Fachhochschule
Weihenstephan

Fakultät Biotechnologie und Bioinformatik

Diplomarbeit
in Bioinformatik

A computational method to reduce RNAi off-target effects by artificially designed siRNAs in mammals

David Langenberger

June 17, 2008

Betreuer:

Prof. Dr. Dmitrij Frishman
Lehrstuhl für genomorientierte Bioinformatik, TU München

Prof. Dr. Frank Lesske
Fakultät für Biotechnologie und Bioinformatik, FH Weihenstephan

Eidesstattliche Erklärung:

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

Freising, den 17.06.2008

.....
David Langenberger

Acknowledgements

This work was carried out in the Frishman lab of the Department of Genome-Oriented Bioinformatics at the Technical University of Munich in Freising. I thank all colleagues for the good working atmosphere and the scientific—and sometimes maybe not so scientific—discussions.
I am grateful to my supervisor Prof. Dr. Dmitrij Frishman for his scientific support, and the opportunity to write this thesis in his lab. Especially, I want to thank Martin Sturm for his supervision. He gave lots of food for thought, scientific support, helped in many discussions and supported my work wherever he could. I also want to thank Thomas Rattei for giving me helpful support and advice about how to structure my thesis.

I express my sincere thanks to Prof. Dr. Bernhard Haubold for extensive comments on an earlier version of this thesis and for his help concerning any kind of questions.

During the time I worked on this thesis, I enjoyed fruitful discussions with many people. In particular, I gratefully acknowledge Oliver Krieg, Patrick Tischler, Philipp Eser, Elisabeth Steidele and Christian Hainzinger for their contributions.
I am also very grateful to Meghan Ho from UBC for correcting my English in this thesis.

Last, but not least, I want to thank my parents for putting me through university and backing me all the way.

Summary

RNA interference (RNAi), mediated by short interfering RNAs (siRNAs), is widely used to silence gene expression and to define gene function in mammalian cells. Initially, this gene silencing via transcript degradation was believed to be highly specific, requiring nearly perfect base-pairing between the siRNA and the target mRNA. However, several recent publications have indicated that siRNAs can influence the expression of unintended genes in a microRNA-like manner. MicroRNAs are endogenous RNAs, approximately 22 nt in length. They play important regulatory roles in animals and plants by targeting mRNAs for cleavage or translational repression. Since microRNAs use presumably the same RNAi pathway as siRNAs, siRNAs can also target other than the intended mRNAs and subsequently repress their translation by acting in a microRNA-like manner. As these off-target effects can lead to measurable phenotypes and hence complicate the analysis of the effects of the actually intended knockdown, it is of up-most importance to minimize them. In this thesis I developed a novel approach for designing functional and efficient siRNAs with minimal off-target effects. Therefore potent siRNA canidates were generated based on the lastes published design rules in a first step and evaluated in terms of off-target effects in a second step. Finally a ranked list with the most potent siRNAs producing least off-target effects on top are returned.

Contents

1 Introduction

1.1 Motivation for this work

After completion of the human genome sequence, the next milestone for researchers is to map all the functions of the encoded genes. Even though a lot of work is done already, little is still known about the functions of most gene products. Hence, the systematic identification of single elements, as well as their intra-cellular interactions, are of major interest for both basic research and drug development. Therefore, methods are developed, which can reveal these functions in a fast and efficient way. RNA interference (RNAi) is one of these tools. RNAi offers researchers the unique possibility to specifically turn of genes of interest, by injecting 21 nt long, short interfering RNAs (siRNAs) that have a perfectly complementary target site on the messenger RNA (mRNA) of a specific gene. After binding that target site, the mRNA is cleaved and degraded, resulting in a gene knock-down and hopefully a change in the phenotype of the gene. But besides some almost solved obstacles, like the safe delivery of the siRNAs to the exact tissues within the human body, or triggering innate immunity mechanism, i.e. a non-specific Type I interferon response, which killed a lot of mammals in the beginning stages of RNAi research, there remains one major problem, the so-called off-target effects. Off-target effects are all the unintended ramifications that are directly caused by the injected siRNA. There are two classes of off-target effects. a) siRNA-sequence independent side effects, like the insertion of too many siRNAs, resulting in saturation of the RNAi pathway and thus blocking the endogenous regulation pathways. And b) siRNA-sequence dependent off-target effects, caused by siRNAs that silence not only the intended gene, but also others. Sequence dependent effects can partly be detected by searching for other genes with perfect complementary sites, which would also lead to cleavage of their mRNA. Several siRNA designing approaches use a simple BLAST search, to find these perfect matches. Some newer findings have shown that there can also be off-target effects, caused by imperfect bindings, targeting in a microRNA-like manner. Since siRNAs seem to use the same pathway as microRNAs, they can most likely regulate in the same way and thus not only trigger the RNAi pathway resulting in mRNA-cleavage, but also inhibit its translation. The latter is activated by imperfect target bindings, containing mismatches and bulges. microRNA off-target effects are ignored in almost all designing tools published yet [1]. The latest upgrad in that field is integrated by Dharmacon [1]. They search for the so-called seed regions, which are perfect complementary bindings of the nucleotides 2-8 of the siRNAs 5' end. These regions are thought to play a major role in microRNA:target recognition. But since the seed regions have been shown to cover only a fraction of microRNA tar-

gets, the Dharmacon siRNA design tool still does not predict all potential microRNA-like off-target effects. In order to find these missed target sites, up-to-date microRNA target prediction knowledge has to be used. Therfore, in this thesis, I present a completely new method of predicting and consequently minimizing not only perfect, but also all imperfect (microRNA-like) sequence dependent off-target effects.

1.2 RNA interference

RNA interference (RNAi) is a potent method using only a few double stranded RNA (dsRNA) molecules per cell to silence the gene expression. The efficiency of repression and ease of use made it a very important tool. For the same reasons is has been one of the most newsworthy topics in molecular biology in the last few years [2, 3, 4]. Prior to RNAi, scientists conducted gene knock out by using antisense, dominant negative or knockout techniques which were expensive and time consuming, but the discovery of RNAi has enabled to knock out genes in any organism efficiently and instantly [5].

1.2.1 Discovery of RNAi

Not until some recent discoveries [6, 7], only very little was known about the biology of small, non-coding RNAs (sncRNAs). Antisense RNA technologies were developed in 1988 to knock down genes in petunia and tomato [5, 8]. To this time, no one knew that the regulation was actually performed by small RNAs originated from the antisense RNA. Two years later, Jorgensen and his group discovered that the expression of a gene was suppressed, when adding extra copies of the same gene [9]. They called this effect "co-suppression", but again, they did not envisage that tiny RNAs where actually behind this phenomenon. Even as antisense RNA techniques became established as a tool for studying gene function in animals, the biology of small RNA was still not understood. 1995, Gao and Kemphues performed an experiment to analyse the downregulation of specific genes. For this purpose they injected worms with antisense RNA and the respective sense strand as control [10]. Actually, in that experiment they simultaneously repeated both traditional procedures, antisense RNA and "co-suppression". They got the result they hoped to get, namely the knockdown of the par-1 transcript from *C. elegans*. However, injecting the sense strand of the RNA into the embryos as a control, curously produced the same knockdown effect. This extraordinary experiment in worms caught the attention of Fire and Mello. Puzzled by the question, why both strands of the RNA, the sense and the antisense have similar gene silencing effects, they designed their own experiments and finally realized that not the single stranded antisense or the control sense RNA were the real trigger for the gene silencing, but the traces of double stranded RNAs

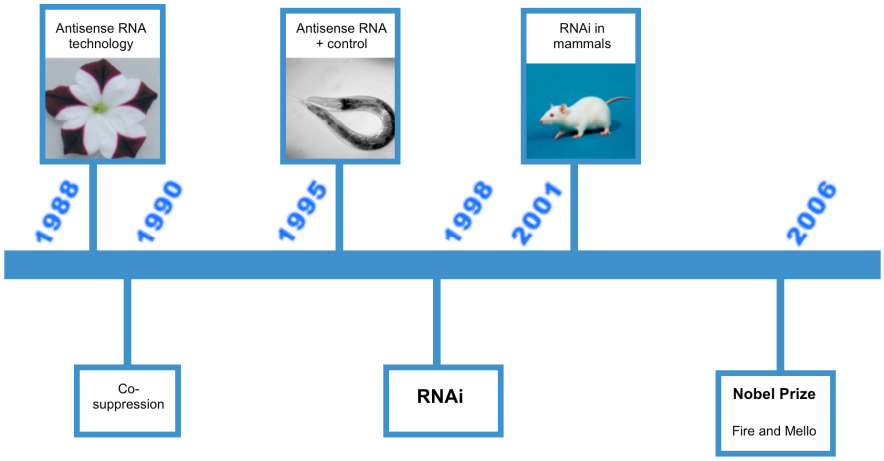

Figure 1: Discovery of RNAi

(dsRNAs) that had contaminated the antisense and sense RNAs used for injection. Their exceptional insights and speculations led to the discovery of RNA interference (RNAi) in 1998 [11, 12]. This phenomenon in which small amounts of dsRNA induce silencing of genes that share the same sequence of the dsRNA, set one important milestone for modern biology. It gave researchers a tool, to knock down genes of interest in a highly specific and instant way. But unfortunately that method did not work in mammals, since the long dsRNAs that were used as triggers for the RNAi, led to the induction of a non-specific Type I interferon response rather than sequence-specific silencing. This interferon response results in widespread changes in protein expression, masking any sequence-specific effects, eventually leading to apoptosis. One year after the discovery of RNAi, Tuschl and Zamore diverged from their previous research and started to establish a biochemical *in vitro* system to study the RNAi mechanisms [13]. Soon they discovered that small RNAs of 21-23 nucleotides (nt), termed short/small interfering RNAs (siRNAs), are the key players in mediating specific RNA degradation [14, 15]. Finally, Fire and Mello won the 2006 Nobel Prize in Physiology or Medicine for their work leading to the discovery of RNA interference (Figure 1).

But there was actually another part of the small RNA world, that emerged parralel to the RNAi discovery. In 1993, Ambros and colleagues discovered that an endogenous short RNA of 22 nt, lin-4, transcribed from a non-protein-coding region of the genome, controls aspects of developmental timing in worms. After discovering another microRNA in worms, let-7, the hunt for these , called microRNAs led to the discovery of hundreds of new small non-coding RNAs. The following years, researchers payed a lot of intention to microRNAs and started to understand the mechanism of how microRNAs regulate their targeted genes. Quickly it became clear that siRNAs share the same endogenous pathway

as microRNAs.

1.2.2 Small RNAs

Different classes of small RNAs, which are based on their distinct biogenesis, can influence several levels of gene regulation. The most prominent classes are siRNAs and microRNAs.

1.2.2.1 siRNAs

siRNAs are short pieces of dsRNAs, approximately 21-25 nt in length, which are central for RNAi. Most of the time, this double stranded RNA is composed of a 21 nt sense and a 21 nt antisense strand that are paired resulting in a 2 nt 3' overhang. siRNA directly induces the RNAi pathway and by binding to an almost perfect complementary region of the targeted transcript, cleaves the mRNA using the endonuclease Argonaute 2 (Ago2). In the majority of cases, siRNAs are synthesized chemically and then introduced into the cytoplasm by transfection or electroporation. The RNAi effect thus achieved is transient, lasting typically for 3–7 days. Other methods for producing synthetic siRNAs are:

- ***In Vitro* Transcription (IVT)**
  The required sense and antisense strands are transcribed from *in vitro* synthezised DNA oligonucleotide templates (Figure 2)

- ***In Vitro* Transcription of Long dsRNA Followed By Cleavage With Dicer or RNase III**
  Synthetic DNA oligonucleotide templates are used to transcribe the required sense and antisense strands, *in vitro* (Figure 2).

- **Expression of shRNA from a Plasmid or Viral Vector**
  Short hairpin RNA (shRNA), expressed from a plasmid or viral vector within the cell, can trigger RNAi. Although vector constructs are more laborious to use than chemically synthesized siRNAs, and vector-encoded siRNA design rules are not as well established, this method does provide a viable alternative when chemically synthesized siRNAs cannot be used. In addition, viral-based vectors permit delivery by infection, which can be beneficial if your cell system is very difficult or impossible to transfect with siRNAs (Figure 2).

Using siRNAs and the RNAi pathway for deliberate 'knockdown' of a gene of interest opened up the basis for reverse genetics. While forward genetics seeks to find the genetic basis of a phenotype or trait, RNAi makes it possible to find phenotypes of a specific gene.

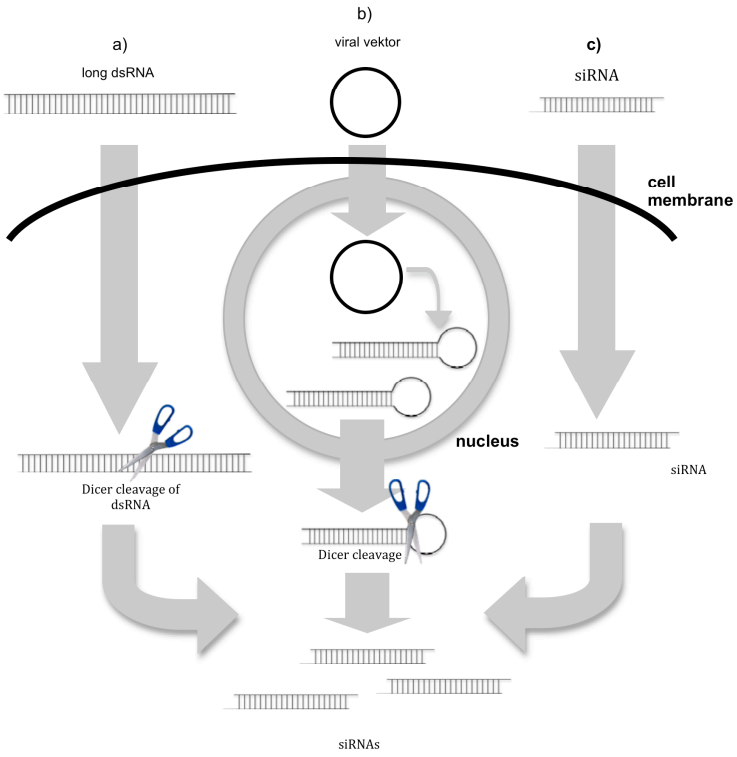

Figure 2: Three ways to trigger the RNAi pathway. a) *in vitro* transcription of long dsRNA followed by cleavage with Dicer, b) expression of shRNA from a plasmid or viral vector c) *in vitro* transcription (IVT).

Another exciting feature of this mechanism is that it forms the basis of first possible direct treatment for viral infections. It can also be used to selectively knock down the expression of specific genes that cause diseases like cancer.

1.2.2.2 microRNAs

A microRNA can be defined as a single stranded small ∼22 nt RNA processed from a long transcript stem-loop like precursor present in the nucleus and cytoplasm. MicroRNAs are endogenously encoded within the genome. Thus far, most of the microRNA genes identified are located in the intergenic regions that are quite distant from previously annotated regions. This suggests that these genes are transcribed as independent transcription units. But there are also some microRNAs that are located in intronic regions and thus they are transcribed together with their host-genes. These intronic microRNAs exhibit similar expression patterns as the mRNAs that encode them. In most cases, microRNAs are found in clusters. It is quite possible that clustered microRNAs may target the same gene resulting in robust regulation, or may target genes belonging to the same biological pathway. A microRNA gene is typically transcribed by RNA polymerase II

into a primary microRNA (pri-miRNA), and includes a 5' cap and a 3' poly(A) tail. Then the dsRNA-specific ribonuclease Drosha digests the pri-miRNA in the nucleus to release hairpin, precursor miRNA (pre-miRNA). That pre-miRNA is approximately 70 nt long with a 1-4 nt 3' overhang, a 25-30 bp stem region and a relatively small loop. Exportin-5 (Exp5) seems to be responsible for the export of the pre-miRNA from the nucleus to the cytoplasm (Figure 3) [16]. To date, more than 3200 microRNAs have been identified

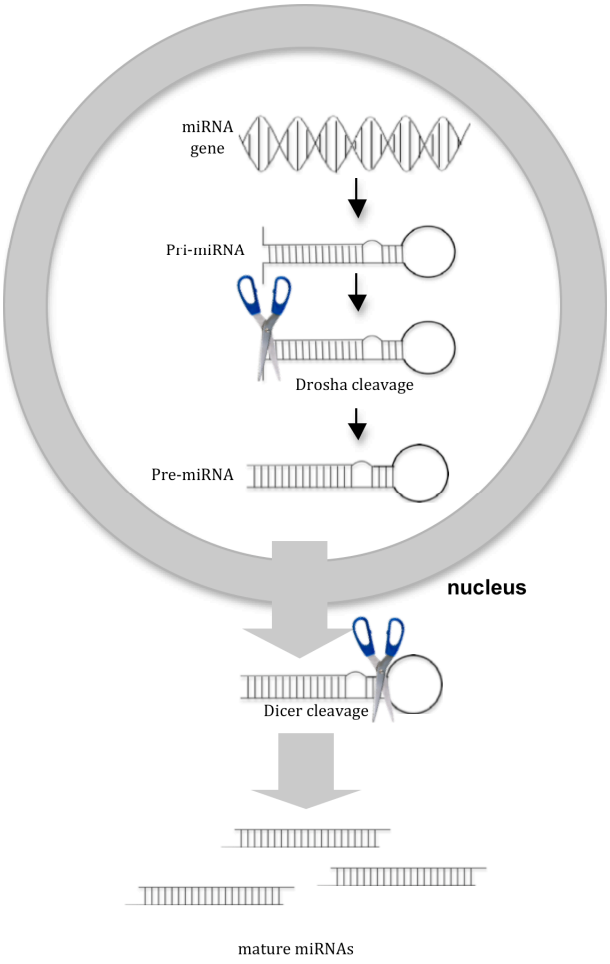

Figure 3: microRNA maturation

in mostly all eukaryotic organisms like Drosophila, mouse, or human and various plant species [17]. Also the understanding of the biogenesis and the diversity of the functions of microRNAs have grown. MicroRNAs function as post-transcriptional repressors of their target genes when bound to specific sites in the 3' UTR of the target mRNA. The level of target mRNA does not change significantly, which indicates that the inhibition occurs at the level of translation, although the mechanism of inhibition remains obscure. Recent studies have revealed a wide variety of microRNA functions, including suppres-

sion of apoptosis, stimulation of cell proliferation, tissue formation during development, left-right asymmetry of neuronal chemoreceptor expression, control of haematopoiesis and developmental regulation (Table 1) [18].

Table 1: List of some known microRNA functions
(Cel, Caenorhabditis elegans; Dme, Drosophila melanogaster; Hsa, Homo sapiens; Mmu, Mus musculus)

microRNA	Function	Known targets	Species
lin-4	Developmental timing	lin-14, lin-28	Cel
let-7	Developmental timing	lin-41, hbl-1	Cel
Lsy-6	Neuronal patterning	cog-1	Cel
miR-273	Neuronal patterning	die-1	Cel
bantam	Cell death, proliferation	hif	Dme
miR-14	Cell death, fat storage	N/A	Dme
miR-181	Haematopoiesis	N/A	Mmu
miR-196	Development	HoxB8, HoxC8, HoxD8, HoxA7	Mmu
miR-143	Adipocyte differentitation	N/A	Hsa
miR-375	Insulin decretion	Myotrophin	Mmu

1.2.3 RNAi pathway

The RNAi pathway is a central part of the gene silencing machinery. It is used by both, siRNAs and microRNAs and starts with the cleavage of long double-stranded RNAs into small dsRNA fragments of 21-23 nt.

1.2.3.1 Preprocessing by the Dicer Containing Complex

Dicer, a class III type RNase III enzyme, plays a significant role in the biogenesis pathway of microRNAs in the Cytoplasm and is also able to cut long double stranded RNAs into siRNAs. The Dicer protein is about 220 kDa and has a PAZ domain in the middle. The PAZ has its name, because it contains the proteins Piwi, Argonaut and Zwille. It can bind to nucleic acids with low-affinity, and interacts probably with the 3' two nucleotide overhang structure of the pre-miRNA [19]. This type of binding might ensure the proper orientation of the pre-miRNA for efficient Dicer processing. Dicer then cleaves the long dsRNA to a $\sim$22 nucleotide dsRNA in an ATP-independent manner [20]. In human, Dicer can process both pre-miRNAs to double stranded mature miRNAs, as well as long dsRNAs to siRNAs.

