung der drei Klassen von pre-MicroRNAs.

Tabelle 3.3: Struktur der konservierten Regionen in der Datenbank. Eine Zeile steht für eine Sequenz in einem bestimmten Organismus. Über die contig_id lässt sich der Organismus und das Chromosom einer Sequenz ermitteln. Die Anfangsposition und Endposition einer Sequenz sind in den Spalten posfrom und posto notiert. Zusammengehörende Sequenzen sind über die pair_id verknüpft.

pair_id	contig_id	posfrom	posto
1	658.618	56.210	56.310
1	698.953	110.510	110.610
2	658.618	67.415	67.615
2	698.953	110.510	110.710
	.		
	.		
	.		

Tschebyschev-Ungleichung angewendet.

3.3.2 Konservierungsgrad

Der Konservierungsgrad gibt die Anzahl der Organismen an, in denen die betreffende Sequenz vorkommt. Die konservierten Regionen sind als Sequenzpaare in der DB abgelegt. Das heißt, eine Sequenz von „Organismus 1“ ist über eine id („pair_id“) mit einer Sequenz von „Organismus 2“ verknüpft. Tabelle 3.3 auf Seite 15 veranschaulicht dies. Die Information, in wievielen Organismen eine Sequenz konserviert ist, ist nicht direkt zugänglich. Zum Bestimmen des Konservierungsgrades einer Sequenz werden zuerst alle Zeilen ermittelt, in denen die gesuchte Sequenz vorkommt. Dies geschieht durch Abgleich der genomischen Possitionen, bezogen auf den jeweiligen Organismus (contig_id, posfrom und posto müssen übereinstimmen). Die Anzahl der so gefundenen Zeilen kann größer sein als die Anzahl der verwendeten Organismen, da eine Sequenz von „Organismus 1“ durchaus an verschiedenen Stellen in „Organismus 2“ vorkommen kann. Mehrfaches Vorkommen einer Sequenz in einem Organismus wird beim Konservierungsgrad allerdings nicht berücksichtig.

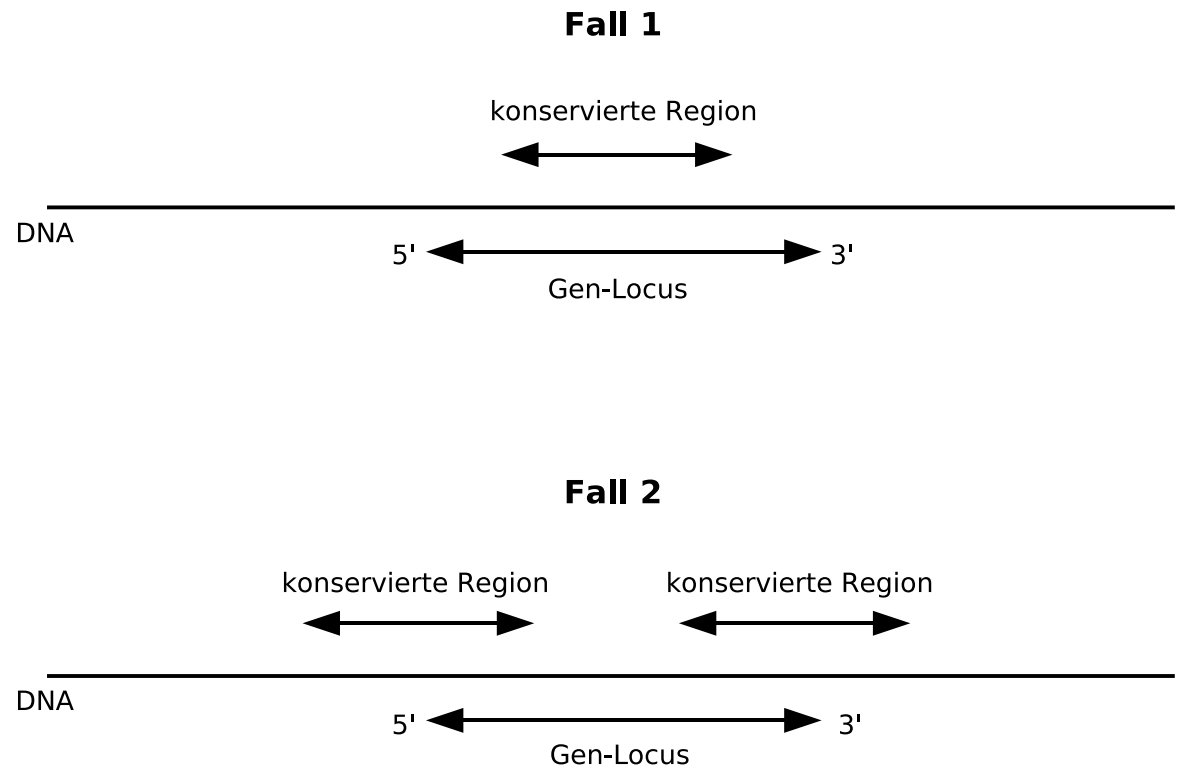


Abbildung 3.2: Schematische Darstellung der Positionsvergleiche zwischen konservierten Regionen und Gen-Loci.

3.3.3 Konservierte Regionen und Gen-Loci

Für die Korrelation von konservierten Regionen und Gen-Loci wurden einfache Positionsvergleiche vorgenommen. Die Einteilung der konservierten Regionen in die bereits bekannten 3 Klassen wurde ebenfalls beibehalten. Anfangs- und Endposition der konservierten Regionen wurden mit den Anfangs- und Endpositionen von bekannten Gen-Loci verglichen. Dabei wurden wiederum drei Fälle unterschieden. Eine schematische Darstellung der ersten beiden Fälle ist in Abbildung 3.2 zu sehen. Der erste Fall beinhaltet konservierte Regionen, die komplett innerhalb eines Gen-Locus liegen, d. h. die Anfangsposition einer konservierten Region ist größer oder gleich der Anfangsposition eines Gen-Locus und die Endposition der konservierten Region ist kleiner oder gleich der Endposition des Gen-Locus. Der zweite Fall umfasst konservierte Regionen, die eine Überlappung mit einem Gen-Locus aufweisen. Hier muss die Anfangsposition einer konservierten Region kleiner als die Anfangsposition eines Gen-Locus sein und die End-

position der konservierten Region zwischen Anfangs- und Endposition des Gen-Locus liegen. In diesem Fall hat die konservierte Region eine Überlappung mit dem 5'-Ende des Gen-Locus, analog dazu ist auch eine Überlappung mit dem 3'-Ende möglich. Beim dritten Fall gibt es keine Übereinstimmung, die konservierte Region liegt also zwischen zwei Gen-Loci. Wenn eine konservierte Region komplett in einem Gen-Locus enthalten war, so wurde zusätzlich noch eine Korrelation zwischen der konservierten Region und den Exons (bzw. Introns) des Gen-Locus vorgenommen. Wurde festgestellt, dass die konservierte Region komplett in einem Exon liegt, so wurde überprüft, ob es sich bei dem betreffenden Gen um ein Ein-Exon-Transkript handelt. Diese Information ist interessant, da Ein-Exon-Transkripte zwar exprimiert werden, häufig aber nicht für Proteine kodieren (intern kommuniziert durch Andreas Klingenhoff, Genomatix Software GmbH). Befindet sich eine konservierte Region der Klasse 1 (enthält eine bekannte MicroRNA) in einem Ein-Exon-Transkript, so kann man davon ausgehen, dass das Produkt diese Ein-Exon-Transkriptes kein Protein, sondern eine MicroRNA ist.

3.4 Ermittlung von Sekundärstruktur und minimaler freier Energie

Zur Auswertung der konservierten Regionen bezüglich ihrer Sekundärstruktur wurde das Programm RNAfold aus dem Vienna RNA Package verwendet [40]. RNAfold errechnet die Struktur mit minimaler freier Energie (MFE) einer Sequenz und gibt einen MFE-Wert sowie die Struktur in Bracket-Notation zurück. Alle konservierten Regionen mit einer Länge von 75 bis 125 Nukleotiden wurden mit RNAfold analysiert. Die Entscheidung, ob eine konservierte Region möglicherweise einen MicroRNA-Precursor enthält, sollte nun anhand des MFE-Wertes geschehen. Vorher muss man allerdings einen sinnvollen cut-off-Wert festlegen, der bestimmt, ob eine konservierte Region zu den Kandidaten für einen MicroRNA-Precursor zählt oder nicht. Hierfür wurde eine zweite Analyse mit konservierten Regionen, die eine bereits bekannte MicroRNA enthalten, durchgeführt. Durch den Vergleich der Erwartungswerte für die MFE der beiden Datensätze wurde dann ein cut-off-Wert bestimmt. Da andere RNAs aber durchaus auch eine Sekundärstruktur einnehmen können, welche einen mit MicroRNAs vergleichbaren MFE-Wert hat, aber keine stem-loop-Struktur ist, reicht das Kriterium der minimalen freien Energie allein nicht aus, um MicroRNAs vorhersagen zu können. Deshalb wurde die von RNAfold

