NTNU
Norwegian University of
Science and Technology

MicroRNAs and Transcriptional Control

Even Skaland

Master of Science in Computer Science
Submission date: June 2009
Supervisor: Arne Halaas, IDI
Co-supervisor: Pål Sætrom, IDI

Norwegian University of Science and Technology
Department of Computer and Information Science

Problem Description

Cis natural antisense transcripts (cis-NATs) are RNAs that are complementary to an RNA on the opposite strand in the DNA. Research has suggested that such cis-NATs have regulatory roles in the genome, although the mechanisms behind it is not well understood at this point.

The hypothesis to be tested is that MicroRNAs (miRNAs), which is a class of regulating RNAs, might bind to the non-coding side of such cis-NAT pairs. The end result of this binding is that the coding side of the cis-NAT pair is being activated. Thus, the goal of this project is to test this hypothesis.

Assignment given: 15. January 2009
Supervisor: Arne Halaas, IDI

Preface

This report represents the master thesis of my 5th year in the Computer Science Master program at NTNU in Trondheim. The project has been done at the St. Olavs Hospital. I would like to thank Pål Sætrom for being available for help and discussions, even when being busy with other work.

Abstract

Background: MicroRNAs are small non-coding transcripts that have regulatory roles in the genome. Cis natural antisense transcripts are transcripts overlapping a sense transcript at the same loci in the genome, but at the opposite strand. Such antisense transcripts are thought to have regulatory roles in the genome, and the hypothesis is that miRNAs might bind to such antisense transcripts and thus activate the overlapping sense transcript. **Aim of study:** The following two aims have been identified during this project: (1) investigate whether the non-coding transcript of cis-NATs show significant enrichment for conserved miRNA seed sites, and (2) to correlate miRNA expression with expression of the sense side of targeted cis-NAT pairs. **Results:** Seed sites within such antisense transcripts gave significant enrichment, suggesting that miRNAs might actually bind to such antisense transcripts. There is a significant negative correlation between the expression of mir-28 and the expression of its targeted antisense transcripts, whereas the other miRNAs have no significant correlations. Also, the 3'UTR of the sense side of cis-NAT pairs is longer and more conserved than random transcripts. **Conclusion:** This work has strengthened the hypothesis that miRNAs might bind to such antisense transcripts.

Contents

List of Tables

List of Figures

Chapter 1

Background

This chapter will review the biological background for the study, and give a brief introduction to how miRNA and siRNA based regulation works. There will also be a quick introduction to cis natural antisense transcripts (cis-NATs).

1.1 MicroRNAs regulate protein translation

MicroRNAs (miRNAs) are conserved non-coding RNAs (ncRNAs) that are transcribed from the genome, but not further translated into a protein as many other genes are [1]. MicroRNAs are first transcribed (in the cell nucleus) into a structure called the primary transcript, or pri-miRNA (see figure 1.1). Then the miRNA is further processed into a ~70 nucleotide hairpin-loop called pre-miRNA. In animals, the Microprocessor complex, which consists of the two enzymes Drosha and Pasha (also called DGCR8), is doing this process. Plants, on the other hand, lack a Drosha homologue, and use a Dicer homologue to process the pri-miRNA into a duplex form in the nucleus [1].

After this pre-miRNA has been formed, it is transported into the cytoplasm with the help of the carrier protein Exportin 5 [1]. In the cytoplasm, this pre-miRNA is further processed by an enzyme called Dicer. This process involves cleaving the hairpin-loop, and two complementary short strands (~19-23 nt) are formed. After this step, one of the two strands is selected to be integrated into the RNA-Induced Silencing Complex (RISC-complex). The selected strand is called the ”guide strand”, and the protein Argonaute is selecting this strand, whereas the other strand is degraded. After integrating into the RISC-complex, the miRNA guides RISC to target RNAs by binding to partially complementary messenger RNAs (mRNAs). This whole process of translational gene silencing is known as RNA interference (RNAi).

Although there exist examples of perfect complementarity between miRNAs and mRNAs [2], most miRNAs have only partial complementarity to mRNAs. It has also been shown that complementarity between mRNAs and the positions 2-8 (from the 5’ side) of miRNAs are most crucial for miRNAs to

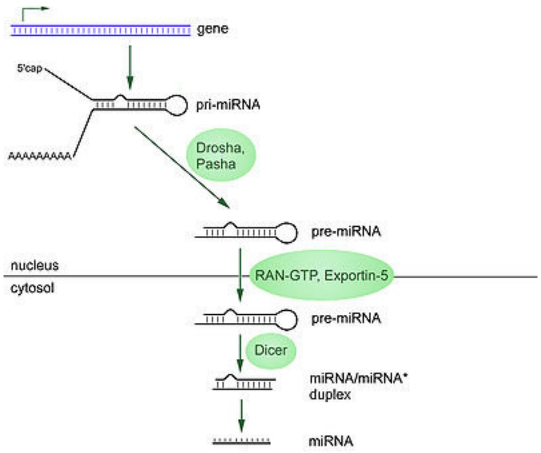


Figure 1.1: The RNAi pathway (from wikipedia).

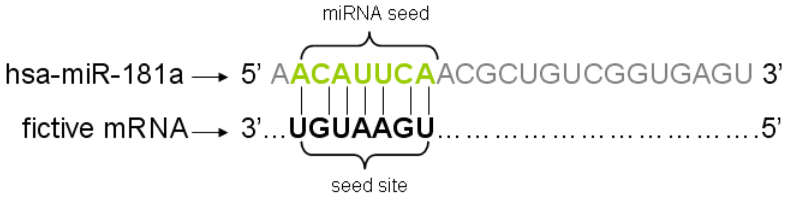

miRNA seed
hsa-miR-181a → 5' AACAUUCAACGCUGUCGGUGAGU 3'
fictive mRNA → 3'...UGUAAGU...............................5'
seed site

Figure 1.2: The mature sequence of the miRNA hsa-miR-181a is shown on top. Seed positions 2-8 (7mer-m8) is highlighted in this sequence. This miRNA binds to a fictive mRNA. The complementary site in the fictive mRNA is called seed site. The miRNA sequence is listed from 5' end to 3' end, whereas the mRNA is listed in the opposite direction.

induce mRNA-regulation. Even if positions 2-8 are most important for mRNA-regulation, complementarity between mRNAs and miRNA positions 2-7 might induce mRNA-regulation as well, but at a significantly lower rate [3].

Positions 2-8 of miRNAs are referred to as the seed sequence, whereas sites with perfect complementarity to the seed sequence in mRNAs are called seed sites (see figure 1.2). Sequences, for example inside miRNAs, of length 7 are commonly called heptamers, whereas sequences of length 6 are called hexamers. Research has also shown that most miRNA target sites in animals are located in the 3'UTR of mRNAs [3, 4], but some sites in coding sequences (CDS) can also be functional.

1.2 Short interfering RNAs

Short interfering RNAs (siRNAs) are ~20-25 nt double stranded RNA (dsRNA) molecules. Whereas miRNAs are transcribed from the genome itself, siRNAs are typically exogenous, meaning that they are introduced into cells through for example laboratory manipulations. Since siRNAs are exogenous, they are not processed by the Microprocessor, but are directly incorporated into the RISC complex. Short interfering RNAs can therefore act as miRNAs, and cause translational suppression of genes with imperfect complementarity. Most siRNAs are designed to have perfect complementarity to a single target.

1.3 MicroRNAs may also regulate protein transcription

MicroRNAs are known to regulate protein translation by binding to partially complementary 3'UTR or CDS of mRNAs. Small RNAs might also regulate mRNAs transcriptionally, either through silencing or activation [5].

Gene silencing has been shown to be a result of small RNAs targeting the promoter of the coding gene [6]. In a previous work, it was shown that miRNAs may silence protein transcription through epigenetic mechanisms [2]. In this work, Kim et al. provided evidence of a cis-regulatory role for miR-320 in transcription of the gene POL3D.[1] They also report that conserved miRNA seed sites were enriched by ~5% within promoters. This result suggests that mammalian miRNAs might regulate transcription *in trans*.

Recently, it has been shown that small RNAs also might activate gene transcription [5]. In this work, they showed that small RNAs targeting a non-coding antisense transcript that is complementary to the targeted coding gene might induce gene activation.

Even though small RNAs can regulate transcription, current literature shows just a few specific examples of this actually happening, and it is unclear whether RNA-based transcriptional regulation is a general mechanism that can affect all genes or, alternatively, what rules determine if a gene can be affected. Furthermore, it is unclear to what extent seed-based targeting drive RNA-based transcriptional gene regulation.

1.4 Cis natural antisense transcripts

Cis natural antisense transcripts (cis-NATs) are RNAs that are perfectly complementary to an RNA on the opposite strand in the DNA. These transcript pairs are referred to as sense-antisense (SA) gene pairs [7]. Also, these SA gene pairs can be further divided into different categories depending on where they are overlapping with each other (figure 1.3). In addition, cis-NATs has been suggested to have regulatory roles [8]. The mechanisms behind this regulation

[1] Known as Polymerase (RNA) III, and is located on chromosome 8 in the human genome.

