TECHNISCHE UNIVERSITÄT MÜNCHEN

Fachgebiet für Bioinformatik

Prediction and analysis of microRNAs and their targets in eukaryotic genomes

Martin Sturm

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. K. Schneitz

Prüfer der Dissertation: 1. Univ.-Prof. Dr. D. Frischmann

2. Univ.-Prof. Dr. H.-W. Mewes

Die Dissertation wurde am 16.02.2010 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 01.07.2010 angenommen.

to my family

Abstract

MicroRNA research is not only a very young and fascinating discipline (intensive research began in 2000), but also highly topical.

MicroRNAs (miRNAs) are small, ~22 nucleotide long RNA molecules that play a critical role in gene regulation of eukaryotes by pairing to the mRNAs of protein-coding genes to direct posttranscriptional repression. In the human genome as many as ~700 microRNAs have been identified yet. According to recent proteomic studies a single microRNA is capable of directing repression of hundreds of genes and fine tune protein production of thousands of genes. Consequently, microRNA regulation is associated with many fundamental cellular processes and consistently also with a number of most severe diseases.

To systematically analyze the function of microRNAs and to understand their regulatory role, there are two major challenges to take. First, all microRNAs of an organism have to be detected and second, the targets of each of the microRNAs have to be identified. The focus of this thesis is to provide machine learning based bioinformatics tools that contribute to both fields.

Next generation sequencing technology allows the detection of hundreds of thousands of small RNA molecules in a single experiment. To extract information from the enormous data generated, we have developed *miRanalyzer*. This allows both the detection of known microRNAs together with their expression levels and the discovery of novel microRNAs. For the latter we have developed a prediction model that is based on the *Random Forest* learning scheme and automatic feature selection of a wide spectrum of features. The high levels of accuracy achieved by this

approach is partly based on the fact that we harness the footprint of Dicer processing that is visible in deep-sequencing data for the first time.

For the prediction of microRNA target sites many computational approaches have been developed in recent years. Due to limited biological knowledge about the characteristics of these cis-regulatory sites, the majority of tools focus on the identification of the most prominent feature, the seed match. To keep false-positive rates low those tools additionally require its evolutionary conservation. Naturally many functional sites will therefore be missed. A fraction that might be as large as 40%, according to recent studies.

With *TargetSpy* we have developed a completely different approach. It is based on a machine learning approach that considers several characteristics selected by automatic feature selection. We further utilized a recently published set of high quality binding site data (HITS-CLIP) derived in deep-sequencing experiments to train our model. Extensive evaluations suggest that *TargetSpy* offers high prediction accuracy and may be applicable to a broad taxonomic range of organisms. Further it allows the identification of species-specific target sites.

Zusammenfassung

Die microRNA-Forschung ist nicht nur eine sehr junge und spannende Disziplin (intensive Forschungsarbeiten begannen erst im Jahr 2000), sondern ebenso hochaktuell.

MicroRNAs (miRNAs) sind kleine, etwa 22 Nukleotid lange RNA Moleküle, die durch Bindung an die mRNA von proteinkodierenden Genen post-transkriptionelle Repression bewirken und damit eine wesentliche Rolle in der eukaryotischen Genregulation spielen. Allein im menschlichen Genom wurden bereits etwa 700 microRNAs identifiziert. Laut neuester proteomischer Studien ist eine einzelne microRNA dazu befähigt, hunderte Gene zu reprimieren und die Proteinproduktion tausender Gene in Feinabstimmung zu regulieren. In Folge ist die microRNA-Regulation in vielen grundlegenden zellulären Prozessen involviert und daraus resultierend an einer Reiher schwerwiegender Krankheiten beteiligt.

Um die Funktion von microRNAs systematisch zu untersuchen und ihre regulative Rolle zu verstehen, müssen zwei große Herausforderungen bewältigt werden. Einerseits gilt es alle microRNAs eines Organismus zu entdecken und andererseits die Zielgene jeder microRNA zu identifizieren. Der Schwerpunkt dieser Dissertation liegt darin, auf Maschinellem Lernen basierende Bioinformatik-Tools zu entwickeln, die zu beiden Bereichen beitragen.

Moderne Tiefensequenzierungstechnologie ermöglicht, in einem einzigen Experiment, die Detektierung hunderttausender kleiner RNA Moleküle. Um aus der Datenfülle verwertbare Informationen zu extrahieren, entwickelten wir das Tool *miRanalyzer*. Dieses ermöglicht einerseits die Identifizierung bekannter microRNAs

zusammen mit den jeweiligen Expressionsleveln und andererseits die Entdeckung bisher unbekannter microRNAs. Für Letzteres haben wir ein Vorhersagemodell entwickelt das auf dem Lerner *Random Forest,* sowie der automatischen Feature-Selektion eines breiten Spektrums von microRNA-Eigenschaften basiert. Die Verwendung der Tiefensequenzierung ermöglicht erstmalig die Nutzung der Fußabdrücke aus der Dicer-Prozessierung. Diese werden in den Sequenzierungs-datensätzen erstmalig sichtbar und verhelfen mitunter zu der hohen Vorhersage-genauigkeit dieser Methode.

Zur Vorhersage der microRNA-Bindestellen wurden in den letzten Jahren eine Vielzahl von Ansätzen entwickelt. Aufgrund des begrenzten biologischen Wissens hinsichtlich der Eigenschaften von microRNA-Bindestellen konzentrieren sich diese Ansätze überwiegend auf das prominenteste Charakteristikum, dem Seed-Match. Um die Fehlerraten auf annehmbarem Niveau zu halten, verlangen diese Methoden zudem dessen evolutionäre Konservierung. In Folge dessen werden etliche funktionale Bindestellen übersehen. Ein Anteil, der laut neuester Studien sich in einer Größenordnung von bis zu 40% bewegen kann.

Mit *TargetSpy* entwickelten wir einen gänzlich neuen Ansatz. Dieser basiert auf Maschinellem Lernen und berücksichtigt verschiedenste Eigenschaften, die in einer automatischen Feature-Selektion ausgewählt wurden. Hinzu verwendeten wir die kürzlich publizierten, qualitativ hochwertigen Bindestellen (HITS-CLIP) aus Tiefensequenzierungs-Experimenten um unser Modell zu trainieren. Umfangreiche Evaluierungen lassen die Schlussfolgerung zu, dass *TargetSpy* über hervorragende Vorhersagegenauigkeit verfügt und auf eine große taxonomische Bandbreite von Organismen angewendet werden kann. Zudem ermöglicht es die Identifizierung von Spezies-spezifischen Bindestellen.

Acknowledgements

This work would have not been possible without the help of various people who supported me during this time.

In particular I would like to thank my advisor Prof. Dr. Dmitrij Frishman for his support, guidance, ideas and patience. He constantly motivated and inspired me to develop own attempts to a solution and gave me the necessary trust and space to accomplish it.

Special thanks go to my coworker and friend Dr. Michael Hackenberg. He was an important motivator during the whole endeavor and he constantly and constructively challenged me, resulting in improved work.

I would also like to thank Dr. Philipp Pagel for always helping me out whenever I had questions in statistics and Dr. Andreas Kirschner for his support in machine learning topics. I further thank Dr. Thomas Rattei for keeping the computer system in Weihenstephan in such a good shape, as my research heavily relied on it.

I thank Angelika Fuchs, Claudia Luksch, Dr. Erik Granseth, Frauke Moeller-Beau, Léonie Corry, Nadia Latif, Patrick Tischler, Qibin Luo, Roland Arnold, Sheng Zhao, Sindy Neuman and Stefka Tyanova for many relaxing hours in the kitchen.

I would also like to thank David Langenberger and Oliver Krieg for doing their diploma theses and Christopher Huptas for doing his bachelor thesis under my supervision. I also learned a lot during that time.

I thank Prof. Dr. Hans-Werner Mewes for giving me the opportunity to work at his Bioinformatics Chair.

I am deeply grateful to my parents, my sister, my uncle and my friends who always supported me in their special way and on whom I can always count, although my work is still all Greek to them.

My deepest and most sincere thanks go to Ramona Weiss. She constantly supported me under all circumstances.

Thanks and good luck to everyone!

Table of contents

Chapter 1

Introduction

First the current understanding of the biology of the microRNA pathway will be detailed, followed by an overview of the current state-of-the-art prediction approaches for microRNA genes and microRNA target sites. Subsequently the claim of this work and the contribution to the field will be laid out. Finally the chapter closes with an outline of this thesis.

1.1. Biological background

1.1.1. The history of microRNAs

Albeit their significance, microRNAs escaped detection until 1993. At that time, Victor Ambros and his colleagues Rosalind Lee and Rhonda Feinbaum investigated the function of lin-4, a gene essential for normal temporal control of postembryonic developmental events in *Caenorhabditis elegans*. Several lines of evidence indicated that lin-4, known to negative control LIN-14, does not encode for a protein but instead produces two small RNA transcripts, lin-4S and lin-4L. The smaller molecule, lin-4S, was approximately 22 nucleotides (nt) long, whereas lin-4L was found to be 61 nt in length. They also discovered that the 5' regions of the two RNA molecules were identical and that they had antisense complementarity to seven sites in the 3' untranslated region (3'UTR) of LIN-14. Subsequently they demonstrated

that these sites were crucial for the regulation of LIN-14 and proposed the hypothesis in which the small RNA lin-4S, now recognized as the founding member of the microRNA class, pairs to the 3'UTR of LIN-14 to repress its translation (Lee et al. 1993).

Today it is known, that microRNAs are endogenous ~22 nt long single stranded RNA molecules, found to play a fundamental role in the regulation of gene expression in eukaryotes. With several hundred microRNAs identified in human and several thousand known in total (Griffiths-Jones 2004), the microRNA gene family is one of the most abundant classes of gene regulators in multicellular organisms (Bartel 2004). Considering most recent experimental findings, microRNAs can act by mRNA destabilization and by directly repressing translation of hundreds of genes. The impact of microRNAs on the proteome observed in this context suggests that a microRNA may act as a rheostat, making fine-scale adjustments to protein synthesis from thousands of genes (Baek et al. 2008; Selbach et al. 2008).

1.1.2. Significance of microRNA regulation

So far, however, the biological roles of microRNAs have been elucidated only for a small fraction. Though there are several lines of evidence suggesting that microRNAs play critical roles in most, if not all, physiological processes. Among others microRNAs have been proven to be involved in tissue differentiation, cell growth and proliferation, fat metabolism, cellular signaling, embryonic development and apoptosis (Esquela-Kerscher and Slack 2006).

Therefore dysregulation of microRNAs or their targets, dysfunctions in the microRNA biogenesis or mutations in the mature microRNA or their target site have been shown to lead to various severe diseases (Esquela-Kerscher and Slack 2006). Cancer for example is caused by runaway proliferation of defective cells in combination with their spurious survival. Usually those processes are highly regulated in a coordinated fashion to ensure proper function and to safeguard against defects. Though damage to those genes that drive this complicated regulation, generally referred to as oncogenes and tumor-suppressor genes, may lead to oncogenesis. Since microRNAs have been shown to be involved in those critical

biological processes, it is not surprising that impaired microRNA expression is involved in the formation of cancer (Esquela-Kerscher and Slack 2006; Lu et al. 2008). Other diseases that are associated with microRNA dysregulation are schizophrenia, neurodegenerative diseases like the Parkinson's disease and diabetes. In total the human miRNA-associated disease database (HMDD) currently lists 70 diseases and its size is constantly growing (Esquela-Kerscher and Slack 2006; Lu et al. 2008).

In case of human cancer it was shown that the expression profile of microRNAs reflects the developmental lineage and differentiation stage of tumors. The same publication nicely demonstrated that while mRNA expression profiles were inappropriate to distinguish between poorly differentiated tumors, microRNA expression profiles successfully classified the samples (Lu et al. 2005). Also in case of diabetes, several microRNA expression levels are reported to be impaired in different animal models for type-2 diabetes. Further, two key insulin-responsive proteins, Insig1 and cav2, are validated as direct targets of microRNAs. Moreover microRNAs are generally found to be involved in diabetes-associated diseases (Kolfschoten et al. 2009). Altogether, microRNAs with their critical role in the physiology of living cells and consequently their implications in diseases might prove also useful in the diagnosis and treatment of diseases.

1.1.3. MicroRNA biogenesis and maturation

According to recent findings, microRNA genes occur as distinct transcriptional units as well as polycistronic units in microRNA gene clusters and more than half of all known mammalian microRNAs reside within the introns of protein coding genes or within either the introns or exons of non-coding genes (Lagos-Quintana et al. 2001; Lau et al. 2001; Reinhart et al. 2002). Intronic microRNAs are found to be usually in the same orientation as the protein-coding transcript and are therefore co-expressed with it, as they share the same primary transcript (Baskerville and Bartel 2005; Rodriguez et al. 2004). Only about one tenth are located within the exons of long non-protein coding transcripts (Rodriguez et al. 2004). Very few microRNA are also found to be within untranslated regions of protein-coding genes (Cullen 2004).

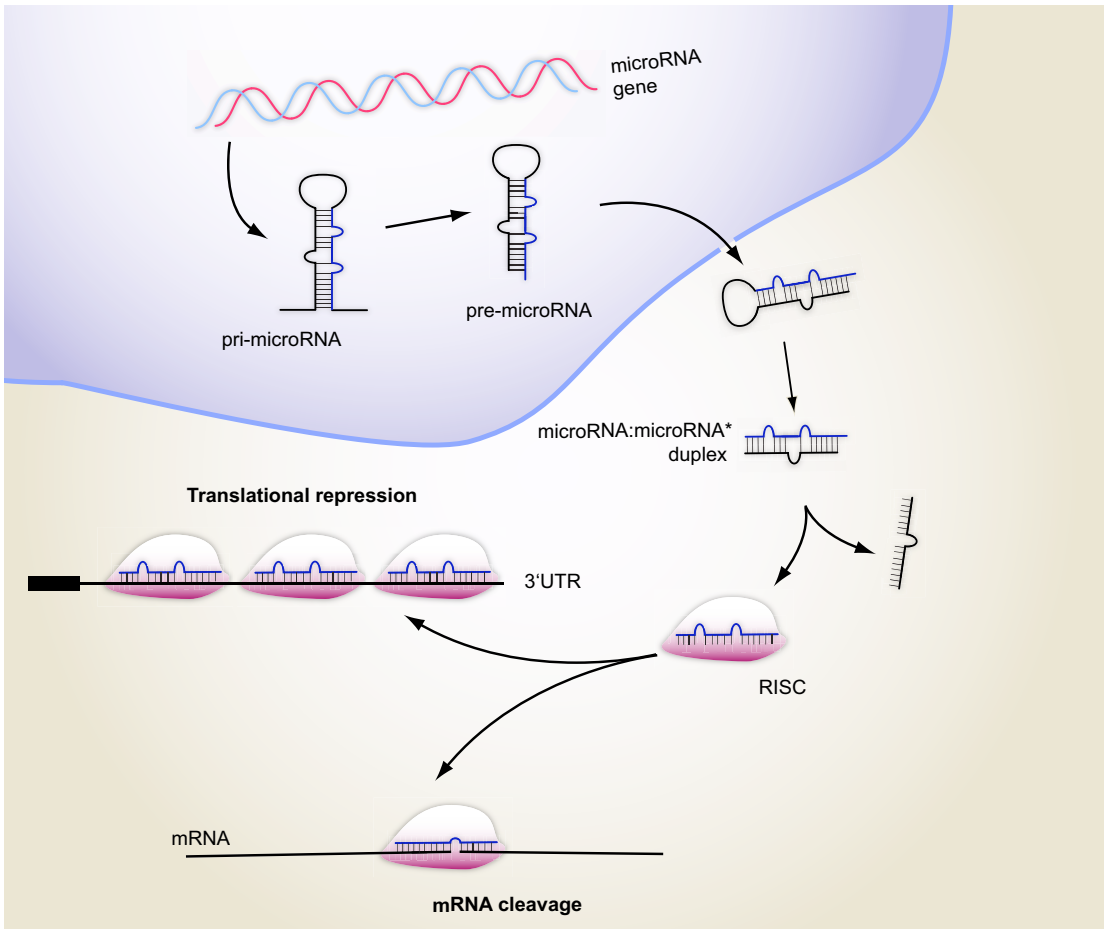


Figure 1: Pathway from microRNA biogenesis to mRNA regulation.
The microRNA gene is transcribed by RNA polymerase II (pol II) into the primary transcript (pri-microRNA). Still in the nucleus Drosha mediates the 'cropping' step, a procedure that removes flanking sequences, resulting in the ~70 nucleotide long pre-microRNA. After the relocation into the cytoplasm by exportin-5, Dicer, a cytoplasmatic RNase III, performs the second cleaving step called 'dicing' to produce the microRNA:microRNA* duplex. Subsequently the duplex is separated and one strand gets incorporated into the RISC, while the other strand is degraded. Finally the microRNA loaded RISC is potent for regulating protein production, either by translational repression or mRNA cleavage.

Current models suggest that microRNA biogenesis and maturation is a stepwise process (see Figure 1) that starts in the nucleus and ends in the cytoplasm. First, microRNAs are transcribed from RNA polymerase II (in rare cases also by RNApol III) to primary transcripts (pri-microRNAs) ranging from several hundred to thousands of nucleotides in length (Cai et al. 2004; Lee et al. 2004). Subsequently the microprocessor complex, consisting of the double-stranded RNA binding protein Pasha and the ribonuclease III (RNase III) endonuclease Drosha, processes the transcript by cleaving the flanking sequences of the pri-microRNA. An

approximately 70-nucleotide stem-loop structure named precursor microRNA (pre-microRNA) with two nucleotide 3' single-stranded overhanging ends, typical for RNase III, as well as an 5' phosphate and 3' hydroxy termini is released (Denli et al. 2004; Han et al. 2004; Lee et al. 2003; Lee et al. 2002). The cleavage site marks already the mature microRNA that can either reside in the 5' or in the 3' arm of the pre-microRNA.

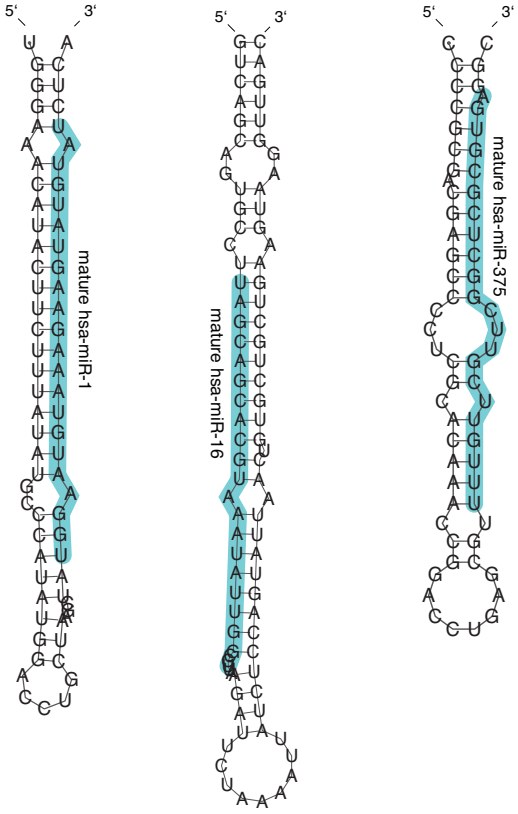


Figure 2: Secondary structure of three pre-microRNAs predicted by RNAduplex with the mature microRNA sequences highlighted in green.
The two mature microRNA sequences, miR-1 and miR-375 are located on the 3' arm of the pre-microRNA hairpin structure; miR-16 is on the 5' arm. According to our current understanding there seems to be no rigid pattern in terms of which strand is chosen for the mature microRNA. The only evidence we know of is that the thermodynamic stability at the microRNAs 5' end seems to play a critical role in the selection process in that the strand with the lower stability is usually associated with RISC.

The characteristic hairpin structure (Figure 2) is recognized by Exportin-5 that subsequently exports the pre-microRNA from the nucleus into the cytoplasm (Zeng and Cullen 2004). Later on it is processed by Dicer, again a RNase III, removing the loop from the stem, releasing a ~22 nucleotide long microRNA:microRNA* duplex (Bernstein et al. 2001; Forstemann et al. 2005; Hutvagner et al. 2001).

Finally one strand (mature microRNA) of that duplex is incorporated into the Argonaute protein of the RNA-induced silencing complex (RISC), while the other strand (microRNA*) is degraded. Supposedly, the strand that is thermodynamically less stable paired at its 5'end is chosen as the mature microRNA (Khvorova et al. 2003; Schwarz et al. 2003). However a few cases are reported where both strands of the duplex seem to be chosen to enter the RISC (Lagos-Quintana et al. 2002; Schwarz et al. 2003).

1.1.4. Mechanism of microRNA-mediated repression

MicroRNA directed posttranscriptional regulation of gene expression may be exerted by the two different mechanisms *mRNA cleavage* and *translational repression* (Carthew 2006; Pillai et al. 2007).

Upon binding of a microRNA to its target, the RISC functions as an endonuclease and cleaves the mRNA between the 10^{th} and 11^{th} nucleotide if the target site exhibits perfect or near perfect Watson-Crick base-parings to the full microRNA sequence. Although such binding sites are found both in the coding sequence and in the untranslated region of protein coding mRNAs, most sites reside in the coding region. Since cleavage leads to degradation of the mRNA, the impact of microRNAs can be observed in reduced mRNA expression levels. Cleavage has been shown to be the predominate mechanism in plants (Hutvagner and Zamore 2002; Llave et al. 2002; Tang et al. 2003).

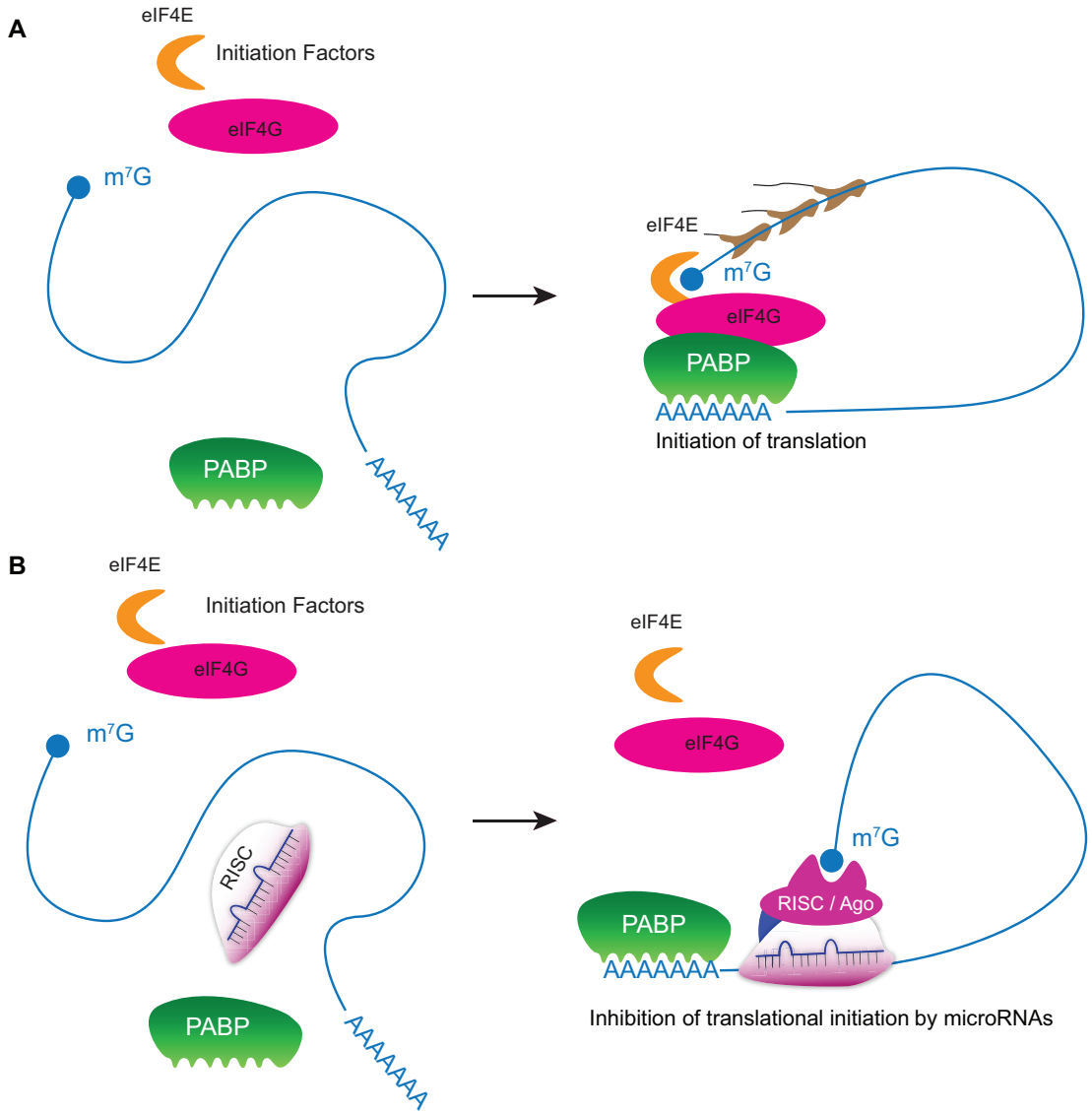


Figure 3: Schematic illustration of microRNA directed repression of translational initiation.
A) Normal translational initiation in the absence of microRNAs. First, the translational initiation factor eIF4E binds to the 7-methyl-guanine (m^7G) cap of the transcript. Second, eIF4G binds to eIF4E and the poly(A)-binding protein (PABP) to build a closed loop for efficient translation. B) In case the microRNA loaded RISC complex has bound to a target site in the 3'UTR of the transcript, Ago proteins, present in the RISC, compete with eIF4E for binding to the m^7G cap. That way the ribosome is not capable of initiating the translation process and hence protein production is prevented.

In animals however only few microRNAs have sufficient complementarity to their targets for mRNA cleavage. Therefore the prevailing regulation type is translational repression, a mechanism that requires less complementarity. Instead of slicing the mRNA, the effective translation of mRNA into protein is hindered (Brennecke et al. 2005; Lai 2004; Lewis et al. 2005; Lewis et al. 2003).

Yet, the underlying mechanistic of the repression and the actual determinants for target recognition have not been reliably unveiled. Regarding the mechanistic, recent studies in mammals strongly suggest that by blocking the translational initiation process protein production is prevented (Figure 3). The loaded RISC binds to a target site in the 3'UTR of the transcript and the Ago protein subsequently competes with the translational initiation factor eIF4E for binding the m^7G cap. Hence the initiation factors cannot build the initiation complex necessary for the ribosomes to bind to the mRNA and start translation. (Kiriakidou et al. 2007; Mathonnet et al. 2007; Meister 2007; Thermann and Hentze 2007; Wakiyama et al. 2007). Though there is also experimental support for the hypothesis, that protein production is prevented by blocking translational elongation (Nottrott et al. 2006; Olsen and Ambros 1999; Peters and Meister 2007). Consensus, however, exists, in that the required cis-regulatory sites reside mostly in the 3'UTR.