1.2.3.2 The RISC complex

Among the many proteins that are associated with the RNA Induced Silencing Complex

(RISC), Argonaut (AGO) family proteins are the only small RNA binding proteins. Most of the AGO proteins are about 100 kDa and highly basic and have two unique signature domains: the PAZ domain shared by Piwi, Argonaut and Zwille/Pinhead proteins, and the PIWI domain initially found in the Piwi protein from Drosophila. The PAZ domain is about 130 amino acids long and has been identified in both AGO proteins and Dicer-like enzymes, while the PIWI domain is about 300 amino acids in size and is located at the highly conserved C-terminus of the AGO protein [21]. The PAZ domain has been shown to be directly involved in the interaction between the small RNA and the AGO protein, while the PIWI domain functions as a catalytic domain to interact with the down-stream RNA targets. The PAZ domain can indiscriminately bind any single stranded small RNAs or double stranded RNA duplexes with 3' overhangs, whereas blunt-ended dsRNAs and dsRNAs with 5' overhangs are poorly recognized by the PAZ binding surface [22, 23].

1.2.3.3 RISC Assembly

RISC assembly is a process in which the small RNAs (siRNAs and microRNAs) recognize and interact with the RISC complex. Both Dicer and AGO proteins contain a PAZ domain, but it is not yet known, if the PAZ domains of the Dicer and AGO proteins could play a role in passing siRNA duplexes between the two proteins. Recent studies indicate that RISC assembly is a sequential process that starts from the biogenesis of the small RNAs and requires multiple steps of RNA-protein and protein-protein interactions. According to the nature of small RNA biogenesis, RISC assembly can be divided into two sub-types: the siRNA-induced RISC (siRISC) assembly and the miRNA-induces RISC (miRISC) assembly [24, 25, 26]. The RLC contains some unidentified proteins and two dsRNA binding proteins, DCR2 and R2D2, which form a stable heterodimer (DCR2-R2D2). The key RISC component AGO proteins seem not to be required for the formation of the RISC Loading Complex (RLC). The siRNA or miRNA duplexes must bind to the DCR2-R2D2 heterodimer to initiate the subsequent RISC assembly. RISC maturation starts from the strand separation of the small RNA duplex on RLC and ends with the recruitment of AGO proteins by the single-stranded small RNA. The exact process is still not fully understood [16]. The structure of the small RNA duplex is important in selecting the strand of the duplex that gets assembled into the RISC. The preference of the double stranded small RNA by RISC can be partly explained by the thermodynamic properties of the two strands. The strand with a less thermodynamical stable 5' end is the one that is incorporated into RISC. Unfortunately, this still does not explain the release of the guide-strand from the duplex. It is widely believed that a yet to be identified ATP-dependent helicase, unwinds the duplex to release the guide-strand and to form an active RISC [27]. The passenger strand of the duplex is recognized as a RISC substrate and is

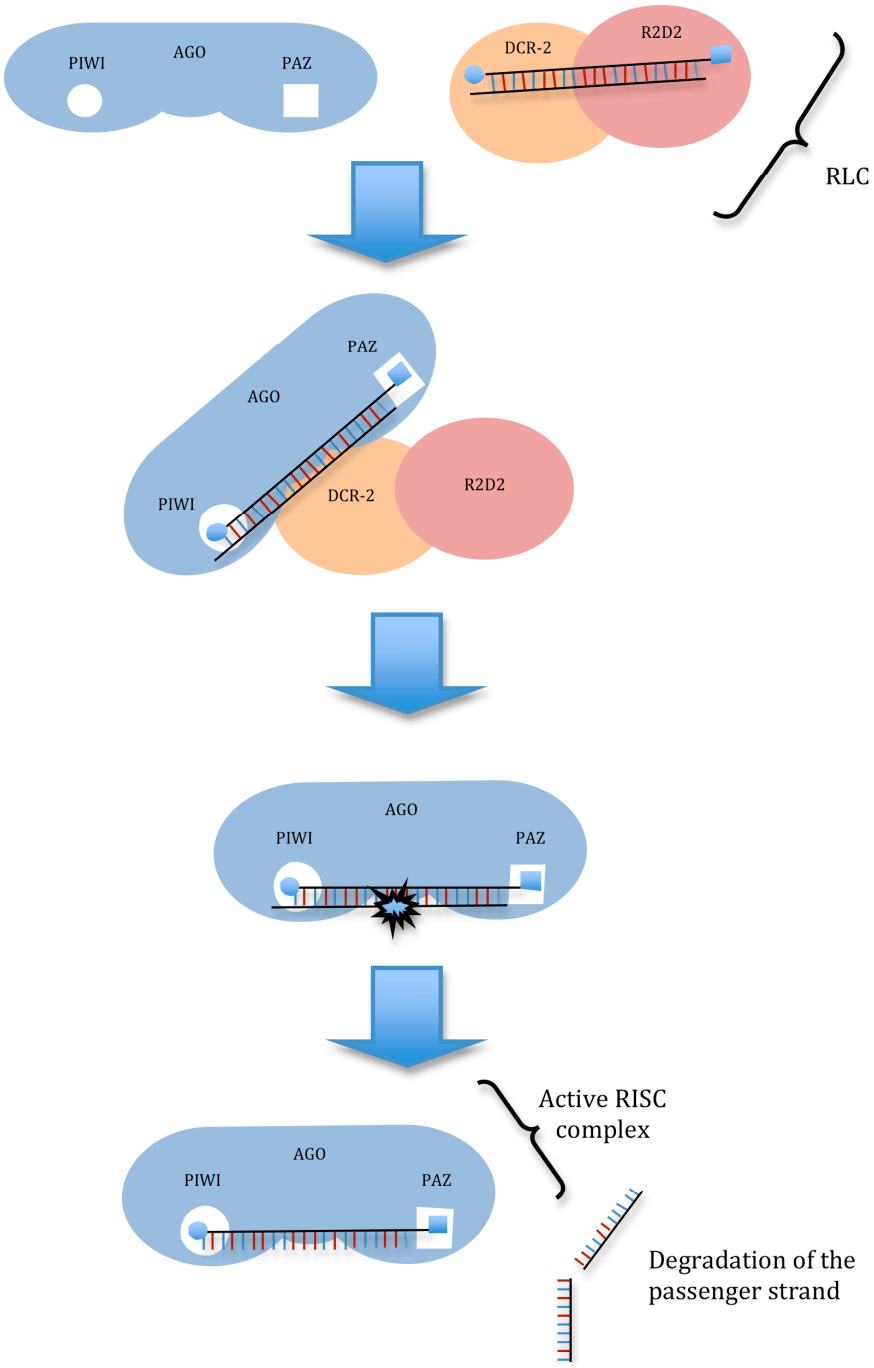

Figure 4: RISC loading

cleaved (Figure 4).

1.2.4 microRNA:target site interaction

The interaction of the small RNA with the target site in the 3'UTR of the regulated gene is still not fully understood. The importance of target mRNA sequence recognition through sites of imperfect complementarity was demonstrated early in the history of the discovery of microRNAs [28, 29, 30]. Unlike in plants, where a high level of complementarity between microRNAs and their targets makes the search relatively straightforward, the presence of a variable level of mismatches in animal microRNA-target pairs makes the identification of targets more difficult. In the beginning of the microRNA-target search, only five microRNA-mRNA pairs were biological validated, offering merely 23 examples of microRNA recognition sites [31, 32]. Using the few available targets, a set of rules

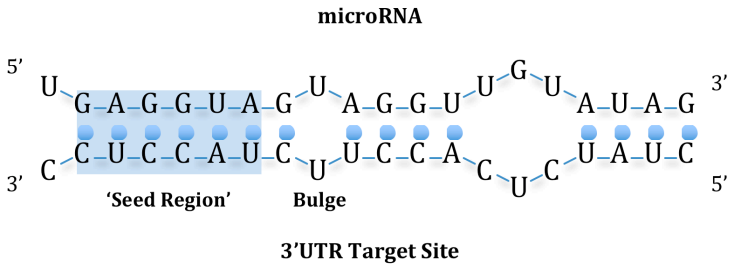

Figure 5: Structure of a microRNA binding site showing the seed region and bulge of a typical binding site. In this case the microRNA is shown 5' to 3' for clarity.

describing the mechanism of target recognition by microRNAs was created. Most notabe among these was the requirement for a perfect match between residues 2 and 8 at the 5' end of the microRNA and the 3' end of the complementary site on the target mRNA (also refereed to as "seed", "core" or "nucleus") [31, 33]. Several observations supported this assumption: (1) Most of the available biologically validated examples shared this feature; (2) residues 2-8 are the most conserved among orthologous microRNAs [31, 34]; (3) sequence motifs previously known to be able to mediate posttranscriptional regulation of gene expression, the K box and Brd box, were complementary to the 5' end of microRNAs [35]. The importance of a perfect match in the "seed" region was directly verified through experiments carried out in tissue culture cells transfected with luciferase reporters [36]. Nonetheless, the emphasis on the importance of the perfect match in the "seed" region might obscure some aspects of the mechanism of target sequence recognition by microRNAs. In the well-characterized example of let-7-mediated repression of lin-41, the complementary "seed" region is not perfect reverse complementary, but

contains a mismatch, that is required for function in one of the two let-7 complementary sites (LCS). The integrity of the region between the two sites is also required for

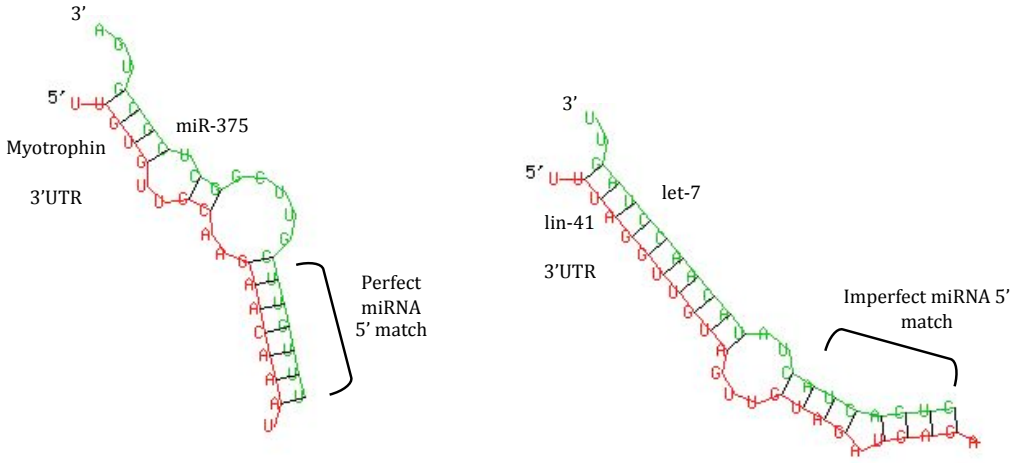

Figure 6: Two classes of miRNA target sites. a) Class one targets have perfect, censecutive Watson-Crick base pairings between the 5' end of the microRNA and the 3'UTR target sites but insignificant complementarity in the remainder of the miRNA sequence. b) Class two targets have an imperfect miRNA 5' match, but significant complementarity of the remainder of the microRNA sequence.

regulation, hinting a possible role for secondary of the lin-41 3'UTR or involvement of other trans factors [37, 38]. Then, a computational search for new targets of let-7 that allowed extended complementary regions in the 3' end of let-7 to compensate for mismatches in the "seed" region yielded 12 new targets that were validated both genetically and by analysis of reporters in trangenic animals [39, 40]. —n 2005 a study rigorously assesed the requirements of sequence complementarity between microRNAs and their targets using transgenic reporters in the context of Drosophila wing imaginal disc [41]. Two classes of microRNA-mRNA interactions were defined (Figure 6): "5' dominant", with an interupted stretch of seven nucleotides corresponding to the 5' end of the microRNA, followed by a variable degree of complementarity in the 3' end (class I); and "3' compensatory" sites, which have weak 5' base pairing and depend on strong compensatory pairing to the 3' end of the microRNA (class II) [41]. Class I bindings are estimated to outnumber class II bindings by aproximately one order of magnitude [41].

1.2.5 Gene regulation

Gene regulation by siRNAs and microRNAs can be divided into two general types

- mRNA cleavage
  If the target site is perfectly complementary to the small RNA, or has only very few

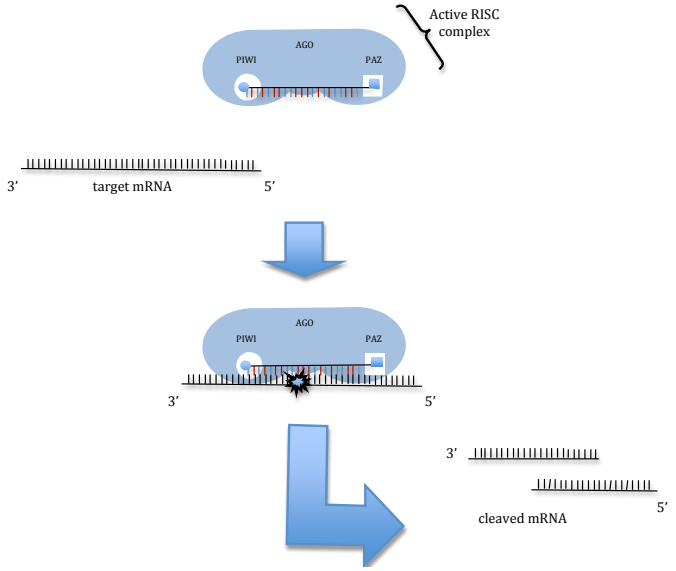

Figure 7: Active RISC complex binds to its target mRNA and cleaves it

(one or two) mismatches, an AGO2-mediated mRNA cleavage is activated, which cuts the mRNA 10 nt upstream of the siRNAs 5'end (Figure 7). The mRNA level in the cell goes down and consequently the protein level. Well designed siRNAs can knock down genes in a range from 70% up to 100% [42].

- translational repression
  If the target site is not perfectly complementary to the small RNA, the small RNA binds to a target sequence in the 3' UTR, but does not cleave the mRNA. It is thought that the some Argonaut proteins of the RISC complex interact with the Cap-structure of the mRNA and thus stops the translation of the mRNA (Figure 8). A decline in the protein level is observed, but the mRNA level stays constant. The gene ist downregulated.

1.2.6 Applications of RNAi

Since the demonstration by Tuschl and colleagues in 2001 [15] that synthetic siRNAs could be used in mammalian cells for gene silencing, siRNA-induced RNAi has become a key strategy for investigating gene function [43, 44, 45, 46, 47, 48, 49, 50, 51, 52]. The rapid adoption of RNAi technology resulted primarily from the ease of use of siRNAs and the strong need for a method to reduce the expression of individual genes in mammalian cells in order to establish a link between gene identity and gene function. Since siRNA-mediated RNAi results in knockdown of gene expression, the observed phenotype depends on how much remaining gene expression is required to cause measurable func-

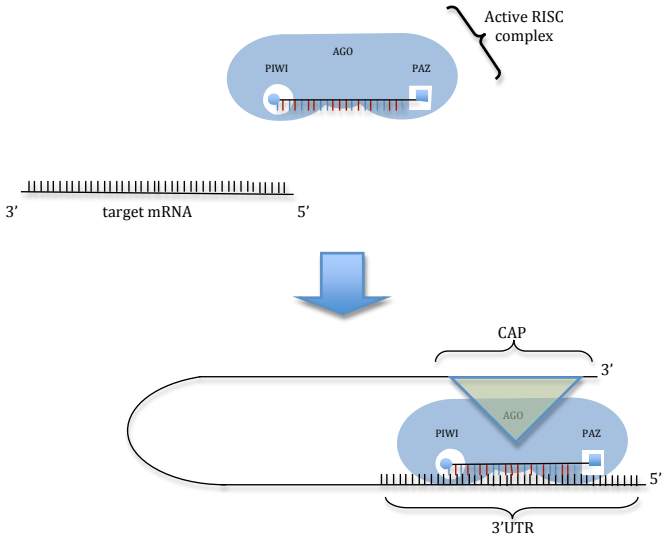

Figure 8: Active RISC complex binds to its target mRNA and represses the translation

tion in the assay used. Other prominent applications for RNAi in mammalian systems are:

- Testing hypotheses about gene function
- Functional screening and target identification
- Target validation
- siRNAs as therapeutics

1.3 siRNA off-target genes

Several labs have demonstrated that introduction of siRNA into cells can induce unintended gene regulation. One explanation for these interferences is that siRNAs bind non-target genes in a microRNA-like way (binding with bulges and mismatches) (see Section 1.2.4). These genes are called off-target genes. It has been shown that artificially designed siRNAs can affect up to hundreds of off-target genes [53, 54]. These off-targets can only be seen, if the siRNA and the off-target gene are expressed simultaniously in the cell. Therfore it is hard to say how many off-targets a given siRNA really has. Most of the companies which sell siRNAs perform very expensive and extensive high-throughput experiments in order to find these off-targets. But since these experiments have to be done for each gene and siRNA separately for each tissue, highly specific siRNAs with low off-target effects are expensive.

2 Basics of siRNA design

Several studies dealing with siRNA design have noted that position specific features (presence or absence of specific nucleotides in certain positions within the siRNA), thermodynamic properties and secondary structures of the target site are important determinants of regulatory efficiency [15, 55, 56, 57, 58]. The following sections briefly summarize the results of a few commonly used design techniques.

2.1 Tuschl Rules

The first technique for designing effective and efficient siRNAs was developed by Elbashir et al. [15]. That group suggested that synthesizing a siRNA duplexes with a 23nt sense strand and a 21nt antisense strand, paired in a manner to have a 2nt 3' overhang, mediates the efficiency of target RNA cleavage. The collected rules for siRNA design are summarized below:

- The targeted region starts 50 to 100 nucleotides downstream of the start codon of a given transcript.
- The 5' UTR is avoided, assuming that UTR-binding proteins and/or translation initiation complexes may interfere with binding of the siRNP or RISC endonuclease complex.
- The 3' UTR is functional for siRNA knowdown.
- All 23nt sequences with the motif AA(N19)TT (N, any nucleotide) are potential siRNA target sites. If no suitable sequences are found, the search is extended using the motif NA(N21).
- The G/C content of target sites has to range between 30% and 70%.
- The target site's secondary structure has no effect on silence efficiency.
- A blast search against an EST library is recommended, to ensure that only the gene of interest is targeted.
- siRNA target sequences with SNPs should be avoided.

2.2 Reynolds Rules

Reynolds et al. [55] analyzed a set of 180 siRNAs. They divided the siRNAs into different groups, based on their repression efficiency. They asked, if siRNAs with high functionality have any similarities in their sequence. As a result they provided six rules of how to design highly potent siRNAs:

- G/C content has to be between 30% and 52%
- Presence of nucleotide A at position 3 and 19
- Presence of U at position 10
- Absence of G or C at position 19
- Absence of G at position 13
- Presence of A/U in positions 15 through 19

Their algorithm assigns a score based on the number of rules satisfied. Each siRNA exceeding a specific threshold is predicted to be functional.

2.3 Ui-Tei Rules

Ui-Tei et al. [56] analyzed 62 targets in mammalian and Drosophila cells. They came up with four features siRNA should simultaneously satisfy to cause efficient silencing:

- A/U at the first nucleotide of the 5' end of the antisense strand
- G/C at the first nucleotide of the 5' end of the sense strand
- At least five A/U nt in the 5' terminals first-third of the antisense strand
- Absence of any 'GC' stretch of more than 9 nt in length

These rules were found to be applicable to mammalian cells but did not apply to Drosophila cells.

2.4 Stockholm Rules

This prediction algorithm by Chalk et al. [57] incorporates the thermodynamic properties of the siRNA. The following rules are called the "Stockholm rules".