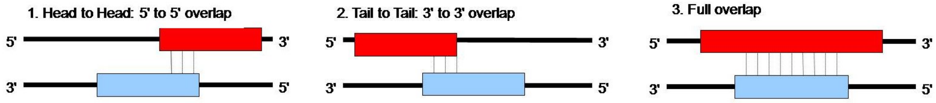

1. Head to Head: 5' to 5' overlap
5'
3'
3'
5'
2. Tail to Tail: 3' to 3' overlap
5'
3'
3'
5'
3. Full overlap
5'
3'
3'
5'

Figure 1.3: The different orientations of SA gene pair overlap (from wikipedia).

are not well understood at this point. Cis-NAT pairs are also relatively common, and 4-9% of all human gene transcripts are overlapping [9]. Zhang et al. searched for cis-NAT pairs in humans, and found 3915 such sense-antisense pairs [8], whereas Engström et al. found 6141 of them [10]. Also, Engström et al. found that only 17% of cis-NAT pairs were conserved between human and mouse, indicating frequent species specificity of antisense arrangements.

There has also been done some GO analyses (for a brief description, see section 3.7) of cis-NAT pairs, and they found that for humans, (1) cellular physiological process is the most enriched BP term, (2) catalytic activity is the most enriched MF term, and (3) intracellular is the most enriched CC term [11].

Chapter 2

Aim of the study

We already mentioned that miRNAs might activate coding genes by binding to antisense non-coding transcripts. Therefore, the hypothesis is that miRNAs might regulate gene protein expression indirectly by binding to the non-coding transcript of cis-NAT pairs. The aim of this study is to investigate if this hypothesis might be true. The subgroup of cis-NAT pairs that will be further examined are the ones that contain exactly one protein coding transcript and one non-coding transcript. In that case, a miRNA might bind to the non-coding transcript that further binds to the coding transcript (as illustrated in figure 2.1). Since previous work has shown that small RNAs binding to non-coding antisense transcripts direct gene activation, we would expect a positive correlation between the expression of the miRNA and expression of the regulated coding gene itself.

The two main aims (research questions) of this project are:

- **RQ1:** to investigate whether the noncoding transcript of cis-NAT pairs, similar to coding mRNAs, show significant enrichment for conserved miRNA seed sites.
- **RQ2:** to correlate miRNA expression with expression of the coding side of targeted cis-NAT pairs.

Each of these two research questions is described in depth in the next two sections.

2.1 RQ1 - investigate whether the non-coding transcript of cis-NATs show significant enrichment for conserved miRNA seed sites

First of all, what do we mean by enrichment? Enrichment is a way to measure the difference ratio between two values. In this case, we are measuring the difference between the number of targets for real miRNA seeds versus control

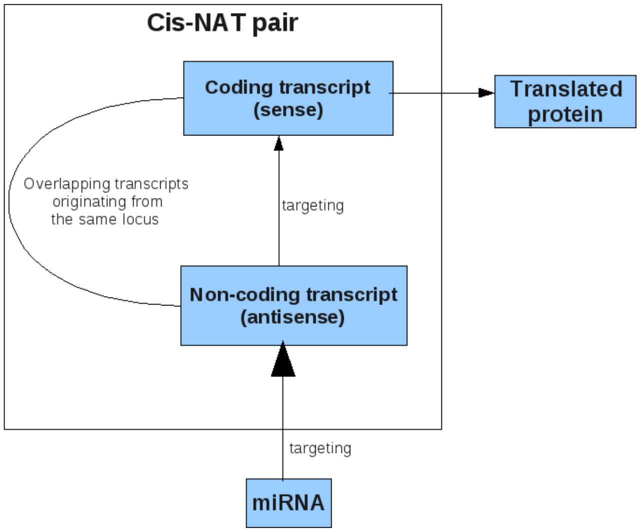

Cis-NAT pair
Coding transcript (sense)
Translated protein
Overlapping transcripts originating from the same locus
targeting
Non-coding transcript (antisense)
targeting
miRNA

Figure 2.1: The subgroup of cis-NAT pairs to be analyzed in this study. The miRNA might bind to the non-coding antisense transcript of cis-NAT pairs, and thus regulate expression of the sense coding transcript.

seeds. When miRNA seeds have more hits within a group of transcripts than control seeds have, we have found indications of predicted targets. Thus, the control seeds represent the false positives, whereas the miRNA seeds represent the predicted targets. Therefore, the control seeds should be made in such a way that they represent what would be expected by just the chance alone. This is a common way to estimate if miRNAs are targeting a transcript [12]. So, for this research question, why would it be interesting to observe an enrichment for conserved miRNAs within cis-NAT pairs? Because we would expect a significant enrichment within cis-NAT pairs if they are actually being targeted. Thus, a positive enrichment would mean that we have found predicted targets.

This research question may also be divided into two sub questions. With respect to the non-coding transcript of cis-NAT pairs,

1. are seeds from highly conserved miRNAs more enriched than seeds from less conserved miRNAs?
2. are highly conserved seed sites more enriched than less conserved seed sites?

The above questions are relevant for several reasons. First, seeds from highly conserved miRNAs are expected to give a higher enrichment than less conserved seeds. The reason for this is that highly conserved miRNAs would be under greater selective constraint, and therefore less likely to change during the course of evolution[4]. Second, conserved seed sites are more likely to have an important function in the body than the less conserved ones. Therefore, an enrichment for conserved seed sites means that the miRNAs are targeting an actual function in the genome.

2.2 RQ2 - correlate miRNA expression with expression of the coding side of cis-NAT pairs

If and when miRNAs target the non-coding side of cis-NAT pairs, we would expect a significant positive correlation between the miRNA expression and expression of the targeted coding gene itself. The following experiment could answer if there is such a correlation: for each tissue in the human body, correlate the miRNA expression to expression of both targeted and non-targeted cis-NAT pairs. We would expect a difference between the correlations in the targeted and non-targeted group if miRNAs are actually regulating genes through binding to cis-NAT pairs. We would expect the correlation between a miRNA and its non-targets to be zero or close to zero, whereas correlation to the target group would be expected to be farther away from zero than the non-target group. This difference in correlation values could be tested with a statistical method, which could answer if the difference is strong enough to be significant.

Chapter 3

Methods

This chapter will explain the choice of algorithms, implementations and statistical methods for answering RQ1 and RQ2 mentioned in chapter 2. This chapter will review the functionality needed for answering both RQ1 and RQ2 (section 3.1). There will also be a review of functionality needed only for answering RQ1 (section 3.2), and functionality needed only for answering RQ2 (section 3.3). The remaining sections review functionality used for answering both RQ1 and RQ2.

3.1 Functionality needed for answering both research questions

The needed functionality that is shared for both research questions is listed here:

- **conservation filter -** it should be possible to apply a conservation filter when receiving sequences from transcripts or genes. For example, it should be possible to get all nucleotides conserved among human, mouse and rat. This can be achieved by using the external source Multiz17way, which is described more in depth in section 3.5.

- **region filter** - it should be possible to apply a region filter when receiving a sequence from a transcript or gene. For example, it should be possible to extract only specific parts of a cis-NAT sequence, for example the 5'UTR, 3'UTR or CDS.

- **get transcript sequence** - for a specific transcript, it should be possible to extract its sequence. Also, it should be possible to apply the conservation filter or region filter described above.

- **get the transcripts of a cis-NAT pair -** given a cis-NAT pair, it should be possible to extract its two transcripts (both the coding and non-coding transcript).

- **filter cis-NATs -** it should be possible to filter cis-NAT pairs based on where the cis-NAT transcripts are overlapping with each other. The three different categories are convergent, divergent, and complete overlaps.
- **cache decorator -** it should be possible, in an easy way, to cache the result from intensive methods, and thus make these methods faster to execute at a later point.
- **reverse complement a genomic sequence -** it should be possible to get the reverse and/or complement of a sequence.

3.2 Functionality needed for answering RQ1

This system is going to try to answer the two sub questions regarding seed enrichment of cis-NATs as mentioned in section 2.1. This is the basic functionality needed in the system:

- **get conserved miRNA seeds -** retrieve conserved miRNA seeds according to a given conservation threshold. It should be possible to retrieve specific sub sequences of miRNAs, for example the seed positions 2-8.
- **generate comparable miRNA control sequence -** make a list of comparable control sequences for each mirna sequence.
- **calculate the enrichment and p-value -** count the number of predicted targets for both real miRNAs and controls. By comparing these two values, it is possible to calculate the enrichment. Also, by doing this comparison multiple times (resamplings), it is possible to estimate the p-value. This p-value makes it possible to see if the enrichment is significant.

Figure 3.1 shows a simple overview of how the system is calculating enrichments and p-values for conserved seed sites. The top three components represent the external data sources.

The left component at the second row (from the top) retrieves conserved cis-NAT sequences. We may for example ask for all nucleotides that is conserved among human, mouse and rat. The middle component at the second row retrieves unconserved cis-NAT sequences. The right box at the second row retrieves conserved and unique miRNA seeds. We may for example ask for miRNA seeds by specifying a conservation threshold of 20. This component returns a list of seed sequences where duplicated seed sequences are removed, so as to eliminate any biases toward the most occurring seed sequences.

The left component at the third row calculates the number of matches for controls and miRNAs within conserved cis-NAT sequences. The right component on the third row generates control seeds. This component creates controls based on the actual miRNAs and the unconserved cis-NAT sequences, as the pointers suggests.

The bottom box estimates enrichment and p-values of miRNAs within cis-NAT sequences.