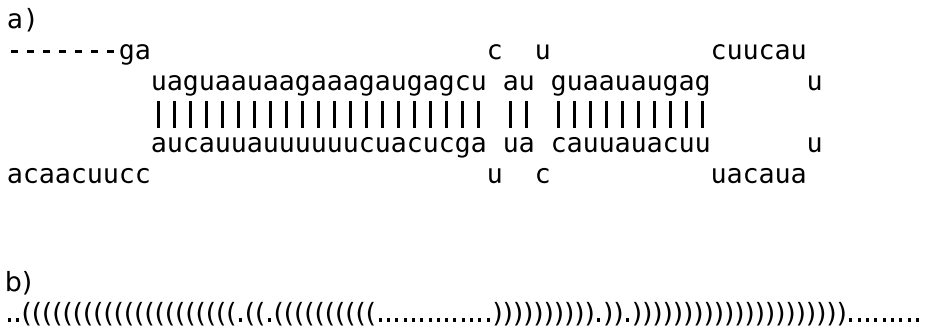


Abbildung 3.3: Darstellung der typischen Struktur eines MicroRNA-Precursors (a) und der entsprechenden Bracket-Notation (b).

bestimmte Sekundärstruktur selbst als weiteres Kriterium herangezogen. Diese wird von RNAfold in Form einer Bracket-Notation ausgegeben. Abbildung 3.3a zeigt eine stem-loop-Struktur, wie sie typischerweise bei MicroRNA-Precursorn vorkommt. Abbildung 3.3b zeigt die dazugehörige Bracket-Notation. Diese Notation besteht aus zusammengehörenden Klammern und Punkten. Ein Basenpaar bestehend aus einer Base an der Stelle i und einer Base and der Stelle j, wird durch eine öffnende Klammer an Position i und eine schließende Klammer an Position j repräsentiert. Ungepaarte Basen werden durch einen Punkt dargestellt. Durch einen einfachen Test dieser Notation könnnen Sequenzen, die nicht in das Muster einer stem-loop-Struktur passen, leicht herausgefiltert werden. Es werden nur Strukturen erlaubt, die eine Abfolge von sich öffnenden Klammern und dann eine Abfolge von sich schließenden Klammern aufweisen. Die Abfolgen dürfen durch Punkte unterbrochen sein. Für die Schleife wurde eine maximale Länge von 20 Nukleotiden festgelegt. Außerdem darf der Stamm maximal 16 ungepaarte Nukleotide enthalten. Diese Kriterien werden von etwa 90% der pre-MicroRNAs aus der miRBase erfüllt, stellen also akzeptable Beschränkungen dar.

4 Ergebnisse

4.1 Korrelation zwischen bekannten pre-MicroRNAs und konservierten Regionen

Diese Arbeit beschränkt sich auf das Identifizieren phylogenetisch konservierter MicroRNAs. Um herauszufinden, wie viele MicroRNAs mit diesem Ansatz von vornherein ausgeschlossen werden und ob MicroRNAs tatsächlich gehäuft in konservierten Regionen vorkommen, wurde ein Abgleich zwischen den pre-MicroRNAs aus der miRBase und den konservierten Regionen vorgenommen. Zum Vergleich wurde dieser Vorgang mit Zufallssequenzen anstelle der pre-MicroRNAs wiederholt (Erläuterungen siehe Abschnitt 3.3.1). Die Ergebnisse dieser Korrelationen sind in Tabelle dargestellt. Von insgesamt 1029 pre-MicroRNAs aus der miRBase liegen 328 (31,88%) innerhalb von konservierten Regionen. Von den Zufallssequenzen dagegen erwartet man nur 11,34 (1,10%) innerhalb von konservierten Regionen (Standardabweichung: +/- 3,42). Mit Hilfe der Tschebyschev-Ungleichung 4.1 wurde die Signifikanz dieses Ergebnisses ermittelt.

$$P[|X - \mu| \geq k] \leq \frac{\sigma^2}{k^2} \tag{4.1}$$

Tabelle 4.1: Korrelation bekannter pre-MicroRNAs und Zufallssequenzen mit konservierten Regionen.

	gesamt	in konservierter Region enthalten	mit konservierter Region überlappt	Abstand von max. 1000 BP zu konservierter Region
bekannte pre-MicroRNAs	1029	328 (31,88%)	441 (42,86%)	159 (15,45%)
Zufallssequenzen	1029	11,34 (1,10%)	77,18 (7,50%)	462,01 (44,90%)

Setzt man nun die entsprechenden Werte im Fall der pre-MicroRNAs, die innerhalb einer konservierten Region liegen, in 4.1 ein, erhält man folgendes:

$$P[|X - 11,34| \geq 316,66] \leq \frac{3,42^2}{316,66^2} \tag{4.2}$$

$$P[|X - 11,34| \geq 316,66] \leq 0,00012 \tag{4.3}$$

Die Wahrscheinlichkeit, dass von 1029 erzeugten Zufallssequenzen 328 oder mehr in konservierte Regionen fallen, ist kleiner oder gleich 0,00012, also etwa 0,1 Promille. Die tatsächlichen pre-MicroRNAs liegen also signifikant häufiger in konservierten Regionen als die Zufallssequenzen. Von den 1029 pre-MicroRNAs aus der miRBase weisen weitere 441 (42,86%) eine Überlappung mit den konservierten Regionen auf. Bei den Zufallssequenzen trifft dies nur auf 77,18 zu (Standardabweichung: +/- 7,50). Auch hier ist der Unterschied zwischen den pre-MicroRNAs aus der miRBase und den Zufallssequenzen signifikant (P=0,00043). 159 (15,45%) pre-MicroRNAs aus der miRBase befinden sich in der Umgebung (maximal 1000 Nukleotide Abstand) von konservierten Regionen. Für die Zufallssequenzen liegt diese Anzahl bei 462,01 (Standardabweichung: +/- 16,46). Es liegen also signifikant (P = 0,00295) mehr Zufallssequenzen als pre-MicroRNAs in der Umgebung von konservierten Regionen.

Die Ergebnisse der Korrelation zwischen bekannten pre-MicroRNAs und konservierten Regionen machen deutlich, dass MicroRNAs gehäuft in konservierten Regionen vorkommen. Dies unterscheidet sich signifikant von den Zufallssequenzen. Da in dieser Arbeit nach pre-MicroRNA-Kandidaten innerhalb konservierter Regionen gesucht wird, gehen etwa 68% aller MicroRNAs verloren, unter der Annahme, dass die momentan bekannten MicroRNAs repräsentativ für alle MicroRNAs sind.

4.2 Ermittlung verschiedener Eigenschaften von konservierten Regionen, die mit einer bekannten MicroRNA in Relation gebracht werden können

In diesem Abschnitt werden verschiedene Eigenschaften der konservierten Regionen untersucht. Es wird mit konservierten Regionen gearbeitet, die eine bekannte MicroRNA enthalten, oder in deren näherer Umgebung sich eine bekannte MicroRNA befindet.

Der Grund hierfür ist, dass es am wahrscheinlichsten ist, für eben diese konservierten Regionen geeignete Kriterien für die Suche nach neuen MicroRNAs zu entdecken. Die konservierten Regionen wurden dabei in drei Klassen eingeteilt, ähnlich wie die MicroRNAs in Abschnitt 4.1. Die erste Klasse umfasst konservierte Regionen, die eine bekannte MicroRNA komplett enthalten. Der zweiten Klasse wurden konservierte Regionen zugeteilt, die eine Überlappung mit einer MicroRNA von mindestens einem Basenpaar aufweisen. Für die dritte Klasse wurden schließlich die Kriterien so erweitert, dass konservierte Regionen aufgenommen wurden, die eine Entfernung von maximal 1000 Basenpaaren zu einer MicroRNA aufweisen. In Klasse 1 befinden sich 675, in Klasse 2 2.480 und in Klasse 3 5.132 konservierte Regionen. Keine dieser drei Klassen enthält Mitglieder aus einer der anderen Klassen.

4.2.1 Konservierungsgrad

In einer ersten Analyse wurden die konservierten Regionen auf ihren Konservierungsgrad hin untersucht. Der Konservierungsgrad gibt an, in wievielen von den fünf Organismen eine konservierte Region vorkommt. In Abbildung 4.1 sind die Konservierungsgrade dargestellt. Die Aufteilung der konservierten Regionen in die bereits bekannten drei Klassen wurde beibehalten. Auf der x-Achse wurde der Konservierungsgrad und auf der y-Achse die Anzahl der konservierten Regionen (für alle fünf Organismen) aufgetragen. Bei konservierten Regionen der Klasse 1 kann man deutlich erkennen, dass der Großteil (etwa 89%) über mindestens vier Organismen hinweg konserviert ist. Ähnlich stellt sich dies für die konservierten Regionen der Klasse 2 dar, etwa 85% haben einen Konservierungsgrad von 4 oder 5. Für die konservierten Regionen der Klasse drei kann eine solche Aussage nicht getroffen werden. Zwar kommt auch hier der größte Anteil in vier oder fünf Organismen vor (etwa 55%), aber es ist auch der Anteil relativ groß, der nur über zwei oder drei Organismen hinweg konserviert ist (etwa 45%). Der Konservierungsgrad ist damit ein gutes Kriterium für die Suche nach MicroRNA-Kandidaten, wenn man sich bei der Suche auf MicroRNA-Kandidaten beschränkt, die in konservierten Regionen liegen, oder eine Überlappung mit konservierten Regionen aufweisen. In diesen Fällen sind Konservierungsgrade von 4 oder 5 die Regel.