1.1.5. Target site recognition for translational repression

On the basis of experimental validation of microRNA binding sites in their physiological context parameters that are responsible for proper recognition of microRNA target sites are elucidated. Though as more analyzes were performed contradictory statements on the determinants were given leading to increasing complexity of the topic (Didiano and Hobert 2008).

The most prominent feature, however, was found to be perfect Watson-Crick base pairings to the microRNA 5' end. Moreover in this region, generally referred to as the microRNA "seed match", no G:U base-pairs were observed. Additionally, point mutations in the 5' region causing loss-of-function supported the seed as an important factor (Lee et al. 1993; Reinhart et al. 2000). Functional seed matches were found to be between 6 and 8 nt in length (Figure 4).

The 6-mer seed match exhibits perfect base pairings to the microRNA nucleotides 2-7 and is considered as the least effective. The 7-mer seed match can either be a perfect base pairing between microRNA nucleotides 1-7 or 2-8, while the second, also referred to as a 7mer m8 seed match, seems to be more effective.

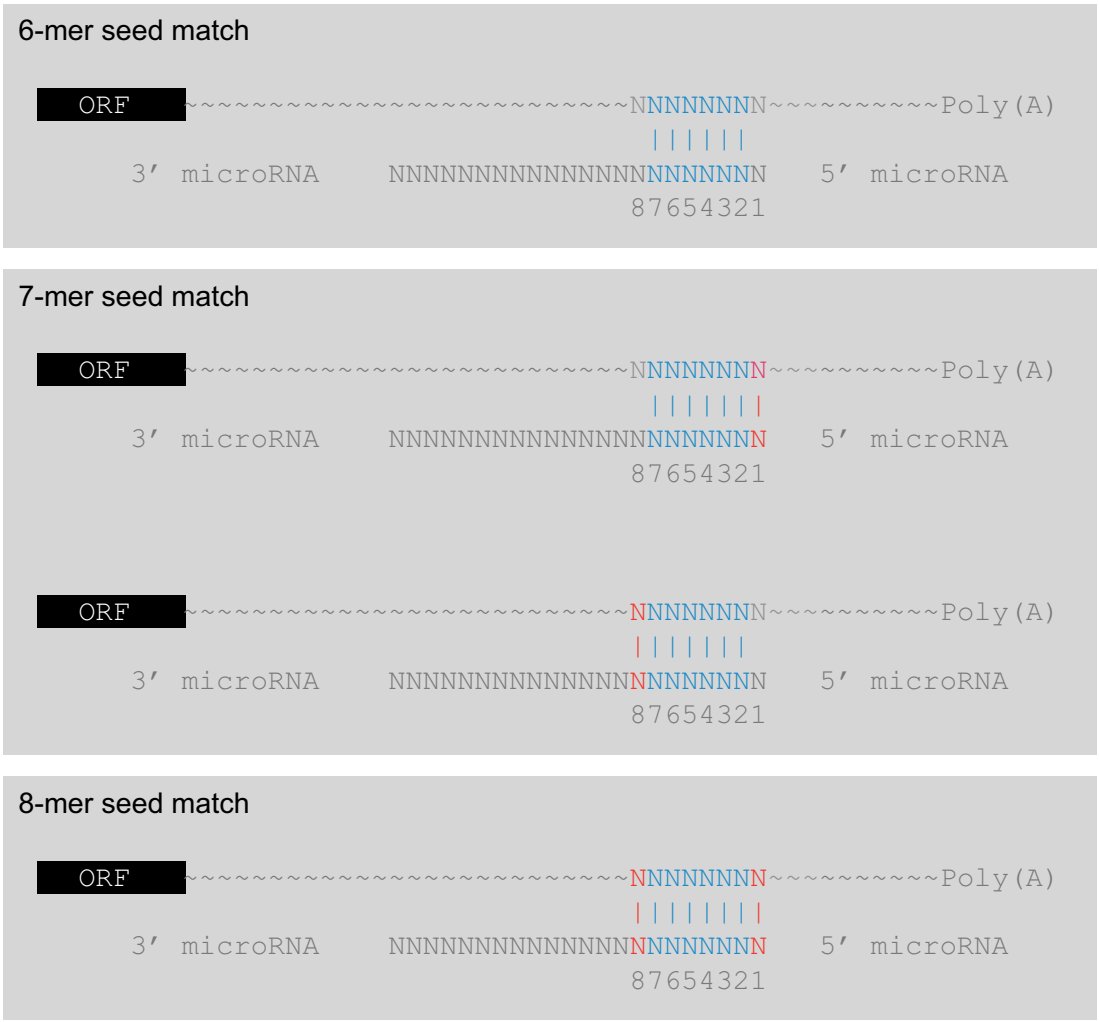


Figure 4: Types of microRNA seed matches.

The 6-mer seed (nucleotides 2-7) is colored blue, the additional match to either microRNA position 1 or 8 is colored red. The 6-mer seed match is a six nucleotide long stretch in the mRNA that shows perfect reverse complementarity to the 6-mer seed in the microRNA. If in addition to that match either the first or the eighth nucleotide pairs as well, one speaks of a 7-mer seed match. An 8-mer seed match exists if both, the first and the eighth nucleotide match.

Due to experimental findings, indicating a higher efficiency when the first nucleotide in the target site is an Adenosine (A), no matter if a base pairing occurs at that position, such 7-mer sites are called 7mer-A1 sites. Finally the most effective seed match found is the 8-mer that exhibits base pairings to the microRNA nucleotides 1-8. Again, those sites showing an A opposite to the first microRNA nucleotide seem to be even more effective (Grimson et al. 2007).

However, not all biological functional binding sites do exhibit such a perfect pairing to the microRNA seed, like for example binding sites in the gene LIN-14 for lin-4, the founding member of the microRNA class (He and Hannon 2004). Also the work of Didiano and Hobert clearly demonstrate that a seed is not a general requirement of functional target sites (Didiano and Hobert 2006; Didiano and Hobert 2008). The current view on this topic is that up to approximately 60% of all regulated binding sites exhibit a perfect seed match (Selbach et al. 2008) and therefore 40% or more of all microRNA targets do not show this perfect Watson-Crick base pairing to the microRNA 5' end. Based on systematic mutation experiments as well as extensive bioinformatics analyzes, it was affirmed, that this other class of target sites exists. These target sites show imperfect matches to the microRNA seed but additional matches to the 3' end of the microRNA. It was therefore proposed that target sites of this class compensate the disruptions in the seed region by the base pairings observed to the microRNA 3' end (Brennecke et al. 2005; Doench and Sharp 2004; Lim et al. 2005).

Although it has been reported that in some cases perfect pairing to an 8-mer seed appear to be sufficient for repression (Brennecke et al. 2005; Doench and Sharp 2004; Lai et al. 2005), seed sites alone do not always guarantee functionality. Interestingly, the magnitude of repression has been found to be highly variable depending on the UTR context (Farh et al. 2005). So far experimental evidence for determinants beyond pure seed pairing has been found for the accessibility of target sites to the RISC complex (Kertesz et al. 2007). Further it was demonstrated that closely spaced sites often show synergetic action and that additional base pairings to microRNA positions 12-17 were strongly correlated to down-regulation. Additionally there is experimental support that functional target sites often reside in AU rich context, a determinant that is certainly connected to the local RISC accessibility. Analyzes on the position of target sites further suggest that functional sites seem to reside preferentially in the 3'UTRs, but not too close to the stop codon and that the most effective regulation emanates from sites near both ends of the 3'UTR (Grimson et al. 2007). However not all of these finding could be affirmed. Didiano and Hobert analyzed the RISC accessibility and also the local AU content, but could not detect any strong correlation between those criteria and the present and

strength of the measured regulation. Also when they considered the position in the 3'UTR and the context of microRNA target sites they found that a relocation of a functional site towards the beginning of the 3'UTR but with a clear distance to the stop codon showed loss-of-function. Thought it could be verified that a certain distance between two sites is mandatory for regulation. Moreover an AU rich 25bp long region downstream of the target sites was determined as being necessary for proper regulation (Didiano and Hobert 2008).

1.2. Computational prediction approaches

Due to the difficulty of detecting microRNAs and their targets systematically by experimental techniques, models for predicting microRNA genes and microRNA target sites were built shortly after the microRNA pathway was discovered. These models are based on very limited experimental data, however the basic principles that were used to build these are still found on the core of most current state-of-the-art predictions approaches.

1.2.1. MicroRNA gene prediction

As described in Section 1.1.3, mature microRNAs stem from pre-microRNA molecules, showing a characteristic stem-loop hairpin structure. However when scanning the human genome for pre-microRNA-like hairpins, one identifies about 11 million structures (Bentwich et al. 2005). Therefore most of the existing prediction approaches utilize comparative genomics information next to structural features.

The first approach MiRscan, published in 2003, is based on the observations of seven components of 50 conserved pre-microRNA hairpin structures corresponding to the known microRNAs in *Caenorhabditis elegans* at that time. From the ~40,000 conserved hairpin structures identified in a sliding window based genome scanning approach in *C. elegans*, a total number of 35 stem-loop structures additionally to those that are already known were predicted with high confidence. Subsequently 16 of the 35 were experimentally verified, while the rest were conjectured to be false positives. For the human genome 107 microRNA candidates were predicted (Lim et al. 2003). Later, miRseeker (Lai et al. 2003) was developed for *Drosophila melanogaster* by recognizing microRNA specific conservation patterns. This

approach predicted a total of 48 microRNA candidates. In a similar approach, the conservation patterns of known microRNAs were used to predict novel candidates and to estimate that about 1,000 microRNAs might exist in vertebrate genomes (Berezikov et al. 2005). For plants MIRcheck (Jones-Rhoades and Bartel 2004) and MIRFINDER (Bonnet et al. 2004) were developed.

Instead of searching for conserved sequences folding to hairpin structures, Xie et al. analyzed 3'UTR sequences for conserved 8-mer sequences for overrepresentation. They found that many of those corresponded to the reverse complement of the seed region of know microRNAs. In total 129 microRNA candidates were predicted in human by considering those conserved 8-mer sequence motifs for which no corresponding microRNA was known (Xie et al. 2005). A further strategy for predicting homologs of known microRNAs was developed based on genomic alignments to microRNA sequences and structures (Legendre et al. 2005; Nam et al. 2005; Wang et al. 2005).

All these approaches so far make use of evolutionary conservation and are therefore obviously unsuited for the detection of novel, unconserved, species-specific microRNA candidates (Berezikov et al. 2006a). The number of non-conserved microRNAs, however, was shown to be tremendous (Bentwich et al. 2005). As a consequence several *ab initio* methods based on machine-learning approaches have been developed to fill that segment (Huang et al. 2007; Jiang et al. 2007; Nam et al. 2006; Sewer et al. 2005; Sheng et al. 2007; Xue et al. 2005).

Bentwich et al. constructed a method, PalGrade, in that they generate all possible hairpin structures from the genome and assign them a stability score according to the tendency to appear in many folding configurations. Further structural and sequence-based features are extracted and condensed into a single score. Based on their probabilistic approach they estimated the number of conserved microRNAs in human to be ~400-500. Including non-conserved microRNAs they approximated the number human microRNA to ~800 and therewith proposed that the world of microRNAs is larger than initially believed (Bentwich et al. 2005).

All computational prediction approaches that analyze genomic DNA structures that resemble known microRNA precursors are affected by sensitivity problems and sizeable false positive rates. Subsequently experimental test of newly predicted microRNAs are required. This however is a fundamental issue, since the bandwidth of microRNA expression is enormous, ranging from a just few molecules per cell to tens of thousands, and the detection of lowly expressed molecules is reaching the limits of what is technically feasible (Friedlander et al. 2008).

With the availability of next-generation sequencing platforms such as those from Solexa/Illumina and 454 Life Sciences/Roche, DNA can be sequenced orders of magnitude faster and cheaper compared to standard Sanger sequencing. This novel technology allows the detection and profiling of known and novel microRNAs at unprecedented sensitivities. To analyze the enormous output that is generated by this new technology, several computational challenges have to be taken. MiRDeep is the first tool that is capable of detecting known microRNAs from these data. By accounting for traces left by dicer processing, visible by this technology for the first time, miRDeep identifies novel microRNAs with high accuracy and robustness. In summary, this approach reports ~230 previously unknown microRNAs in dog, human and worm (Friedlander et al. 2008).

1.2.2. MicroRNA target site prediction

The discovery of microRNAs in multicellular organisms raised various absorbing questions. Probably the most fascinating and challenging question is what the function of these small non-coding RNA molecules might be. The answer to that problem lies hidden in their targets. Once we know all the actual targets of all microRNAs we will be able to integrate these interactions with other regulatory information to build huge networks. Ultimately we will then be able to infer functional information for each microRNA from these networks.

In plants, many targets can be confidentially predicted by simply searching for extensive sequence complementarity (see 1.1.4) to the microRNA (Rhoades et al. 2002). In contrary, extensive complementarity leading to cleavage of the targeted mRNA occurs only occasionally in animals (Yekta et al. 2004). Hence the challenge

in the accurate detection of metazoan microRNAs is to perform a genome-wide computational search to find most of the regulatory targets without concurrently introducing too many false positives (Bartel 2009).

Based on the very early observations of reverse sequence complementarity of microRNA lin-4 to multiple conserved sites in the 3'UTR of mRNA lin-14 first models for predicting microRNA target sites have been developed. Today umpteen approaches are published. These can be classified in many different ways. For example one may draw a distinction depending on the methodology of the methods. In that case there are traditional approaches relying on specific base-pairing rules and machine learning based approaches that try to identify patterns within true and false microRNA-target duplexes. Another possible arrangement in groups relies on the requirements a target site has to fulfill to be selected as a potential target site candidate. Here, meaningful classification criteria are the necessity of a seed match, a seed match and the usage of conservation or none of both requirements. In the following, we used both measures to categorize widely used prediction approaches (Table 1).

	Traditional approach	**Machine Learning approach**
Seed match requirement and usage of cross-species information	PicTar PITA TOP EIMMo MiRBase Targets MiRanda DIANA-microT TargetScanS TargetRank	MirTarget2 TargetMiner
Seed match requirement	PITA ALL TargetScanS non-conserved MicroInspector	NBmiRTar
No seed match requirement	RNA22 MirWIP	

Table 1: Overview of the most currently used microRNA target prediction approaches.

The majority of the tools are designed as traditional approaches requiring a seed match. As described earlier (see 1.1.5), the formation of a seed match is the most prominent determinant of target site detection. However, almost every approach is

using a slightly different definition of a seed match. The different levels of observed sensitivity and specificity of a particular seed definition and the major goals the authors pursued with their approaches mainly drive this. In the end it is also influenced by the knowledge at the time of publication. For instance, PicTar uses 7-mer seed matches that are either pairing to nucleotides 1-7 or 2-8 of the microRNA 5' end. TargetScanS demands a perfect 8-mer seed match pairing to the microRNA nucleotides 1-8 with a mandatory 'A' opposite to the first nucleotide. And PITA TOP also requires the 8-mer seed match, however it does not call for the 'A' opposite to the first microRNA nucleotide. In contrast PITA ALL even allows for 6-mer seed matches.

It was shown that especially in cases where the target site or at least the seed match is conserved in other species the false-positive rate is markedly reduced. Since as a consequence the prediction reliability improves (Brennecke et al. 2005; Krek et al. 2005; Lewis et al. 2005; Lewis et al. 2003) many of the approaches are making use of cross-species information.

However, conservation is not always applied the same way. Many different approaches have been developed. PicTar offers two configurations - target sites that are conserved in human, chimp, mouse, rat and dog (4-way) and the more confident prediction set (5-way) that shows additional conservation in chicken. Interestingly, only the middle 6-mer of the two overlapping 7-mer seed matches needs to be perfectly conserved in these species. The 7^{th} conserved nucleotide can be opposite to microRNA nucleotide 1 or 8 and does not necessarily be the same in all species. Early versions of TargetScanS started with a similar usage and called the resulting prediction sets 'highly conserved', 'conserved' and 'poorly conserved' depending on the number of species in which the full length 8-mer seed match is conserved. Today TargetScanS is using phylogenetic trees to calculate cross-species conservation. MiRanda predictions make use of the precalculated phastCons conservation score.

Besides these two criteria for predicting microRNA target sites several other intrinsic information is used to rank target site candidates. For instance, PITA calculates, what the authors call, a site accessibility score (Kertesz et al. 2007). PicTar computes a maximum likelihood score by using an HMM model to rank target sites according to

combinations of microRNAs that are cooperatively targeting that transcript (Krek et al. 2005). As outlined in Section 1.1.5, Grimson et al. 2007 analyzed determinants beyond seed pairing. The outcome of that work was subsequently integrated into TargetScanS in terms of a context score. Finally, minimum free energy (MFE) calculations for identification of energetically stable hybrids can be found in almost any approach.

In several research fields as for example the transcription factor binding site prediction, the application of machine learning approaches significantly improved prediction performance. Recently, efforts were taken to improve microRNA target site prediction performance of traditional approaches by extracting several seed-based and seed-independent features from target site candidates to classify them with the help of machine learning techniques (Bandyopadhyay and Mitra 2009; Wang and El Naqa 2008; Yousef et al. 2007). All of those approaches however suffer from a lack of sufficient training data. As especially non-functional binding sites are as good as not existent, scrambled or randomized sequences were utilized for negative set generation.

Two recent proteomic studies indicate that only 30–45% of proteins associated with microRNA regulation contain perfectly matched, conserved seed elements in the 3'UTRs of their transcripts. In *C. elegans* it was found, that 40% of the verified target sites reside within 3'UTRs that poorly align between *C. elegans* and *C. briggsae*. The proteomic studies further showed that only up to 60% of the down-regulated proteins exhibit a perfect seed match in the 3'UTR of their messenger RNA (Baek et al. 2008; Hammell et al. 2008; Selbach et al. 2008). In conclusion, a significant fraction of functionally important target sites are thought to not show a perfect seed match and can therefore not be predicted by the approaches discussed so far.

In 2006, Miranda et al. presented a novel pattern-based approach (RNA22) for the identification of microRNA binding sites that does neither use cross-species information, nor requires a perfect seed match (Miranda et al. 2006). Another tool that falls into that class is mirWIP. Unfortunately predictions are only available for *C. elegans* (Hammell et al. 2008).

Altogether, currently available prediction methods are diverse and all have room for improvements (Bartel 2009) as the key factors for proper target site detection have not been revealed yet.

1.3. Contribution of this dissertation

The objectives of this thesis are to develop methods that contribute to the two major problems in field of microRNA regulation.

The first method, *miRanalyzer*, was developed to analyze data of deep-sequencing experiments. The main focus of this work was to identify unknown microRNAs. For accomplishment, a machine learning approach was set up and trained on experimental data to learn the characteristics of pre-microRNA hairpin structures.

The objective of the second method, *TargetSpy*, is to generate a general model for the prediction of microRNA target sites, including those that do not show a perfect seed match and are not conserved in other species. TargetSpy is a machine learning approach, based on target site-specific features and a training set of high quality experimentally determined Ago-binding sites. As will be seen, this general approach achieves superior performance compared to state-of-the-art prediction methods.

1.4. Thesis Outline

With the availability of next generation sequencing technology, vast amounts of RNA molecules can be sequenced in a single experiment. To convert the extensive amount of data into knowledge, advanced bioinformatics tools are necessary to process the data and to extract significant information. Consequently, we have developed *miRanalyzer*, a pipeline for the comprehensive analysis of deep sequencing experiments of small RNA molecules. Beside the detection of known microRNAs, *miRanalyzer* is capable of predicting novel microRNAs, too. Chapter 2 of this thesis will explain in detail, how the pipeline works and how we extract and utilize valuable information like the Dicer footprint for a more accurate machine learning based prediction model.

Chapter 3 focuses on the analysis of predicted microRNA target sites. Commonly the core of nowadays prediction approaches is to search for reverse complementary seed regions that are further conserved in various species. As a consequence the location and the amount of predicted target sites are highly dependent on the nucleotide composition of the seed sequences. Concretely, microRNAs and transcripts are therefore examined in the light of their GC content and their conservation. In addition background models are generated to distinguish between characteristics unique to microRNA target sites and those that are a consequence of the employed prediction model. Lastly we reexamined target sites in terms of position specific occurrences of single nucleotide polymorphisms.

Subsequently, Chapter 4 is dedicated to *TargetSpy*, our approach to the prediction of microRNA targets sites, disregarding microRNA seed match and conservation requirements. The model is based on machine learning, automatic feature selection and a wide range of features covering current biological knowledge. It is further trained on HITS CLIP data that currently constitute the most accurate experimentally derived (deep sequencing) evidence of microRNA target sites. Altogether, *TargetSpy* is perfectly suited for predicting seed based and seed free target sites, a sub class of substantial size that is missed by most currently available mainstream tools (see Table 1). The abdication of conservation features *TargetSpy* to analyze species-specific microRNA-target interactions also in unconserved genomic sequences. In order to estimate prediction accuracy, we performed extensive evaluations and benchmarked our method with state-of-the-art approaches. In total, the results suggest that *TargetSpy* performs excellent, especially on the human dataset revealing fold-change in protein production for five selected microRNAs, where it shows superior performance in all three target site classes (see 1.2.2). In conclusion, this method contributes significantly to the field of computational prediction of microRNA target sites.

In the final Chapter 5 we will summarizes the work presented in this thesis and provide and outlook on promising extensions to our work for future research.

Chapter 2

MiRanalyzer: Analysis and prediction of microRNAs in deep sequencing data

Next generation sequencing is revolutionizing genomics since these new techniques allow now the sequencing of even small RNA molecules and the estimation of their expression levels. Hence, microRNA sequence data is expected to increase notably during the next years. Consequently, there will be a high demand of bioinformatics tools to cope with the several gigabytes of sequence data generated in each single deep-sequencing experiment.

Given this scene, we developed miRanalyzer, a user-friendly web server tool for the analysis of deep-sequencing experiments for small RNAs. MiRanalyzer broadens the scope of currently existing standalone tools by adding new types of analyzes. The web server tool requires a simple input file containing a list of sequence reads and the number of times each read has been obtained (expression levels). Using these data, miRanalyzer 1) detects all known microRNA sequences annotated in miRBase, 2) finds all perfect matches against other libraries of transcribed sequences as mRNA, RepBase and RFam and, 3) identifies unknown microRNAs. The prediction of new microRNAs is an especially important point as there are many species with very few known microRNAs.

Therefore, we implemented a machine-learning algorithm for the prediction of new microRNAs that reaches area under the curve values of 97.9% and recall values of up to 75% on unseen data. The web tool summarizes all the described steps in a single output page, which provides a comprehensive overview of the analysis, adding a link to more detailed output pages for each analysis module.

MiRanalyzer is available at http://web.bioinformatics.cicbiogune.es/microRNA/.

2.1. Background

The recent years witnessed a profound change in our understanding of the regulation of gene expression. Small non-coding RNA especially came into focus as it became clear that they are key players in many cellular processes by post-transcriptionally regulating gene expression via either degradation, translational repression, or both (Kim and Nam 2006; Lagos-Quintana et al. 2001). MicroRNAs, belonging to the family of small non-coding RNAs, are endogenous in many animal and plant genomes and are now recognized to be one of the major regulatory gene families in eukaryotic cells. They are believed to regulate the expression of around one third of all genes in the human genome, involved in many fundamental processes like metabolism, development and regulation of the nervous and immune systems (Bagasra and Prilliman 2004; Ouellet et al. 2006). Furthermore, it has been reported that some microRNAs are actively involved in the development of pathologies like cancer (Lu et al. 2005).

The traditional experimental approach to measure the expression levels of microRNAs involves cloning and Sanger sequencing. This is an expensive and time-consuming procedure, and as a consequence, relatively little expression data is currently available (see (Landgraf et al. 2007) for a microRNA expression atlas). Moreover, the huge range of microRNA expression from tens of thousands to just few molecules per cell complicates the detection of microRNAs expressed at low copy numbers. Hence many undetected microRNA may exist even in well-explored species. Recently, microRNA expression profiling panels became available for measuring expression levels by means of hybridization. These panels allow a high-

throughput detection of microRNA expression. However, they do not allow the detection of new microRNAs.

Next generation sequencing platforms like Genome Analyzer (Illumina Inc.) or Genome SequencerTM FLX (454 Life ScienceTM and Roche Applied Science) became recently available for the sequencing of small RNA molecules that allow both the detection of expression levels and new microRNA sequences at high speed and sensitivity and low cost. However, each sequencing experiment produces up to three Gbp of sequence data, whose analysis represents an important bioinformatics challenge.

Given the importance of microRNAs in the regulation of gene expression, in the coming years many deep-sequencing experiments will be carried out to detect and measure their expression. Therefore, user-friendly tools are required for the processing of the enormous amount of data that will be generated. To our knowledge, so far there is only one standalone tool available for the analysis of deep sequencing microRNA data: miRDeep published by Friedländer et al. (2008).

On the other hand, the prediction of microRNA genes has been extensively employed over the past years and several distinct approaches have been developed. Some of the methods used in the purely computational detection approaches were, for example, conservation of certain regions - phylogenetic shadowing (Berezikov et al. 2005), different machine learning methods like support vector machines using structure-sequence features (Xue et al. 2005), random forest models (Jiang et al. 2007) or probabilistic co-learning models (Nam et al. 2006). Bentwich et al. (2005) used further features like the stability of the hairpin together with an experimental validation. The main drawbacks of these approaches are that they are either limited to conserved microRNAs or that they tend to have a high rate of false positive predictions. However, new sequencing experiments open new possibilities in the prediction of microRNAs, allowing the generation of previously unavailable characteristics like, for example, the traces left by dicer processing.

Consequently, we have developed miRanalyzer, a web server tool that implements all necessary methods for a comprehensive analysis of deep sequencing experiments

of small RNA molecules. It detects known microRNAs annotated in miRBase and matches in other transcribed sequences (RNA, RFam and RepBase). Furthermore, miRanalyzer performs highly accurate (Area Under the Curve - AUC - value of 97.9%) predictions for new microRNAs, by utilization of machine learning techniques and novel features that capitalize the additional information provided by deep-sequencing data. Furthermore our approach directly learned from nature, as we trained on experimental data, in contrary to training sets that are artificially generated by shuffling sequences. This high accuracy is important for the identification of novel microRNAs, a process that usually results in high false positive rates. The tool also includes a Perl script for the proper generation of the input file using the Genome Analyzer (Illumina Inc.) pipeline results. Currently, miRanalyzer works for nine frequently used model species (human, mouse, rat, fruit-fly, round-worm, zebrafish, dog, chicken and the protozoan *Giardia lamblia*).