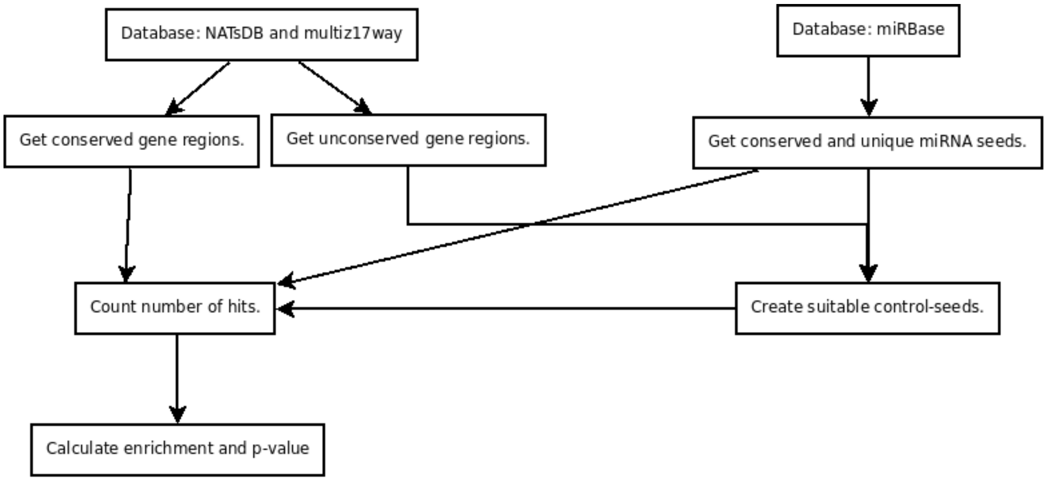

Database: NATsDB and multiz17way
Database: miRBase
Get conserved gene regions.
Get unconserved gene regions.
Get conserved and unique miRNA seeds.
Count number of hits.
Create suitable control-seeds.
Calculate enrichment and p-value

Figure 3.1: How the enrichments and p-values of conserved seed sites within non-coding side of cis-NATs are calculated.

Here is a more detailed description of how the various components of the system work:

- **get conserved miRNA seeds** - this function finds conserved miRNA seeds based on a conservation-threshold. Conserved miRNA seeds are found by looking for seeds that is shared among many different species. For example, the 7mer-m8 seed with the sequence GGAAGAC is shared among miRNAs from 40 different species (as of miRBase v12.0). The idea is that highly conserved miRNA seeds has an important function in the genome compared to the less conserved ones.

- **Build a list of suitable shuffled seed-sites.** In this project, we are counting the number of predicted targets for different miRNA seed sites. Since these hits also can contain some false positives, it is important to have a control to compare to. This component builds a list of control seed sequences that share the same properties as the original miRNA seed sequences, as described more in depth in section 3.2.1. It is very important that these controls share the same properties as the miRNAs, since otherwise the controls may give an artificially high or low number of predicted targets (i.e. artificial enrichment).

- **Count the number of hits within a sequence for both miRNAs and controls.** This function counts the number of hits for both groups. The hits are found by looking for perfect Watson-Crick (W-C) seed matches. The number of hits for controls represent the random hits (i.e. the noise), and the miRNAs represents the predicted targets (i.e. the signal).

- **Estimate enrichment and p-value.** The enrichment is found by comparing the number of hits for miRNA sequences against the number of hits for control sequences. The number of hits for controls are averaged using resampling. This resampling involves counting hits for controls several times (n=1000), and then estimate the mean value and squared error. By dividing the number of hits for miRNAs by the average number of hits for controls, we get the enrichment. A high and significant enrichment indicates that we have found authentic targets for miRNAs within human genes. The p-value is calculated by finding the proportion of controls that is located above the miRNAs, in terms of hits. A p-value lower than a given confidence interval (e.g. $p<0.05$) means that the observed enrichment is significant.

3.2.1 How the shuffled controls are built

To estimate enrichments and p-values, the system needs to build a list of control seeds. The way to make such control seeds is to find permutations of the miRNA-seeds that share the same properties as the original miRNAs. The properties that are used are the 0- and 1-order Markov properties, as well as the number of hits for each miRNA seed site within the sequence that we are searching for targets within.

To illustrate, assume that we have the sequence "ACCGTCG". Table 3.1 shows the 0-order Markov matrix for the mentioned sequence. This table basically shows the probability of observing a specific nucleotide in the sequence. We can for example see that the sequence has got more C's than A's. This matrix can be used to estimate the probability of observing a sub sequence in the sequence. For example, the chance of observing the sub sequence "AC" in the mentioned sequence is:

$$P_{0-order}("AC") = P("A") * P("C") = 0.143 * 0.429 = 0.061$$

For the above formula to be valid, the probability distribution of each position needs to be independent from the other positions. It is important that the generated controls share this property. We see that the sub sequence „TC“ has the same 0-order probability of occurring in the sequence "ACCCGTCG" as the sub sequence "AC". Therefore, "TC" is a valid control with respect to the 0-order Markov properties.

In table 3.2, we see the 1-order Markov properties for the sequence "ACCCGTCG". We see that the probability of observing for example the sequence „ACC“ in the sequence "ACCCGTCG" is:

$$P_{1-order}("ACC") = P("AC") * P("CC") = 0.17 * 0.17 = 0.028$$

All valid controls need the same probability of occurring in the sequence as well, so a valid control sequence would be for example „GTC“.

The last property that is taken into account, is the number of hits; i.e. how many non overlapping matches does each seed site have in the sequence.

Nucleotide	A	C	G	T
Probability	0.143	0.429	0.286	0.143

Table 3.1: The 0-order Markov properties for the sequence „ACCGTCG“.

Nucleotide	A	C	G	T
A	0	0.17	0	0
C	0	0.17	0.33	0
G	0	0	0	0.17
T	0	0.17	0	0

Table 3.2: The 1-order Markov properties for the sequence „ACCGTCG“

For example, the sub sequence „AC“ has got 1 hit in the mentioned sequence „ACCGTCG“. This means that for example the control sequence „GA“ is not a valid control seed, since it has no hits in the mentioned sequence.

In addition to evaluate the Markov properties and the number of hits, we also need to introduce a slack when choosing comparable controls. This slack is necessary to be able to find enough comparable control seeds for each miRNA seed.

3.2.2 Slack-variable for the generated controls affects the signal

As stated in section 3.2.1, the controls are made so that they share the same properties as the original miRNA seeds in terms of occurrences, 0-order- and 1-order Markov properties in addition to the number of hits. To find enough suitable controls for each miRNA, it is necessary to introduce a slack-variable. This slack-variable means that controls may be accepted even if they are not a "perfect candidate“, as long as all their properties lay in between this slack-margin.

A problem with using this slack-variable is that it is also affects the signal. Assume that we use a slack-value of 7.5% when creating suitable controls. For a miRNA that has 100 seed hits, a control must lay in between the interval 93 to 107 hits to be considered suitable. For controls that are skewed toward for example the lower side of the interval 93-107, the actual miRNA seeds may give up to 7.5% enrichment without being considered significant. So, a higher slack-value gives rise to more noise in the measurements. Off course, this depends on the distribution of the controls, but this is just to illustrate the worst case scenario. In that sense, it is wise to keep the slack-variable as low as possible, so that we are not drowning the signal in this noise. However, a too small slack-variable may give too few suitable controls for each miRNA seed. So in practice, +/- 7.5% seems like a reasonable slack-value, and is the one used in this system.

3.3 Functionality needed for answering RQ2

The specific functionality needed only for RQ2 is listed here:

- **extract the expression profile of cis-NATs and miRNAs -** it should be possible to extract the expression profile of both cis-NAT pairs and miRNAs.

- **correlate miRNA expression to expression of targets and non-targets**

- **determine if a gene belongs to a cis-NAT pair** - given a gene, it is interesting to know if it belongs to a cis-NAT pair or not. This functionality might help to answer if gene that belongs to cis-NAT pairs has some other properties than regular genes.

In figure 3.2, one can see how the system is correlating expression of miRNAs with expression of transcripts. In this case, the transcript to be correlated with is the coding side of cis-NAT pairs.

3.4 Statistical analysis

The statistical methods used in this study are:

- **Wilcoxon rank sum test**. This test was used to compare the correlation between gene expression and miRNA expression of two different groups of genes. The first group is genes having a miRNA target, and the second group is the non-target group. In this case, our hope is that the noise is low enough so that the Wilcoxon test can pick up on the difference between these two groups. This test was run in R, using default settings, except the ”less” parameter.

- **Pearson and Spearman correlation**. These tests was used to correlate the expression of genes and expression of miRNAs over all human tissues. The tests was run in R, using default settings.

- **Multiple correction adjustment.** When analyzing multiple samples in terms of p-values, it is necessary to adjust the confidence interval accordingly. Sidak correction has been used to measure if p-values were significant or not. For example, lets say that we chose a confidence interval of 0.05. In that case, a p-value below 0.05 would be considered significant. If we are doing 20 such tests, we would expect that on average one of these 20 tests have a p-value below or equal to 0.05 even if the null-hyphotesis is true (Type 1 error). Said in another way, multiple comparisons are increasing the chance of doing a Type 1 error, meaning rejecting the null hypothesis when it should not be rejected. That is why we need to use such correction methods when doing multiple hypothesis testing.