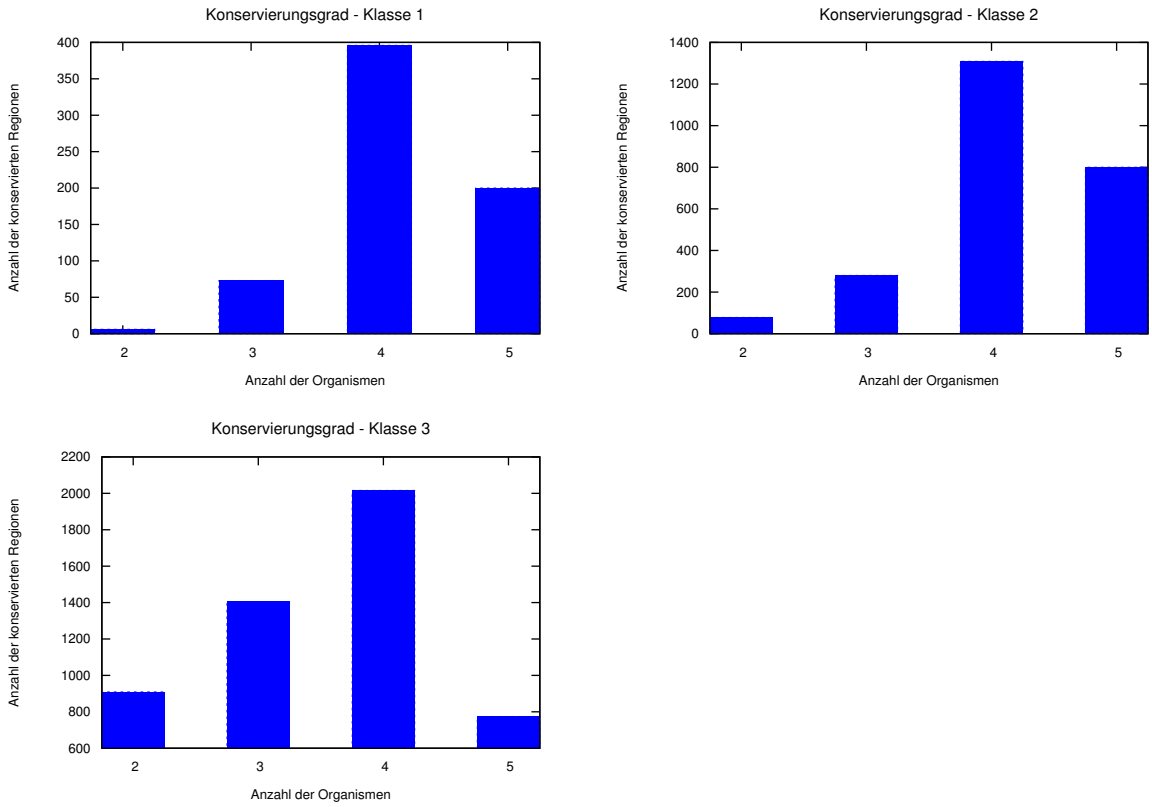


Abbildung 4.1: Darstellung der Konservierungsgrade für die drei Klassen der konservierten Regionen aller fünf Organismen.

4.2.2 Länge

Die zweite untersuchte Eigenschaft konservierter Regionen ist die Länge. Hier wurde die Aufteilung in die drei Klassen ebenfalls beibehalten, zusätzlich wurde die Längenverteilung von allen konservierten Regionen ermittelt. Das beste Ergebnis bekommt man bei Klasse eins, dargestellt in Abbildung 4.2. Hier kommen Längen zwischen 75 und 125 Nukleotiden am häufigsten vor, was sehr gut mit den Längen der precursor-Transkripte von MicroRNAs korreliert (ca. 65-115 Nukleotide), siehe Abbildung 4.3. Die konservierten Regionen der Klasse 1 sind dadurch definiert, pre-MicroRNAs komplett zu enthalten, d. h. sie sind auch mindestens so lang wie pre-MicroRNAs. Daraus ergibt sich eine Grenze auf der linken Seite der Längenverteilung für Klasse 1. Auf der rechten Seite der Verteilung gibt es keine solche scharfe Begrenzung, allerdings sind nur vereinzelte konservierte Regionen deutlich länger als pre-MicroRNAs. Dies lässt den Schluß zu, dass im

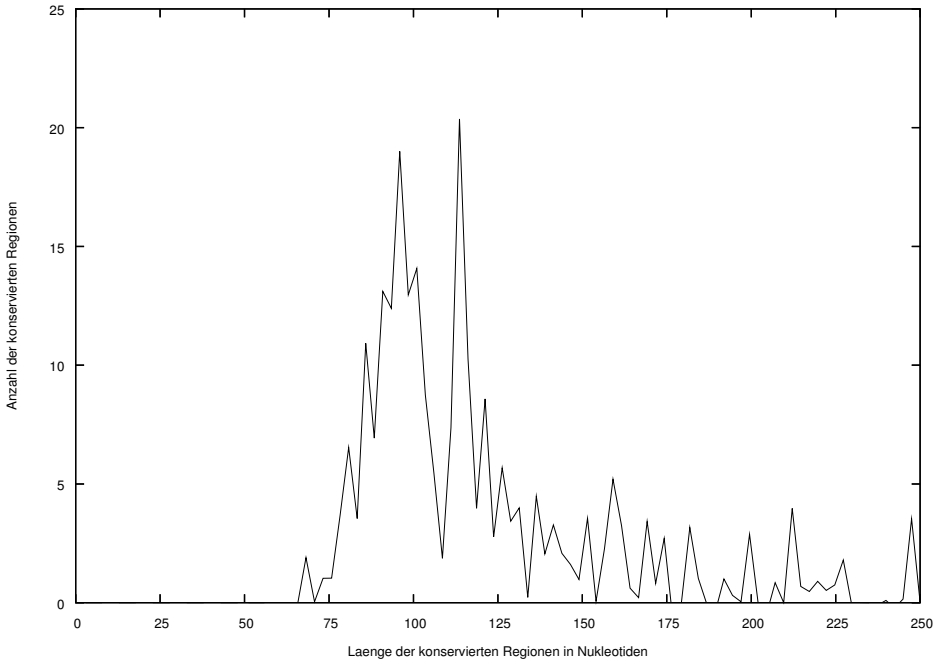


Abbildung 4.2: Längenverteilung konservierter Regionen der Klasse 1.

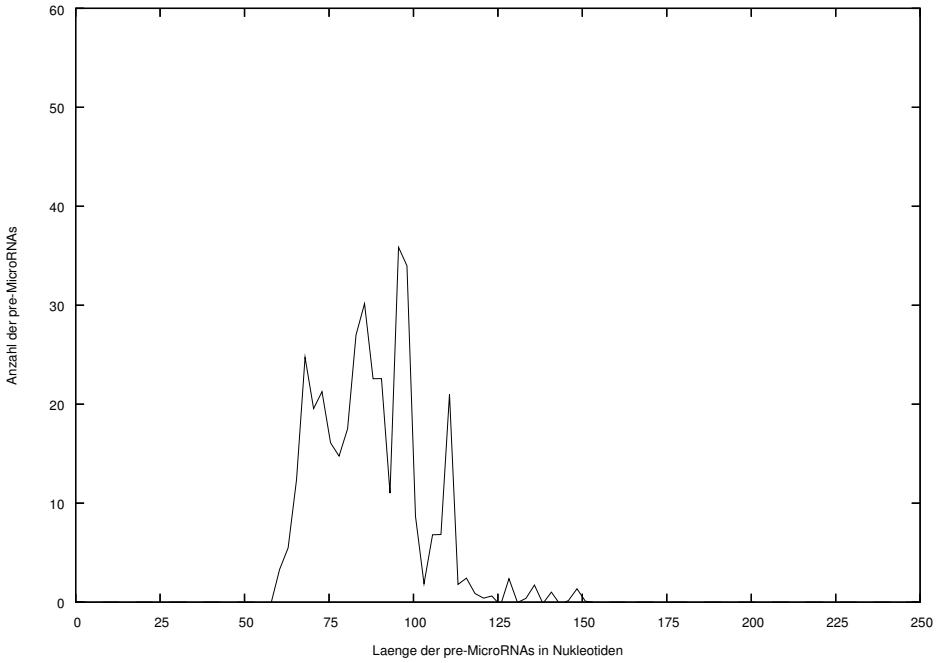


Abbildung 4.3: Längenverteilung der pre-MicroRNAs.