2.2. Material and Methods

2.2.1. Sequence data

MiRanalyzer uses the newest genome assembly of each species, downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/downloads.html): *Homo sapiens* (hg18, NCBI 36.1), *Mus musculus* (mm8, NCBI 36), *Rattus norvegicus* (rn4, version 3.4), *Drosophila melanogaster* (dm3, BDGP Release 5), *Caenorhabditis elegans* (ce6, WUSTL School of Medicine GSC and Sanger Institute version WS190), *Canis familiaris* (canFam2, v2.0) , *Danio rerio* (danRer5), *Gallus gallus* (galGal3). Data for *Giardia lamblia* (gli1) were derived from the GiardiaDB (http://giardiadb.org/giardiadb/).

The mRNA sequence data were derived from different databases: *H. sapiens*, *M. musculus*, *R. norvegicus* and *D. rerio* from the NCBI RefSeq database (ftp://ftp.ncbi.nih.gov/refseq/), *D. melanogaster* from FlyBase (http://flybase.org/) and *C. elegans* from WormBase (http://www.wormbase.org/). The mRNA sequences for *C. familiaris* were extracted from the genomic sequence using the Galaxy platform (Giardine et al. 2005).

In addition, mature microRNA sequences were derived from miRBase version 12.0 (http://microrna.sanger.ac.uk/sequences/); RNA sequences included in RFam version 9.0 (Gardner et al. 2009) were downloaded from http://rfam.sanger.ac.uk/; and RepBase version 10.10 (Jurka et al. 2005) were obtained from http://www.girinst.org/. Annotations and genomic coordinates of RepeatMasker and PhastCons elements where downloaded from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables?command=start).

We used deep–sequencing data from three different experiments: a) the combined *C. elegans* data (accession no. GSE6282 and GSE5990 from GEO database at NCBI), which have been used also in (Friedlander et al. 2008) with a total of 205,575 unique reads, b) data from human HeLa cells (Friedlander et al. 2008) with accession no. GSE10829 and 319,939 unique reads, and c) data from rat hepatocytes with 22,086 unique reads generated in the CIC bioGune lab, publicly available on our website (http://web.bioinformatics.cicbiogune.es/microRNA/defaultReads.txt).

2.2.2. Generating 'unknown mature-star' sequences

We generated the unknown star sequences by means of the miRBase precursor and mature sequences. First, we calculate the secondary structures for all hairpins using RNAfold (Hofacker 2003) with parameters '-noLP'. Then, we detect the coordinates of the mature microRNAs within the pre-microRNA hairpin. By means of these coordinates, the information of the secondary structure and the characteristic "2-nt 3' overhang" caused by Dicer, we extracted the corresponding sequence pairing with the mature microRNA.

2.2.3. Read Alignment

Read sequences often contain adapter sequences (see standard protocol of small RNA sample preparation at http://www.illumina.com/) at its 3' ends. Therefore, miRanalyzer has two alignment options depending on whether the reads have adapter sequences or not. In general, the tool generate a prefix tree of all input reads and subsequently walk in a single run over the genome to detect the reads. By default, miRanalyzer assumes the existence of adapter sequences and therefore, first detects matches of a subsequence of 16 bp starting at the 5' end of the read. When

miRanalyzer detects an initial match, it expands the subsequence as long as a perfect match is given. Finally, only matches of the longest subsequence are retained.

2.2.4. **Ontological analysis**

We used a recently published tool, Annotation-Modules (Hackenberg and Matthiesen 2008), to pre-calculate the significant annotations of all target gene lists for all microRNAs in the miRBase (12.0). Currently, the user can choose between two different target site prediction methods: miRBase target site predictions by miRanda software (Enright et al. 2003) and TargetScan (Lewis et al. 2005).

2.2.5. Secondary structure prediction

For predicting the secondary structure and its minimum free energy (MFE) we utilized the Vienna RNA package (Hofacker 2003).

2.2.6. Training and test sets

For the machine learning approach we created three data sets, one from each of the three species: *Homo sapiens*, *Caenorhabditis elegans* and *Rattus norvegicus*. First, we extracted all pre-microRNA candidates from the experimental dataset that could be mapped to a known microRNA and labeled them as positive instances. Second, we selected an equal amount of pre-microRNA candidates from the same dataset by random selection with the known microRNAs removed and labeled them as negative. In total we obtained a dataset of 612 instances in human, 468 instances in worm and 376 instances in rat.

2.2.7. **Features**

We created a broad variety of features associated with nucleotide sequence, structure and energy. Table 2 lists all the features used in this work.

Feature Name	Description of the feature
Read count	Number of reads mapping to the pre-microRNA
Length	The length of the longest hairpin structure
Stem length	The length of the longest hairpin structure stem
Mfe	The mean free energy of the hairpin
Loop length	The number of bases in the loop of the hairpin
Loop GC	The GC-content of the loop
GC	The GC-content of the small hairpin
Asymmetric bulges	The number of asymmetric bulges and mismatches regarding the stem
Symmetric bulges	The number of symmetric bulges and mismatches regarding the stem
Bulges	The number of bulges in the stem
Longest bulge	The number of non-pairing nucleotides of the longest bulge
Mismatches pre-microRNA	The number of single mismatches in the hairpin
Mismatches microRNAs	The number of single mismatches in the mature microRNA region of the hairpin
Stability	The smallest hairpin harbouring the read is extended 10 times for 10bp at both ends. The stability is given as the frequency the original structure is found in the elongated structures (see Figure 5 for an illustration)
Alternating stability	Reports whether a structure disappears in the stability calculation when extending the sequence, but reappears again (see Figure 5 for an illustration)
Triple-SVM features	All features that were proposed by Xue *et al*. (Xue et al. 2005)
Bindings	The number of bindings in the stem divided by the hairpin length

Table 2: Features calculated for the generation of the classifier.

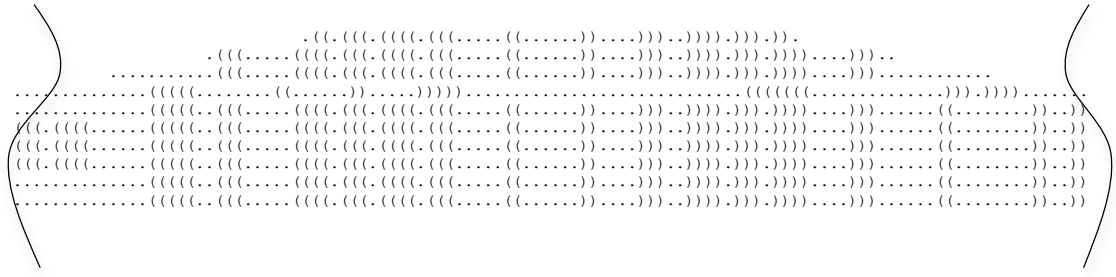


Figure 5: Aligned hairpin structures from the stability calculation displayed in dot-bracket notation.
The margins left and right are trimmed for proper illustration. As nine of ten structures contain exactly the structure of the smallest hairpin covering the read sequence, the value of the stability feature is 9/10. The 4th elongation leads to a varied structure. Because the following elongations, however, fold identical again, the alternating stability feature is allocated with “true”.

2.2.8. **Classifier**

To detect new MicroRNAs we set up a machine learning approach based on the standard WEKA (Witten and E 2005) implementation of the random forest learning scheme (Breiman L 2001), with the number of trees set to 100. To find the features with the highest prediction power, on which the learning scheme is subsequently trained, we performed feature selection by evaluating the features by means of their information gain. We then ranked the features according to their discrimination power. The top ten features used for building the final classifier are: stability, mfe, bindings, stem length, read count, longest bulge, mismatches microRNA, mismatches pre-microRNA, alternating stability and the Triple-SVM feature "A…".

2.2.9. **Pre-processing**

In order to check the reads for putative new microRNAs we perform a pre-processing of the data that contains the following steps:

1. Reads that overlap in the genome are clustered together.

2. Due to sequencing errors in reads, dicer products (mature, mature-star and loop) could be grouped together such that they appear as non-microRNA products (for example producing a long cluster that overlaps the loop of the precursor). To avoid such a situation, we walk along the cluster sequences and test whether the start of the current read overlaps less than 3 nucleotides with the end positions of previous reads. In that case the cluster is split at the current read start position. This way, clusters may contain exactly one non-dicer product or the mature microRNA or the mature-star microRNA, but not more than one theoretical product.

3. Clusters of more than 25bp length are discarded.

4. Since the microRNA can be located either on the 5' arm or the 3' arm of the hairpin, we extract the cluster sequence twice from the genomic location, with 60bp upstream and 10bp downstream flanking areas and vice versa. For both sequences the secondary structure is predicted via RNAfold, but only the energetically favourable is retained.

5. Non-hairpin structures are discarded.

6. Structures where the cluster sequence is not fully included or spans the loop and a part of the stem cannot be dicer products are consequently discarded.

7. Finally, since our analysis showed that virtually all known microRNAs show more than 14 bindings in the microRNA:microRNA-star duplex, we considered this as a mandatory requirement.

Having applied the pre-processing step to the three experimental data sets, we receive 6,967 candidate precursors for rat, 12,233 for worm and 43,905 for human.

2.2.10. Post-processing

After classification of the deep-sequencing data in form of the clusters created in the pre-processing step, clusters containing the mature and mature-star microRNA are merged such that one cluster represents one microRNA precursor.

2.2.11. Input file description

A usual next-generation sequencing experiment produces several hundred million base pairs of output corresponding to hundreds of megabyte or several gigabytes of data when stored into a file. That is by far too many data to send over the web to analyze it using a web server tool. However, some reads (tags) obtained in microRNA sequencing experiments can be found multiple times in the raw data output. The number of copies detected for a unique read is proportional to its expression level. Given this redundancy, the only information needed for the analysis of microRNAs are the sequences of the reads and the number of times each unique read was encountered in the experiment. This reduces the size of the input file drastically to a few megabytes, which is an acceptable size for a web server tool.

MiRanalyzer accepts two different input formats:

1. A tab separated file with the read sequences and its counts (number of times each read has been obtained in the experiment)

```
GAGGTAGTAGGTTGTA        49862
ACCCGTAGAACCGACC        15490
```

```
GGAGCATCTCTCGGTC          13762
...                       ...
```

Figure 6: Sample for a tab separated input file

2. A multi-FASTA file with the copy number of the unique reads (read count) as the description in the header (e.g. >ID 'count').

```
>ID 49862
GAGGTAGTAGGTTGTA
>ID 15490
ACCCGTAGAACCGACC
>ID 13762
GGAGCATCTCTCGGTC
```

Figure 7: Sample for a multi-FASTA input file

Along with this web-tool, we supply a Perl script that counts the reads of a Genome Analyzer (Illumina Inc.) deep-sequencing experiment, producing the tab separated input format needed by the miRanalyzer. The script allows averaging of several lines, filtering for low quality reads and a simple analysis of differential expression (log2 ratios between different lines). To counts the sequence reads the script needs to have the s_L_sequence.txt files (being L the lane on the flowcell) supplied as input. Two quality measures to filter out low quality reads have been implemented.

A more detailed description of the Perl script can be found on the tutorial page (http://web.bioinformatics.cicbiogune.es/microRNA/manual.html).

2.2.12. Input parameters

Apart from the file with the read sequences, several other input options are available:

Species and genome assembly

In the current version one of the following nine species genome assemblies can be chosen:

- *Homo sapiens* (hg18, NCBI 36.1)
- *Mus musculus* (mm8, NCBI 36)
- *Rattus norvegicus* (rn4, version 3.4)

- *Drosophila melanogaster* (dm3, BDGP Release 5)
- *Caenorhabditis elegans* (ce6, WUSTL School of Medicine GSC and Sanger Institute version WS190)
- *Canis familiaris* (canFam2, v2.0)
- *Danio rerio* (danRer5)
- *Gallus gallus* (galGal3)
- *Giardia lamblia* (gli1)

Number of mismatches

Next generation sequencing data are normally characterized by a higher sequencing error than the Sanger sequencing, but this error rate is balanced by a much higher redundancy. Therefore, the user should carefully choose the number of allowed mismatches (0, 1 or 2) to assign a sequence-read to a microRNA (default value is 1). Note that for the detection of new microRNAs and the overlapping with repetitive sequences, only perfect matches to the genome are considered.

Target gene table

The program makes available the putative target genes for each detected microRNA and direct links to their ontological analysis. Thus, the user must select a set of predicted target genes. We offer two different prediction methods:

- Predictions from miRBase hosted at the Sanger Institute
- Predictions from TargetScan (conserved family – conserved target)

Posterior probability threshold

The posterior probability of a random event or an uncertain proposition is the conditional probability that is assigned after the relevant evidence is taken into account. One of the available values should be selected (default is 0.9).

Do not consider adapter sequences

By default, the tool tries to detect if adapter sequences exist by searching for the perfect matches in subsequences of the input reads. If this option is selected, the tool will only look for the perfect matches of the input reads complete sequences.

Detect just new microRNAs

This option will skip the detection of known microRNAs.

Remove reads detected in mRNA database

Reads detected in a RefSeq transcript are removed from the input. Note that reads that map perfectly to a mRNA sequence may correspond to degradation products. If this option is not chosen, the tool automatically eliminates all reads that map to more than 5 mRNA sequences

Remove reads detected in RFam

This option removes reads that correspond to other known RNAs sequences as these might be easily confused with microRNAs (in the prediction of new microRNAs).

Removes reads detected in RepBase

This option removes all reads found in a RepBase sequence (transposons)

Predict only conserved microRNAs

All read clusters that do not overlap with a phylogenetically conserved element (PhastCons) are removed.

2.3. Results and Discussion

Our tool *miRanalyzer* follows three internal analysis steps (see Figure 8): (i) detection of known microRNAs, (ii) mapping against libraries of transcribed sequences (mRNA, ncRNA, etc.) and (iii) prediction of new microRNAs. After each of these three steps, the detected reads are removed from the input data following the options set by the user (see Table 7).

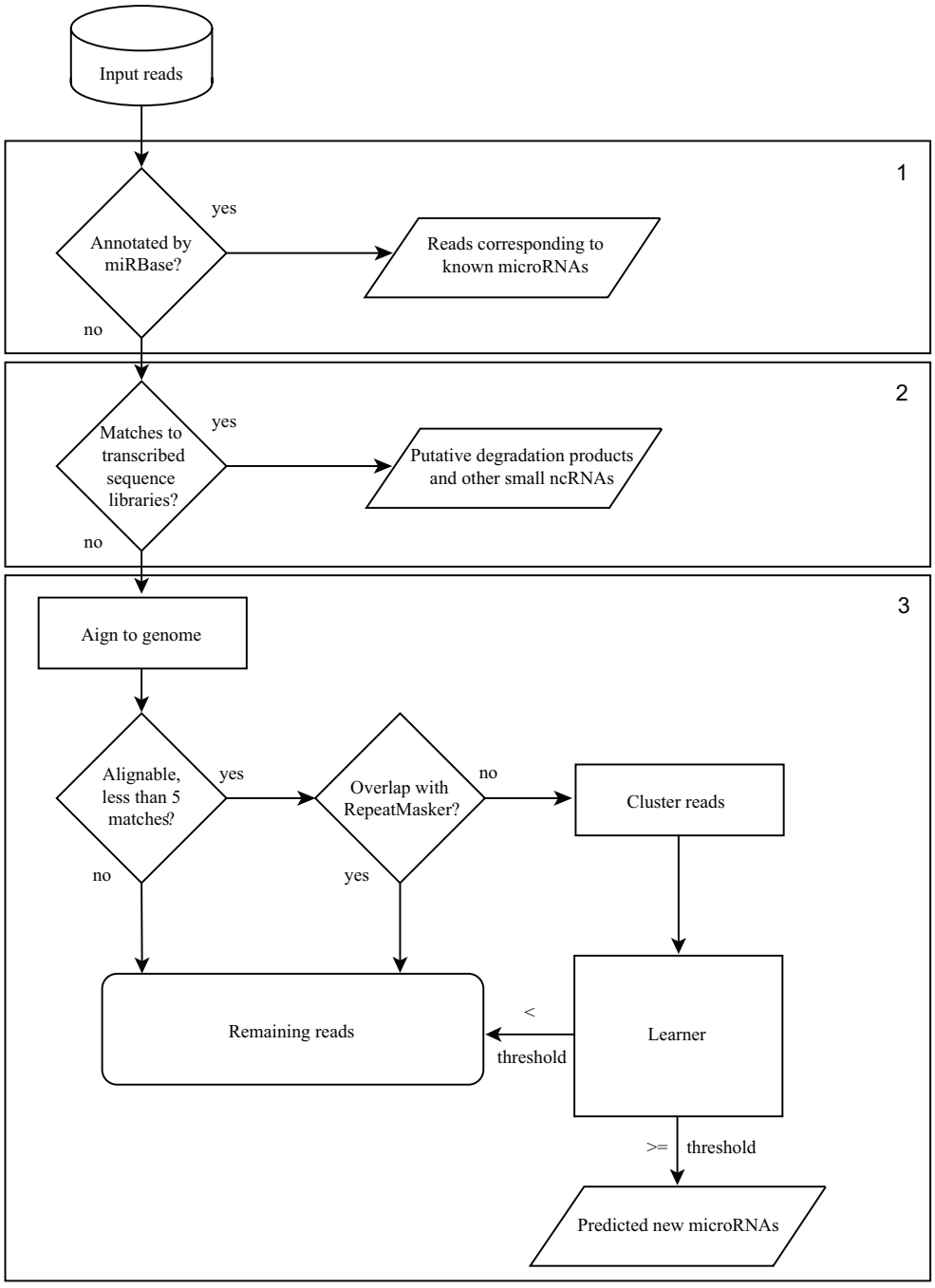


Figure 8: Workflow of miRanalyzer processes.
The analysis stream can be divided into three steps: detection of known microRNAs, detection of reads in other expressed sequences like mRNA, ncRNA etc. (to estimate the sample quality and remove reads which are prone to give false positives in the prediction of new microRNAs) and the prediction of new microRNAs. Finally, the program outputs all reads (remaining reads) that have not been assigned to any entity or have been filtered out from the beginning (reads with ambiguous characters in their sequence).

2.3.1. Detection of known microRNAs

In many of the microRNA experiments, the main purpose will be the detection of the expression levels of known microRNAs or frequently the differential expression of microRNAs between two samples. Therefore, as the first analysis step, miRanalyzer detects the reads that correspond to known microRNAs. To carry out the detection of known microRNAs, we used the miRBase repository (Griffiths-Jones 2006) which offers mature (the mature sequences of known microRNAs), mature-star (the sequence which pairs with the mature microRNA in the pre-microRNA secondary structure) and precursor microRNA sequences (sequence of the hairpin). For some of the microRNA precursors, it is unclear which of the two sequences (mature or mature-star) is biologically functional. In the case where both sequences are found to be expressed and the predominant product can be clearly detected, the minor product is labeled with a * (mature-star). Apart from the known mature-star sequences we generated a library with all other theoretically possible mature-star sequences. This also allows the detection of functional mature-star microRNAs whose expression has not been observed previously.

Many microRNA sequences, especially those belonging to the same microRNA family, exhibit a high degree of sequence similarity. Given that sometimes the read might be rather short (16 bp), non-unique matches might occur. A non-unique match exists if a read maps with the same quality (i.e. the same number of mismatches) at different positions or to more than one sequence in the library. Often, alignment programs such as ELAND (included in Illumina Inc. pipeline), do not report these ambiguous matches. However, this might result in a loss of important information. Therefore, miRanalyzer reports these ambiguous matches, stating all microRNAs where matches have been found. Note that the groups of microRNAs that have been detected by the same read will normally belong to the same family.

The exact order of mapping against known microRNAs is: mature, mature-star, unknown mature-star and precursors/hairpin. Both unique matches (a read matches just to one known microRNA) and ambiguous matches (a read matches several microRNAs with the same quality) are detected and removed from the input at each

step. The removal is important as otherwise the reads would be detected again in the precursor sequences (hairpins).

After known microRNAs detection, the corresponding target genes (those genes which are predicted to be regulated by the detected microRNA) are extracted (see *Material and Methods*) and pre-calculated ontological analyzes are made available. In the case of ambiguous matches where the set of target genes is made up of a combination of various microRNAs, a link to Annotation-Module (Hackenberg and Matthiesen 2008) is offered to launch the ontological analysis with the obtained gene list.

2.3.2. Mapping against transcribed sequences

After detecting reads that correspond to known microRNAs, miRanalyzer maps the remaining reads to databases of transcribed sequences as mRNA, non-coding RNA (RFam) and (retro)-transposons. Only perfect matches are considered in this analysis. These alignments are performed to achieve several aims:

First, the mapping against the transcriptome should not yield any matches except for exonic microRNAs (Kim and Nam 2006). Therefore, the number of matches can be viewed as a sample quality parameter, i.e. contamination of the RNA sample with degradation products and poly-A tails.

Second, the mapping to RFam (and other libraries of ncRNA) and RepBase has two goals: 1) it might be interesting to see which other known small ncRNAs are in the sample and 2) the removal of these reads will lower the number of false positives in the prediction of new microRNAs as those small ncRNA might be confused with actual microRNAs. The removal of those sequences is optional (see *Material and Methods*).

Third, we also used the genomic annotation of repeats and transposons derived by RepeatMasker (http://www.repeatmasker.org). After aligning all reads with the genome, miRanalyzer checks if the read coordinates overlap with those of the RepeatMasker annotation. In this way we can detect reads that overlap with

‘degraded’ transposons whose expression might indicate ‘domestication’ (acquired function).

2.3.3. Prediction of microRNAs

The last step of the pipeline is the detection of novel, previously unreported microRNAs. This constitutes a very important analysis step in the miRanalyzer tool as 1) a controversy exists over the real number of microRNAs (Berezikov et al. 2006b) and therefore it is important to mine sequencing experiments for new previously undetected microRNAs and 2) for many species there are none or just a few microRNAs known. Consequently, the analysis of sequencing experiments in these species relies almost completely on the prediction of new microRNAs.

As described in 1.1.3 microRNA precursors form a very characteristic hairpin structure. After the relocalization into the cytosol, the pre-microRNAs are cleaved at the loop end of the hybrid by an endonuclease called Dicer. The location of the cleavage (see Figure 9A) is relatively well determined (Filipowicz et al. 2008; Grishok et al. 2001; Hutvagner et al. 2001). As a consequence, the precursor can be divided into exactly three regions, the 5’ arm, the loop and the 3’ arm.

Reads stemming from pre-microRNAs should therefore correspond to one of these regions. They must not overlap the cleavage site, as this would only be the case if they were not a product of Dicer. Since all microRNAs are however processed by Dicer, non-compliant reads (see Figure 9B) indicate other cleavage or degradation events that are not associated with the microRNA biogenesis pathway.

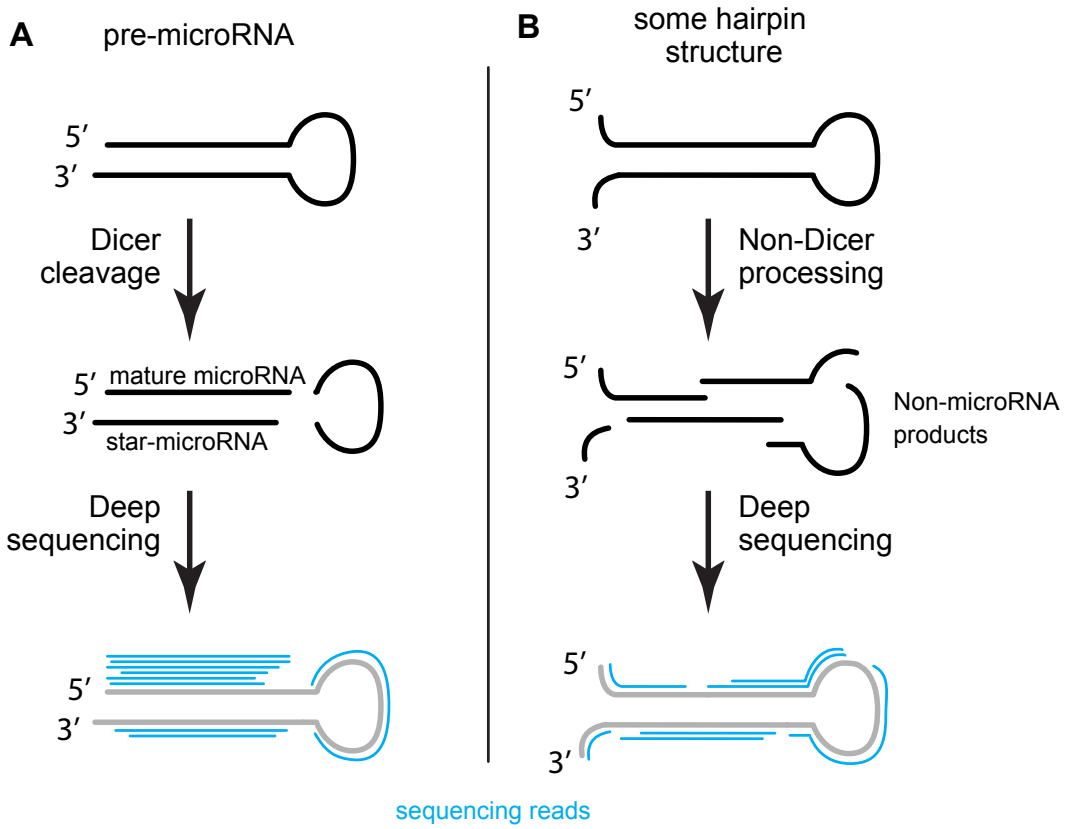


Figure 9: Traces of Dicer processing detectable in deep-sequencing experiments.

Generally, a very high number hairpin structures exist within a genome. When the tens and hundreds of thousands of reads that are produced in a deep-sequencing experiment are mapped to the genome, the overwhelming majority is associated with hairpin structures. Therefore testing the reads for the Dicer footprint serves as a powerful first-step filter in the identification of functional microRNAs (Friedlander et al. 2008).

Subsequently, we extracted several biologically motivated features (see Table 2 for a full list). One of these features, for instance, is the stability of the sub-structure that covers the mature microRNA sequence. The Idea behind this is that the structure of a functional microRNA should persist, independent of the surrounding context. We thus compute the secondary structure of the mature sequence with increasing flanking sizes and report the fraction of structures that maintain the exact substructure of the mature sequence (see *Material and Methods*; Figure 5).

As it is expected that each features is different in its predictive power, we employed a feature selection approach (see *Materials and Methods* for detail) to receive a ranked list with the best features on top. To train only on the most relevant features we restricted the classifier to the best ten. As learning scheme we used the random forest method (Breiman L 2001) .

