A methodical analysis of target structure characteristics in RNA-RNA interactions

Bachelor Thesis
Chair of Bioinformatics
Department of Computer Science
Faculty of Engineering
Albert-Ludwigs-Universität Freiburg

submitted by
Kyanoush Seyed Yahosseini
13.09.2010

Assessor:	Prof. Dr. Rolf Backofen
Supervisor:	M.Sc. Sita Lange

Declaration

I hereby declare, that I am the sole author and composer of my thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work.
I hereby also declare, that my thesis has not been prepared for another examination or assignment, either wholly or excerpts thereof.

place, date　　　　　　　　　　Signature

Abstract

The prediction of mRNA target sites is indisputably a big challenge in bioinformatics. This investigate the influences of accessibility and structural stability at the RNAi interaction site. Therefore, the unpaired probability and the positional entropy for five different datasets, with experimentally validated target sites are analysed. These datasets each contain between 60 and 290 endogenous miRNA and synthesised siRNA interactions targeting the mRNA of *Homo sapiens*, *Arabidopsis thaliana* and *Photinus pyralis*. The focus is on the surrounding areas down- and upstream of the target site using a sliding window approach.

Using statistical measurements comparing functional with non-functional interaction sites, this work shows the existence of high accessible regions. For each of the datasets two high accessible regions exist, surprisingly not directly next to the target site or even within. The high accessible region downstream is the more significant one, located between 10 and 180 nucleotides downstream. The size of this region is at most 100 nucleotides for each dataset and for three of the datasets it was split into two. The significant upstream region is located between 170 and 80 nucleotides upstream. Several statistical tests were applied to validate these results. The structural stability as measured in this work, seems to have no influence on the ncRNA targeting in RNAi.

Danksagungen

Acknowledgement

Mein besonderer Dank gilt: Prof. Dr. Rolf Backofen für die Möglichkeit an seinem Lehrstuhl diese Bachelorarbeit schreiben zu dürfen. Sita Lange für die wunderbare Betreuung und das geduldige Beantworten all meiner Fragen. Meiner Familie die mir das Studium erst ermöglicht hat. Und natürlich meiner Freundin für ihren Beistand, vor allem in Zeiten in denen nichts so läuft wie es laufen sollte. Natürlich gilt mein Dank auch den Korrektoren die sich die Zeit genommen haben, diese Arbeit zu lesen.

Contents

1. Introduction

This thesis involves the analysis of specific characteristics of RNA-RNA interaction data with the aim to help improve the current knowledge of the underlying mechanism and thus to improve current prediction methods.

First the background in biology required to understand the specific problems of RNA-RNA interaction prediction is introduced. This involves a general introduction to RNA and RNA structure, followed by RNA-RNA interactions and the motivation for this work.

1.1. RNA Biology

Ribonucleic acids (RNA) are in most cases single stranded molecules, which consist of a chain of nucleotides. There are four different kinds of nucleotides in RNA, called adenine (A), guanine (G), cytosine (C) and uracil (U). Each of these nucleotides consists of phosphate, a ribose sugar, with its carbons enumerated from 1' to 5', and a base attached to the 1' position. Henceforth, nucleotides are also simplified to just bases. In comparison to deoxyribonucleic acids (DNA), it is not only found in the nucleus, but also in the cytoplasm of a biological cell.

1.1.1. Structure

The primary structure of an RNA is the sequence of the single nucleotides. The primary structure is noted as a string with the nucleotides from the 5' to the 3' end. When looking at a region of interest, bases to the direction of the 5' end are locaded upstream and bases in the 3' direction downstream.

The secondary structure consists of the base pairings between the nucleotides of the sequence. Base pairings are formations of hydrogen bonds between the bases of different nucleotides. The most common base pairs are A with U and G with C, called Watson-Crick pairs. Moreover, other possible pairs exist, like G with U, which appear less frequent. In this thesis, secondary structure and structure will be used synonymously.

A secondary structure is a set of pairs, with each nucleotide involved in at most one pair. These pairs can build up different structures as shown in Figure 1.1. For computational reasons pseudoknots are forbidden in most cases. For the formal definition of secondary structure and pseudoknots see Section 2.3.

The tertiary structure is defined by the three-dimensional structure of the RNA, which consists of the atomic coordinates of the nucleotides. This tertiary structure is formed by

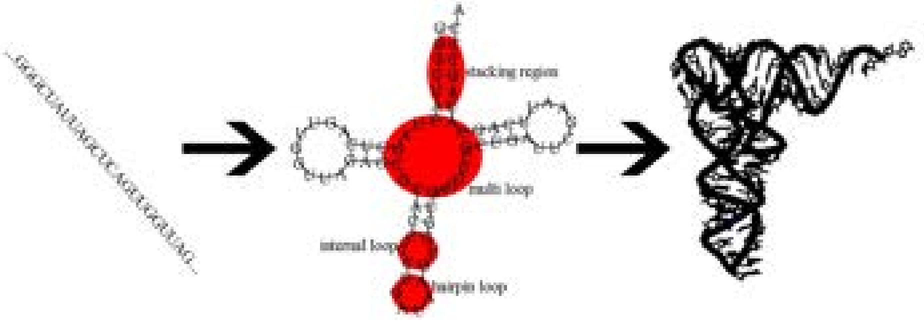


Figure 1.1.: Primary, secondary and tertiary structure of RNA. For the definitions of the different secondary structure elements see, Section 2.3.

additional bonds between the nucleotides. Because the greatest energy contribution to the stability of the tertiary structure originates from the secondary structure and because of the difficulty of tertiary structure prediction, this form of RNA structure is ignored in most prediction models and in this thesis.

Figure 1.1 visualises the different kinds of RNA structure and secondary structure elements.

1.1.2. General Function

In addition to the primary Central Dogma of Molecular Biology[1], it is shown that RNA not only acts as a carrier of genetic information, but also has several other functions. This leads to a classification of RNA molecules into two different groups, the coding and the non-coding RNA.

Messenger RNA (mRNA) is the coding RNA, this type of RNA is the product of transcription from DNA that encodes for a protein. The nucleotide sequence is translated into the amino acid sequence of a protein, called polypeptide chain, using transfer RNAs and ribosomes [32].

There are also several other RNAs, called non coding (ncRNA), they perform a multitude of functions that assist and regulate translation and transcription. Some examples of ncRNAs are:

- **Transfer RNA (tRNA)** are short, 73 to 94 nucleotides long, RNA synthesised in the nucleus. They transfer a specific amino acid to the mRNA and ribosome complex for the assembly of the polypeptide chain and thus the protein [32].
- **Ribosomal RNA (rRNA)** is assembled into ribosomes and decodes mRNA into

[1] The primary Central Dogma of Molecular Biology says, that if information gets into a protein, it cannot flow back to DNA or RNA.

amino acids and interacts with the tRNA and is also necessary to enable and support the translation process [32].

- **MicroRNA (miRNA)** and small interfering RNA (siRNA) are very short, about 21 to 24 nucleotides long, ncRNAs responsible for post transcriptional gene regulation called RNA interference [7].
- **Small bacterial RNA (sRNA)** are a class of 50-500 nucleotides long RNAs in bacteria and are, amongst other functions, responsible for the regulation of gene expression [6].

1.1.3. RNA-RNA Interaction and RNA Interference

RNA-RNA interaction is the interaction between two RNAs, in most cases the interaction between a mRNA and a ncRNA. These interactions are nessesary for many gene regulatory processes.

RNA interference (RNAi) is an example of post transcriptional gene regulation by RNA-RNA interaction. By insertion of an RNA-induced silencing complex (RISC) and the hybridisation of siRNA or miRNA with the mRNA, translation is repressed or mRNA is degraded [34]. Figure 1.2 clarifies the different biogenesis pathways or origins of siRNA and miRNA, while displaying a generally similar function. SiRNA is often introduced by viruses or in *in vitro* experiments, while miRNA is endogenous.

Another example for RNA-RNA interaction is directed by antisense RNA. In this interaction the non coding strand of the DNA, directly opposite to the gene, is transcribed into antisense RNA. This antisense RNA binds to the mRNA corresponding to it, i.e. antisense to it. After this process the mRNA is no longer accessible for proteins and ribosomes and cannot be translated.

1.2. Requirements and Aims

Basically three different steps are required to be realised:

1. Gather experimentally validated interaction data, for different species and types of ncRNAs
2. Implement scripts to analyse data and evaluate them statistically
3. Analyse and discuss the results and search for similarities within the gathered data

With this approach this work statistically evaluates the accessibility and structural stability of RNAi. This evaluation is not only limited to the target site and its immediate flanking area, but also regions down- and upstream of the target site. The aim is to find characteristics with significant and consistent results. These results should help to

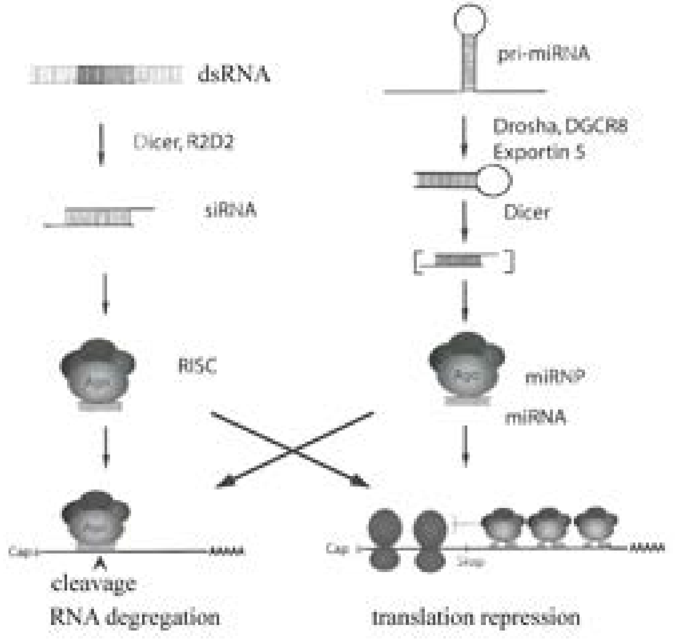


Figure 1.2.: Origin and function of siRNA and miRNA [34].

improve current knowledge of the underlying mechanisms of RNAi and eventually the prediction of such specific RNA-RNA interaction sites.

1.3. Motivation

The efficient prediction of RNA-RNA interactions, such as the exact target sites of miRNA or siRNA is a challenge in bioinformatics. Recent research has shown that miRNA are linked to human diseases, including cancer [30] and viral infections [35], e.g. playing an anti-viral role [25]. Even drug addiction seems to be regulated by miRNA [18]. MiRNA acts also as a key regulator in various cell processes like cell proliferation, cell death [5] and cell differentiation [10]. The ability to predict target sites on mRNA is necessary to understand the complex post transcriptional gene regulation in cells.

While the general biological function of RNA-RNA interaction e.g., RNAi starts to become clear, the prediction of valid target sites and the influences of different structural characteristics at the target site is far from being solved (see Section 3.2).

Although there is the possibility to predict target sites by experiments, these approaches are slow and expensive [19]. Furthermore, prediction of target sites within developed algorithms still lead to many false positive target sites and has an overall low sensitivity (see Section 3.1). Even though several researchers have tried to improve the prediction of specific target sites by exploring the exact influence of accessibility and other characteristics like positional entropy, the mechanism of how the ncRNA correctly identifies its target site is still not completely understood.

1.4. Thesis Structure

This chapter gives a short introduction about the biological background necessary which is for this work. In the next chapter, the scientific background and bioinformatic algorithms are introduced that are used to deal with the target structure characteristics in RNA-RNA interactions. In the third chapter some recent researches are introduced, describing the current knowledge about RNA target site characteristics. Chapter 4 shows the methods which are used to generate the results, displayed in Chapter 5. Afterwards, Chapter 6 discusses the results and gives a brief outlook to possible further work.

2. Scientific Background

For a more thorough understanding of the presented work, this chapter introduces some necessary background knowledge in the bioinformatics of RNA secondary structure and interaction prediction.

2.1. Secondary Structure Prediction

Although there are biophysical experimental methods to find out the secondary structure but they are too expensive for practical use. To be able to predict target sites on mRNAs and for RNA-RNA interaction in general, it is first of all necessary to predict the secondary structure of RNA. Secondary structure is a fundamental determinant of the function of ncRNA, usually more important than the primary sequence [27]. This is shown by the high amount of base-pair conservation across diverse species.

All RNA secondary structure prediction algorithms mentioned in this work are based on the principles of dynamic programming[1]. Also all introduced algorithms can not calculate pseudoknots, for the definition of pseudoknots see Section 2.3 Definition 2. Some programs do exists which are especially designed for the prediction of pseudoknots, such as pknotsRG [37]. But these methods cannot compute base pair probabilities, as introduced in Section 2.1.3, except for Dirks and Pierce [8], which is too slow in practice.

2.1.1. Energy Model

To predict RNA structure, an energy model for RNA, Gibbs Free Energy[2], is defined. When RNA structures are formed there is an associated loss of energy caused by the formation of hydrogen bonds between the bases [32]. To be able to compute the energy of a complete RNA structure, first it is necessary to get the energy of all substructures. By experiments, the change of free energy due to the folding into a substructure, like those introduced in Section 2.3, Definition 2, starting from the unfolded sequence can be identified [21]. These thermodynamic parameters are used to calculate the free energy for a complete RNA sequence by summing up the enthalpic and entropic terms for the single substructures [11].

[1] A method of solving complex problems by breaking them down into small subproblems.

[2] Gibbs Free Energy $G = H - TS$, with H the enthalpy, T the temperature and S the entropy.

2.1.2. Minimal Free Energy

The structure with the minimal free energy (MFE) ΔG has been assumed to be the best biologically structure in the past citemount04. The goal of the Zuker algorithm is to compute this MFE structure.

Zucker uses dynamic programming to calculate the best substructures with a recursion formula and stores them into two matrixes. One of them containing the optimal substructure with unpaired ends and the other the best closed substructure, consisting of an internal loop, a hairpin, a stacking region or a multiloop, with the lowest free energy.

The overall free energy is calculated by the sum over all structural elements and the resulting structure is calculated by backtracking while one recursion represents the minimal energy of a general substructure the other one represents the minimal energy of a closed substructure. For further information and exact definition read [45]. Programs implementing this idea are e.g. mfold [45] and RNAfold [13,17].

2.1.3. Base Pair Probabilities

The MFE structure has been shown to not necessarily be the best biological structure for a given RNA [29]. There can be a large number of different secondary structures that are close or even equal to the MFE structure in terms of free energy. Also in most cases, due to the fact that RNA structure often conserves the function of the sequence, the best structure is a stable one, which is not much influenced by the change of single nucleotides [27].

The system of structures is called ensemble. To be able to calculate the base pair probabilities efficiently, it is assumed that the lower the energy of a structure the more probable it is. Therefore, the different structures are Boltzmann distributed.

So the probability of a given structure P with the free energy ΔG is calculated with help of the Boltzmann factor and is

$$Pr[P|S] = \frac{e^{\frac{-\Delta G}{kT}}}{Z}. \tag{2.1}$$

k is the Boltzmann constant, T the temperature, S the structure ensemble and Z the partition function $Z = \sum_S e^{\frac{-\Delta G}{kT}}$. The probability of a base pair (i,j) is calculated by

$$Pr[(i,j)|S] = \sum_{P \ni (i,j)} Pr[P|S] \text{ [29]}. \tag{2.2}$$

The recursion of the algorithm is very similar to Zuker's, but uses multiplication instead of summation to calculate the probability of a base pair or a structure [29]. RNAfold [13] also implements this approach, when given the option -p.

2.1.4. Global and Local Folding

Global folding is the prediction of a secondary structure for the complete RNA. But there are several restrictions, which make global folding for long RNA sequences pointless.

First of all, proteins and other elements like ribosomes disrupt the secondary structure and additionally to that, the mRNA is partially unfolded to be ready for translation [41]. Also base pairs over a long distance are very improbable. Thus, the energy data gathered for the minimum free energy structure is not very accurate to predict such structures [9].