- Total hairpin energy $<$ 1 kcal/mol

- 5’ end binding energy < 9 kcal/mol in the antisense strand
- 5’ end binding energy in the range 5-9 kcal/mol exclusive in the sense strand
- G/C Content between 36% and 53%
- Middle area (7-12) binding energy < 13 kcal/mol
- Energy difference between antisense and sense 5’ energies < 0 kcal/mol
- Energy difference between antisense and sense 5’ energies within -1 kcal/mol and 0 kcal/mol

A scoring method that adds one for each rule satisfied. Effective siRNA have to exceed a threshold of six.

2.5 Amarzguioui Rules

Amarzguioui et al. [58] created the following siRNA design rules, based on their study of 46 siRNAs with a knockdown rate of more than 70%:

- asymmetry in the stability of the duplex ends (A/U differential of the three terminal base pairs at either end of the duplex)
- Presence of G/C at position 1
- Presence of A at position 6
- Absence of U at position 6
- Absence of U at position 1
- Absence of G at position 19
- Presence of A/U at position 19

Each rule either adds a point, if satisfied and subtracts one, if failed, respectively. siRNAs with a score of more than 3 are considered to be efficient.

2.6 Review of the design rules

Since most of the design rules of the five groups mentioned above coherently recommend the usage of a low G/C content, I assume this to be the most important feature. All the several rules dealing with repeats, GC stretches and target regions that have to be considered. The rest of the proposed rules mostly relies on position specific features. Since the groups who developed these rules performed only few *in vivo* or *in vitro* experiments (e.g. Reynolds: 180; Ui-Tei: 62) to measure the efficiencies of the designed siRNAs, the obtained rules have most likely a strong bias. Indeed, the overlap of the position specific features is extremely small and thus I decided not to use them in my method. Nevertheless, I integrated features mimicing these posititon specific rules as optional, user definable filters.

2.7 siRNA design tools

Figure 9 shows a list of some popular siRNA design tools and highlits the selections user can make, the basic design methods and other options, worth to be mentioned.

Tools		Seletions							Search Methods							Other	
Company	Name	Identifier / Sequence	Species	Region	GC-content	Off-Target search	pattern	other	Seed count	SNP avoidance	Internal folding	motif avoidance	repeats	microRNA seed avoidance	ranking	comment	website
Dharmacon	siDESIGN Center	Identifier, Sequence	-	Selection between 5'UTR, ORF and 3'UTR (default: ORF)	Min, Max (default: 30, 64)	Blast	-	-	x	x	x	x	5 or more of a single base in a row; GC/stretches; A/U stretches	x	x	-	http://www.dharmacon.com/DesignCenter
Ambion	siRNA Target Finder	Sequence	-	-	Max (default: All)	-	End sequence with TT or UU	-	-	-	-	-	-	-	-	Blast search possible, but only by pressing the Blast button for each sequence separately	http://www.ambion.com/techlib/misc/siRNA_finder.html
Novartis	BIOPREDsi	Identifier, Sequence	human, mouse, rat	-	-	-	-	number of predicted siRNA; negative control	-	-	-	-	-	-	x	time consuming	http://www.biopredsi.org
Invitrogen	Block-iT RNAi Designer	Identifier, Sequence	-	Selection between 5'UTR, ORF and 3'UTR (default: ORF)	Min, Max (default: 35, 55)	Blast	default, Tuschl	-	-	-	-	-	-	-	x	specificity check is unspecified	https://rnaidesigner.invitrogen.com/rnaiexpress/setOption.do?designOption=sirna
Whitehead Institute	siRNA Selection Server	Identifier, Sequence	-	-	Min, Max (default: 30, 70)	Blast	custom, Tuschl; TT end	Avoidance of 4 or more of a single base (G, A, T) in a row; GC/stretches;	-	-	-	-	-	-	x	Requires pre-registration	http://jura.wi.mit.edu/bioc/siRNAext/
University of Tokyo	siDirect	Identifier, Sequence	human, mouse, rat, chicken, dog	Selection of exact positions possible (default: ORF)	Min, Max (default: All)	Blast	custom, Ui-Tei, Reynolds	-	-	-	-	-	x	-	-	No Score/Rank	http://genomics.jp/sidirect

Figure 9: A List of siRNA Design Tools

3 Applied bioinformatics tools

There are several bioinformatics tools available for predicting secondary structure and predicting microRNA target sites. The following programs were used for siRNA design, off-target prediction and analysis, performed for this thesis.

3.1 Vienna RNA secondary structure package

A single RNA can potentially form many different structures. The physical approach to this problem is to look for the most stable structure of the molecule. Thermodynamic measurements are available that give a good idea of the energy and entropy change of formation of helices and loops. Thus, a free energy can be assigned to any structure and the minimum free energy (MFE) represents the most stable structure of these. But that approach gives no information, if the real molecule folds to that MFE structure, or is trapped in a meta-stable state. Nevertheless, the Vienna RNA secondary structure package makes use of the MFE structure approach, since the RNA structure model is known to be sufficiently realistic to be able to predict structures of real biological sequences [59]. In principle, the MFE can be obtained by considering every single base pairing pattern and calculating the free energy for each binding, by using a set of experimentally determined energy rules. But the number of potential structures increases exponentially with the length of the molecule, N. Using dynamic programming, that calculation can be done in a time $O(N^3)$. This method works by writing a recursion relation that breaks down the structure of a large sequence into a sum of smaller parts. The Vienna RNA Secondary Structure package [60] is one of several tools that uses this method to predict structures with MFE. The recursion relations used in the programs of the Vienna RNA package are considerably more complicated because they have to account for penalties of formation of loops of different types and there are many special cases to be considered. Nevertheless the algorithms remain efficient, and still scale as $O(N^3)$ for the full energy parameters. In this work, I used RNAfold, RNAduplex and RNAcofold, which are part of the Vienna RNA secondary structure package.

3.1.1 RNAfold

RNAfold [60] is a secondary structure prediction algorithm for RNA sequences. The tool takes a single RNA sequences as input and returns the MFE structure of the input sequence in bracket notation with its Gibbs free energy as output (Figure 10). In addition, a plot of the predicted structure of the RNA sequence in a postscript-formatted file named ”rna.ps” is given (Figure 11).

```
[langenberger@gepard ~]$ RNAfold

Input string (upper or lower case); @ to quit
....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,....8
CCCGGGCUGAGGUAGGAGGUUGUAUAGUUGAGGAGGACACCCAAGGAGAUCACUAUACGGCCUCCUAGCUUUCCCCAGG
length = 79
CCCGGGCUGAGGUAGGAGGUUGUAUAGUUGAGGAGGACACCCAAGGAGAUCACUAUACGGCCUCCUAGCUUUCCCCAGG
(((.(((..(((.(((((((((((((((((.(((...(((....)))........)))))))))))))))))).)))..)))).)))
 minimum free energy = -37.80 kcal/mol
```

Figure 10: RNAfold output for the has-let-7e microRNA precursor sequence.

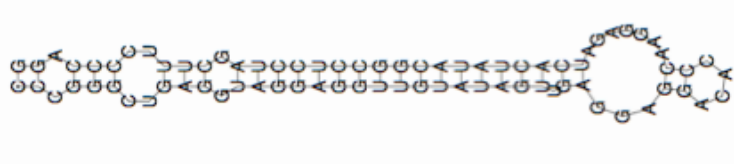

Figure 11: RNAfold output file (rna.ps), illustrating the predicted secondary structure graphically.

3.1.2 RNAduplex

RNAduplex [60] predicts hybrids formed between a short and a long RNA sequence. The tool takes these two RNA sequences as input and returns optimal and, if required, also the suboptimal hybrid structures. The results are presented as one structure per line (Figure 12). Each line consist of a) the structure in dot bracket format with a ”&” separating the two strands, b) the range of the hybridized structure for each sequence in the format ”from,to : from,to” and c) the energy of the duplex structure in kcal/mol. To create suboptimal structures, the –e option has to be used. This option allows the user to define a range of MFE of the optimum he wants to be returned.

```
[langenberger@gepard ~]$ RNAduplex

Input two sequences (one line each); @ to quit
....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,....8
AAAGAGAACAAGAAAAGUACAACUACAACAUUGCUUAAGUCCUACCUCAAUUGCUUAAGUCCUACCUCA
UGAGGUAGGAGGUUGUAUAGUU
lengths = 69,22
.(((((((((..............((((((((((.&))))))))))))))))))).  16,50  :   1,19  (-23.76)
```

Figure 12: RNAduplex output for has-let-7 and the lin-41 homolog target site in human.

3.1.3 RNAcofold

RNAcofold [60] works much like RNAfold, but it considers these parts of the two sequences that don't hybridize, as they can form secondary structures on their own. The tool takes two RNA sequences, concatenated by using the '&' character as separator, as input and returns the MFE hybrid structure. By using the –C option, the user has the option to fix the exact binding structure of the two RNAs by entering its structure in bracket notation (Figure 13).

```
[langenberger@gepard ~]$ RNAcofold -C
Input constraints using the following notation:
| : paired with another base
. : no constraint at all
x : base must not pair
< : base i is paired with a base j<i
> : base i is paired with a base j>i
matching brackets ( ): base i pairs base j

Input sequence(s); @ to quit
Use '&' as spearator between 2 sequences that shall form a complex.
....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,....8
AAAGAGAACAAGAAAAGUACAACUACAACAUUGCUUAAGUCCUACCUCAAUUGCUUAAGUCCUACCUCA&UGAGGUAGGAGGUUGUAUAGUU
.................(((((((((.............(((((((((.....................&)))))))))))))))))....
length1 = 69
length2 = 22
AAAGAGAACAAGAAAAGUACAACUACAACAUUGCUUAAGUCCUACCUCAAUUGCUUAAGUCCUACCUCA&UGAGGUAGGAGGUUGUAUAGUU
.................(((((((((.............(((((((((.....................&)))))))))))))))))....
 minimum free energy = -23.76 kcal/mol
```

Figure 13: RNAcofold output for has-let-7 and the lin-41 homolog target site in human.

3.1.4 Kerror

Kerror [61] is a tool that efficiently solves the problem of finding a pattern P (length m) in a text T (length n) by allowing k errors (mismatches, insertions or deletions). The main underlying of Kerror is that by division of P in $k+1$ substrings, there is at least one substring of P with a perfect match in T. For example, when $k = 5$, P will be divided into 6 segments of equal length r.

$$r = \frac{m}{k+1} \tag{1}$$

Now, at least one of these segments remains error free (Figure 14). This can easily be tested by randomly throwing 5 errors into pattern P. By making use of that trick, the algorithm of Kerror starts searching for the first fragment of P, using a local alignment. If a match is found, the matched site is extended to the size of P. Applied to our example, if the first segment of P matches in T, the algorithm extracts the match and the 5 times r nucleotides to its right plus additional k bases on both borders. If the second subsequence

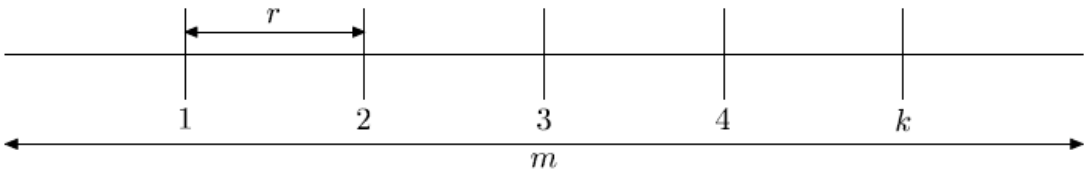

Figure 14: Divide of P (length m) in k fragments of length r (as seen in [62]).

matches, r bases to the left and 4 times r nt to the right plus the additional k bases at both borders, and so on. Subsequently, a global alignment using the elongated sequence and P is calculated to get the score and the exact information about errors, which occur. Since this algorithm runs in time $O(m^2 * n)$, the number of exact matches of the length r have to be small, for a fast search result.

3.2 Blast

The Basic Local Alignment Search Tool (BLAST) [63] searches in a database for the best hits, based on a local alignment score, using a heuristic approach that approximates the Smith-Waterman algorithm. Although the used heuristic approach is less accurate than Smith-Waterman, it is over 50 times faster. To create such a database as needed by BLAST, the sequences have to be preprocessed formatdb. Due to the algorithm's design, search sequences must not be shorter than 20 nt in length.

3.3 microRNA off-target prediction tools

So far, a number of methods have been developed to predict miRNA targets [64]. They can be devided into two main categories, those primarily searching for sequence complementarity and those searching for favourable microRNA:target duplex thermodynamics. Overall, the approaches are often quite similar [64]. Most methods require conservation of binding sites and the presence of multiple sites is used to rank certain sites higher. Moreover, some approaches even demand strict complementarity between the 'seed-region' [65, 31] of the microRNA (nt 2-7) and any predicted target to further improve the signal to noise ratio, or to reduce the amount of false positives.

3.3.1 PicTar

This approach was first presented by Krek et al. [66]. Multiple sequence alignments of all available 3'UTRs are scanned for those with perfectly conserved seed matches to miRNAs and filtered according to their predicted thermodynamic stability. The approach

then uses a Hidden Markov Model (HMM) to score each predicted target, based on the combinatorial effects of multiple microRNA targeting a 3'UTR sequence. Of the targets predicted in mouse, a total of 7 out of 13 were experimentally validated.

3.3.2 miRanda

This algorithm is implemented in a standalone package for finding microRNA binding sites in genomic sequence [67]. The approach is similar to that of Stark et al. [32], as it first identifies potential binding sites with a high degree of complementarity to a microRNA by using dynamic programming. To account for the observation of perfect seed matches, the scoring matrix is designed to favor complementary nucleotides binding at the 5' end of the microRNA over those binding at the 3' end. Potential sites identified in this manner are also evaluated thermodynamically using customized version of the Vienna RNA folding package [68]. There is also a precompiled online version of miRanda mircoRNA target sites available, including target site conservation for better specificity.

3.3.3 RNAhybrid

The methods described so far use shuffling analysis after target scanning in order to estimate global false-positive rates. The RNAhybrid method [69] developed by Rehmsmeier et al., was the first method to incorporate a robust statistical model similar to those used for large-scale sequence comparison. The method is also based on a dynamic programming algorithm to identify regions in a 3'UTR with the potential to form a thermodynamically favourable duplex with a specific microRNA. This is more accurate than forcing potential duplex sequences through Mfold [70] or Vienna [68], as these tools are designed for single sequence folding and not duplexes, hence the energies produced are skewed by artificially added linker sequences. Each microRNA is initially profiled by scanning it against a set of shuffled 3'UTR sequences with maintained dinucleotide frequencies and taking the maximum free energy for that microRNA in every UTR. These energies are length normalised for both UTR and microRNA length. Random energies derived in this manner should exhibit an Extreme Value Distibution (EVD). Using the derived distribution from shuffled sequences, the parameters of the EVD that best describe the data for a given microRNA is empirically calculated. These parameters can now be used to directly calculate a P-Value for any hit for that microRNA to any 3'UTR. Hence, at the scanning stage microRNAs are scanned against a database of real 3'UTRs and each hit is compared to the expected distribution and assigned a P-Vlaue. The method has been used to predict known targets successfully in *D. melanogaster* and appears to have a very low false-positive rate [71]. Unlike ranking and filtering approaches, the assignment of

a P-value (or E-value) to a predicted target is attractive as it allows individual sites to be evaluated and compared to other sites directly.

4 Evaluation of different design strategies

The opinions of how to design highly efficient and specific siRNAs differ strongly (see Section 2). Each of the above mentioned group assumes that their derived rules achieve the best regulation efficiency. Since the focus of this work is not only on high siRNA efficiency but also on high specificity, i.e. producing little off-target effects, I evaluated design strategies in order to find the best adjustment.

4.1 siRNA length

The available siRNA design tools (see Section 2.7) use three different lengths for their siRNA design,

- 19 nt: siDesign Center (Dharmacon), Block-iT (Invitrogen)
- 21 nt: BIOPREDsi (Novartis), siRNA Target Finder (Ambion)
- 23 nt: siRNA Selection Server (Whitehead Institiute), siDirect (University of Tokyo)

I therefore had to decide which default length to use for my siRNA design. There are also two *in silico* experiments, which conclude that target specificity and low probability of off-target effects are optimally balanced for siRNAs of length 21 [72, 73]. Furthermore, in gene silencing studies it hast been verified that siRNAs ranging from 19 to 29 nucleotides in length are functional.
Considering all this information, I decided to design a test to set the amount of off-target effects in relation to the length of the siRNA. Off-target effects due to siRNA induced mRNA cleavage occur less often for longer siRNAs, since perfect target site complementarity is required. Since these off-target effects are very rare anyway, I concentrated on microRNA-like off-target effects. To predict these off-targets, I used miRanda (see 3.3.2). miRanda is a microRNA target prediction tool, that can be used without the conservation-constraint, which is non-applicable for siRNAs. Thus, it is perfectly suited for this test case. First I randomly created siRNA ranging from 19nt up to 29 nt in length. Second, for each length I designed 100 different siRNAs and used miRanda to predict the amount of off-targets for each. Third I calculated the average over all predictions of each length and plotted the result into a histogram (Figure 15). Figure 15 shows that the probability of off-targets increases with the length of the siRNAs.
At first glance, designing siRNAs of length 19 seems to be the best choice with the view to minimize off-target effects. Nevertheless, a length of 19 is the lower limit for siRNAs to be functional and due to their small size, their attraction to their target site is less strong than these of longer siRNAs. Considering that the binding of the siRNA:target site

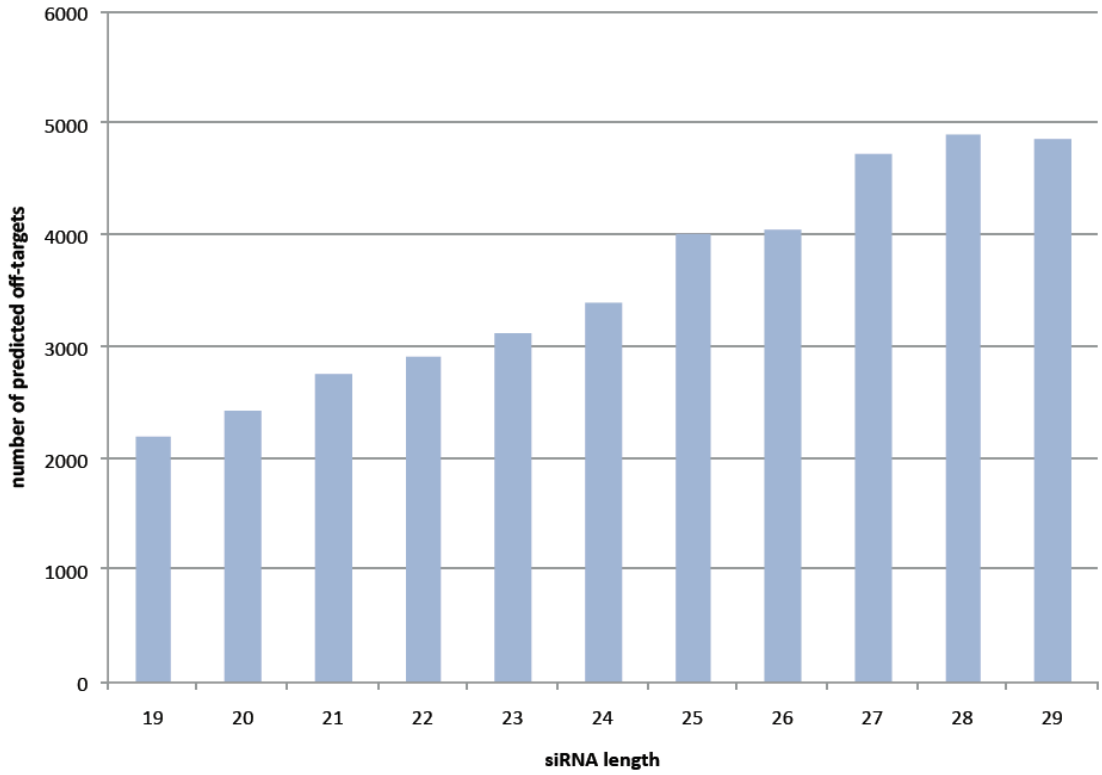

Figure 15: The frequency of predicted target sites against the siRNA length

hybrid is most likely just strong enough for valid binding, but has no chance to absorb innercellular influences without loosing the connection, I decided not to use siRNAs of this short length. Even though siRNAs of length 23 have a high target specificity and a strong binding, they are much more susceptible to target unwanted genes (see Figure 15) and thus not appropriate for this approach. Alltogether, by considering the higher target specificity of longer siRNAs, and the smaller amount of off-target-effects of smaller siRNAs, candidates of length 21 seem to yield the best trade-off. Using a length of 21 also coincides with the two publications, mentioned above.