2.3.4. Evaluation of prediction model

We used three different data sets from human (hsa), rat (rno) and worm (cel, see *Material and Methods*) for building the final prediction model. The results shown in Table 3 suggest that the classifier is highly sensitive and specific not only according to a standard ten-fold cross-validation, but also in a cross-species test on completely unseen test data. The results shown in the upper part of Table 3 depict the outcome when learning with one of the species (training set) and predicting the remaining ones (test data). For evaluation of prediction power in the same species, we applied a ten-fold cross validation approach.

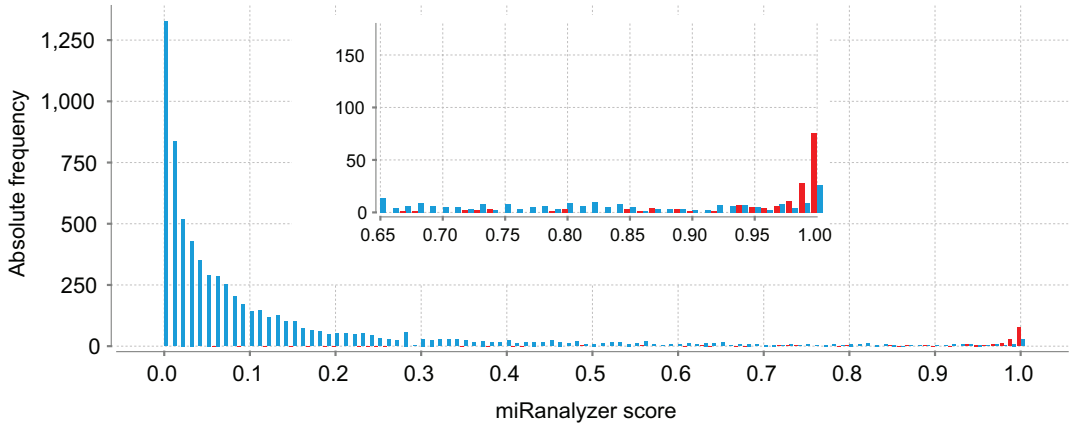


Figure 10: Histogram of miRanalyzer scores.
Known microRNAs are colored in red, all other data are colored in blue. The insert is a close-up for candidates with scores better than 0.65.

It can be seen that while the cross-validated results are high, the recall is moderate predicting on unseen data. We highlighted (yellow) the worst prediction values on

the different test sets, which are 0.66 (cel/rno), 0.48 (rno/cel) and 0.64 (rno/hsa). To check whether we can improve prediction power for those in particular, we merged two datasets and evaluated against the third set (values highlighted in green). It can be seen that the prediction improved significantly, especially for *C. elegans*. While trained solely on rat or human and evaluated on worm a recall of only 0.48 and 0.67, respectively, could be reached.

False negative rate (threshold: 0.9)

Training set \ Test set	rno	cel	hsa	rno-cel	rno-hsa	cel-hsa	rno-cel-hsa
rno	0.01^{CV}	0.008	0.009	0.004	0.008	0.001	0.005
cel	0.005	0.004^{CV}	0.003	0.002	0.01	0	0.005
hsa	0.005	0.004	0.01^{CV}	0.01	0.01	0.005	0.005
rno-cel	0.02	0.008	0.01	0.009^{CV}	0.01	0.007	0.01
rno-hsa	0.02	0.01	0.01	0.01	0.01^{CV}	0.01	0.01
cel-hsa	0.005	0.004	0.009	0.004	0.01	0.003^{CV}	0.01
rno-cel-hsa	0.01	0.004	0.003	0.01	0.01	0.009	0.007^{CV}

True positive rate (threshold: 0.9)

Training set \ Test set	rno	cel	hsa	rno-cel	rno-hsa	cel-hsa	rno-cel-hsa
rno	0.74^{CV}	0.48	0.64	0.66	0.73	0.57	0.65
cel	0.66	0.77^{CV}	0.69	0.80	0.68	0.79	0.76
hsa	0.74	0.67	0.77^{CV}	0.70	0.84	0.81	0.79
rno-cel	0.89	0.91	0.75	0.79^{CV}	0.80	0.82	0.84
rno-hsa	0.91	0.71	0.93	0.80	0.78^{CV}	0.84	0.86
cel-hsa	0.74	0.91	0.91	0.83	0.84	0.81^{CV}	0.86
rno-cel-hsa	0.89	0.91	0.90	0.91	0.91	0.92	0.79^{CV}

Table 3: The true positive rates (top part) and false positive rates (bottom part) for different classifiers at a posterior probability threshold of 0.9.
The superscripted "CV" denotes that this value was achieved in a standard ten-fold cross-validation approach. The highlighted values in yellow indicate the worst prediction performances when trained on a single data set. The worst prediction performances when trained on two merged data sets and evaluated on the third are highlighted in green.

The merged training set, however, achieves a recall of 0.71, suggesting synergetic effects when integrating instances from different species into the training set. To benefit most from this effect, we trained the final classifier on all three data sets. Thus we obtain an area under the curve (AUC) value of 97.9% with a true positive rate of 0.79 and a false positive rate of 0.007 for the fixed threshold at 0.9. To test for

robustness, we repeated the cross validation on ten different negative sets, which resulted in a mean AUC value, true positive rate and false positive rate of 97,9%, 0.79 and 0.0077 with the standard deviations of 0.001, 0.01 and 0.003, respectively.

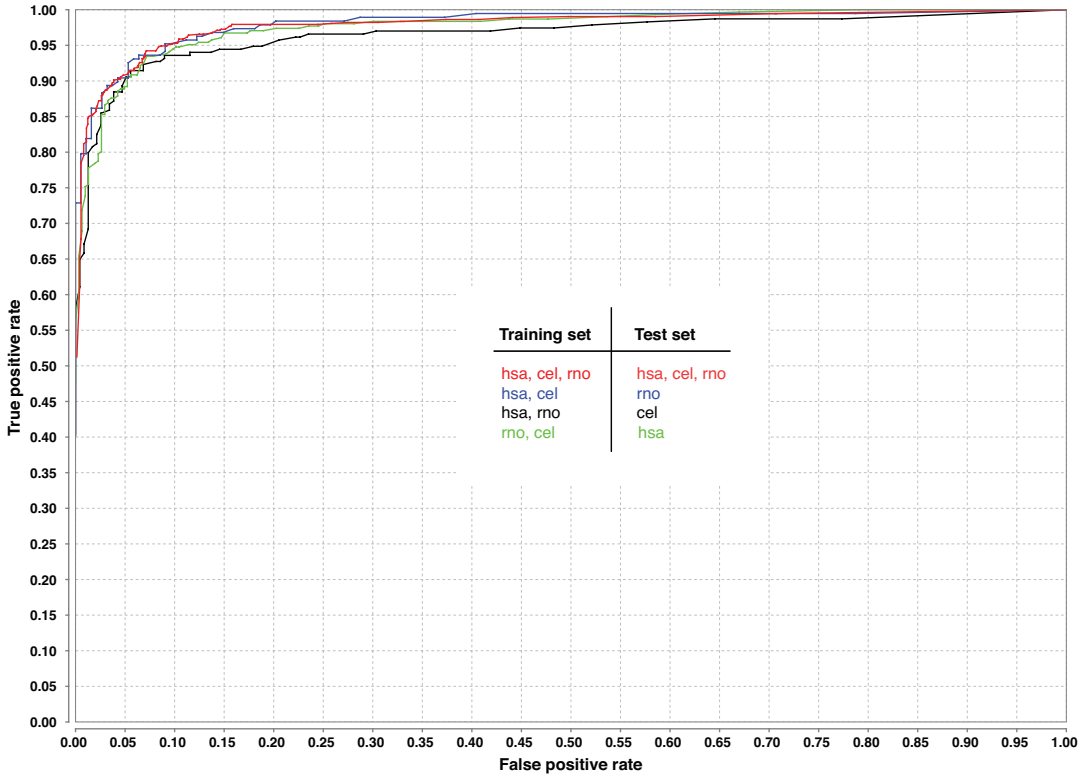


Figure 11: Receiver operating characteristics for the classifier trained on human, *C. elegans* and mouse (red), *C. elegans* and human (blue) and rat and C. elegans (green). The latter three classifiers were evaluated each with the one dataset that was not used for training, while the final classifier (red) was evaluated in a standard ten-fold cross-validation.

Figure 10 shows a cross-species evaluation of miRanalyzer trained on human and *C. elegans* and evaluated on rat. Obviously, most of the data have very low scores (the posterior probability assigned by the classification model to each instance) assigned. We build a close-up for the range between 0.65 and 1 to better visualize the high scoring predictions. It can be seen that the known rat microRNAs are strongly accumulated towards scores of 1, demonstrating the high predictive power of our

approach and the good ability to generalize. Note that the classifier has never seen data from rat before.

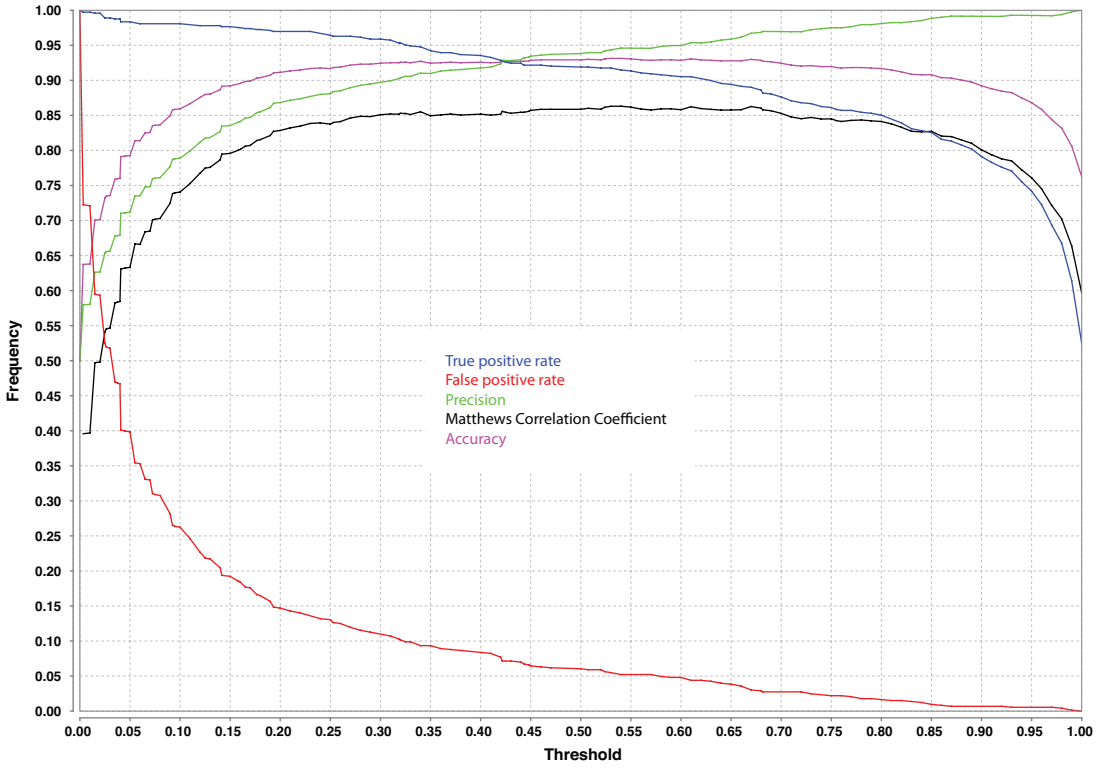


Figure 12: Various performance measures for the final classifier trained on human, *C. elegans* and rat and evaluated in a standard ten-fold cross-validation approach.

2.4. Application

As a working example we used data derived from an experiment carried out in the laboratory of the functional genomic unit at CIC bioGune with rat hepatocytes following standard protocols for smallRNA sample preparation and deep-sequencing (http://www.illumina.com/). Figure 13 shows the summary output page of miRanalyzer run on these data. The page is made up of five boxes that reveal the intrinsic workflow of miRanalyzer:

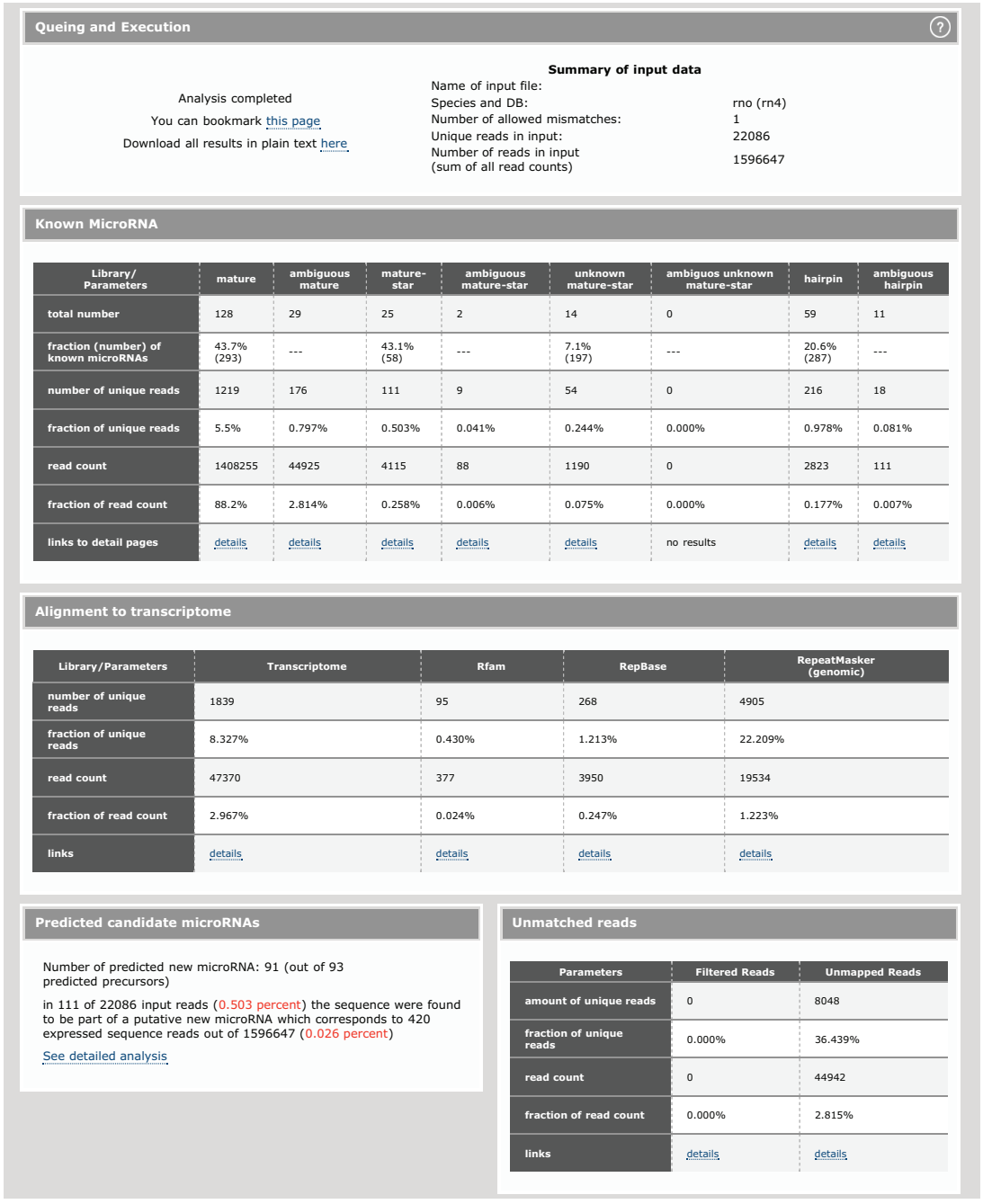
The first box shows the current state of the process (executing, pending, etc.) on the left side and depicts a short summary of the process (input data and options) on the right side.

The second box shows the summary of the analysis of known microRNAs. Each column corresponds to the mapping against a different set of sequences (mature, mature-star, etc.). The last row provides a link to detailed output for each of the columns. For example, the analysis of unknown mature-star sequences shows that miR-423-star is moderately expressed (744 copies) while the sequence that is annotated in miRBase (mature miR-423) has less than 10 copies.

The third box summarizes the matching of reads to several sets of transcribed sequences. For example the fraction of reads mapped to the transcriptome may give a good estimate on the sample quality. It can be seen that around 8.3% of all reads in this sample originate from mRNA but this corresponds just to 3% of transcription amount (number of mRNA reads/total number of reads).

The fourth box shows the summary of the detection of new microRNAs. In addition, a link is given for further information on each read cluster that has been predicted to be a novel microRNA. A link is also provided to a detailed output page with information on the chromosomal coordinates, the long hairpin structure and a verification if the reads have been detected before in the experiment (for example if matched against RepBase, etc.).

Finally, the last box gives a summary of the filtered and unmapped reads.



Queing and Execution

Analysis completed

You can bookmark this page

Download all results in plain text here

Summary of input data

Name of input file:	
Species and DB:	rno (rn4)
Number of allowed mismatches:	1
Unique reads in input:	22086
Number of reads in input (sum of all read counts)	1596647

Known MicroRNA

Library/ Parameters	mature	ambiguous mature	mature-star	ambiguous mature-star	unknown mature-star	ambiguos unknown mature-star	hairpin	ambiguous hairpin
total number	128	29	25	2	14	0	59	11
fraction (number) of known microRNAs	43.7% (293)	---	43.1% (58)	---	7.1% (197)	---	20.6% (287)	---
number of unique reads	1219	176	111	9	54	0	216	18
fraction of unique reads	5.5%	0.797%	0.503%	0.041%	0.244%	0.000%	0.978%	0.081%
read count	1408255	44925	4115	88	1190	0	2823	111
fraction of read count	88.2%	2.814%	0.258%	0.006%	0.075%	0.000%	0.177%	0.007%
links to detail pages	details	details	details	details	details	no results	details	details

Alignment to transcriptome

Library/Parameters	Transcriptome	Rfam	RepBase	RepeatMasker (genomic)
number of unique reads	1839	95	268	4905
fraction of unique reads	8.327%	0.430%	1.213%	22.209%
read count	47370	377	3950	19534
fraction of read count	2.967%	0.024%	0.247%	1.223%
links	details	details	details	details

Predicted candidate microRNAs

Number of predicted new microRNA: 91 (out of 93 predicted precursors)

in 111 of 22086 input reads (0.503 percent) the sequence were found to be part of a putative new microRNA which corresponds to 420 expressed sequence reads out of 1596647 (0.026 percent)

See detailed analysis

Unmatched reads

Parameters	Filtered Reads	Unmapped Reads
amount of unique reads	0	8048
fraction of unique reads	0.000%	36.439%
read count	0	44942
fraction of read count	0.000%	2.815%
links	details	details

Figure 13: The summary page of miRanalyzer.
Five boxes are shown which correspond to summary & state of the process, analysis of known microRNA, matches against transcribed sequences, detection of new microRNAs and summary of unmatched sequences.

2.5. Conclusion

MiRanalyzer is a powerful and easy to use web server tool for the integral analysis of next generation sequencing data of small RNA molecules. It allows both the detection of known microRNAs and the prediction of new microRNAs. For the prediction of new microRNAs a new sensitive machine learning based approach was developed that reaches an AUC of 97.9% in our tests.

Furthermore, the tool detects matches of the reads against other libraries of transcribed sequences such as mRNA, RFam (RNA) and RepBase (Transposons). Currently, the tool works for nine species, but can easily be extended to further species in future.

Chapter 3

Examination of microRNAs target sequences

As pointed out earlier, a perfect match to the microRNA seed region is considered as a critical factor in the detection of target sites. It was further demonstrated, that the false positive rate in the prediction of target sites could be significantly reduced if the seed match is also conserved across several species. Therefore a simple but effective strategy of detecting target site with a high likelihood of biological impact is to search for conserved reverse complimentary seed sequences in the 3'UTRome (Bartel 2009).

This approach is solely based on nucleotide sequence and cross-species information and does not include any secondary information, like hybrid structure, stability or local influences to the target site. Consequently the detection process is highly dependent on the nucleotide composition of the microRNA seed. We therefore study the GC content distributions of microRNAs and 3'UTRs in this chapter. In addition, we analyze the structure of all known human genes (RefSeq) – with intronic sequences removed – regarding their GC content and conservation. Finally we built microRNA target site background models. We found that highly conserved microRNAs target significantly more transcripts than expected.

With the availability of genome-wide SNP data, extensive studies of cis-regulatory sites became possible. Recently, it was proposed, that with respect to these data,

there is negative selection acting on predicted microRNA binding sites (Chen and Rajewsky 2006). These authors detected a significant gap in SNP density between the seed region and the rest of the target site. Although their random control showed a gap at the same position, too, the associated p-value for the background gap was an order of magnitude higher. Hence the authors proposed that the conserved seed matches of ultra conserved microRNAs are under negative selection pressure.

Driven by our findings regarding the highly conserved microRNAs, we reassessed the SNP density at predicted microRNA target sites. Our results show that the SNP density is not reduced in the seed match of highly conserved microRNAs compared to conserved random 8-mer sequences. We further demonstrate that the position of the gap in SNP density is dependent on the seed length used for predicting target sites.

3.1. Materials and Methods

3.1.1. Alignment data and conservation

We downloaded the 17-way multiZ alignments and the human RefSeq transcript annotations form the University of California, Santa Cruz (USCS) browser and assembled alignments for 5000nt upstream, 5'UTR, CDS, 3'UTR and 5000nt downstream into continuous multiple alignments per RefSeq transcript annotation using the Galaxy Webservice (Giardine et al. 2005).

To calculate the conservation profile, we divided the alignments into 100 bins of equally long substrings. Each nucleotide position was then tested for perfect conservation in human, chimp, rat, mouse and chicken. Subsequently the fraction of conserved nucleotides of a bin was returned as degree of conservation.

3.1.2. MicroRNA data

All mature microRNA sequences were downloaded from the microRNA registry version 12.0 (Griffiths-Jones 2004). Ultra conserved microRNAs were obtained from (Krek et al. 2005). Background seed sequences were generated by recursively adding one nucleotide at a time such that all combinatorial possible 6-mer and 8-mer

sequences are covered. The *shifted set* was created by extracting 8-mers from the 56 ultra conserved microRNA sequences, starting at positions 6, 8 and 10.

3.1.3. Conserved target site prediction

For the prediction of microRNA target sites, three approaches were used in this work. They vary in the seed definition and way conservation is applied. In the first approach we identified target sites by searching for 6-mer seed matches that are perfectly conserved in the species human, chimpanzee, mouse, rat, and dog. In the second approach we simply changed from 6-mer seeds to 8-mer seeds. The last approach, similar to the core PicTar algorithm, uses 7-mer seed matches and a slightly different way of conservation usage. First, 6-mer sites with perfect Watson-Crick complementarity to the microRNA bases 2-7, counted from the microRNA 5' end are searched for. These 6-mers have to be perfectly conserved in human, chimpanzee, mouse rat and dog. Moreover either a perfect match to microRNA position 1 or 8 is required for each species.

3.1.4. SNP data

All human SNP data used in this analysis were downloaded from the dbSNP (build 128) track of the UCSC genome browser. From this set we discarded all SNPs originating from insertion and deletion events as well as all SNPs with more than two alleles and those with only one allele (monomorphic). SNPs with more than one assigned loci (release hg18) were excluded as well. Genotype data were retrieved from the International HapMap Project (build 23a, NCBI build 36) and from Perlegen Science (version 1, NCBI build 34). We used liftOver to map Perlegen data to hg18. In total we retrieved 9,657,322 SNPs from dbSNP, from which 4,086,708 SNPs are genotyped by HapMap and 1,545,504 SNPs are genotyped by Perlegen. HapMap genotypes correspond to 90 CEPH individuals with ancestry from northern and western Europe (European population), 90 Yoruban individuals (African population), 45 Han Chinese in Beijing (Chinese population) and 44 Japanese in Tokyo (Japanese population). Perlegen genotype data were obtained from 24 European Americans (European population), 23 African Americans (African population) and 24 Han Chinese (Chinese population). Taken together, we retrieved

1,240,720 SNP data, genotyped from both, HapMap and Perlegen. From these we were able to map 1,700 to the 5'UTR, 16,700 to CDS and ~13,500 to 3'UTR.

3.2. Results and Discussion

3.2.1. MicroRNAs and transcripts in the light of GC content

We first investigated the human 5' untranslated regions (5'UTRs), coding sequence (CDS) and 3' untranslated regions (3'UTRs) with respect to their GC content. Therefore we calculated the GC content of all sequences, with the exception of very short (<100nt) ones. The resulting histogram is shown in Figure 14A. As can easily be observed, there is a significant difference between the three groups. While the 5'UTRs are Gaussian like distributed, but shifted towards GC richness, the CDS are almost uniformly distributed in the range between 35% and 65% GC content. The 3'UTRs are clearly the GC poorest sequences in human, with a global maximum at 33%. The distribution however is bimodal and a local maximum appears at approximately 50%.

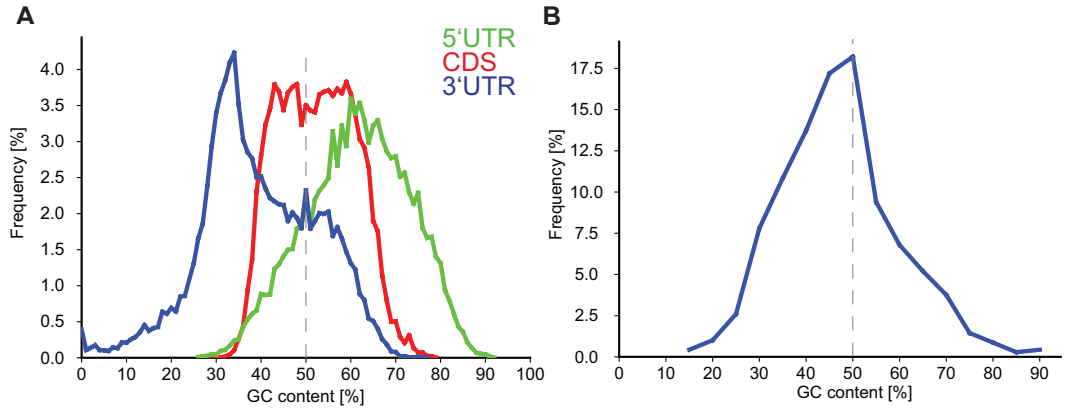


Figure 14: GC content distributions of A) human microRNAs and B) human 3'UTRs.

In a second step, we analyzed all known human microRNAs according to their GC content (Figure 14B). We see 1) the peak at 50% GC content and 2) a bias towards the region 30%-50%. This is particularly interesting, as we have seen that the majority of 3'UTRs fall into that range. As a pure matter of chance, microRNA target sites will therefore occur in the 3'UTRs much more often than in the CDS or the 5'UTRs. A simple search for non-conserved microRNA seed matches (Table 4)

reveals for the 3'UTRs a 20% (16.4 matches/kb versus 13.7 matches/kb) density enrichment compared to 5'UTRs.