One approach to avoid these problems is to use local folding. Local folding calculates the secondary structure for a part of the sequence independent of the whole sequence. E.g. the program RNAlfold [13] allows only interactions between base pairs with a given local distance L to each other. While RNAplfold [4] calculates the structure with independent and sliding windows of length W given a maximal distance between base pairs L. It averages over all windows that include each base pair. This program also has the positive side effect that it uses much less storage space and works faster than global folding programs [4].

2.2. RNA-RNA Interaction Prediction

The methods to predict RNA-RNA interaction or target sites experimentally are very expensive. Therefore, methods predicting these interactions computationally are necessary for all types of RNA-RNA interactions to pre-filter possible interactions and reduce the number of interactions that need to be experimentally verified. Due to the fact that a nucleotide can only pair with one other nucleotide[3], the knowledge of secondary structure is necessary to predict RNA-RNA interactions, to evaluate the influence of the previous single structures before the interaction occurs. Figure 2.1 clarifies two main approaches.

2.2.1. Concatenation Approach

The concatenation approach for RNA-RNA interaction prediction is based on the idea to connect both RNA sequences and to calculate their combined secondary structure. The connection site is commonly flagged and handled as an internal loop with specific energy parameters. The combined sequence is folded using a modified version of RNA folding algorithms. The folding of the concatenated sequences produces a combined structure and therefore a possible interaction between both RNA sequences.

Due the limitation of the Zuker algorithm, this approach can not predict pseudoknots between both RNAs, as shown in Figure 2.2 (B), but in nature this kind of structure is a very common interaction between two RNA sequences [3]. Therefore, this method shows serious flaws. Programs implementing this approach are e.g. PairFold and RNAcofold [3].

[3] Actually this is only true in the secondary structure, but not in the tertiary structure.

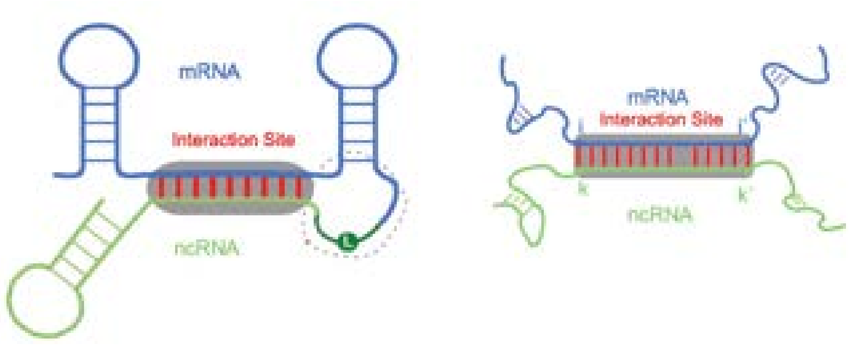


Figure 2.1.: Schematic comparison between the concatenation approach on the left and the ensemble approach on the right [3].

2.2.2. Ensemble Approach

The ensemble approach is based on two different steps, first the hybridisation energy between both RNA sequences is calculated. This is the energy of the most favourable hybridisation site, i.e. the hybridisation $H(i, i', k, k')$ with the lowest free energy, between the RNA sequences. Then the accessibility, which is measured as the energy to make the target site and the binding RNA single stranded, $ED[i, i']$ and $ED[k, k']$[4], at the hybridisation site is calculated. By combining the energy contributions the most favourable interaction is predicted, as

$$H(i, i', k, k') = E^{hybrid} + ED(i, i') + ED(k, k') \text{ [6]}. \tag{2.3}$$

The first step is implemented separately e.g. in RNAhybrid [24], but because it does not consider the structure of the RNAs, and their internal base pairings, these results are often inaccurate, as shown in Figure 2.2 (A).

Through the separate calculation of the two sequences, this approach overcomes the limitations of the concatenation approach and allows pseudoknots between the RNA sequences, but it only allows one interaction site. So all interactions interrupted by intra-molecular base pairs cannot be predicted. Programs implementing this approach are e.g. IntaRNA, RNAup and RNAplex [6, 13, 31]. Both approaches cannot predict more complex interactions like double kissing hairpins, shown in Figure 2.2 (D) [3, 6].

2.2.3. Problem Specific

Problem specific approaches try to predict special kinds of RNA-RNA interactions, e.g. miRNA-mRNA interactions or only interactions in one species. The approaches that

[4] For the formal definition of ED check Equation (2.4) in Section 2.3

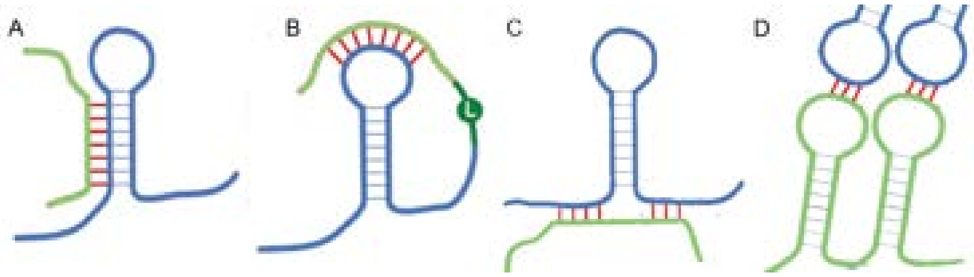


Figure 2.2.: (A) Biologically unlikely structure possibly predicted by RNAhybrid. (B) External pseudoknots, which can not be predicted by the concatenation approach. (C) Double interaction site, which can not be predicted by the ensemble approach. (D) Double kissing hairpins [3].

predict specific interactions, mostly use the ensemble approach and add some special features, specific to the binding mechanism.

The most problem specific programs focus on miRNA-mRNA interactions. For example PicTar uses additionally to the thermodynamic model, cross-species comparisons to filter out false positives. Using a thermodynamic model, IntaRNA also handles seed regions, as defined in Section 3.2.1. TargetScan searches for seed matches and expands them. This program calculates the free energy and based on that assigns a score for each possible binding site. For a detailed compare of miRNA prediction programs see [31].

2.3. Definitions

This section gives some necessary formal definitions. First of all the definition of secondary structure and secondary structure elements, useful for the understanding of the structure prediction algorithms, are given.

Definition 1 (*secondary structure*)

Given a sequence S, a secondary structure is a set of pairs $P = \{(i,j) : S_i$ and S_j form a valid pair $\}$.
Also a valid secondary structure must satisfy:

1. $\forall (i,j) \in P : 1 \leq i < j \leq |S|$
2. $\forall (i,j),(i',j') \in P : (i = i') \leftrightarrow (j = j')$
3. $(\forall (i,k),(i',k') \in P : i < i' \rightarrow (k < i') \vee (k' < k))$ *i.e. there are no pseudoknots*

Definition 2 (*Secondary structure elements*)

Given a RNA structure P, we call (i, j) *a pair iff,* $(i, j) \in P$

- $(i, j) \in P$ *is closing a hairpin loop, iff* $\forall i < i' < j' < j : (i', j') \notin P$
- $(i, j) \in P$ *is a stacking, iff* $(i + 1, j - 1) \in P$
- $(i, j) \wedge (i', j') \in P$ *are closing an internal loop, iff*
  1. $i < i' < j' < j$
  2. $(i, j) \wedge (i', j') \in P$
  3. $\neg\exists(k, l) \in P$ *: between* (i, j) *and* (i', j')

In most works accessibility is defined as the energy difference between the paired structures and the unpaired.

Definition 3 (*Energy difference [16, 22]*)

The energy necessary to unfold a region a to b is:

$$ED_{a,b} = E^{all} - E_{a,b}^{unpaired} \tag{2.4}$$

E^{all} *is the free energy of the ensemble of all structures,* $E_{a,b}^{unpaired}$ *the free energy of the ensemble of all structures with the complete substring unpaired.*

With the knowledge of the $ED(a, b)$ it is easy to calculate the probability that a substring a to b is unpaired using the Boltzmann distribution.

Definition 4 (*Probability that a substring a to b is unpaired [16]*)

The Probability of a region a to b to be unpaired is:

$$PU_{a,b} = e^{E^{all} - E_{a,b}^{unpaired}/RT} = e^{\frac{ED_{a,b}}{RT}} \tag{2.5}$$

R the gas constant and T the temperature. The size of the region a to b is called u.

Positional entropy shows the entropy of a base. While a low value indicates a reliable prediction a high positional entropy stands for a high amount of alternative structures.

Definition 5 (*Positional entropy [13]*)

The positional entropy of a base i is:

$$PE_i = -\sum_{i \neq j} P_{i,j} \cdot log(P_{i,j}) - P_i^u \cdot log(P_i^u) \tag{2.6}$$

where $P_i^u = 1 - \sum_{i \neq j} P_{i,j}$*, the probability of i being unpaired.*

3. Related Work

This chapter introduces the current knowledge of target site characteristics for the different kinds of ncRNAs. It also discusses the open questions which remains.

3.1. Performance of RNA-RNA Interaction Prediction

As the development of special ncRNA-mRNA interaction prediction programs proceed, the predictive power of these programs increase. By the implementation of special features like seed regions and cross species comparisons, state of the art programs predicting miRNA-mRNA interaction have a lower false positive rate. For example TargetScan and PicTar have an estimated false positive rate between 22% and 30 % with a sensitivity about 80 % [31]. This value has originated from data of well explored species, like *Drosophila melangogaster* or *Homo sapiens.*

Marín *et al.* [26] rate the prediction performance of state of the art programs as very low. Hybridisation energy and total free energy at the target site have, although used in most programs as a criteria to predict interactions, a very low predictive power. Actually randomly taken regions with a matching seed site have a higher possibility to be a true interaction site than PITA, IntaRNA, miRanda and RNAhybrid predictions, with their default values. For both fruit fly and human genes. Even with optimised values the best introduced prediction program only finds 21 of 137 miRNA-mRNA interaction sites in the first 1000 predictions or 48 interactions in the first 10000 predictions for the fruit fly. The prediction performance at human genes is even worse [26].

3.2. Influences of Accessibility

Although there is the general assumption that accessibility has an influence on RNA-RNA interactions, the accurate influence is still not completely clear. In most cases it is shown that higher accessibility leads to a higher repression for siRNA, miRNA and sRNA [3, 22, 38]. The details of how to use accessibility for target site prediction remain unclear.

There are some different opinions about the minimal high accessible region and the different parameters to predict accessibility. Also in literature many different terms have been used for accessibility. The concrete influence of accessibility seems to be different for diverse ncRNA classes. The influence and predictive power of the target site structure and flanking regions are also unclear.

The following sections introduce some previous work done on various target sites, considering their accessibility.

3.2.1. miRNA Target Sites

Target prediction in plants is fairly uncomplicated, because plant miRNAs bind to their target with perfect or nearly perfect complementarity. For animal miRNA there seem to be different characteristics at the target site. The behaviour of a 7 to 8 nucleotide long sequence starting from the 5' end of the miRNA, called seed, seems to be important. Either a perfect 5' seed region or a seed region including mismatches with a long stretch of base parings at the 3' end is characteristic for a target site [28].

Kertesz *et al.* found out that in *D. melangogaster* regions with a high ED at the target site lead to high gene repression. The result shows, in this particular case, that there is a high correlation between the accessibility for the whole target site and the degree of regulation [22].

In contrast to this Hausser *et al.* showed that accessibility is not always a good indication of a target site, as shown in Figure 3.1. Firstly, the accessibility of the seed region seems to have only a minimal predictive power. Secondly, the accessibility for the complete target site or the flanking region seems to be a good indication for a target site only for some datasets; this could be an indicator of different regulatory mechanism or of poor data quality. [15]. This analysis is done on *human* data.

3.2.2. siRNA Target Sites

SiRNA are often designed to fit perfectly to a given mRNA target site, therefore not the prediction of their target sites is the main problem, but the correlation between repression efficiency, specific sequence and structural characteristics. Thus, the prediction of efficient siRNA is the tricky task.

For siRNA target sites there also seems to be a correlation between stacking regions and regulation efficiency. Schubert *et al.* found out that gene expression is up to 60 % lower in a region with fully unstacked nucleotides in comparison to a fully stacked target region, using VsiRNA1 to silence the expression of the *rat vanelloid receptor*. There was also evidence for a nearly linear correlation between the local free energy and the protein expression, independent of the accurate target structure [38].

Using RNAplfold, Tafer *et al.* have shown that there is a significant dependency between silencing activity and target site accessibility. The most significant results were achieved with a window of 80 nucleotides and maximal span between base pairs of 40 nucleotides. These results emphasise the benefits of local over global structure prediction. . The nucleotides at position 6-8 and 12-16 seems to work as a seed region. Also, due to

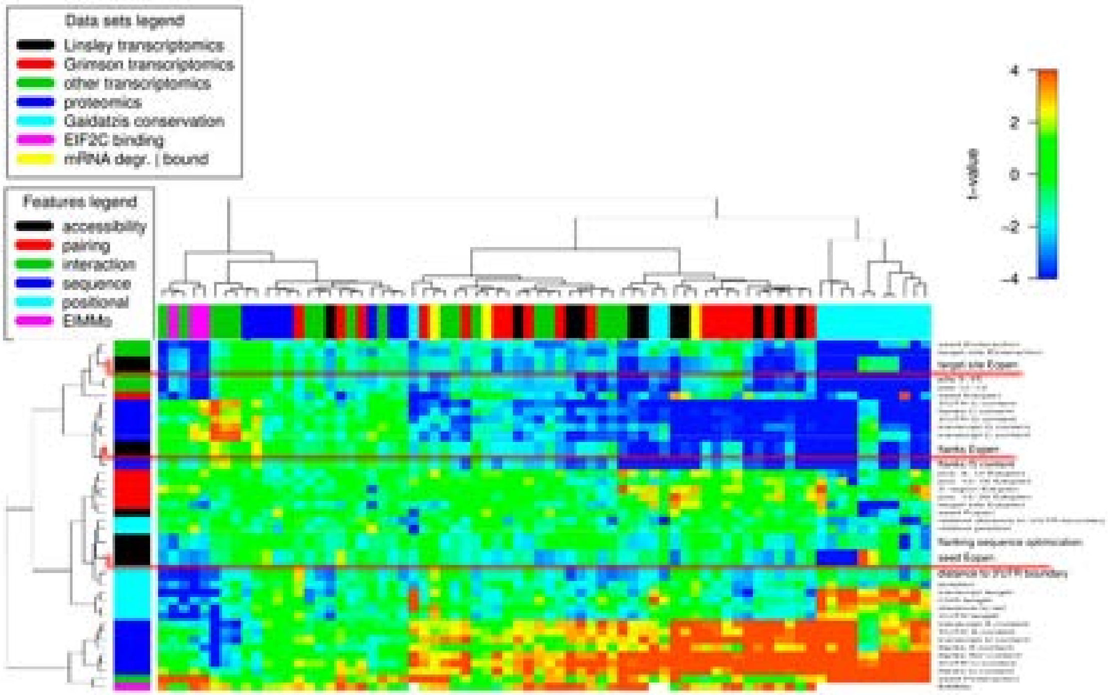


Figure 3.1.: The predictive power of the different structural characteristics [15]. Relevant for this work are the red marked characteristics (from top to bottom): accessibility target site, accessibility flank region and accessibility seed region.

the high complementary between siRNA and mRNA the hybridisation energy seems to be negligible [40].

3.2.3. sRNA Target Sites

A wide field of different sRNA in bacteria exists and thus, the prediction of their target sites seem to be even more complex and flexible than miRNA target prediction.

SRNA sequences can be between 50 and 200 nucleotides long with various functions and therefore have different requirements to their target site. In addition there is no known single significant feature. Although the precise requirements are unknown, it is assumed that accessibility and mainly the hybridisation energy play a major role [3, 6].

Furthermore, the number of experimentally verified interactions in this class of RNAs is still very small.

3.2.4. Open Questions

Although many works state a high accessibility at the target site is a good indicator for all kinds of RNA-RNA interactions. The exact influence is unclear, as many analyses give confusing or contradictory results [15, 22, 38, 40]. Additional to that, a review rates the accessibility of the flanking regions higher than the target site [40]. One of the open questions is which region around the target site needs to have a high accessibility. And do similarities between different species or ncRNAs exists. So far, the influence of positional entropy on RNA-RNA interactions and its predictive power has not been analysed at all.