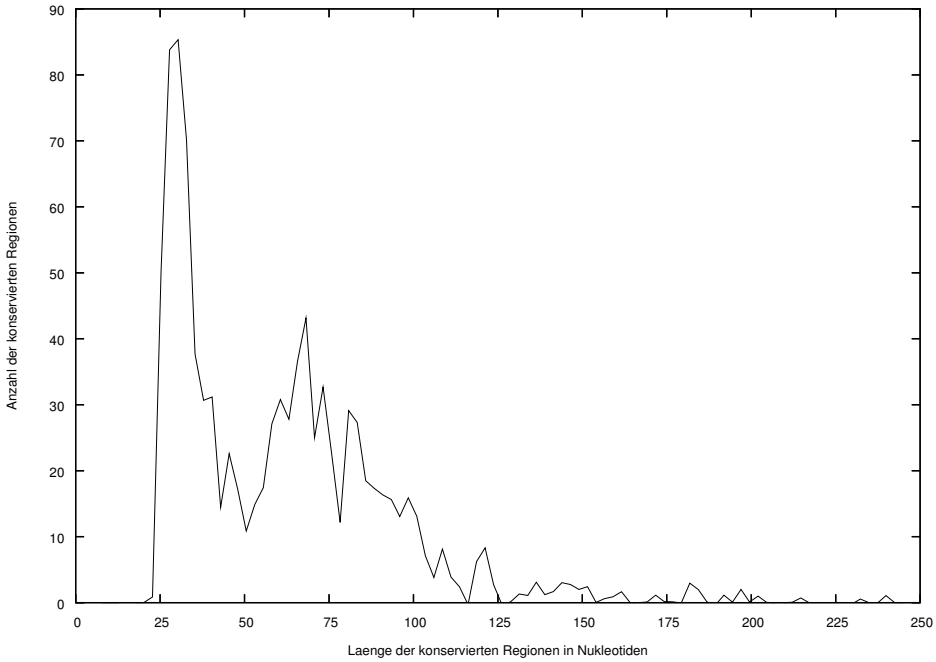


Abbildung 4.4: Längenverteilung konservierter Regionen der Klasse 2.

Regelfall nur die Sequenz der pre-MicroRNA phylogenetisch konserviert ist, nicht aber die pri-MicroRNA (Primärtranskript). Die Längenverteilung der konservierten Regionen der Klasse 2 ist in Abbildung 4.4 dargestellt. Hier sind die Längen in zwei Hauptgruppen aufgespaltet. Die erste Gruppe liegt zwischen 25 und 50 Nukleotiden. Dies lässt sich leicht erklären, wenn man sich die Längenverteilung aller konservierten Regionen in Abbildung 4.5 ansieht. Die meisten konservierten Regionen liegen genau im selben Bereich. Die zweite Gruppe liegt in einem Längenbereich von 50 bis 100 Nukleotiden, hat also wieder eine teilweise Übereinstimmung mit dem Längenbereich der pre-MicroRNAs. Die Längenverteilung von Klasse drei (Abbildung 4.6) dagegen kann man nicht für eine Vorhersage von MicroRNA-Kandidaten gebrauchen, da keine relevanten Unterschiede zur Längenverteilung aller konservierten Regionen zu erkennen ist.

Abschließend kann gesagt werden, dass man sich bei der Suche nach konservierten MicroRNAs auf konservierte Regionen mit einer Länge von 75 bis 125 Nukleotiden konzentrieren kann, da die pre-MicroRNAs selten in längeren konservierten Regionen vorkommen.

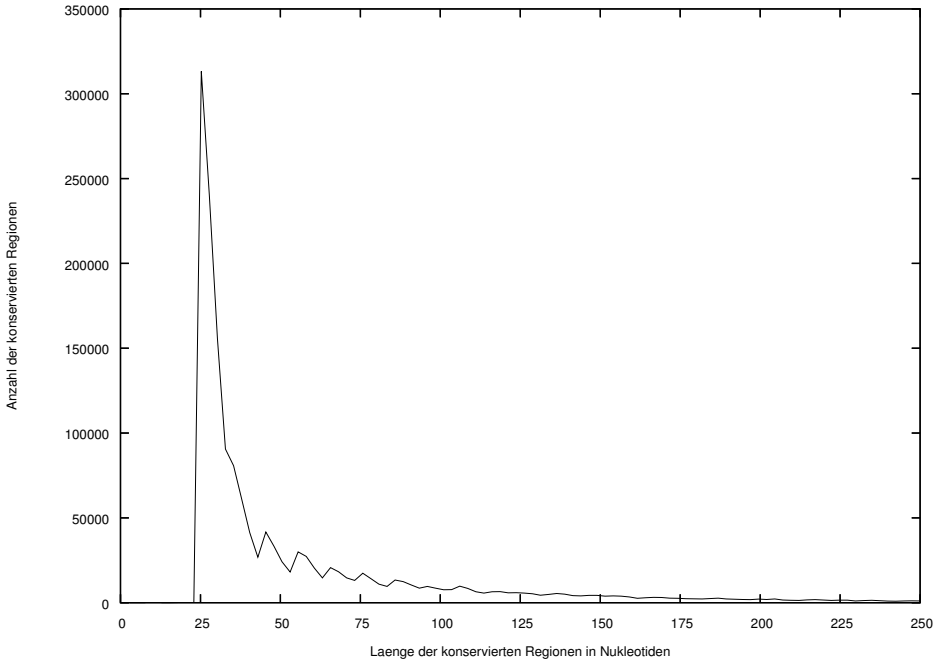


Abbildung 4.5: Längenverteilung aller konservierter Regionen.

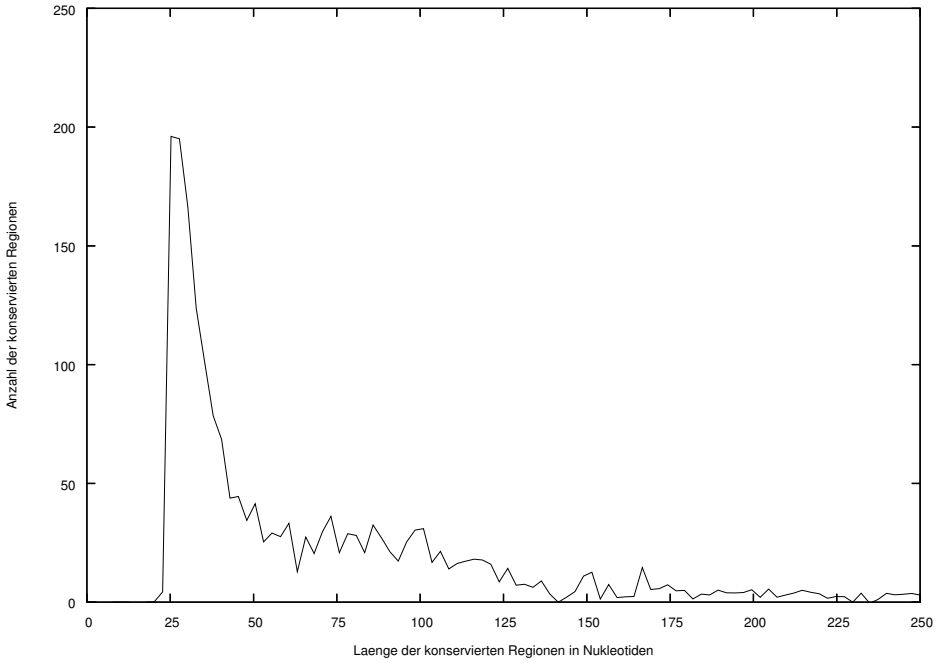


Abbildung 4.6: Längenverteilung konservierter Regionen der Klasse 3.

Tabelle 4.2: Korrelation von konservierten Regionen mit Gen-Loci.

	gesamt	in Gen-Loci enthalten	mit Gen-Loci überlappt	zwischen Gen-Loci
Klasse 1	675	319 (47,26%)	31 (4,59%)	325 (48,15%)
Klasse 2	2.480	972 (39,19%)	44 (1,77%)	1.464 (59,03%)
Klasse 3	5.132	2.430 (47,35%)	86 (1,68%)	2.616 (50,08%)

Tabelle 4.3: Korrelation von konservierten Regionen in Gen-Loci mit Exons.

	gesamt	in Exons	mit Exons überlappt	in Introns
Klasse 1	319	84 (26,33%)	24 (7,52%)	234 (73,35%)
Klasse 2	972	222 (22,84%)	51 (5,25%)	756 (77,78%)
Klasse 3	2.430	833 (34,28%)	471 (19,38%)	1.291 (53,13%)

4.2.3 Korrelation mit Gen-Loci

In einem weiteren Schritt wurde überprüft, inwieweit die konservierten Regionen der drei Klassen mit Gen-Loci übereinstimmen. Wenn eine konservierte Region in einem Gen-Locus liegt, wurde auch die Lage bezüglich Exons und Introns festgestellt. Liegt eine konservierte Region in einem Exon, so wurde zusätzlich geprüft, ob dies das einzige Exon im entsprechenden Transkript ist. In Tabelle 4.2 sind die Ergebnisse für die Korrelation von Gen-Loci und den konservierten Regionen dargestellt. Von Klasse 1 und 3 befinden sich jeweils etwa die Hälfte der Sequenzen in Gen-Loci, bei Klasse 2 sind es etwa 40%. Hier lässt sich also kein für die MicroRNA-Vorhersage relevanter Unterschied zwischen den Klassen feststellen.