	Number of perfect seed matches	Total number of nucleotides	Seed match density (matches/kb)
5'UTR	126,731	9,268,795	13.7
CDS	728,013	49,094,793	14.8
3'UTR	468,403	28,453,035	16.4

Table 4: Overview of the number of seed matches in the 5'UTR, CDS and 3'UTR.

As the GC content distributions of the three groups 5'UTR, CDS and 3'UTR are so different, we focused on the position specific analysis of those (Figure 15A). In addition, we considered 5000 nucleotides up- and downstream of the start and end position of the genes. Interestingly, one sees that the GC content is extreme at the transcription start (GC rich) and stop site (GC poor), a phenomenon that was earlier reported as *genomic punctuation* (Zhang et al. 2004). Between these two sites, a continuous decay in GC content can be observed. While the decay is linear in the 5'UTR, it is almost stopped in the CDS. Here the GC content remains at around 52%. In the 3'UTR, the decay reinitiates, while it is extreme towards the beginning and ending. The up- and downstream flanking regions, illustrating the general genomic level, show a constant GC content of about 45%.

In addition, we have plotted the conservation of these regions in a position specific manner, too. As expected (Figure 15B) the highest conservation is observed in the CDS with a mean of around 70% followed by the 5' UTR and 3' UTR with a conservation level of approximately 35%. It is interestingly, however, that the 3' UTRs show strong conservation peaks at the boundaries, forming a U-shape. The 5'UTRs also show a strong peak at the end. This is however narrow and likely attributed to the translation start site which is directly following. For the 3'UTR U-shape we do not have such a conventional explanation.

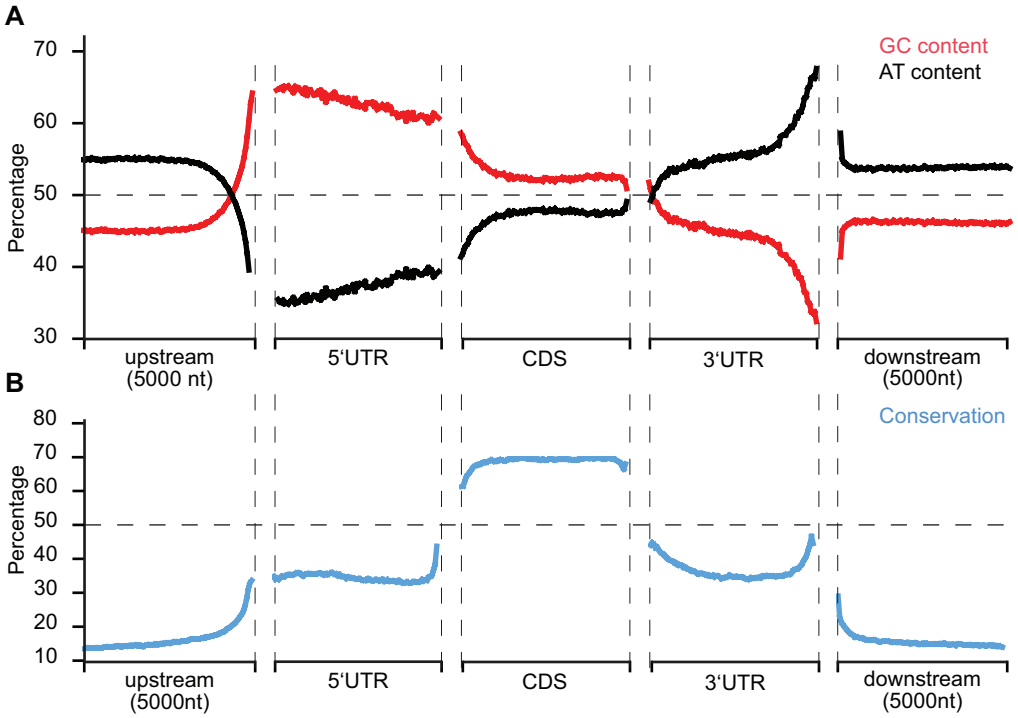


Figure 15: Structure of known human genes with intronic sequences removed in reference to GC content and conservation.

However, as pointed out in 1.1.5, it could be demonstrated, that microRNA target sites that are positioned at the margins of the 3'UTRs show a higher functionality compared to those that are located in the middle (Grimson et al. 2007). Therefore microRNAs might be the driving force that shaped the GC content distribution of 3'UTRs.

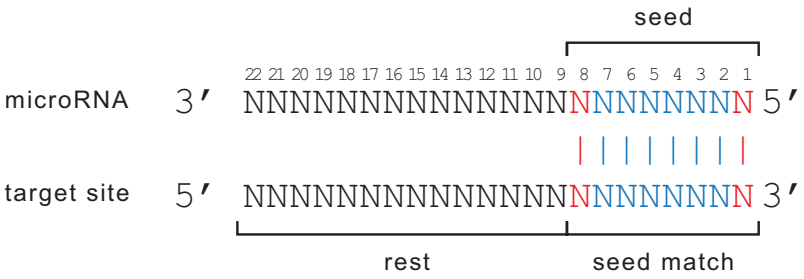


Figure 16: Compartmentation of microRNA target site

3.2.2. The more target sites a microRNA has the AT richer its seed region

As we say in the last chapter, that 3'UTRs are higher conserved at their margins and that target sites in these regions have been demonstrated to show a higher functionality, we subsequently focused on those highly conserved target sites.

Therefore we considered both, seeds of human microRNAs as well as any 8-mer sequences. We clustered all target sites, i.e. a perfectly conserved 6-mer (nucleotide 2-7) seed match and an additionally conserved match to either microRNA nucleotide 1 or 8, of a microRNA and an 8-mer sequence together such that we receive a simple relation table with the 8-mer sequences in the first column and the number of conserved sites in the 3'UTR in the second. Subsequently we clustered entries together by employing a binning on the number of target sites. For each bin, we then calculated the mean GC content both of the seed match and of the target site rest (see Figure 16 for an explanation).

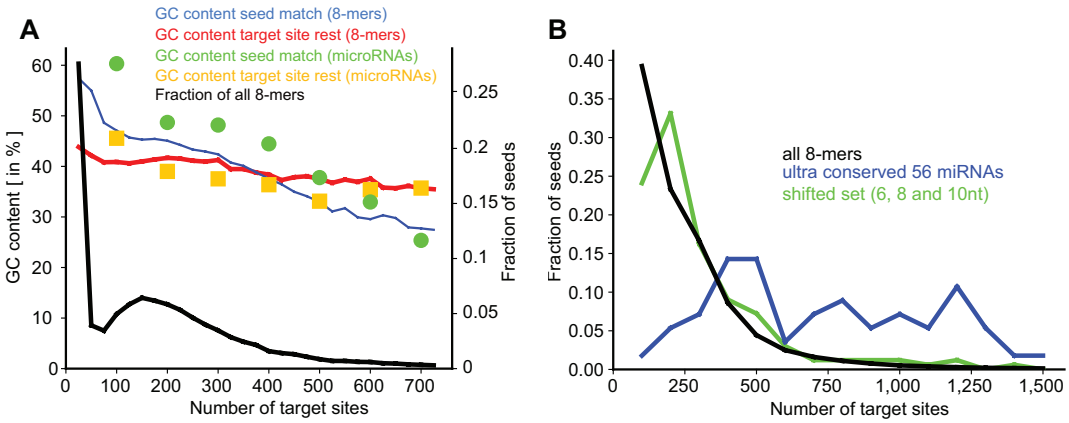


Figure 17: GC content of the seed and the target site rest relative to the number of conserved target sites detected in the 3'UTRome.

As can be seen in Figure 17A, the GC content of the target site rest stays almost constant, independent on the number of targeted sites. On the contrary, the GC content of the seed matches constantly declines with the number of target sites observed. In consequence, this leads to an asymmetry within target sites for 8-mers that match either extremely often or rare. Target sites of 8-mers that match in the range between 250 and 500 are approximately balanced. For microRNAs the intersection is slightly shifted towards higher numbers (~ 600) of target sites. Additionally we observe that the fraction of 8-mer seeds is bimodal distributed with a significant fraction (> 40%) of all seeds having equal or less than 70 target sites. The distribution for microRNA seeds follows the same general trend (data not shown).

We then questioned whether this is also the case when only ancient, i.e. highly conserved microRNAs are considered. We therefore used the list of 56 microRNAs, that was initially compiled by (Krek et al. 2005) and we additionally created a non-microRNA background (shifted set) as used by (Chen and Rajewsky 2006). Since the data are significantly sparser than the analysis of all 8-mers, we increased the bin size to 100. Therefore the bimodal effect vanishes (Figure 17B). The remarkable point, however, is that the ultra conserved set of 56 microRNAs follows a distinct distribution. All these microRNAs seem to be biased towards high amounts of conserved target sites, a clear break with all 8-mers, all microRNAs and the shifted set.

3.2.3. Negative selection on predicted conserved microRNA target sites is equal to other conserved sites

According to recent studies, a single mutation in the seed match of a target site can be sufficient to destroy the regulatory function of microRNAs (Brennecke et al. 2005). Further it was shown that single nucleotide polymorphisms (SNPs) have significant impact on the functionality of microRNA regulation (Abelson et al. 2005; Clop et al. 2006). As we have just found that ultra conserved microRNAs tend to target much more transcripts than the 8-mer sequences, we turned towards the question, whether this has influence on the SNP density and therefore might affect formerly drawn conclusions on negative selection acting on microRNA target sites.

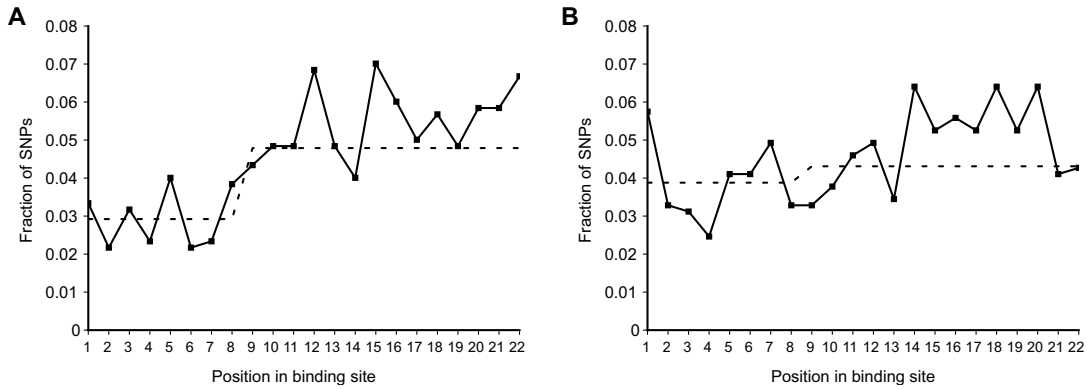


Figure 18: Position specific fraction of SNPs in conserved microRNA target sites.

The dashed lines indicate the average SNP fraction in the seed region (bases 1-8) and the rest of the target site (bases 9-22). In (A) the SNP fractions of predicted target sites according to the ultra conserved microRNAs are shown. Conserved random controls (shifted set) are displayed in (B).

We therefore extracted all human genotyped SNP data (Perlegen and HapMap) from dbSNP (Sherry et al. 1999; Sherry et al. 2001) and mapped them on the 3'UTRs to get a crosslink between the target sites and the SNPs.

In a first analysis, we computed the fraction of SNPs according to the position within the target site, following the protocol of (Chen and Rajewsky 2006). In total we retrieved 581 SNPs within 25.228 target sites. Figure 18A shows the distribution for target sites corresponding to the ultra conserved microRNAs, while B shows the distribution for the random controls (shifted set as before). In compliance with Chen et al. we found for the target site of the ultra conserved microRNAs the difference between the seed match and the rest to be highly significant ($\chi^2 - test, P = 2.1 \times 10^{-6}$). In the control set the difference was still significant ($\chi^2 - test, P = 1.06 \times 10^{-3}$), though by far not as significant as for the microRNAs.

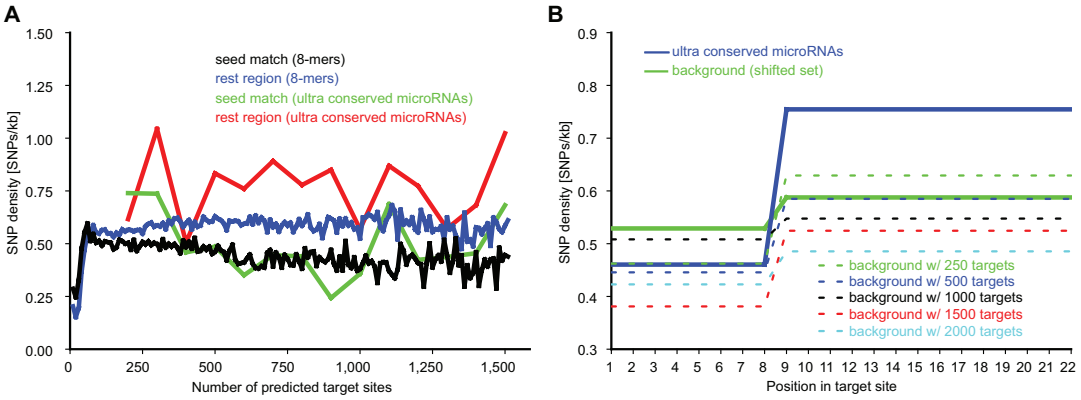


Figure 19: SNP density of predicted target sites
A) SNP density of the seed match and rest region binned according to the number of conserved predicted target sites in the 3'UTRome. The number of target sites is discretisized into equally wide bins of size 10 for all 8-mers and of size 200 for those belonging to the ultra conserved microRNAs. The difference between the SNP density of the seed match and the rest region of all 8-mers is constantly increasing as the number of target sites per seed raises. An exception to that are seeds with 70 or less target sites. At 80 target sites per seed, the SNP density is approximately equal in both (seed match and rest) distributions. In terms of SNP density, the rest region of all 8-mers (blue) seems to be independent of the number of seed matches, while the SNP density in the seed region (black) declines. Superimposing the ultra conserved microRNAs, the seed region (green) follows the seed region of all 8-mers (black), while the rest region (red) is above the rest of all 8-mers. B) Position specific SNP density in conserved seed matches, averaged in the seed match (bases 1-8) and the rest of the site (bases 9-22).

Subsequently we investigated whether the observed gap between the seed match and the target site rest (see Figure 18) could be due to the fact that the ultra conserved microRNAs show significant higher amounts of conserved target sites per microRNA. As before, we clustered the microRNAs and 8-mer sequences by the number of conserved target sites they exhibit and binned them accordingly. For each bin the mean SNP density was then calculated (Figure 19A). Apparently, the SNP density of the 8-mer target site rest seems uncorrelated to the number of targeted sites, while the seed match declines with the number of conserved target sites per microRNA. We also see that the seed match of ultra conserved microRNAs follows the general trend of the 8-mer counterpart. The rest region of the target sites from the ultra conserved microRNAs, however, shows much higher SNP densities than those from the 8-mer sequences. This is intriguing, as it 1) explains the observed step in the SNP fraction and 2) indicates that rest region shows a higher flexibility, i.e. fewer evolutionary constrains than expected.

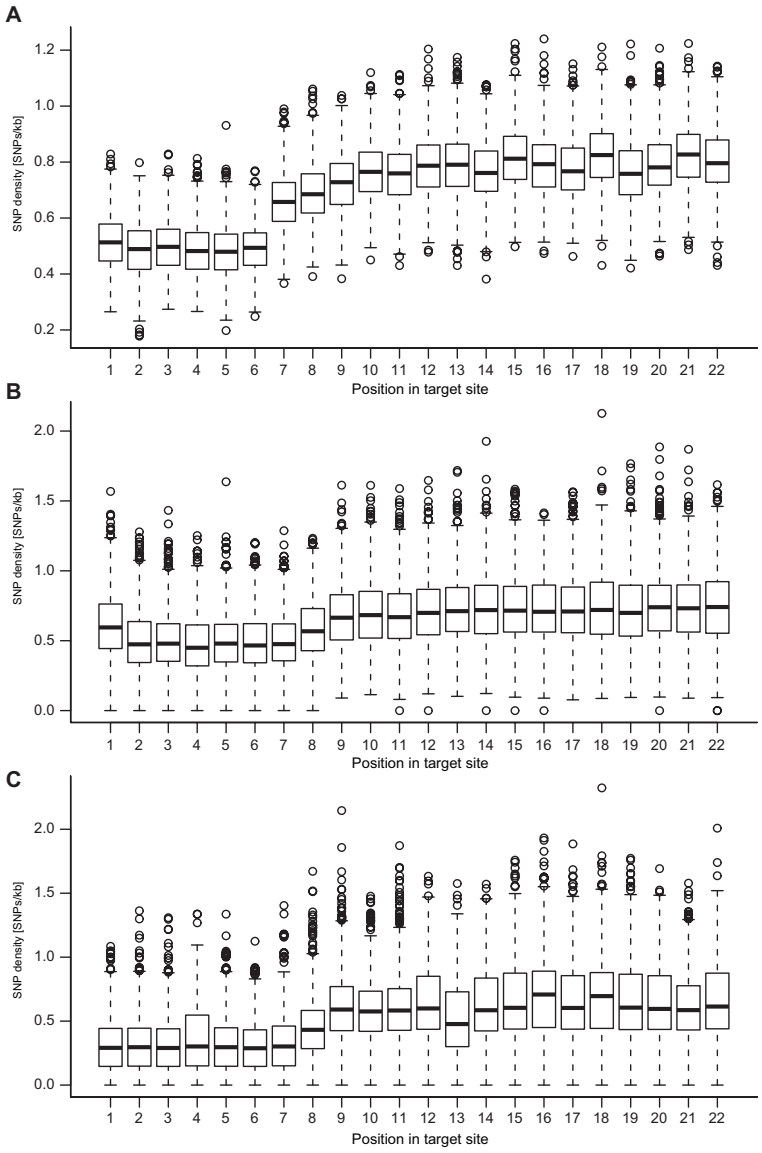


Figure 20: Position specific a SNP density in conserved seed matches for A) 6mer, B) core PicTar and C) 8mer seeds.
Indicated by the mean marks of the box and whisker elements, the gap between seed and rest region is for 6mer seeds at position 6, for core PicTar seeds (7-mer; nucleotide 1-7 or 2-8) at position 8 and for 8mer seeds at position 8. Further the mean SNP density is decreasing for the whole target site from 6mer seeds to 8mer seeds.

To prove it we repeated the plot of Figure 18 with the modification of using actual SNP densities instead of the fraction. Additionally we created several background distributions by drawing target sites at various bins, but only from the same bin each (Figure 19B). As can be easily seen, the SNP density of the seed matches of the ultra conserved microRNAs is never below background, but the SNP density of the rest

region is vastly above the background sets, irrespective of the chosen amount of targeted sites of the background.

We also modified the definition of the seed match requirement. By applying not only the 8-mer, but also the 7-mer and 6-mer seed definition, we repeated the last analysis for all possible x-mer sequences. As result, we received Figure 20A-C. One sees, that on a 6-mer background the step is located at position 6. With a 7-mer seed match the step is at nucleotide 8, while the first nucleotide has a slightly higher SNPs density, too. In the last scenario, when 8-mer matches are requested, also the first nucleotide is at the same low density as the rest of the seed match.

3.3. Conclusion

We have demonstrated that the GC content of the microRNAs and the 3'UTRs is of major interest for the field of computational prediction of target sites. Generally, both ends of the 3'UTR are higher conserved than the middle. The 3'UTRs show a pronounced gradient in GC-content, being GC-rich at its 5'-end and AT-rich at its 3'-end. Though, generally, the more conserved target sites a microRNA exhibits, the AT richer is the seed.

The GC content of the rest of the target site, however, is not correlated to the number of target sites of the respective microRNA. Interestingly these findings hold true not only for microRNAs, but for all possible 8-mer pseudo microRNAs. Therefore this is complicating the search for real microRNA target sites.

We have further analyzed SNPs at target sites and were able to reproduce the significant gap in SNP density between the seed match and the rest. In contrast to the findings of Chen and Rajewsky (2006), we have support that the observed step is not due to negative selection on the seed match of ultra conserved microRNAs, but due to a higher SNP density in the rest.

In addition, we have shown that the position of the step in SNP density is dependent on the length of the used seed definition. We therefore hypothesize that the observed step may be a GC dependent phenomenon introduced by the usage of conservation for identifying target sites. We currently have no explanation for the higher SNP

density in the rest of the target sites of ultra conserved microRNAs compared to background. The sample size of mappable SNPs to predicted target sites is larger than which was used by Chen and Rajewsky (2006), though it is still rather small. With the availability of more accurate SNP data more robust analyzes will be possible.

Chapter 4

TargetSpy: Analysis and prediction of microRNA target sites

Virtually all currently available microRNA target site prediction algorithms require the presence of a (conserved) seed match to the 5' end of the microRNA. Recently however, it has been shown that this requirement might be too stringent, leading to a substantial number of missed target sites.

We developed *TargetSpy*, a novel computational approach for predicting target sites regardless of the presence of a seed match. It is based on machine learning and automatic feature selection using a wide spectrum of features covering current biological knowledge. Our model does not rely on evolutionary conservation, which allows for the detection of species-specific interactions and makes *TargetSpy* suitable for analyzing unconserved genomic sequences.

In order to allow for an unbiased comparison of *TargetSpy* to other methods, we classified all algorithms into three groups: I) no seed match requirement, II) seed match requirement, and III) conserved seed match requirement. Appropriate post filtering generates *TargetSpy* predictions for classes II and III. On a human dataset revealing fold-change in protein production for five selected microRNAs our method shows superior performance in all classes. In *Drosophila melanogaster* not only our class II and III predictions are on par with other algorithms, but notably the class I (no-seed) predictions are just marginally less accurate. We estimate that *TargetSpy*

predicts between 26 and 112 functional target sites without a seed match per microRNA that are missed by all other currently available algorithms.

Only a few algorithms can predict target sites without demanding a seed match and *TargetSpy* demonstrates a substantial improvement in prediction accuracy in that class. Furthermore, when conservation and the presence of a seed match are required, the performance is comparable with state-of-the-art algorithms. *TargetSpy* was trained on mouse and performs well in human and drosophila, suggesting that it may be applicable to a broad range of species. Moreover, we have demonstrated that the application of machine learning techniques in combination with upcoming deep sequencing data results in a powerful microRNA target site prediction tool (www.targetspy.org).

4.1. Background

The discovery of microRNAs in 1993 (Lee et al. 1993) introduced a totally new dimension in our understanding of how gene expression is regulated. Animal and plant genomes contain hundreds of microRNA genes (Bartel 2009; Bentwich et al. 2005) that control fundamental cellular processes and are implicated in severe diseases. Incorporated into a protein complex named RISC, microRNAs perform posttranscriptional gene regulation either through perfect binding to a cis-regulatory target site in the 3'UTR that is subsequently cleaved, leading to mRNA degradation, or by imprecise binding preferably of the microRNA 5' end to a target site, leading to possibly reversible repression of protein production. While posttranscriptional cleavage is prevalent in plants, translational repression is the predominant type of regulation in animals. Our current knowledge about the function of specific microRNAs, their targeted messenger RNAs, and the exact location of binding sites is limited.

Experimental detection of microRNA target sites is a costly and time-consuming process. While recent estimates suggest that more than 50% of human protein-coding genes may be regulated by microRNAs and that each microRNA may bind to 300-400 target genes, the latest release of the TarBase database contains information on only 995 human *in vivo* microRNA-gene interactions involving 103 distinct

microRNAs and 825 distinct genes, a far cry from the actual extent of microRNA targeting (Bartel 2009; Sethupathy et al. 2006).

Computational prediction of microRNA/gene interactions is a valuable tool for guiding wet-lab experiments, and it remains the only option for systematic genome-wide reconstruction of the complex combinatorial picture of microRNA-mediated target binding. It is also a challenging task because of the daunting difficulty of distinguishing true microRNA-mRNA hybrids against the noisy background of millions of possible microRNA-gene combinations and, more generally, because the basic mechanisms of microRNA target recognition remain largely unknown.

Over the recent years many target prediction algorithms have been developed based on different principles (see Bartel 2009 for a review). However, the two recurring parameters used by the available methods are i) the existence of a seed match (continuous base pairing between a 3'UTR and the first 6-8 bases of a microRNA 5' end), and ii) evolutionary conservation of the target site across multiple species. Utilization of these powerful constraints in prediction algorithms leads to more reliable detection of those functional duplexes that contain them, but at the same time limits our ability to identify biologically relevant microRNA target sites that do not fulfill these requirements.

By definition, organism-specific or simply poorly conserved sites cannot be predicted at all if the conservation filter is applied. It has also been suggested that the seed match requirement may be too stringent, and that at least a second "type" of target sites - the so called 3' compensatory target sites – exists that cannot be detected by the seed match based methods. On the other hand, many potential microRNA-target interactions that do involve conserved seed regions may be non-functional in a physiological context (Didiano and Hobert 2006). Furthermore, new biological insights into the mechanisms of target binding have been obtained which could be used for predicting target sites. For example, target site accessibility to the RISC complex has been suggested as an important determinant of functional interactions.

We sought to develop a computational technique free from both the seed requirement and the conservation filter and thus capable of predicting species-specific and 3' compensatory target sites. Our method, *TargetSpy*, incorporates current biological knowledge in form of multiple sequence and structure features evaluated in the framework of an objective machine-learning prediction scheme.

However, since the (conserved) seed match is a strong determinant of target site detection, even though not the only one, we additionally generate predictions for sites with conserved and unconserved seed matches by post-filtering *TargetSpy* results.

We carried out extensive benchmark tests of the *TargetSpy* performance in human and *Drosophila melanogaster*. Our results suggest that *TargetSpy*, although trained on mouse, achieves the same performance as the best state-of-the-art methods in *D. melanogaster*, implying that the method can be applied to a broad taxonomic range of species for which no experimentally validated target sites are known. Furthermore, on the recently published experimental human dataset, describing changes in protein synthesis mediated by microRNAs, our method shows the highest accuracy among all tested prediction algorithms.

4.2. Methods

4.2.1. Dataset of 3' UTR sequences

We retrieved 3'UTR sequences from the UCSC Genome Database (Karolchik et al. 2008) using the UCSC Table Browser. For human (hg18, March 2006), mouse (mm8, July 2007), rat (rn4, November 2004) and chicken (galGal2, May 2006) we used the RefSeq Genes Track, for fly (dme, April 2006) we took the FlyBase annotations. For generating target site predictions considering conserved seeds, we used Galaxy (Giardine et al. 2005) to extract 3'UTR alignments for human, chimp, mouse, rat and dog from the 17-way human whole genome alignment and *D. melanogaster*, *D. yakuba*, *D. ananassae*, and *D. pseudoobscura* from the 15-way *d. melanogaster* whole genome alignment.