Therefore, it is necessary to carry out further analysis of the influence of structural characteristics on the target site on various datasets from multiple sources.

4. Methods

After the related background information have been given in the previous chapters, this chapter introduces the programs, the interaction data and the statistical measurements used to carry out a methodical analysis of RNAi target site structural characteristics. Also a description of the implementation is given.

4.1. Data

To analyse the characteristics of RNA-RNA interaction sites it is necessary to collect the exact positions of validated interaction sites. Gathering experimentally validated target sites was complicated, because validated interaction pairs in databases usually do not have validated target sites within the mRNA. Often it is only known that the mRNA is regulated by the given miRNA, trough analysing of the expression of the mRNA. To validate the target site usually mutation experiments at this region needed to be performed. SiRNAs are designed to fit to one exact target site, but it is not tested whether the siRNA binds to other locations in the mRNA, although given the large amount of sequence complementary, this is very unlikely.

The search for validated target sites results in the five different datasets summarised in Table 4.1 and described in the following sections. These datasets contain RNAi interaction sites for siRNA and miRNA, three different organisms far apart on a scale of evolution and consists of many target genes.

4.1.1. Artificial siRNA Data

All artificial datasets have a quality measurement, calculated by the knockdown efficiency of the ncRNA as the average measured mRNA repression. These quality values were normalised to values between zero and one with a linear interpolation:

$$f(x) = \frac{x - min(X)}{max(X) - min(X)}. \tag{4.1}$$

Where $max(X)$ and $min(X)$ are the highest and lowest quality value within all observations X, respectively. Additionally, each dataset is split up into a functional and a non-functional group, by taking a fixed amount of the best and the worst interactions according to their quality value.

name	ncRNA	species	repression measure	source
Tafer02	artificial siRNA	*Homo sapiens*	quality values	[40]
Tafer03	artificial siRNA	*Homo sapiens*	quality values	[40]
Firefly	artificial siRNA	*Photinus pyralis*	quality values	[40]
AtmiR	endogenous miRNA	*A. thaliana*	functional & non-functional	[12]
Human	endogenous miRNA	*Homo sapiens*	functional & random	[44]

Table 4.1.: Names and details of the datasets used in this work.

Three datasets were extracted from the Tafer *et al.* paper [40], who in turn gained the data from [20]. Their knockdown efficiencies were verified by analysing mRNA and protein levels.

Tafer02 are artificial siRNAs synthesised to target against arbitrary regions of the coding sequences of human genes. The targeted genes are MAP2K1, GAPDH, PPIB, and LMNA. This dataset contains 294 interaction sites. The functional and non-functional group contains the best respectively worst 70 interactions.

Tafer03 are also artificial siRNAs, originally synthesised to evaluate the performance of an RNA interaction prediction program targeted against human genes. The targeted genes are cyclophilin, ALPPL2 and DBI. Originally, the **Tafer03** dataset also contains, the Firefly dataset, which here is considered separately because it is a different organism. After the seperation this dataset contains 270 interaction sites. The functional and non-functional group contains the best respectively worst 70 interactions.

Firefly is a subset of the original **Tafer03** dataset containing 89 artificial siRNAs, targeting the gene of firefly (*Photinus pyralis*) luciferase. Due to the small size of the overall interactions, the functional and non-functional group consists only of the best and worst 30 interactions.

4.1.2. Endogenous miRNA Data

The endogenous miRNA data is divided into functional and non-functional interactions. Because the human dataset does not contain non-functional interactions, random data, as described in Section 4.1.3, is used as the non-functional group.

The dataset **AtmiR** consists of 110 functional and 114 non-functional miRNA target sites in *Arabidopsis thaliana*, where the functionality is based on experimental evidence. The functional set was taken from a cleavage analysis performed by German *et al.* [12]. They have performed deep sequencing to identify cleavage products of miRNA degradation of target mRNAs in two cell lines. A more in depth description of this dataset is given in Appendix A.1.

The dataset **Human** consists of 67 functional miRNA target sites in 36 *Homo sapi-*

ens mRNAs taken from miRecords[1]. Entries in miRecord were only taken if mutation experiments were performed and, because of the incomplete version classification of the mRNAs, the target site could be located in the given mRNA. So the functionality and the actual position of the target sites is based on mutational experiments [44]. As mentioned before, 67 random miRNA target sites in the same mRNA target sequences as the non-functional group were generated.

4.1.3. Random Data

For each dataset, a corresponding set of random data was generated. To generate the random data, for each interaction in each dataset a random position over the whole unchanged mRNA with the same size as the original target site was chosen. The result is a set of interactions with the same size as the original dataset within the same target mRNA, so no structural bias of a different sequence is introduced.

4.1.4. Data Format

Every dataset is stored in a flat file, where each line contains one interaction. Each line contains an ID, the target mRNA, the start and stop positions of the interaction in the mRNA and a quality or functionality value, separated by tabs. The ncRNA ID is not stored at all, because the sequence and structure of these is not considered. The mRNA sequences are stored in separate files in fasta format.

4.2. Statistical Methods

There are several statistical methods to evaluate whether the data shows significant differences between the characteristics at functional and non-functional interactions. Random data, as introduced in Section 4.1.3, is used as a null model, to verify the significance of the results. This means the same test is repeated but only with the random data, where you expect the results to be less significant than the real data results. For further information about statistical hypothesis test and correlation please, as given in this Section read any standard statistic textbook or [23, 39].

4.2.1. Correlation

The linear correlation is used to analyse the relationship between two continuous variables. Here the **Spearman's rank correlation**, which is a non-parametric measure of statistical dependence between two variables, is applied. In this work, the correlation

[1] Extraced from http://mirecords.biolead.org/download.php, release 5 May 2010.

coefficient is used to evaluate the linear dependence between the quality value of the artificial siRNA datasets and the specific characteristic measure. The p-values[2] calculated by this correlation test are used to show the statistical significance of the correlation coefficient.

4.2.2. Hypothesis Tests

Two-sample hypothesis tests are used to compare two sets of data. Almost always the tests are using null-hypothesis tests, i.e. the tests are calculating the probability that results as extreme as the given ones occur by chance. Because only two-sample tests are introduced, the goal is to check if both sets of data are drawn from the same distribution.

The **Two-sample Student's t-test** assumes that the given data is approximately normal distributed. Therefore to check if the null hypothesis can be rejected it is enough to compare the means of the sets. And use the result to calculate the t-value [39]. With a given t-value and the size of a dataset the p-value can be calculated or looked up in a table.

The **Wilcoxon signed-rank test** which is equivalent to the Man-Whitney u-test is a nonparametric test, making no assumption over the distribution of the data. Here the sample values are ranked and the sum for each of the samples is calculated. After normalisation the difference between the rank sums is calculated, which is called u-value [39]. So in principle this test is a Student's t-test with a data ranking over the combined samples.

The **Kolmogorow-Smirnow test** is a very stable nonparametric test. This test basically compares the difference between each value in the set. To reject the null hypothesis this difference must not to exceed a given limit.

For this work one of the used measurements is the two-sample Student's t-test, because this test implies normal distributed data points and this is however not always correct. E.g. the boxplots[3] in Figure 5.1 show not normal distributed but significant results. The Wilcoxon signed-rank test is used to calculate the significance of a result. The u-values are not used because they are not comparable between different datasets. Therefore the p-values calculated by the Wilcoxon signed-rank test are used as a measurement for significance and the t-values of Student's t-test are used to show the direction of the difference. To minimise the possibility of statistic errors depending on a specific test, the Kolmogorov-Smirnov test is used to check the results of the Student's t-test and the Wilcoxon signed-rank. All tests introduced in this subsection are summarised with the name hypothesis tests.

[2] P-value is the probability, assuming that the null hypothesis is true, of observing a test statistic at least as extreme as the one that was actually observed. It describes a Type I error of the null hypothesis.

[3] Boxplots are introduced in the next Section

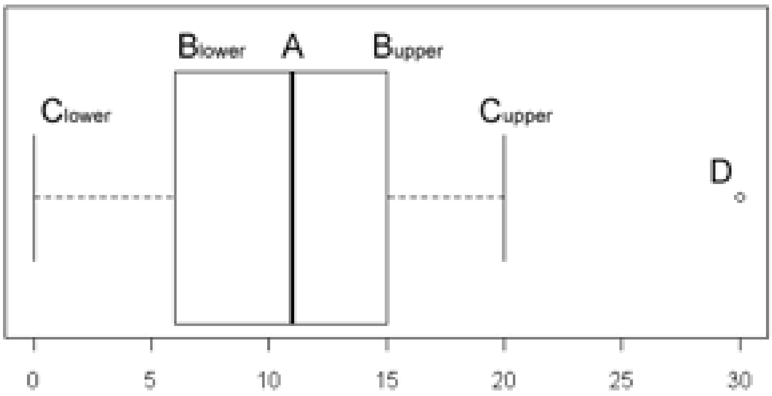


Figure 4.1.: A horizontal boxplot showing the (A) median (B) lower and upper quantile with an interquantile rage that encompasses 50% of the data (C) minimum and maximum observation still within 1.5 times the interquantile range of the lower and upper quartile (D) outlier, represented as a separate dot.

4.2.3. Data Summary

To visualise significant results boxplots can be used. Boxplots are diagrams to summarise data. They give a fast and easy overview of the distribution of the data. The spacing between the different parts indicates the degree of dispersion and skewness in the data. In contrast to the pure p-value, boxplots give a visual overview of the nature of data and the difference between the variables. Figure 4.1 shows a boxplot and defines the different parts of it.

4.3. Applied Programs

4.3.1. RNAplfold

RNAplfold[4] [4] has been chosen as the secondary structure prediction program for the mRNA. This prediction is necessary to calculate the structural characteristics in Section 4.4.2.

RNAplfold is used as the structure prediction program because of its benefits over most other structure prediction programs. It calculates the local structure [5] and has a relatively fast run time. Also RNAplfold can calculate the probability of a region u to be unfolded, see Definition 2.5 at Section 2.3.

[4] RNAplfold is provided as part of the Vienna Package 1.8.4 and can be downloaded at http://www.tbi.univie.ac.at/ivo/RNA/index.html.

[5] For the benefits of local folding check Section 2.1.4.

The parameters for RNAplfold used in this work are u=10, to calculate the probability that regions of 1 to 10 base pairs are unfolded, c=0 to disable the cutoff and to enable the calculation of positional entropy. Window length W is set to 100 and the maximal distance between base pairs L to 50, these values have proven to give good results when a large variety of parameter combinations were tested [15]. Also the parameter d=2 is used to specify a realistic energy model.

4.3.2. R

The calculation of the statistical measurements is done by the program R version 2.11.1[6] [36]. For the visualisation of the output R is extended by the library ggplot2 [43]. The statistical measurements are explained in Section 4.2.

4.4. Analysed characteristics

Two main structural characteristics were analysed:

1. The probability that a subsequence or region of the target is single-stranded PU, i.e. accessible to binding molecules.
2. The positional entropy PE, which measures the structural stability of a given nucleotide. A high PE value means the nucleotides can be directly involved in many alternative structures, a low value suggests that the given prediction is more likely to be correct.

Definition of PU and PE are given in Section 2.3.

4.4.1. Regions

For a thorough investigation of the target site, not only the interaction site itself is assessed, but also the flanking regions to the left and the right of the target site.

The flanking regions are analysed by a sliding window approach. Therefore all sequences are aligned accordingly to the 3' end of the target site and this position is 0, as shown in Figure 4.2. A window with the size w beginning at position i is called

$$R_w(i) = S(i, i + w). \tag{4.2}$$

$S(i, i + w)$ is a subsequence from i to $i + w$. Each window is plotted at position i, where i represents the leftmost border of the subsequence, i.e. the 3' end as shown in Figure 4.2. The window $R_{20}(-20)$ corresponds to the interaction site, if this is exactly 20 nucleotides long. The sequences are analysed starting from $R_w(-200)$ to $R_w(+200)$. To simplify matters this thesis Because the results of the separate target site were not significant, they were not considered.

[6] R can be downloaded at http://www.r-project.org/.

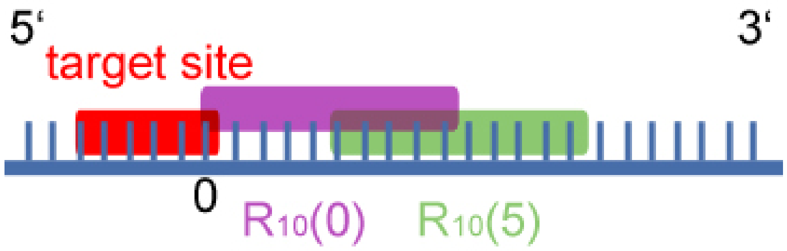


Figure 4.2.: Visualisation for different windows. The red rectangle marks the interaction site, while the violet and green rectangles flag $R_{10}(0)$ and $R_{10}(5)$. The interaction site is six nucleotides long, i.e. $R_6(-6)$ corresponds to the interaction site.

4.4.2. Accessible Regions

The accessibility of a region is defined as follows:

$$\begin{align}
maxPU_{k,l}(u) &= max\ \{PU_{a,b}|a < b, a \geq k, b \leq l, b-a=u\} \tag{4.3}\\
minPU_{k,l}(u) &= min\ \{PU_{a,b}|a < b, a \geq k\, b \leq l, b-a=u\} \tag{4.4}\\
meanPU_{k,l}(u) &= mean\ \{PU_{a,b}|a < b, a \geq k, b \leq l, b-a=u\} \tag{4.5}\\
&= \sum_{\substack{a,b\\ a<b,a\geq k\\ b\leq l,b-a=u}} \frac{PU_{a,b}(u)}{l-k+1}
\end{align}$$

Where $k \leq l$ and $S(k,l)$ is the subsequence of S from position k to l.

The PU values are calculated with RNAplfold with u values from 1 to 10. The window size for a region is set to 20 for the calculation of accessible regions. Figure 4.3 shows a schematic approach for the calculation of the accessibility in these regions.

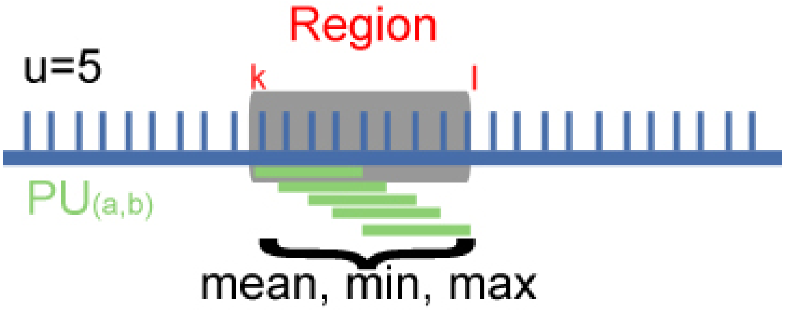


Figure 4.3.: Schematic visualisation of the calculation of accessible regions. The grey rectangle visualise the current window and the green lines are the different calculated $PU_(a,b)$ values.

4.4.3. Structural Stability

The structural stability analysis was done in regions, analogous to accessibility:

$$maxPE_{k,l} = max\ \{PE_a | k \leq a \leq l\} \tag{4.6}$$
$$minPE_{k,l} = min\ \{PE_a | k \leq a \leq l\} \tag{4.7}$$
$$meanPE_{k,l} = mean\ \{PE_a | k \leq a \leq l\} \tag{4.8}$$
$$= \sum_{\substack{a \\ k \leq a \leq l}} \frac{PE_a}{l - k + 1}$$

Where $k \leq l$ and $S(k,l)$ is the subsequence of S from position k to l.

The window size for a region is set to 5 for the calculation of structural stability.