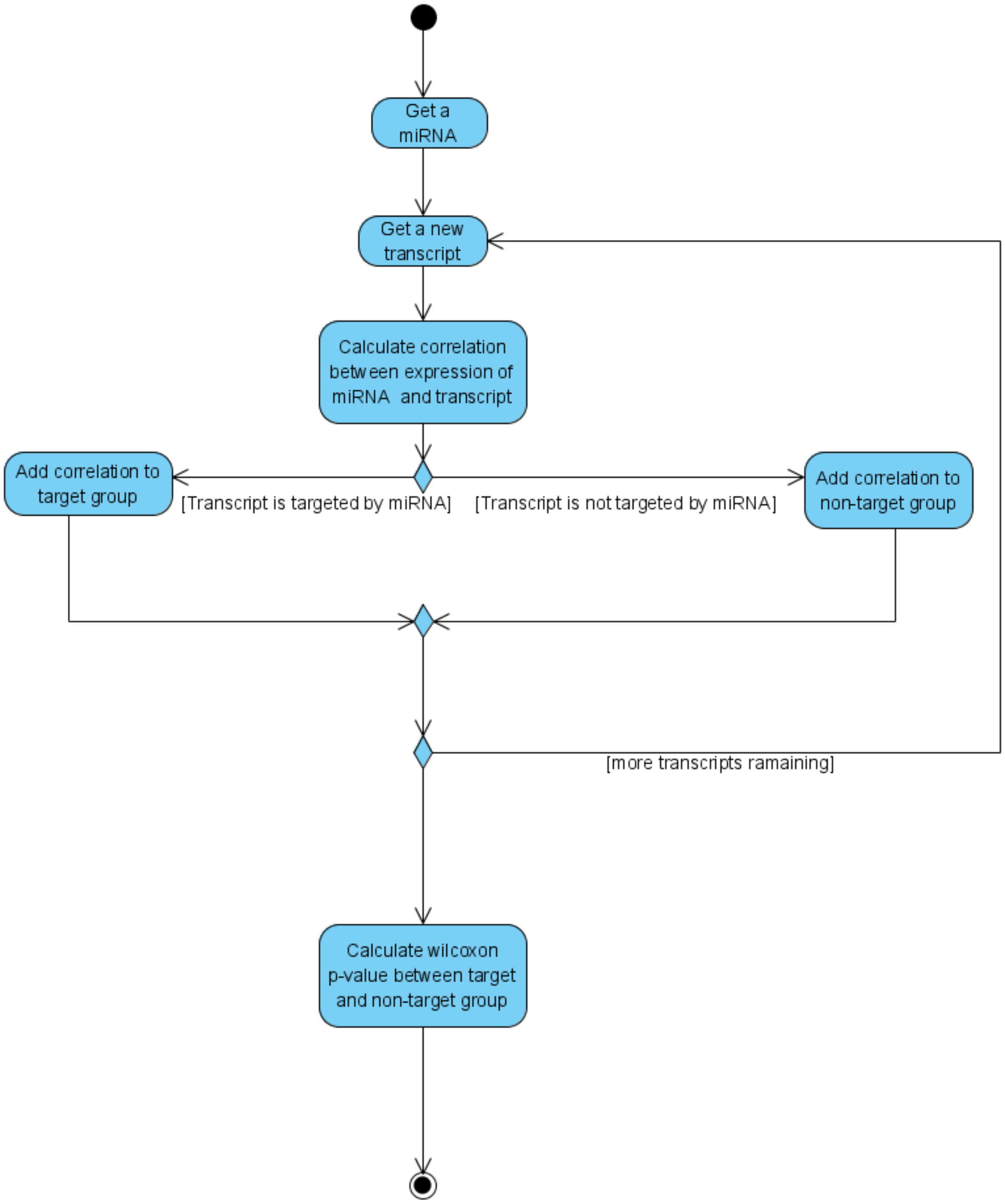

Get a miRNA
Get a new transcript
Calculate correlation between expression of miRNA and transcript
Add correlation to target group
[Transcript is targeted by miRNA]
[Transcript is not targeted by miRNA]
Add correlation to non-target group
[more transcripts ramaining]
Calculate wilcoxon p-value between target and non-target group

Figure 3.2: Activity diagram showing the basic way of correlating expression of miRNAs with expression of targeted transcripts.

- **Sidak correction.** This is one correction method used in this project. The formula for adjusting a p-value with this method is: $Sidak(p) = 1 - (1 - \alpha)^{1/n}$, where the alpha is the confidence interval, and ”n” is the number of tests.
- **Benjamini and Hochberg correction.** This is another correction method made by Benjamini and Hochberg [13]. This method adjusts for the false discovery rate, and might therefore be a better choice than Sidak. This method will be used during this project as an alternative to the Sidak correction. The Benjamini and Hochberg correction method was run through the R function ”p.adjust”, using ”BH” as the method parameter.

3.5 Implementation details

This section will describe implementation details, such as the programming language, revision control external packages, external data sources, database, unit testing and design patterns used in the system.

Python as a programming language

The programming language chosen for this project is Python. There are two main reasons for using Python in this system. The first reason is that Python has a long tradition in being used for solving biology types of problems, which makes it a lot easier to extend the system with packages made by others. The second reason is that it is an interpreted language, which means that less code is generally needed to do a specific job in Python than in other languages like C or Java, partly because there is no need for datatype declaration.

As mentioned above, many packages for solving biology-tasks in Python already exists. One example of such a package is BioPython, which contains functionality for handling sequences and parsing different formats containing biology data. Other examples of such Python-packages are FindSuitableRandomSeeds and GeneLoader (see section 3.5).

One drawback of using Python in many applications (including mine) is that it may be slower (in terms of computational time) than other languages when doing very intensive calculations. This is because Python is an interpreted language, and therefore has some overhead compared to faster languages like C or Java. Benchmarks show that C is ~5 times faster than Python when doing reverse-complement calculations, and may also be up to 371 times faster than Python on Mandelbrot-calculations when using an Intel Q6600 on the Debian OS [14]. However, it should be mentioned that Python can be extended with precompiled packages from other faster languages like C, and thus lower the computational time of the most intensive calculations. One example of such a package is NumPy, which contains code for doing fast matrix operations.

Subversion as the revision control

Revision control system makes it possible to revert back to a previous version of the code at any time, and also acts as a backup for the source code in case of an accident. Subversion was used in this system.

External packages used in this system

A package basically means a collection of files containing source code. The external packages used in this system, are:

- **FindSuitableRandomSeeds**. Pål Sætrom has made this package in Python. This package can find suitable shuffled seeds based on a list of miRNA-seeds and a so called "flatfile" which contains the sequences that we want to count targets within. The output is a list of suitable shuffled seeds that shares the same properties (see section 3.2.1) as the original seeds. To use this package, it was necessary to do some small adjustments in the code. Because in my system I needed not just a long list of suitable controls, but also the mapping to their corresponding miRNA seeds. In that way, I am eliminating any possibility for biases toward some of the miRNA seeds.

- **GeneLoader.** Erik Gorset has made this package in Python. This package can retrieve nucleotides for one or many genes based on specified options. It is also possible to retrieve specific regions, such as the coding region (CDS), and 5' and 3'UTRs. GeneLoader can also retrieve nucleotides conserved in a specified number of species. The output from this program can be written to file.

External data sources

This project used multiple external data sources. The data sources RefSeq, Multiz17way, and mRNA expression profile were available in SQL, making it easy to incorporate them into the existing system. NATsDB and the miRNA expression profile, on the other hand, were only available in Microsoft Excel files, and is was therefore necessary to convert these two files to SQL databases before they could be used in the system.

- **UCSC RefSeq (HG18)[1].** This is a freely available database that contains positional information about human genes. This database is updated regularly, but the results obtained in this report was gathered using the as of version 14 November 2008 of this database.

- **Multiz17way.** This database gives information about the evolutionary conservation among 17 different vertebrates for each distinct region of the chromosome. It is giving a specific conservation-score to each distinct region. This database has been made with help of the tools Multiz [15],

[1] **genome.ucsc.edu**

Blastz [16, 17], and PhastCons. More details about the creation of Multiz17way can be found on the UCSC-site.

- **miRBase, v12.0.** This database contains information about the currently known miRNAs, such as their names, family, stem loop structure and mature sequences [18, 19, 20]. Some of the miRNAs contain several alternative mature sequences, and in that case each one of them is shown in the database. A screenshot from this site is shown in figure 3.3.
- **NATsDB.** A excel file that contains cis-NAT pairs [11, 21]. Since NATsDB is originally stored in the old hg17 format, it was necessary to convert it to the newer hg18 format with the tool liftover first. Liftover is a tool online (`http://genome.ucsc.edu/cgi-bin/hgLiftOver`) that converts genome coordinates and genome annotation files between different releases of the UCSC genome database.
- **miRNA expression.** A excel file with expression profiles of miRNAs was gathered from a previously published paper [22]. This file contains information about the expression of miRNAs across 172 different human cell lines and tissues. Since this file is stored in a excel file, it was necessary to convert it to a database table first.
- **mRNA expression.** This data was gathered from the database GnfAtlas2, which can be found on the UCSC site online (`http://genome.ucsc.edu/`). This database contains expression profiles of mRNAs across 79 different human tissues.
- **Merged mRNA and miRNA expression.** To correlate mRNA and miRNA expressions, it was necessary to merge them together first. Of the 79 tissues inside the mRNA profile and the 172 tissues inside the miRNA profile, only a total of 19 tissues were overlapping (see table 3.3). Therefore, only these 19 overlapping tissues were used in the correlation study.

