HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Electronics, Communications and Automation

Kalle Karhu

EXPLORING THE EFFECT OF DIFFERENT MICRORNA TARGET PREDICTION TECHNIQUES

Thesis submitted for examination for the degree of Master of Science in Technology

Espoo 11.12.2009

Thesis supervisor:

Professor Jorma Tarhio

Thesis instructor:

Professor Harri Lähdesmäki

Tekijä: Kalle Karhu

Työn nimi: MikroRNAiden kohdepaikkojen ennustaminen eri menetelmin

Päivämäärä: 11.12.2009 Kieli: Englanti Sivumäärä: 7+57

Tiedekunta: Elektroniikan, tietoliikenteen ja automaation tiedekunta

Professuuri: Ohjelmistojärjestelmät Koodi: T-106

Valvoja: Professori Jorma Tarhio

Ohjaaja: Professori Harri Lähdesmäki

MikroRNAiden kohdepaikkojen ennustaminen on verrattain uusi tutkimuskenttä, jossa tarkoituksena on ennustaa noin 22 nukleotidin mittaisten RNA-sekvenssien käyttäytymistä. On osoitettu, että mikroRNAt säätelevät geenien ekspressiotasoja eläimissä ja kasveissa sitoutumalla lähetti-RNA:ihin ja siten estämällä niiden translaatiota proteiineiksi. Ensimmäiset mikroRNAiden kohdepaikkoja ennustavat työkalut esiteltiin vuonna 2003.

Kohdepaikkojen ennustaminen on erityisen vaikeaa eläinkuntaan kuuluvissa organismeissa. Ongelmat aiheutuvat osin jo tunnettujen mekanismien monimutkaisuudesta ja osin siitä, että kaikkia mekanismeja, jotka liittyvät mikroRNA:n sitoutumiseen lähetti-RNA:han, ei vielä täysin tunneta.

Tässä diplomityössä esitellään seitsemän käytössä olevaa mikroRNAiden kohdepaikkoja ennustavaa työkalua ja yksi uusi työkalu. Esitelty uusi työkalu käyttää geneettistä algoritimia mikroRNAn ja lähetti-RNA:n rinnastuksessa käytettävien parametrien optimointiin.

Tätä työkalua verrataan tunnettuun ja tunnustettuun miRanda työkaluun. Saadut tulokset osoittavat, että GA-pohjainen työkalu saavuttaa mikroRNA–lähetti-RNA luokittelussa samoilla spesifisyys arvoilla tasaisesti korkeampia sensitiivisyys arvoja, kuin miRanda.

Lisäksi esitetyt tulokset tukevat hypoteesia vähintään kahden eri tyyppisen mikroRNA–lähetti-RNA dupleksin olemassaolosta.

Avainsanat: mikroRNAt, mikroRNAiden kohdepaikat, geneettiset algoritmit

Author: Kalle Karhu

Title: Exploring the effect of different microRNA target prediction techniques

Date: 11.12.2009 Language: English Number of pages: 7+57

Faculty: Faculty of Electronics, Communications and Automation

Professorship: Software Systems Code: T-106

Supervisor: Professor Jorma Tarhio

Instructor: Professor Harri Lähdesmäki

MicroRNA target prediction is a relatively new field, predicting the actions of approximately 22 nt long RNA sequences, which are shown to cause translational repression in animals and plants. First prediction tools were introduced in 2003.

In animals, the prediction is computationally extremely challenging. This is due to the various complex mechanics involved and the fact that the biological phenomenon of miRNA-mRNA binding is still largely unknown.

This thesis provides a condensed overview of the field, by presenting seven existing and one new animal miRNA target prediction tool. The new prediction tool uses a genetic algorithm to optimize affine gap alignment parameters used in the prediction.

This tool is compared with a popular and established tool, miRanda. The results suggest that the GA-based tool produces more accurate target predictions. It is shown that the GA-based target predictor outperforms the miRanda tool in the classification of potential miRNA-mRNA interactions, consistently resulting in higher sensitivity values with identical specificity values. It is additionally shown that the affine gap alignment parameters produced by the GA result in better performance than a set of hand tuned parameters used by the miRanda target prediction tool.

The results presented in this thesis additionally give rise to the hypothesis of the existence of at least two different types of miRNA-mRNA duplexes.

Keywords: microRNAs, microRNA targets, genetic algorithms

Acknowledgements

The work reported in this thesis was carried out in the Department of Computer Science and Engineering of the Helsinki University of Technology. The research was supported by the Academy of Finland.

First, I want to thank Professors Jorma Tarhio, Sami Khuri and Harri Lähdesmäki for help and guidance regarding this work and the world of microRNAs in general. I would like to thank my former colleague Juho Mäkinen for the productive time we spent on various subjects, including miRNA target prediction.

I would also like to thank my friends and colleagues for bearing me and helping me out during the previous years and this work.

Finally, I wish to thank my brother and my parents for all their love and support throughout my life, and my common-law wife Hanna for being her wonderful self, bearing my ups and downs.

Otaniemi, 11.12.2009

Kalle A. V. Karhu

Contents

Abbreviations

3' UTR	Three prime untranslated region
5' UTR	Five prime untranslated region
dsRNA	Double-stranded RNA
FPR	False positive rate
GA	Genetic algorithm
MFE	Minimum free (folding) energy
miRNA	MicroRNA
miRNP	MicroRNA ribonucleoprotein particle
mRNA	Messenger RNA
RAN-GTP	RAs-related nuclear protein bound to guanosine triphosphate
RISC	RNA induced silencing complex
ROC	Receiver operating characteristic
snoRNA	Small nucleolar RNA
SVM	Support vector machine
TPR	True positive rate

1 Introduction

1.1 Problem setting

MicroRNAs (miRNAs) are approximately 22 nucleotide long, non-coding ribonucleic acids, which do their part in controlling the translation from a messenger RNA to a protein in animals and plants [43]. The first miRNAs have been discovered in the year 1993 [31]. Even though the miRNAs have been a very hot topic of study, especially in the 21st century, a large variety of open questions still exist in the field of miRNA study.

From the computational point of view, there are multiple problems related to miRNA research. The most notable two fields are the *miRNA gene finding* and the *miRNA target prediction*. The focus of this work is in the field of microRNA target prediction.

Currently there are about ∼9500 known miRNAs, but only ∼1300 experimentally verified *miRNA targets*, meaning the mRNAs (messenger ribonucleic acids) involved in the miRNA-mRNA interactions [17, 38]. Experimentally verified microRNA targets are mRNAs whose expression levels are affected by the presence of a microRNA. It has been shown that a single miRNA can have multiple target mRNAs and a single mRNA can be targeted by multiple different miRNAs [38]. This suggests that a very large number of miRNA-mRNA interactions are yet to be found. The number of possible miRNA-mRNA pairs is so large that target prediction tools are needed to guide the search for new miRNA targets.

The miRNA binding, which results in the regulation of the translation, happens quite differently in animals and plants, especially from a computational point of view. In plants, the binding is usually perfectly complementary for 7 nucleotides near the 5' end of the miRNA [23]. Because of the relatively simple mechanics, the miRNA target prediction in plants is quite straightforward. In animals however, the binding of this 5' end of the miRNA is often imperfect, allowing some mismatches and other anomalies. The exact regularities and rules controlling the animal miRNA binding are not extensively known. Allowed gaps, mismatches and other variable properties of the target sites make the computational problem of animal miRNA target prediction more challenging and interesting, compared to plants. The focus of this thesis is particularly on the animal miRNA target prediction.

Animal miRNAs are involved in a large number of various functions. MiRNAs have been associated with cell proliferation, cell death and fat metabolism [32]. MiRNA induced gene regulation has also been shown to play a role in neurodegenerative diseases, such as Parkinson's disease, Alzheimer's disease, cerebellar neurodegeneration and ataxia, and fragile X mental retardation syndrome [36]. It has also been shown that microRNAs have an important role in the antigenic variation the parasite *Trypanosoma brucei* uses, in order to escape protective counter-mechanisms of its victim organism [34].

So far it seems like the miRNA targets can have any kind of functions in an organism. MiRNAs are just a general translation regulation system. As such, the improvements in the comprehension of miRNA-mRNA interactions aid a very broad range of biological and medical research.

The animal miRNA target prediction has been an active topic in the field of bioinformatics, especially during the last six years. Numerous target prediction tools have been published and new ones are still being published quite actively. The results and performance of these tools vary a lot as do the methods used in them. This thesis presents seven recognized prediction tools, providing a condensed view of the state of the miRNA target prediction field. Additionally, design and results of a new target prediction tool are presented.

1.2 Contributions of the thesis

In this thesis, a new miRNA target prediction tool using genetic algorithms is presented. This GA-based predictor is compared with a popular and recognized target prediction tool, miRanda [22]. Parts of the results are based on the article [25], while most are unpublished. Moreover, this thesis presents a broader analysis of the presented results. The main novelty of the presented tool lies within the implementation of the genetic algorithm. Additionally, the tool uses the genetic algorithm to optimize alignment parameters, which has not been done before in miRNA target prediction. In addition, this thesis provides a condensed overview of the current key tools of miRNA target prediction.

1.3 Structure of the thesis

This thesis is organized as follows. Section 2 gives biological background information of miRNA forming and miRNA-mRNA binding. This background information helps greatly in understanding the workings of miRNA target prediction algorithms and the challenges existing in this field.

In Section 3, seven miRNA target prediction tools, divided into traditional target prediction tools and machine learning tools, are presented. Subsections 3.1.1 and 3.2.1 include some necessary background information about *dynamic programming* and *genetic algorithms*, which help to understand the prediction tools described.

Section 4 introduces a new miRNA target prediction tool, which uses a genetic algorithm to find optimal scoring and classification of miRNA-mRNA pairs.

Lastly, the conclusions in Section 5 wraps up the work. Section 5.1 summarizes the thesis, and Section 5.2 discusses the results of the presented tools and possible future work.

2 Biological background

Micro ribonucleic acids, or miRNAs, are relatively short, non-coding RNA sequences. As in all RNAs, the nucleotides in a miRNA sequence contain a ribose sugar, with carbons numbered 1' through 5'. A base is attached to the 1' position, generally adenine (A), cytosine (C), guanine (G) or uracil (U). A phosphate group is attached to the 3' position of one ribose and the 5' position of the next, connecting a chain of nucleotides to form a sequence. The common 5' to 3' direction of reading is carried over from DNA-data to RNA and miRNA data, if not explicitly specified otherwise. The length of miRNA sequences varies from 21 to 23 nucleotides. [43]

In 1993 the first one of these RNAs, named lin-4 after the gene it was transcribed from, was identified in the worm *Caenorhabditis elegans*. The responsible researchers discovered that this tiny RNA sequence had a role in regulating the activity levels of the lin-14 gene to produce its corresponding protein. This miRNA was binding to the *three prime untranslated region* (3' UTR) of the mRNA transcribed from the gene, repressing the probability of the mRNA to be translated into a protein. This sort of regulation, where the regulating factor originates from the same organism is called endogenous regulation. [31]

After the first findings in 1993, more endogenous miRNAs regulating the gene expression levels in numerous animals and plants have been found. In September 2009, there were 9499 miRNA entries from 103 species in the miRNA database [17] and the database of experimentally verified miRNA targets, TarBase, contained 1333 miRNA targets [38]. The amount of these verified targets has risen slowly — in 2003 there were not enough of them to be used as a real verification measure for the first wave of target prediction algorithms designed in that era [32, 14]. Typical verification of predicted (or randomly chosen) targets commonly means observation of the gene expression levels coupled with under- or over-presentation of a miRNA. This means that only the miRNA targets are verified, rather than the actual sections of the mRNAs where the miRNA binds, commonly referred to as *miRNA target sites*.

2.1 MiRNA formation

MicroRNAs either originate from genes dedicated to their creation or are processed from a variety of different RNA species via the enzyme Dicer. These various RNA species include 3' UTRs, long noncoding RNAs, transposons, snoRNAs and introns. [46]

The phases of gene originated miRNA formation in animals are shown in Figure 1. It is possible for both sense and anti-sense strands to encode a miRNA [49]. MiRNA genes tend to occur either individually or in clusters of 2–7 genes that are co-expressed [12].

In plants the formation of miRNAs is slightly different due to the lack of the nuclease Drosha. The endoribonuclease Dicer is involved in a larger number of steps in plants, filling the role of the lacking Drosha nuclease. [30]

MiRNAs are first transcribed as primary transcripts (pri-miRNA) containing a cap and a poly-A tail. The pri-miRNA is required to have single stranded RNA in both 5' and 3' ends in order to be further processed. The enzymes Drosha and Pasha, together forming the Microprocessor complex, process the pri-miRNA to the pre-miRNA stage. Drosha complex cleaves the RNA molecule ∼22 nucleotides away from the terminal loop. The resulting pre-miRNA form is about 70 nucleotides long stem-loop structure. [9]

Most pre-miRNAs do not have a perfect double-stranded RNA (dsRNA) structure topped by a terminal loop. There are few possible explanations for such selectivity. One could be that dsRNAs longer than 21 base pairs activate interferon response and anti-viral machinery in the cell. Another plausible explanation could be that the thermodynamic profile of pre-miRNA determines which strand will be incorporated into Dicer complex. [18]

If the miRNA is originating from an intron, no pri-miRNA phase is involved. The intron is transformed to a pre-miRNA via the help of spliceosome and debranching enzymes. The exact debranching enzymes involved vary. As an example, in the case of *Drosophila melanogaster* the responsible enzyme is lariat debranching enzyme (Ldbr). [50]

The phases from gene to pri-miRNA and onwards to pre-miRNA occur inside the nucleus of the cell. The pre-miRNA hairpin is exported from the nucleus to cytosol by the help of Exportin-5 and RAN-GTP [9, 50].

The transformation from the pre-miRNA to a mature miRNA happens in the cytosol and is catalyzed by the endoribonuclease Dicer. The stem-loop of the pre-miRNA is removed, resulting in two complementary short RNA molecules. Argonaute protein selects one of these strands, based on the stability of the 5' end of the two sequences and connects it to the RNA induced silencing complex (RISC). The remaining sequence called miRNA* or passenger strand, is degraded as RISC substrate. [9]

The protein complex that is involved with the mature miRNA has been named both RISC and miRNP (miRNA ribonucleoprotein particle) complex in several studies. It is not certain if these variable protein complexes are actually the same complex or not [47]. The two complexes have some components in common and are of similar size [21]. In this thesis, the protein complex which accompanies a mature miRNA will be referred to as RISC.