4.5. Implementation

Several scripts were implemented for an automated calculation of the characteristics for the different datasets and regions. An overview of the scripts is given in Table 4.2.

name	input	output	function
create	fasta file of mRNA sequences	PE and PU data over complete sequences	calls RNAplfold and calculate PE
calculate	PE, PU and interaction data	characteristic results at interaction sites for one region	calculates characteristics of one region for all interactions
test	characteristic results	boxplots and test results	calculates statistical hypothesis tests and correlations with R
plot	test results	plots of correlation or hypothesis test results	plots the test results, differentiates between characteristic, region and u

Table 4.2.: Overview over the function of the different Perl scripts

Additionally, a wrapper script was implemented. The wrapper calls the scripts in the right order. First calling the **create** script to predict the secondary structure, then for each interaction file and each region the **calculate** script is called, followed by the **test** script. After all calculations for one dataset are done, the **plot** script plots the results. Therefore all scripts support different input and output file names via parameter. The **test** script has two calculation modes, one is calculating the correlation and the other the hypothesis tests. Additionally the start and stop position of the window relatively to the interaction sites can be given via parameters. Not used in this work but also implemented is the support for different secondary structure prediction parameters, like window size, programs in the **create** script and the support for different window sizes in the **calculate** script. The implementation of these scripts is done in Perl 5.10.1.1.

4.6. Analysis Procedure

After collecting the data and bringing it into the format shown in Section 4.1.4 for each interaction, all characteristics as described in Section 4.4.2 were calculated. This procedure is repeated for each region described in Section 4.4.1, so from $R_w(-200)$ to $R_w(+200)$ at a resolution of five nucleotides, i.e. the region is shifted 5 nucleotides for each calculation. Figure 4.4 summarises the calculation of the results.

The results were separated for each dataset. For datasets with quality values, correlation is used as the main test and statistical hypothesis tests for the functional and nonfunctional subsets, to approve the correlation results.

For datasets without quality values, Student's t-test, Wilcoxon ranked sum test and Kolmogorow-Smirnow test are used as the main tests, and correlation is used to approve

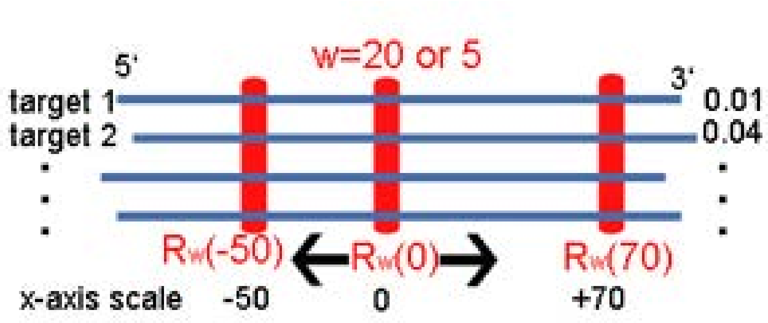


Figure 4.4.: Diagram of the analysis. The red rectangles flag three different windows and their corresponding window name. The blue lines visualise different interactions.

the results. The combination of Student's t-test and Wilcoxon ranked sum test is chosen to combine the profits of both tests: the comparable t-values and the p-values from the Wilcoxon ranked sum test which does not assume normal distribution. In both cases random data as described in Section 4.1 is used to cross check the results as a statistical null model. Results were stated as significant if their p-value is below 0.01.

5. Results and Discussion

The results generated using the methods explained in the last chapter are presented in this chapter. Headed by a general overview of highly significant regions the results for the different datasets and characteristics are described. All main tests are shown in full length in Appendix A.

5.1. Accessible Regions

Figure 5.2[1] shows the final results of the analysis of accessible regions. The analysis was done within u[2] values from 1 to 10. But low u values between 1 and 4, produce very inconsistent and noisy results. For u values between 6 and 10 the results do not vary essentially. Therefore u=8 is picked for a more in-depth analysis.

All datasets have a high accessible region downstream and all sets, except **Tafer02**, a weaker but still significant high accessible region upstream; located between $R_{20}(-170)$ and $R_{20}(-100)$, i.e. 170 to 80 nucleotides upstream to the 3' end of the target site. The high accessible region downstream is located between $R_{20}(10)$ and $R_{20}(180)$, i.e. 10 to 180 nucleotides downstream. These region is for the AtmiR, Human and Tafer03 datasets somehow split into two parts. Surprisingly the target site at approximately $R_{20}(-20)$ shows no significant results. In this regions the functional group has not only a higher accessibility than the non-functional group, but also than as random target sites like the t-values for the **Human** dataset show. These results are somehow supported by Hausser *et al.* [15], their paper rates the predictive power of flanking regions higher than the actual target site. Also Wang *et al.* [42] rate the accessibility up- and mainly downstream to be more critical than at the actual target site. They predict that these are the regions where the Argonaute protein binds.

MeanPU, maxPU and minPU show in most cases the same tendencies, if a broad region has significant results then mostly all accessibility measurements show significant results. The results with the highest significance are mostly produced by maxPU or minPU, while significant meanPU results are more rare. This observation fits to the definition of the characteristics, due to averaging meanPU is not as pronounced for a window of 20 nucleotides. Significant minPU results represent a relative low frequency of paired bases. While maxPU represents a region of size u with a high probability to be

[1] A detailed description of the plots is given in Appendix A.

[2] u is the number of bases to be unpaired for the unpaired probability PU.

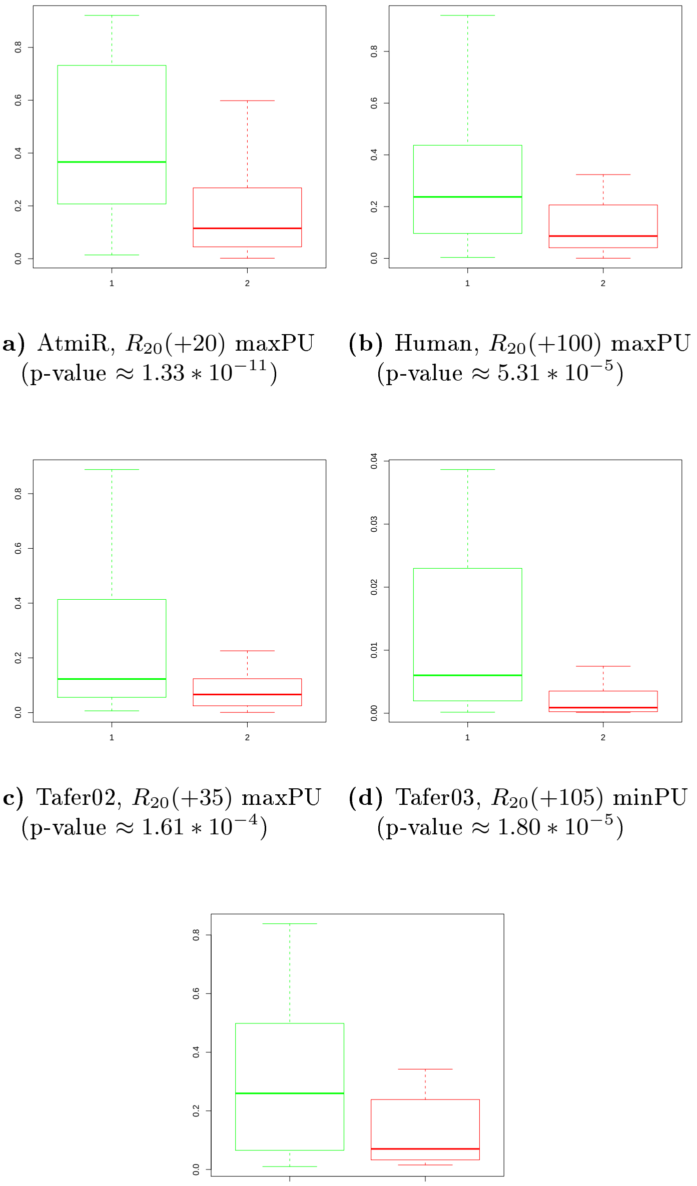
unpaired. Both are more likely to be extreme than the average over all PU within the region.

AtmiR produces the highest amount of significant results of all datasets. The highly accessible regions are $R_{20}(-160)$ to $R_{20}(-150)$, $R_{20}(0)$ to $R_{20}(50)$ and $R_{20}(80)$ to $R_{20}(120)$. The first downstream regions have the slightly higher significant results. The best result is shown in Figure 5.1 (a), with a Wilcoxon ranked sum test p-value $\approx 1.33 * 10^{-11}$ at $R_{20}(+20)$, i.e. 20 to 40 nucleotides downstream.
The most significant region for the **Human** dataset is at $R_{20}(95)$ to $R_{20}(125)$. There are also two less significant regions from $R_{20}(-200)$ to $R_{20}(-190)$ and $R_{20}(175)$ to $R_{20}(185)$. These weaker significant regions can be a result of the fact that this dataset contains interaction validated by different experiments and therefore probably a higher amount of noise.
Using the **Tafer02** dataset only one broad region generates significant results. $R_{20}(25)$ to $R_{20}(35)$ has a significant high accessibility. Altough this dataset shows lower level of significance, it shows a regions consistence with the other datasets, see Figure 5.1 (c).
The **Tafer03** has three equally significant high accessible regions, $R_{20}(-135)$ to $R_{20}(-105)$, $R_{20}(35)$ to $R_{20}(65)$ and $R_{20}(95)$ to $R_{20}(110)$.
As the only dataset **Firefly** has significant low accessible regions. E.g. $R_{20}(-5)$, $R_{20}(85)$ to $R_{20}(95)$ and $R_{20}(185)$ to $R_{20}(195)$, differing from the previous results this could be caused by the small dataset with only one target gene and the largly overlapping target sites. Additionally the high accessible regions $R_{20}(-175)$ to $R_{20}(-165)$ and $R_{20}(150)$ to $R_{20}(155)$ are very small but highly significant.

The results for regions with the lowest Wilcoxon ranked sum test p-values for each dataset are shown as boxplots in Figure 5.1. To be able to produce boxplots for the synthesised siRNA, datasets hypotheses tests used for validation are used.

5.2. Positional Entropy

The results for positional entropy are very inconsistent. As shown in Figure 5.3, significant results do exists, but they are inconsistent both within a single dataset and between all datasets. The only observable trend is that positional entropy may be lower for functional interactions. Significant low results are about 3.5 times more likely than significant high ones.



(a) AtmiR, $R_{20}(+20)$ maxPU (p-value $\approx 1.33 * 10^{-11}$)

(b) Human, $R_{20}(+100)$ maxPU (p-value $\approx 5.31 * 10^{-5}$)

(c) Tafer02, $R_{20}(+35)$ maxPU (p-value $\approx 1.61 * 10^{-4}$)

(d) Tafer03, $R_{20}(+105)$ minPU (p-value $\approx 1.80 * 10^{-5}$)

(e) Firefly, $R_{20}(+65)$ maxPU (p-value $\approx 2.46 * 10^{-4}$)

Figure 5.1.: Boxplots, excluding outlier, of the most significant hypothesis test results for each dataset with u=8. Green (1) is the functional and red (2) the non-functional group. P-values are taken from the Wilcoxon ranked sum test. All boxplots besides (d) are scaled from 0.0 to 0.8, (d) is scaled from 0.0 to 0.04 as it shows a different measurement.

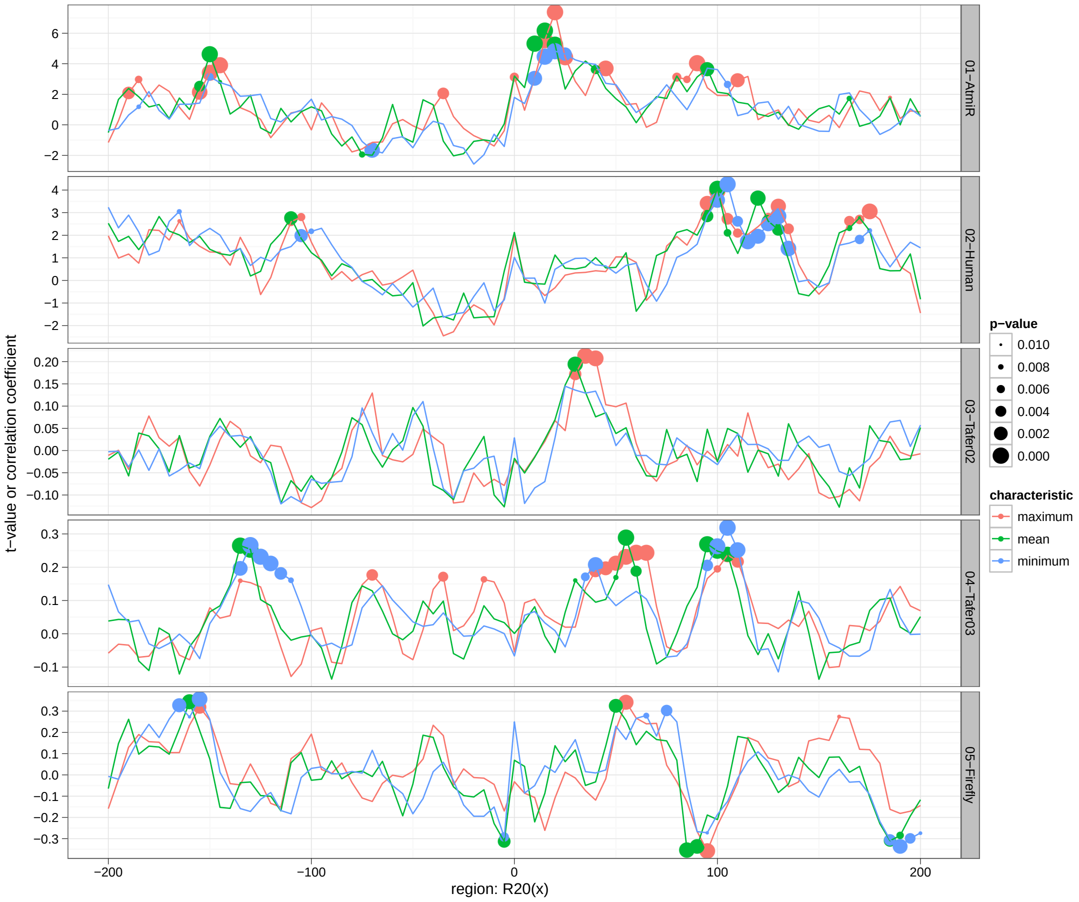


Figure 5.2.: Results for PU with u=8, Student's t-test t-values on the x-axis and Wilcoxon signed-rank test p-values indicated by the size of the dots for 1 and 2 and Spearman's rank correlation coefficient on the x-axis and p-values indicated by the size of the dots for 3 to 5. The y-axis shows the position relative to the 3' end of the target site.