MySQL as a database

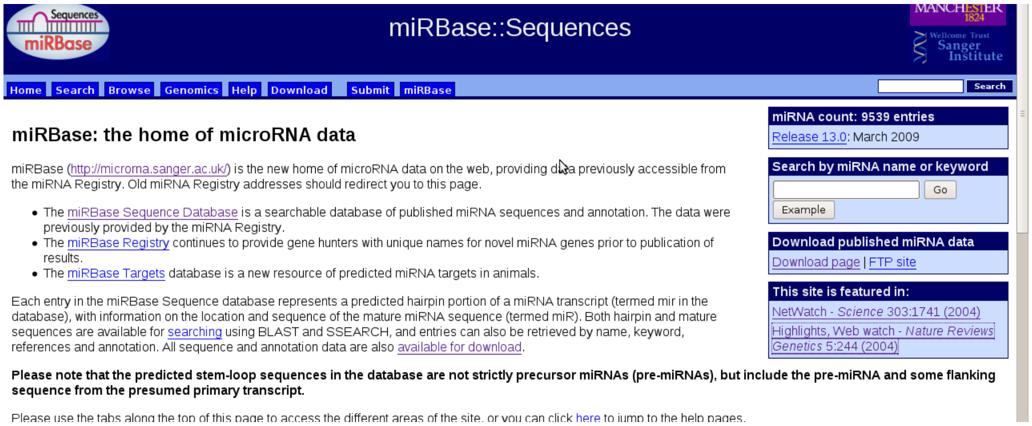
My application should be able to connect to different external sources containing biological information, such as the mentioned RefSeq, Multiz17Way and miRBase database. In this case, a MySQL database seems to fit perfectly for serving my application with biological information. MySQL is open-source, and in addition it is easy to find answer to questions you get on your way, since it is a popular and widely used database system.

The statistical software R

This software has been used extensively for doing statistical calculations and making plots of the results.

miRNA expression id	mRNA expression id
hsa_B-cell-CD19	PC-CD19+ Bcells
hsa_Burkitt_Raji	lymphoma Burkitts Raji
hsa_Cerebellum-adult	cerebellum
hsa_Heart	heart
hsa_HSC-CD34	BM-CD34+
hsa_liver	liver
hsa_medullobl-DAOY	medulla oblongata
hsa_Monocytes-CD14	PC-CD14+ monocytes
hsa_NK-CD56	PB-CD56+ NKCells
hsa_Ovary	ovary
hsa_Pancreatic-islets	pancreatic islets
hsa_Pituitary	pituitary gland
hsa_Placenta	placenta
hsa_Prostate	prostate
hsa_T-cell-CD4	PB-CD4+ Tcells
hsa_T-cell-CD8	PB-CD8+ Tcells
hsa_Testis	testis
hsa_Thyroid	thyroid
hsa_Uterus	uterus

Table 3.3: The 19 tissues that were overlapping between the miRNA profile and the mRNA profile. These are the tissues that were used in the correlation study.



miRBase: the home of microRNA data

miRBase (http://microrna.sanger.ac.uk/) is the new home of microRNA data on the web, providing data previously accessible from the miRNA Registry. Old miRNA Registry addresses should redirect you to this page.

- The miRBase Sequence Database is a searchable database of published miRNA sequences and annotation. The data were previously provided by the miRNA Registry.
- The miRBase Registry continues to provide gene hunters with unique names for novel miRNA genes prior to publication of results.
- The miRBase Targets database is a new resource of predicted miRNA targets in animals.

Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript (termed mir in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature sequences are available for searching using BLAST and SSEARCH, and entries can also be retrieved by name, keyword, references and annotation. All sequence and annotation data are also available for download.

Please note that the predicted stem-loop sequences in the database are not strictly precursor miRNAs (pre-miRNAs), but include the pre-miRNA and some flanking sequence from the presumed primary transcript.

Please use the tabs along the top of this page to access the different areas of the site, or you can click here to jump to the help pages.

miRNA count: 9539 entries

Release 13.0: March 2009

Search by miRNA name or keyword

Download published miRNA data

Download page | FTP site

This site is featured in:

NetWatch - *Science* 303:1741 (2004)

Highlights, Web watch - *Nature Reviews Genetics* 5:244 (2004)

Figure 3.3: The miRBase site.

RPy2

RPy2 is a Python interface to the statistical software R. RPy2 has been used extensively during this project.

Unit testing

Unit testing (xUnit) was used extensively during the development of this application. Unit testing makes it possible to test the system after each modification, so as to make sure that the modification did not introduce any new bugs to the system. This is especially important in this system, since it is a subject of rapid changes over time.

Test driven development

Test driven development has been used during this project. The usage of this development style has made it easier to refactor the code while ensuring the correctness of the system. Code coverage tools has also been used. With these tools, it is possible to find regions within the code that is not well covered with unit tests. These regions of no coverage is very often a weak spot, so it is beneficial to improve the coverage of the code as much as possible.

Design patterns

In this project, it has been crucial to take advantage of design-patterns, since so many different components is communicating together. The usage of design patterns also makes it a lot easier to make modification or extensions to the system at any point.

Decorator pattern - used to cache the output from intensive methods

Many parts of the system contains very cpu-intensive functions, which often takes many minutes to complete. In addition, these functions are often called with the same parameters on two different runs. Therefore, caching the results from these functions to disk would save much time.

One way to cache the output from intensive functions is to just modify the involved functions, so as to incorporate cache functionality into them. But this will eventually evolve into a maintenance nightmare, and it turns out that there exists a design pattern that fits nicely to this situation. The name of this design pattern is Decorator pattern. This pattern can decorate all intensive functions with cache functionality without altering the original function itself. Thus, this pattern ensures a clear separation between an intensive function and the cache-functionality itself.

One example of a situation in my system where the decorator is used is when retrieving sequences of genes. There is often the need to return these sequences over and over again, and in this case the decorator pattern fits nicely.

GRASP principles

The GRASP principles are a collection of design principles on how to do structured development [23]. This system has taken advantage of these principles, since they are making it easier to keep the code structured and well organized. Some of these principles are:

- **Low coupling**, which means that we want have as few dependencies between classes or modules as possible. Low coupling is making it easier to maintain, read, understand and modify the source code.

- **High cohesion**, which means that we want to keep the task of each class or module as centered as possible. For example, a class which contains 10.000 lines of codes is probably doing too much unrelated work, and therefore should be divided into smaller classes that is doing more specific tasks.

- **Protected variations**, this is also sometimes referred to as „encapsulate what varies". It means that whenever we see a part of the system that is going to change frequently, it is a wise idea to place a protecting "frame" around that part. In that way, other parts of the system is not affected by these changes.

- **Information expert.** This principle answers which class should be assigned a new task. It states that a task should be assigned to the class which has the necessary information to perform it. The usage of this principle is, as a bi effect, supporting low coupling and high cohesion.

- **Creator.** This principle solves the problem of who should be responsible for the creation of new instances of a class.

- **Controller.** This principle assigns the responsibility of dealing with system events to a non-UI (User Interface) class that represent the overall system or a use case scenario.

3.6 Database schema and SQL Alchemy

Figure 3.4 shows a plot of the different databases that were used during this project, and how they connect together. Here is a detailed description of the tables used:

- **Mirna, MirnaToMature, and Mature** - these are the tables from the miRBase site online.

- **ExperimentIds -** gives a list of the 79 different tissues of the expression profiles.

- **MirnaToExperiment -** basically a many-to-many jointable between the Mature table and the ExperimentIds table. Gives information about the expression of miRNAs in the corresponding tissues.

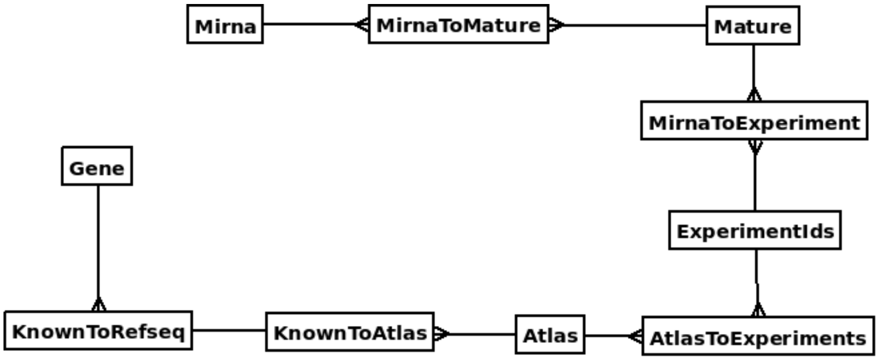


Figure 3.4: The database tables used in this project, and how they connects together.

- **Atlas -** this table contains the GNF 2 expression profile of mRNAs.
- **AtlasToExperiments -** another many-to-many join table, which gives the expression score of mRNAs in each of the 79 different tissues.
- **KnownToAtlas -** a table that maps microarray probes to a known gene id.
- **KnownToRefSeq -** a table that maps known gene ids to refseq ids.
- **Gene -** a table that contains all the refseq genes.

The object-relational mapper SQL Alchemy (`www.sqlalchemy.org`) has been used extensively during this project. SQL Alchemy makes it possible to communicate with the database by using objects instead of through native SQL queries. Thus, SQL Alchemy is hiding much of the complexity involved in doing SQL queries.

3.7 Gene ontology analysis

Gene ontology (`www.geneontology.org`) is a structured gene annotation database that groups different genes together based on whether they are related or not. Genes can be related in three different ways:

1. **Cellular component.** Classifies the product of genes according to which part of the cell they would be a part of. Examples of such cellular components is membranes or organelles.
2. **Molecular function.** Classifies the product of genes according to which elemental activity it would have. Examples of such molecular functions is binding and catalysis.

3. **Biological process.** Classifies the product of genes according to which operations or sets of molecular events they are a par of.