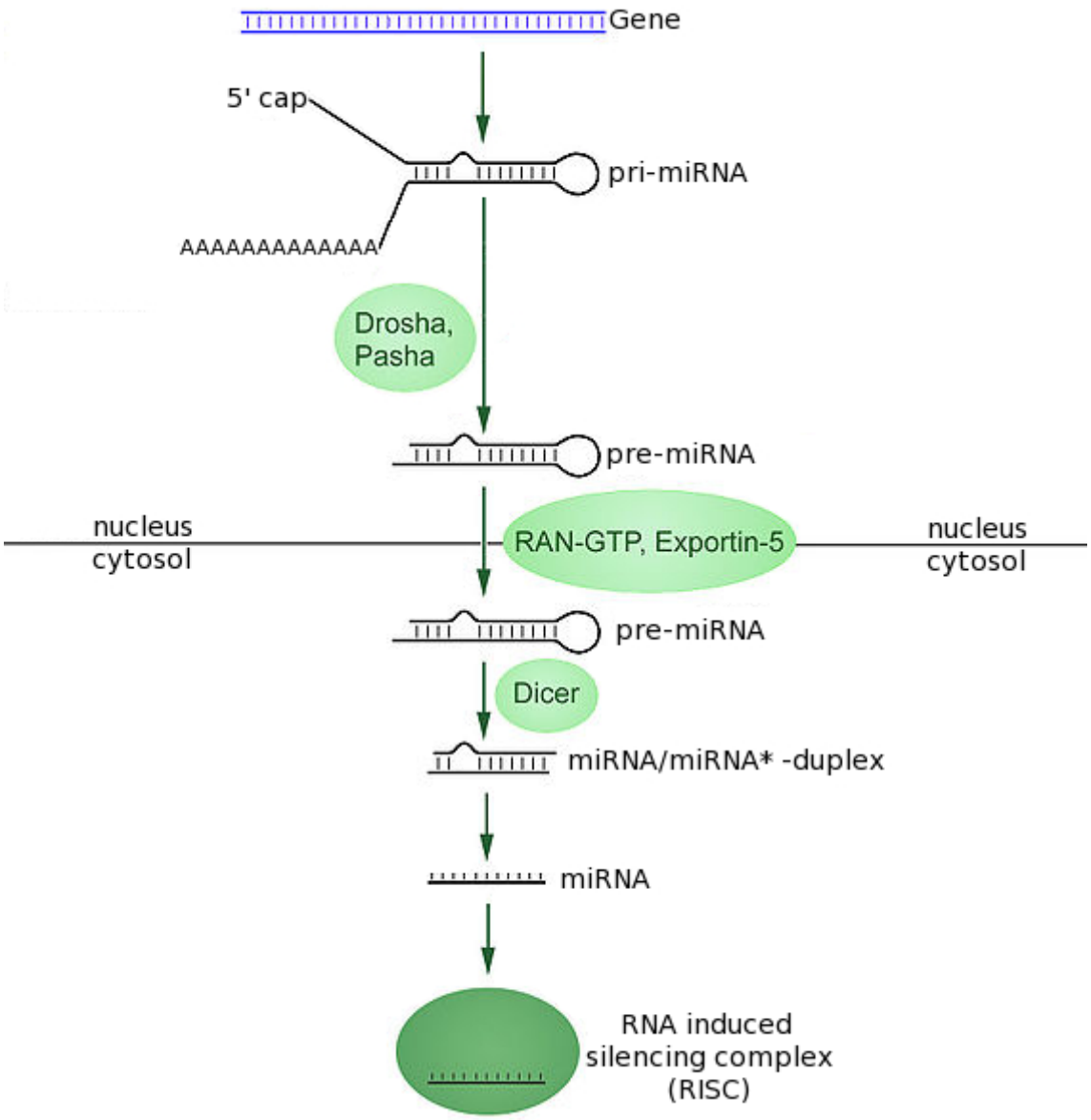

Figure 1: Common miRNA formation pathway in animals.

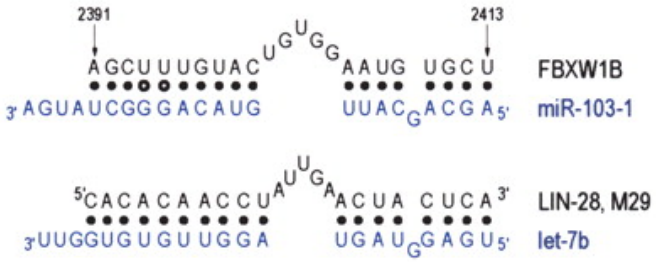

Figure 2: Predicted miRNA-mRNA binding by the algorithm Diana MicroT for miRNA let-7 on the 3' UTR of LIN-28 gene, and for miRNA miR-103-1 on the 3' UTR of FBXW1B gene [28].

2.2 MiRNA binding to mRNA

The exact regularities of the miRNA binding to mRNA are largely unknown. The basic concept is that the the mature miRNA in the RISC guides the complex to bind to a target site, which has a complementary sequence with the mature miRNA. When talking about complementarity or matching related to miRNA-mRNA binding, both the terms stand for Watson-Crick base pairing. In the case of RNA, this means that nucleotide base Adenine (A) is a complementary match to Uracil (U) and Cytosine is a complementary match to Guanine (G). The complementary matches are the pairs which form the most energetically stable *hybridization* between two sequences. Additionally, the nucleotide pairs can form non-Watson-Crick base pairs, or *wobble pairs*, which are less energetically stable [8]. A common wobble pair, which is taken into account by many miRNA target prediction algorithms, is the G:U wobble pair.

Typically, there is a shorter sequence within the miRNA that has strong complementarity with the target site. Such sites tend to vary in length, being typically between 6 and 9 nucleotides long. Often the sites start from the very 5' end of the miRNA. These sites, in their variable representations and forms, are commonly referred to as *seed sites*, or *seeds* for short, and this term will be often used in the latter sections of this thesis. The site in the mRNA that matches the seed site and is of similar length is referred to as a *seed match*.

There is a typical example of a predicted animal miRNA binding to the 3' UTR of mRNA in Figure 2, taken from the paper [28]. The miRNA-mRNA interaction is verified by gene expression data, but the exact binding is simply predicted using an algorithm called Diana MicroT [28]. The figure shows the predicted binding of miRNA let-7 on the 3' UTR of LIN-28 gene, and of miRNA miR-103-1 on the 3' UTR of FBXW1B gene. Both of the predicted bindings have 9nt long seeds in the 5' of the miRNAs, containing 1 gap. Additionally, there are also long matching parts in the 3' ends of the miRNAs.

There are three distinctive differences between miRNA-mRNA binding in animals and plants. In plants, there are extremely often seeds with perfectly complementary seed matches in the predicted mRNAs [23]. In these cases, which form the large majority, binding of miRNA to a mRNA initiates cleaving of the whole mRNA, which fully prevents any translation [23, 26]. In animals, and very rarely in the case of plants, there is no seed match in the predicted mRNA that would be perfectly complementary to the miRNA seed site. This means that there are some mismatches or wobble pairs present in the binding between the seed and the seed match. Such binding more often results in inhibition of protein translation by the RNA induced silencing complex RISC [51]. The inhibition of protein synthesis has been shown to be more effective when a miRNA is predicted to have multiple potential target sites in a mRNA [11]. The incomplete binding can also result to deadenylation [15], causing mRNAs to be degraded sooner, or DNA methylation [19] of promoter sites, which also affects the expression of targeted genes. The third distinctive difference is that plant miRNAs tend to have target sites in the coding regions of the mRNA [39], while for animal miRNAs this is extremely rare. Animal miRNAs target the 3' UTR sequence most commonly and many algorithms consider this as the only potential target area for new miRNA target sites.

As stated in Sections 1.1 and 1.2 of this thesis, the focus of miRNA target prediction tools in general, and the focus of this work is in the target prediction in animals. There are some additional hypothesis and assumptions about the regularities of miRNA binding in animals in addition to the common seed complementarity, which was mentioned in this section. All the hypothesis, including the concept of a seed site, are simply characteristics that fit the verified miRNA-mRNA interaction data well. There are currently no experimentally verified miRNA target sites, meaning nucleotide-by-nucleotide examples of actual miRNA binding to mRNA.

Many of the target prediction tools use secondary structure prediction of a connected miRNA - putative seed match -structure. This is based on the basic assumption that the hybridization to be formed should be energetically stable [42]. There are also hypotheses that the target site should be easily accessible in the predicted secondary structure of the mRNA [33].

Predicted regions of the mRNAs involved in miRNA-mRNA interactions have been shown to be highly conserved across species [5]. This has given rise to the hypothesis that cross-species conservation would be a good sign for a site to be potential miRNA target site, if not compulsory [29]. No specific role has been explained for the 3' ends of miRNAs even though they tend to be evolutionarily conserved over their entire lengths [27].

More experimentally verified miRNA-mRNA interactions would be needed to gain better knowledge of the actual miRNA-mRNA binding. Naturally it would be of revolutionary importance and help, if one could verify the actual nucleotide-by-nucleotide binding in numerous miRNA-mRNA interactions.

3 Existing tools and methods to predict miRNA targets

Reliable identification of miRNA targets is drastically different from standard sequence similarity analysis. In traditional sequence similarity analysis, one tries to assess the likelihood of a hypothesis, for an example, whether similarity between two sequences is due to common ancestry or a chance occurrence. However, when predicting possible targets for miRNA induced regulation, the idea is to consider the likelihood of an actual physical interaction between two molecular species, assuming the molecules in question are present in a cell at the same time at sufficient concentrations. Because of this important background feature of the problem, other methods on top of simply aligning two sequences have been used by various previous approaches attempting to find potential target sites for miRNAs.

MiRNA target prediction is a relatively new field of bioinformatics. The first prediction tools have been published in 2003, and at that time, there were not any known vertebrate miRNA-mRNA interactions available [32]. Because of the lack of verified interactions, the quality of the predictions the new miRNA prediction tools produced could not be measured. Lacking ways to measure their quality, the tools from this era simply predicted new potential targets in the articles they were introduced in.

As the number of verified miRNA-mRNA interactions has increased, it has become possible to measure the performance of the prediction tools to some extent. However, different tools produce quite different predictions, and even the most accomplished tools still have hard time to distinguish verified miRNA-mRNA pairs from other random pairs, especially in a case where the mRNAs are known to be involved in some miRNA-mRNA interactions, as can be seen later in Section 4.2. Wetlab experiments are the only way to verify the exact quality of the predictions done by each tool.

In this section, seven miRNA target prediction tools are presented. These tools are divided into two groups: traditional target prediction tools and machine learning target prediction tools. Traditional target prediction tools typically use some kind of local sequence alignment together with folding free energy-, orthology-, or some other type of thresholds. Machine learning target prediction tools try to classify miRNA-mRNA pairs using a set of experimentally verified positive and negative interactions.

Some of the tools use very similar methods with each other, but typically the small details of implementation cause differences in the prediction results. It is noteworthy, that the ways of scoring the interactions and the form of output used by the tools tends to differ quite a lot from tool to tool, which makes the comparison of tools quite challenging.

3.1 Traditional target prediction

MicroRNA target prediction tools that do not have trainable parameters affecting the predictions, are classified in this thesis as traditional target prediction tools. Typically these tools take into account the fact that miRNA targets are extremely often in the 3' UTR region of animal mRNA.

All of the traditional target prediction algorithms presented here use dynamic programming to align the miRNA with the 3' UTR. This alignment is done by either calculating the RNA secondary structure of the miRNA and the 3' UTR combined, or by using a local sequence alignment algorithm to align the two sequences. Basic examples for both of these cases are shown in Section 3.1.1. Additionally, traditional target prediction tools typically use some other filtering methods, such as requiring some level of cross species conservation of the potential target site.

3.1.1 Dynamic programming

Dynamic programming is a term that covers a certain way of reducing a larger problem to sub problems and solving them. Originally dynamic programming was just an impressive name used to shield the work of Richard Bellman from the US Secretary of Defense, Charles Wilson, who was known to be hostile to mathematics research [13]. Bellman, who worked with time series and planning, figured that dynamic programming would sound like "something not even a Congressman could object to" [13]. The term has been confusing students of bioinformatics ever since.

In the two following subsections, two common examples of dynamic programming are shown. The first example is the Smith-Waterman local alignment algorithm, and the second example is the Nussinov algorithm, which is used to predict secondary structures of an RNA. Both of these algorithms, or further improved variations of them, are widely used in miRNA target prediction tools.

Smith-Waterman algorithm

The Smith-Waterman algorithm [48] is a method to solve optimal local alignment, given certain alignment parameters. Commonly, the alignment is done between two DNA or RNA sequences. However, it is possible to handle the sequences to be aligned codon by codon, efficiently aligning letters corresponding to amino acids. It is also possible to align any type of symbol sequences with the algorithm, as long as there exist an extensive set of parameters for the alignment.

The general goal of an alignment in this sense is to find a way to arrange two sequences so that they have same nucleotides or amino acids in the matching positions, penalizing for gaps between nucleotides and mismatches between two nucleotides at matching positions. The exact balancing between matches, mismatches and gaps

are defined by the alignment parameters. This means that the resulting alignment, and the score the alignment gets are both fully defined by the alignment parameters.

Common use for Smith-Waterman algorithm is the approximate search of a RNA or DNA sequence from a larger database. One scenario would be the search of a certain gene from a genome. The gene can be from another species or subject to mutations, which is why exact search is out of the question. Different builds of genomes of the same species also have slight differences between them, so local alignment is really the way to go when searching any sort of sequence from any genetic sequence database. It is not reasonable to simply try out and score all possible alignments between two sequences. In the case of global alignments, there are about $2^{2N}/\sqrt{2\pi N}$ ways to align two sequences of length N [13]. This number gets out of hand very quickly with real sequences.

All dynamic programming algorithms consist of four basic parts: a recursive definition of the optimal score; a dynamic programming matrix for remembering optimal scores of subproblems; a bottom-up approach of filling the matrix by solving the smallest subproblems first; and a traceback of the matrix to recover the structure of the optimal solution that gave the optimal score [13].

For a simple example, let us consider two RNA sequences x and y of length M and N respectively, to be aligned. The i^{th} nucleotide of the sequence x is x_i and the j^{th} nucleotide of y is y_i. The alignment parameters for two RNA sequences consist of the scores $\sigma(a, b)$ for all 10 possible nucleotide pairings (A-A, A-C, A-G, A-U, C-C, C-G, C-U, G-G, G-U and U-U) and the gap penalty γ. Let us denote the optimal score for the alignment from the beginning to nucleotides x_i and y_i as $S(i, j)$ The recursive definition for this optimal score is:

$$S(i,j) = \max \begin{cases} S(i-1, j-1) + \sigma(x_i, y_i), \\ S(i-1, j) + \gamma, \\ S(i, j-1) + \gamma, \\ 0 \end{cases} \tag{1}$$

where the third row corresponds to aligning a gap with y_j, the second row corresponds to aligning a gap with x_i and the top row corresponds to aligning x_i with y_j. Let us select simple alignment parameters $\gamma = -6$, $\sigma(a, b)$ is 5 for a match, 1 for G-U wobble pair (stands for pairings A-G and U-C in this alignment) and -2 for other mismatches. Example local alignment with Smith-Waterman algorithm is shown in Figure 3.

In the case of Smith-Waterman algorithm, the practical application of the bottom-up approach of filling the matrix, using the smaller subproblems to solve bigger ones is rather simple. The filling of the grid starts from upper left corner, going from left to right row by row, using equation 1. Whenever a cell ends up with a positive number, an arrow is drawn to state how this score was obtained.

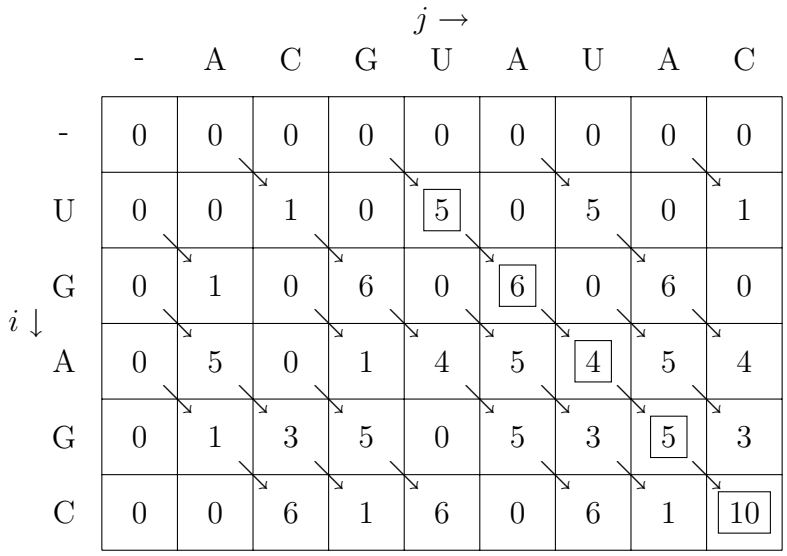

Figure 3: Smith-Waterman algorithm local alignment of RNA sequences x = 'UGAGC' and y = 'ACGUAUAC'.