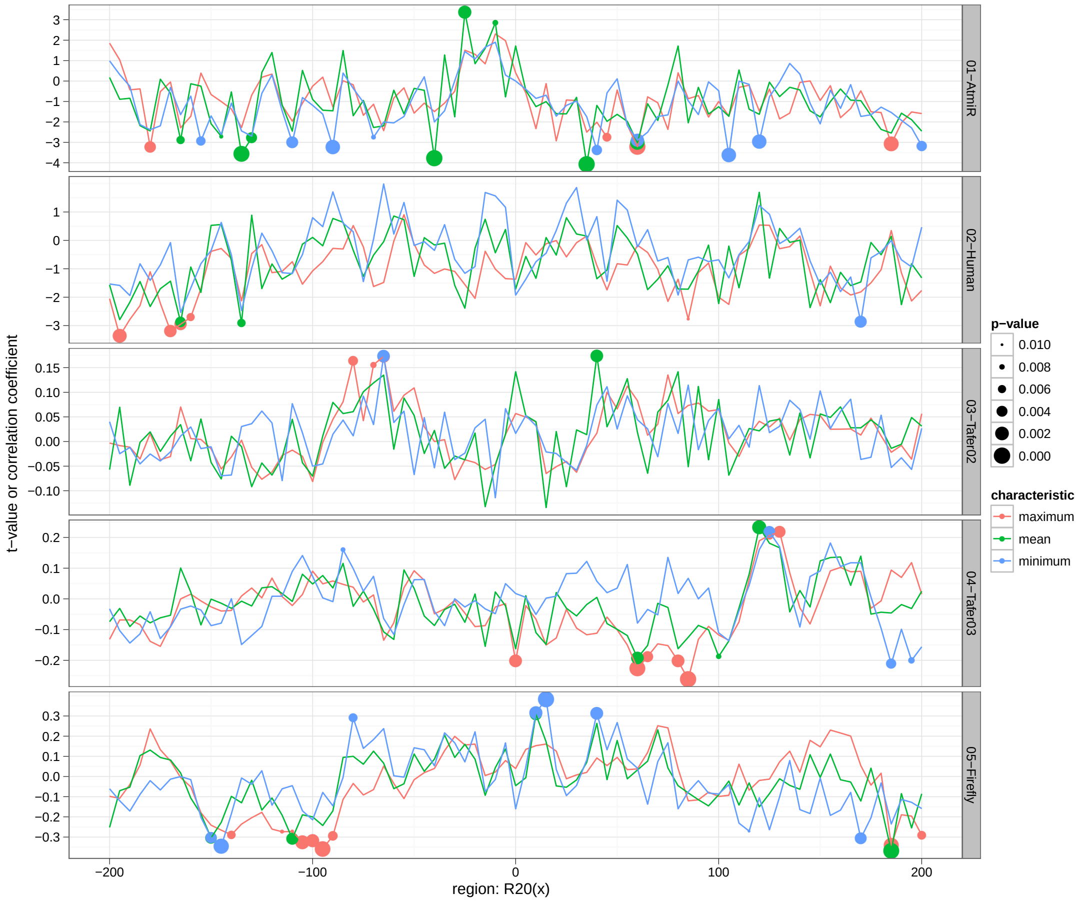


Figure 5.3.: Results for PE, Student's t-test t-values on the x-axis and Wilcoxon signed-rank test p-values indicated by the size of the dots for 1 and 2 and Spearman's rank correlation coefficient on the x-axis and p-values indicated by the size of the dots for 3 to 5. The y-axis shows the position relative to the 3' end of the target site.

5.3. Statistical Evaluation of Significance

To validate the significance of the results, various statistical tests were applied as described in Section 4.6. The differences between the hypothesis tests are marginal. Regions showing significant results for one of the hypothesis tests in most cases also show significant results in both other hypothesis tests. Some special properties of the tests are e.g. that the Kolmogorov-Smirnov test generates more noisy results and the Student's t-test generates more significant results than the other tests. As an example, the results for the different tests are shown for **AtmiR** in Figure 5.4, the other results are not shown.

The verification by correlation for endogenous datasets leads nearly to the same results as the hypothesis tests, as can be seen in Appendix A.

The verification of the synthetic datasets by hypotheses tests is more complicated. While the tendency is always observable, the exact p-values are often weaker than the original ones. The results are shown in Appendix A. **Tafer03** and **Tafer02** show good results, the **Firefly** dataset lacks in some originally significant regions. This may be caused by the small difference of interaction quality originating from the small amount of interactions.

5.3.1. Null Model Tests

The null model tests , where only random target sites were tested, resulted in very few significant results. As p-values below 0.01 are taken to be significant, random data is estimated to give one significant result every 100 tests. A single accessibility test for one u value contains 81 tests, therefore about one significant results can be expected. On average the null model tests have slightly more often significant results. But without the consistency shown if the real targets are tested.

Figure 5.5 shows the null model tests for all datasets for accessibility u=8 . The complete null model tests are not shown.

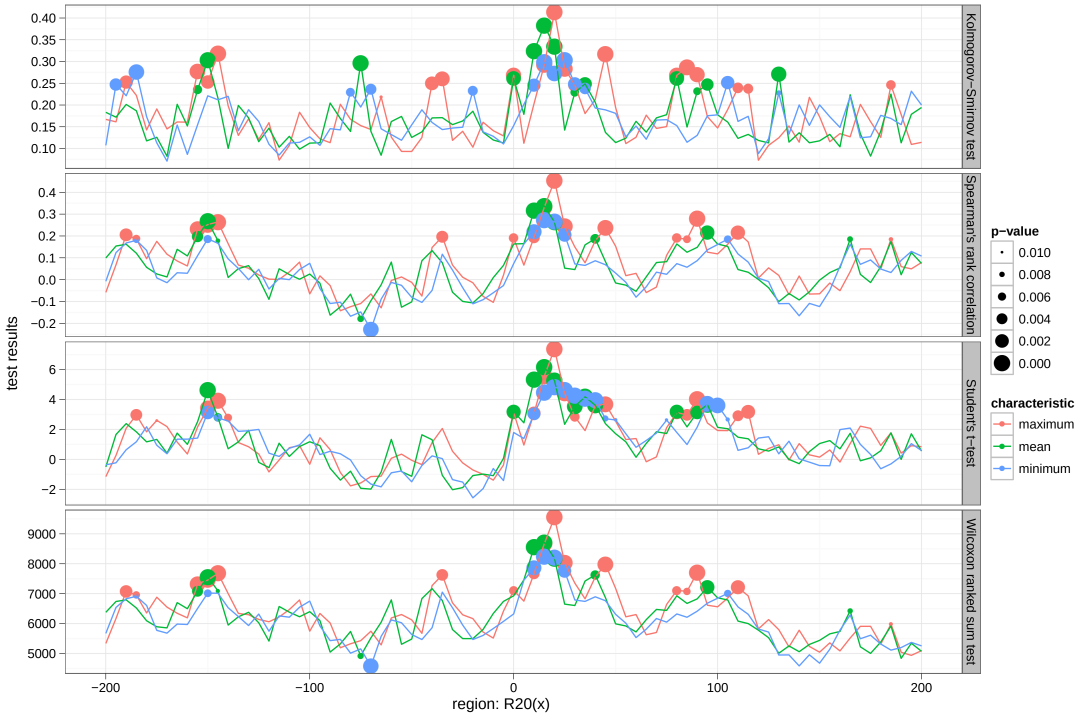


Figure 5.4.: Different statistical measurements (Kolmogorov-Smirnov, Spearman's rank correlation, Student's t-test, and Wilcoxon ranked sum test and their corresponding p-values) for the **AtmiR** dataset, PU with u=8.

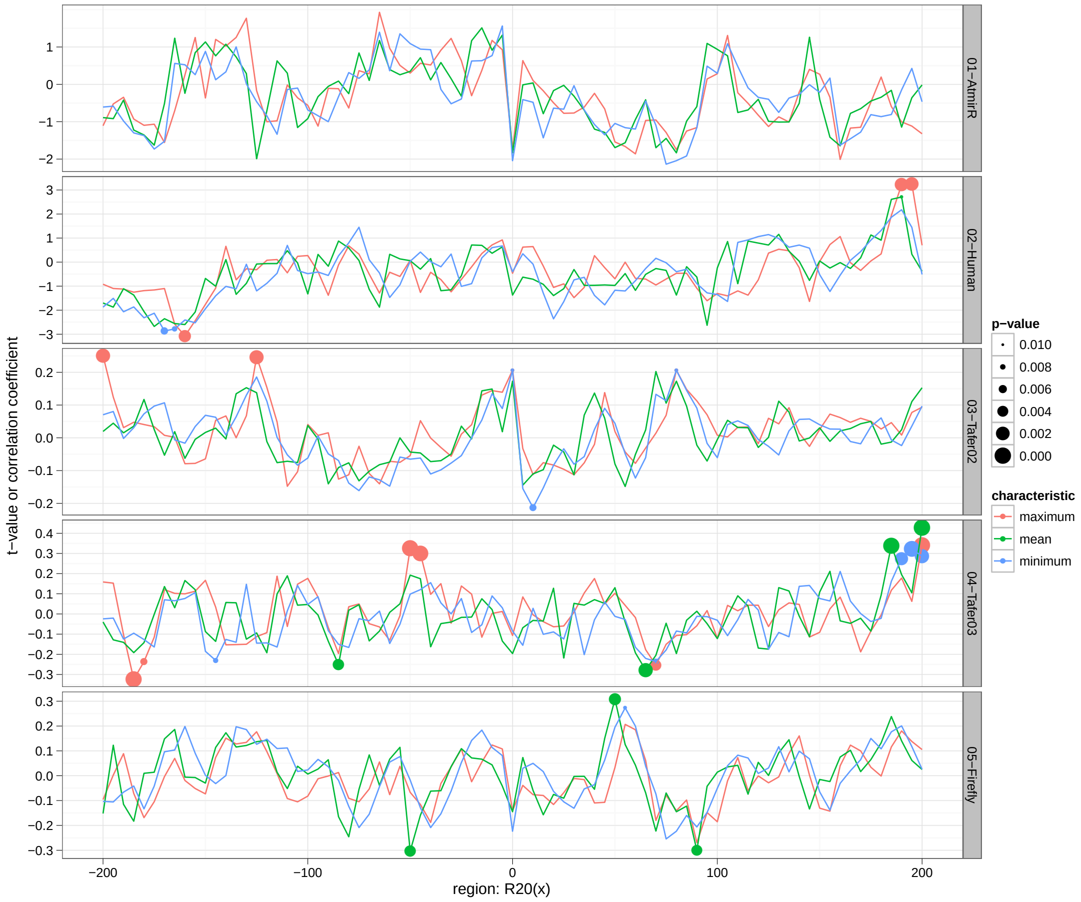


Figure 5.5.: Null model tests with random target sites for PU with u=8, Student's t-test t-values on the x-axis and Wilcoxon signed-rank test p-values indicated by the size of the dots for 1 and 2 and Spearman's rank correlation coefficient on the x-axis and p-values indicated by the size of the dots for 3 to 5.

6. Conclusion and Outlook

Using the predicted secondary structure of mRNAs and several scripts the accessibility and structural stability of the mRNAs is calculated. These calculations are used to compare the characteristics of functional and non-functional interactions and their surrounding sequences. Five datasets, three synthesised siRNA datasets two of them targeting human genes and one a firefly gene and two endogenous miRNA targeting human or *Arabidopsis thaliana* mRNA, are gathered and analysed.

The results indicate two or three high accessible regions for each dataset. These regions are actually not directly next to the target site, but between 10 and 180 nucleotides downstream or 170 to 80 nucleotides upstream. Although this work do not consider multiple testing, the different statistical measurements showing the same results confirm the presented results. Thus the goal to find characteristics which give significant and consistent results between different types of ncRNAs and species was successful. But this work could not find any influence of structural stability on the RNA-RNA interactions.

Due to time limitations some details are not checked in this work: The influences of different parameters for the secondary structure prediction, such as other values for the window size or maximal distance between base pairs. Also the influence of different window sizes for the calculation of the regions are not considered at all, altough previous work has shown no major difference in response to these factors. The measurement for structural stability probably needs to be calculated differently, as the positional entropy for each base pair is considered separately in this work. Maybe structural stability should be calculated by considering local structures as a unit. Some statistical measurements are also not tested on, for example the already mentioned multiple testing problem and z-scores which indicate the distance of the means in comparison to the background. These measurements could result in a clearer signal.

It was possible to show a significant difference in the accessibility of certain regions, but the actual values of accessibility at the target site were not characterised. Therefore these initial results can not yet be used as a filter for predicted target sites, but first the underlying biological mechanism behind these results needs to be understood. For this further computational and especially experimental analyses are required.

.

Bibliography

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[2] Tyler W. H. Backman, Christopher M. Sullivan, Jason S. Cumbie, Zachary A. Miller, Elisabeth J. Chapman, Noah Fahlgren, Scott A. Givan, James C. Carrington, and Kristin D. Kasschau. Update of asrp: the arabidopsis small rna project database. *Nucl. Acids Res.*, 36(suppl_1):D982–985, January 2008.

[3] Rolf Backofen and Wolfgang R. Hess. Computational prediction of srnas and their targets in bacteria. *RNA Biology*, 7(1):33–42, January 2010.

[4] Stephan H. Bernhart, Ivo L. Hofacker, and Peter F. Stadler. Local rna base pairing probabilities in large sequences. *Bioinformatics (Oxford, England)*, 22(5):614–615, March 2006.

[5] Julius Brennecke, David R. Hipfner, Alexander Stark, Robert B. Russell, and Stephen M. Cohen. bantam encodes a developmentally regulated microrna that controls cell proliferation and regulates the proapoptotic gene hid in drosophila. *Cell*, 113(1):25–36, April 2003.

[6] Anke Busch, Andreas S. Richter, and Rolf Backofen. Intarna: efficient prediction of bacterial srna targets incorporating target site accessibility and seed regions. *Bioinformatics (Oxford, England)*, 24(24):2849–2856, December 2008.

[7] Peter Clote and Rolf Backofen. *Computational Molecular Biology: An Introduction.* Wiley, 1 edition, September 2000.

[8] R. M. Dirks and N. A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.*, 24(13):1664–1677, October 2003.

[9] Kishore Doshi, Jamie Cannone, Christian Cobaugh, and Robin Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for rna secondary structure prediction. *BMC Bioinformatics*, 5(1):105+, August 2004.

[10] Josée Dostie, Zissimos Mourelatos, Michael Yang, Anup Sharma, and Gideon Dreyfuss. Numerous micrornps in neuronal cells containing novel micrornas. *RNA (New York, N.Y.)*, 9(2):180–186, February 2003.

[11] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of rna duplex stability. *Proceedings of the National Academy of Sciences of the United States of America*, 83(24):9373–9377, December 1986.

[12] Marcelo A. German, Manoj Pillay, Dong-Hoon Jeong, Amit Hetawal, Shujun Luo, Prakash Janardhanan, Vimal Kannan, Linda A. Rymarquis, Kan Nobuta, Rana German, Emanuele De Paoli, Cheng Lu, Gary Schroth, Blake C. Meyers, and Pamela J. Green. Global identification of microrna-target rna pairs by parallel analysis of rna ends. *Nature Biotechnology*, 26(8):941–946, June 2008.

[13] Andreas R. Gruber, Ronny Lorenz, Stephan H. Bernhart, Richard Neubock, and Ivo L. Hofacker. The vienna rna websuite. *Nucl. Acids Res.*, 36(suppl_2):W70–74, July 2008.

[14] Adam M. Gustafson, Edwards Allen, Scott Givan, Daniel Smith, James C. Carrington, and Kristin D. Kasschau. Asrp: the arabidopsis small rna project database. *Nucl. Acids Res.*, 33(suppl_1):D637–640, January 2005.

[15] Jean Hausser, Markus Landthaler, Lukasz Jaskiewicz, Dimos Gaidatzis, and Mihaela Zavolan. Relative contribution of sequence and structure features to the mrna binding of argonaute/eif2c-mirna complexes and the degradation of mirna targets. *Genome research*, 19(11):2009–2020, November 2009.

[16] Michael Hiller, Rainer Pudimat, Anke Busch, and Rolf Backofen. Using rna secondary structures to guide sequence motif finding towards single-stranded regions. *Nucl. Acids Res.*, 34(17):gkl544–e117, September 2006.

[17] Ivo L. Hofacker. Vienna rna secondary structure server. *Nucl. Acids Res.*, 31(13):3429–3431, July 2003.

[18] Jonathan A. Hollander, Heh-In Im, Antonio L. Amelio, Jannet Kocerha, Purva Bali, Qun Lu, David Willoughby, Claes Wahlestedt, Michael D. Conkright, and Paul J. Kenny. Striatal microrna controls cocaine intake through creb signalling. *Nature*, 466(7303):197–202, July 2010.

[19] Xin Hong, Molly Hammell, Victor Ambros, and Stephen M. Cohen. Immunopurification of ago1 mirnps selects for a distinct class of microrna targets. *Proceedings of the National Academy of Sciences*, 106(35):15085–15090, September 2009.

[20] Dieter Huesken, Joerg Lange, Craig Mickanin, Jan Weiler, Fred Asselbergs, Justin Warner, Brian Meloon, Sharon Engel, Avi Rosenberg, Dalia Cohen, Mark Labow, Mischa Reinhardt, Francois Natt, and Jonathan Hall. Design of a genome-wide sirna library using an artificial neural network. *Nature Biotechnology*, 23(8):995–1001, July 2005.

[21] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved predictions of secondary structures for rna. *Proc. Nati. Acad. Sci. USA*, Vol. 86:7706–7710, October 1989.