4.2 Target region

The RNAi mechanism regulates gene expression post-transcriptionally, thus the entire precursor mRNA (pre-mRNA) is basically available for siRNA targeting. Introns and splicing junctions are theoretically possible for siRNA interaction, but they haven't yet been thoroughly investigated as suitable sites. Therefore, they are not used in this approach. The remaining exonic sequences are biologically separated in three relevant sections, the 3'UTR, the CDS and the 5'UTR. Event though the higher conservation und the

consequently lower amount of polymorphism, argues for using only the CDS for siRNA target design, I decided to use also the 3'UTR, by default. The latter represents a well-known region for post-transcriptional regulation performed by microRNAs and there are no objective reasons arguing against it [74], besides the higher SNP rate. To account for this, a filter to avoid SNPs was integrated, in order to delete candidates targeting documented polymorphisms. Since the 5'UTR has never been shown to be available for any post-transcriptional silencing [75], it is completely omitted in this approach.

4.3 G/C Content

Most of the available siRNA design tools recommend a G/C content between ~30% to ~60% (see 2.7). In experimental studies it has been observed that siRNAs with lower G/C content are more efficient than those with a high GC content [42]. One explanation for that behavior may be the strength of the binding between the sense and the antisense strand of the exogenous siRNAs double-strand. Assuming the RISC complex has to open the binding of the perfect complementary double-stranded structure in order to load the siRNA, one might suspect, that there is correlation between the G/C-Content and the potential for the RISC complex to unwind that structure. In case the binding cannot be opened, the siRNA cannot be loaded into the RISC complex and is thus not functional (Figure 16). Hence, my assumption is that there is a specific energy threshold for the double-stranded

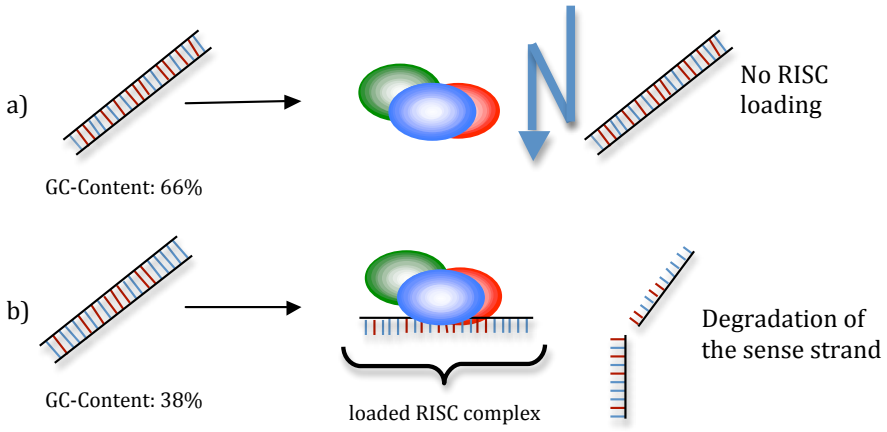

Figure 16: a) Too tight binding of the siRNA sense and antisense duplex makes it impossible for the RISC complex to open the structure for a correct RISC loading. b) For a good RISC loading, a binding with less energy is needed. Blue binding positions stand for an A-T binding and red ones for a G-C binding.

siRNA, that must not be exceeded for proper RISC loading. To test this hypothesis *in*

silico, I used the double-stranded binding of the endogenously encoded microRNA precursors. These precursors form hairpin loops which the protein Dicer subsequently cuts into a double stranded mature form. Since the microRNAs were experimentally verified to reduce gene expression, they have to be correctly loaded into the RISC complex, i.e. the double stranded mature form must have been unwound. I therefore calculated the binding

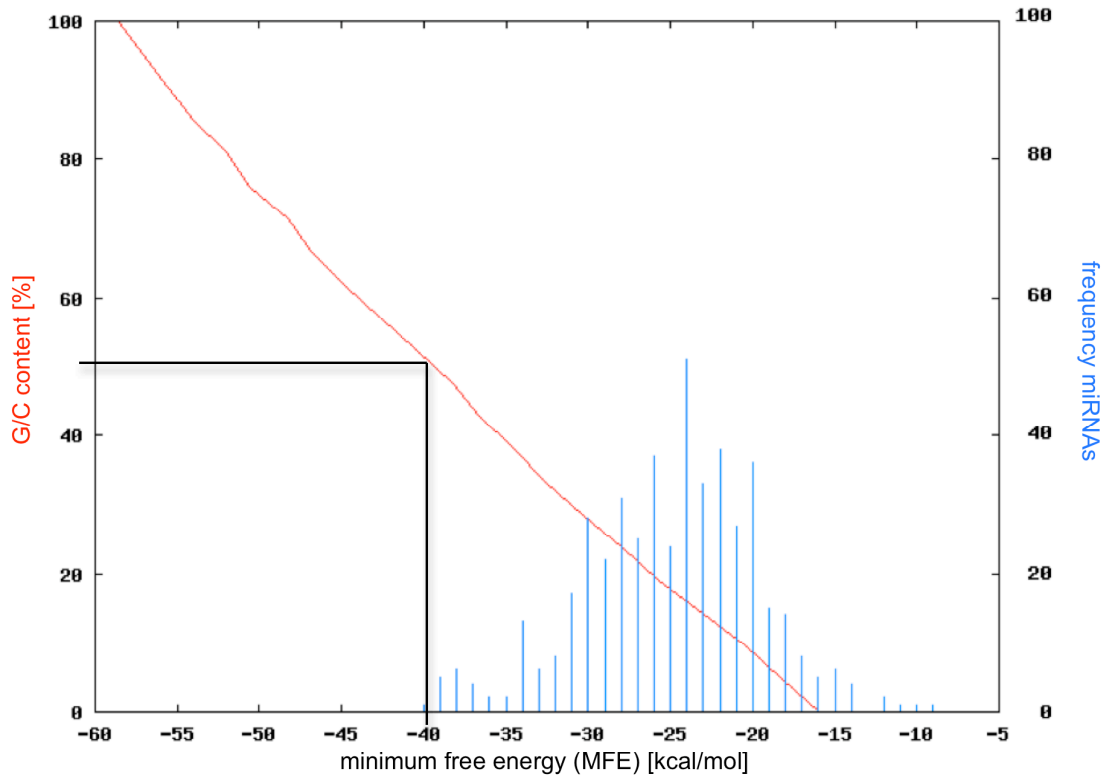

Figure 17: G/C content against the MFE

energies of all human mature microRNA double-strands using RNAduplex. These values reflect the binding energies that can be unwound *in vivo* and thus should not be exceeded by double-stranded siRNAs. Unfortunately, there is no direct correlation between the G/C-Content of the microRNAs and that of the siRNAs, since in microRNAs bulges are allowed and for the siRNA double-strands, perfect complementary is required. So, I created a set of randomly designed siRNAs with a steadyly increasing G/C content, starting with a sequence composed of only A's and T's and mutating step by step one A or T to a G or C until the G/C-Content achieves 100%. This process is repeated 100 times in order to create a representative set for each G/C-Content. Plotting the mean of the resulting energies against the G/C-Content of the sequence, a falling straight line is observed (Figure

17). Higher G/C-Content leads to less free energy and therefore to a stronger binding, wheras lower G/C content leads to a weak binding. The energies of the microRNAs were also added to the plot, to demonstrate, where the functional endogenous binding energies are positioned. The falling line represents the G/C content dependent thermodynamic behavior of all 21 nucleotide long siRNAs and the bars illustrate the energies of microRNAs (Figure 17). I used the energy of the microRNA with the strongest binding as a threshold for a doublestranded small RNA to be functional. By getting the corresponding G/C content for the siRNAs, I decided to use by default a maximum G/C content of 50%.
Another motivation to use smaller G/C-Contents in siRNA design is the decreasing amount of off-targets. To illustrate this, I created a set of randomly designed siRNAs with increasing G/C-Content using the same approach as explained above. I used RNAduplex to search for all possible off-target bindings in a set of 1000 randomly chosen transcript 3'UTRs. All targets with less than ten bindings were deleted, since such bindings are most likely not strong enough for microRNA target binding. I repeated this process ten times to get robust results. Then I plotted the number of hits against the G/C-Content (Figure 18).

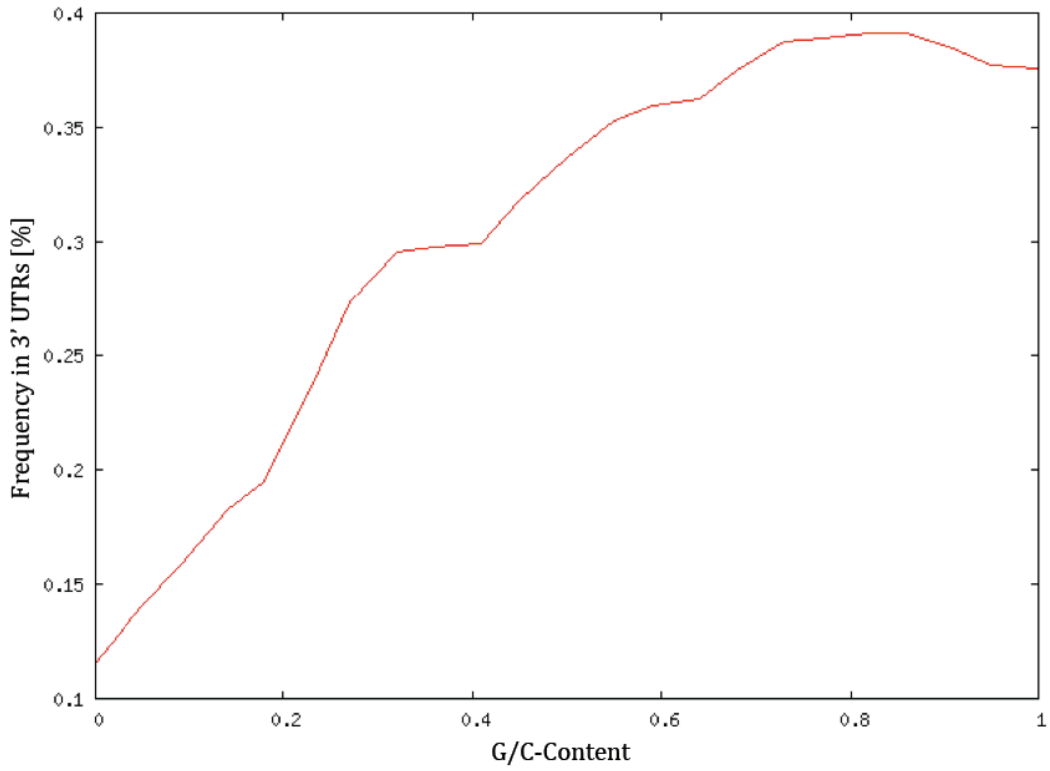

Figure 18: G/C content against the predicted off-targets

In order to minimize off-target effects, siRNAs with smaller G/C-Content should be

ranked higher, since they tend to have less off-target effects.

4.4 Local Accessibility

Recently several labs have demonstrated that the secondary structure of the target sequence and its direct neighborhood is important for a target site to be functional [76, 77]. It is assumed that the target site has to be accessible to the loaded RISC complex for binding. If the local structure of the target site is too stable, it is not visible to RISC and hence cannot function as a target site. Here, I use the term 'target site' to refer to a microRNA-like off-target binding of the siRNA. Since the main target site has to be perfectly complementary to the siRNA, the local accessibility is less important. Two papers have recently been published on microRNA accessibility. The first showed exper-

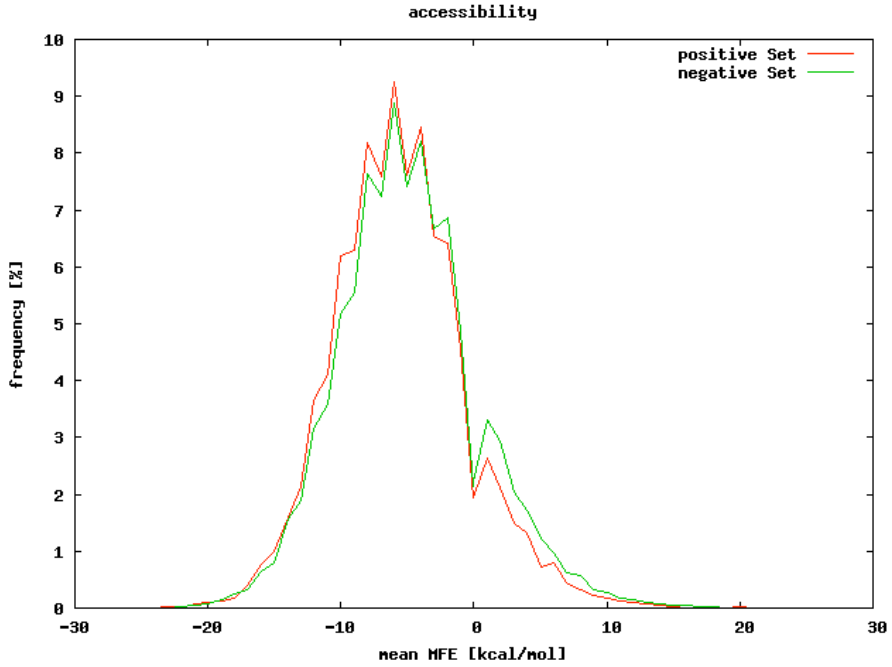

Figure 19: Local Accessibility of the 70 nt flanking the target site.

imentally that the secondary structures of the 70 nucleotides flanking the target site are essential for microRNA binding [77]. The authors calculated the average folding energy of all sequences of length 70 occuring in all 3' UTRs for a species. Then they measured that validated targets for the gene lsy-6 in *C. elegans* have flanking regions of length 70, that have significantly less stable folding energies than the pre-computed *C. elegans* mean energy. Nonvalidated targets were shown to lie in regions with folding energies that are closer to the mean energy. I tried to reproduce their findings *in silico*. To do so, I calculated the average free folding energy for all sequences of length 70 in human 3'UTRs

using RNAfold. Then I defined a conserved and a non-conserved test-set. To create the conserved set, I used 58 microRNAs that are conserved in human, chimpanzee, mouse, rat and dog. As a conserved target site I used a 6-mer in a 3'UTR with exact Watson-Crick complementarity to bases 2-7 from the 5' end of the mature microRNAs defined above. I predicted microRNA targets by searching for these 6-mers that are exactly conserved in the same species as the microRNAs, requiring an additional match to either base 1 or base 8 of the mature microRNA in human. That method is similar to the core PicTar algorithm and is described in [66]. As negative set I used all 6-mers with a conservation of less than 50%. Assuming, that the negative targets should accumulate at a region with lower free energy than the average and the positive at a region with higher free energies, I plotted the shave of the conserved and the non-conserved targets against the mean energies of the upstream and downstream foldings. A small shift between the positive and the negative set can be seen, but it is far away from any hoped separation (Figure 19).

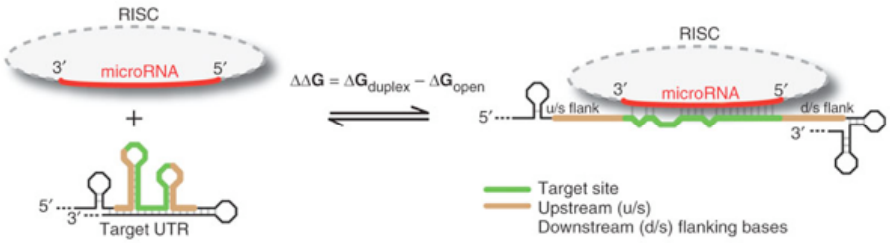

Figure 20: microRNA-target interaction model from [76].

In the other publication the role of site accessibility in microRNA target recognition was analyzed [76]. The authors showed that mutations, which reduce the site accessibility by increasing the self-binding structure of the target site, also reduce microRNA-mediated translational repression. They built a model that explains variability in target strength due to differences in accessibility and showed that this model works for their three experimentally tested binding sites of grim (miR-2), hid (bantam) and rpr (miR-2) in *D. melanogaster*. They introduced an interaction energy $\Delta\Delta G$ for microRNA:target site interaction that is computed as the free energy ΔG_{open} gained by transitioning from the state in which microRNA and target are unbound to the state ΔG_{duplex} in which the microRNA binds its target. The region that needs to be unpaired for the RISC complex to bind includes the target site plus flanking nucleotides (Figure 20). In their experiments, they used 70 nucleotides for that flanking region on each side. To reproduce their approach *in silico*, I utilized the same conserved and non-conserved set as mentioned above. Instead of folding only the flanking regions, I calculated two energy values using RNAcofold. The first energy was the self-folding energy of the target site plus 70 nucleotides upstream and

downstream. The second one was the energy of the hybrid, the target site plus flanking and the according microRNA. Plotting the shave of the conserved and non-conserved targets against the difference of the self-binding energy and the hybrid energy, the observed shift is again not significant (Figure 21).

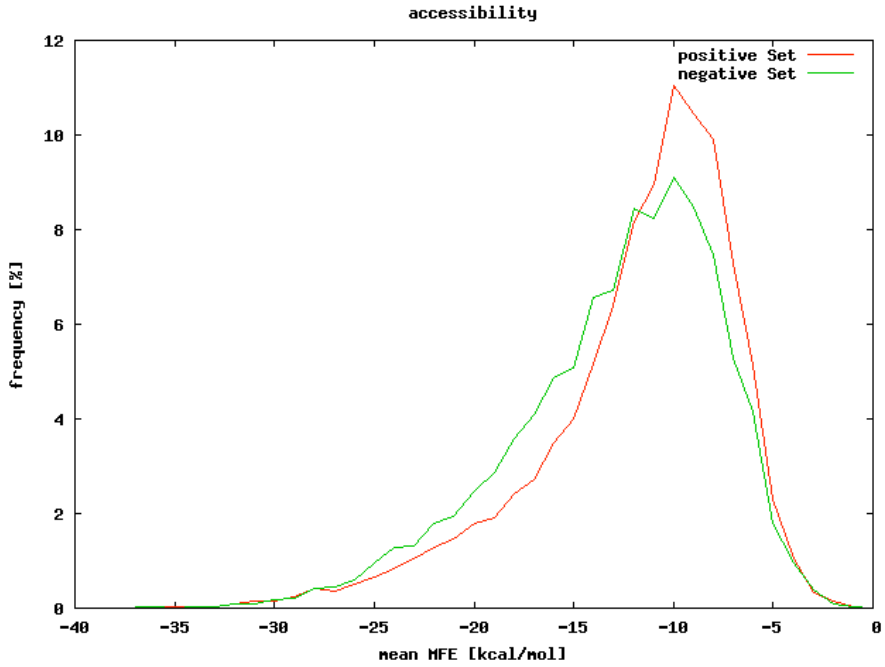

Figure 21: Local Accessibility of the target site and the 70 nt flanking.

Since both methods don't lead to a significant separation of the conserved and non-conserved set, I decided to design my own apporach, similar to that from [76]. I realized, that the length of the flanking region significantly changes the difference between the duplex and the self-folding energies. The self folding secondary structure and therefore the energy increases with the sequence length, while the energy of the duplex remains constant. I decided to vary the flanking regions in length, to figure out, if there is a specific flanking length for my test-set to separate more accurate. For that test, I increased the flankings of the target sites in steps of five. I calculated the self-folding energy using RNAfold and the hybrid energy, using RNAduplex. A hybrid energy greater than the self-folding energy of the target site represents a accessible target site. By counting the conserved target sites that were found as accessible and the non-conserved target sites found as not-accessible, a sensitivity value for each flanking region could be calculated. These results were graphed by a heatmap (Figure 22). With the results showing in Figure 22, I decided to use 15 nt upstream and 10 nt downstram flanking.