The traceback of the matrix to recover the structure of the optimal solution, that gave the optimal score, is done by finding the highest score in the matrix, and extending the alignment to the last positive number. The arrows show the allowed directions for the extension of the alignment. If there is a tie between multiple options in the equation 1, there are also multiple optional local alignments. In the case of this example, the alignment resulting from the traceback was:

```
Seq_Y  ACGUAUAC    8
Seq_X  ...UGAGC    5
```

which contains 1 mismatch, 2 G-U wobble pairs (both A-G) and 2 matches. The score for this local alignment using these parameters is 10.

The Smith-Waterman algorithm is not very commonly used, as faster heuristic yet reliable local alignment methods, such as BLAST [2], have been developed for the task. A common modification of the Smith-Waterman algorithm is the affine gap local alignment [1], which allows different values for gap initiation and gap extension. The affine gap local alignment is done using the same basic principals, even though allowing gap extensions requires three matrices for the subproblems instead of two (arrows and values).

Nussinov algorithm

The Nussinov algorithm, originally introduced in [37], is a RNA secondary structure prediction algorithm, designed to maximize the number of base pairings in a folded RNA sequence. The four basic parts of dynamic programming algorithm (recursive definition of optimal score, dp-matrix, bottom-up approach of filling the matrix, and the traceback of the matrix [13]) can be found from the working frame of the Nussinov algorithm, just as they were implemented in the workings of the Smith-Waterman algorithm.

The recursive definition of optimal score $S(i, j)$ is:

$$S(i,j) = \max \begin{cases} S(i+1, j-1) + 1[\text{if } i,j \text{ base pair}] \\ S(i+1, j) \\ S(i, j-1) \\ \max_{i<k<j} S(i,k) + S(k+1, j) \end{cases} \tag{2}$$

where $S(i, j)$ more precisely stands for the folding of the subsequence of the RNA sequence from index i to j, which results in the highest number of base pairs. The rows correspond to the situations pictured in Figure 4. The framed structures stand for existing structures and the filled black circles correspond to nucleotides

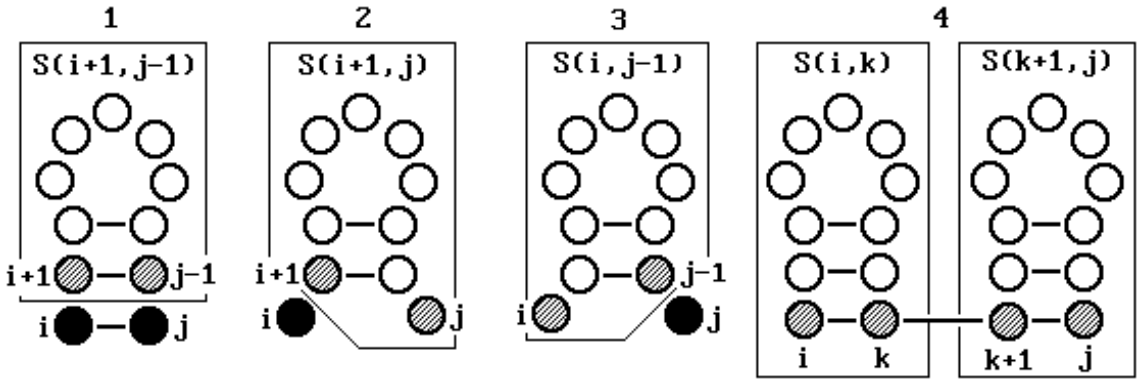

Figure 4: Four possible situations of extending a RNA structure prediction.

to be added to the structure prediction. The first row stands for a situation where nucleotides i and j base pair, meaning that the i, j pair will be added onto best structure found for the subsequence $i + 1, j - 1$. The second row of the equation stands for a situation where i is left unpaired, and the third row corresponds to a situation where j is left unpaired. The fourth row of the equation corresponds to *bifurication*, combining two optimal substructures i, k and $k + 1, j$. [37]

Let us consider an example of predicting the structure of a RNA sequence 'AUCCAU'. The filled DP-matrix for Nussinov algorithm on this example is shown in Figure 5. The bottom-up approach for filling the matrix starts by applying 0s to the two diagonals closest to the bottom left corner. After this, the diagonals are filled using the equation 2, going from the lower left diagonals to the upper right diagonals. The sources of all values are saved, and can be seen as arrows in the figure. The traceback is done by starting from the upper left corner, moving from cell to cell "upstream" the arrows. This example resulted in two optimal solutions, shown in Figure 6.

One downside of the Nussinov's algorithm is that it can predict secondary RNA structures that are not physically possible. The result B in Figure 6, resulting from bifurication, is an example of such non-viable result.

In miRNA target prediction, secondary structure prediction of RNA sequences is done in order to predict accessibility of a sequence. Once the secondary structure prediction has been done, it is possible to additionally calculate the free energy of the folded RNA. Both of these methods are used in some of the tools that will be presented in the following sections. The required computational resources as a function of the sequence length L can be presented as a "big O"-notation $O(L^3)$ in the case of Nussinov's algorithm. More sophisticated folding methods such as RNAfold [55] also tend to have $O(L^3)$ time complexity. This can be improved though, by applying restrictions for the folding.

In the following Subsections 3.1.2–3.1.5, traditional microRNA target (and target

$i \downarrow$ / $j \rightarrow$	A	U	C	C	A	U
A	0	1 →	1 →	1 →	1	[2]
U	0	0	0	0	[1] →	1
C	-	0	0	0	0	1
C	-	-	0	0	0	1
A	-	-	-	0	0	1
U	-	-	-	-	0	0

Figure 5: Nussinov algorithm's secondary structure prediction of RNA sequence 'AUCCAU'.

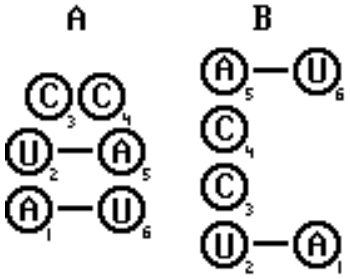

Figure 6: Two RNA structure predictions produced by the Nussinov's algorithm for RNA sequence 'AUCCAU'.

site) prediction tools are presented. All of these methods use dynamic programming methods very similar to the Smith-Waterman and Nussinov algorithms shown here.

3.1.2 TargetScan

The TargetScan [32] miRNA target prediction method can be considered to belong to the first wave of target prediction methods. At the time of its publication in 2003, the mRNAs regulated by vertebrate miRNAs were completely unknown. TargetScan combines thermodynamics-based modeling of RNA:RNA duplex interactions with comparative sequence analysis to predict targets conserved across multiple genomes.

The outline of the operation of the TargetScan tool is shown in Figure 7. Given a miRNA that is conserved in multiple organisms and a set of orthologous 3' UTR sequences from these organisms, TargetScan searches the UTRs for segments of perfect Watson-Crick complementary to bases 2-8 of the miRNA (numbered from the 5' end). Short matching nucleotide sequences like this are commonly referred to as "miRNA seeds", or just "seeds" for short. The 3' UTR counterpart with perfect Watson-Crick complementarity is often referred to as "seed match", as mentioned earlier in Section 2 of this thesis.

The seed matches in the 3' UTR are extended with additional base pairs matching to the miRNA as far as possible in both directions. G:U wobble pairs are allowed here, but the extension is stopped at a mismatch. The RNAfold program, based on [55], is used to find an optimized base pairing between the remaining 3' portion of the miRNA and the 35 bases of the UTR immediately 5' of each seed match. This extended complex will be later on referred to as a "target site". The folding free energy of the miRNA:target site interaction is calculated using RNAeval [20] program. [32]

Based on the folding free energy G_k of the k^{th} target site and the number of matches n in the UTR, a Z score is assigned to each UTR. This Z score is defined as:

$$Z = \sum_{k=1}^{n} e^{-G_k/T} \tag{3}$$

where T is a parameter influencing the relative weighting of UTRs with fewer high-affinity target sites, to those with larger numbers of low-affinity target sites. [32]

TargetScan was initially applied using two sets of miRNAs. The first set was all-mammalian set of 79 miRNAS with homologs in human, mouse and pufferfish and identical sequence in human and mouse. The other set was a vertebrate set of 55 miRNAs that have identical sequence in human, mouse and pufferfish. When multiple miRNAs had identical seed heptamers (seven nucleotides in miRNA positions 2-8), a single representative was chosen. [32]

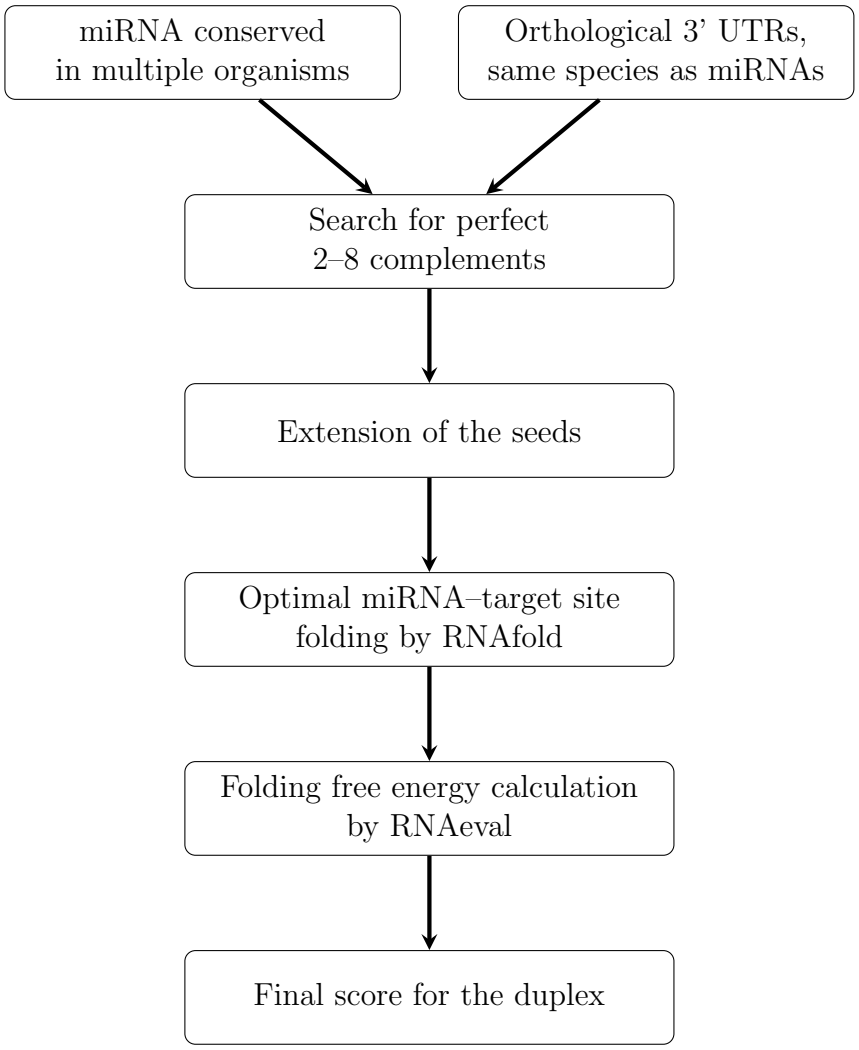

Figure 7: The outline of the operation of the TargetScan [32] miRNA target prediction tool.

For each miRNA in the all-mammalian set, randomly permuted sequences with the same starting base, length, and base composition as the real miRNA were generated until four sequences were found that deviated from the original miRNA by less than 15% in the following properties: the first order Markov probability of the seed match, the first order Markov probability of the antisense of the 3' end of the miRNA, the observed count of seed matches in the UTR dataset, and the predicted folding free energy of a seed:seed match duplex. These shuffled miRNAs form a control set, which can be used to draw a signal to noise ratio to estimate the quality of the parameters used. This signal to noise ratio corresponds to the amount of predicted targets per miRNA reported for the real miRNAs, compared to the amount of predicted targets per miRNA reported for these shuffled miRNAs. The 3' UTR sequences for all human genes, and all mouse, rat and pufferfish genes with a human ortholog were retrieved from the Ensembl database [4] in 2003. [32]

40 miRNAs were randomly chosen from the mammalian set and 27 miRNAs were randomly chosen from the vertebrate set as training sets, while leaving the remaining miRNAs for complementary test sets. Using these training and test sets, suitable values of T and thresholds for required Z score and the rank of the 3' UTR by the score were selected from the following range. T was varied from 5 to 25 in increments of 5, Z score threshold Z_c was varied between 1 and 10 in increments of 0.5, and the rank threshold R_c was varied between 50 and 1000 in increments of 50. The parameters $T = 10$, $Z_c = 4.6$ and $R_c = 350$ were found to give an optimal signal to noise ratio of 3.4:1 for the mammalian training set, while the parameters $T = 20$, $Z_c = 4.6$ and $R_c = 350$ gave an optimal signal to noise ratio of 4.6:1 for the vertebrate training set. [32]

These optimized values were used to obtain new predictions for possible miRNA targets. As there were no experimentally verified miRNA-mRNA interactions in vertebrates at the time, the signal to noise ratios were the only way to estimate the value of the predictions. In essence, TargetScan find parameters within the given range, which give the largest possible number of predicted targets for known miRNAs compared to the shuffled miRNAs, which were used as control.

The method used for finding parameters for scoring in the TargetScan tool could be considered to be a machine learning method. However, no verified miRNA-mRNA interactions are used in the training, and the training is done by simply selecting the parameters, which give optimal signal to noise ratio, from a ready-made range. For these reasons, TargetScan is classified as a traditional target prediction tool in this thesis.

At the time of writing this thesis, the TargetScan tool was available for download at http://www.targetscan.org. The latest version of the tool was TargetScan 5.1 from April 2009. The source code, written in Perl, is fully available. There is also an online version of the tool available on the site, which is convenient for trying out the tool with single miRNAs.

3.1.3 miRanda

The MiRanda microRNA target prediction tool [22] is based on sequence complementarity between the mature miRNA and the target site, binding energy of the miRNA-target duplex, and the evolutionary conservation of the target site sequence and target position in aligned UTRs of homologous genes.

As the starting point, every gene in a given genome is considered to be a potential target site for every miRNA. Originally in 2003, miRanda was designed for handling the miRNAs of *D. melanogaster*, so only the miRNAs of the species known at the time were considered to be potential to target the genes of the species [14]. At its present state however, the method can be used to find targets for any given list of miRNAs from a set of 3' UTRs or mRNAs.