[22] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microrna target recognition. *Nature genetics*, 39(10):1278–1284, October 2007.

[23] Erwin Kreyszig. *Statistische Methoden und ihre Anwendungen.* Vandenhoeck & Ruprecht, January 1991.

[24] Jan Kruger and Marc Rehmsmeier. Rnahybrid: microrna target prediction easy, fast and flexible. *Nucl. Acids Res.*, 34(suppl_2):W451–454, July 2006.

[25] Hongwei Li, Wan X. Li, and Shou W. Ding. Induction and suppression of rna silencing by an animal virus. *Science*, 296(5571):1319–1321, May 2002.

[26] Ray M. Marin and Jiri Vanicek. Efficient use of accessibility in microrna target prediction. *Nucl. Acids Res.*, pages gkq768+, August 2010.

[27] David H. Mathews, Susan J. Schroeder, Douglas H. Turner, and Michael Zucker. *The RNA World, Third Edition (Cold Spring Harbor Monograph Series)*, chapter Predicting RNA Secondary Structure. Cold Spring Harbor Laboratory Press, 3 edition, October 2005.

[28] Pierre Mazière and Anton J. Enright. Prediction of microrna targets. *Drug discovery today*, 12(11-12):452–458, June 2007.

[29] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers*, 29(6-7):1105–1119, May 1990.

[30] Pedro P. Medina, Mona Nolde, and Frank J. Slack. Oncomir addiction in an in vivo model of microrna-21-induced pre-b-cell lymphoma. *Nature*, advance online publication, August 2010.

[31] Hyeyoung Min and Sungroh Yoon. Got target? computational methods for microrna target prediction and their extension. *Experimental & molecular medicine*, 42(4):233–244, April 2010.

[32] David W. Mount. *Bioinformatics: Sequence and Genome Analysis, Second Edition.* Cold Spring Harbor Laboratory Press, 2nd edition, July 2004.

[33] Stephan Ossowski, Rebecca Schwab, and Detlef Weigel. Gene silencing in plants using artificial micrornas and other small rnas. *The Plant journal : for cell and molecular biology*, 53(4):674–690, February 2008.

[34] Christian P. Petersen, John G. Doench, Alla Grishok, and Phillip A. Sharp. *The RNA World, Third Edition (Cold Spring Harbor Monograph Series)*, chapter The Biology of Short RNAs. Cold Spring Harbor Laboratory Press, 3 edition, October 2005.

[35] Sébastien Pfeffer, Mihaela Zavolan, Friedrich A. Grässer, Minchen Chien, James J. Russo, Jingyue Ju, Bino John, Anton J. Enright, Debora Marks, Chris Sander, and Thomas Tuschl. Identification of virus-encoded micrornas. *Science (New York, N.Y.)*, 304(5671):734–736, April 2004.

[36] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2010.

[37] Jens Reeder, Peter Steffen, and Robert Giegerich. pknotsrg: Rna pseudoknot folding including near-optimal structures and sliding windows. *Nucleic acids research*, 35(Web Server issue), July 2007.

[38] Steffen Schubert, Arnold Grünweller, Volker A. Erdmann, and Jens Kurreck. Local rna target structure influences sirna efficacy: systematic analysis of intentionally designed binding regions. *Journal of molecular biology*, 348(4):883–893, May 2005.

[39] Murray Spiegel and Larry Stephens. *Schaum's Outline of Statistics (Schaum's Outline Series).* McGraw-Hill, 4 edition, November 2007.

[40] Hakim Tafer, Stefan L. Ameres, Gregor Obernosterer, Christoph A. Gebeshuber, Renee Schroeder, Javier Martinez, and Ivo L. Hofacker. The impact of target site accessibility on the design of effective sirnas. *Nature Biotechnology*, 26(5):578–583, April 2008.

[41] Seyedtaghi Takyar, Robyn P. Hickerson, and Harry F. Noller. mrna helicase activity of the ribosome. *Cell*, 120(1):49–58, January 2005.

[42] Wang-Xia X. Wang, Bernard R. Wilfred, Kevin Xie, Mary H. Jennings, Yanling Hu, Arnold J. Stromberg, and Peter T. Nelson. Individual micrornas (mirnas) display distinct mrna targeting "rules". *RNA biology*, 7(3), May 2010.

[43] Hadley Wickham. *ggplot2: elegant graphics for data analysis.* Springer New York, 2009.

[44] Feifei Xiao, Zhixiang Zuo, Guoshuai Cai, Shuli Kang, Xiaolian Gao, and Tongbin Li. mirecords: an integrated resource for microrna-target interactions. *Nucleic acids research*, 37(Database issue):D105–110, January 2009.

[45] M. Zuker and P. Stiegler. Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133–148, January 1981.

A. Appendix

The appendix is divided into three sections. The first and second are the tests results for all datasets, the Figures A.1 to A.5 shows all correlation results. While the second part, i.e. the Figures A.6 to A.10, shows all hypothesis test results. Every Figure contains one dataset. The first 10 subplots are the PU characteristics and the u values 1 to 10. The last subplot is the PE characteristic. The x-axis shows the different regions $R_{20}(x)$ respectively $R_5(x)$ for PE, between -200 and 200, relative to the 3' end of the target site. While the y-axis shows the Spearman's rank correlation coefficient or the Student's t-test t-value, depending on the exact statistical test; the p-value is indicated by the size of the dots. Only significant result, i.e. results with a p-value below 0.01, have a dot. The colours indicate the different characteristics, maximal (red), mean (green) and minimal (blue). Related to maxPU, meanPU and minPU respectively maxPE, meanPE and minPE. For the first section always the complete dataset is used, for the second one the functional and non-functional groups as described in Section 4.1 are used. The third section is a detailed description of the AtmiR dataset, as given to me by my supervisor.

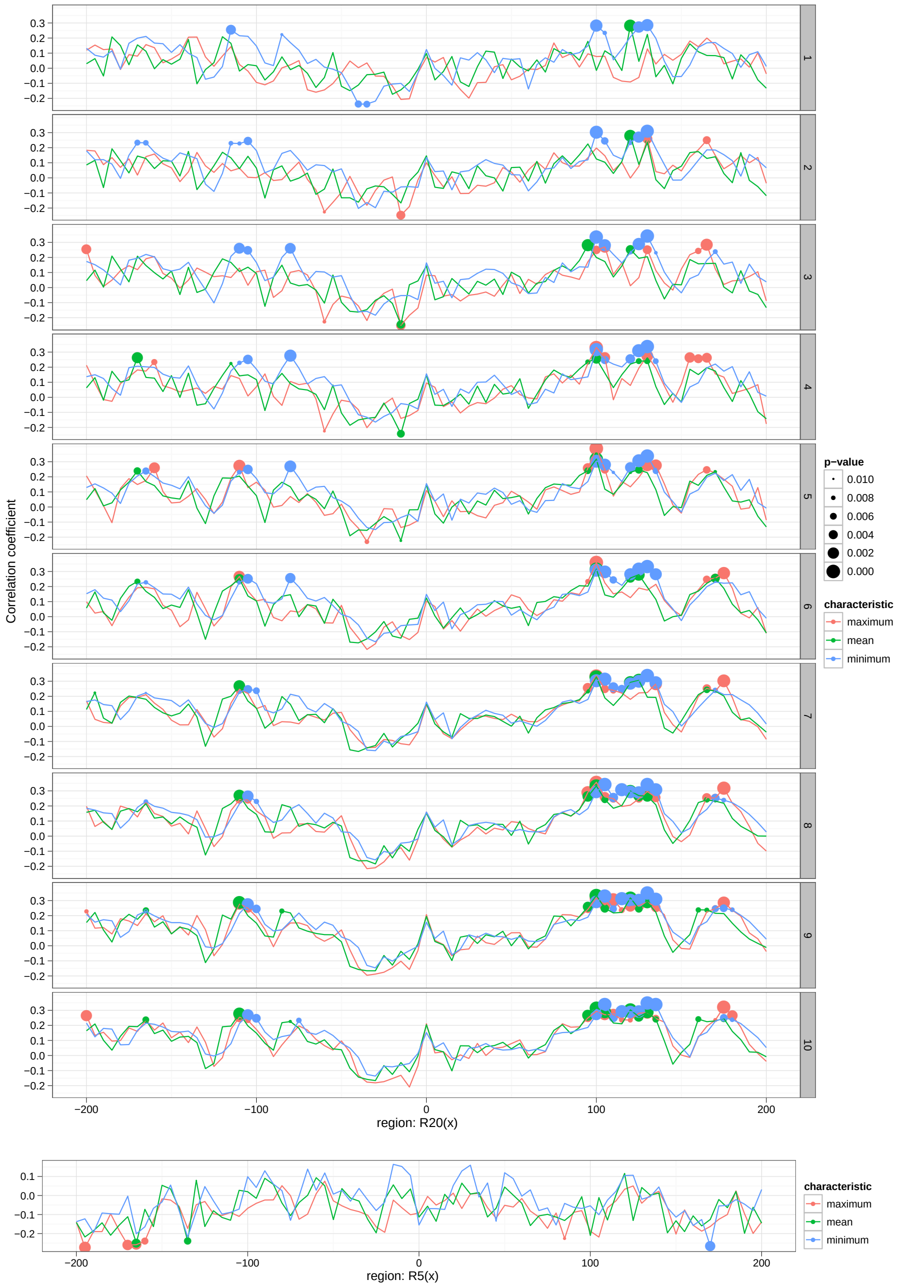


Figure A.1.: Dataset: **Human**, correlation between calculated characteristics (see Section 4.4) and quality, for all regions between $R_{20}(-200)$ and $R_{20}(200)$ (see Section 4.4.1).

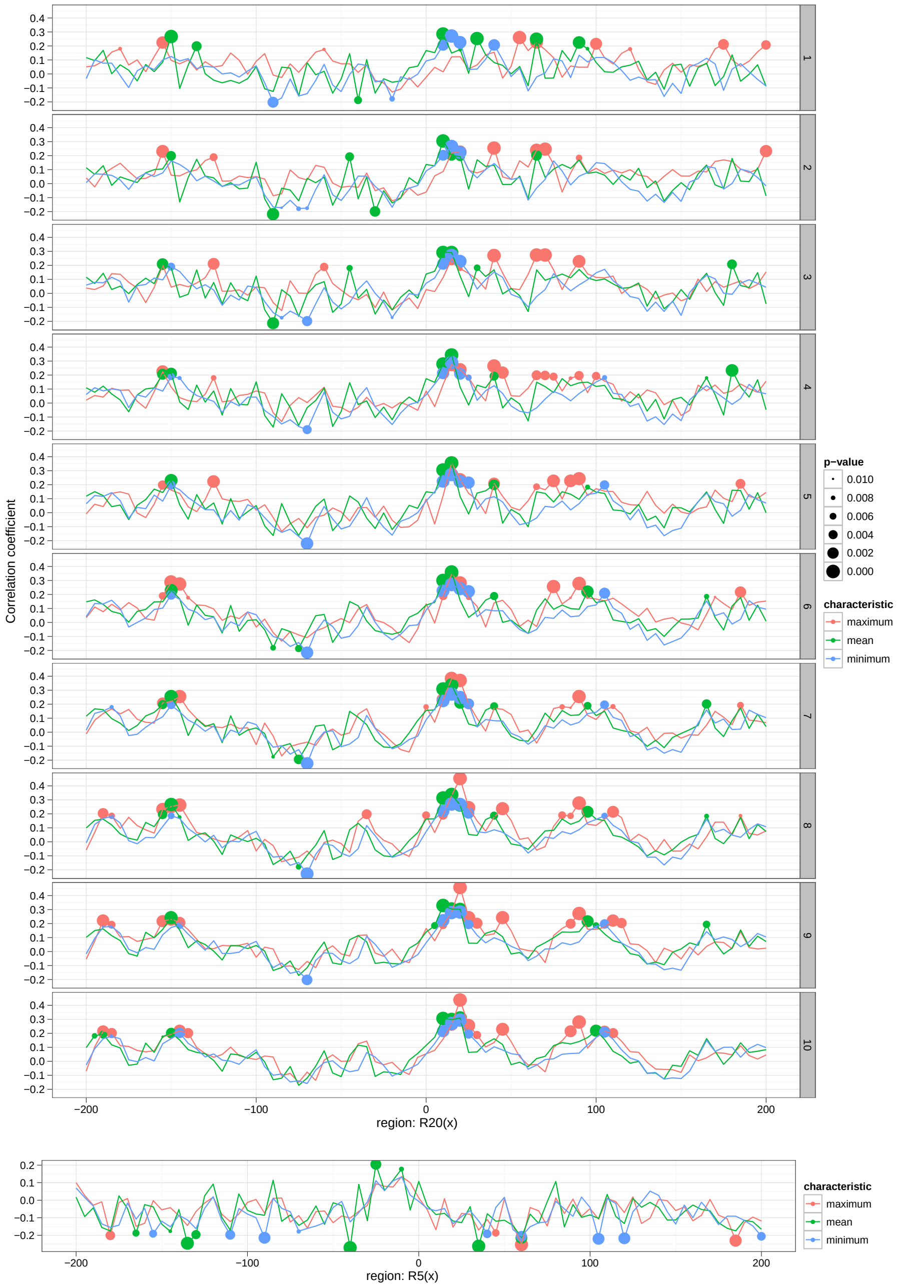


Figure A.2.: Dataset: **AtmiR** Correlation between calculated characteristics (see Section 4.4) and quality, for all regions between $R_{20}(-200)$ and $R_{20}(200)$ (see Section 4.4.1).

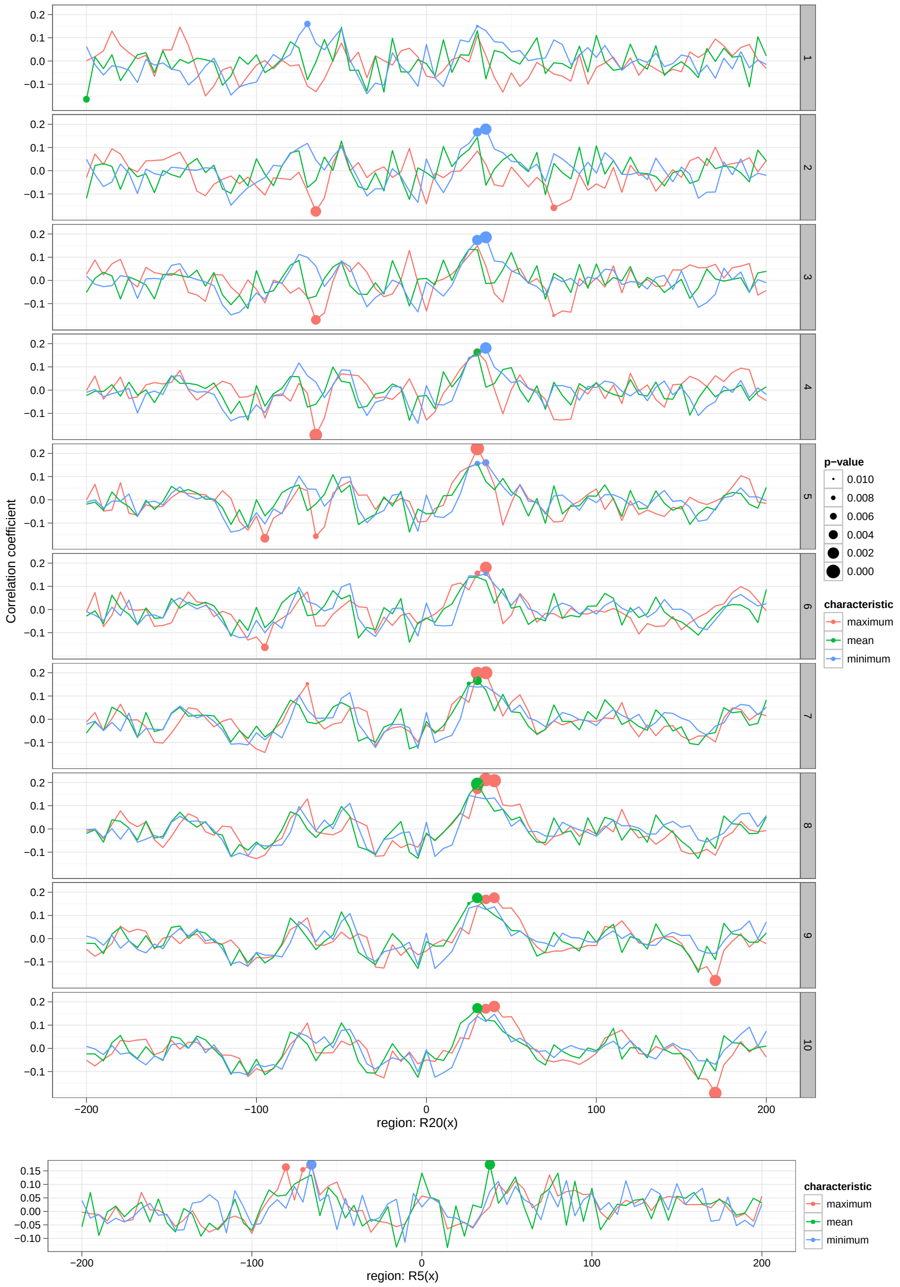


Figure A.3.: Dataset: **Tafer02** correlation between calculated characteristics (see Section 4.4) and quality, for all regions between $R_{20}(-200)$ and $R_{20}(200)$ (see Section 4.4.1).