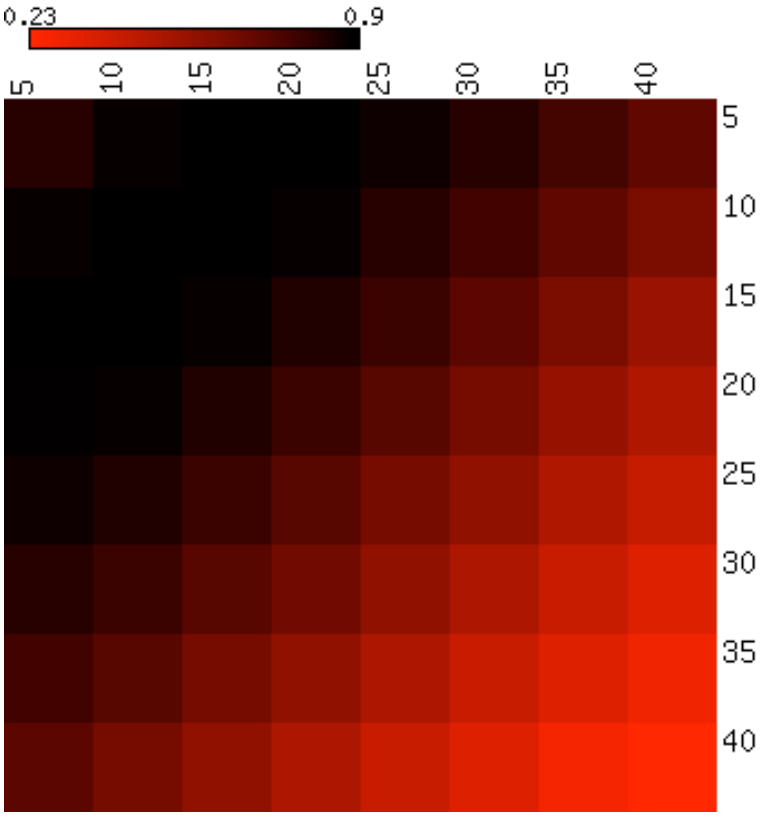

Figure 22: HeatMap of the Local Accessibility for different upstream and downstream flankings.

4.5 Perfect seed match

In 2006, A. Birmingham et al. shed light on off-target effects caused by one or more perfect 3'UTR 6mer, 7mer or 8mer seed matches of the siRNA [78]. They designed 11 siRNAs targeting 3 genes and measured the off target effects using microarray analysis. 347 off-target genes with a 1.5-fold change in the expression rate were observed. For their positive set, they randomly samples 84 genes out of this set. As negative control, they randomly sampled 84 genes out of a set of genes, that showed no change in their expression rate, indicating that there is no interaction between the siRNA and the mRNA. They observed, that 70% of the validated off-target genes have at least one seed region in their UTR, while in the negative control only 30% have one. Unfortunately the group did not note, how exactly they created their negative set. I repeated the same test, using the set of validated off-target genes, provided in the supplementary data as my positive pool and all siRNA:gene pairs, that show no interaction, as my negative pool. I randomly sampled 100 interactions from the positive pool and 100 from the negative pool. Then I searched for reverse complementary 6mers of the siRNAs 5' position 2 to 8, the first and

second 7mer from position 1-7 and 2-8 and the whole 8mer seed from position 1-8. This test was repeated 1000 times and the mean occurrence of at least one seed in the positive and the negative set was calculated (Table 2). According to calculated sensitivity, only

Table 2: Sensitivity, specificity of siRNA 6mer, 7mer and 8mer seed matches

seed	True Positive	False Positive	True Negative	False Negative	Sensitivity	Specificity
6mer	64,60%	24,70%	75,30%	35,40 %	64,6%	75,3%
7mer	28,60%	9,30%	90,70%	71,40%	28,6%	90,7%
8mer	18,20%	2,80%	97,20%	81,80%	18,2%	97,2%

6mer seeds will be used for later off-target analysis.

As mentioned above, a perfect Watson Crick base pairings at microRNAs seed region seem to be the key component for target recognition [32, 33, 31]. I counted all off-target seeds of a set of microRNAs. As Table 3 shows, perfect seed matches range from a few hundred to several thousands occurences for different microRNAs. Indicating that siRNAs with seeds of only a few off-target seed matches in the genome will likely have less real off-target effects, siRNAs should be ranked according to the amount of off-target seeds.

Table 3: miRNAs with their corresponding amount of off-target seeds in all transcript 3'UTRs

miRNA	seeds	miRNA	seeds
hsa-miR126	195	hsa-miR217	4048
hsa-miR-100	353	hsa-miR-218	4165
hsa-miR213	496	hsa-miR-137	4259
hsa-miR-187	730	hsa-miR-138	4344
hsa-miR-184	960	hsa-miR-19b	4431
hsa-miR-196b	2003	hsa-miR-122a	4539
hsa-miR-219	2029	hsa-miR-1	4582
hsa-miR-142-3p	2047	hsa-miR-101	4621
hsa-miR-215	2256	hsa-miR-205	4694
hsa-miR-190	2420	hsa-miR-30b	4714
hsa-miR-21	2526	hsa-miR-199a	4855
hsa-miR-18	2703	hsa-miR-9	4921
hsa-miR-33	2823	hsa-miR-34a	4923
hsa-miR-223	3087	hsa-miR-130a	4936
hsa-miR-375	3168	hsa-miR-22	4996
hsa-let-7a	3205	hsa-miR-26a	5010
hsa-miR-221	3221	hsa-miR-125b	5177
hsa-miR-133a	3221	hsa-miR-7	5426
hsa-miR-10b	3247	hsa-miR-142-5p	5593
hsa-miR-183	3282	hsa-miR-200a	5612
hsa-miR-153	3294	hsa-miR-23b	5937
hsa-miR-140	3613	hsa-miR-211	6163
hsa-miR-135a	3660	hsa-miR-27b	6195
hsa-miR-124a	3748	hsa-miR-203	6449
hsa-miR-29b	3767	hsa-miR-181a	6478
hsa-miR-92	3783	hsa-miR-30a-3p	6513
hsa-miR-216	3825	hsa-miR-15b	6544
hsa-miR-146	3838	hsa-miR-20	6651
hsa-miR-194	4024	hsa-miR-24	6817

5 Methods and Implementation

This section contains details of the implementations of the developed method.

5.1 Objectives

The goal of this thesis is primarily to develop a method to design functional and efficient siRNAs with a high target specificity. For this I turn a particular attention to the microRNA-like off-target effects of these siRNAs. The latter are predicted for each siRNA candidate and the returned siRNAs are ranked according to their amount of off-target genes. During development I also emphasized the possibility for the user to individually change as many settings as possible. This is important, since there are many different and frequently changing views of how to design highly efficient siRNAs. One focus of my work is to provide an easy-to-use method that gives researchers as much influence as possible, by always providing a default value, adjustable by the users to their requirements.

5.2 Applied programming language

The method and the analysis were implemented in the programming language Perl. Perl is a dynamic programming language created by Larry Wall and was first released in 1987 [79]. Perl provides powerful text processing facilities without arbitrary data length limits and it is thus an appropriate language for bioinformtics tools. The Perl modules provide a mechanism for extending the language without modifying the interpreter. There is also a collection of modules available that facilitate the development of Perl scripts for bioinformatics applications. Since the overlap of needed functions for my method and these from BioPerl [80] was small, I decided to designed my own modules, fitted to my needs. The complete written Perl-code can be found on the attached CD.

5.3 Workflow

To develop a computational method for designing highly specific siRNAs by minimizing the *in silico* predicted off-target effects in mammals, I broke down the problem into three central parts: a) siRNA design by following the latest published design rules, b) off-target prediction by using up-to-date microRNA target prediction know-how and c) off-target minimization by ranking the received candidates according to their amount of predicted off-targets. In the following, the single steps pictured in the workflow (Figure 23), are described.

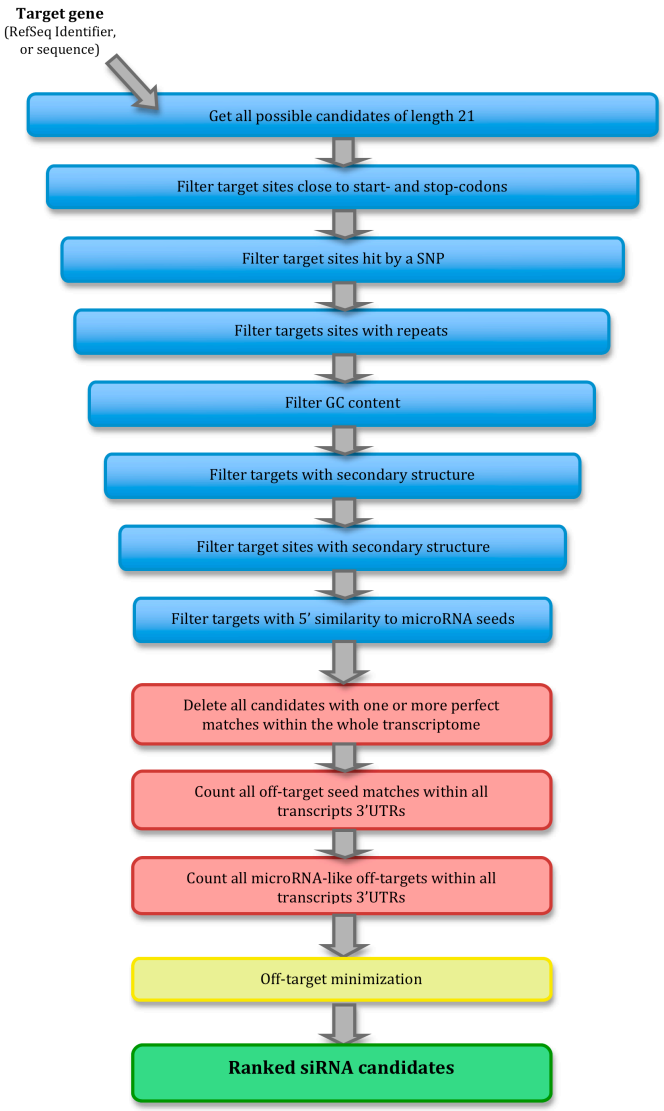

Figure 23: Workflow; The method is partitioned into three central parts: siRNA design (blue), off-target prediction (red) and off-target minimization (yellow).

5.3.1 Design of siRNAs

In order to create functional siRNAs, it is indispensable to use the latest designing rules. Several pharmaceutical companies, specialized on RNA synthesis, as well as a lot of academic groups, regularly publish their experience of how to create siRNAs for getting best knockdown results [55, 42, 1, 15]. They use high-throughput experiments to explore how various design criteria change siRNA's functionality and specificity. Here, I selected the most promising rules extracted on the basis of *in vitro* and/or *in vivo* experiments already performed. I reconsidered commonly used rules, like complete sequence complementar-

ity, G/C content or the relative position within the coding sequence, but I also included more recently detected features, like the local accessibility of the target site and single nucleotide polymorphism within the target sequence.

5.3.1.1 Human genome sequence

Today, several databases are available, that allocate pre-processed gene annotations. The probably most established databases containing that information are RefSeq [81], UCSC [82] and ENSEMBL [83] (Figure 24). I decided to use the latest version (March 2006) of

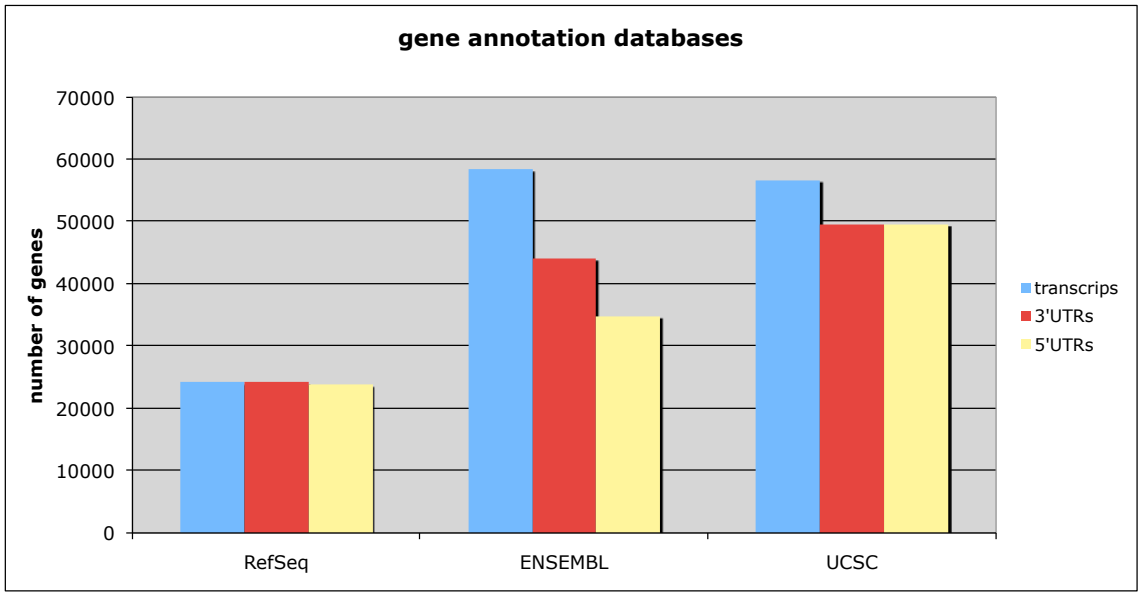

Figure 24: Gene Annotations from RefSeq, ENSEMBL, UCSC.

the NCBI's RefSeq genes [81]. This release contains 24,790 unique transcripts. 23.967 of these have annotated 3'UTR sequences. This database is best for siRNA design, since it is non-redundant, curated, and contains an annotated collection of sequence records for mostly all major model organisms. The human transcriptome was downloaded using the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables) [84]. The table browser provides the option to download just the coding sequences, without introns, and the respective 5' and 3' UTR regions. Moreover it is possible to flag the UTR regions by returning them in lower case letters, making it easy to separate the different regions later. The generated fasta-file contains in his headers the RefSeq-ID (e.g. NM_000942), the chromosome, on which the gene occurs, the absolute positions, where the mRNA starts and ends and the strand. It is important to note that the returned absolute start and stop positions represent the absolute genome region of the whole gene (introns & exons). Hence, the difference between these two positions is not the length of the transcript, which is

intron-free. Furthermore, the received sequences are already converted to the real mRNA 5' to 3' direction. This means that mRNA sequences on the minus strand can only be found on NCBI's whole genome sequence after creating the reverse complement.

5.3.1.2 siRNA length

The length of the siRNAs to be designed is by default set to 21 (see Section 4.1), though I implemented the option to set a user-dependent value. But it has to be mentioned that siRNAs, longer than $\sim$30nt will very likely activate the interferon pathway.

5.3.1.3 Build candidate siRNA list

The candidate siRNA list is generated by simply applying a sliding window approach with a window length l to the target region. l is either set to 21 nt or user-definded (Figure 25).

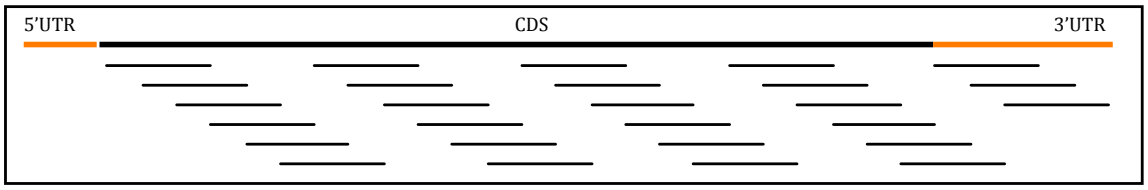

Figure 25: candidate creation; a sliding window generates all possible target sequences of length l, which have a perfect binding with their reverse complement candidate.

5.3.1.4 Paralogous or alternatively spliced transcripts

Paralogous or alternatively spliced genes have high sequence similarity and are thus likely to be co-targeted by a siRNA designed for a single variant. To take account of that, paralogous or alternatively spliced transcripts are searched using BLAST. Transcripts with high sequence similarity (e-value smaller that 10^{-30}) with the target gene are retained for further analysis. There is also an output, containing a list of these potential co-targeted genes.

5.3.1.5 siRNA-target position

Candidates target sites that are close to the start or the stop codon are omitted, because UTR-binding proteins or translation initiation complexes may interfere with binding of the RISC complex [74, 75]. The default exclusion region is set to 100 nucleotides downstream of the start codon and 100 nucleotides up- and downstream of the stop codon. Since the smallest transcript is 146 nt in length (NM_181620), the remaining squence for siRNA targeting could be too short and thus no candidates can be generated. Therefore, an option for a user-definded shorter exclusion region is provided.

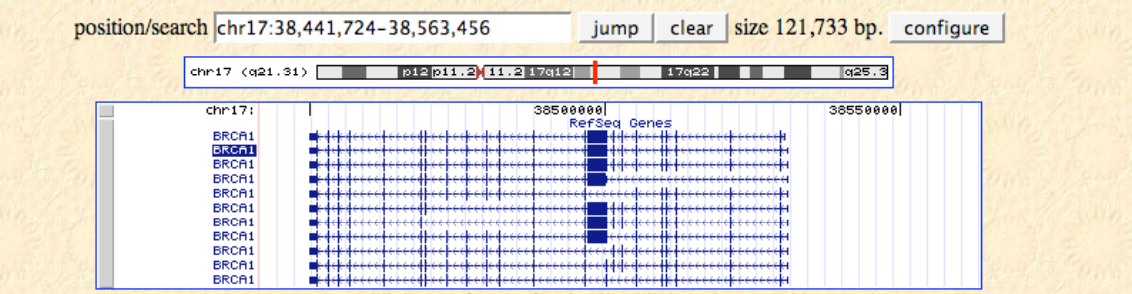

Figure 26: Example from the UCSC Genome Browser [82], for an alternatively spliced gene (BRCA1). All these genes have the same gene name, but different RefSeq-IDs. siRNAs designed for one of these variants have a high possibility to target the other genes, too, since they have a high amount of identical sequences.

5.3.1.6 SNPs in siRNA-targets

It has been demonstrated that single point mutations in siRNA-targets can dramatically change the functionality of the siRNA [1]. For the RISC complex to cleave the targeted mRNA, a near perfect complementary target site is mandatory. As few as a single mutation destroys this perfect complementarity and the siRNA could loose its functionality (Figure 27). Therefore the latest version of the dbSNP database (build 126) was downloaded using the UCSC Table Browser [85, 84]. dbSNP is a database of single nucleotide polymorphisms to address the large-scale sampling designs required by association studies, gene mapping, and evolutionary biology [Ponomarenko 2000]. The database contains all known SNPs for humans. I downloaded a subset of all polymorphisms overlapping with RefSeq genes, using the UCSC table browser's intersection filter to exclude entries with positions that overlap regions of another table. Here, I used the dbSNP data and created an intersection with the RefSeq genes, that are used in this approach. Then I downloaded the gene information file for the used RefSeq transcriptome file. The information in that file includes the absolute position of the respective gene, the positions of each exon and the start and end positions of the CDS. Using that information, I calculated the relative position of each SNP within the transcript sequences and stored it together with a flag, if the SNP is within the 5' UTR, the CDS, or the 3'UTR, respectively. Since the goal of a well-designed siRNA is to target all strains of a species similarly, these data are used to delete all siRNA candidates, targeting sequences containing one or more polymorphisms.