4.2.2. Dataset of MicroRNA sequences

All mature microRNA sequences originate from the miRBase, release 12 (Griffiths-Jones 2004). In total we retrieved 692 microRNAs for human, 513 for mouse, 443 for chicken and 147 for fly.

4.2.3. Target site predictions by previously published methods

The target predictions of *PicTar* (Krek et al. 2005) were downloaded from the UCSC database using the Table Browser and were migrated from hg17 to hg18 by applying the UCSC command line tool *liftover*. We used the predictions conserved in human, mouse, rat, chimp and dog (4-way) as well as the predictions additionally conserved in chicken (5-way). For fly we downloaded the sensitive prediction set (S1) of PicTar that is composed of predictions conserved in *D. melanogaster*, *D. yakuba*, *D. ananassae*, and *D. Pseudoobscura*, also via the UCSC Table Browser. Predictions for the human genome made by *miRanda* (John et al. 2004), release September 2008, were downloaded from http://microRNA.org (Betel et al. 2008). Only predictions for transcripts contained in the RefSeq database were considered. Human and fly predictions made by *miRBase Targets* (Enright et al. 2003), version 5, were downloaded from http://microrna.sanger.ac.uk/targets/v5/. *RNA22* (Miranda et al. 2006) predictions for human 3'UTR sequences were downloaded from http://cbcsrv.watson.ibm.com/rna22.html. Since these predictions were made using Ensembl transcripts, we mapped the predictions to RefSeq genes by applying mapping tables provided by Ensembl and UCSC. Predictions of *PITA* (Kertesz et al. 2007) were downloaded from http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html. We utilized the "*TOP*" and the "ALL" set with 3/15 flankings. *TargetScanS* (Lewis et al. 2005) predictions and the corresponding microRNA family mapping table were downloaded from http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_50.
Predictions made by Gaidatzis et al. (Gaidatzis et al. 2007) were downloaded from the EIMMo server http://www.mirz.unibas.ch/. Targets predicted by *mirTarget2* (version 3) (Wang and El Naqa 2008) were downloaded from http://mirdb.org/miRDB. Human target site predictions of *DIANA-microT* v3.0 (Maragkakis et al. 2009) were retrieved via the web server at

http://diana.cslab.ece.ntua.gr/microT/ for the thresholds loose (score=7.3) and strict (score=19). Finally, we downloaded the human target site predictions of TargetRank (Nielsen et al. 2007) from http://hollywood.mit.edu/targetrank/.

4.2.4. Experimental data for evaluation

Two sets of experimentally verified target sites were used to benchmark target prediction algorithms. For evaluation on *Drosophila melanogaster*, we used the 120 experimentally tested microRNA - gene interactions compiled by Stark et al. (2005) and the 190 interactions published by Kertesz et al. (2007). The former set is composed of 61 functional and 59 non-functional interactions; the latter set consists of 102 functional and 88 non-functional interactions. The appropriate 3'UTR sequences were derived from the FlyBase annotations provided by UCSC. Transcripts for which no 3' UTR was available were discarded. For evaluation on human, we used an experimental dataset that is based on the pSILAC technique and reveals fold changes in protein production caused by five selected microRNAs (Selbach et al. 2008), downloaded from http://psilac.mdc-berlin.de.

Free energy estimates

All duplex structures and energy estimates were calculated by the RNAduplex and RNAcofold programs from the Vienna package version 1.6.1 (Hofacker et al. 1994). We applied the option -noLP to exclude base pairs, which can only occur as lonely pairs and the option -e to retrieve all suboptimal structures instead of just the one with the minimum free. The minimum free energy that can be observed for a microRNA is defined as the energy value calculated for the duplex of the microRNA and its perfect reverse compliment.

4.2.5. Generation of candidate zones

The microRNA - mRNA interaction is typically characterized as an interval within the mRNA sequence that is almost perfectly reverse complementary to the microRNA sequence over a substantial fraction of the microRNA, or at least over a seed region of 6-8 bases. In this work we investigate the possibility to abandon the strict requirement for the presence of a seed region and attempt to find zones of high attraction between the microRNA and its target mRNA independent of seed

occurrence. Such candidate zones cover not just a particular binding site, but a larger stretch of sequence including several potential adjacent binding sites. This approach involves three subsequent steps illustrated in Figure 21A-C.

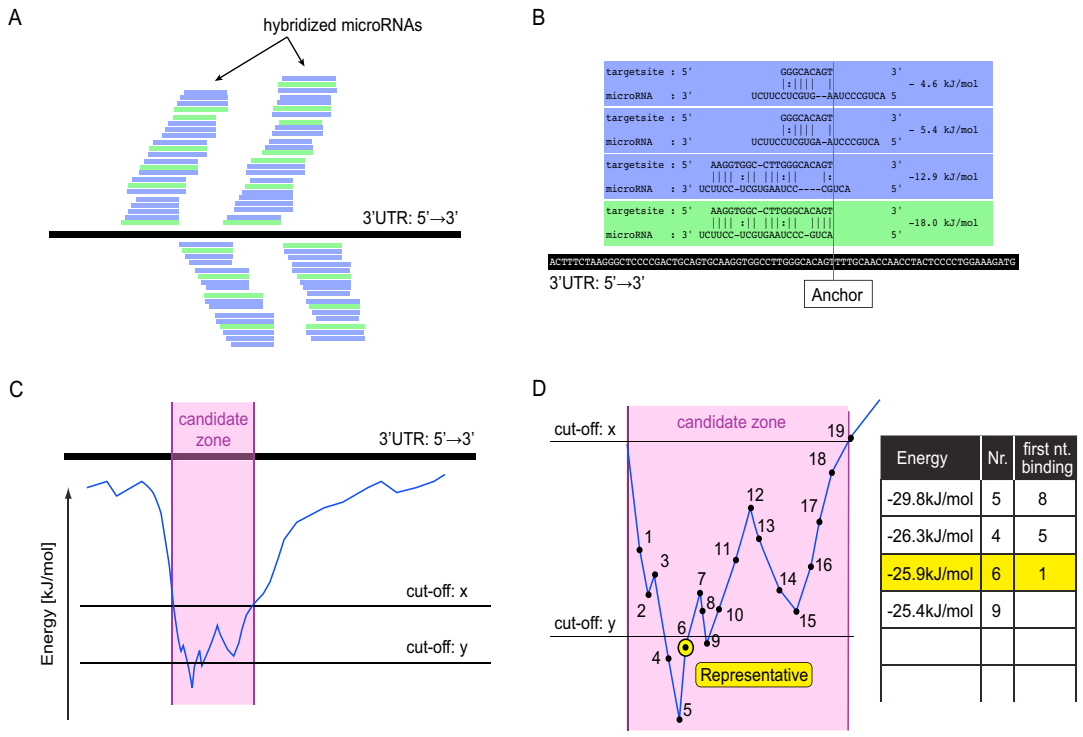


Figure 21: Schematic illustration of the candidate target site generation pipeline.
A) MicroRNA-mRNA duplexes sharing the same anchor position on the mRNA are grouped. Duplexes with the lowest free energy in each group are shown in green color, all others in blue. B) Zoom-in at one group. The anchor of each hybrid (red vertical line) is the first nucleotide of the target site base-pairing with the 5' end of the microRNA. Only the energetically most favorable hybrid, shown in green, is retained for further analysis. C) Smoothed attraction graph of all the retained hybrids. A candidate zone is defined as the stretch of the target sequence (shown in purple) where the smoothed hybrid free energy falls below a certain energy threshold. D) For each candidate zone the energetically most favorable hybrid that shows base pairing within the first two nucleotides counting from the microRNA 5' end is selected as its representative.

First, for a pair of microRNA and mRNA sequences, all possible duplex structures predicted by RNAduplex are ordered according to the sequence position of their anchor (see next section).

In a second step energy values of the selected duplexes are plotted against the respective anchor positions, resulting in a graph reflecting the attraction of individual areas of the mRNA towards the particular microRNA under study, measured in terms of Gibbs free energy values. To reduce local fluctuations the curve is smoothed by taking the average of the energy values for the current position and for its two immediate neighbors (*i.e.* by using a sliding window of length 3).

In the last step those mRNA areas with a particularly strong attraction for a given microRNA are identified based on the requirement that all energy values of predicted duplex structures be below a certain cut-off x, and at least one duplex, we call it the representative, be below a cut-off value y. Based on the current experimental knowledge (Ambros 2004) base pairing for the representative is additionally required to start with the first or second nucleotide of the microRNA counted from its 5'end. We call the areas satisfying these conditions candidate zones. The variables x and y are expressed in terms of the ratios between the observed energy of a duplex and the maximal energy of a given microRNA. For example a value of 0.25 means that the duplex has 25% of the energy of a perfect reverse complementary hybrid. In view of our intention to detect as many potential target sites as possible in the first step of our workflow we set x to 0.24 and y to 0.25, which is well below the energy cutoffs applied by other approaches (Krek et al. 2005).

On average this leads to eight candidate zones for each microRNA and 3'UTR in human (for comparison, seven sites were reported for the classic lin-4:lin-14 in *Caenorhabditis elegans* (Lee et al. 1993)).

4.2.6. Duplex stacking and anchor choice

Since we do not choose duplex structures based on significant free energy values, we obtain for each microRNA-mRNA pair a vast amount of overlapping predicted secondary structures, typically in the order of 5000-20000. Therefore we need to group the resulting structures according to their location on the mRNA. Structures with a G:U match in the first eight base pairings as well as those that show five or more G:U pairings in the entire duplex are discarded. Given that duplexes will

heavily overlap on the mRNA, we need to define anchor points in order to map the duplexes to a specific position on the mRNA sequence.

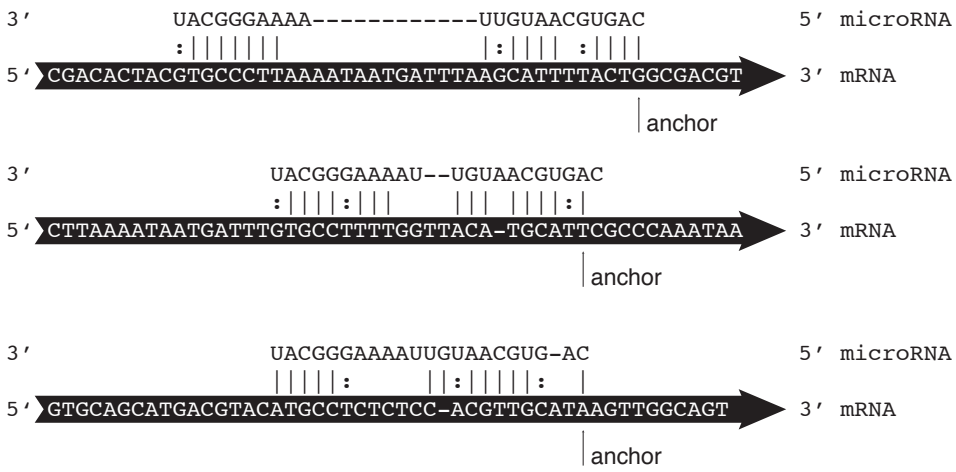


Figure 22: Defining the position of a microRNA-mRNA duplex on the mRNA sequence. The anchor is the position at which the first base pairing occurs, viewed from the mRNA 3' end (or microRNA 5' end).

Different types of anchors can be considered including, for example, the mRNA position where the first nucleotide pairing occurs, counting either from the 3' end or from the 5' end of the corresponding microRNA, or some middle point of the duplex. In view of the findings suggesting that base pairing at the 5' end of the microRNA particularly strongly contributes to target site recognition (Brennecke et al. 2005; Doench and Sharp 2004; Lai 2002; Lewis et al. 2003; Rajewsky and Socci 2004) we chose to define the anchor point as the position on the mRNA sequence where the first base-pairing with the microRNA 5' end occurs (Figure 22). If duplexes with a perfect microRNA 5' end pairing of at least 7 consecutive nucleotides (seed region) are present all other structures with the same anchor point and no seed region are discarded. Subsequently the energetically most favorable one for each anchor is selected as the best candidate for this specific anchor position (Figure 21A,B).

4.2.7. Training set

In order to develop a classifier of high quality, it is essential to obtain a training set that is truly representative of both the positive (actual target sites) and negative (non-target sites) class.

Recently a set of argonaute (Ago) - mRNA binding sites, identified by a novel technique that isolates RNA by crosslinking immunoprecipitation in high-thoughput experiments (HITS-CLIP), was published for the 20 most abundant microRNAs present in the P13 mouse brain (Chi et al. 2009). Argonautes are proteins that upon association with microRNAs form the RNA-induced silencing complex (RISC), which is responsible for the repression of target mRNA expression.

To our knowledge this is the first experimental data set that reports directly microRNA target sites in a large-scale fashion and *TargetSpy* is the first algorithm that is using it for training. We retrieved the data from http://ago.rockefeller.edu/ and removed all sites that did not map to 3'UTRs nor had a RefSeq accession number associated. Since only the microRNA family is specified in this publication, we identified the candidates for all microRNAs belonging to that family. Those candidates that overlapped the experimentally derived sites were retained as positive instances. In cases where several candidates of the same microRNA family overlapped, we took the energetically most favorable one. Target site candidates having no equivalent in the set of experimentally derived Ago binding sites are unlikely to be biologically relevant. We therefore identified the energetically most stable candidate for a reported Ago-mRNA interaction that does not overlap the validated Ago-binding site. Those candidates served as negative instances. In total we obtained 3872 positive and 4540 negative instances.

4.2.8. Features of microRNA - mRNA duplexes

In order to build an accurate microRNA target predictor it is of paramount importance to define a set of characteristics that effectively distinguish real microRNA-mRNA interactions from any other types of hybrids. Numerous properties of such duplexes have been reported in biological literature in recent years, and some of them have been incorporated in target prediction methods

developed earlier. Here, instead of relying on a limited number of empirically selected features we chose to objectively evaluate the performance of a possibly broad spectrum of pairing requirements within the framework of a machine learning approach. Below follows the list of features used in this study.

General extent of microRNA-mRNA binding

- Number of base-pairings to the microRNA 8-mer seed.
- Number of base-pairings to the first eight nucleotides of the microRNA 3' end.
- Number of consecutive base-pairings at the microRNA 3' end with two allowed non-pairing positions, beginning at the first base pairing position.
- Length of the longest stretch of consecutive base-pairings anywhere in the hybrid.
- Length of the target site.
- Binding asymmetry. Here we measured the ratio between the amounts of paired bases in the 3' versus the 5' region of the microRNA. We considered 8 nucleotides on each side.

Extent of G:U base pairing

- Total number of G:U wobble base pairs in the microRNA - mRNA hybrid.

Bulge-related features of duplexes

- Number of bulges on the microRNA.
- Number of bulges on the target site.
- Total bulge length on the microRNA.
- Total bulge length on the target site.
- Number of bulges on the microRNA. We tested the bulge lengths of 1,2,3,4 and 5 bases.

- Number of bulges on the target site. In this cases we tested bulge lengths of 1,2,3,4,5,6 and equal or greater than 7 bases as bulges on the mRNA sequence tend to be larger than those on microRNAs.
- Length of the second largest bulge on the microRNA.
- Length of the second largest bulge on the target site.
- Mean length of bulges on the target site.
- Number of symmetric bulges.

Position specific features

- Position of the target site in the 3'UTR. We split 3'UTRs into 100 bins and returned the index of the bin containing the anchor of the candidate zone's representative as the position of the target site.
- Following the reasoning of Lewis et al. (Lewis et al. 2005) we calculated four features related to the base occurrence at given positions. Specifically we recorded the nucleotides in the target site at microRNA positions 1 (t1 anchor) and 9 (t9 anchor) and the existence of an S (A or U) or W (G or C) base at the same positions (t1 S/W anchor and t9 S/W anchor).

Compositional features

Base composition of both 3' UTRs (Robins and Press 2005) and microRNAs plays an important role in mRNA-microRNA recognition. Here we employ the following compositional features:

- G+C content of the target site.
- G+C content of the 50 nucleotide long region upstream of the target site.
- G+C content of the 50 nucleotide long region downstream of the target site.
- G+C content ratio between the microRNA and the target site.
- Difference in G+C content between the target site and the upstream flanking region.

- Difference in G+C content between the first and the last eight nucleotides of the target site.
- Occurrence of CpG di-nucleotide in the target site sequence as well as in its 3' and 5' flanking regions.

The length of 3' and 5' flanking regions was taken to be 20 nucleotides, unless otherwise stated.

Compactness

We reasoned that hybrids that are more compact, i.e. having only few unpaired nucleotides both in the microRNA and in the target site are more likely to be biologically functional that others. Therefore our goal was to unify these two features into a single measure. We define the compactness of a hybrid as the mean value of the following ratios: *number of basepairs*/*microRNA length* and *number of basepairs*/*target site length*. Compactness values are thus in the range between 0 and 1, with the latter value corresponding to perfect complementarity. If the target site is shorter than the microRNA a penalty is introduced, as this case is not taken into account by the mean of the ratios stated above:

$$\text{compactness} = \frac{1}{2}\left(\frac{\text{basepairings}}{\text{microRNA length}} + \frac{\text{basepairings}}{\text{target site length}} \times f\left(\frac{\text{target site length}}{\text{microRNA length}}\right)\right)$$

$$f(x) = \begin{cases} \sqrt{\cos\left(\pi \times \frac{(1-x)}{2}\right)} & x < 1 \\ 1 & else \end{cases},$$

Accessibility of the target site to RISC

Recent literature (Kertesz et al. 2007; Zhao et al. 2005) suggests that target site accessibility to RISC is a critical factor in microRNA target recognition. Examples of approaches that have been developed to approximate accessibility are RNAup (Muckstein et al. 2006) and IntraRNA (Busch et al. 2008). We applied the definition of Kertesz et al. (2007) and calculated accessibility as the difference between the free energy of the microRNA hybrid and the energy of the local secondary structure of

the target site including 3 nt upstream and 15 nt downstream flanking sequences. We further tested all combinations for upstream and downstream flankings from 0nt to 30nt in 5 nt steps.

4.2.9. Classifier

Using the positive and negative instances we developed a classifier capable of distinguishing microRNA - mRNA duplexes from other hybrids in the feature space described above. The problem is that most of the biologically motivated features implicated in microRNA target recognition, with the exception of the seed match, display only a weak correlation with functionality. A standard approach to enhance the prediction performance in case of weak features is to utilize boosting. We therefore applied the learning scheme based on boosting called MultiBoost (Webb 2000) with decision stumps as base learner. In comparison with other methods we have tried (SVM (Xu et al. 2009), Naive Bayes (George H. John 1995), C4.5 (Quinlan 1993), AdaBoost (Freund 1996) with C4.5, MustiBoost with C4.5) it consistently produced superior results (Figure 23).

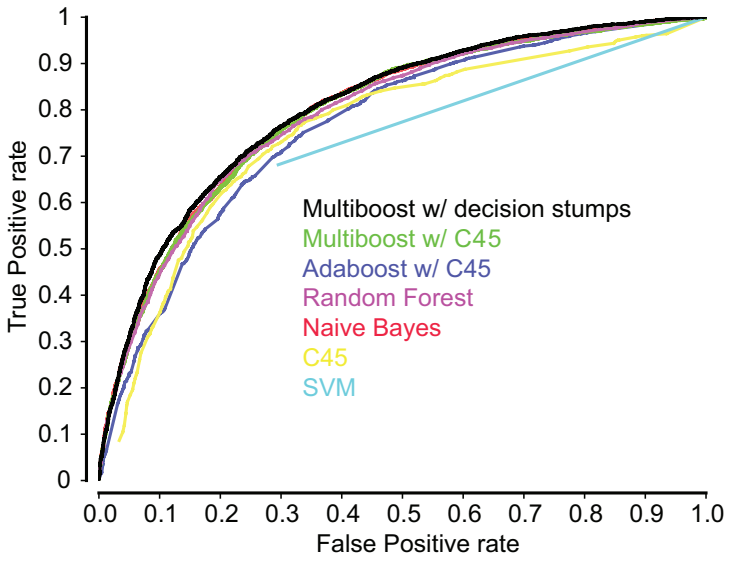


Figure 23: ROC curves generated by various classifiers evaluated in 10-fold cross-validations on the training set.

We used the WEKA (Witten and Frank 2005) implementation of the learning scheme and largely relied on the standard configuration provided by WEKA, with the exception of setting the number of bagging iterations to 200. For estimating the quality of each individual feature and for subsequent ranking of features, we used the Relief-F (Kononenko 1994) algorithm. ReliefF estimates the quality of features according to how well they distinguish between closely neighbored instances of different classes. To find the best possible set of features, we applied the feature evaluation approach called "Correlation-based Feature Selection" (Mark A. Hall 1998) (CFS) together with the best-first search algorithm. Only features from the subset computed by this filter approach are taken for the classifier.

4.2.10. Target site prediction

For each organism considered we predicted microRNA target sites and ranked them according to their score. As explained above *TargetSpy* initially considers every potential candidate zone and assigns a score to it. Overlapping candidate zones were merged together, and the representative with the highest score becomes the representative of the entire merged zone. This permissive approach generates vast amounts of candidates with very low scores. Additional criteria are subsequently imposed in various combinations to narrow down the set of predicted targets (Table 2). The naming of the prediction datasets is based on whether or not the presence of a seed region is required, and whether a permissive (sens) or strict (spec) threshold is applied.

Prediction dataset name	Seed match required	Conservation considered	False-positive rate threshold
TargetSpy no-seed sens	No	No	0.05
TargetSpy no-seed spec	No	No	0.01
TargetSpy seed sens	Yes	No	0.05
TargetSpy seed spec	Yes	No	0.01
TargetSpy cons. seed sens	Yes	Yes	0.05
TargetSpy cons. seed spec	Yes	Yes	0.01

Table 5: Applied threshold and limitation to the prediction subsets.

4.2.11. Evaluation of prediction performance

For assessing the quality of our classifier we used the following performance measures: *sensitivity, specificity*, accuracy and the *Matthews correlation coefficient* (MCC). As in any classification process four different possibilities have to be accounted for: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). For the evaluation on the training set, we obtained these values in form of a confusion matrix by performing a standard 10 fold cross validation followed by plotting a receiver operating characteristic (ROC) curve. From this we calculated the area under the curve (AUC) statistics, a measure that is understood as the probability that the classifier will assign a positive instance a higher score than a negative instance when picking an instance from each class randomly. Given the confusion matrix, sensitivity and specificity, are defined by the following equations:

$$sensitivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{FP + TN}$$

A single measure representing the predictive power of the classifier must account for all those four possibilities listed above. A factor considered to be one of the best performance measures is the Matthew's correlation coefficient (Baldi et al. 2000) given by:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Its value range is [-1,1]. A value of MCC = 1 indicates the best possible prediction, such that any positive instance is correctly predicted and only positive instances are predicted as positive. MCC = -1 indicates the worst possible prediction. A value of MCC = 0 would be expected for a random prediction scheme.

The evaluation on pSILAC data was performed as in Selbach et al. 2008. There, the performance measure (accuracy) was defined as the fraction of predicted mRNA targets with reduced protein production ($\log_2$ fold change < -0.1).

4.2.12. Implementation and availability

We implemented our method as a stand-alone *Java* program called *TargetSpy*. The program relies on the Java Virtual Machine version 1.5 and two freely available third party software packages - the Vienna package for RNA secondary structure prediction (version 1.6.1) and the data mining software WEKA (version 3.5.3). *TargetSpy* is available from our web site (http://www.targetspy.org) along with installation instructions and links to all required third party software packages.

4.3. Results and Discussion

4.3.1. Classification of prediction approaches

As discussed in 1.2.2, current tools for predicting microRNA target sites can be grouped into three distinct classes according to their requirements on target sites (see Table 6 for an overview of all tools, including TargetSpy, used in the following evaluations). Class I is constituted by those approaches that make use of neither the seed match requirement nor conservation. Class II contains all approaches that do require a seed match, but make no use of conservation. Finally, class III is for those predictors that both require a seed match and rely on conservation.

Some methods cannot be perfectly fitted into this scheme. For example, while *miRanda* does actually not require a perfect seed match it weights the seed region so high that on average just around 7% of all predicted target sites show mismatches to the 7-mer seed (microRNA nucleotides 1-7 or 2-8). *miRBase Targets* uses *miRanda* for candidate generation, and permits a single mismatch to the seed region. Since both approaches additionally require the target site to be conserved, we consider them as members of class III.

Organism	Seed match not required	Seed match required	Seed match required and conservation considered
Human	RNA22 TargetSpy no-seed	PITA All 3/15 TargetScanS non-conserved TargetSpy seed	EIMMo MiRBase Targets MiRanda PicTar DIANA-microT TargetScanS PITA TOP MirTarget2 TargetRank TargetSpy cons. seed
Fly	RNA22 TargetSpy no-seed	PITA All 3/15 TargetSpy seed	EIMMo PicTar MiRBase Targets TargetSpy cons. seed TargetScanS

Table 6: Classification of microRNA target site prediction tools.

Note also that *TargetSpy* generally belongs to class I, since our model does not impose a strict seed match requirement and does not consider conservation of target sites. However, we can easily build subsets of our predictions that satisfy the criteria of class II and III. Throughout this work we refer to the subset of *TargetSpy* predictions containing a perfect 7-mer seed match as *TargetSpy seed*. Likewise, *TargetSpy conserved seed* denotes a set of predicted target sites containing a conserved seed match.

4.3.2. Computational pipeline for predicting microRNA target sites

Our intention here is to build a pipeline for predicting microRNA target sites based on the multiple features described in the Methods section. At run time TargetSpy takes two multiple FASTA files as input; one with the 3' UTR sequences and the other one with the microRNA sequences. Note that no other extrinsic information, such as evolutionary conservation needs to be provided.