Gene ontology analysis in this project were done using the R package GOStats, available in Bioconductor (`www.bioconductor.org`). The test named GOHyperGParams was used, together with the Affymetrix Human Genome U133 Plus 2.0 Array annotation data (hgu133plus2.db). The mentioned test GOHyperGParams also takes a ”universe” as an input parameter. A universe is the same as the background set of genes that we are testing against. For the tests done during this work, the default universe was chosen. This means that all the genes of the array annotation data file is the universe.

Chapter 4

Results and discussion

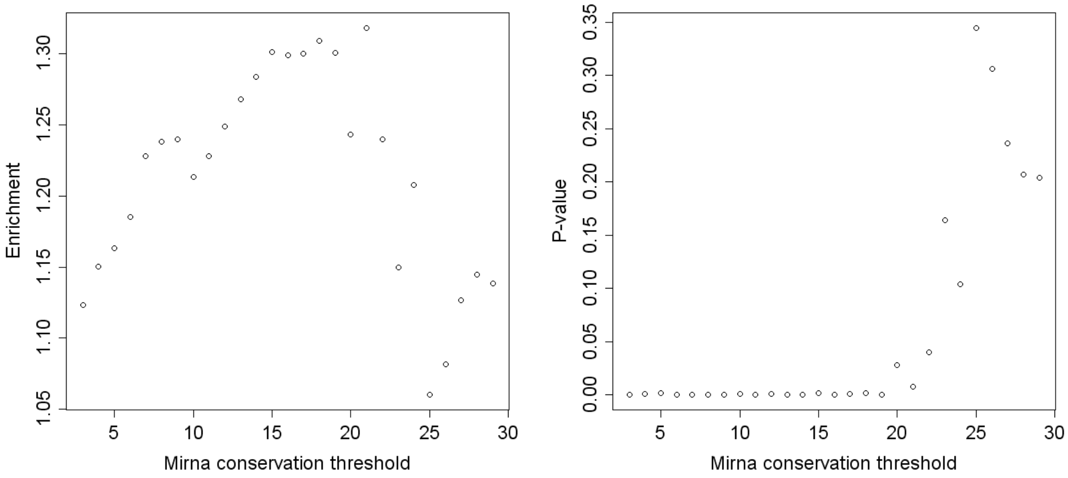
This chapter will describe all the results that were made during the project. This chapter also discuss various aspects of the results that were found.

4.1 Significant enrichment of conserved miRNA seeds within non-coding part of cis-NATs

Previous work [3] revealed that highly conserved miRNAs have a high number of predicted targets within 3'UTR of mRNAs. In this section, I am calculating the enrichment of conserved miRNA seeds within the non-coding side of cis-NAT pairs. The experiment was done by looking for seed sites conserved among human, mouse, rat, dog and chicken. More specifically, we are looking for evidence of predicted targets within such non-coding antisense transcripts.

We see that highly conserved miRNA seeds are more enriched than less conserved miRNA seeds (figure 4.1). This corresponds to what has been found for miRNA seed sites within 3'UTR of mRNAs [4]. One interesting observation is that for very conserved miRNAs, conserved among at least 20 different species, the enrichment starts to decrease again. Thus, it seems like moderately conserved miRNAs, conserved among approximately 15 different species, are the ones that are most significantly enriched.

In total, there are 2094 different non-conserved 7mer-m8 seeds as of miRBase v12.0. Of these 2094 miRNA seeds, 79 are conserved in miRNAs from at least 15 different species. When looking for seed sites conserved among human, mouse, rat, dog, and chicken, 851 hits where found for these 79 miRNA seeds. The control seeds gave 654 hits, which means that the 79 miRNA seeds have an average of 197 predicted targets (851-654). Thus, each miRNA seeds has an average of about 2.5 predicted targets. The predicted targets in this context is the antisense non-coding part of cis-NAT pairs. Thus, each miRNA seed conserved among at least 15 different species is targeting an average of 2.5 non-coding antisense transcripts.



(a)

(b)

Figure 4.1: Enrichments (a) and p-values (b) for seed sites within non-coding side of cis-NAT pairs. Only seed sites conserved among human, mouse, rat, dog and chicken is taken into consideration. P-values are calculated by running multiple (n=1000) resamplings for the control seeds.

GO id	GO term	P-value
GO:0019899	enzyme binding	4.68e-05
GO:0005515	protein binding	7.65e-05
GO:0043167	ion binding	1.96e-04
GO:0005516	calmodulin binding	2.19e-04
GO:0003779	actin binding	3.64e-04
GO:0004725	protein tyrosine phosphatase activity	3.85e-04
GO:0004723	calcium-dependent protein serine/threonine phosphatase activity	5.04e-04
GO:0005509	calcium ion binding	5.48e-04
GO:0016791	phosphatase activity	7.52e-04
GO:0005345	purine transmembrane transporter activity	1.48e-03
GO:0035259	glucocorticoid receptor binding	1.48e-03
GO:0005159	insulin-like growth factor receptor binding	2.13e-03
GO:0019904	protein domain specific binding	2.67e-03
GO:0005262	calcium channel activity	3.60e-03
GO:0005070	SH3/SH2 adaptor activity	3.94e-03
GO:0005543	phospholipid binding	3.98e-03
GO:0016887	ATPase activity	4.11e-03
GO:0031267	small GTPase binding	4.18e-03
GO:0001515	opioid peptide activity	4.81e-03

Table 4.1: Over represented molecular function GO terms with p-value below 0.005.

4.2 Gene ontology analysis of targeted cis-NAT pairs

It would be interesting to analyze the gene ontology (GO) of the coding side of targeted cis-NAT pairs. This information could tell us if some GO terms are over-represented among the targeted cis-NAT pairs, and give some information about the functions of these genes.

In this experiment, the coding transcript of cis-NAT pairs is analyzed for gene ontology functions. To be considered a targeted cis-NAT pair, the pair has to be targeted by at least one conserved miRNA seed. Only miRNAs conserved among at least 15 different species were used in the experiment. Also, only target seed sites conserved among human, mouse, rat, dog and chicken were counted. In table 4.1, we see that ”enzyme binding” is the molecular function with most significant enrichment. The most enriched cellular component ontology is ”intracellular part” (table 4.2). Finally, the most enriched biological process ontology is ”cellular protein complex assembly” (table 4.3).

GO id	GO term	P-value
GO:0044424	intracellular part	3.5e-03
GO:0030027	lamellipodium	1.5e-02

Table 4.2: Over represented cellular component GO terms with p-value below 0.005.

GO id	GO term	P-value
GO:0043623	cellular protein complex assembly	6.77e-05
GO:0034621	cellular macromolecular complex subunit organization	8.81e-05
GO:0006928	cell motility	2.94e-04
GO:0043933	macromolecular complex subunit organization	3.63e-04
GO:0006461	protein complex assembly	6.52e-04
GO:0051179	localization	1.09e-03
GO:0016043	cellular component organization and biogenesis	1.27e-03
GO:0018242	protein amino acid O-linked glycosylation via serine	1.59e-03
GO:0018243	protein amino acid O-linked glycosylation via threonine	1.59e-03
GO:0050794	regulation of cellular process	1.73e-03
GO:0045664	regulation of neuron differentiation	2.46e-03
GO:0007399	nervous system development	2.68e-03
GO:0065007	biological regulation	4.04e-03
GO:0048167	regulation of synaptic plasticity	4.60e-03
GO:0008064	regulation of actin polymerization and/or depolymerization	4.97e-03

Table 4.3: Over represented biological process GO terms with p-value below 0.005.

4.3 3'UTR of sense transcripts within cis-NAT pairs are longer and more conserved than other transcripts

The work done so far in this project has suggested that miRNAs might target cis-NAT pairs as a way of regulating coding gene expression. I also did another experiment, to see if the number of targets for cis-NATs changes when having an additional hit in its 3'UTR region. This experiment could give some answer to whether miRNAs has weaker binding power to a cis-NAT transcript when having this additional hit in the 3'UTR.

The result from this experiment was that the 3'UTR of the coding transcript of cis-NATs has more conserved miRNA seed sites than other genes. Additional experiments showed that this different is caused by a difference in the conservation of the 3'UTR of cis-NATs compared to other genes. In addition, the conserved nucleotides in cis-NAT genes seems to be more collected in blocks, rather than spread evenly through the genome.

The 3'UTR of the sense coding transcript of cis-NAT pairs is also longer than other transcripts. Of 2048 sense coding cis-NAT transcripts, the average 3'UTR length is 1281.49. The average 3'UTR length of 2048 random transcripts is 1093.58. This means that the cis-NAT sense transcripts are on average 17.18% longer than random transcripts ($p<0.001$). The p-value was estimate by doing multiple ($n=1000$) resamplings. An important fact about 3'UTR lengths is that they are not following the normal distribution. Therefore, the median is a better estimator. The median length of the 3'UTR of sense coding transcripts was 865, whereas random transcripts had an median of 668.49. Thus, the median coding cis-NAT transcript is 29.4% longer than a random transcript ($p<0.001$; found through multiple resamplings).

3'UTR of cis-NAT pairs is also more conserved than 3'UTR of other transcripts. Analysis of 3'UTRs from the coding gene of cis-NAT pairs showed that on average, 13% of all nucleotides are conserved among human, mouse, rat, dog and chicken (HMRDC). This conservation-rate of 13% was compared to the conservation rate of random genes through multiple resamplings ($n=1000$). The conservation-rate for random genes was 11% (SD=0.64%). This gave a significant ($p < 0.001$) higher conservation of ~18% for the coding transcript of cis-NAT pairs compared to random genes. As an example, assume that the average coding cis-NAT transcript has a 1000 nt (nucleotide) long sequence. Of these 1000 nts, 130 of them would be expected to be conserved among HMRDC. The average random gene would then be expected to have 110 conserved nts among these species.