5.3.1.7 Repeats in siRNA-targets

Candidates with four or more A's or T's in a row are deleted, as these are thought to cause premature termination of transcription when the siRNA due to be expressed by

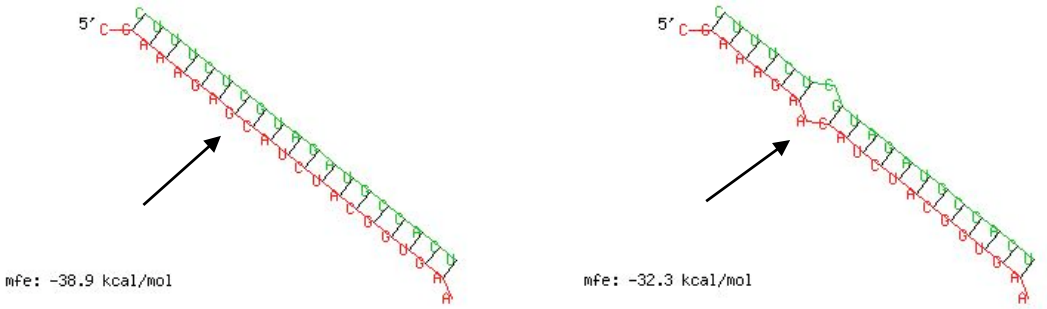

Figure 27: Example for changes in binding energy (mfe: 6.6 kcal/mol ==> 27% more free energy) and structure generated by a single SNP (G/A). Green letters show the candidate siRNA and red letters the target site. This plots are created with RNAhybrid.

a polymerase III [42]. Those with four or more G's and C's are deleted, too, because they have been shown to be one of the strongest negative determinants for siRNA activity [42]. Such regions have pronounced local stability, greatly inhibiting duplex dissociation. Stretches of short repeats have also been shown to reduce functionality and to be less selective. In order to find short repeats in the candidates, a simple search algorithm was implemented. A list with all possible 2mers, as well as all possible 3mers was created. In cases where one of these repeats covers more than 50% of the siRNAs sequence, the candidate is deleted.

5.3.1.8 G/C content

Candidates with a G/C content of less than 30% or greater than 60% are discarded (see 4.3). High overall G/C content is a strong negative determinant of functionality, inhibiting the dissociation of the siRNA duplex, which is necessary for the RISC loading. siRNAs with very low G/C contents (smaller 30%) have been shown to be less functional [42], presumably due to lowered target affinity and specificity. The user may adjust the G/C content, while the default thresholds are set to 30% and 60%.

5.3.1.9 Secondary structure of the siRNAs

Candidates, bearing complementary stretches, leading to an internal structure (e.g. hairpins), are a negative determinant for functionality, possibly through a hindered RISC loading [1]. After the RISC complex unpaired the double-stranded siRNA, the secondary structure of the mature siRNA mustn't be too tight, so that the RISC complex is able to tear it apart (Figure 28). To pitch out structured candidates, the RNA folding tool RNAfold, from the Vienna RNA Secondary Structure Package, is used. Candidates with three or more bindings in a row are deleted.

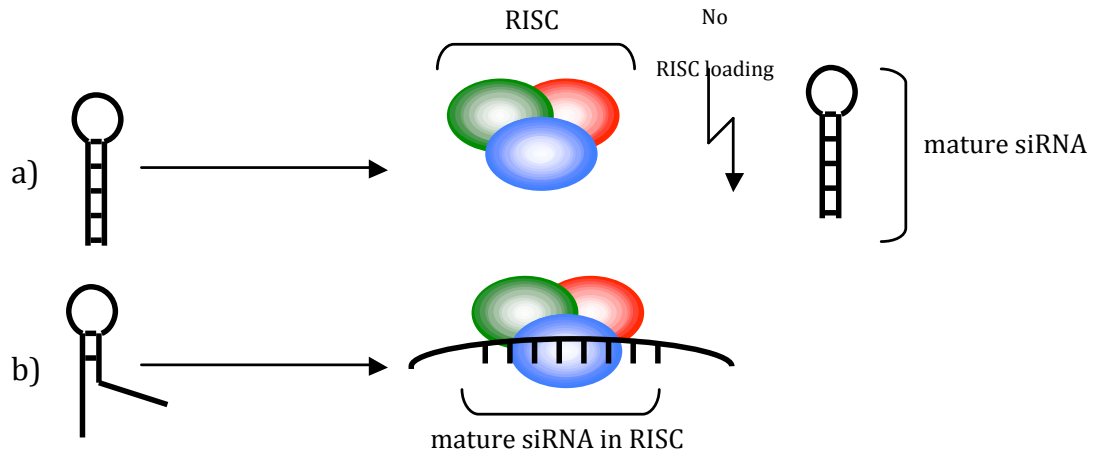

Figure 28: a) Too tight secondary structure of the siRNA makes it impossible for the RISC complex to open the bindings for a correct RISC loading. b) For a good RISC loading, the internal binding has to be less perfect.

5.3.1.10 Secondary structure of the siRNA-targets

The target has to be accessible for the siRNA/RISC complex for binding (see 4.4) [42]. In order to evaluate the accessibility of the target site, the local lowest free energy of the target site, including 30 nucleotides upstream and 20 nt downstream, is calculated, using RNAfold, from the Vienna RNA Secondary Structure Package. Then, RNAduplex is used to get the energy of the target site with the bound siRNA. If the energy of the target's secondary structure alone is lower than the energy of the candidate:target duplex, it is unlikely that the candidate binds to the RNA. The lengths of the flanking regions can also be adjusted, if there will be more accurate experiments and data available in the nearer future.

5.3.1.11 microRNA seed similarity

Several groups have observed that a perfect seed region seems to be enough for microRNA to bind its target mRNA [78, 1]. Those results emphasize the importance of the seed region. For siRNA design, it therefore seems to be a crucial feature. A set of seed regions was extracted by cutting out six nucleotides starting at position two from all known mature microRNA. If one of these 6mer emerges in the candidates first eight bases, the siRNA has a high chance for real off-targets and is subsequently removed.

5.3.2 off-target prediction

In principle off-target prediction is a prediction of microRNA target sites, using siRNAs instead of microRNAs. This is based on the assumption that siRNAs can also function in a microRNA-like manner, performing translational repression, sharing the microRNA pathway as has been shown various times [86, 87, 88]. Since this unintended siRNA:off-target gene interaction may result in phenotypes, unconnected to the actual intended knockdown, the reduction of these off-target bindings is the ultimate goal. To accomplish this, the most common microRNA target prediction methods are applied in this work, in order to predict exactly these target interactions, as accurately as possible.

5.3.2.1 Perfect off-target binding

The most obvious off-target genes are these with a perfect target site, since their mRNAs will most likely be cleaved and degraded. Hence the first step is to mark all candidates that match one or more perfect reverse complementary to one or more sequences in the whole transcriptome, which includes the CDS as well as the UTRs. It has been shown that siRNas that have up to two mismatches can still bind to their target sites to cleave the mRNA [74]. Therefore the second step is to search for sequences that are almost perfectly reverse complementary to the candidate. These sequences will be labeled as well. A blast search would be a fast and easy way to find these sequences. Though the blast algorithm is not designed for such short sequences, fullfilling the additional requirement of accepting one or two mismatches. Instead Kerror (see Section 3.1.4 was integrated for finding these sequences.

5.3.2.2 Perfect seed matches

As shown in 4.5, the seed region seems to be an important anchor for siRNA:target site interaction. So it is important to get these siRNAs, which have less off-target seeds. As seed region the bases 2-7 of the candidates 5' end were used (see 4.5). The frequency of perfect seed matches occuring in all 3' UTRs are counted.

5.3.2.3 microRNA-like off-target prediction

Integration and prediction of microRNA-like off-targets presents one major improvement in siRNA design approaches of this work. For this part only the 3'UTR of the transcriptome is used, since microRNA binding has only shown to be functional in this specific region of the transcripts [75]. Searching for that kind of target is not trivial at all, since experimentally validated microRNA:target interactions are very rare and their underlying principles still remain largly unknown.

siRNA-mRNA binding energy

Most of the existing microRNA target prediction tools use the minimum free energy of the microRNA:mRNA duplex in some way. These energies are then classified by energy thresholds and distributions (like EDV) in order to recieve only the statistically most significant target sites. Though these techniques depend directly on the G/C content of the microRNA. Since a G:C pairing is more stable and consequently has lower free energy, there are much more predicted binding sites for G/C rich microRNA, than for A/T rich ones, introducing a strong bias in target site selection. In this approach, a equitable energy threshold for each siRNA is calculated, based on a unpublished method from Sturm et al. [89]. The basic idea for that strategy is using the best duplex energy of one siRNA candidate at each position of a specific mRNAs 3'UTR. By plotting the energies against the positions of the 3'UTR, a energy-landscape emerges. The minimums of that landscapes represent positions with a sequence, that is highly complementary to the siRNAs sequence. The maximas show sequences, that have little complementarity to the siRNAs sequence. That way, the energy is used as a representer for position specific siRNA attraction. The mean energy μ of all this siRNA:3'UTR duplexes of all genes then stands for the appropriate siRNA attraction to all 3'UTRs. By calculating the standard deviation σ and setting an energy threshold $t = \mu * 3\sigma$, target sites with high attraction to the siRNA can be extracted. Thus the statistically significant target sites are indipendant of the siRNAs G/C content.

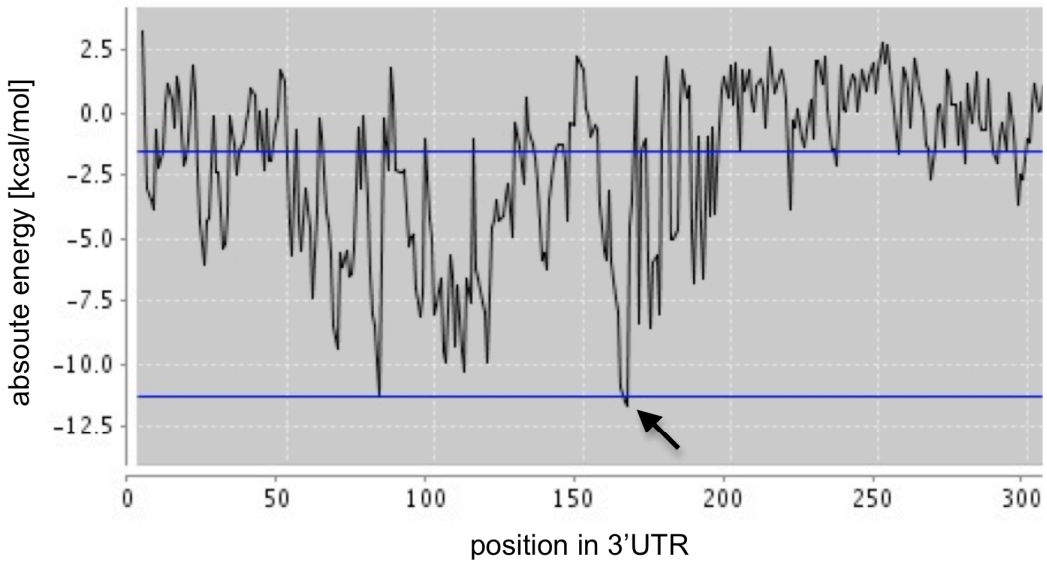

Figure 29: Energy landscape for the 3'UTR of the gene NM_031938 and a siRNA candidate, designed for NM_000942. The upper line represents the mean free energy of the candidate and the lower line the appropriate standard deviation.

Figure 29 shows an example of the individual energy landscape for a siRNA candidate, designed to knockdown the transcript NM_000942, targeting the 3'UTR of the transcript NM_031938. The arrow points to this position that has an energy smaller than the threshold. The target site at the position with the smallest free energy of this small set is the candidate with the highest attraction to the siRNA. This candidate is assumed to be a potential off-target and will be analyzed in the following steps for accessibilty and siRNA:candidate binding structure.

Target-site accessibility

The RISC complex, which has loaded the candidate siRNA must be able to access the mRNA in proximity to the target site (see 4.4). As a measure for accessibility, I chose to use the rati between the free energy of the hybrid and the free energy of the target sites secondary structure. For the latter a flanking region of 15 nt upstream and 10 nt downstream was used. Values greater than 1 indicate accessible target sites.

Binding structure

Like mentioned before, there are two classes of microRNA-like binding. Class one with

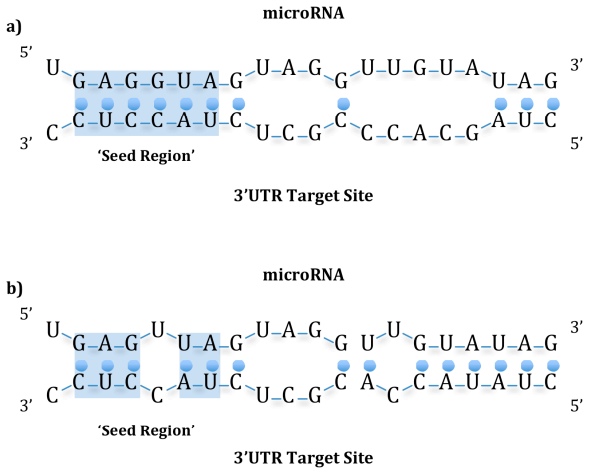

Figure 30: a) Class one microRNA-like off-target binding, with a perfect seed region binding and b) a class two microRNA-like off-target binding with a imperfect seed, but a good 5' binding

a perfect Watson-Crick complementary binding of the 5' ends seed region of the siRNA (Figure 30 a)) and class two with no seed, but additional compensatory bindings at the 3' end of the siRNA (see 1.2.4) (Figure 30 b)). The potential target sites, which pass the previous filters, are now checked for their binding structure. Besides a siRNA binding

that starts at position 1 or 2 of the siRNAs 5' end, one of the following rules has to be applied by a target site to be accepted:

- perfect , consecutive binding of the first 9 bases of the siRNAs 5' end (class one binding)
- perfect binding of the siRNAs seed region, with additional bindings at the siRNAs 3' end (class one binding)
- imperfect binding at the siRNAs 5' end (imperfect seed), with additional compentsatory bindings (minimum 7 of 9 nt) at the siRNAs 3' end (class two)

5.3.3 off-target minimization

First of all, these candidates having perfect or nearly perfect off-target metches are deleted, since these off-targets have the highest probabilty to be real, assuming a siRNA-like off-target transcription cleaveage. After that deletion, siRNA candidates can still range from a handful up to several dozen, according to the transcript length of their target gene. To rank theses siRNAs according to their amount of off-targets, to get this one with the smallest, two steps are implemented. Since the off-target prediction is highly time consuming (up to 3 or more hours per candidate), the siRNAs are firstly ranked according to their amount of off-target seeds. Only an optimized set containing the best five candidates with the smallest amount of off-target seeds will be passed to the second step. After calculating the off-target effects like described above (see 5.3.2, the set of siRNA candidates is ranked by their predicted microRNA-like off-target genes and returned for the user.

6 Results and Discussion

6.1 Workflow

Table 4 shows an example of how candidate siRNA sequences (targeting the transcript of the gene PPIB) are filtered. After designing the candidates, counting the off-target seed regions and achieving the off-target prediction, the user will get a list of siRNA candidates ranked by their predicted target specificity (Figure 31). The process of designing

Table 4: Forty consecutive siRNA sequences from the PPIB gene (NM_000942) and the evaluation at different stages in the siRNA design process.

sequence	close to start-, stop-codon	SNP	Repeats	GC content	secondary structure	local Accessibility	microRNA seeds	perfect matches
AAGTCACCGTCAAGGTGTATT	Passed	Passed	Passed	Passed	....(((((.....)))))....	-	-	-
AGTCACCGTCAAGGTGTATTT	Passed	Passed	Passed	Passed	...(((((.....)))))....	-	-	-
GTCACCGTCAAGGTGTATTTT	Passed	Passed	poly(T)	-	-	-	-	-
TCACCGTCAAGGTGTATTTTG	Passed	Passed	poly(T)	-	-	-	-	-
CACCGTCAAGGTGTATTTTGA	Passed	Passed	poly(T)	-	-	-	-	-
ACCGTCAAGGTGTATTTTGAC	Passed	Passed	poly(T)	-	-	-	-	-
CCGTCAAGGTGTATTTTGACC	Passed	Passed	poly(T)	-	-	-	-	-
CGTCAAGGTGTATTTTGACCT	Passed	Passed	poly(T)	-	-	-	-	-
GTCAAGGTGTATTTTGACCTA	Passed	Passed	poly(T)	-	-	-	-	-
TCAAGGTGTATTTTGACCTAC	Passed	Passed	poly(T)	-	-	-	-	-
CAAGGTGTATTTTGACCTACG	Passed	Passed	poly(T)	-	-	-	-	-
AAGGTGTATTTTGACCTACGA	Passed	Passed	poly(T)	-	-	-	-	-
AGGTGTATTTTGACCTACGAA	Passed	Passed	poly(T)	-	-	-	-	-
GGTGTATTTTGACCTACGAAT	Passed	Passed	poly(T)	-	-	-	-	-
GTGTATTTTGACCTACGAATT	Passed	Passed	poly(T)	-	-	-	-	-
TGTATTTTGACCTACGAATTG	Passed	Passed	poly(T)	-	-	-	-	-
GTATTTTGACCTACGAATTGG	Passed	Passed	poly(T)	-	-	-	-	-
TATTTTGACCTACGAATTGGA	Passed	Passed	poly(T)	-	-	-	-	-
ATTTTGACCTACGAATTGGAG	Passed	Passed	poly(T)	-	-	-	-	-
TTTTGACCTACGAATTGGAGA	Passed	Passed	poly(T)	-	-	-	-	-
TTTGACCTACGAATTGGAGAT	Passed	Passed	Passed	Passed	Passed	Passed	hsa-miR-515	-
TTGACCTACGAATTGGAGATG	Passed	Passed	Passed	Passed	Passed	Passed	hsa-miR-515	-
TGACCTACGAATTGGAGATGA	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed
GACCTACGAATTGGAGATGAA	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed
ACCTACGAATTGGAGATGAAG	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed
CCTACGAATTGGAGATGAAGA	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed
CTACGAATTGGAGATGAAGAT	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed
TACGAATTGGAGATGAAGATG	Passed	Passed	Passed	Passed	Passed	Passed	Passed	NM_024865
ACGAATTGGAGATGAAGATGT	Passed	Passed	Passed	Passed	Passed	Passed	Passed	NM_024865
CGAATTGGAGATGAAGATGTA	Passed	Passed	Passed	Passed	Passed	Passed	Passed	NM_024865
GAATTGGAGATGAAGATGTAG	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed
AATTGGAGATGAAGATGTAGG	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed
ATTGGAGATGAAGATGTAGGC	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed
TTGGAGATGAAGATGTAGGCC	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed
TGGAGATGAAGATGTAGGCCG	Passed	Passed	Passed	52,38%	-	-	-	-
GGAGATGAAGATGTAGGCCGG	Passed	Passed	Passed	57,14%	-	-	-	-
GAGATGAAGATGTAGGCCGGG	Passed	rs1155859	-	-	-	-	-	-
AGATGAAGATGTAGGCCGGGT	Passed	rs1155859	-	-	-	-	-	-
GATGAAGATGTAGGCCGGGTG	Passed	rs1155859	-	-	-	-	-	-
ATGAAGATGTAGGCCGGGTGA	Passed	rs1155859	-	-	-	-	-	-
TGAAGATGTAGGCCGGGTGAT	Passed	rs1155859	-	-	-	-	-	-

functional, specific siRNAs and minimizing their predicted off-target effects *in silico* has been automated by my method. The siRNA design is thus completed in just a few seconds compared to the longer time (hours or days) required for manual calculations of the relevant steps. However, the prediction of the off-target effects is time-consuming, since for each siRNA candidate, the mean binding energy has to be calculated separately. This step can take up to several hours for each siRNA, depending on the available computer power. In my approach, *in silico* off-target effects are compared with *in silico* off-target effects and it is assumed that the error prediction rate is equal for each prediction. Meaning that although the absolute value of off-target effects may be in-accurate, the ratios between

```
get 19-mers:                                1008
SNP Filter:                                 735
Position Filter:                            523
Poly Filter:                                401
Repeat Filter:                              387
GC-Content Filter:                          124
miRNA-seed Filter:                          102
Secondary structure Filter:                 51
Seed rating:                                51
perfect matches Filter:                     9
predict off-target bindings:                5
nr|RefSeq ID|position|co-regulated genes (paralogs)|siRNA-sequence|target_sequence|GC content|off-target seeds|off-target genes
1|NM_000942|316|NM_024798|CTTCATCTCCAATTCGTAG|CTACGAATTGGAGATGAAG|42.10526315789471|4458|1243
2|NM_000942|443|NM_024798|TTGATTACACGATGGAATT|AATTCCATCGTGTAATCAA|31.5789473684211|2776|1804
3|NM_000942|438|NM_024798|TACACGATGGAATTTGCTG|CAGCAAATTCCATCGTGTA|42.10526315789471|752|2019
4|NM_000942|435|NM_024798|ACGATGGAATTTGCTGTTT|AAACAGCAAATTCCATCGT|36.84210526315791|726|2093
5|NM_000942|439|NM_024798|TTACACGATGGAATTTGCT|AGCAAATTCCATCGTGTAA|36.84210526315791|700|2122
```

Figure 31: Returned results for NM_000942; the siRNAs are ranked based to their amount of off-target genes.

the in silico off-target effects, will very likely be representative. Therefore this in silico approximation of the off-target effects is a very good indicator, making expensive and time consuming experiments at this stage dispensable.