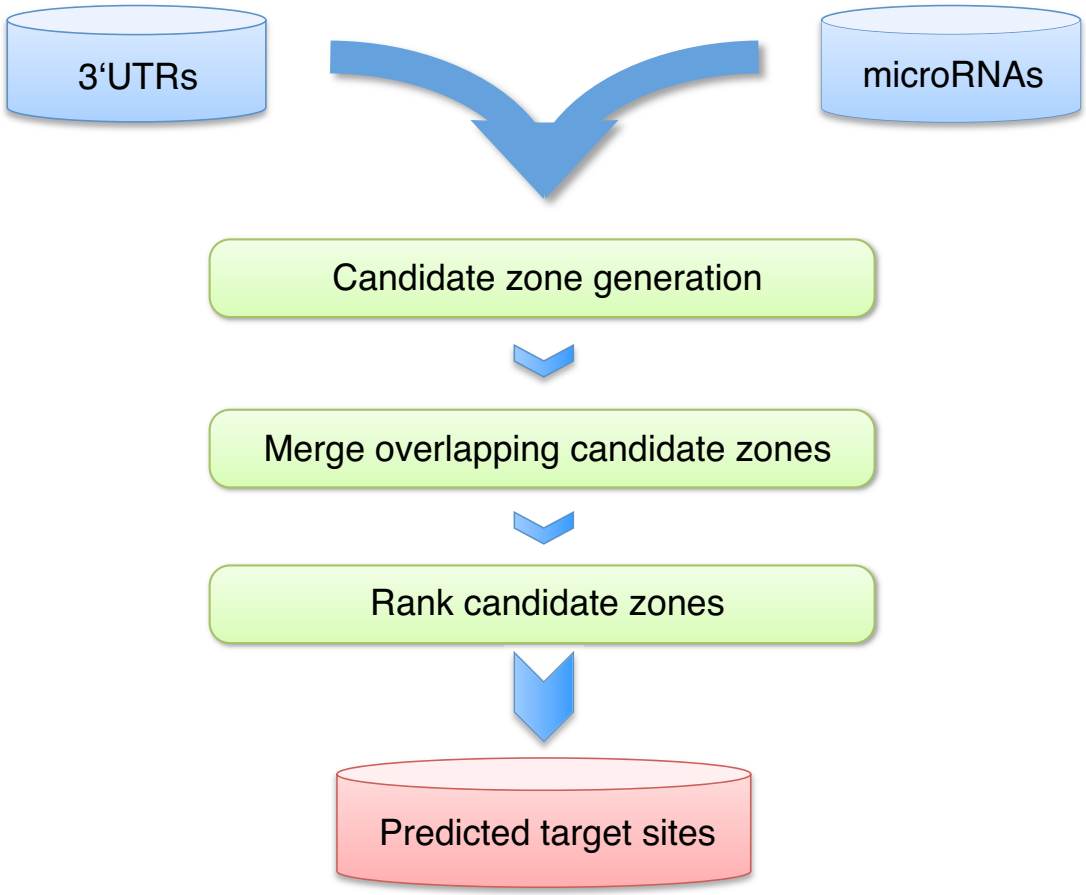


Figure 24: A schematic overview of the TargetSpy prediction pipeline.
Annotated 3’UTR sequences and all known microRNAs from a given species serve as input. MicroRNAs are matched against 3’UTRs to generate potential candidate zones. The resulting candidate zones are classified and ranked according to their score, with overlapping zones being merged together.

For each input microRNA TargetSpy identifies candidate zones (stretches of DNA sequence potentially harboring a target site) in all 3’UTR sequences. It calculates the score for the representative of each candidate zone, merges overlapping candidate zones, and ranks the predictions according to their scores (Figure 24, see Methods for details). Using this protocol target sites were predicted for *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus* and *Drosophila melanogaster* (Table 7).

	Number of 3'UTRs	Average 3'UTR length	Number of microRNAs	Number of predicted target sites			
				TargetSpy no-seed sens	TargetSpy no-seed spec	TargetSpy seed sens	TargetSpy seed spec
Human	26161	1210	692	4837k	1023k	829k	339k
Mouse	18694	1082	513	1906k	407k	340k	137k
Rat	11859	760	292	535k	113k	91k	36k
Chicken	3676	927	443	372k	80k	59k	24k
Fly	15884	471	147	247k	54k	50k	20k

Table 7: Number of target sites predicted in each species by different versions of the *TargetSpy* method. See *Methods* for more detail.

4.3.3. Target site candidates

A usual starting point of a prediction workflow is the search for perfect seed matches in the 3'UTR of transcripts of interest. Since our goal is to develop a model that does not rely on the presence of a seed match we had to redefine the rules for selecting initial candidate target sites. Following the reasoning that a functional site is more attracted by the loaded RISC complex than its surrounding area, we identify candidates by searching for areas in the target sequence where the predicted Gibbs free energy of the microRNA-target duplex is below a certain microRNA-specific energy threshold (see *Methods* for detail). To ensure a high coverage of functional binding sites we have chosen a conservative cut-off. With this candidate definition at hand, we identified about 150 million target site candidates for all microRNAs in human.

4.3.4. Selection of informative features and classifier evaluation

As described in the *Methods* section we evaluated a wide range of target site features by applying the Relief-F (Kononenko 1994) technique (see Table 8 for the ranked list of features). Some features generally considered to be highly relevant for target

site recognition by microRNA, such as the number of base parings to the microRNA seed, performed very well. On the other hand, the feature *accessibility* with 3 nt upstream and 15 nt downstream flankings, reported in Kertesz et al. (2007) to be strongly discriminative, was evaluated as poorly performing. To analyze whether this is due to the chosen flanking sequences, we tested other flanking settings and found 30 nt upstream and 30 nt downstream to perform slightly better than the 3/15 setting; however the improvement was marginal (data not shown). Interestingly, the feature *compactness* (combining the length of the target site and the number of nucleotides binding to the microRNA, see *Methods*), introduced in this work performs among the best.

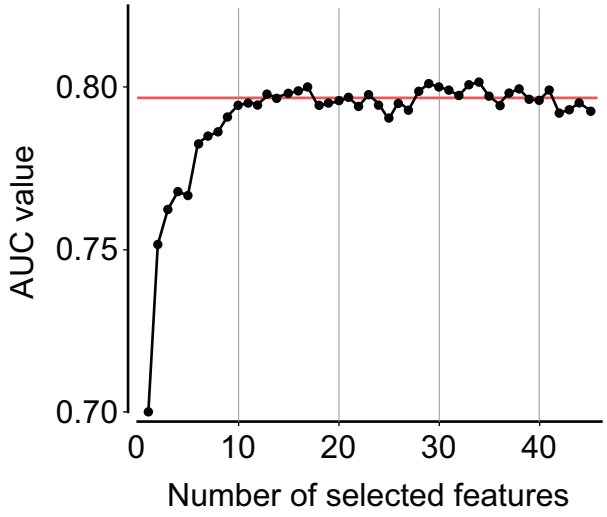


Figure 25: Classifier performance as a function the feature set size.
The classifier was evaluated in an iterative process where one feature was added in each step. Features were selected according to the ranked feature list (**Table 5**), beginning with the best feature. In black the AUC values (y-axis) for the corresponding feature set size (x-axis) are shown. The red line indicates the AUC value of the feature set that was achieved by the feature subset selection approach.

Subsequently we evaluated the performance of the classifier with respect to the features used for training. We started with the single best feature and incrementally added features from Table 8 one by one, according to the ranking. Each classifier was evaluated on the training set by a standard 10-fold cross-validation procedure, as implemented in WEKA (Witten and Frank 2005). Figure 25 shows the number of features used for building the classifier and the corresponding area under the curve

(AUC) values. Apparently, the AUC value increases up to the 14th added feature. Upon adding further features the performance begins to oscillate around the AUC value 0.79 and does not improve further.

Since Relief-F only considers one feature at a time and does not take into account the correlation between features, we additionally applied the *Correlation-based Feature Selection* (CFS) (Mark A. Hall 1998) for identifying the best feature subset. This approach returned an AUC value of 0.79 (Figure 25, red line) with a set of only seven features, namely compactness, G+C content ratio between microRNA and target site, length of the longest stretch of consecutive base-pairings anywhere in the hybrid, binding asymmetry, G+C content of the target site, number of base-pairings to the microRNA 8-mer seed, and the position of the target site in the 3'UTR. The first four features are used here for the first time, while the latter three have been proposed before (Bandyopadhyay and Mitra 2009; Kim et al. 2006; Lewis et al. 2005; Yousef et al. 2007). Interestingly, several of these features evaluated individually are classified merely as weak performers, while in combination the classifier exploits synergetic effects between features, making it the smallest set of all with a comparable performance.

Following the common practice of selecting from all competing models of equal performance the simpler one, we chose the feature set generated by CFS for our machine learning technique.

Since *TargetSpy* tries in the first step to identify as many potential target sites as possible and subsequently ranks those according the classifier score, enormous amounts of target sites are produced from which only a fraction, the top predictions, are of interest. In order to make application and benchmarking of *TargetSpy* more transparent, we created a subset of predictions with high sensitivity and high specificity. The recognition thresholds were set such that the target sites with a false-positive rate lower than 5% (as evaluated in a 10-fold cross-validation) were assigned to the sensitive subset and those with a false-positive rate of 1% or less to the specific set (Table 5).

Rank	Features	Score
1	Number of base parings to the microRNA 8-mer seed	0.03175
2	G+C content of target site	0.01263
3	Number of base pairings to the first 8 nucleotides of the microRNA 3' end	0.01038
4	Number of consecutive base-pairings to the microRNA 3' end with two allowed non-pairing positions	0.00995
5	Occurrence of CpG in target site	0.00799
6	G+C content ratio between the microRNA and the target site	0.00642
7	Compactness	0.00619
8	T9 anchor	0.00556
9	Longest stretch of consecutive base-pairings in the hybrid	0.00513
10	Number of bulges in the microRNA of size three	0.00498
11	T1 S/W anchor	0.00491
12	Total number of base-pairings	0.00475
13	Number of bulges on the target site of size seven or greater	0.00442
14	T1 anchor	0.00434
15	Number of bulges in the microRNA of size two	0.00433
16	Occurrence of CpG in the upstream flanking area	0.00383
17	Number of bulges in the target site of size one	0.00374
18	Total bulge length of the target site	0.00362
19	Length of the target site	0.00336
20	Total bulge length of the microRNA	0.00334
21	Target site position within the 3'UTR	0.00333
22	Number of symmetric bulges	0.00290
23	G+C content upstream of the target site	0.00287
24	Number of bulges on the target site	0.00286
25	Length of the second largest bulge on the target site	0.00268
26	Mean length of bulges on the target site	0.00263
27	T9 S/W anchor	0.00261
28	Binding asymmetry	0.00255
29	Number of bulges in the target site of size two	0.00240
30	Total number of G:U wobble base pairs	0.00227
31	Local RISC accessibility 30/30	0.00220
32	Local RISC accessibility 3/15	0.00215
33	Number of bulges in the target site of size four	0.00210
34	Difference in G+C content between the first and the last nt of the target site	0.00201
35	Occurrence of CpG in downstream flanking area	0.00179
36	Number of bulges in the microRNA of size one	0.00179
37	Length of the second largest bulge on the microRNA	0.00174
38	Number of bulges on the microRNA	0.00153
39	Number of bulges in the microRNA of size five	0.00128
40	Number of bulges in the target site of size three	0.00113
41	Difference in G+C content between the target site and the 20 nt upstream and downstream flanking region	0.00112
42	Number of bulges in the target site of size five	0.00100
43	Number of bulges in the microRNA of size four	0.00084
44	G+C content downstream of the target site	0.00084
45	Number of bulges in the target site of size six	0.00021

Table 8: A ranked list of all features used in this work.
The score is calculated by the ReliefF method**.**

4.3.5. Evaluation on experimentally verified data

We next set out to evaluate the quality of the learning scheme implemented in *TargetSpy* on experimentally verified data and to benchmark *TargetSpy* against commonly used methods. This task is challenging as published methods are based on different principles, which makes it hard to compare them in a fair fashion. Our current knowledge about microRNA target sites is almost exclusively drawn from a handful of experiments exploring the targeting of a minority of the most highly expressed microRNAs (Baek et al. 2008). These experiments may have a strong selection bias in that they usually analyze the impact of microRNA overexpression or depletion on conserved molecular mechanisms. In addition, experimentally identified targets are often biased towards computational prediction approaches used to identify the initial pool of candidates (Stark et al. 2005).

Recently the impact of microRNA overexpression and knockdown was analyzed in large-scale proteomic studies (Baek et al. 2008; Selbach et al. 2008) not suffering from the selection bias discussed above. A further advantage is that none of the prediction approaches were trained on these data. However the data were generated by mass spectrometry and are therefore prone to an expression bias, although the authors state that this bias is mild. An additional complication is that the changes in protein expression of only a few microRNAs were measured, but the precise location of the target site in the respective transcript was not determined.

However, until high quality deep sequencing data like those from (Chi et al. 2009) become available in large amounts, these data constitute the current gold standard. Hence we use the fly dataset (Kertesz et al. 2007; Stark et al. 2005) and the human pSILAC dataset (Selbach et al. 2008) for evaluation.

Performance comparison in *Drosophila melanogaster*

In 2005 Stark et al. (Stark et al. 2005) conducted a broad comparison of widely used target prediction approaches. A set of 133 experimentally tested functional (61) and non-functional (59) microRNA-gene interactions was compiled, from which 120 were used for the actual comparison (Stark et al. 2005). This dataset served as the standard of truth to evaluate the evolutionary approach to microRNA target

prediction published by Gaidatzis et al. (Gaidatzis et al. 2007) and was later extended to 190 interactions by Kertesz et al. (2007). Note that this latter set also includes the 13 interactions that were excluded by Stark and colleagues since their respective 3'UTRs were not annotated. To assess the predictive power of our method and compare it with other methods, we applied it first to the original set and then to the extended set of targets. Since we assign each candidate zone a score, we are able to quantify the performance by a receiver operating characteristic (ROC) curve, making the comparison to other approaches more transparent.

When comparing the performance of methods on the original dataset of Stark et al. (2005) (Figure 26A,C) EIMMo achieves the best results, showing a high true-positive rate coupled with a low false-positive rate. Then follow PicTar, *TargetSpy seed,* PITA ALL 3/15, TargetScanS and *TargetSpy conserved seed* in the order of decreasing AUC values. The next best approach is *TargetSpy no-seed*, followed, with significant distance, by miRanda and finally RNA22, that is performing marginally better than random.

Despite the benefit of being able to compare all methods by just one value, looking at the ROC curve progression is even more enlightening, especially for those methods that are clustered closely together by the AUC value. Particularly interesting is the characteristic of the curves at low false-positive rates as for many experiments the amount of samples may be strongly limited. *TargetSpy conserved seed* shows the lowest false-positive rate (FPR) in the test up to a true-positive rate (TPR) of 48%. *TargetSpy seed* shows the second lowest FPR, but offers slightly better TPR, comparable to that of PicTar. Note that *TargetSpy no-seed* shows a performance that is close to class II and III methods, especially for its top predictions that cover more than 50% in TPR.

Benchmarking on the extended set of 190 experimentally verified microRNA-target interactions (see *Methods*) produces several interesting observations (Figure 26B,D). First, the AUC values are generally lower compared to the original set. Second, PITA, specifically fitted to this set, is far ahead of all other approaches. Third, the ranking of the other approaches has not changed except that i) *TargetSpy seed* performs ahead of PicTar and EIMMo, *TargetSpy conserved seed* outperforms

TargetScanS and miRBase Targets performs better than PITA TOP 3/15 and ii) the relative distance between *TargetSpy no-seed* and TargetScanS is reduced. Finally, the specificity in particular that of EIMMo and TargetScanS, suffered strongly especially for their top predictions.

In summary, the evaluation on experimental fly data suggests that *TargetSpy*, which was trained on mouse data, performs as good as current state-of-the-art algorithms when enforcing the seed match criterion. Furthermore, the no-seed prediction is notably better than RNA22, the other tested algorithm that does not require a perfect seed match.

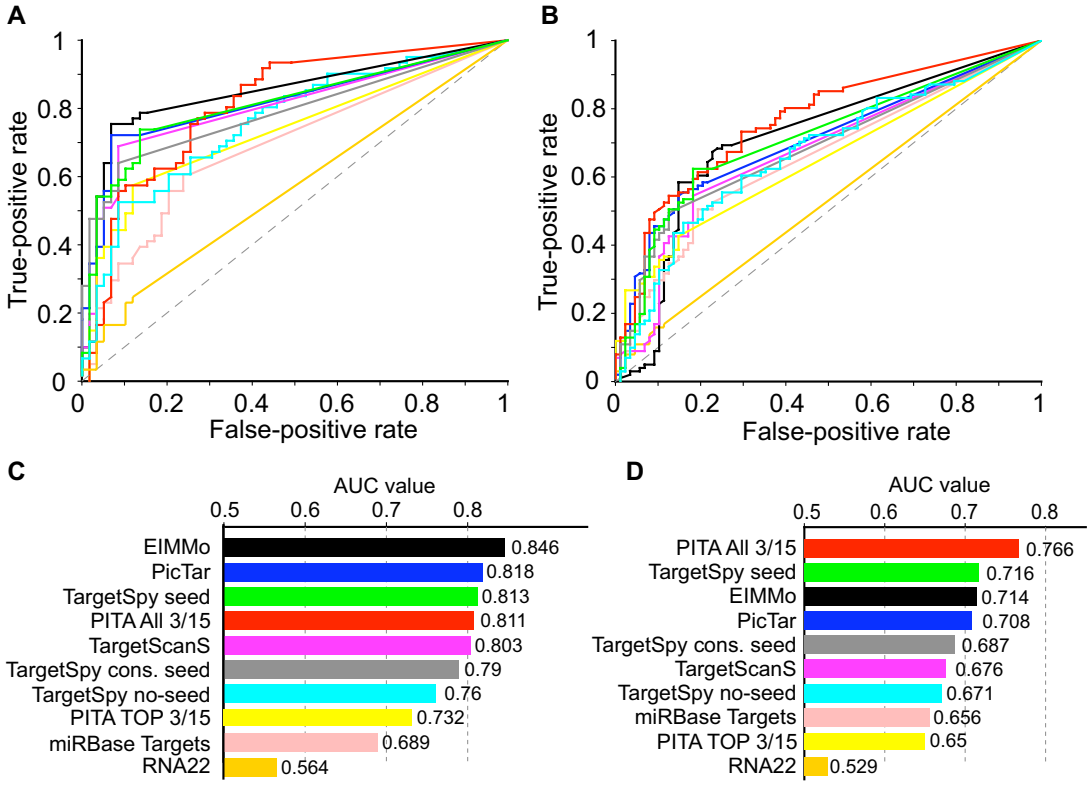


Figure 26: Performance comparison of target prediction approaches.
A) and C) refer to the dataset compiled by Stark et al. (Stark et al. 2005)). B) and D) refer to the dataset compiled by Kertesz et al. (2007). A) and B) show the ROC curves of the tested approaches, C) and D) the AUC values. The gray line indicates the performance of random guessing.

Evaluation on pSILAC data

Selbach et al. (Selbach et al. 2008) performed a comparison of the most widely used approaches by measuring the fraction of predicted target sites associated with proteins that are more strongly down-regulated than -0.1 $\log_2$ fold change. They generated two background (random) sets: i) a set where all mRNAs present are considered as targets, and ii) a set of all mRNAs that have a 6-mer seed match in their sequence, further referred to as the *Selbach background*. In order to compare class III predictors to the random expectation, we additionally introduced a background dataset for conserved seed matches as proposed by (Bartel 2009). Specifically, we searched for 6-mer (positions 2-7) seed matches that are perfectly conserved in human, chimp, mouse rat and dog that show additionally a match to either base 1 or 8 (Chen and Rajewsky 2006). This way we have background sets produced by trivial prediction strategies in place for each of the three classes of prediction tools.

As seen in Figure 27 for the first class (no seed/no conservation) a completely random selection of target sites would yield a ~27% intersection (background accuracy) with down-regulated proteins in pSILAC. Both *RNA22* and *TargetSpy no-seed* perform better than random. *TargetSpy no-seed* attains the accuracy of 34.2%, a significant improvement compared to random predictions. RNA22 shows 36.2% accuracy, however at a sensitivity that is more than 6.6 times lower than *TargetSpy no-seed sens*. In the specific setting *TargetSpy no-seed contains* still more than 2.3 times as many target sites as RNA22, but achieves an accuracy of 42.9% and is thus on the same level as the 6-mer seed background of class II.

The background accuracy for class II is at 42.6% when using 6mer seeds. As PITA covers all target site candidates with seed matches beginning at the size of 6 nt and subsequently ranks them according to their accessibility, it is necessary to consider only its top ranking predictions. Following Selbach et al. 2008 (Selbach et al. 2008) we took the top 1000 predictions per microRNA and found a 42.3% overlap with down-regulated proteins demonstrating an accuracy below the background level. TargetScanS, predicting non-conserved target sites with at least a 6-mer seed match, shows a higher accuracy (47.9%) than PITA and noticeably out-performs the 6-mer

seed background, although it does not pass the accuracy of the 7-mer seed background. For 7mer seeds, which are used by *PicTar* and *TargetSpy seed*, the corresponding background accuracy was 48.4%. Both *TargetSpy* seed *sens* and *TargetSpy seed spec* perform clearly better than this trivial prediction, showing accuracies of 52.8% and 55.3%, respectively, and thus perform best in class II.

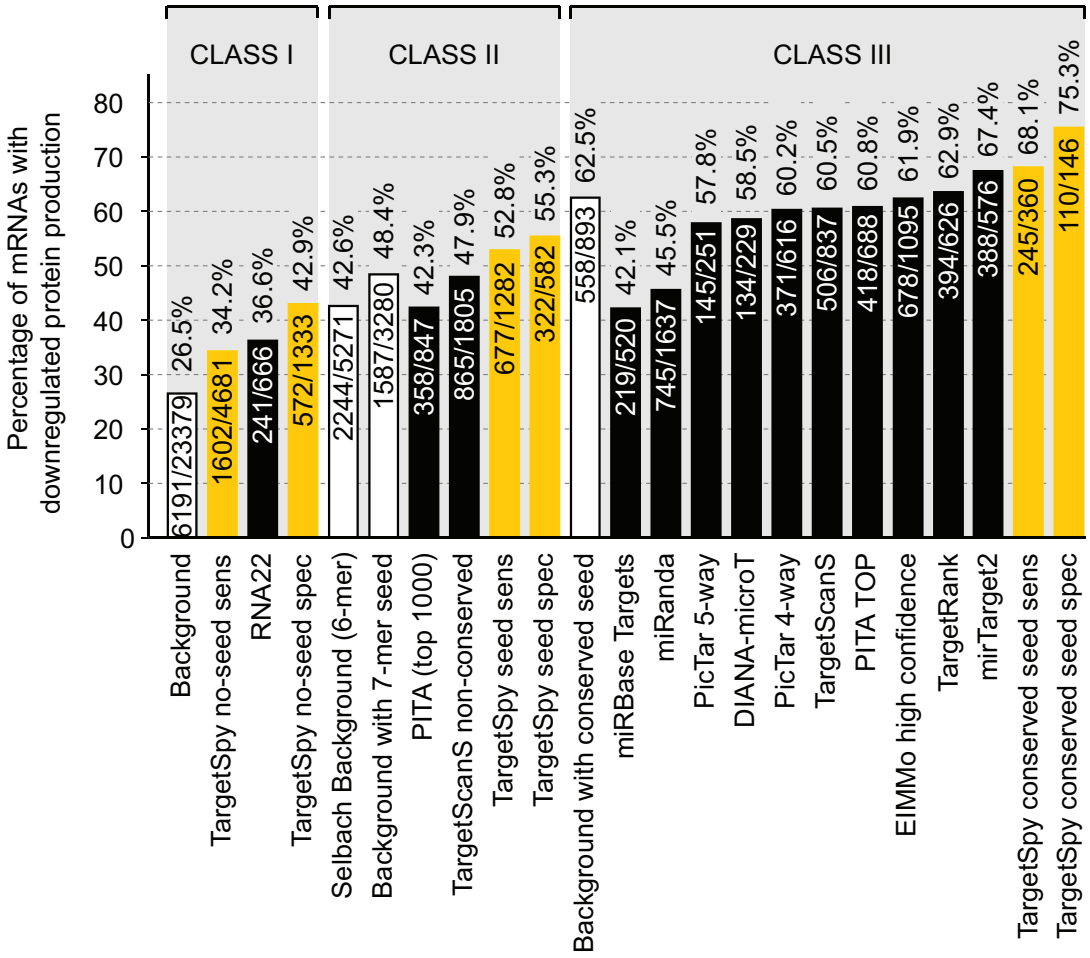


Figure 27: Performance evaluation of various prediction approaches on the pSILAC data set.
This set contains changes in protein production caused by the five microRNAs miR-1, miR-16, miR-155, miR-30a-5p and let-7b. The first value in each bar represents the number of microRNA-target interactions that are associated with down-regulation (log_2-fold change < -0.1) and the second value reports the total number of interactions predicted for the pSILAC set. The value on top of each bar displays the accuracy. White bars with black outlines display the trivial predictors, *TargetSpy* is represented in orange and other approaches are displayed in black.

The final class of target prediction approaches (seed/conservation) shows an overlap of 62.5% with random predictions. Both *miRanda* and *miRBase Targets* predictions include a small fraction of conserved target sites with imperfect seed regions, and the corresponding accuracies (45.5% and 42.1%, respectively) are below the background and close to that of the trivial seed predictor of class II. *PicTar 4-way* (those predictions where the seed match is conserved among 4 species*)* reaches an accuracy of 60.2%, which is also below the background. Interestingly the more stringent PicTar 5-way that additionally requires conservation in chicken performs worse that PicTar 4-way. Also DIANA-microT, TargetScanS, PITA TOP, an official subset of PITA with conserved 8-mer seed matches required, and EIMMo with the high confidence setting (score >= 0.5) are performing slightly below background. The first approach performing above the trivial predictor of class III is TargetRank (62.9%), followed by mirTarget2 (67.4%) and the sensitive subset of *TargetSpy conserved seed* (68.1%). Finally, *TargetSpy conserved seed spec* achieves the highest accuracy of all methods (75.3%). It should be noted, however, that although *TargetSpy* achieves superior performance in terms of accuracy and sensitivity in classes I and II, the sensitivity of *TargetSpy* in class III is lower compared to other approaches. The higher sensitivity of some approaches might be attributed to the choice of seed match that is enforced. EIMMo, for example, integrates several different seed match definitions (including also short 6-base long ones) and shows a sensitivity that is 2.7 times higher than our approach at the sensitive threshold and more than 6 times higher when compared to our specific setting.

To exclude the possibility that the performed evaluation is only valid for the chosen threshold of -0.1 $\log_2$ fold change, we also investigated the cumulative fraction of predicted target sites as a function of the protein $\log_2$ fold change. Figure 28 shows the distributions for each of the three classes. It becomes apparent that the relative performance of computational approaches remains practically unchanged for each fold change value in each class. However, as seen in Figure 28C, the advance of *TargetSpy conserved seed spec* (green line) is particularly pronounced for low $\log_2$ fold change values. Since low fold change values correspond to stronger protein down-regulation this may imply that our approach performs even better for highly efficient target sites.

In general our results on the pSILAC data suggest that *TargetSpy* performs best in each class, showing furthermore a constant gain in accuracy from the sensitive to the specific threshold. This finding implies that the prediction quality increases with the score and therefore the ranking of target sites imposed by the score of our model seems to have biological relevance.