4.4 Expression of miRNAs often correlates with expression of targeted mRNAs

The main goal is to correlate expression of miRNAs to expression of cis-NAT pairs. But it would we wise to correlate miRNA expression to mRNAs first, since mRNAs have been proved to actually be regulated by miRNAs. If the correlation between miRNAs and mRNAs is not making any sense, it would probably be even more difficult to get any sense from the cis-NAT correlations, since cis-NATs were shown to yield much weaker enrichments than mRNAs. In my work, I got a signal:noise ratio of up to about 1.3 for cis-NAT pairs (according to section 4.1), whereas 3'UTR of mRNAs yields a much stronger signal:noise ratio of 3.8 [3].

In this experiment, miRNA expression is correlated to expression of mRNAs targeted within the 3'UTR. The targeted mRNAs for each miRNA was found by searching for mRNAs having perfectly complementary 6-mer seed sites (positions 2-7) conserved among human, mouse and rat. The list of targeted mRNAs was then compared to the non-target group, yielding a ranked-sum Wilcoxon p-value for each miRNA. This Wilcoxon test was done in the statistics package R, using the one-sided ”less” parameter. Other than that, only default settings were used.

Previous work [24] revealed that the expression of some miRNAs correlate positively to their target genes, whereas other miRNAs have negative correlation. My work has supported this fact. My work shows that some miRNAs have positive correlation (figure 4.2), whereas others have negative correlation (figure 4.3). Also, many miRNAs showed no significant correlations at all. My result gave much stronger p-values than what was found by Liu et al. [24] , and the reason for this might be that they did not consider the expression of each miRNA when doing the correlation calculations.

Significant correlations were found by setting the confidence interval at 0.05. This confidence interval was adjusted with the Sidak formula, yielding a new adjusted confidence interval of $1 - (1 - 0.05)^{1/282} = 0.000182$. The number 282 was used since there were 282 different miRNAs tested. Of these 282 miRNAs, 40 (about 14%) had significant ($p<0.000182$) negative correlations, whereas 72 (about 26%) had significant positive correlations. This all means that about 40% of all miRNAs had significant correlations of their expression to the targeted mRNAs after adjusting the confidence interval using the Sidak formula.

The reason that the remaining 60% of miRNAs were not giving significant correlations may be the following: (1) the target prediction algorithm is not powerful enough, since it has been shown that for example requiring additional matches toward the 3' region of miRNAs may improve the predictions [12], or (2) that I only considered the 19 tissues that were overlapping between the mRNA and miRNA expression databases. Thus, using more than these 19 tissues would probably improve the predictions.

It is also interesting to note that the amount of miRNAs with positive correlations is higher than the amount of negative correlations. This observation con-

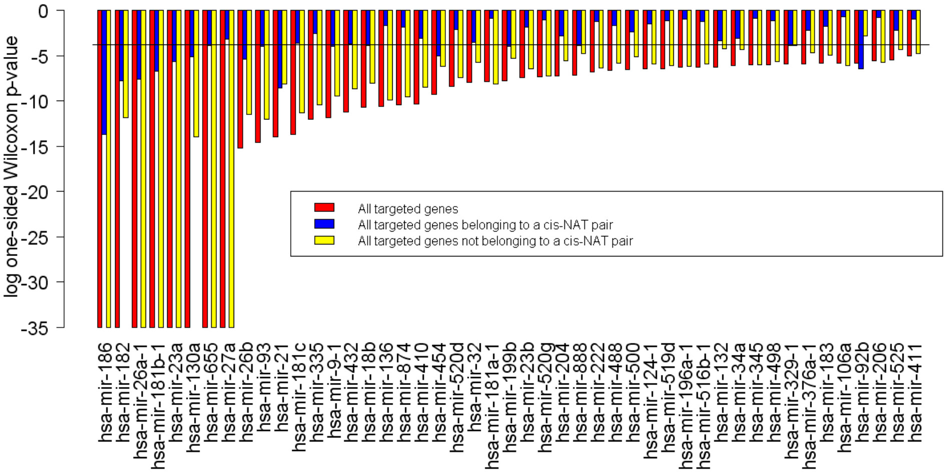


Figure 4.2: MicroRNAs with positive (p<0.00001) correlation to the expression of their target genes. Note that a log p-value of -35 in this graph equals a p-value of exactly zero. The line at -3.74 represents the confidence interval adjusted according to the Sidak formula.

tradicts the fact that miRNAs is known to downregulate their targets, meaning that we would expect a higher percentage of negative correlations than positive correlations. Therefore, this might indicate that some of these targets are not targeted directly, but indirectly through some other transcripts.

I also did another experiment, to see if targeted mRNAs that belong to a cis-NAT pair has a more significant p-value than other mRNAs. From this graph, we see that the targeted mRNAs that belong to a cis-NAT pair often has much less significance power that targeted mRNAs that not belongs to cis-NAT pairs. This might partly be explained by the fact that the number of transcripts that belongs to cis-NAT pairs is lower than the number of transcripts that does not. More experiments should be done before one can answer why there is such a difference between the cis-NAT group and the non cis-NAT group.

4.5 Expression of mir-28 has significant negative correlation to the expression of targeted cis-NAT pairs

We will now try to answer RQ2, which was mentioned in chapter 2. Since we found a significant correlation between some miRNAs and the targeted mRNAs, it would be interesting to see if cis-NAT transcripts show such correlation as well.

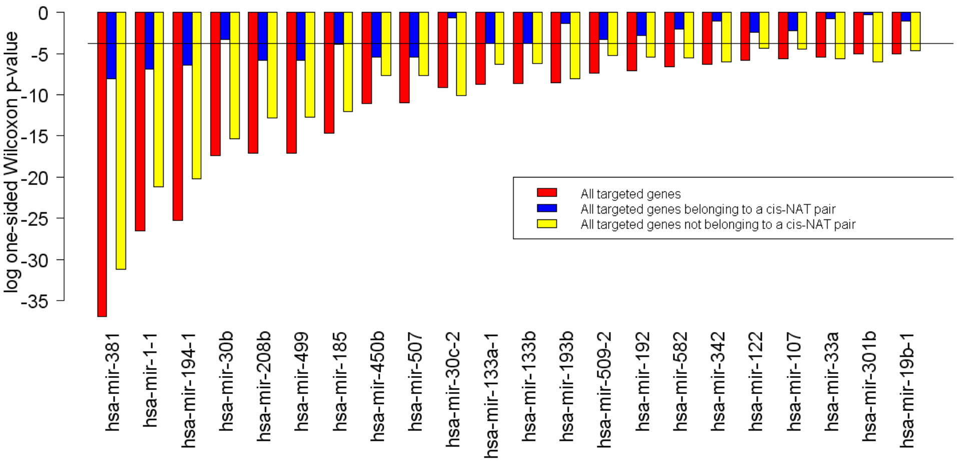


Figure 4.3: MicroRNAs with negative (p<0.00001) correlation to the expression of their target genes. The line at -3.74 represents the confidence interval adjusted according to the Sidak formula.

To test this, I followed the procedure shown in figure 3.2. I first tried to search for non-coding transcripts inside cis-NAT pairs that have perfectly complementary 7-mer-m8 (positions 2-8) miRNA seed sites conserved among human, mouse, and rat. It turned out that using 7-mer-m8 seeds gave to few hits to give any good results, therefore I decided to search for 6mer (positions 2-7) seed sites instead. The list of targeted cis-NAT transcripts was then compared to the non-target group, yielding a ranked-sum Wilcoxon p-value for each miRNA. As in the last section, the experiment was done through the statistics package R, using the one-sided ”less” parameter. I also restricted the considered miRNAs to the ones having seeds conserved among at least 15 different species, yielding 117 human miRNAs. The reason for this restriction is that we have shown previously (section 4.1) that these are the most enriched miRNAs within cis-NAT pairs.

From this experiment, I found mir-28 with a significant p-value (p=0.00028). This is a significant p-value, since it is lower than the confidence interval adjusted with the Sidak formula $(1-(1-0.05)^{1/117} = 0.00044)$. The number of targeted transcripts for mir-28 are 309, and 1532 transcripts in the non-target group.

One source of noise in this experiment may be that some of the protein-coding cis-NAT transcripts are also targeted directly inside their 3’UTR. Therefore, I did a new experiment where I took this possibility into consideration. I excluded cis-NAT pairs that were targeted by 7-mer-m8 seeds within 3’UTR of the coding side. This new setup gave a better p-value (p = 0.0002) for mir-28. The number of transcripts in the target group were 305, and 1516 in the non-target group. This result strengthens the assumption that targeting of 3'UTR is weakening our signal for cis-NAT pairs.

To further improve the p-value, I tried to restrict the considered cis-NATs to the ones not being targeted by 6-mer seeds within 3'UTR of the coding side. This gave a p-value of 0.00027 for mir-28. The target group had 292 transcripts, whereas the non-target group had 1471 transcripts. I also tried this same experiment, but with 5-mer seeds instead. This setup gave a p-value of 0.000088 for mir-28. And, when looking for 4-mer seeds, I got a p-value of 0.000043.