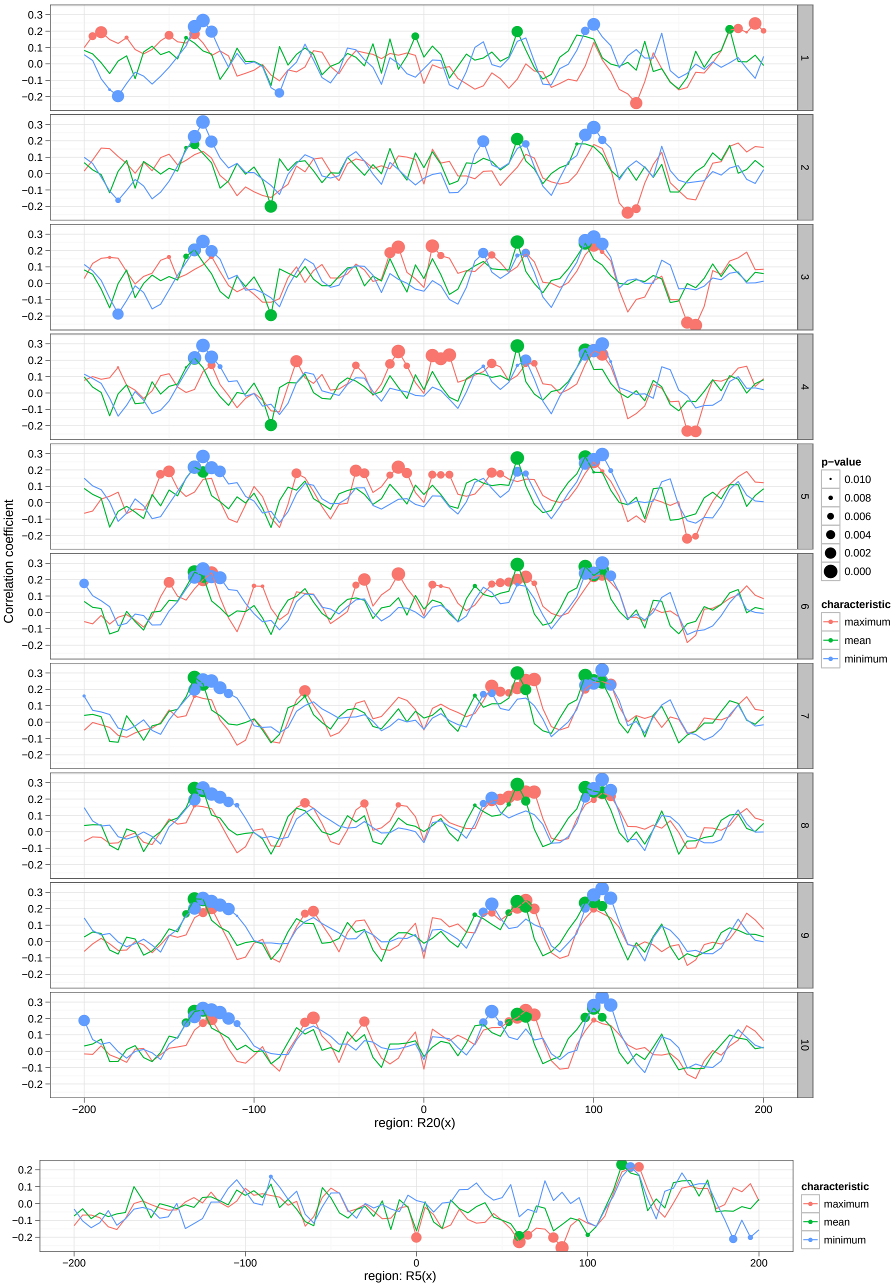


Figure A.4.: Dataset: **Tafer03**, correlation between calculated characteristics (see Section 4.4) and quality, for all regions between $R_{20}(-200)$ and $R_{20}(200)$ (see Section 4.4.1).

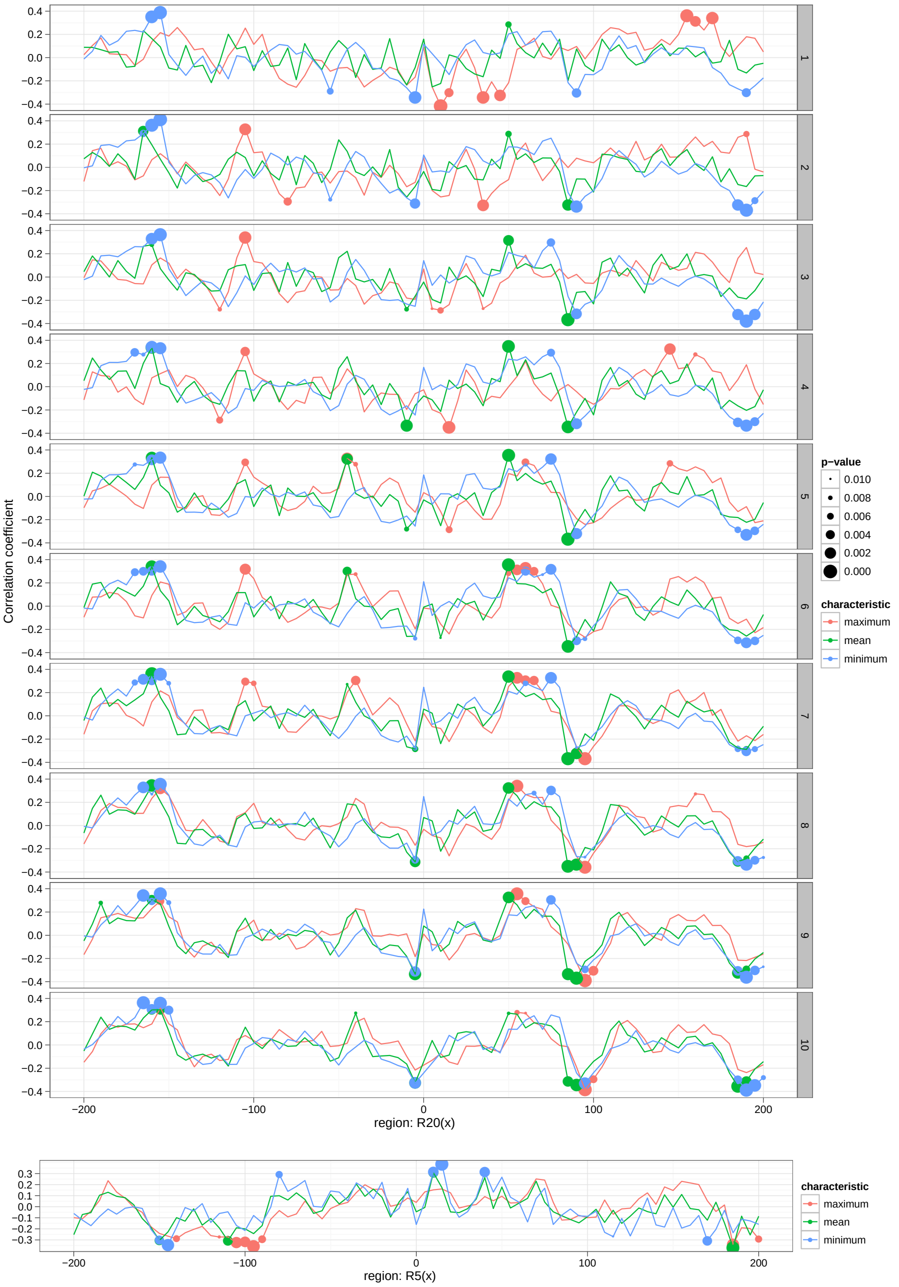


Figure A.5.: Dataset: **Firefly**, correlation between calculated characteristics (see Section 4.4) and quality, for all regions between $R_{20}(-200)$ and $R_{20}(200)$ (see Section 4.4.1).

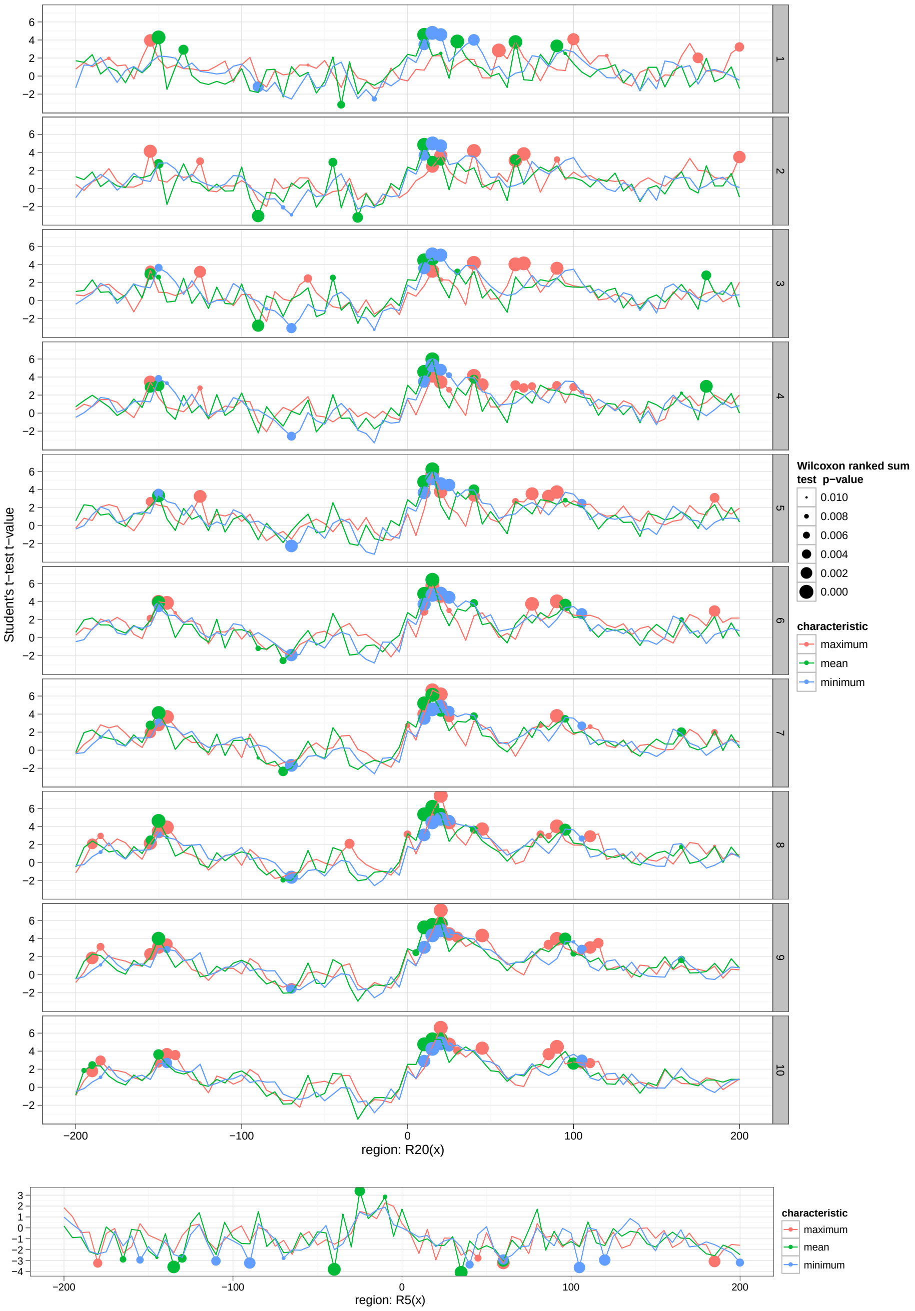
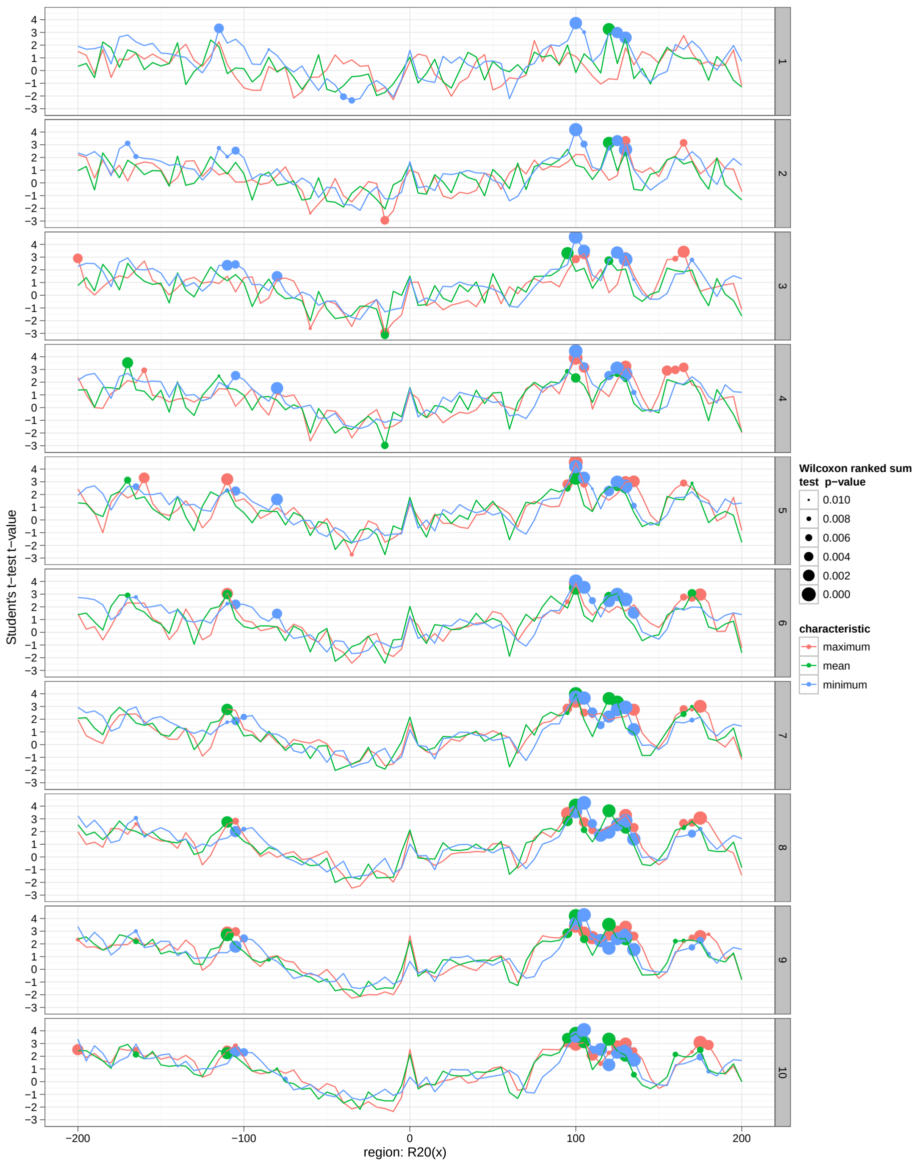


Figure A.6.: Dataset: **AtmiR**, Student's t-test between the calculated characteristics (see Section 4.4) of the functional and non-functional group, for all regions between $R_{20}(-200)$ and $R_{20}(200)$ (see Section 4.4.1).