6.2 Comparison with commonly used siRNA designing tool

Since my approach ranks the candidate siRNAs on the basis of their individual amount of off-target effects, the designing tools, which I want to compare, should also return a ranked list of their siRNA candidates. That way it is possible to directly compare my best siRNAs to these from each tool. For this purpose I selected the following tools:

- ”siDesign Center” from Dharmacon
- ”BIOPREDsi” from Novartis
- ”BlockIT RNAi Designer” from Invitrogen
- ”siRNA Selection Server” from the Whitehead Institute

All these tools return a ranked list of siRNAs. I picked two genes for this test (NM_000942, NM_002046). I used each of the four siRNA designing tools with their default designing criterias to create the best siRNA candidates. All these tools can be used online, by pasting either the RefSeq Identifier, or the transcript sequence. If there was any choice to define which region to use (5’UTR, ORF, 3’UTR), I choose the ORF and the 3’UTR. If no menu item was given, I pasted the sequence of that specific region, since that is the region were I think siRNAs can bind. Most of the designing programs are quite fast. Only BIOPREDsi took several hours to return its results. Then I used my approach to predict siRNAs for the same three genes and created a table with all the results. The amount of

overlapping sequences was low and if there was an overlap, the ranking positions were absolutely different. By taking a closer look to the designed siRNA candidates from the other tools, it became clear that most of their candidates do not match my design criteria (e.g. secondary structure of the siRNA, repeats, microRNA seed similarity, etc.) and are thus not present in my list of candidates. Thus, these siRNAs have a much higher possibility to be less efficient than these, designed by using my very strict design rules.
To compare the results in a reliable way, I decided to use two different off-target prediction tools, miRanda and RNAhybrid. These two tools provided the opportunity to directly compare all the different designed siRNAs to their amount of microRNA-like off-target effects. Since the frequency of returned siRNAs of each program was very variable, I just used the best 5 candidates for this test. Since RNAhybrid predicts one off-target for each 3'UTR together with a p-value, I introduced a cutoff at a p-value of 0.5 to reduce the amount of predicted target sites and keep only these siRNAs with the highest probability to be real. That way, I still get too many predicted target sites, but the amount of predictions goes constantly up with the p-value (test was performed - unpublished) and since I used the same p-value for each candidate, the results are directly comparable. For miRanda I used the default settings. I plottet the results for each gene in a histogram sorted by their ranked position (Figure 32, 33).
The five siRNA candidates I designed for NM_000942 have significantly less off-target effects than these designed by using the online design tools. Only the top ranked siRNA from the Novartis tool has less predicted off-targets, when miRanda is used for target prediction. The rest of my siRNAs show very low sequence complementarity with the 3' UTR sequences and thus have very few off-targets. Hence, if RNAhybrid and miRanda predict these off-target effects correctly, four out of five siRNAs designed by using my method have a much higher target specificity than the top ranked candidates from the frequently used design tools.
Unfortunately, the tool from the Whitehead Institutes returned just one siRNA candidate for NM_002046. By looking at the highest ranked siRNAs, again, my approache returned four out of five candidates, which have a smaller amount of predicted off-target effects using miRanda and RNAhybrid. Interestingly, the siRNAs designed with the Dharmacon tool 'siDesign Center' seem to have quite a lot off-target genes despite their use of a seed-count for the candidate ranking. Especially miRanda scores the seed region very strong and thus one should expect that the Dharmacon siRNAs perform much better. This observation can be an indication that the count of the seed regions is not enough for minimizing of off-target effects. A recent publication by Didiano et al. [90] is also supporting this observation, saying that the perfect seed match is generally not a reliable predictor for microRNA binding.

My algorithm perfomes very well for these two genes, as the four graphs illustrate quite nicely. There are some outliers for my best candidates, that have more predicted off-targets than some of the siRNA designed by the other tools, but overall, based on these data, my approach returns the best choice of candidates. In this approach, my design criteria are chosen quite strict and none of the other tools use all of these design criteria together. The results can be improved, when these criteria would be relaxed, giving more siRNA candidates the chance to be checked for off-target genes. The higher amount of candidates would most likely lead to more siRNAs with a higher target specificity.

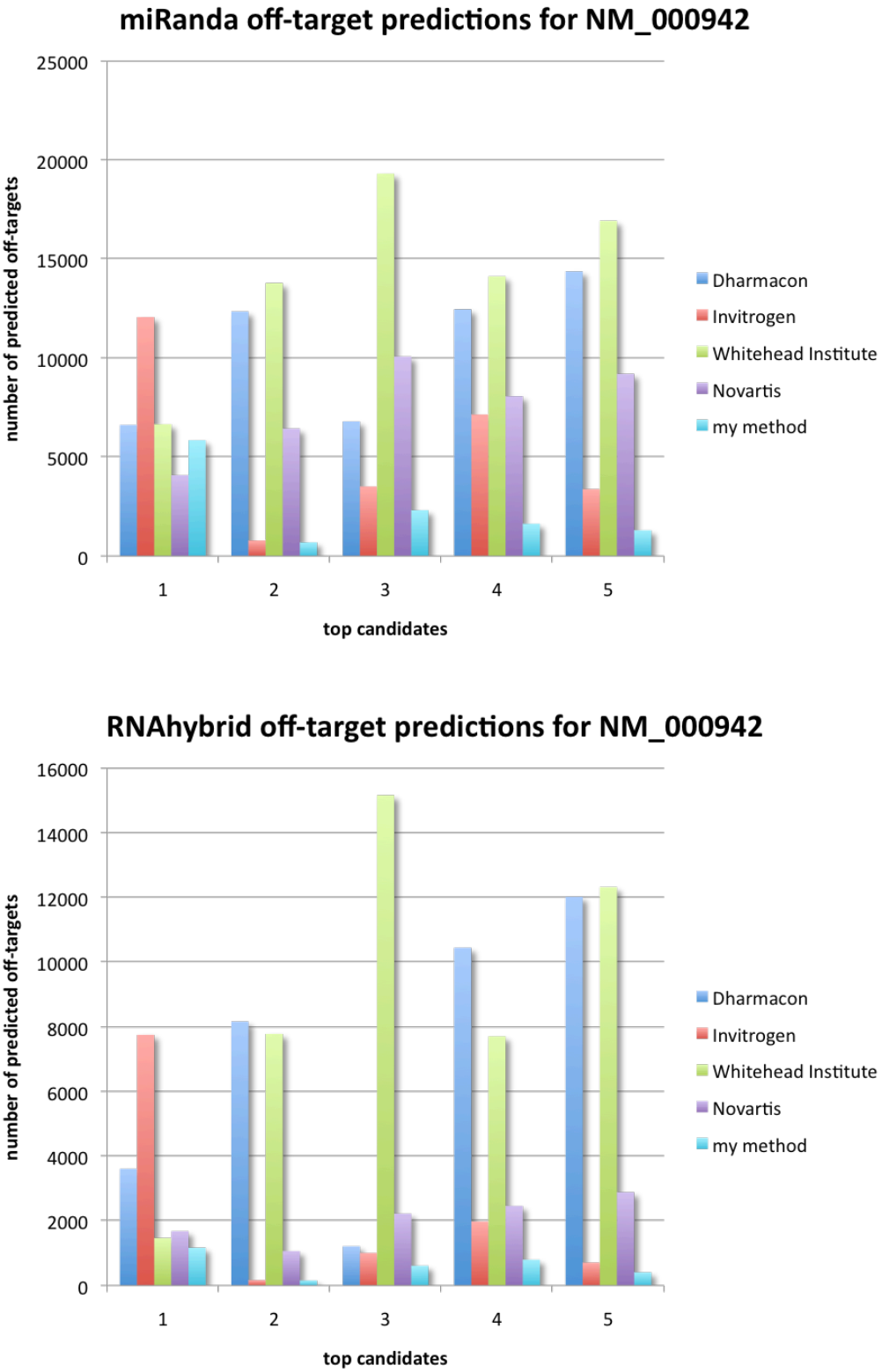

Figure 32: predicted off-target genes of siRNAs designed for NM_000942 using miRanda and RNAhybrid

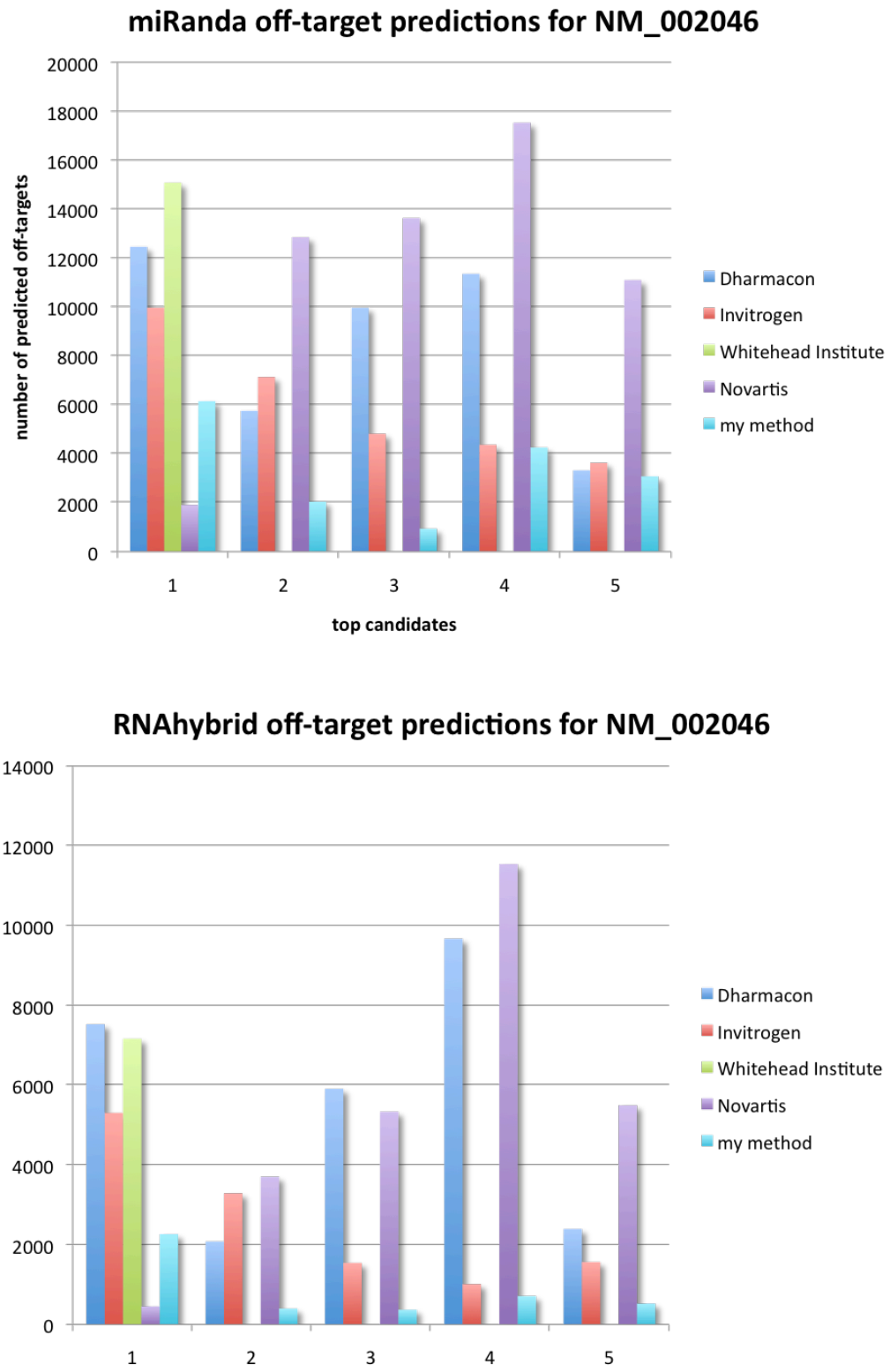

Figure 33: predicted off-target genes of siRNAs designed for NM_002046 using miRanda and RNAhybrid

6.3 Comparison with commonly used microRNA target prediction tools

In order to evaluate my off-target prediction, I compared my tool with some microRNA target prediction tools. I fed the microRNA-like off-target algorithm of my tool with microRNA sequences instead of siRNA candidates. In this way I generate predictions for real microRNAs. For my comparison I used two well known microRNA prediction tools:

- PicTar (set with targets conserved between five species)
- miRanda

I used the web based prediction results from miRanda, since there they use conservation, and thus the results are more reliable and the amount of target sites much lower than using the stand-alone tool. I randomly chose two microRNAs

- hsa-miR-221
- hsa-let-7

for that test and plottet Venn-diagrams representing the preserved overlap (Figure 34, 35).

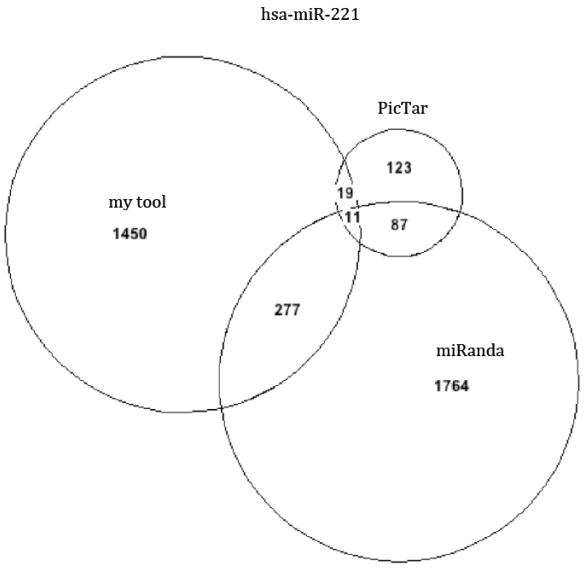

Figure 34: Predicted off-target effects for hsa-miR-221.

The two Venn-diagrams give a good insight into one of the major *in silico* target prediction problems. Most of the target prediction tools return quite different results, with a poor

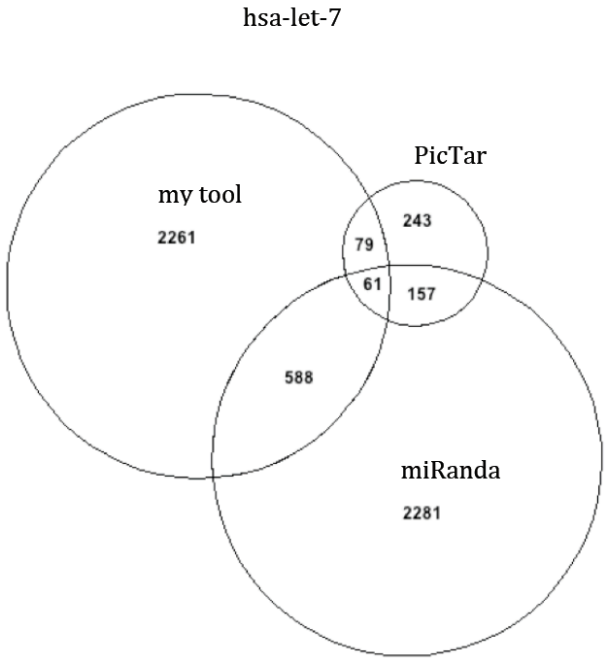

Figure 35: Predicted off-target effects for hsa-let-7.

overlap of smaller than 20%. My prediction joins this observation and has a quite small overlap with the tested microRNA target prediction tools. My method predicts 1.450 targets for hsa-miR-221, with a ∼20% overlap with the miRanda predictions and only a ∼2% overlap with the PicTar targets (see Figure 34). Interestingly, the overlap of all three prediction methods covers only 11 genes, which are ∼0.7% of my predictions. Thus, the major amount of targets predicted by my method, overlapping with PicTar, is not predicted by miRanda and the major amount of the miRanda predictions that overlap with PicTar are not predicted by my tool. Even though, all three methods make use of similar features, like the seed region, the difference between the results is quite big. That can have several reasons. For example do both microRNA target prediction tools, PicTar as well as miRanda use target site conservation for their prediction. My approach ignores this constraint at all, since conservation obviously does not improve siRNA target sites prediction. Moreover, PicTar and miRanda score the seed region quite high, whereas my method rates it only limited. For hsa-let-7 the results look similar as for miR-221 (see Figure 35). Here, my method predicted 2.261 and these have an overlap with miRanda of almost 30% and ∼6% with PicTar. miRanda has an overlap with PicTar of almost 10%.
It is hard to make any statement about the existence of the predicted target sites, since not enough experimentally validated target sites are available. It can be observed that the amount of predicted target sites my method returns is always somewhere between the Pic-

Tar results and miRanda results. Based on the facts, that my tool is used for minimization only and it performs always like that, it can be assumed that the siRNA candidates of my approach will be ranked according to their increasing microRNA-like off-target effects.

6.4 Recall of experimentally validated off-targets

To see how good my method performs on real off-target effects, I used a set of experimentally validated off-targets from Dharmacon [78]. They tested 11 siRNAs for off-target effects and found 315 off-target genes. I used my microRNA-like off-target algorithm, the stand-alone version of miRanda and RNAhybrid to predict the off-targets. Then I calculated the pecision value for each tool using the recall of validated targets (True Positve) and the targets that were predicted by the tools, but not by the experiment (False Positive).

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{2}$$

Table 5 shows the results and the according precision of each tool over all 11 siRNAs.