Für einen Gen-Locus kann es mehrere alternative Transkripte geben. Diese alternativen Transkripte unterscheiden sich vor allem durch Anzahl und Position der Exons. Dadurch kann es vorkommen, dass eine konservierte Region, bezogen auf ein bestimmtes Transkript, in einem Exon liegt, bezogen auf ein alternatives Transkript aber in einem Intron. Bei der Korrelation der konservierten Regionen mit den Exons bzw. Introns, wurden solche Fälle doppelt gezählt, d. h. einmal als in einem Exon liegend, und ein zweites Mal als in einem Intron liegend. Die Ergebnisse sind in Tabelle 4.3 aufgelistet. Man erkennt, dass konservierte Regionen aus Klasse 1 und Klasse 2 häufiger in Introns vorkommen als die konservierten Regionen der Klasse 3. Als nächstes wurden die konservierten Regionen, die in einem Exon liegen, genauer betrachtet. Für etwa 50% der konservierten Regionen aus Klasse 1, die in einem Exon liegen, wurde festgestellt, dass

es sich hierbei ausschließlich um Ein-Exon-Transkripte handelt. Da solche Transkripte häufig nicht für ein Protein kodieren (kommuniziert durch Andreas Klingenhoff, Genomatix Software GmbH), liegt es nahe, dass sie MicroRNA-Gene darstellen. Für Klasse 2 und Klasse 3 wurden Werte von 44% und 18% ermittelt. Konservierte Regionen aus Klasse 1 und Klasse 2 kommen also wesentlich häufiger in Ein-Exon-Transkripten vor als konservierte Regionen aus Klasse 3.

4.3 Vorhersage neuer MicroRNAs

Zur Vorhersage neuer MicroRNAs wurden alle konservierte Regionen systematisch klassifiziert. Die Kriterien hierfür waren die Länge der konservierten Regionen, ihre Sekundärstruktur mit zugehöriger minimaler freier Energie und der Konservierungsgrad. Anschließend wurden Hinweise auf die Expression der untersuchten konservierten Regionen gewonnen, indem Positionsvergleiche mit Gen-Loci und Expressed Sequence Tags durchgeführt wurden. Am Ende steht ein Datensatz von konservierten Regionen, die gute Kandidaten für pre-MicroRNAs sind.

4.3.1 Vorauswahl der pre-MicroRNA-Kandidaten anhand der Länge der konservierten Regionen

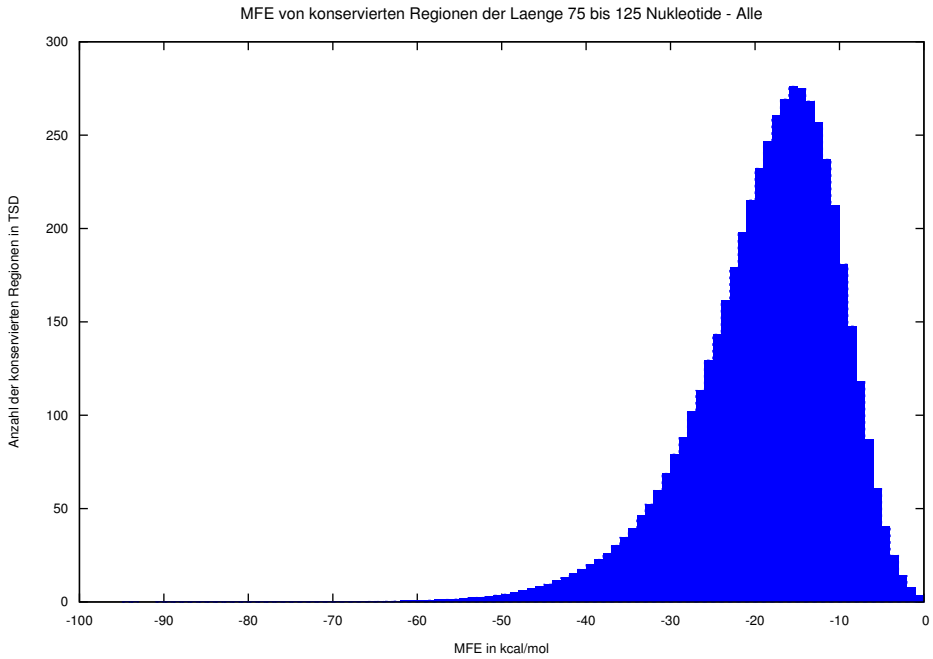
In Absatz 4.2.2 wurde festgestellt, dass konservierte Regionen, die eine pre-MicroRNA enthalten, sehr häufig eine Länge zwischen 75 und 125 Nukleotiden aufweisen. Dies bedeutet nichts anderes, als dass die stem-loop-Strukturen phylogenetisch konserviert sind, die zugehörigen Primärtranskripte (pri-MicroRNAs) aber nicht. Diese Information reicht natürlich nicht aus, um sagen zu können, dass konservierte Regionen mit einer Länge zwischen 75 und 125 Nukleotiden pre-MicroRNAs darstellen. Allerdings wurde diese Information genutzt, um eine erste Vorauswahl der konservierten Regionen zu treffen. Die Anzahl der zu analysierenden Sequenzen kann durch die Längenbeschränkung von etwa 24 Millionen auf etwa 2,5 Millionen reduziert werden. Dadurch wird der nachfolgende Schritt, die Sekundärstrukturanalyse, wesentlich beschleunigt.

4.3.2 Verfeinerung der Auswahl der pre-MicroRNA-Kandidaten anhand der Sekundärstruktur und zugehörigen MFE

Die thermodynamische Stabilität der Sekundärstruktur ist ein gutes Charakteristikum, um pre-MicroRNAs zu identifizieren. Bonnet u. a. [5] haben gezeigt, dass pre-MicroRNAs, im Gegensatz zu tRNAs und rRNAs, eine freie Faltungsenergie haben, die signifikant niedriger ist als bei Zufallssequenzen. Diese Tatsache wurde genutzt, um die Auswahl der konservierten Regionen, die möglicherweise eine MicroRNA enthalten, weiter zu verfeinern. Um zwischen konservierten Regionen mit potentieller MicroRNA und dem Rest unterscheiden zu können, musste ein cut-off-Wert für die MFE festgelegt werden. Zur Bestimmung eines möglichst günstigen MFE-Wertes wurden zwei Datensätze mit RNAfold analysiert (Beschreibung siehe Methodenteil). Abbildung 4.7a zeigt die Verteilung der MFE für alle konservierten Regionen der Länge 75 bis 125 und in Abbildung 4.7b ist die MFE-Verteilung der konservierten Regionen der Länge 75 bis 125, die eine bekannte MicroRNA enthalten (Klasse 1), dargestellt. In Abbildung 4.8 wurden zusätzlich zu den Verteilungen jeweils noch die zugehörigen Summenverteilungskurven geplottet. Anhand dieser Summenverteilungskurven wurde dann ein cut-off-Wert von -30 kcal/mol festgelegt, d. h. alle konservierte Regionen mit einem MFE-Wert kleiner -30 kcal/mol werden als MicroRNA-Kandidaten betrachtet. Etwa 95% der konservierten Regionen der Länge 75 bis 125, die eine bekannte MicroRNA enthalten, erfüllen dieses Kriterium. Bei einem cut-off-Wert von -30 kcal/mol gehen nur etwa 5% als falsch negative verloren. Betrachtet man alle konservierten Regionen der Länge 75 bis 125, so haben etwa 90% einen MFE-Wert größer als -30 kcal/mol, d. h. etwa 10% werden fälschlicherweise als MicroRNA-Kandidaten betrachtet. Ein cut-off-Wert von -30 kcal/mol stellt also einen guten Kompromiss zwischen richtig positiven und falsch positiven Entscheidungen dar.