The latest version of miRanda algorithm, updated in September 2008 [3], works as follows. The outline of the workings of the algorithm is also shown in Figure 8. Using the given miRNAs as probes, miRanda starts by scanning the 3' UTR or mRNA sequences given for possible complementary matches using dynamic programming algorithm, similar to the Smith-Waterman algorithm [48]. The algorithm used takes into account G-U wobble pairs, allows moderate insertions and deletions and uses a weighting scheme that rewards complementarity at the 5' end of the miRNA by a factor of 4. Complementarity parameters at individual alignment positions are +5 for G-C, +5 for A-U, +1 for G-U and -3 for all other nucleotide pairs. The algorithm uses affine penalties for gap opening (-9) and gap-extension (-4). In addition, the sequence match phase of miRanda uses some position-specific, empirically defined rules. These rules are presented here, with positions counted starting at the 5' end of the miRNA: no mismatches at positions 2 to 4; fewer than five mismatches between positions 3-12; at least one mismatch between positions 9 and L-5 (where L is total alignment length); and fewer than two mismatches in the last five positions of the alignment. In addition, first two nucleotides from the 5' end and the last nucleotide in the 3' end of the miRNA are ignored in the alignment [3]. The resulting score (S) of this sequence matching phase is used to discard possible miRNA-mRNA duplexes not exceeding a given threshold. If there are multiple putative miRNA target sites in a single mRNA, the scores are summed together before comparing to the threshold. [14]

In order to estimate the thermodynamic properties of the remaining duplexes, the algorithm uses folding routines from the Vienna RNA secondary structure programming library (RNAlib) [52]. The folding free energy of the duplex is also used to discard potential miRNA targets, which do not produce an energy value low enough. [14]

As the two phases described before, the sequence conservation phase is essentially done in a manner to balance between false positives and false negatives. In the first miRanda paper, which focuses on predicting potential miRNA targets for *D. melanogaster*, the sequence conservation was done comparing the sequences with

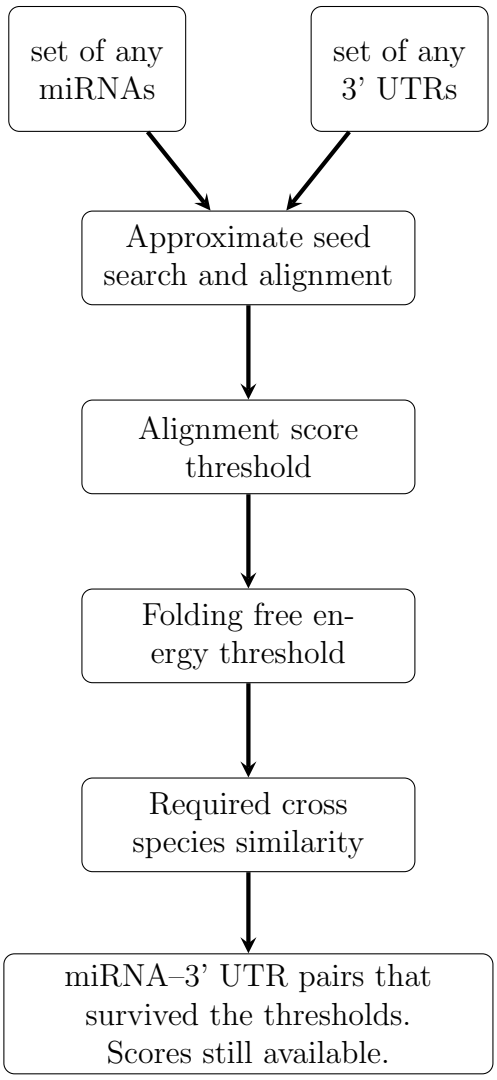

Figure 8: The outline of the operation of the miRanda [22] miRNA target prediction tool.

two close relatives. Sequence identity of 80% was required with *D. pseudoobscure* and sequence identity of 60% was required with *A. gambiae* [14]. In the second paper focusing on human miRNA targets, the target sites were required to have an identity of 90% with mouse and rat target sites [22].

The latest versions of miRanda miRNA target predictions for human, mouse and rat are available online [3]. These predictions have also been updated on September 2008. For each potential miRNA-mRNA pair exceeding the set thresholds and fulfilling the conditions, the results include precise alignment score, target conservation over 30 vertebrates and folding free energy value. At the September 2008 result sets the thresholds are: alignment score cutoff $S >= 140$, conservation cutoff of .566 and energy cutoff $E <= -7.0$.

At the time of writing, the microRNA.org resource [3] provided the source code for miRanda, which is written in C++, at http://www.microrna.org. This naturally makes it possible to alter and experiment with all the parameters freely.

As the miRanda package and results are continuously updated and maintained, miRanda is an excellent method to compare other, new methods with. Additionally, the tool is relatively fast and the output it produces can be modified easily. Results produced by miRanda are compared with the results by a genetic algorithm -based miRNA target prediction tool, originally introduced in [25], in Section 4.2 of this thesis.

3.1.4 RNAhybrid

RNAhybrid, originally published in 2004 by Marc Rehmsmeier et al. [42], represents the first generation of microRNA target prediction tools. The program finds the energetically most favorable hybridization sites of a small RNA in a large RNA. In this problem setting, the small RNA is typically microRNA and the large RNA is the 3' UTR of a messenger RNA.

RNAhybrid is an extension of the classical RNA secondary structure prediction algorithm for two sequences by Zuker and Stiegler in 1981 [55]. The microRNA is hybridized to the target in an energetically optimal way, which means that it is the way yielding minimum (folding) free energy (MFE). The MFE is calculated using dynamic programming technique, essentially similar to the method originally designed by Nussinov et al. [37].

All possible starting positions for the hybridization are considered in both the miRNA and the potential target mRNA's 3' UTR. RNAhybrid does not allow hybridizations which would happen completely between target nucleotides or between miRNA nucleotides. Stretches of unpaired nucleotides, which are also called *bulge loops*, are restricted to a constant maximum length in either sequence. This maximum length is 15 by default.

The time consumption of the RNAhybrid is of the order $O(c^2mn)$, where c is the maximum length of the bulge loops, m is the length of the 3' UTR and n is the length of the microRNA. As the length of microRNAs is rather static at around 20 to 22 nucleotides and the maximum length of the bulge loops is also fixed, the time consumption of RNAhybrid is linear in the target length m.

The resulting minimum free energy for each miRNA-mRNA pair is normalized with two different methods. The score is first normalized by the length of the alignment. The length normalized minimum free energy result is considered to follow extreme value distribution and the statistical significance of the result is considered. Only results with significance exceeding a certain threshold are considered to correspond to potential miRNA-mRNA interactions.

At the time of the publication of the RNAhybrid article, there were very few experimentally verified miRNA-mRNA interactions available. The RNAhybrid method was used for searching 78 miRNAs, found from *D. melanogaster*, from the 3' UTRs of *D. melanogaster*, *D. pseudoobscura* and *A. gambiae*. Results reported in the paper focus on predicting new potential interactions. Some later methods are compared to RNAhybrid in the latter sections of this thesis.

At the time of writing this thesis, the RNAhybrid tool was available for download at http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/. The latest version, RNAhybrid 2.1, was from November 2004. The source code, which is written in C++, is fully available though.

3.1.5 Sfold

Sfold is a traditional target prediction method, designed by Dang Long, Chi Ye Chan and Ye Ding from Wadsworth Center, New York State Department of Health, in 2008 [33]. Their method is based on handling the miRNA-target binding as a two-step hybridization reaction. The first step is considered to be the initial binding, or nucleation, of the miRNA to an accessible target site. The second step is considered to be hybrid elongation to disrupt local target secondary structure and form the complete miRNA-mRNA duplex. The Sfold does not involve seed match conservation, unlike some other traditional target prediction methods. The rough outline of the workings of the Sfold tool for a single miRNA and a single mRNA is shown in Figure 9. Sfold can be run with multiple miRNAs or mRNAs.

The main innovative feature of Sfold is that instead of considering only a single secondary structure for a potential target site, Sfold considers a whole structure population in a dynamic equilibrium, taking into account the likelihoods of each single structure [33]. If one wants to address the secondary structure, the method that the Sfold tool is using is more extensive and hence superior to a method where only a single structure is considered. It is important to remember though, that it is currently not possible to accurately and reliably predict the tertiary or more complex structures of the target site [53, 54]. As the more complex structures actually determine whether or not the potential microRNA target site is truly accessible, the relative accessibility proposed by the secondary structure alone may very well be misleading.

The Sfold requires perfect base pairing of four consecutive complementary nucleotides in the miRNA and 3' UTR in a site considered accessible by the secondary structure prediction of the mRNA. The nucleation potential is the stability of this 4-bp block in the putative, accessible binding site. For a certain potential pair formed by the miRNA and one mRNA structure of the population, the four base pair long block with the highest nucleation potential is considered to represent the pair in question. Since the method is considering a population of secondary structures, the final nucleation potential is received by averaging over the whole structure population.

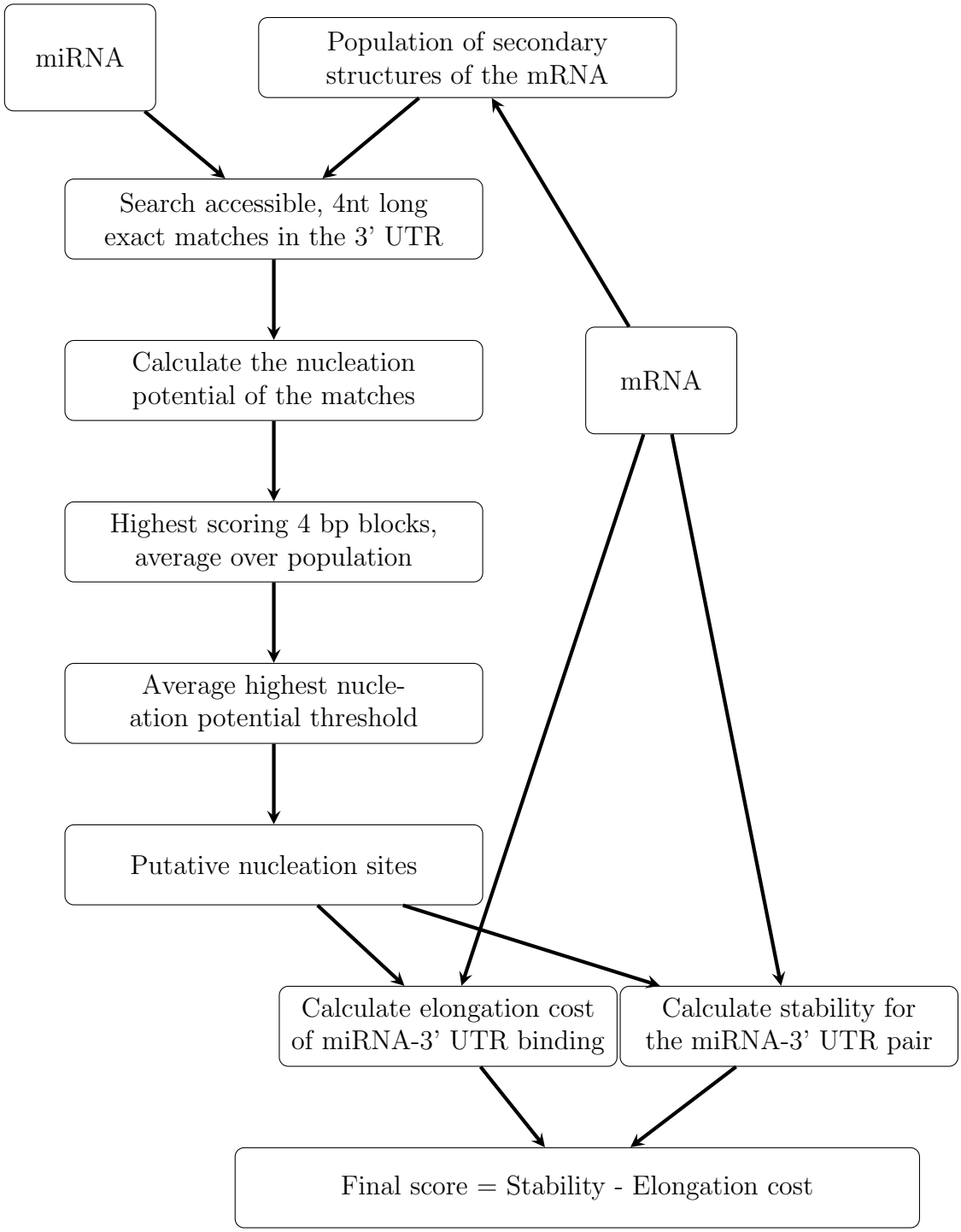

Figure 9: The rough outline of the workings of the Sfold tool [33] for a single miRNA and a single mRNA.

A typical size for a structure population is approximately 1000 structures. If the nucleation potential does not exceed a set threshold, this mRNA population is not considered to be potential for the miRNA in question, and is not taken into account in any further processing, i.e. it is discarded.

For each of the putative nucleation sites proposed by the previous step, energy cost caused by the elongation of the miRNA-3' UTR biding for the local secondary structure is calculated, and this is subtracted from the stability the RNAhybrid program [42] predicts for the miRNA-3' UTR pair. Efficiently, by requiring the nucleation site, Sfold adds a prefiltering phase to modified results of the RNAhybrid tool.

The paper introducing the method had results for 19 experimentally verified interactions. Sfold was able to find 16 of these on the thresholds used. Unfortunately, there are no predictions on genome wide data for Sfold, and no results showing the amount of total miRNA-mRNA interactions Sfold would predict as positive with the used nucleation potential and hybridization energy thresholds. However, the method introduces an insightful way of bringing secondary structure to miRNA target prediction. At the time of writing this thesis, an online version of the Sfold tool was available for use at http://sfold.wadsworth.org/. No source code or downloadable executables are available, though.

3.2 Machine learning target prediction

The target prediction tools that have trainable parameters affecting the predictions are classified as machine learning miRNA target prediction tools in this thesis. Typically machine learning methods have a training set, which is used to find out the best possible set of parameters, which are then applied to do predictions on a separate test set.

In the training, the usual goal is to maximize the amount of correct predictions and minimize the amount of false predictions, while avoiding overfitting. This requires at least known positive data in the training set, which in the case of miRNA prediction translates to verified miRNA-mRNA interactions. It would also be preferable to have access to sufficiently comprehensive set of known negative data, which in this case would be experimentally verified miRNA-mRNA pairs, where the miRNA causes no regulation of the mRNA translation to protein. Unfortunately, there is very limited amount of this kind of verified negative data available, which causes definite challenges for machine learning miRNA target prediction tools.

In Subsections 3.2.2–3.2.4, three different machine learning tools are presented. Additionally, basic concept of genetic algorithms are presented in Section 3.2.1. This concept will also provide useful background information for Section 4, where a miRNA target prediction tool using a genetic algorithm is introduced.

3.2.1 Genetic algorithms

A *genetic algorithm* (GA) is considered to be a heuristic problem solving technique, used to find solutions for optimization problems [35]. The general form of a genetic algorithm presented in this section is not the only possible implementation of a machine learning method that would be classified as a GA. The goal of this section is to give a general idea how genetic algorithms work, not an extensive overview of the whole field.

The GA evolves a population of candidate solutions, called *individuals*, towards better solutions. One individual holds one full set of parameters that are being optimized. The individuals can be presented as simple binary strings of 0s and 1s, but other encodings are also possible and can allow the GA to converge to the global optimum solution in a faster and more reliable manner. The allowed ranges that are built in the presentation of the individual define the allowed search space of the parameter optimization. [35]

Especially when using bit string representations of integers, *Gray coding* is a common choice for the encoding. The method was originally presented in the patent [16] and has been later on named after the author Frank Gray. In Gray coding, the bitwise presentations for two successive values differ by a single bit. Because of this principal property, mutations in the code allow for mostly incremental changes, but occasionally a single bit-change can cause a big leap, leading to new properties of the individual. As there can be multiple ways to implement the Gray coding for a set of values, the exact implementation has the last word on the relative impact of mutations in different positions of the code. These characteristics of Gray coding make it especially suitable for genetic algorithms. It is noteworthy that the usage of Gray coding does not necessary limit the accuracy of parameters to the level of integers. [16]

The quality of the individuals is measured by a *fitness function*. The fitness function gives a single comparable value that should correspond to the goal of the parameter optimization. Efficiently this fitness function determines the direction of the development of the population. The function itself can have any form, as long as the values of the fitness function depend on each of the parameters. [35]

The genetic algorithm starts its work with a certain initial population. The easiest way to create the initial population is just to randomly generate the desired amount of individuals. All these individuals are given a fitness value, using the fitness function. From this initial population, a new population will be created using three operators generally used in genetic algorithms: *selection*, *crossover* and *mutation*. [35]

The selection phase picks a number of individuals, based on their fitness values, to breed a new generation. These individuals are commonly referred to as *parents* [35]. Typically the individuals that have a high fitness value have a high probability

of being selected to be parents, but there are numerous different ways to do the exact implementation of this step. Most methods are stochastic in their nature and designed so that there is always a small probability, that less fit individuals are selected as parents. Wheel selection [24] and tournament selection [6] are two much used solutions, and as such there is a lot of experience on the results received from their usage.