It would also be interesting to correlate the expression of mir-28 and the excluded transcripts (the ones being targeted within 3'UTR). I did this experiment, and the result was a significant positive correlation (p=0.97), and this result is consistent with the improving p-value for for mir-28 just mentioned.

In addition, I tried to use the Spearman correlation method instead of Pearson. In all cases, the spearman correlation formula gave worse results than the Pearson formula, indicating that there is a linear correlation between the expression of miRNAs and expression of targeted transcripts.

My last step was to confirm the significant p-value for mir-28 by using the Benjamini and Hochberg method mentioned earlier in this report. The analyzation was done with the R function named "p.adjust". The result confirmed to the results I have got so far, in that mir-28 was the only miRNA with significant p-value (p = 0.0063; Benjamini and Hochberg adjusted). No other miRNAs gave significant p-values. This run was done by excluding transcripts having 4mer seed sites within 3'UTR conserved among HMR.

It is interesting to notice that there is a negative correlation between the expression of mir-28 and the expression of the coding transcript of the targeted cis-NAT pairs. According to the theory outlined in section 1.3, we would expect a positive correlation instead of a negative correlation, and further experiments could reveal the reason to why this happens.

As a conclusion, this shows that the expression of mir-28 has significant negative correlation to expression of the targeted non-coding transcripts. When also considering the fact that the coding side may be targeted within its 3'UTR, the p-value got even stronger. However, I found no other miRNAs with significant correlations. Despite of a lack of significant p-values for other miRNA, there might still be correlations between the other miRNAs and their targets. It may just be that the noise is too high compared to the signal. As a conclusion, it is hard to tell if there actually is a correlation between the other miRNAs and their targeted cis-NAT pairs. Even though my experiment only showed significant correlations for mir-28, it can not exclude the possibility that there might still be a significant correlation for the other miRNAs.

Chapter 5

Conclusion

This work has suggested that miRNAs might bind to the non-coding transcript of cis-NAT pairs, and thus regulate expression of the coding transcript. Seed sites from conserved miRNAs were enriched within these non-coding transcripts, suggesting that miRNAs are actually taking advantage of this way of regulating a coding gene. Even though there was an enrichment of seed sites within non-coding transcripts, it was a lot weaker than within 3'UTR of mRNAs. The miRNA mir-28 also has significant negative correlation to the targeted cis-NAT pairs, whereas the other miRNAs gave no significant correlation. Despite that mir-28 was the only miRNA with significant correlations, this result does not exclude the chance that there might still a be correlation between the other miRNAs and their targeted non-coding transcripts. It was also shown that the 3'UTR of the coding transcript within cis-NAT pairs are longer and more conserved than other transcripts.

Chapter 6

Further work

- **Improvement to the way of correlating miRNA expression with cis-NAT expression -** this experiment has already been done in this project, but it would be interesting to try to improve the results. There are a number of ways to possibly improve the results: (1) We could use a more powerful target prediction algorithm when looking for miRNA targets. (2) We could use the NCI-60 microarray data, that contains expressions of 60 human cancer cell lines from several distinct tissues[25]. The usage of this microarray data could possibly give some improvements over the 19 tissues analyzed in my study.

Bibliography

[1] Pål Saetrom, Ola Snøve, and John J Rossi. Epigenetics and micrornas. *Pediatr Res*, 61(5 Pt 2):17R–23R, May 2007.

[2] Daniel H Kim, Pål Sætrom, Ola Snøve, and John J Rossi. Microrna-directed transcriptional gene silencing in mammalian cells. *Proc Natl Acad Sci U S A*, Oct 2008.

[3] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120(1):15–20, Jan 2005.

[4] Benjamin P Lewis, I hung Shih, Matthew W Jones-Rhoades, David P Bartel, and Christopher B Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, Dec 2003.

[5] Kevin V Morris, Sharon Santoso, Anne-Marie Turner, Chiara Pastori, and Peter G Hawkins. Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. *PLoS Genet*, 4(11):e1000258, Nov 2008.

[6] Kevin V Morris, Simon W-L Chan, Steven E Jacobsen, and David J Looney. Small interfering rna-induced transcriptional gene silencing in human cells. *Science*, 305(5688):1289–1292, Aug 2004.

[7] Xiu-Jie Wang, Terry Gaasterland, and Nam-Hai Chua. Genome-wide prediction and identification of cis-natural antisense transcripts in arabidopsis thaliana. *Genome Biol*, 6(4):R30, 2005.

[8] Yong Zhang, X. Shirley Liu, Qing-Rong Liu, and Liping Wei. Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-nats) in ten species. *Nucleic Acids Res*, 34(12):3465–3475, 2006.

[9] Xuefeng Zhou, Ramanjulu Sunkar, Hailing Jin, Jian-Kang Zhu, and Weixiong Zhang. Genome-wide identification and analysis of small rnas originated from natural antisense transcripts in oryza sativa. *Genome Res*, 19(1):70–78, Jan 2009.

[10] Pär G Engström, Harukazu Suzuki, Noriko Ninomiya, Altuna Akalin, Luca Sessa, Giovanni Lavorgna, Alessandro Brozzi, Lucilla Luzi, Sin Lam Tan, Liang Yang, Galih Kunarso, Edwin Lian-Chong Ng, Serge Batalov, Claes Wahlestedt, Chikatoshi Kai, Jun Kawai, Piero Carninci, Yoshihide Hayashizaki, Christine Wells, Vladimir B Bajic, Valerio Orlando, James F Reid, Boris Lenhard, and Leonard Lipovich. Complex loci in human and mouse genomes. *PLoS Genet*, 2(4):e47, Apr 2006.

[11] Yong Zhang, X. Shirley Liu, Qing-Rong Liu, and Liping Wei. Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-nats) in ten species. *Nucleic Acids Res*, 34(12):3465–3475, 2006.

[12] David P Bartel. Micrornas: target recognition and regulatory functions. *Cell*, 136(2):215–233, Jan 2009.

[13] Hochberg Y Benjamini Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B*, 1995.

[14] http://shootout.alioth.debian.org. The computer language benchmarks game.

[15] Mathieu Blanchette, W. James Kent, Cathy Riemer, Laura Elnitski, Arian F A Smit, Krishna M Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D Green, David Haussler, and Webb Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, 14(4):708–715, Apr 2004.

[16] Scott Schwartz, W. James Kent, Arian Smit, Zheng Zhang, Robert Baertsch, Ross C Hardison, David Haussler, and Webb Miller. Human-mouse alignments with blastz. *Genome Res*, 13(1):103–107, Jan 2003.

[17] F. Chiaromonte, V. B. Yap, and W. Miller. Scoring pairwise genomic sequence alignments. *Pac Symp Biocomput*, pages 115–126, 2002.

[18] Sam Griffiths-Jones. The microrna registry. *Nucleic Acids Res*, 32(Database issue):D109–D111, Jan 2004.

[19] Sam Griffiths-Jones, Russell J Grocock, Stijn van Dongen, Alex Bateman, and Anton J Enright. mirbase: microrna sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34(Database issue):D140–D144, Jan 2006.

[20] Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J Enright. mirbase: tools for microrna genomics. *Nucleic Acids Res*, 36(Database issue):D154–D158, Jan 2008.

[21] Yong Zhang, Jiongtang Li, Lei Kong, Ge Gao, Qing-Rong Liu, and Liping Wei. Natsdb: Natural antisense transcripts database. *Nucleic Acids Res*, 35(Database issue):D156–D161, Jan 2007.

[22] Pablo Landgraf, Mirabela Rusu, Robert Sheridan, Alain Sewer, Nicola Iovino, Alexei Aravin, Sébastien Pfeffer, Amanda Rice, Alice O Kamphorst, Markus Landthaler, Carolina Lin, Nicholas D Socci, Leandro Hermida, Valerio Fulci, Sabina Chiaretti, Robin Foà, Julia Schliwka, Uta Fuchs, Astrid Novosel, Roman-Ulrich Müller, Bernhard Schermer, Ute Bissels, Jason Inman, Quang Phan, Minchen Chien, David B Weir, Ruchi Choksi, Gabriella De Vita, Daniela Frezzetti, Hans-Ingo Trompeter, Veit Hornung, Grace Teng, Gunther Hartmann, Miklos Palkovits, Roberto Di Lauro, Peter Wernet, Giuseppe Macino, Charles E Rogler, James W Nagle, Jingyue Ju, F. Nina Papavasiliou, Thomas Benzing, Peter Lichter, Wayne Tam, Michael J Brownstein, Andreas Bosio, Arndt Borkhardt, James J Russo, Chris Sander, Mihaela Zavolan, and Thomas Tuschl. A mammalian microrna expression atlas based on small rna library sequencing. *Cell*, 129(7):1401–1414, Jun 2007.

[23] Craig Larman. *Applying UML and Patterns: An Introduction to Object-Oriented Analysis and Design and Iterative Development.* Prentice Hall PTR, 2004.

[24] Hongye Liu and Issac S Kohane. Tissue and process specific microrna-mrna co-expression in mammalian development and malignancy. *PLoS ONE*, 4(5):e5436, 2009.

[25] Paul E Blower, Joseph S Verducci, Shili Lin, Jin Zhou, Ji-Hyun Chung, Zunyan Dai, Chang-Gong Liu, William Reinhold, Philip L Lorenzi, Eric P Kaldjian, Carlo M Croce, John N Weinstein, and Wolfgang Sadee. Microrna expression profiles for the nci-60 cancer cell panel. *Mol Cancer Ther*, 6(5):1483–1491, May 2007.