Von etwa 2,5 Millionen konservierten Regionen mit einer Länge von 75 bis 125 Nukleotiden konnte der Datensatz, durch Berücksichtigung des MFE-Wertes und der Sekundärstruktur, auf 22.331 veringert werden. Diese konservierten Regionen erfüllen die wichtigsten strukturellen Kriterien für pre-MicroRNAs (Länge, Sekundärstruktur und MFE) und stellen nun schon recht gute pre-MicroRNA-Kandidaten dar.

a)



b)

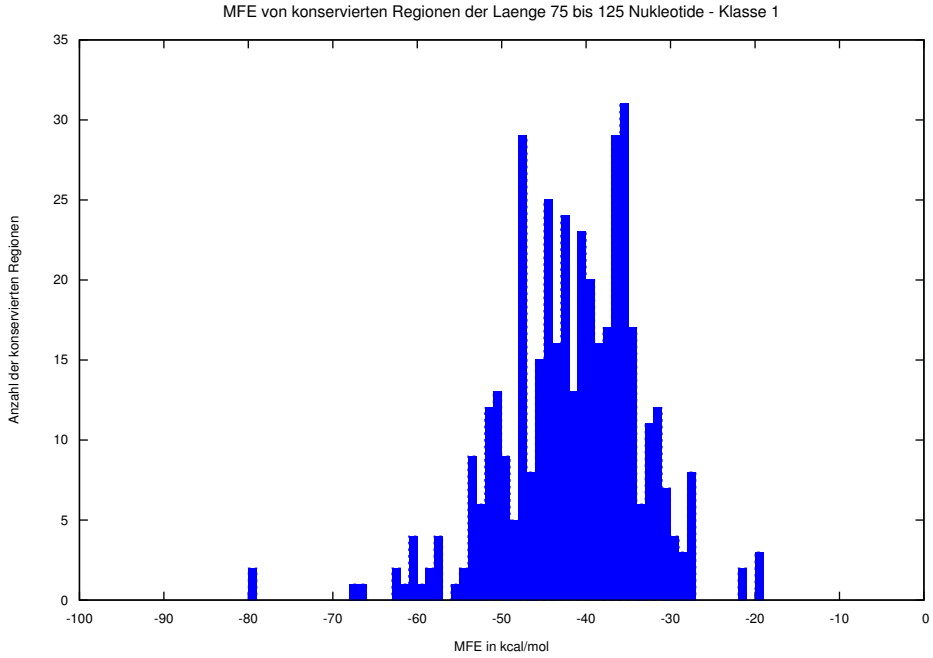
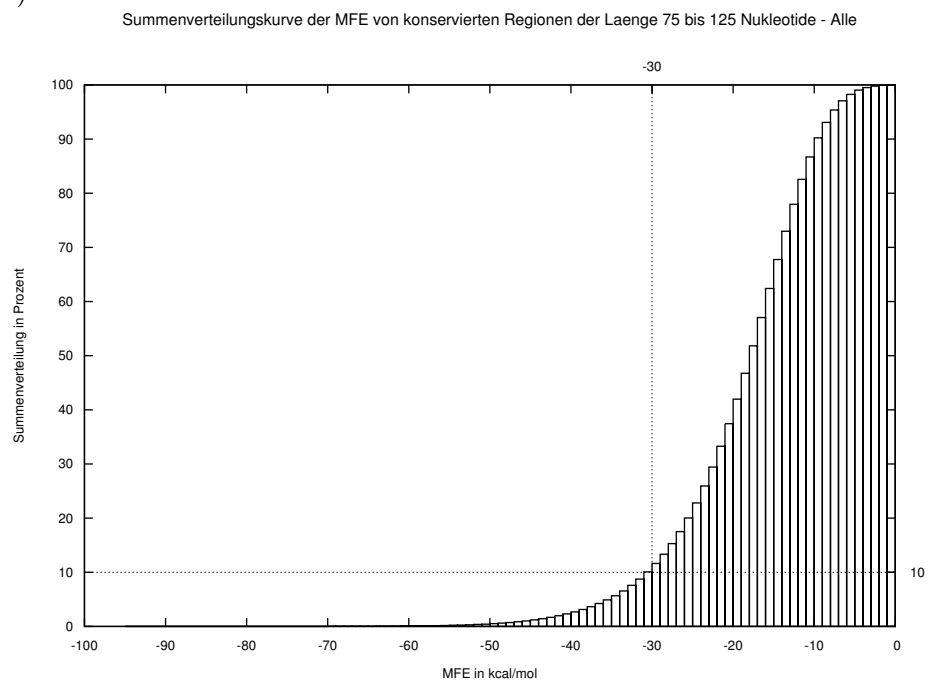


Abbildung 4.7: MFE-Verteilungen für alle konservierte Regionen der Länge 75 bis 125 Nukleotide (a) und für konservierte Regionen der Klasse 1 (b).

a)



b)

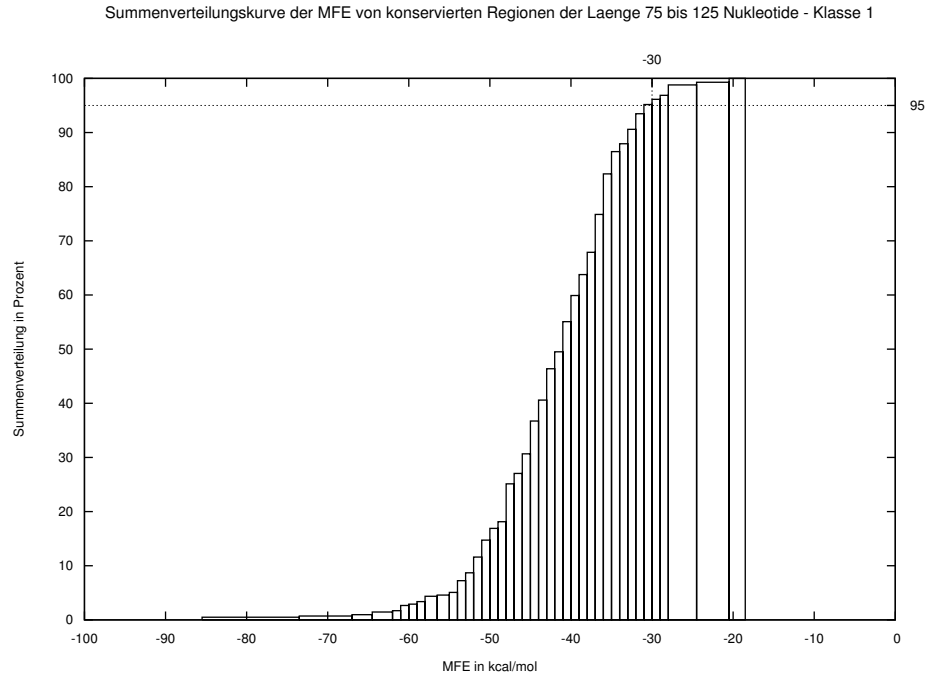


Abbildung 4.8: Summenverteilungskurven der MFE-Verteilungen für alle konservierte Regionen der Länge 75 bis 125 Nukleotide (a) und für konservierte Regionen der Klasse 1 (b).

4.3.3 Weitere Verfeinerung der Auswahl der pre-MicroRNA-Kandidaten anhand des Konservierungsgrades

Vom logischen Ablauf her wäre es sinnvoll gewesen, diesen Schritt mit der Vorauswahl der pre-MicroRNAs anhand der Länge der konservierten Regionen vorzunehmen. Auf Grund des großen Datensatzes der konservierten Regionen für die fünf Vertebraten (Mensch, Maus, Hund, Rind, Zebrafisch) hätte die Ermittlung des Konservierungsgrades für alle konservierten Regionen allerdings unverhältnismäßig lange gedauert. Selbst nach der Verkleinerung des Datensatzes aufgrund des Längenkriteriums auf etwa 2,5 Millionen konservierter Regionen hätte die Analysedauer mehr als 30 Tage betragen. Deshalb wurde der Konservierungsgrad erst nach dem Schritt der Sekundärstrukturanalyse ermittelt, da dann nur noch 22.331 Sequenzen betrachtet werden mussten. In Abschnitt 4.2.1 wurde festgestellt, dass der Großteil (etwa 89%) der konservierten Regionen, die eine bekannte MicroRNA komplett enthalten, einen Konservierungsgrad größer 3 aufweisen. Deshalb wurden von den 22.331 pre-MicroRNA-Kandidaten aus dem vorangegangenen Schritt diejenigen beibehalten, die ebenfalls einen Konservierungsgrad von größer 3 haben. Von den 22.331 bleiben so 13.258 sehr gute pre-MicroRNA-Kandidaten übrig.