Parents chosen in the selection phase breed the new generation in the crossover phase. The simplest crossover methods are the one-point crossover and the two-point crossover. In one-point crossover with two parents, a single point in the string representing the individual is chosen to be the crossover point. All data beyond this point is swapped between the parent strings, creating a new generation of two strings. In two-point crossover, two points are selected and everything between these two points are swapped between the parents. One can use any kind of selection operators, as long as the new generation consists of valid individuals. [35]

The mutation operator is the final way of adding variation to the new generation. The classic example of a mutation operator involves a probability that an arbitrary bit in the newly bred individual will be changed from its original state to the opposite. [35]

The genetic algorithm carries on creating new generations of individuals, always using the previous generation to breed the new one. It is common to leave a portion of the most fit individuals from the previous generation to accompany the new generation. This is called *elitism* [35]. Some methods do this for the whole last generation, growing the size of the population endlessly.

The termination of the loop can be done using various stopping conditions. One option is to simply stop the GA after certain amount of new generations or individuals have been created. Other option is to somehow examine the convergence of the parameters being optimized. For an example, one could set a range for each parameter, where the last hundred best individuals have to fit in. It is also possible to use more complex stopping conditions, such as requiring the variance of the parameters in the last n individuals to be below a certain value. Lastly, there is always the possibility of letting the GA continue the search for the optimal parameters until manually stopped. This can be a reasonable choice especially if the GA is meant to be left running for long times, possibly overnight or over weekends and one wants to avoid unintentional termination of the optimization. In this case, it is necessary to monitor the obtained parameter values and the changes in them to be able to say when the GA has been running long enough. After the termination of the GA, the most fit individuals of the last generation should give the most optimal parameters originally desired.

All of the phases of the workings of a genetic algorithm have a lot of small details that fully define the outcome of the optimization. Because of this, the genetic algorithm should not be understood as a black-box kind of tool, that one can simply plug in

to a problem and get some answers out of it. It is also noteworthy that it is very probable, that the GA will not converge to the same result every time because of the stochastic nature of its operators. In the following Subsection 3.2.2, a miRNA prediction tool using these genetic algorithms as part of it is presented.

3.2.2 TargetBoost

TargetBoost [44] uses machine learning on a set of validated microRNA targets in lower organisms to create *weighted sequence motifs* that capture the binding characteristics between microRNAs and their targets. The method combines a genetic algorithm with *boosting*.

The genetic algorithm used in TargetBoost evolves individual patterns from a population of candidate patterns. The importance of each pattern is guided by the boosting algorithm, which assigns weights for each of the patterns, based on their performance on the training set.

The patterns are general expressions that describe the common properties of target miRNA sites. The syntax of the patterns allows two separate sites in the 3' UTR sequence, that match parts of the miRNA sequence. Examples of the syntax for two separate patterns are shown in Figure 10. The syntax used here is a clarified version of the syntax presented in the original paper. P_i stands for a nucleotide in the 3' UTR sequence, that is complementary to the nucleotide in miRNA position i from the 5' end of the miRNA. Positions 1 to 21 are allowed in the syntax. W stands for a wildcard, meaning any nucleotide in the 3' UTR matches it. In the syntax, there are always two matching sequences of variable length within a variable distance of each other. The total length of the pattern is limited to 30, including gaps.

The line labeled $Q1$ describes a pattern starting with a match of length 3 or 4. There have to be at least three nucleotides complementary to the miRNA positions 21, 20, 19, 18 or 17, which of 3 have to be present in the 3' UTR. $(P_{19}|P_{18})$ means either P_{19} or P_{18}. The following combinations of complementary nucleotides are allowed for this first match: $P_{21}P_{20}P_{19}$, $P_{21}P_{20}P_{18}$, $P_{21}P_{20}P_{17}$, $P_{21}P_{19}P_{17}$, $P_{21}P_{18}P_{17}$, $P_{20}P_{19}P_{17}$, $P_{20}P_{18}P_{17}$, $P_{21}P_{20}P_{19}P_{17}$ and $P_{21}P_{20}P_{18}P_{17}$. After this first match there has to be a gap of 8-15 nucleotides before the second match. Second match is of exact length 6, including one wildcard between the nucleotides complementary to miRNA positions 3 and 2.

The line labeled $Q2$ describes a pattern, where the first match is composed of 5 nucleotides matching miRNA positions 16-15 and 13-12, while allowing the middle nucleotide to be A, C, G or U. There has to be a gap of 4-14 nucleotides in the 3' UTR before the second match, with at least 2 nucleotides complementary to miRNA positions 11-7. The order of the nucleotides in the 3' UTR must be the order stated in the pattern.

	First match	Variable distance	Second match
Q1:	$P_{21}P_{20}(P_{19}\|P_{18})P_{17} : p \geq 3$	$r = 8, d = 7$	$P_7P_5P_4P_3WP_2 : p \geq 6$
Q2:	$P_{16}P_{15}WP_{13}P_{12} : p \geq 5$	$r = 4, d = 10$	$P_{11}P_{10}P_9P_8P_7 : p \geq 2$

Figure 10: Example of the pattern syntax used in the TargetBoost prediction method.

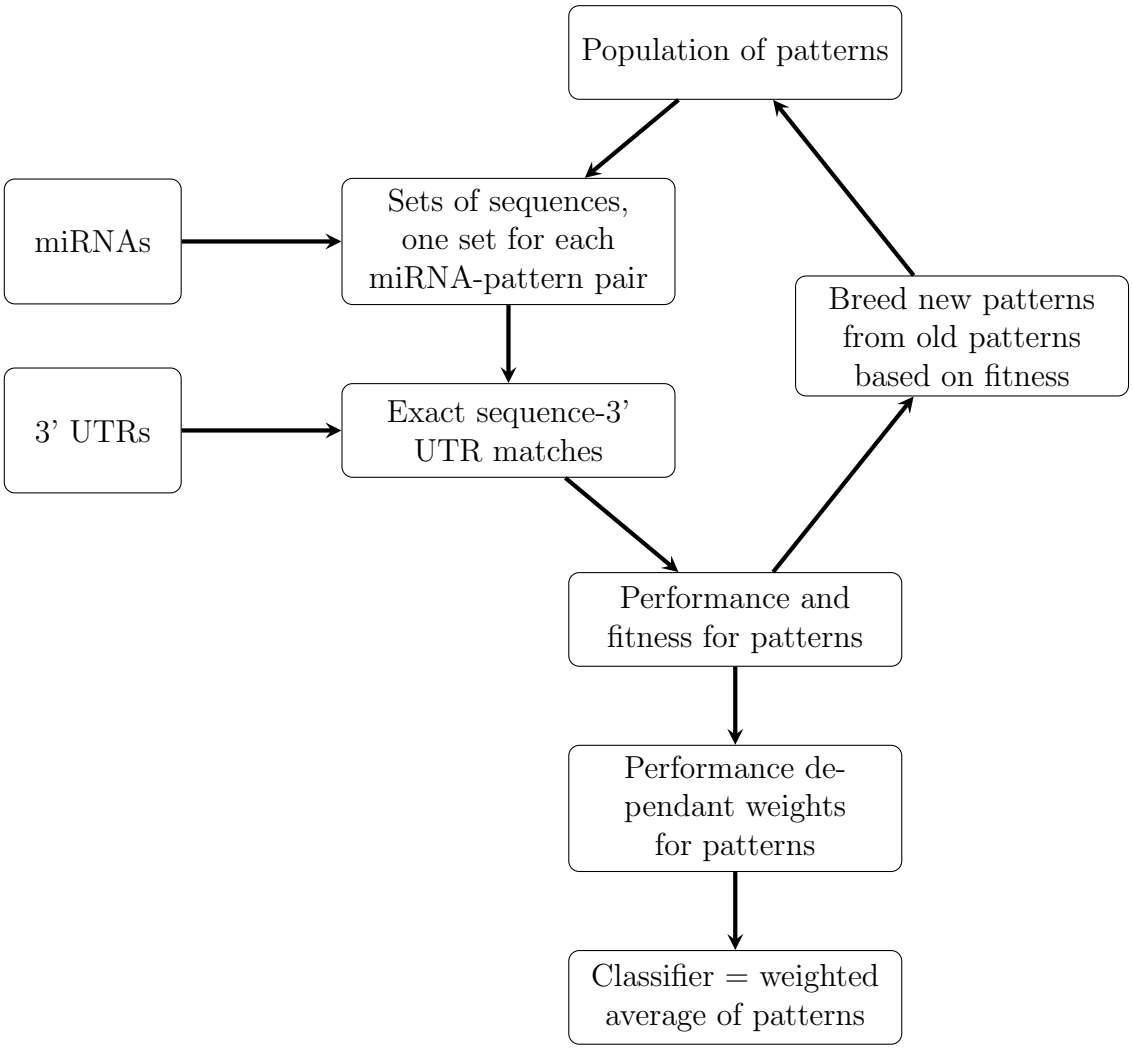

Figure 11: The basic outline of the mechanics behind TargetBoost [44] classifier formation.

The outline of the creation of the classifying pattern is shown in Figure 11. For each miRNA, all possible sequences allowed by the pattern in question are created. Simple search for exact matches is done to find matches for these sequences in the 3' UTR. If an exact match for a sequence created by the pattern is found, the site found is classified as a positive miRNA target site. There is no typical weighted alignment step involved.

The boosting technique assigns a weight to each pattern, depending on its performance on the training set. The genetic algorithm component evolves new patterns, using two well performed patterns as parents for the new one. Exact methods and details of the boosting and genetic algorithm phases are shown in the paper [45]. The final classifier is the weighted average of a large number of patterns resulting from the GA.

More precise details are not really necessary here. One of the main points to capture from the workings of TargetBoost is that the genetic algorithm optimizes the pattern of required matching nucleotide pairs, where as the GA-based predictor, which is more comprehensively presented in Section 4 of this thesis, uses genetic algorithm to optimize the alignment parameters.

The TargetBoost uses no additional filters, such as requiring conservation of the target sites or the presence of multiple target sites in the 3' UTRs. The reasoning behind this is, that these sort of filters can be used independently of the initial method to predict the target sites. This means that that improving the quality of the initial candidates will also improve the final predictions. [44]

Using 10-fold cross-validation, TargetBoost was compared with two other methods: Nucleus by Rajewsky and Socci in 2004 [40] and the RNAhybrid [42], which has been earlier described briefly in Section 3.1.4 of this thesis.

The positive data set used consisted of 36 experimentally verified targets, and their predicted target sites for the miRNAs let-7, lin-4, miR-13a and bantum in *C. elegans* and *D. melanogaster*. Each target site was padded with their respective sequences, such that the length of the sequences was 30 nt. Target sites longer than 30 nt were discarded from the data set. [44]

The negative data set consisted of 3000 random strings, all 30 nt long. The frequencies used in the generation of the random strings were $P_A = 0.34$, $P_C = 0.19$, $P_G = 0.18$ and $P_U = 0.29$. These frequencies correspond to the nucleotide composition of D. melanogaster 3' UTRs [40]. [44]

The overall performance of the three algorithms was compared computing the ROC-score for each algorithm on each miRNA. The resulting receiver operating characteristic curves are shown in Figure 12. Additionally, for each individual miRNA, the authors tested whether the best algorithm was significantly better than the other algorithms. Figure 13 shows the exact ROC-scores for each algorithm and miRNA. The scores that are not significantly different from the highest score (90% confidence

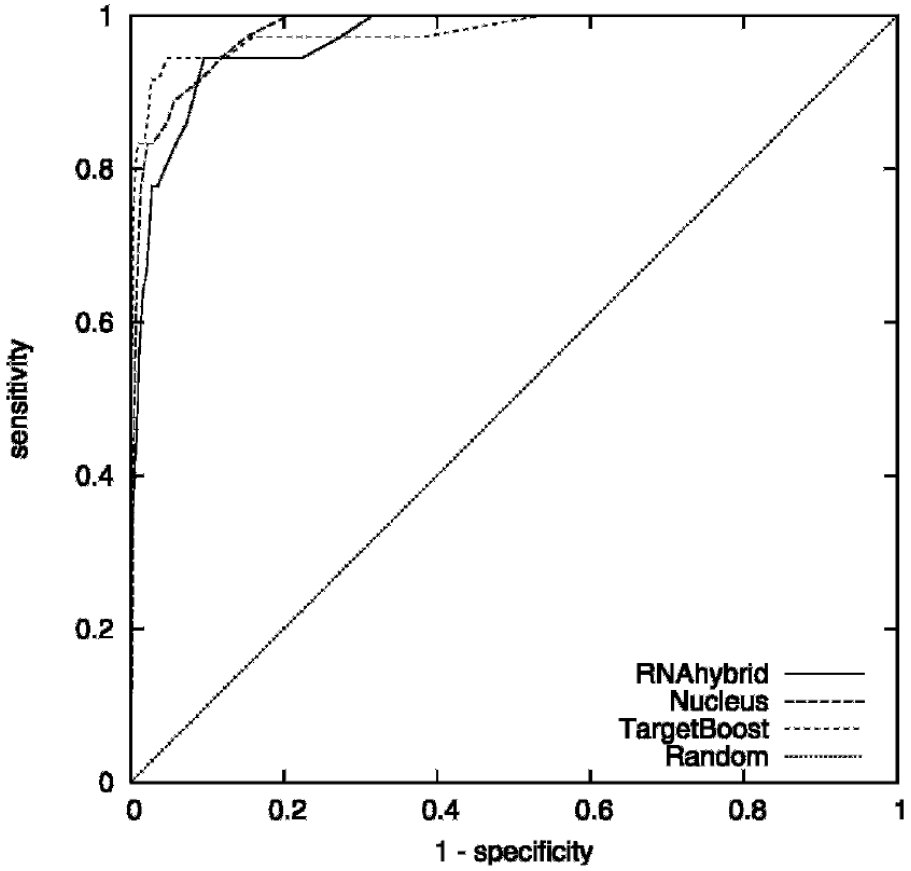

Figure 12: Overall ROC curves for each algorithm compared in the TargetBoost [44] paper.

Algorithm	let-7	lin-4	miR-13a	bantam	all
TargetBoost	**0.997**	**0.944**	**0.972**	**0.998**	0.979
RNAhybrid	**0.989**	0.931	**0.979**	**0.991**	0.967
Nucleus	0.988	**0.962**	0.928	**0.998**	0.973

Figure 13: ROC-scores for algorithms TargetBoost [44], RNAhybrid [42] and Nucleus [40]. Scores that are not significantly different from the highest score on a particular miRNA are in boldface (90% confidence level).

level) on a particular miRNA are boldface. [44]

The results show, that TargetBoost was more consistent and accurate in this data set than the RNAhybrid and nucleus algorithms. However, on very high specificity values, the nucleus algorithm outperformed the other two. This is not a trivial feat, as when working with genome wide data, the performance with very high specificity values is quite important.

Additionally, the authors found that the optimization result of the genetic algorithm

converged to solutions, where the six of the first eight bases of a miRNA had to have a complementary match in the target site. [44] This is beautifully in line with the earlier concepts of a "seed match" matching the 5'end of the miRNA.