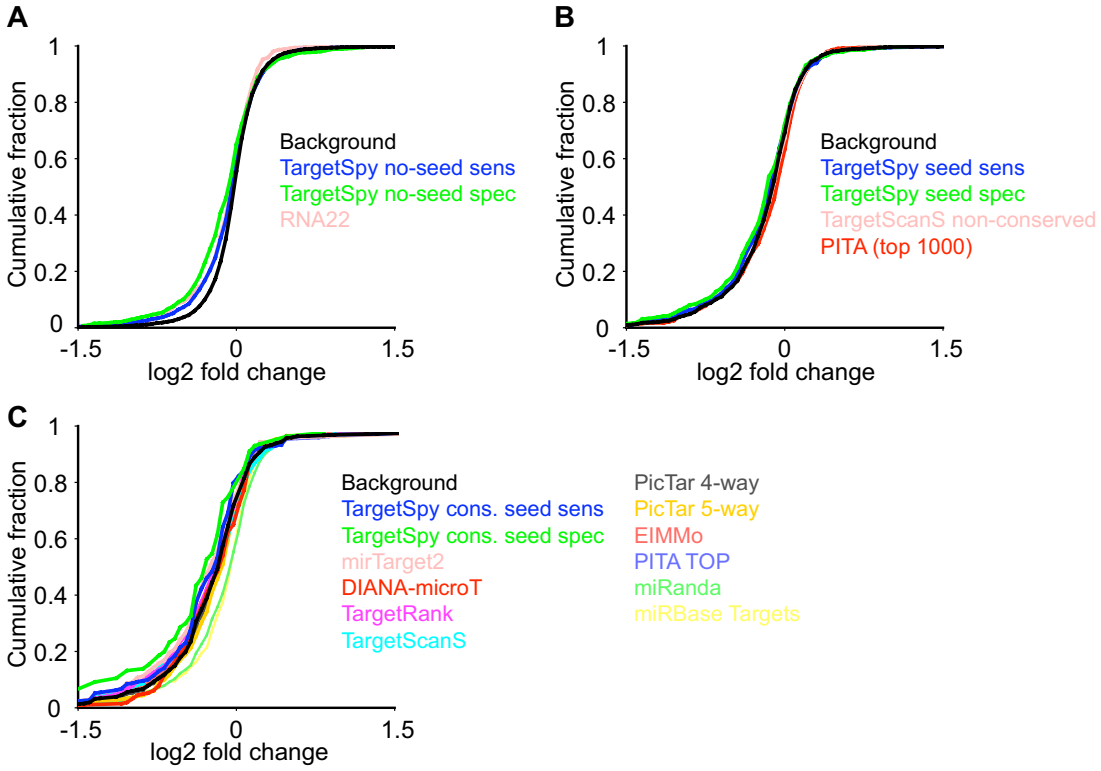


Figure 28: Cumulative fraction of predicted target sites of down-regulated proteins according to the measured fold change.
The distributions are given for A) approaches not requiring a seed (class I), B) approaches requiring a seed (class II), and C) approaches requiring a seed and considering site conservation (class III).

Finally, we determined the number of functional target sites (down-regulated proteins) not possessing seed regions, which have been correctly predicted solely by *TargetSpy* but not by any other algorithm. To avoid putatively spurious seedless target sites, we excluded from consideration those gene-microRNA interactions where both potential target sites containing seed regions and those not containing them are present. After removing such ambiguous gene-microRNA interactions we obtained 564 unique, target sites without seed region in the sensitive set and 134 in

the specific set. This means that *TargetSpy* reports between 26 (spec) and 112 (sens) functional target sites showing no seed match per microRNA that could not be detected by any other tool.

4.4. Conclusion

In summary, we have developed a novel computational approach for predicting microRNA target sites that neither implies the existence of a seed nor utilizes phylogenetic footprinting. Instead of using rigid rules and/or arbitrarily selected target site features, we objectively derived a set of discriminative features to be used for machine learning. Due to these important advantages *TargetSpy* is able to predict species specific (*i.e.* unconserved) target sites, is suitable for processing poorly conserved/low quality genomic sequences for which methods that rely on conservation and species specific information will not work, and allows for analyzing differences in microRNA targets between various species.

We grouped computational prediction approaches into three classes, depending on their usage of a seed match criterion in order to provide a comparison of their performance among each other and against the background level. On an experimentally derived microRNA-target interaction set in *Drosophila*, our method is on par with the best available approaches. In a further benchmark on human microRNA-target data generated by the pSILAC technology, *TargetSpy* not only reported the highest accuracies in class I, but also in the other two classes for which our predictions where post filtered according to the class definition. Given that *TargetSpy*, trained on experimentally derived Ago binding sites in the 3'UTR of mouse transcripts, showed very good performance when evaluated in fly and human, we suggest that our algorithm can be applied to a broad taxonomic range of organisms.

Finally, we have shown that even on a small high quality target site specific data set, derived by a deep sequencing experiment (Chi et al. 2009), machine learning techniques show high potential in the prediction of microRNA target sites. We assume that advances in this direction will become even more pronounced as more data of this kind become available.

Chapter 5

Conclusion and Outlook

The microRNA research is a very young scientific discipline. Although back in 1993 the first cornerstone was laid as the (microRNA) gene lin-4 was discovered in *C. elegans*, the observed phenomenon was believed to be a special case in nematodes. Initial in the year 2000 with the discovery of microRNA let-7 in *C. elegans* as well as homologs in fly, human and eleven other bilateral animals, the microRNA pathway was considered as a general concept on gene regulation in eukaryotes. The knowledge we have about microRNAs is therefore the result of intensive research of less than a decade.

Today, a considerable amount of microRNA genes in various species is known. However, the end of the flagpole is not reached, as the number is constantly growing. Further, only for very few microRNAs their function is at least partly understood. Consequently, the very next major milestone in microRNAs research is the functional annotation of all microRNAs of an organism. Therefore the two goals towards that milestone are the detection of all microRNAs and the identification of their targets.

With the technology progression of deep-sequencing platforms, vast amounts of actually expressed RNA sequences become available. Our pipeline *miRanalyzer* solves two major bioinformatics challenges:

First, the multitudinous amounts of sequence reads are clustered and mapped to the genome. Consequently, annotations from various databases like RFam, RepBase and

miRBase can be associated with the mapped reads. Hence, *miRanalyzer* provides an analysis of deep-sequencing experiments in the sense that known microRNAs are identified and reported together with their read copy number, an estimate for the expression level.

Second, clustered reads that could not be mapped to known annotations are tested if they constitute a functional but unknown microRNA. The two major key points here are the exploit of the Dicer footprint, visible due to next generation sequencing technologies for the first time, and the usage of a machine learning technique trained on experimentally derived data. The combination of this highly distinctive model and the fact that deep-sequencing data actually contain sequence reads from expressed RNA molecules only, led to a prediction accuracy that is strongly ahead of traditional approaches. Besides, we completely abstained from conservation such that *miRanalyzer* is suitable for the detection of species-specific microRNAs.

The other method, *TargetSpy*, we have developed in this thesis predicts microRNA target sites without requiring a seed match and conservation. It is based on machine learning techniques, a broad spectrum of biologically motivated features and an automatic feature selection approach. *TargetSpy* is trained on recently available deep-sequencing data of Ago binding sites, a dataset that constitutes the current gold standard in means of actual target site information. Due to these important advantages we are able to predict species-specific target sites with and without a seed match, process poorly conserved genomic sequences and allow for analyzing differences in microRNA targets between various species.

Depending on the usage of a seed match criterion and conservation, we have structures the field of prediction approaches into three distinct classes. To perform meaningful evaluations, we post-filtered the predictions of TargetSpy to meet the minimum requirements of the class considered. Conditioned on the evaluations, *TargetSpy* performs either on par with the best available approaches or even superior in every prediction class. As *TargetSpy* was trained on mouse data, but evaluated on human and fly data, we suggest that our algorithm can be applied to a broad taxonomic range of organisms.

Altogether we have provided two approaches in this work, one for each of the targeted issue. Either of them contributes significantly to the research field. Both methods utilize next-generation sequencing data and both are based on machine learning techniques. One additional critical factor is the usage of biologically motivated features that are evaluated in an automatic feature selection approach.

Ultimately, the ideas and methods presented in this thesis may be extended and supplemented in various directions in future research.

First and foremost, the list of features used in both approaches can be easily extended and adapted such that novel experimental findings can be accounted for in future releases. *MiRanalyzer* for instance is highly focused on features intrinsic to the read sequence data from deep-sequencing experiments. Integrating novel features that are based on characteristics of the mapped genomic location for example promotor regions or other microRNA loci (gene cluster) may improve prediction accuracy. Also the features currently used are a source for inspiration and improvements. On the one side, one may combine two features by expert knowledge to create a more powerful third feature. Promising candidates therefore are certainly GC content related features and RISC accessibility. On the other side features themselves could be improved. Again, RISC accessibility is a promising candidate, as currently only the energies of the microRNA-target hybrid and the self-folded target sites are considered. Improvements could be achieved when the actual structure change around the target site is measured between the unbound and bound state.

We have developed a whole new approach in the identification of microRNA target site candidates, a necessary step since we abandoned the usage of a seed match requirement. We have chosen conservative parameters to lose as few functional target sites as possible. However, the reverse of this approach is that we end up with enormous amounts of apparently improper candidates that clutter the successive classification step. As new biological insights become available that allow a more restrictive screening for candidates without losing actual functional sites, those criteria may be integrated into the *TargetSpy* pre-processing step. Consequently the prediction accuracy may improve while simultaneously the time needed for prediction will decrease as the amount of improper candidates is reduced.

Ultimately, the usage of deep-sequencing technology is at the outset. As more of these high quality data will become publicly available, both of our developed methods will directly profit, as they just have to be retrained on the new data. Further, more extensive and unbiased evaluations will be possible such that the benefits of individual features and prediction approaches will become visible in a much more detailed fashion.

In the end it will be the tight interplay of experimental design and bioinformatics tools that ensure a continuous increase in microRNA knowledge and model advancements. Through these we will gain a more sophisticated understanding of organism complexity that will lead to important applications and improvements of disease therapies.

Appendices

List of Tables

List of Figures

Bibliography

Abelson, J.F., Kwan, K.Y., O'Roak, B.J., Baek, D.Y., Stillman, A.A., Morgan, T.M., Mathews, C.A., Pauls, D.L., Rasin, M.R., Gunel, M. et al. 2005. Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science (New York, N.Y* **310:** 317-320.

Ambros, V. 2004. The functions of animal microRNAs. *Nature* **431:** 350-355.

Baek, D., Villen, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. 2008. The impact of microRNAs on protein output. *Nature* **455:** 64-71.

Bagasra, O. and Prilliman, K.R. 2004. RNA interference: the molecular immune system. *Journal of molecular histology* **35:** 545-553.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., and Nielsen, H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics (Oxford, England)* **16:** 412-424.

Bandyopadhyay, S. and Mitra, R. 2009. TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics (Oxford, England)* **25:** 2625-2631.

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

Bartel, D.P. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* **136:** 215-233.

Baskerville, S. and Bartel, D.P. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA (New York, N.Y* **11:** 241-247.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E. et al. 2005. Identification of hundreds of conserved and nonconserved human microRNAs. *Nature genetics* **37:** 766-770.

Berezikov, E., Cuppen, E., and Plasterk, R.H. 2006a. Approaches to microRNA discovery. *Nature genetics* **38 Suppl:** S2-7.

Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R.H., and Cuppen, E. 2005. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120:** 21-24.

Berezikov, E., van Tetering, G., Verheul, M., van de Belt, J., van Laake, L., Vos, J., Verloop, R., van de Wetering, M., Guryev, V., Takada, S. et al. 2006b. Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome research* **16:** 1289-1298.

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

Betel, D., Wilson, M., Gabow, A., Marks, D.S., and Sander, C. 2008. The microRNA.org resource: targets and expression. *Nucleic acids research* **36:** D149-153.

Bonnet, E., Wuyts, J., Rouze, P., and Van de Peer, Y. 2004. Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. *Proceedings of the National Academy of Sciences of the United States of America* **101:** 11511-11516.

Breiman L. 2001. Random Forests. *Machine Learning* **45:** 28.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. 2005. Principles of microRNA-target recognition. *PLoS biology* **3:** e85.

Busch, A., Richter, A.S., and Backofen, R. 2008. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics (Oxford, England)* **24:** 2849-2856.

Cai, X., Hagedorn, C.H., and Cullen, B.R. 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA (New York, N.Y* **10:** 1957-1966.

Carthew, R.W. 2006. Gene regulation by microRNAs. *Curr Opin Genet Dev* **16:** 203-208.

Chen, K. and Rajewsky, N. 2006. Natural selection on human microRNA binding sites inferred from SNP data. *Nature genetics* **38:** 1452-1456.

Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. 2009. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460:** 479-486.

Clop, A., Marcq, F., Takeda, H., Pirottin, D., Tordoir, X., Bibe, B., Bouix, J., Caiment, F., Elsen, J.M., Eychenne, F. et al. 2006. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nature genetics* **38:** 813-818.

Cullen, B.R. 2004. Transcription and processing of human microRNA precursors. *Molecular cell* **16:** 861-865.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F., and Hannon, G.J. 2004. Processing of primary microRNAs by the Microprocessor complex. *Nature* **432:** 231-235.

Didiano, D. and Hobert, O. 2006. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nature structural & molecular biology* **13:** 849-851.

Didiano, D. and Hobert, O. 2008. Molecular architecture of a miRNA-regulated 3' UTR. *RNA (New York, N.Y* **14:** 1297-1317.

Doench, J.G. and Sharp, P.A. 2004. Specificity of microRNA target selection in translational repression. *Genes & development* **18:** 504-511.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. 2003. MicroRNA targets in Drosophila. *Genome biology* **5:** R1.

Esquela-Kerscher, A. and Slack, F.J. 2006. Oncomirs - microRNAs with a role in cancer. *Nat Rev Cancer* **6:** 259-269.

Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B., and Bartel, D.P. 2005. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science (New York, N.Y* **310:** 1817-1821.

Filipowicz, W., Bhattacharyya, S.N., and Sonenberg, N. 2008. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature reviews* **9:** 102-114.

Forstemann, K., Tomari, Y., Du, T., Vagin, V.V., Denli, A.M., Bratu, D.P., Klattenhoff, C., Theurkauf, W.E., and Zamore, P.D. 2005. Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. *PLoS biology* **3:** e236.

Freund, Y.S., R. E. 1996. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 148-156. Morgan Kaufmann, Bari, Italy.

Friedlander, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. 2008. Discovering microRNAs from deep sequencing data using miRDeep. *Nature biotechnology* **26:** 407-415.

Gaidatzis, D., van Nimwegen, E., Hausser, J., and Zavolan, M. 2007. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC bioinformatics* **8:** 69.

Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R. et al. 2009. Rfam: updates to the RNA families database. *Nucleic acids research* **37:** D136-140.

George H. John, P.L. 1995. *Estimating Continuous Distributions in Bayesian Classifiers*. Morgan Kaufmann, San Mateo.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J. et al. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome research* **15:** 1451-1455.

Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic acids research* **32:** D109-111.

Griffiths-Jones, S. 2006. miRBase: the microRNA sequence database. *Methods in molecular biology (Clifton, N.J* **342:** 129-138.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell* **27:** 91-105.

Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G., and Mello, C.C. 2001. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell* **106:** 23-34.

Hackenberg, M. and Matthiesen, R. 2008. Annotation-Modules: a tool for finding significant combinations of multisource annotations for gene lists. *Bioinformatics (Oxford, England)* **24:** 1386-1393.

Hammell, M., Long, D., Zhang, L., Lee, A., Carmack, C.S., Han, M., Ding, Y., and Ambros, V. 2008. mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *Nature methods* **5:** 813-819.

Han, J., Lee, Y., Yeom, K.H., Kim, Y.K., Jin, H., and Kim, V.N. 2004. The Drosha-DGCR8 complex in primary microRNA processing. *Genes & development* **18:** 3016-3027.

He, L. and Hannon, G.J. 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nature reviews* **5:** 522-531.

Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M., and Schuster, P. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly* **125:** 167--188.

Hofacker, I.L. 2003. Vienna RNA secondary structure server. *Nucleic acids research* **31:** 3429-3431.

Huang, T.H., Fan, B., Rothschild, M.F., Hu, Z.L., Li, K., and Zhao, S.H. 2007. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC bioinformatics* **8:** 341.

Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Balint, E., Tuschl, T., and Zamore, P.D. 2001. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science (New York, N.Y* **293:** 834-838.

Hutvagner, G. and Zamore, P.D. 2002. A microRNA in a multiple-turnover RNAi enzyme complex. *Science (New York, N.Y* **297:** 2056-2060.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., and Lu, Z. 2007. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic acids research* **35:** W339-344.

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. 2004. Human MicroRNA targets. *PLoS biology* **2:** e363.

Jones-Rhoades, M.W. and Bartel, D.P. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Molecular cell* **14:** 787-799.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110:** 462-467.

Karolchik, D., Kuhn, R.M., Baertsch, R., Barber, G.P., Clawson, H., Diekhans, M., Giardine, B., Harte, R.A., Hinrichs, A.S., Hsu, F. et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic acids research* **36:** D773-779.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. 2007. The role of site accessibility in microRNA target recognition. *Nature genetics* **39:** 1278-1284.

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

Kim, S.K., Nam, J.W., Rhee, J.K., Lee, W.J., and Zhang, B.T. 2006. miTarget: microRNA target gene prediction using a support vector machine. *BMC bioinformatics* **7:** 411.

Kim, V.N. and Nam, J.W. 2006. Genomics of microRNA. *Trends Genet* **22:** 165-173.

Kiriakidou, M., Tan, G.S., Lamprinaki, S., De Planell-Saguer, M., Nelson, P.T., and Mourelatos, Z. 2007. An mRNA m7G cap binding-like motif within human Ago2 represses translation. *Cell* **129:** 1141-1151.

Kolfschoten, I.G., Roggli, E., Nesca, V., and Regazzi, R. 2009. Role and therapeutic potential of microRNAs in diabetes. *Diabetes Obes Metab* **11 Suppl 4:** 118-129.

Kononenko, I. 1994. Estimating Attributes: Analysis and Extensions of RELIEF, pp. 171--182. Springer Verlag.

Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M. et al. 2005. Combinatorial microRNA target predictions. *Nature genetics* **37:** 495-500.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. 2001. Identification of novel genes coding for small expressed RNAs. *Science (New York, N.Y* **294:** 853-858.

Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., and Tuschl, T. 2002. Identification of tissue-specific microRNAs from mouse. *Curr Biol* **12:** 735-739.

Lai, E.C. 2002. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nature genetics* **30:** 363-364.

Lai, E.C. 2004. Predicting and validating microRNA targets. *Genome biology* **5:** 115.

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

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

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M. et al. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129:** 1401-1414.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. 2001. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science (New York, N.Y* **294:** 858-862.

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

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S. et al. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425:** 415-419.

Lee, Y., Jeon, K., Lee, J.T., Kim, S., and Kim, V.N. 2002. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J* **21:** 4663-4670.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., and Kim, V.N. 2004. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* **23:** 4051-4060.

Legendre, M., Lambert, A., and Gautheret, D. 2005. Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics (Oxford, England)* **21:** 841-845.

Lewis, B.P., Burge, C.B., and Bartel, D.P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15-20.

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. 2003. Prediction of mammalian microRNA targets. *Cell* **115:** 787-798.

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433:** 769-773.

Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. 2003. The microRNAs of Caenorhabditis elegans. *Genes & development* **17:** 991-1008.

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

Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A. et al. 2005. MicroRNA expression profiles classify human cancers. *Nature* **435:** 834-838.

Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., and Cui, Q. 2008. An analysis of human microRNA and disease associations. *PLoS ONE* **3:** e3420.

Maragkakis, M., Reczko, M., Simossis, V.A., Alexiou, P., Papadopoulos, G.L., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K. et al. 2009. DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic acids research* **37:** W273-276.

Mark A. Hall, L.A.S. 1998. Feature Subset Selection: A Correlation Based Filter Approach. In *Thesis submitted in partial fulfilment of the requirements of the degree of Doctor of Philosophy at the University of Waikato*.

Mathonnet, G., Fabian, M.R., Svitkin, Y.V., Parsyan, A., Huck, L., Murata, T., Biffo, S., Merrick, W.C., Darzynkiewicz, E., Pillai, R.S. et al. 2007. MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F. *Science (New York, N.Y* **317:** 1764-1767.

Meister, G. 2007. miRNAs get an early start on translational silencing. *Cell* **131:** 25-28.

Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.S., Tam, W.L., Thomson, A.M., Lim, B., and Rigoutsos, I. 2006. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126:** 1203-1217.

Muckstein, U., Tafer, H., Hackermuller, J., Bernhart, S.H., Stadler, P.F., and Hofacker, I.L. 2006. Thermodynamics of RNA-RNA binding. *Bioinformatics (Oxford, England)* **22:** 1177-1182.

Nam, J.W., Kim, J., Kim, S.K., and Zhang, B.T. 2006. ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucleic acids research* **34:** W455-458.

Nam, J.W., Shin, K.R., Han, J., Lee, Y., Kim, V.N., and Zhang, B.T. 2005. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic acids research* **33:** 3570-3581.

Nielsen, C.B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., and Burge, C.B. 2007. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA (New York, N.Y* **13:** 1894-1910.

Nottrott, S., Simard, M.J., and Richter, J.D. 2006. Human let-7a miRNA blocks protein production on actively translating polyribosomes. *Nature structural & molecular biology* **13:** 1108-1114.

Olsen, P.H. and Ambros, V. 1999. 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.

Ouellet, D.L., Perron, M.P., Gobeil, L.A., Plante, P., and Provost, P. 2006. MicroRNAs in Gene Regulation: When the Smallest Governs It All. *Journal of biomedicine & biotechnology* **2006:** 69616.

Peters, L. and Meister, G. 2007. Argonaute proteins: mediators of RNA silencing. *Molecular cell* **26:** 611-623.

Pillai, R.S., Bhattacharyya, S.N., and Filipowicz, W. 2007. Repression of protein synthesis by miRNAs: how many mechanisms? *Trends Cell Biol* **17:** 118-126.

Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*. CA: Morgan Kaufmann, San Mateo.

Rajewsky, N. and Socci, N.D. 2004. Computational identification of microRNA targets. *Developmental biology* **267:** 529-535.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. 2000. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403:** 901-906.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. 2002. MicroRNAs in plants. *Genes & development* **16:** 1616-1626.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. 2002. Prediction of plant microRNA targets. *Cell* **110:** 513-520.

Robins, H. and Press, W.H. 2005. Human microRNAs target a functionally distinct population of genes with AT-rich 3' UTRs. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 15557-15562.

Rodriguez, A., Griffiths-Jones, S., Ashurst, J.L., and Bradley, A. 2004. Identification of mammalian microRNA host genes and transcription units. *Genome research* **14:** 1902-1910.

Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115:** 199-208.

Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. 2008. Widespread changes in protein synthesis induced by microRNAs. *Nature* **455:** 58-63.

Sethupathy, P., Corda, B., and Hatzigeorgiou, A.G. 2006. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA (New York, N.Y* **12:** 192-197.

Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S., Brownstein, M.J., Tuschl, T., van Nimwegen, E., and Zavolan, M. 2005. Identification of clustered microRNAs using an ab initio prediction method. *BMC bioinformatics* **6:** 267.

Sheng, Y., Engstrom, P.G., and Lenhard, B. 2007. Mammalian microRNA prediction through a support vector machine model of sequence and structure. *PLoS ONE* **2:** e946.

Sherry, S.T., Ward, M., and Sirotkin, K. 1999. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome research* **9:** 677-679.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29:** 308-311.

Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. 2005. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell* **123:** 1133-1146.

Tang, G., Reinhart, B.J., Bartel, D.P., and Zamore, P.D. 2003. A biochemical framework for RNA silencing in plants. *Genes & development* **17:** 49-63.

Thermann, R. and Hentze, M.W. 2007. Drosophila miR2 induces pseudo-polysomes and inhibits translation initiation. *Nature* **447:** 875-878.

Wakiyama, M., Takimoto, K., Ohara, O., and Yokoyama, S. 2007. Let-7 microRNA-mediated mRNA deadenylation and translational repression in a mammalian cell-free system. *Genes & development* **21:** 1857-1862.

Wang, X. and El Naqa, I.M. 2008. Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics (Oxford, England)* **24:** 325-332.

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

Webb, G.I. 2000. MultiBoosting: A Technique for Combining Boosting and Wagging. *Machine Learning* **40:** 159-196.

Witten, I. and E, F. 2005. Data Mining: Practical machine learning tools and techniques. *Morgan Kaufmann, San Francisco* **2nd Edition**.

Witten, I.H. and Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques*. 2nd Edition edition. Morgan Kaufmann, San Francisco.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434:** 338-345.

Xu, Z., Dai, M., and Meng, D. 2009. Fast and efficient strategies for model selection of Gaussian support vector machine. *IEEE Trans Syst Man Cybern B Cybern* **39:** 1292-1307.

Xue, C., Li, F., He, T., Liu, G.P., Li, Y., and Zhang, X. 2005. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC bioinformatics* **6:** 310.

Yekta, S., Shih, I.H., and Bartel, D.P. 2004. MicroRNA-directed cleavage of HOXB8 mRNA. *Science (New York, N.Y* **304:** 594-596.

Yousef, M., Jung, S., Kossenkov, A.V., Showe, L.C., and Showe, M.K. 2007. Naive Bayes for microRNA target predictions--machine learning for microRNA targets. *Bioinformatics (Oxford, England)* **23:** 2987-2992.

Zeng, Y. and Cullen, B.R. 2004. Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. *Nucleic acids research* **32:** 4776-4785.

Zhang, L., Kasif, S., Cantor, C.R., and Broude, N.E. 2004. GC/AT-content spikes as genomic punctuation marks. *Proceedings of the National Academy of Sciences of the United States of America* **101:** 16855-16860.

Zhao, Y., Samal, E., and Srivastava, D. 2005. Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature* **436:** 214-220.

Publications

Parts of this thesis have appeared in the following publications:

Hackenberg M*, Sturm M*, Langenberger D, Falcon-Perez JM, Aransay AM (2009) miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res 37: W68-76.
(* Joint first authors)

Sturm M, Hackenberg M, Langenberger D, Frishman D (2009) TargetSpy: a supervised machine learning approach for microRNA target prediction. BMC Bioinformatics (revised version submitted)

Supervised theses

Diploma theses:

"A computational method to reduce RNAi off-target effects by artificially designed microRNAs in mammals" (*David Langenberger*)

"The role of RNAi in HIV-1 infection" (*Oliver Krieg*)

Bachelor thesis:

"Human polymorphisms at predicted microRNA target" (*Christopher Huptas*)