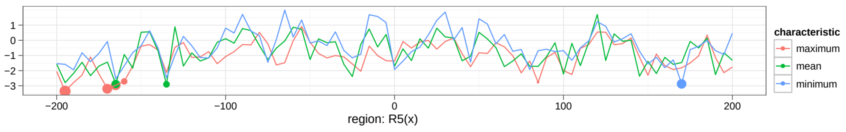


Figure A.7.: Dataset: **Human**, Student's t-test between the calculated characteristics (see Section 4.4) of the functional and non-functional group, for all regions between $R_{20}(-200)$ and $R_{20}(200)$ (see Section 4.4.1).

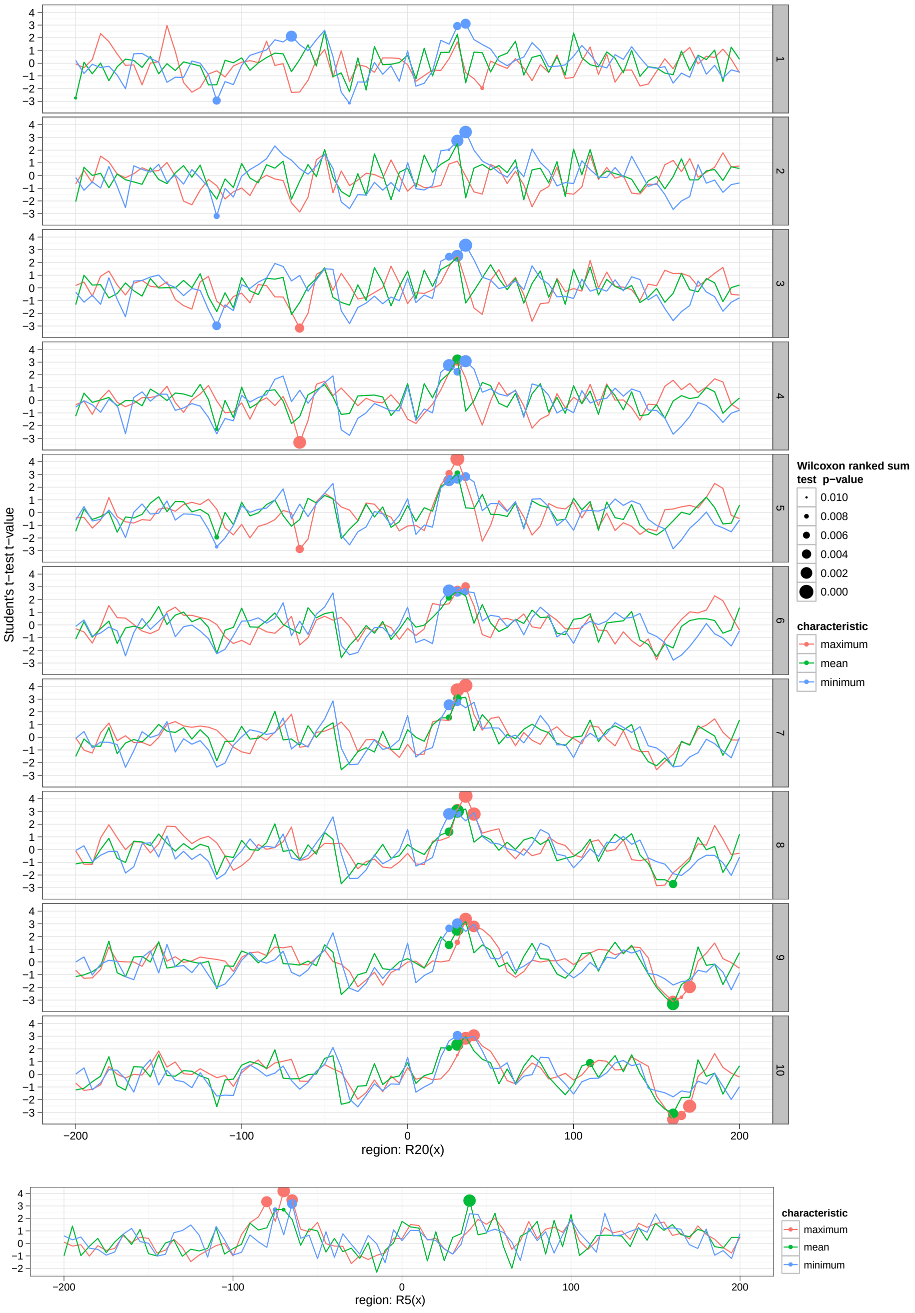


Figure A.8.: Dataset: **Tafer02**, Student's t-test between the calculated characteristics (see Section 4.4) of the functional and non-functional group, for all regions between $R_{20}(-200)$ and $R_{20}(200)$ (see Section 4.4.1).

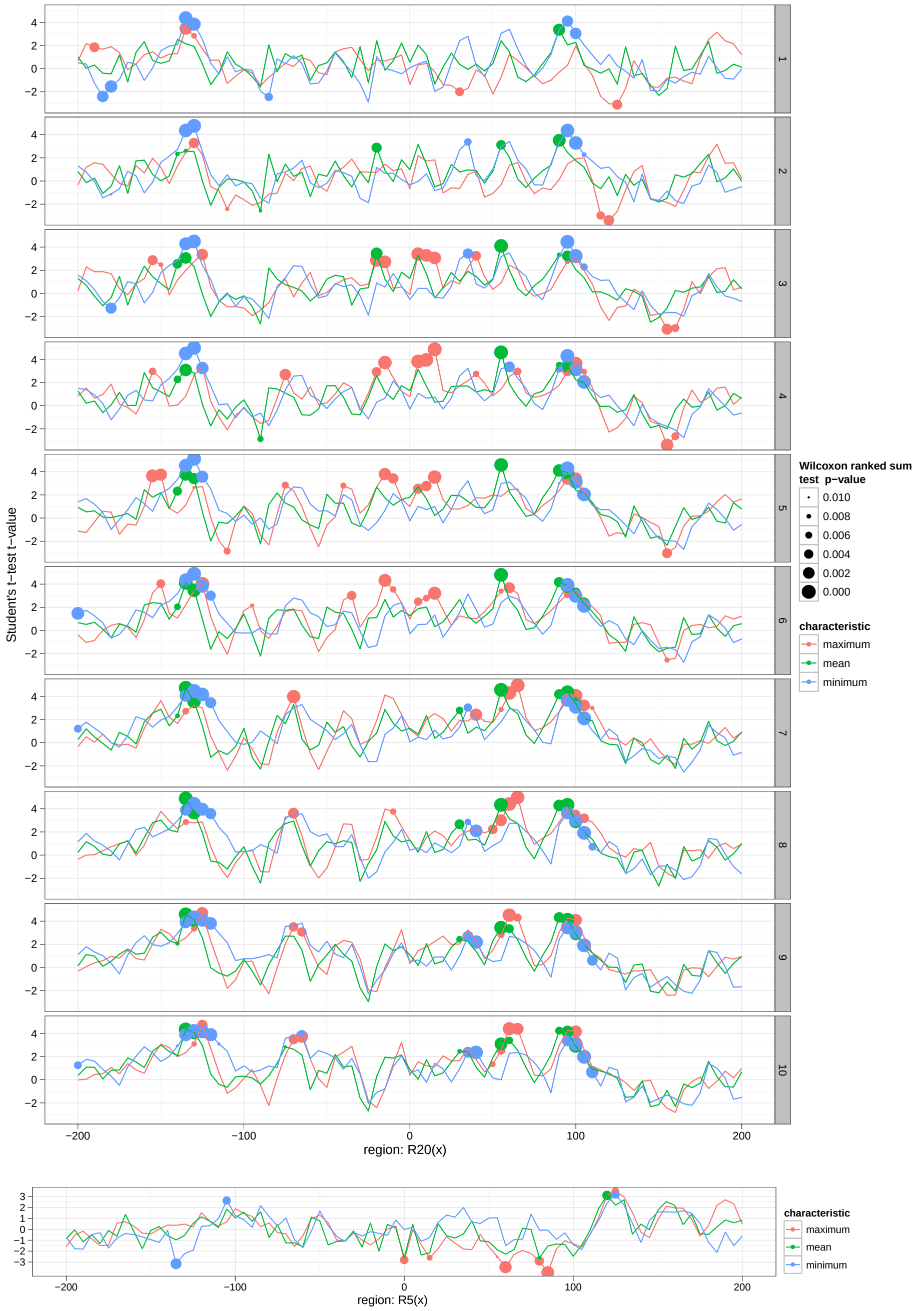


Figure A.9.: Dataset: **Tafer03**, Student's t-test between the calculated characteristics (see Section 4.4) of the functional and non-functional group, for all regions between $R_{20}(-200)$ and $R_{20}(200)$ (see Section 4.4.1).

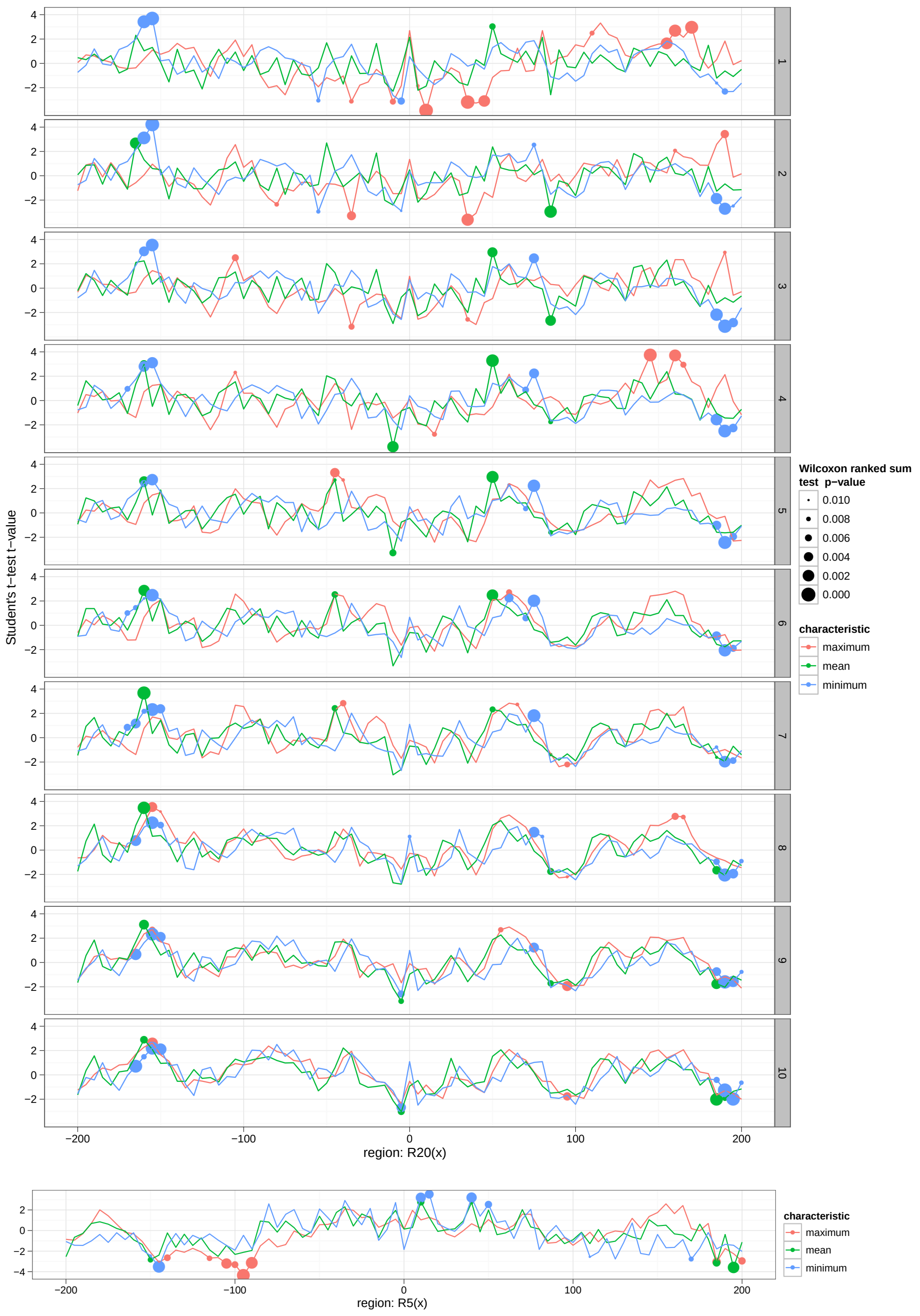


Figure A.10.: Dataset: **Firefly**, Student's t-test between the calculated characteristics (see Section 4.4) of the functional and non-functional group, for all regions between $R_{20}(-200)$ and $R_{20}(200)$ (see Section 4.4.1).

A.1. AtmiR dataset

The dataset AtMiR consists of 110 functional and 114 non-functional miRNA target sites in Arabidopsis thaliana, where the functionality is based on experimental evidence.
The functional set was taken from a cleavage analysis performed by German *et al.* [12]. They have performed deep sequencing to identify cleavage products of miRNA degradation of target mRNAs in two cell lines: wild type col-0 and the mutant xrn4-/-. An abundance of cutting points that lie within the reverse complement sequence of known mature miRNAs are considered as evidence for a target site[1]. German *et al.* provide the miRNA and the target mRNA accession numbers, the SBS signature containing half of the miRNA recognition site, the cutting position in the cDNA, and the abundance of SBS signatures found in the two cell lines. The following steps were performed to filter and extend this data to provide more detail about the the exact hybridisation of each interaction. (1) All miRNA targets were removed that did not contain an SBS signature for *both* cell lines to maintain a high quality of the data. (2) Most of the miRNA are found in more than one locus and therefore exist in different variants that differ in sequence at the 3' end. Thus, a new entry was made for each variant(s) with identical mature sequence(s). (3) To identify the target site on the mRNA, a BLAST [1] search was performed with the reverse complement of each miRNA sequence. The best hits that coincided with the given cutting points were used to identify the exact target site and its position in the cDNA. (4) Finally, a prediction was made of the hybridisation between miRNA and mRNA target site by IntaRNA [6]. Due to the fact that the two sequences are largely complementary, these hybrid predictions should be fairly accurate.
It is a very difficult task to gather a set of verified non-functional miRNA target sites and thus no datasets, large enough for a statistical analysis, exist so far. Most non-functional sites found in the literature are due to mutation experiments and are therefore not native. The task is to gather a set of potential target sites of known miRNAs that *look* functional, but experimental evidence suggests that they are not. Therefore, the results from the Target Search prediction method, which is part of the Web MicroRNA Designer WMD3 [33], was used to predict potential target sites and these were filtered according to two criteria. (1) All mRNA targets given for each miRNA from the ASRP [2, 14] database were removed and (2) the expression data given by the ASRP database was used to delete those mRNAs from the set that showed more than 5 % knock down in the dicer mutant dcl1-7 in comparison to the wild type col-0. Also all pairs were removed that showed no expression of either the miRNA or the mRNA. For the GEO accession numbers of the expression data, see Table A.1.

[1] Only includes target sites of miRNA that result in mRNA cleavage and not inhibition.

Table A.1.: GEO accession numbers for expression data from ASRP [2, 14]

Small RNA 454 sequencing	
`col-0`	GSM154336
`dcl1-7`	GSM154361
Gene expression micorarrays	
`col-0`	GSM47011, GSM47012, GSM47013, GSM47020, GSM47021, GSM47022, GSM47049, GSM47050, GSM47051
`dcl1-7`	GSM47023, GSM47024, GSM47025, GSM47026, GSM47027