At the time of writing, no executables or source code were available for TargetBoost. However, there is an online demo at https://demo1.interagon.com/targetboost/.

3.2.3 PicTar

The rough outline of the workings of PicTar [29] algorithm, designed in 2005, is shown in Figure 14. Input to PicTar consists of multiple alignments of RNA sequences (typically 3' UTRs) and a search set of mature (coexpressed) microRNA sequences.

The program nuclMap locates positions in the 3' UTR that match a seed of 7 nucleotides, starting at position 1 or 2 of the 5' end of the microRNA. Mutations in the seed matches (the complementary matching 3' UTR sequences) are allowed, as long as the folding free energy of the entire miRNA:mRNA duplex does not increase and there are no G:U wobble pairs. The folding free energy of the duplex is required to be under a set threshold value. This threshold value is set to be 33% of the optimal folding free energy of the entire mature microRNA binding to a perfectly complementary target site. [29]

If there is a perfect 7 nucleotide match and the duplex survives the threshold value checks, the putative miRNA target site is counted as an *anchor*. The number of anchors in an UTR determines whether the putative target sites will be scored by PicTar. [29]

A score for a possible miRNA target site is gotten from a probability calculated using Ahab [41] algorithm. Essentially, this algorithm maximizes the likelihood of the model to generate a training set of known duplexes by configuring the weights of premade 7 nucleotide long patterns. The likelihood Z(S) to maximize is calculated as follows:

$$Z(S) = \sum_{T} P(T) \tag{4}$$

$$P(T) = \prod_{k=1}^{N(T)} p_{w_k} m(s|w_k) \tag{5}$$

$$m(s|w_k) = \prod_{j=1}^{l} f_j(n|w_k) \tag{6}$$

Here $f_j(n|w_k)$ are the normalized frequencies of nucleotide n at position j for the

pattern w_k. This means that $m(s|w_k)$ is the probability of the pattern w_k to generate a certain sequence s. T stands for a certain segmentation proposing sequence s to be a potential miRNA target site. $P(T)$ is the product of all probabilities $m(s|w_k)$, weighted by the probability p_{w_k} of each individual pattern w_k. In other words $P(T)$ is the probability that a sequence s is created by the weighted set of patterns in question. The values p_{w_k} are the weights that are configured to maximize the likelihood $Z(S)$. [41]

The segmentation T can be of such form, that there are multiple target sites in the 3' UTR. In this kind of situation, the miRNAs compete with each other for binding. The model accounts for synergistic effects of multiple binding sites of one miRNA or several microRNAs acting together. The weights that maximize the likelihood can be used to score a test set of potential target sites. The score obtained from this usage of the Ahab algorithm, is measured against a threshold score value. [29]

Usage of Ahab algorithm, which requires at least a small set of samples for training, can be argued to cause PicTar to be classified as a machine learning method. This is why the PicTar is under the "machine learning prediction tools" section, even though none of the other steps of the algorithm have anything to do with machine learning.

After the scoring, there is a final filter requiring a minimal (user defined) amount of matching target sites found in matching positions in cross-species multiple sequence alignment of the 3' UTRs. The end result is a list of potential target sites, ranked by their PicTar score. [29]

At the time of writing, PicTar predictions on various organisms were available at http://pictar.mdc-berlin.de/. However, no source code or executables were made publicly available.

3.2.4 miTarget

MiTarget [27], designed in 2006, is a machine learning miRNA target prediction method, based on support vector machines. These support vector machines, or SVMs are a set of related supervised learning methods used for classification and regression [7]. A support vector machine constructs a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification, regression or other tasks [7]. In the case of miTarget, there are only two SVM dimensions, corresponding to two parameters, which have an effect on the way of comparing two different data points with each other. These data points correspond to know positive or negative miRNA-mRNA interactions.

Generally the efficiency and reliability of any machine learning algorithm depends on choosing relevant data and specific features. There are 152 positive and 246 negative miRNA-mRNA interaction data points in the miTarget study. The positive data

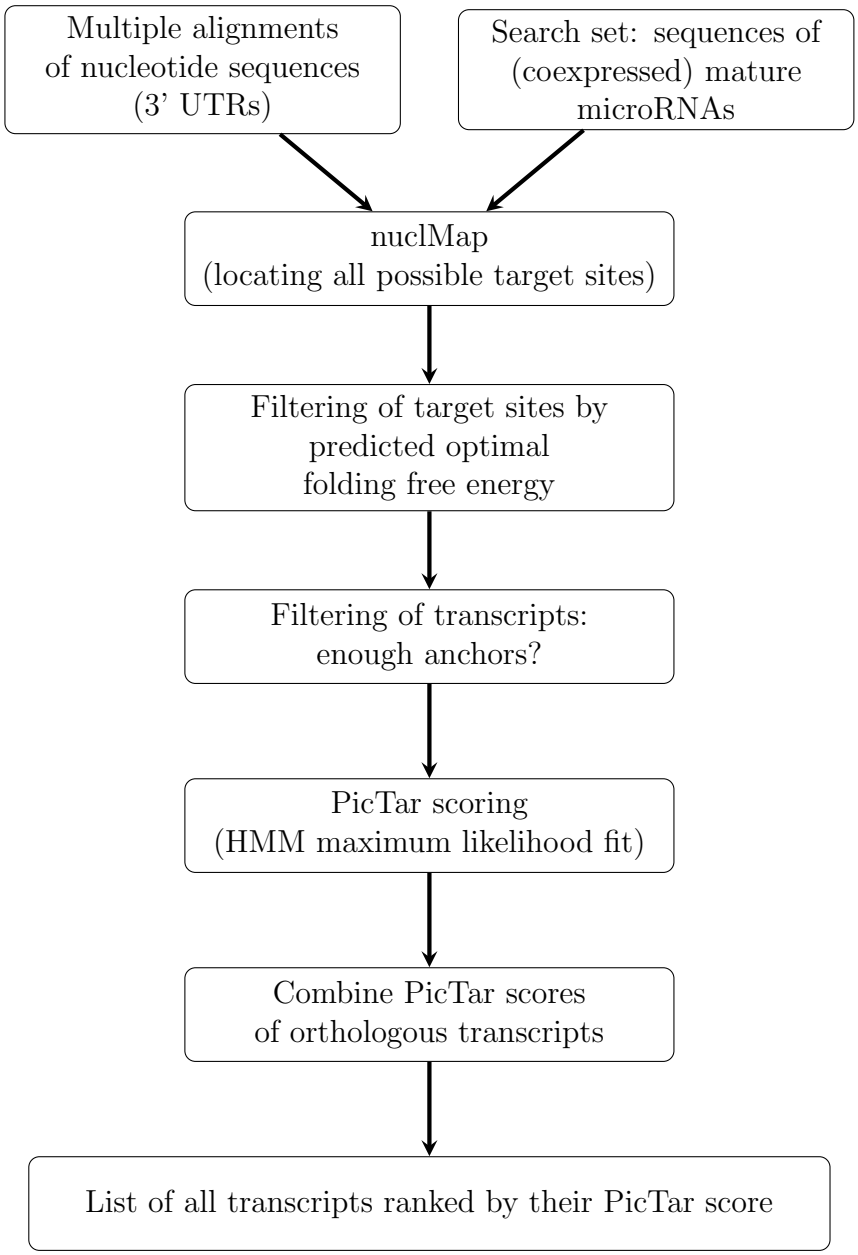

Figure 14: Schematic overview of the workings of PicTar [29] algorithm.

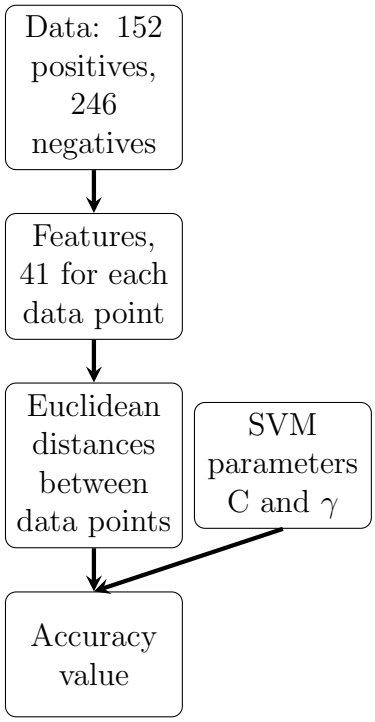

Figure 15: Flowchart roughly describing the crucial steps of miTarget [27] algorithm.

points corresponds to known wetlab experiences in multiple species. The binding sites in the data are putative, based on complementary between the miRNA in question and the mRNA translationally repressed by the miRNA.

Minority of the negative data, 83 sites to be precise, are computationally predicted target sites by earlier methods, such as PicTar [29], for which later wetlab studies show that no miRNA induced translational repression of the mRNA occurs. The majority of the negative sites, meaning 163 sites, are inferred by noting that the deletion of the predicted target sites in the 3' UTR can give a large number of negative examples. The predicted binding sites of let-7 miRNA on lin-41 and lin-28 mRNAs were masked and all the other remaining sites with favorable seed pairings were considered to be negative examples. [27]

The flowchart shown in Figure 15 roughly describes the steps involved in the workings of the miTarget prediction method. 41 different features are drawn from each of the data points described above. These features can be divided into three different categories: structural features, thermodynamic features and position based features. Of these, the two first ones the type of features that have been used in a large number of earlier methods, such as miRanda [14] and RNAhybrid [42]. However, miTarget was the first method to use position-based features to predict miRNA-mRNA interactions. [27]

For structural and thermodynamic features, the secondary alignment (which is the data point) is divided to three parts: the 5' part, the 3' part and the total alignment. For each of these three parts, the count value of matches, mismatches, G:C matches, A:U matches, G:U wobble pairs and the folding free energy of the part in question are calculated. Position-based features correspond to one of the four nominal values of positions 1 to 20 in the miRNA: G:C match, A:U match, G:U wobble pair and a mismatch. All values are normalized to have real values in the interval (0,1). [27]

Using these values of the features, the euclidean distance $||x_1 - x_2||$ between two points can be calculated. Essentially, the final difference value between the two data points is decided by this euclidean distance and the two SVM parameters γ and C. Here γ stands for similarity level of the features and C stands for the trade-off between margin maximization and training error minimization. These values of these two SVM parameters determine the outcome of the classification. [27]

In the miTarget paper, the data was divided into 10 different training and test sets through random sampling. Then a tenfold cross validation was performed with the training data to train a classifier and to optimize parameters. The optimization in the cross validation was maximization of the accuracy and the parameter ranges used were from 0.01 to 2.0 in steps of 0.01 for γ and from 1 to 200 in steps of 1 for C. Finally, the parameters maximizing the accuracy were used to plot a receiver operating characteristic curve on the separate test set, averaging the results over the 10 different sets. The miTarget paper compared ROC curves resulting from three different sets of features: the entire set, position-based features only and with-

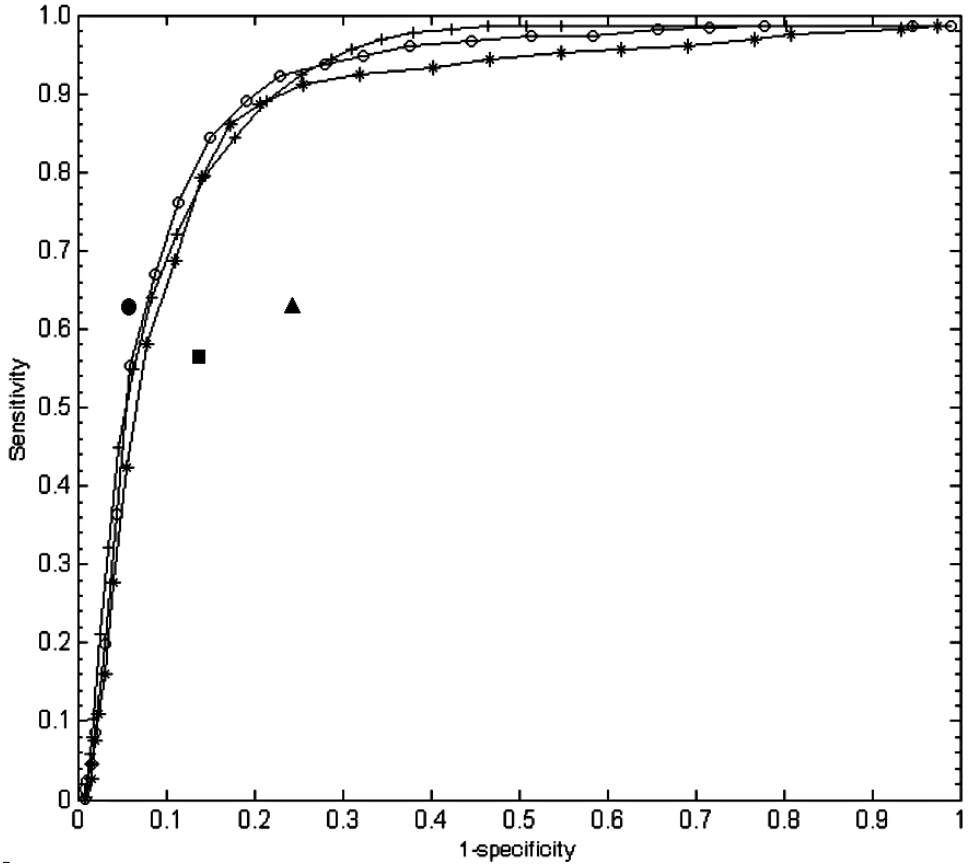

Figure 16: The ROC curves of classifiers created by miTarget on three combinations of features: an entire set (circles), position-based features only (asterisks), and without position-based features (plus symbols). The filled triangle denotes the performance of TargetScan, the filled square shows the performance of RNAhybrid, and the filled circle shows the performance of miRanda. [27]

out position-based features. These along with single results from TargetScan [32], RNAhybrid [42] and miRanda [14] are shown in Figure 16.

Just by looking at Figure 16 the results given by miTarget do seem quite good. When plotting a miRNA target prediction related ROC curve, the choice of the negative set, corresponding to the X axis, has a huge impact on the final results. The amount of known negative examples was rather small in the time of writing of miTarget, and unfortunately this is still the case today. If the known negative examples would extensively cover all types of sites that a miRNA will not bind to, then it would be reasonable to use it as a measure to compare true positive rate against. In a real situation one would generally want to use the miRNA target prediction algorithm to find potential targets in a set of 3' UTRs, while minimizing the amount of wetlab experiments needed to confirm the putative interactions. In this problem frame, it is naturally favorable to pursue a maximal amount of known positive examples the classifier labels as positive in the test set. However, when realistically considering

the state of known negative examples, it would be more favorable to minimize the amount of total interactions classified as positive, instead of minimizing the amount of known negative examples classified as positive in the test set.

In the case of miTarget, the inferred 163 negative examples make the aforementioned situation even more severe. These negatives are just sequences from lin-41 and lin-28 that have 4 nucleotides matching with let-7, once the actual binding sites of let-7 are removed. It would seem unlikely that these sequences, taken from a single 3' UTR, could represent the full set of sequences that no miRNA would bind to. If the sequences do not represent this set well enough, the specificity shown in the ROC curves of Figure 16 does not correspond to the real specificity of the methods.

Keeping this in mind, it is noteworthy though, that miRanda is actually outperforming all the other methods by a notable margin, as can be seen from the results shown in Figure 16.

At the time of writing, there was a web interface for miTarget available online at http://cbit.snu.ac.kr/∼miTarget/. No executables or source code were available.

4 GA-based target predictor

In this section of the thesis a miRNA target prediction tool based on genetic algorithms is presented. The text here is largely based on the paper [25]. The design of this tool has been done by the authors of the aforementioned paper, including the author of this thesis.