4.3.4 Korrelation von pre-MicroRNA-Kandidaten mit Gen-Loci

Bewertet man die konservierten Regionen anhand ihrer Länge, ihrer Sekundärstruktur, der zugehörigen MFE und des Konservierungsgrades, so bleiben 13.258 Sequenzen als pre-MicroRNA-Kandidaten übrig. Diese Kriterien sagen allerdings nichts über eine Expression der betreffenden Sequenzen aus. Deshalb wurde für jede einzelne Sequenz geprüft, ob sie in einem Gen-Locus liegt. Stellt man beispielsweise fest, dass ein pre-MicroRNA-Kandidat in einem Ein-Exon-Transkript liegt, so wäre dies ein deutlicher Hinweis auf die Expression. Aber nicht nur die Lage in einem Ein-Exon-Transkript wäre interessant. Der Ursprungsort einer MicroRNA kann auch in einem Intron eines kodierenden Transkriptes liegen. Beim Splice-Vorgang der mRNA eines solchen Transkriptes wird dann die pre-MicroRNA freigesetzt [17]. Die Expression eines solchen Gens kann daher gleichzeitig zur Repression eines anderen Gens führen. Die Ergebnisse der Korrelation von pre-MicroRNA-Kandidaten mit Gen-Loci und den entsprechenden Exon-Intron-Strukturen sind in Tabelle 4.4 und Tabelle 4.5 dargestellt. Von den 13.258 pre-

Tabelle 4.4: Korrelaton der pre-MicroRNA-Kandidaten mit Gen-Loci.

gesamt	in Gen-Loci enthalten	mit Gen-Loci überlappt	zwischen Gen-Loci
13.258	4.567 (34,45%)	61 (0,46%)	8.630 (65,10%)

Tabelle 4.5: Korrelation der pre-MicroRNA-Kandidaten in Gen-Loci mit Exons. Aufgrund der Tatsache, dass ein pre-MicroRNA-Kandidat in mehreren Transkripten und somit sowohl in einem Exon als auch in einem Intron vorkommen kann, ist die Summe über die drei rechten Spalten größer als die Anzahl der pre-MicroRNA-Kandidaten in einem Gen-Loci.

gesamt	in Exons	mit Exons überlappt	in Introns
4.567	3.344 (73,22%)	197 (4,31%)	1.380 (30,22%)

MicroRNA-Kandidaten liegen 4.576 (34,45%) in Gen-Loci. Davon wiederum liegen 1.380 (30,22%) in Introns. Von den 3.344 pre-MicroRNA-Kandidaten, die in einem Exon liegen, liegen 380 (11,36%) in einem Ein-Exon-Transkript.

4.3.5 Korrelation von pre-MicroRNA-Kandidaten mit ESTs

Eine weitere Möglichkeit, einen Hinweis auf die Expression der pre-MicroRNA-Kandidaten zu finden, ist die Korrelation mit Expressed Sequence Tags (ESTs). ESTs sind kurze Sequenzen (normalerweise 200 bis 500 Nukleotide lang), die durch die Sequenzierung von einem oder beider Enden von cDNAs erstellt wurden. Somit gibt es ESTs von bekannten Genen, aber auch von exprimierten Sequenzen, die zwischen diesen Genen liegen. ESTs stellen ein Abbild der exprimierten Sequenzen in einer bestimmten Gewebeart bzw. in einem bestimmten Entwicklungsstadium eines Organismus dar. Allerdings gibt es nicht für alle Gewebearten bzw. Entwicklungsstadien ESTs. Mit Hilfe der ESTs kann also unabhängig davon, ob ein pre-MicroRNA-Kandidat in einem Gen-Locus liegt, ein Hinweis auf die Expression gefunden werden. Die Korrelation zwischen pre-MicroRNA-Kandidaten und den ESTs wurde anhand einfacher Positionsvergleiche vorgenommen.

Von den 8.630 pre-MicroRNA-Kandidaten, die zwischen Gen-Loci liegen, kommen 4.015 (46,52%) in ESTs vor. Es bleiben 4.615 pre-MicroRNA-Kandidaten übrig, für die mit dieser Methode kein Hinweis auf eine Expression gefunden werden konnte. Dies bedeutet aber nicht, dass eine Expression dieser Sequenzen ausgeschlossen ist.

4.3.6 Abgleich von pre-MicroRNA-Kandidaten mit den stem-loop-Strukturen aus der miRBase

Der Vergleich zwischen den in dieser Arbeit gefundenen pre-MicroRNA-Kandidaten und den pre-MicroRNAs aus der miRBase soll aufzeigen, wieviele der bekannten MicroRNAs mit dem Ansatz dieser Arbeit identifiziert werden. Dadurch kann eine Aussage bezüglich der Güte der in dieser Arbeit verwendeten Methode gemacht werden. Von den 1029 bekannten pre-MicroRNAs mit eindeutiger Position im entsprechenden Genom liegen 328 in einer konservierten Region. Davon wiederum liegen 244 (74,39%) pre-MicroRNAs in einer konservierten Region mit einer Länge zwischen 75 und 125 Nukleotiden. Durch die Verwendung der Länge als Ausschlusskriterium, werden 84 (25,61%) konservierte pre-MicroRNAs fälschlicherweise nicht erkannt. Berücksichtigt man die Sekundärstruktur und MFE als weitere Kriterien, so bleiben von ursprünglich 328 bekannten und konservierten pre-MicroRNAs noch 220 (67,07%) übrig, die mit dem Ansatz dieser Arbeit korrekt als pre-MicroRNAs erkannt werden. Schließt man nun noch pre-MicroRNAs aus, die einen Konservierungsgrad kleiner als 4 haben, dann bleiben noch 203 (61,89%) übrig. Geht man davon aus, dass die Eigenschaften der MicroRNAs in der miRBase repräsentativ für alle MicroRNAs gelten, dann werden etwa 62% der konservierten MicroRNAs von Mensch, Maus, Hund, Rind und Zebrafisch mit dem Ansatz dieser Arbeit gefunden.

5 Diskussion

5.1 Was in dieser Arbeit erreicht wurde

In dieser Arbeit wurde ein in silico Ansatz zum Identifizieren neuer MicroRNAs entwickelt. Dabei wurde sich nicht auf reife MicroRNAs, sondern auf die pre-MicroRNAs konzentriert. Diese Vorgehensweise wurde gewählt, da pre-MicroRNAs wesentlich mehr Unterscheidungsmerkmale gegenüber Zufallssequenzen aufweisen als reife MicroRNAs. Ein Merkmal, das sowohl auf pre-MicroRNAs als auch auf reife MicroRNAs zutrifft, ist die phylogenetische Konservierung. Etwa 3/4 der pre-MicroRNAs aus der miRBase weisen zumindest eine Überlappung mit den in dieser Arbeit verwendeten phylogenetisch konservierten Regionen auf. Aufgrund dieser Tatsache beschränkt sich diese Arbeit bei der Suche nach neuen MicroRNAs auf konservierte Regionen. Durch die Betrachtung der Längenverteilung der konservierten Regionen, die eine bekannte MicroRNA enthalten, wurde festgestellt, dass in der Regel nur die pre-MicroRNAs konserviert sind, nicht aber die Primärtranskripte. Deshalb konnte der Suchbereich weiter eingeschränkt werden, nämlich auf konservierte Regionen mit zu pre-MicroRNAs vergleichbaren Längen. Außerdem wurde herausgefunden, dass konservierte pre-MicroRNAs meistens einen Konservierungsgrad von 4 oder 5 aufweisen. Eine weitere markante Eigenschaft der pre-MicroRNA ist ihre Sekundärstruktur. Pre-MicroRNAs bilden eine stem-loop-Struktur aus. Diese besitzt im Regelfall eine minimale freie Energie, die signifikant niedriger ist als bei Zufallssequenzen [5]. Mit all diesen Eigenschaften von pre-MicroRNAs wurde nun in den phylogenetisch konservierten Regionen von Mensch, Maus, Hund, Rind und Zebrafisch nach neuen pre-MicroRNA-Kandidaten gesucht. Mit diesem Ansatz wurden etwa 13.000 neue pre-MicroRNA-Kandidaten identifiziert und etwa 62% der bekannten und konservierten pre-MicroRNAs aus der miRBase wiedergefunden. Da es wichtig ist, zu wissen, ob diese pre-MicroRNA-Kandidaten exprimiert werden, wurden Positionsvergleiche mit Gen-Loci und ESTs vorgenommen. So konnte für 8.643 pre-MicroRNA-

Kanditaten ein Hinweis auf die Expression gefunden werden. Für die restlichen 4.615 Kandidaten muss eine Expression aber deshalb nicht ausgeschlossen werden, da die Datensätze von Gen-Loci und ESTs nicht vollständig sind.

In der miRBase sind für die betrachteten fünf Organismen insgesamt 1.330 (Stand: Juni 2007) MicroRNAs registriert. In dieser Arbeit wurde etwa 13.000 neue MicroRNA-Kandidaten gefunden, d. h. die Annahme, dass erst ein Bruchteil aller MicroRNAs bekannt ist, wurde bestätigt. Die in dieser Arbeit gefundenen MicroRNA-Kandidaten werden von Genomatix für die Genomannotation verwendet.

5.2 Verbesserungsmöglichkeiten

Verbesserungsmöglichkeiten des in dieser Arbeit entwickelten Ansatzes zur Vorhersage von MicroRNAs gibt es vor allem bei der Sekundärstrukturanalyse. Momentan wird geprüft, ob der pre-MicroRNA-Kandidat eine stem-loop-Struktur ausbildet, wobei die Schleife maximal 20 Nukleotide lang sein darf und der Stamm maximal 16 ungepaarte Nukleotide enthalten darf. Zusätzlich könnten strukturelle Eigenschaften wie die Länge des längsten perfekten Stammes, die Länge von symmetrischen und asymmetrischen Schleifen im Stamm, oder die Nukleotidzusammensetzung verwendet werden. Eine weitere Verbesserungsmöglichkeit bestünde darin, die Werte für die Ermittelte minimale freie Energie zu normalisieren, da diese von der Länge der jeweiligen Sequenz abhängt. Man müsste den MFE-Wert also noch durch die Länge der Sequenz teilen, um die Sequenzen besser miteinander vergleichen zu können. In dieser Arbeit wurde bisher darauf verzichtet, da durch die Längenbegrenzung der MicroRNA-Kandidaten (75 bis 125 Nukleotide) eine gewisse „Normalisierung“ von vornherein gegeben ist.