Table 5: The overall performance of RNAhybrid, miRanda and my method in predicting validated targets

Prediction Method	Number of Predicted Targets	True Positives	False Positives	Precision
RNAhybrid	$40,492$	59	$40,433$	$1.45 * 10^{-3}$
miRanda	$31,757$	141	$31,616$	$4.44 * 10^{-3}$
my method	$27,478$	51	$27,427$	$1.86 * 10^{-3}$

Figure 36 illustrates the precision value of each tool for each single siRNA separately. Here, miRanda looks better than my prediction method and RNAhybrid, but on closer inspection, none of the tools performed satisfiable, since the precision values are all very small (10^{-3}). My method predicts the smallest amount of off-target hits, but has a recall of only 51 out of 315 validated off-target genes, which are $\sim$16%. miRanda has a recall of $\sim$45%, but predicts $\sim$4,300 more targeted genes for all 11 siRNAs than my method. RNAhybrid has the worst result, by having a recall of $\sim$19%, but predicting more than 13,000 more off-target genes than my approach. These results can have several reasons, like the fact that the experiments were performed in Hela-cells and thus a lot of genes are not expressed correctly, or not expressed at all. The predictions of the tools were performed by using the whole transcriptome, instead of using a list of expressed genes in Hela-cells. Thus, all three tools, off course, predict too many targets. That explains

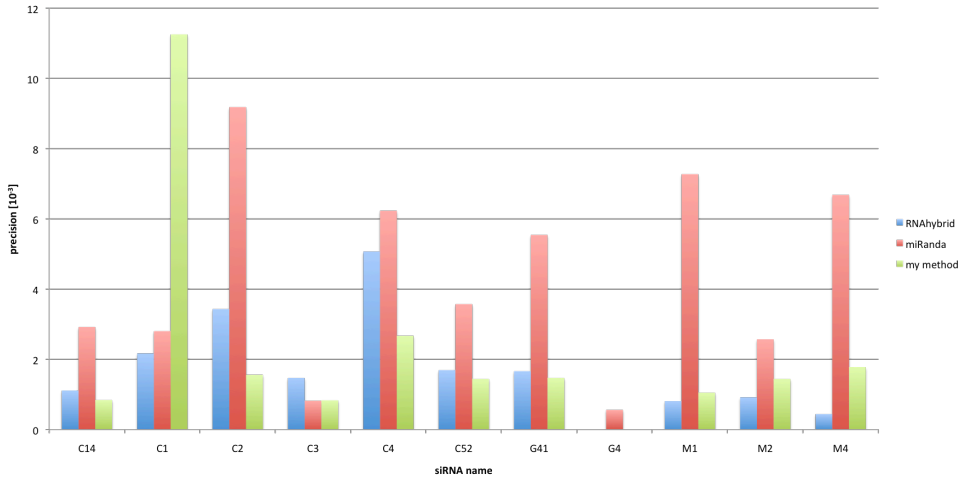

Figure 36: Precision of my tool, miRanda and RNAhybrid for prediction of validated off-target effects for 11 designed siRNAs from Dharmacon.

the very low precision value and the consequently high amount of off-targets. These off-target are thus not necessarily wrong predictions, perhaps the gene is just not expressed and thus the siRNA can not regulate it. Connecting the results with expression data from this cell-line, would probably enhance the result and decrease the amount of predicted off-target genes of each tool. By relaxing some of the off-target prediction criteria, my approach would most likely get a better precision value. But the main exercise of this method is to minimze off-target effects rather than predict real off-targets. Thus I decided to have less predicted off-targets than changing any of my prediction features.

7 Conclusion and outlook

The siRNA design method, developed for this work, gives the user the possibility to control the design critera, by highlighting the obtimal values for each step as default. Unfortunately, there was no chance to test the designed siRNAs for functionality and efficiency by performing any *in vitro* or/and *in vivo* experiment. However, they are designed using all important design rules that were available at this time.

The local accessibilty of target sites was one of the most difficult problems in this work. The results of the analysis were not satisfactory at all. I decided to integrate the method for predicting the accessibility in my approach, but give the user the option to deactivate it. It has been shown that the accessibility affects the siRNA binding, but until now, it remain unknown, how RISC really binds the target site and whether secondary structure and accessibility are important or not. In addition the secondary structure prediction tools prediction tools as well and may not model nature in this specific case correctly.

The off-target prediction algorithm, developed in this work, performes quite well, as long as the goal is to minimize the off-target effects. The prediction of real off-targets is suboptimal, since not enough real experimental data for microRNA:target binding is available. Several recently published features for microRNA target prediction like an A/U rich flanking of the target site [65], or the relative position of the target site within the 3' UTRs [75] have been ommited in my method, since the results of thoroughly designed *in silico* tests have shown no signal at all (unpublished results). But recently, there was a publication that confirmed my results with *in vivo* experiments that were interpreted using bioinformatics tools. Didiano et al. [90] showed in *C. elegans* that target sites of the microRNA *lsy-6* in the 3' UTR of the *cog-1* gene for example do not need A/U rich flanking regions to be functional. He even observed that microRNA seed regions are not neccessary for *lsy6-cog1* binding. Since it throws light on the main problem with microRNA target predictions, namely the lack of validated real target sites. Still not enough is known about this highly complex interaction between the microRNA and the target site. Here one example is the microRNA seed region. In the beginning of microRNA target prediction research, *in silico* experiments revealed that these seven nucleotides of the target sites 3' end seem to be highly conserved. Thus most bioinformaticians used this feature for their predictions and then validated their target sites by *in vitro* or *in vivo* experiments. Consequently, most of the validated target sites have conserved seed regions leading to a strong bias towards the seed region.

Altogether it should be clear in everyones mind that all microRNA target prediction tools are just based on limited knowledge and consequently their results just provide an indication of the microRNA behaviour. Since the goal of my off-target prediction was to

evaluate the relative performance of off-target effects for the designed siRNAs and return the siRNAs having only a few, it was possible for me to design the prediction method less strict than most of the microRNA target prediction tool are. And it successfully returns these siRNAs that have an overall low probabiltity of off-targets and are thus high specific.

Since most of the siRNAs are only used in single tissues, one goal for the future is to integrate tissue specific expression data, for a more accurate off-target forecast. To obtain a better specificity and sensitifity for off-target prediction, more experimental data are neccesary. It may then even be possible to design sets of microRNAs that fine tunes specific genes in a very well defined scope in order to create drugs that have less off-target effects and are more biocompatible than recent siRNAs

References

[1] A. Birmingham, E. Anderson, K. Sullivan, A. Reynolds, Q. Boese, D. Leake, J. Karpilow, and A. Khvorova. A protocol for designing siRNAs with high functionality and specificity. 2007.

[2] N. Shrivastava and A. Srivastava. RNA interference: An emerging generation of biologicals. *BIOTECHNOLOGY JOURNAL*, 3(3):339, 2008.

[3] Z. Racz and P. Hamar. Can siRNA Technology Provide the Tools for Gene Therapy of the Future? *Current Medicinal Chemistry*, 13(19):2299–2307, 2006.

[4] A.M. Gewirtz. On future's doorstep: RNA interference and the pharmacopeia of tomorrow. *Journal of Clinical Investigation*, 117(12):3612, 2007.

[5] CJS Smith, CF Watson, J. Ray, CR Bird, PC Morris, W. Schuch, and D. Grierson. Antisense RNA inhibition of polygalacturonase gene expression in transgenic tomatoes. *Nature*, 334(6184):724–726, 1988.

[6] P.D. Zamore. Ancient Pathways Programmed by Small RNAs. *Science's STKE*, 296(5571):1265, 2002.

[7] G. Hutvágner and P.D. Zamore. RNAi: nature abhors a double-strand. *Current Opinion in Genetics & Development*, 12(2):225–232, 2002.

[8] AR Krol, PE Lenting, J. Veenstra, IM Meer, RE Koes, AGM Gerats, JNM Mol, and AR Stuitje. An antisense chalcone synthase gene in transgenic plants inhibits flower pigmentation. *Nature*, 333:866–869, 1988.

[9] S. References, C. Napoli, C. Lemieux, and R. Jorgensen. Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. *Plant Cell*, 2(4):279–289, 1990.

[10] S. Guo and K.J. Kemphues. par-1, a gene required for establishing polarity in C. elegans embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. *Cell*, 81(4):611–620, 1995.

[11] H. Tabara, A. Grishok, and CC Mello. RNAi in C. elegans: soaking in the genome sequence. *Science*, 282(5388):430–1, 1998.

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

[13] T. Tuschl, P.D. Zamore, R. Lehmann, D.P. Bartel, and P.A. Sharp. Targeted mRNA degradation by double-stranded RNA in vitro. *Genes & Development*, 13:3191–3197, 1999.

[14] P.D. Zamore, T. Tuschl, P.A. Sharp, and D.P. Bartel. RNAi Double-Stranded RNA Directs the ATP-Dependent Cleavage of mRNA at 21 to 23 Nucleotide Intervals. *Cell*, 101(1):25–33, 2000.

[15] S.M. Elbashir, W. Lendeckel, and T. Tuschl. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes & Development*, 15:188–200, 2001.

[16] D.P. Bartel. MicroRNAs Genomics, Biogenesis, Mechanism, and Function. *Cell*, 116(2):281–297, 2004.

[17] S. Griffiths-Jones, R.J. Grocock, S. van Dongen, A. Bateman, and A.J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*.

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

[19] J.J. Song, J. Liu, N.H. Tolia, J. Schneiderman, S.K. Smith, R.A. Martienssen, G.J. Hannon, and L. Joshua-Tor. The crystal structure of the Argonaute 2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. *Nature Structural Biology*, 10(12):1026–1032, 2003.

[20] H. Zhang, F.A. Kolb, V. Brondani, E. Billy, and W. Filipowicz. Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. *The EMBO Journal*, 21:5875–5885, 2002.

[21] M.A. Carmell et al. The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes & Development*, 16(21):2733–2742, 2002.

[22] JB Ma, K. Ye, and DJ Patel. Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature*, 429(6989):318–22, 2004.

[23] K.S. Yan, S. Yan, A. Farooq, A. Han, and L. Zeng. Structure and conserved RNA binding of the PAZ domain. *Nature*, 426:469–474, 2003.

[24] Y.S. Lee, K. Nakahara, J.W. Pham, K. Kim, Z. He, E.J. Sontheimer, and R.W. Carthew. Distinct Roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA Silencing Pathways. *Cell*, 117(1):69–81, 2004.

[25] G. Tang. siRNA and miRNA: an insight into RISCs. *Trends in Biochemical Sciences*, 30(2):106–114, 2005.

[26] S.L. Lin, D. Chang, and S.Y. Ying. Asymmetry of intronic pre-miRNA structures in functional RISC assembly. *Gene*, 356:32–38, 2005.

[27] E.P. Murchison and G.J. Hannon. miRNAs on the move: miRNA biogenesis and the RNAi machinery. *Current Opinion in Cell Biology*, 16(3):223–229, 2004.

[28] B. Wightman, TR Burglin, J. Gatto, P. Arasu, and G. Ruvkun. Negative regulatory sequences in the lin-14 3'-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development. *Genes & Development*, 5(10):1813, 1991.

[29] B. Wightman, I. Ha, G. Ruvkun, et al. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5):855–862, 1993.

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

[31] B.P. Lewis, I. Shih, M.W. Jones-Rhoades, D.P. Bartel, and C.B. Burge. Prediction of Mammalian MicroRNA Targets. *Cell*, 115(7):787–798, 2003.

[32] A. Stark, J. Brennecke, R.B. Russell, and S.M. Cohen. Identification of Drosophila microRNA targets. *PLoS Biol*, 1(3):E60, 2003.

[33] N. Rajewsky and N.D. Socci. Computational identification of microRNA targets. *Developmental Biology*, 267(2):529–535, 2004.

[34] L.P. Lim, N.C. Lau, E.G. Weinstein, A. Abdelhakim, S. Yekta, M.W. Rhoades, C.B. Burge, and D.P. Bartel. The microRNAs of Caenorhabditis elegans. *Genes & Development*, 17(8):991, 2003.

[35] EC Lai. Micro RNAs are complementary to 3'UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet*, 30(4):363–4, 2002.

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

[37] M.C. Vella, E.Y. Choi, S.Y. Lin, K. Reinert, and F.J. Slack. The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR. *Genes & Development*, 18:132–137, 2004.

[38] M.C. Vella, K. Reinert, and F.J. Slack. Architecture of a Validated MicroRNA:: Target Interaction. *Chemistry & Biology*, 11(12):1619–1623, 2004.

[39] H. Großhans, T. Johnson, K.L. Reinert, M. Gerstein, and F.J. Slack. The Temporal Patterning MicroRNA let-7 Regulates Several Transcription Factors at the Larval to Adult Transition in C. elegans. *Developmental Cell*, 8(3):321–330, 2005.

[40] S.M. Johnson, H. Grosshans, J. Shingara, M. Byrom, R. Jarvis, A. Cheng, E. Labourier, K.L. Reinert, D. Brown, and F.J. Slack. RAS Is Regulated by the let-7 MicroRNA Family. *Cell*, 120(5):635–647, 2005.

[41] J. Brennecke, A. Stark, R.B. Russell, and S.M. Cohen. Principles of microRNA-target recognition. *PLoS Biol*, 3(3):e85, 2005.

[42] Ambion. siRNA Design Guidelines. http:www.ambion.comtechlibtbtb_506.html.

[43] M.P. Butler, J.A. Hanly, and P.N. Moynagh. Pellino3 Is a Novel Upstream Regulator of p38 MAPK and Activates CREB in a p38-dependent Manner. *Journal of Biological Chemistry*, 280(30):27759, 2005.

[44] J. Dai, S. Sultan, S.S. Taylor, and J.M.G. Higgins. The kinase haspin is required for mitotic histone H3 Thr 3 phosphorylation and normal metaphase chromosome alignment. *Genes & Development*, 19(4):472, 2005.

[45] K. Krysan, H. Dalwadi, S. Sharma, M. Pold, and S. Dubinett. Cyclooxygenase 2-Dependent Expression of Survivin Is Critical for Apoptosis Resistance in Non-Small Cell Lung Cancer, 2004.

[46] L. Pelkmans, E. Fava, H. Grabner, M. Hannus, B. Habermann, E. Krausz, and M. Zerial. Genome-wide analysis of human kinases in clathrin-and caveolae/raft-mediated endocytosis. *Nature*, 436(7047):78–86, 2005.

[47] L. Pelkmans and M. Zerial. Kinase-regulated quantal assemblies and kiss-and-run recycling of caveolae. *Nature*, 436(7047):128–33, 2005.

[48] J.P. MacKeigan, L.O. Murphy, and J. Blenis. Sensitized RNAi screen of human kinases and phosphatases identifies new regulators of apoptosis and chemoresistance. *Nature Cell Biology*, 7:591–600, 2005.

[49] M. Yang, W.W. Zhong, N. Srivastava, A. Slavin, J. Yang, T. Hoey, and S. An. G protein-coupled lysophosphatidic acid receptors stimulate proliferation of colon cancer cells through the {beta}-catenin pathway. *Proceedings of the National Academy of Sciences*, 102(17):6027, 2005.

[50] J. Shen, R. Samul, R.L. Silva, H. Akiyama, H. Liu, Y. Saishin, S.F. Hackett, S. Zinnen, K. Kossen, K. Fosnaugh, et al. Suppression of ocular neovascularization with siRNA targeting VEGF receptor. *Gene Therapy*, 13:225–234, 2006.

[51] J. Soutschek, A. Akinc, B. Bramlage, K. Charisse, R. Constien, M. Donoghue, S. Elbashir, A. Geick, P. Hadwiger, J. Harborth, et al. Therapeutic silencing of an endogenous gene by systemic administration of modified siRNAs. *Nature*, 432(7014):173–178, 2004.

[52] C. Raoul, T. Abbas-Terki, J.C. Bensadoun, S. Guillot, G. Haase, J. Szulc, C.E. Henderson, and P. Aebischer. Lentiviral-mediated silencing of SOD1 through RNA interference retards disease onset and progression in a mouse model of ALS. *Nature Medicine*, 11:423–428, 2005.

[53] A.L. Jackson and P.S. Linsley. Noise amidst the silence: off-target effects of siRNAs? *Trends in Genetics*, 20(11):521–524, 2004.

[54] A.L. Jackson, S.R. Bartz, J. Schelter, S.V. Kobayashi, J. Burchard, M. Mao, B. Li, G. Cavet, and P.S. Linsley. Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnology*, 21(6):635–637, 2003.

[55] A. Reynolds, D. Leake, Q. Boese, S. Scaringe, W.S. Marshall, and A. Khvorova. Rational siRNA design for RNA interference. *Nature Biotechnology*, 22(3):326–330, 2004.

[56] K. Ui-Tei, Y. Naito, F. Takahashi, T. Haraguchi, H. Ohki-Hamazaki, A. Juni, R. Ueda, and K. Saigo. Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Research*, 32(3):936–948, 2004.

[57] A.M. Chalk, C. Wahlestedt, and E.L.L. Sonnhammer. Improved and automated prediction of effective siRNA. *Biochemical and Biophysical Research Communications*, 319(1):264–274, 2004.

[58] M. Amarzguioui and H. Prydz. An algorithm for selection of functional siRNA sequences. *Biochemical and Biophysical Research Communications*, 316(4):1050–1058, 2004.

[59] I.L. Hofacker, M. Fekete, and P.F. Stadler. Secondary Structure Prediction for Aligned RNA Sequences. *Journal of Molecular Biology*, 319(5):1059–1066, 2002.

[60] IL Hofacker, W. Fontana, PF Stadler, LS Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, 1994.

[61] B. Haubold. Compute k-error matches between 2 sequences. 2005.

[62] B. Haubold and T. Wiehe. *Introduction to Computational Biology: An Evolutionary Approach*. Birkhauser, 2006.

[63] G. Altschul and M. Miller. Lipman," Basic local alignment search tool,". *Journal of Molecular Biology*, 215:403–410, 1990.

[64] E.C. Lai. Predicting and validating microRNA targets. *feedback*, 2005.

[65] B.P. Lewis, C.B. Burge, and D.P. Bartel. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell*, 120(1):15–20, 2005.

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

[67] A.J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D.S. Marks. MicroRNA targets in Drosophila. *feedback*, 2004.

[68] S. Wuchty, W. Fontana, I.L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–165, 1999.

[69] M. REHMSMEIER et al. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, 2004.

[70] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406–3415, 2003.

[71] J. Kruger and M. Rehmsmeier. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Research*, 34(Web Server issue):W451, 2006.

[72] J.F. Rual, N. Klitgord, and G. Achaz. Novel insights into RNAi off-target effects using C. elegans paralogs. *BMC Genomics*, 8(1):106, 2007.

[73] S. Qiu et al. A computational study of off-target effects of RNA interference. *Nucleic Acids Research*, 33(6):1834–1847, 2005.

[74] T. Tuschl, S. Elbashir, J. Harborth, and K. Weber. The siRNA user guide. *URL: http://www. rockefeller. edu/labheads/tuschl/sirna. html*, 2003.

[75] A. Grimson, K.K.H. Farh, W.K. Johnston, P. Garrett-Engele, L.P. Lim, and D.P. Bartel. MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Molecular Cell*, 27(1):91–105, 2007.

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

[77] Y. Zhao, J.F. Ransom, A. Li, V. Vedantham, M. von Drehle, A.N. Muth, T. Tsuchihashi, M.T. McManus, R.J. Schwartz, and D. Srivastava. Dysregulation of Cardiogenesis, Cardiac Conduction, and Cell Cycle in Mice Lacking miRNA-1-2. *Cell*, 129(2):303–317, 2007.

[78] A. Birmingham, E.M. Anderson, A. Reynolds, D. Ilsley-Tyree, D. Leake, Y. Fedorov, S. Baskerville, E. Maksimova, K. Robinson, J. Karpilow, et al. 3'UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nat Methods*, 3(3):199–204, 2006.

[79] L. Wall and R. Schwartz. Perl. *Perl Man page, Perldoc website, visited September*, 2001.

[80] J.E. Stajich, D. Block, K. Boulez, S.E. Brenner, S.A. Chervitz, C. Dagdigian, G. Fuellen, J.G.R. Gilbert, I. Korf, H. Lapp, et al. The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*, 12(10):1611, 2002.

[81] K.D. Pruitt, T. Tatusova, and D.R. Maglott. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(Database Issue):D501, 2005.

[82] D. Karolchik, R. Baertsch, M. Diekhans, TS Furey, A. Hinrichs, YT Lu, KM Roskin, M. Schwartz, CW Sugnet, DJ Thomas, et al. The UCSC Genome Browser Database. *Nucleic Acids Research*, 31(1):51–54, 2003.

[83] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, et al. The Ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41, 2002.

[84] D. Karolchik, A.S. Hinrichs, T.S. Furey, K.M. Roskin, C.W. Sugnet, D. Haussler, W.J. Kent, and O. Journals. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(90001):493–496.

[85] E.M. Smigielski, K. Sirotkin, M. Ward, S.T. Sherry, and O. Journals. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Research*, 28(1):352–355, 2000.

[86] D. Baulcombe. Overview of RNA interference and related processes.

[87] D. Murphy, B. Dancis, and J.R. Brown. The evolution of core proteins involved in microRNA biogenesis. *BMC Evolutionary Biology*, 8(1):92, 2008.

[88] A. Khvorova, A. Reynolds, and S.D. Jayasena. Functional siRNAs and miRNAs Exhibit Strand Bias. *Cell*, 115(2):209–216, 2003.

[89] M. Sturm. Targetspy. in progress.

[90] D. Didiano and O. Hobert. Molecular architecture of a miRNA-regulated 3'UTR. *RNA*, 2008.

List of Figures

List of Tables