The GA-based algorithm was developed with the idea of rating known miRNA-mRNA interactions as high as possible within a set of $I * M$ possible miRNA-mRNA pairs, where I is the number of miRNAs and M is the number of mRNAs in the set. As the amount of known miRNA-mRNA pairs where the miRNA does not repress mRNAs, which would form the negative set in machine learning, is very low, it does not seem reasonable to consider the available examples of such cases to cover the characteristics of all real negative cases extensively and in right proportions. With this reasoning, our tool aims for maximal sensitivity while minimizing the total number of predictions in order to reduce the relative amount of false positives.

The GA-based target predictor tool combines ideas from both traditional algorithms and machine learning techniques. It is implemented as a two-step classifier. Rough outline of the workings of the tool are pictured in Figure 17. In the first main step, a training data set is used to find the best *seed pattern pair*, that predicts the smallest total number of hits (meaning total number of mRNA-miRNA interactions predicted to be potential out of all the possible $M * I$ interactions), while missing no more than one verified miRNA-mRNA interactions in the data set. In the second main step a genetic algorithm optimizes affine gap alignment [1] parameters, which are used to align the miRNA-mRNA pair, extending the seeds found in the first main step. This is the general frame on top of which the following subsections will be adding more details.

4.1 Methods

4.1.1 Building a seed pattern

The aim of this step is to identify a pattern that represents the rules of seed - seed match alignments in known miRNA-mRNA interactions. As a quick recap, as mentioned in Section 2.2 of this thesis, *seed* is the short part of the miRNA, often located in the very 5' end of the miRNA, that has strong complementary with a *seed match* section of the target 3' UTR. As a prefiltering step, it is important to avoid excessive loss of sensitivity, which comes down to avoiding exclusion of known miRNA-mRNA interactions in the used training set.

The *seed patterns* used are tuples of four integers, which characterize the parameters of thought seed - seed match alignments in experimentally verified miRNA-mRNA interactions. These four parameters and their ranges are the length of the seed (from

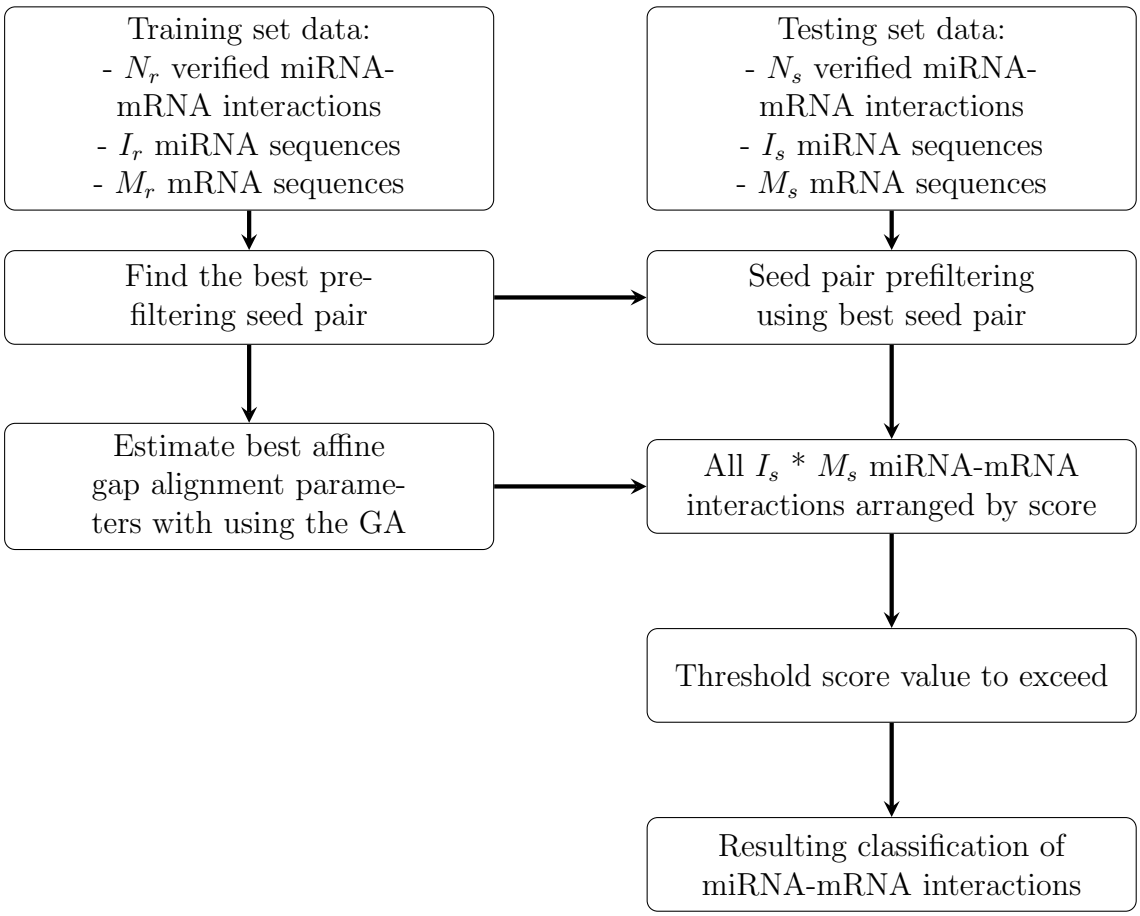

Figure 17: Flowchart describing the workings of the GA-based target predictor.

6 to 8 nt), the offset of the seed from the 5' end of the miRNA (0 to 2 nt), maximum allowable number of G:U wobble pairs in the alignment (0 to 3), and maximum allowable number of errors in the alignment (0 to 2). The aforementioned errors can be substitutions, deletions and insertions, of which the latter two have on impact on the actual length of the seed sequence.

As an example, a seed pattern (6, 2, 2, 1) represents 6 base pair long alignment, starting from position 3 from the 5' end of the miRNA. This alignment is allowed to have up to two G:U wobble pair instead of perfect Watson-Crick nucleotide pairs. Up to one error of type substitution, insertion or deletion is allowed to occur in the seed.

The algorithm starts by generating all possible seed patterns using the ranges described above. A seed pattern is then used to find potential seed matches in the 3' UTRs of the training set and the results for each seed pattern are saved into a table.

The search for each miRNA - seed pattern combination is done by generating all sequences the pattern allows from the miRNA. With a pattern (6, 2, 0, 0) and the miRNA let-7 (the sequence of which is 'UG**AGGUAG**UAGGUUGUAUAGUU'), as a simple example, the algorithm would search for all exact complementary matches to 'AGGUAG' in the 3' UTR database of the set. It is noteworthy that allowed G:U wobble pairs and errors add to the total number of exact matches to be searched for per miRNA. To improve the run times of the search, a simple index structure is used.

To increase the discriminative power of this phase, all possible pairs of seed patterns are generated. These pairs are called *seed pairs*. As an example, a seed pair (7, 1, 0, 1) or (8, 2, 0, 0) means that a 3' UTR sequence is considered as a hit if it contains a match to either or both of the seed patterns. If a seed pair misses more than one verified miRNA-mRNA interaction, the pair in question is discarded. For each seed pair, the total number of hits is recorded. The seed pair reporting smallest number of total hits (miRNA-mRNA pairs predicted to be potential interactions) in the training set is considered to be the best seed pair.

The ranges for the seed pattern parameters are based on rules used by earlier miRNA target prediction methods. The number of seed patterns combined was chosen to be two patterns due to no gain in adding an additional third pattern. More details about this testing are shown in Subsection 4.2. Limiting the ranges and the number of patterns improves the run times of this prefilterative seed pattern step, which in turn makes it possible to use this tool on large sets of data. Comparisons between two and three seed patterns, and the run times for different settings are shown in Subsection 4.2 of this thesis.

Only the miRNA-mRNA pairs that have passed this prefilterative seed pattern step will be further processed. This further processing will be described in the next subsection.

Parameters	Range	Number of bits
Gap opening penalty	-16 – 0	7
Gap extension penalty	-16 – 0	7
Match score	-10 – 10	8
Mismatch score	-10 – 10	8
G:U wobble pair score	-10 – 10	8
5'-end multiplier	0 – 10	7
Middle multiplier	0 – 10	7
3'-end multiplier	0 – 10	7

Figure 18: Affine gap alignment parameters optimized by the genetic algorithm with their allowed ranges and the amount of bits required to store each parameter.

4.1.2 Finding alignment parameters

After the potential miRNA-mRNA pairs have been found with the best seed pair in the previous step, the seed extension is performed to align the rest of the miRNA with its target. Most target predictors use default values for the match/mismatch scores and gap penalties. While some tools will allow the user to change these parameters, the choice of which parameters are the best to use, is not an easy one.

The tool presented here uses a genetic algorithm to find the best alignment parameters. It is noteworthy that even though genetic algorithms have been used before in miRNA target prediction [44], these methods have been only applied to phases that are more similar to the seed pattern phase of the method presented here, instead of the alignment phase. Additionally it is important to realize and remember that genetic algorithms are not a "black-box" type of method one can simply plug in to a tool. The exact choices of mechanisms, fitness functions and GA-parameters used have drastic effect on the efficiency and the final result of the GA. The general terms and properties of genetic algorithms were more widely discussed in Section 3.2.1 of this thesis.

In the genetic algorithm used here, a GA individual represents a set of parameters for an affine-gap pairwise alignment [1]. These parameters are encoded as bit strings using Gray encoding [16]. Each individual is represented by 59 bits. To eliminate illegal solutions, the range for each parameters are limited. These limitations and the number of bits required to store each parameter are shown in Figure 18. Note that the positional multipliers are used to score the beginning (5' end of the miRNA), middle and end (3' end of miRNA) of the alignment. Each of these regions is one third of the entire alignment.

Each match found by the seed pair in the previous step is extended to form a potential miRNA-mRNA duplex and is scored using the parameters encoded by the GA individual. A separate score, s_i is assigned to each verified miRNA-mRNA

interaction in the training set, by determining the number of alignment scores that are lower than the score of this verified miRNA-mRNA interaction, and adjusting it for the sample size. The highest alignment score of a miRNA-mRNA pair is chosen to represent the pair in question. The fitness function is defined as the sum of all s_i's divided by the number of verified miRNA-mRNA interactions in the training set. The objective is to find the individual with the largest fitness. Thus the genetic algorithm attempts to find the parameters that match the known miRNA-mRNA interactions as well as possible, ranking the verified interactions as high as possible in the group of all possible miRNA-mRNA pairs.

In summary the GA works as follows.

1. Generate a random individual with each bit set randomly to one or zero.

2. Calculate the fitness of this individual and place it into the population list ordered by the fitness.

3. Randomly select two individuals from the population, so that the probability of the individual with the highest fitness to be selected is σ. The initial value for σ is set to 0.25.

4. Generate the offspring of these two individuals by selecting a random crossover point and perform a one-point crossover.

5. Swap each bit of this new individual with probability of 0.005.

6. Calculate the fitness of the new individual and place it at the population list.

7. If the individual is not the best individual in the population, reduce σ by a small amount to place larger emphasis to exploration. Otherwise reset σ to its initial value. Return to step 2.

There is no stopping condition mentioned for the GA, as it was manually stopped when no significant improvement in the best individual had happened for a large amount of new individuals created. This was due to the long run times of the tool with real data. Unwanted termination of the GA would have caused large delays in getting results.

It is noteworthy, that the seed pattern step needs to be done only once for the training set in question, whereas the genetic algorithm runs several iterations to find the best alignment parameters. Because of this, these two different operations are separated as completely different steps in the workings of the tool.

4.2 Results

4.2.1 Data sets

A list of 104 experimentally verified human miRNA-mRNA interactions were downloaded from TarBase [38] to be used as positives in the training and test data. All of the miRNAs in this set translationally repress their mRNA targets. The 53 miRNA sequences in the aforementioned interactions were obtained from the miRBase database [17]. Human 3' UTR data used in the experiments consisted of 19,347 3' UTRs and was obtained from the Ensembl database [4].

The aforementioned data was divided into a total number of 13 training and test set pairs. All of these divisions were done by giving each of the 104 verified interactions a probability of 0.60 to belong to the training set. If the verified interaction was not chosen for the training set, it would belong to the test set. This resulted in 13 different training and test sets, where the training set sizes varied from 55 to 76 and the test set sizes varied from 49 to 28 accordingly. More specific sizes will be listed in the following subsection describing the experiments in detail.

No specific negative examples of miRNA-mRNA interactions were used in the training. As mentioned in the beginning of Section 4, our tool aims for maximal sensitivity while minimizing the total number of predictions in order to reduce the relative amount of false positives. In our data set, the group of verified interactions form the group of positives, which will be classified as either positive or negative by the classifier, splitting the group of positives efficiently in to True Positives (TP) and False Negatives (FN), accordingly. Even though there is no confirmed negative data, all the miRNA-mRNA pairs, whose relationship is unknown are labeled as negatives in terms of machine learning. The classifier will label these miRNA-mRNA pairs in the negative set either positive or negative, splitting the negative group in to False Positives (FP) and True Negatives (TN), accordingly.

The number of negatives in each test and training set is $N_{neg} = N_{tot} - N_{pos}$. Here the total amount of miRNA-mRNA pairs $N_{tot} = M * I$, where I is the number of miRNAs and M is the number of mRNAs in the training or test set in question. Some of the experiments in the following section were done with only the 3' UTRs related to the verified miRNA-mRNA pairs, while others were done with the whole collection of available 19,347 3' UTRs. The choice of the 3' UTR collection has a strong influence to the amount and nature of negative data, and onwards to the resulting specificity values and ROC curves. It is important to note and remember the negative data used in the following experiments differs drastically from the random and shuffled negative examples used in some methods mentioned in Section 3.

Set	N_tot=l*M	2 patterns:: TP/(TP+FN)	Hits	TP/ (TP+FN)	FP/ (FP+TN)	3 patterns:: TP/(TP+FN)	Hits	TP/ (TP+FN)	FP/ (FP+TN)
1	1140	(6,0,1,0) or (8,2,0,1):: 39/41	886	0.95	0.77	(6,0,0,0) or (6,0,1,0) or (8,2,0,1):: 39/41	886	0.95	0.77
2	1044	(6,0,2,0) or (6,2,1,0):: 36/40	805	0.9	0.77	(6,0,0,0) or (6,0,2,0) or (6,2,1,0):: 36/40	805	0.9	0.77
3	792	(6,0,2,0) or (8,2,0,1):: 33/34	648	**0.97**	**0.81**	(6,0,0,0) or (6,1,1,0) or (6,2,1,0):: 30/34	582	**0.88**	**0.73**
4	1280	(6,0,2,0) or (8,2,0,1):: 43/44	1109	0.98	0.86	(6,0,0,0) or (6,0,2,0) or (8,2,0,1):: 43/44	1109	0.98	0.86
5	780	(6,0,1,0) or (8,2,0,1):: 32/34	597	**0.94**	**0.76**	(6,2,1,0) or (7,0,1,0) or (8,1,0,1):: 30/34	564	**0.88**	**0.72**
b1	928	(6,0,2,0) or (6,2,1,0):: 33/37	725	0.89	0.78	(6,0,0,0) or (6,0,2,0) or (6,2,1,0):: 33/37	725	0.89	0.78
b2	1260	(6,2,1,0) or (8,1,0,1):: 40/49	862	0.82	0.68	(6,2,0,0) or (6,2,1,0) or (8,1,0,1):: 40/49	862	0.82	0.68
b3	690	(6,2,1,0) or (7,0,0,1):: 33/33	609	**1**	**0.88**	(6,0,2,0) or (6,2,1,0) or (8,2,0,1):: 33/33	553	**1**	**0.79**
b4	924	(6,2,1,0) or (7,1,0,1):: 36/37	845	**0.97**	**0.91**	(6,0,2,0) or (6,2,1,0) or (8,2,0,1):: 37/37	790	**1**	**0.85**
b5	841	(6,0,1,0) or (8,2,0,1):: 34/36	680	**0.94**	**0.8**	(6,2,1,0) or (7,0,1,0) or (8,0,0,1):: 32/36	668	**0.89**	**0.79**
b6	600	(6,2,1,0) or (7,0,0,1):: 28/28	556	**1**	**0.92**	(6,0,1,0) or (6,2,1,0) or (8,2,0,1):: 27/28	502	**0.96**	**0.83**
b7	1476	(6,0,2,0) or (6,2,1,0):: 42/46	1070	0.91	0.72	(6,0,0,0) or (6,0,2,0) or (6,2,1,0):: 42/46	1070	0.91	0.72
b8	1656	(6,0,1,0) or (8,2,0,1):: 45/47	1314	**0.96**	**0.79**	(6,2,1,0) or (7,0,1,0) or (8,2,0,1):: 42/47	1280	**0.89**	**0.77**

Figure 19: Best seed pairs and triplets from 13 training sets and their results on complementary test sets.