Mit der Zunahme der Kriterien, die für die MicroRNA-Suche herangezogen werden, wird die Entscheidung für oder gegen die Einstufung als MicroRNA-Kandidat immer komplexer. Deshalb sollte überlegt werden, ob die Verwendung einer Support Vector Machine[1] sinnvoll wäre.

[1] Support Vector Machine ist eine Methode aus dem Bereich des maschinellen Lernens und wird verwendet um eine Menge von Objekten in zwei Klassen zu unterteilen.

5.3 Ausblick

Der nächste logische Schritt wäre die Target-Vorhersage für die in dieser Arbeit gefundenen MicroRNA-Kandidaten. Vorher müsste allerdings noch ein Ansatz zur Bestimmung der reifen MicroRNAs entwickelt werden. Momentan werden nur die pre-MicroRNAs betrachtet, da es nicht einfach ist, die genaue Lage einer reifen MicroRNA im Stamm der pre-MicroRNA zu bestimmen. Es könnte beispielsweise ein Ansatz des maschinellen Lernens entwickelt werden, mit dem sich die exakte Position der reifen MicroRNA ermitteln läßt. Hat man die Sequenz der reifen MicroRNA, so kann man bestehende Programme/Methoden für die Vorhersag der Targets verwenden. Eines dieser Programme wäre miRanda [35], welches auch vom Sanger-Institut für die Vorhersage von Targets zu den MicroRNAs aus der miRBase verwendet wird.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] miRBase [http://microrna.sanger.ac.uk/].

[2] V. Ambros. The functions of animal microRNAs. *Nature*, 431:350–355, 2004.

[3] E. Bernstein, A. A. Caudy, S. M. Hammond, and G. J. Hannon. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409:363–366, 2001.

[4] M. T. Bohnsack, K. Czaplinski, and D. Gorlich. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA*, 10:185–191, 2004.

[5] E. Bonnet, J. Wuyts, P. Rouze, and Y. Van de Peer. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20:2911–2917, 2004.

[6] J. Brennecke, A. Stark, R. B. Russel, and S. M. Cohen. Principles of MicroRNA-Target Recognition. *PLoS biology*, 3:e85, 2005.

[7] X. Cai, C. H. Hagedorn, and B. R. Cullen. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, 10:1957–1966, 2004.

[8] G. A. Calin and C. M. Croce. MicroRNA-Cancer Connection: The Beginning of a New Tale. *Cancer Research*, 66:7390–7394, 2006.

[9] Y.-L. Chiu and T. M. Rana. RNAi in human cells: basic structural and functional features of small interfering RNA. *Molecular cell*, 10:549–561, 2002.

[10] J. G. Doench, C. P. Petersen, and P. A. Sharp. siRNAs can function as miRNAs. *Genes and Development*, 216:438–442, 2004.

[11] J. G. Doench and P. A. Sharp. Specificity of microRNA target selection in translational repression. *Genes and Development*, 18:504–511, 2004.

[12] S. Griffiths-Jones. The microRNA Registry. *Database Issue*, 32:D109–D111, 2004.

[13] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Database Issue*, 34:D140–D111, 2006.

[14] B. Haley and P. D. Zamore. Kinetic analysis of the RNAi enzyme complex. *Nature structural and molecular biology*, 11:599–606, 2004.

[15] A. Khvorova, A. Reynolds, and S. D. Jayasena. Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115:209–216, 2003.

[16] C. A. Kidner and R. A. Martienssen. The developmental role of microRNAs in plants. *Current Opinion in Plant Biology*, 8:38–44, 2005.

[17] Young-Kook Kim and V Narry Kim. Processing of intronic microRNAs. *EMBO Journal*, 26:775–783, 2007.

[18] S. W. Knight and B. L. Bass. A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. *Science*, 293:2269–2271, 2001.

[19] E. C. Lai, B. Tam, and G. M. Rubin. Pervasive regulation of Drosophila Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs. *Genes and Development*, 19:1067–1080, 2005.

[20] E. C. Lai, P. Tomancak, R. W. Williams, and G. M. Rubin. Computational identification of Drosophila microRNA genes. *Genome Biology*, 4:R42, 2003.

[21] N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294:858–862, 2001.

[22] R. C. Lee, R. L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75:843–854, 1993.

[23] M. Legendre, A. Lambert, and D. Gautheret. Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, 21:841–845, 2005.

[24] L. P. Lim, M. E. Glansner, S. Yekta, C. B. Burge, and D. P. Bartel. Vertebrate microRNA genes. *Science*, 299:1540, 2003.

[25] J. Liu, M. A. Valencia-Sanchez, G. J. Hannon, and R. Parker. MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. *Nature Cell Biology*, 7:719–723, 2005.

[26] D. J. Luciano, D. Mirsky, N. J. Vendetti, and S. Maas. RNA editing of a miRNA precursor. *RNA*, 10:1174–1177, 2004.

[27] E. Lund, S. Guttinger, A. Calado, L. E. Dahlberg, and U. Kutay. Nuclear export of microRNA precursors. *Science*, 303:95–98, 2004.

[28] J. Martinez and T. Tuschl. RISC is a 5' phosphomonoester-producing RNA endonuclease. *Genes and Development*, 18:975–980, 2004.

[29] P. H. Olsen and V. Ambros. The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. *Developmental biology*, 216:671–680, 1999.

[30] J. S. Parker, S. M. Roe, and D. Barford. Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. *Nature*, 434:663–666, 2005.

[31] G. L. Sen and H. M. Blau. Argonaute 2/RISC resides in sites of mammalian mRNA decay known as cytoplasmic bodies. *Nature Cell Biology*, 7:633–636, 2005.

[32] A. Grishok u. a. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, 106:23–34, 2001.

[33] A. Mallory u. a. MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region. *EMBO*, 23:3356–3364, 2004.

[34] A. Sewer u. a. Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, 6:267, 2005.

[35] B. John u. a. Human microrna targets. *PLoS Biology*, 2:e363, 2004.

[36] C. Xue u. a. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6:310, 2005.

[37] D. S. Schwarz u. a. Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, 115:209–216, 2003.

[38] G. Hutvagner u. a. A cellular function for the RNA interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, 293:834–838, 2001.

[39] I. Bentwich u. a. Identification of hundreds of conserved and nonconserved human microRNAs. *Nature genetics*, 37:766–770, 2005.

[40] I. L. Hofacker u. a. Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chemie*, 125:167–188, 1994.

[41] J. B. Ma u. a. Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. *Nature*, 434:666–670, 2005.

[42] J. W. Nam u. a. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Research*, 33:3570–3581, 2005.

[43] L. P. Lim u. a. The microRNAs of Caenorhabditis elegans. *Genes and Development*, 17:991–1008, 2003.

[44] P. W. Hsu u. a. miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes. *Nucleic Acids Research*, 34:D135–D139, 2006.

[45] R. F. Ketting u. a. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes and Development*, 15:2654–2659, 2001.

[46] S. M. Johnson u. a. RAS is regulated by the let-7 microRNA familiy. *Cell*, 120:635–647, 2005.

[47] S. Pfeffer u. a. Identification of microRNAs of the herpesvirus family. *Nature methods*, 2:269–276, 2005.

[48] V. Ambros u. a. A uniform system for microRNA annotation. *RNA*, 9:277–279, 2003.

[49] W. Yang u. a. Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nature structural and molecular biology*, 34:667–675, 2006.

[50] X. Wang u. a. MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, 21:3610–3614, 2005.

[51] X. Xie u. a. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434:338–345, 2005.

[52] Y. Lee u. a. The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425:415–419, 2003.

[53] Y. Lee u. a. MicroRNA genes are transcribed by RNA polymerase II. *EMBO Journal*, 23:4051–4060, 2004.

[54] S. Washietl, I. L. Hofacker, and P. F. Stadler. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. U.S.A.*, 102:2454–2459, 2005.

[55] B. Wightman, T. R. Burglin, J. Gatto, P. Arasu, and G. Ruvkun. Negative regulatory sequences in the lin-14 3'-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development. *Genes and Development*, 5:1813–1824, 1991.

[56] B. Wightman, I. Ha, and G. Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75:855–862, 1993.

[57] Z. Yang, Y. W. Ebright, B. Yu, and X. Chen. HEN1 recognizes 21-24 nt small RNA duplexes and deposits a methyl group onto the 2' OH of the 3' terminal nucleotide. *Nucleic Acids Research*, 34:667–675, 2006.

[58] R. Yi, Y. Win, I. G. Macara, and B. R. Cullen. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes and Development*, 17:3011–1016, 2003.

[59] Z. Yu, T. Raabe, and N. B. Hecht. MicroRNA Mirn122a reduces expression of the posttranscriptionally regulated germ cell transition protein 2 (Tnp2) messenger RNA (mRNA) by mRNA cleavage. *Biology of Reproduction*, 73:427–433, 2005.