4.2.2 Experiments

Using all of the 13 training sets, optimal seed pattern combinations with two and three seed patterns were sought. With three seed patterns, the best pattern combination was considered to be the one returning the smallest amount if miRNA-mRNA pairs to be potential, while missing no more than one verified interaction, just as was the case with seed pairs. The best seed pairs and seed pattern triplets were used to find potential miRNA-mRNA pairs in the corresponding 13 test sets. The sizes of test sets, the best seed pairs and triplets, and the test set results are shown in Figure 19. In these tests, only the 3' UTRs which occur in the verified interactions of the set in question are used.

The first column labeled *Set* states the designated name of the set in question, consisting of a test and a training set. The second column shows the total number of miRNA-mRNA pairs one can form from the miRNAs and 3' UTRs of the test set. The third column specifies the seed pair with best performance on the training set, followed by the number of verified miRNA-mRNA interactions found out of all verified miRNA-mRNA interactions in the test set. Fourth column shows the number of miRNA-mRNA pairs the seed pair considers to be potential interactions, i.e. the hits. Fifth column declares the fraction of verified miRNA-mRNA interactions found by this seed pair, i.e. the sensitivity of the seed pair in this test set. The sixth column declares the fraction of all non-verified miRNA-mRNA pairs labeled as hits, i.e. the false positive rate (FPR) of the seed pair. Seventh column specifies the seed pattern triplet with best performance on the training set, followed by the number of verified miRNA-mRNA interactions found out of all verified miRNA-mRNA interactions in the test set. The eighth, ninth and tenth column declare the same results for the seed pattern triplet, that the columns four, five and six declared for the seed pair.

If there are differences between the results of pattern pairs and pattern triplets on the same sets, the values in columns "$TP/(TP + FN)$" and $FP/(FP + TN)$

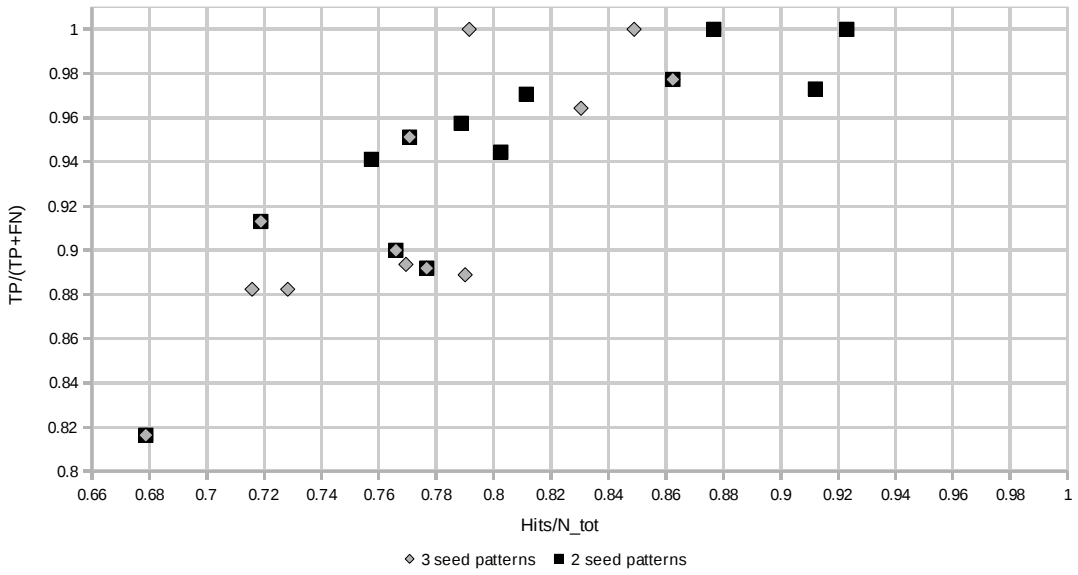

Figure 20: The sensitivities and false positive rates using seed pattern pairs compared to using seed pattern triplets with 13 test sets.

(columns 5, 6, 9 and 10) are boldface. The higher values are underlined in the case of sensitivity, and the lower values are underlined in the case of false positive rates. The sensitivity and false positive rates of the test set results are plotted in Figure 20.

In the case of most sets (sets 1–b2, b6 and b7) no improvement was gained by using three seed patterns instead of two, when comparing the sensitivity to the false positive rates. In the case of sets b3 and b4, the sensitivity remained the same or was improved and the false positive rate was decreased at the same time, when three seed patterns were used instead of two. In the case of sets b5 and b8, the addition of third allowed pattern decreased the sensitivity drastically, while decreasing the false positive rate only a small amount. The general trend of the effect the addition of third allowed pattern had, was decreased false positive rate at the expense of decreased sensitivity. As the purpose of the seed pattern is to act as a prefilterative step, this sort of trend is not something one would hope for. Moreover, the usage of an additional seed pattern increases the run times of this step by a very notable amount, as the amount of possible seed pattern combinations is multiplied by the number of possible seed patterns.

The seed pair phase was run using the miRNAs and verified interactions from the sets 1–5 and the full 19,347 3' UTRs. These training and test set pairs with the full amount of 3' UTRs were named 1F–5F. The best seed pairs resulting from the training set runs and the results from using these pairs on the test sets are shown in Figure 21.

Set	N_tot=I*M	2 patterns:: TP/(TP+FN)	Hits	TP/ (TP+FN)	FP/ (FP+TN)
1F	580410	(6,0,1,0) or (8,2,0,1):: 39/41	354897	0.95	0.61
2F	561063	(6,0,2,0) or (6,2,1,0):: 36/40	341091	0.9	0.61
3F	464328	(6,0,2,0) or (8,2,0,1):: 33/34	295036	0.97	0.64
4F	619104	(6,0,2,0) or (8,2,0,1):: 43/44	404185	0.98	0.65
5F	503022	(6,0,1,0) or (8,2,0,1):: 32/34	304501	0.94	0.61

Figure 21: Best seed pairs from 5 training sets and their results in corresponding test sets, using the full set of 19,347 3' UTRs.

Results of sets 1F–5F are very similar to the results of the sets 1–5. The resulting seed pairs and the amount of verified miRNA-mRNA pairs found from the test set were identical to the runs with sets 1–5. The false positive rates in 1F–5F are considerably lower: the average FPR dropped from 0.794 to 0.624. This means that the seed pair phase found relatively larger amount of new potential miRNA targets within the mRNAs that were already known to act as targets for some miRNAs. Additionally, the average number of seed hits found in the 3' UTRs of verified interactions was 4.1, while the average for all 3' UTRs containing at least one seed hit was 3.2. This was tested on a sixth set resulting in the same seed pair as sets 3F and 4F, using all 3' UTRs. The results suggest that the seed pair phase has been able to capture some collection of properties, that are more common within the known verified interactions than other miRNA-mRNA pairs in general. This infers that the designed prefilterative step is functioning as desired.

The sets 1–5 and 1F–3F were further refined using the genetic algorithm described above in Subsection 4.1.2. The GA was run, using the training sets of each set, until approximately 100,000 individuals had been created. After this, the runs were manually stopped. The choice of the number of individuals to be created before termination of the algorithm was based on multiple earlier test runs. The run times were between 6 and 12 hours for sets 1–5, and between 70 and 120 hours for sets 1F–3F. All runs were performed on a computer running Fedora 9 with two 2.0Ghz AMD Opetron^{TM}246 processors and 5.8GiB memory. The algorithm was implemented using the C++ language.

Alignment parameters corresponding to the best individual of each run are listed in the table in Figure 22. The last column of the table contains the alignment parameters used by the algorithm miRanda [22]. The values in the table are rounded to the nearest hundredths. Exact values are determined by the ranges and number of bits per value defined in the table in Figure 18.

Parameter	Set 1	Set 2	Set 3	Set 4	Set 5	Set 1F	Set 2F	Set 3F	miRanda
Gap opening	-8.63	-15	-14.38	-7.88	-13.25	-9.5	-15.88	-6.88	-8
Gap extension	-2.88	-14.13	-11.5	-10.5	-3.38	-15.88	-11.25	-15.63	-2
Match Score	5.23	2.97	9.38	2.03	2.11	2.89	3.83	3.36	5
Mismatch Score	-5.08	-6.02	-0.63	-3.98	-5.55	-3.52	-7.42	-3.67	-3
G:U score	0.47	-4.14	-10	-5.39	-6.41	-5.94	-0.86	-6.48	2
5'-end multiplier	2.73	5.7	3.91	7.27	5.23	4.61	-3.91	5.47	2
Middle positional multiplier	9.61	3.05	9.92	1.88	1.02	3.13	2.11	3.59	1
3'-end multiplier	3.91	0.55	5.94	0	6.33	1.25	0	1.41	1

Figure 22: Alignment parameters corresponding to the best individual of GA runs with 8 different training sets, and the alignment parameters used by the miRanda tool.

There is very notable variation of single parameter values between the tests. Some strong hypothesis about different types of miRNA-mRNA binding exist. The duplexes can be divided into canonical duplexes that pair well along the entire length of the miRNA and seed duplexes that have little or no pairing at the 3' end of the miRNA [25]. As the GA-based target predictor searches for a single set of parameters to fit well with all the miRNA-mRNA pairs of a training set, the distribution of the type of duplexes in the training set most likely has strong influence on the final values of the alignment parameters. There is indeed strong variation in the positional multipliers, that can be seen in the bottom three rows of Figure 22. This supports the hypothesis of the existence of different types of miRNA-mRNA duplexes.

The performance of the alignment parameters found by the genetic algorithm were compared to the miRanda parameters for test sets 1F, 2F and 3F. Using the parameters in question, the best affine gap alignment score for a miRNA-mRNA pair was chosen the represent the pair. Varying a score threshold, the receiver operating characteristic (ROC) curves for each parameter set were obtained. These ROC curves are shown in Figure 23. As the data has already gone trough the prefilterative seed pair step, the sensitivity does not reach 1 and the false positive rate is below 0.65 for all classifications.

As can be seen from the figure, the alignment parameters discovered by the genetic algorithm produce more accurate classifications than the parameters used by miRanda. In majority of cases, the parameters obtained from the GA optimization give higher sensitivity than the miRanda parameters on the same false positive rate. This is especially true for set 2F. These results show that the alignment parameters optimized by the GA are superior to the hand tuned alignment parameters the miRanda miRNA target prediction tool uses.

Considering the fact that the optimized parameter values do vary quite drastically from set to set, it is certainly interesting that miRanda parameters still seem to be worse for every set. The differences between GA-parameters and miRanda parame-

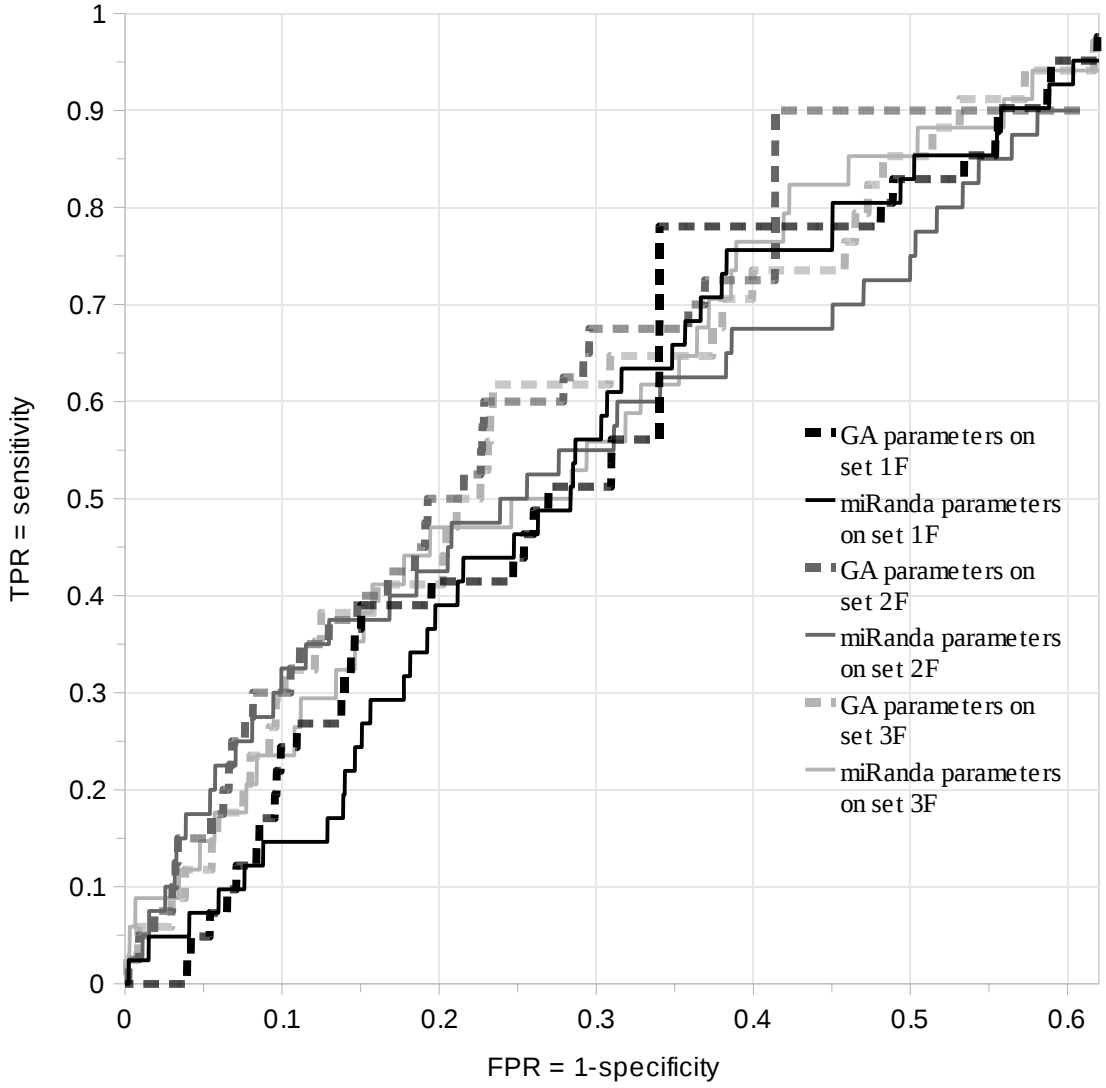

Figure 23: Receiver operating characteristic (ROC) curves for GA optimized alignment parameters and miRanda parameters on test sets 1F, 2F and 3F.

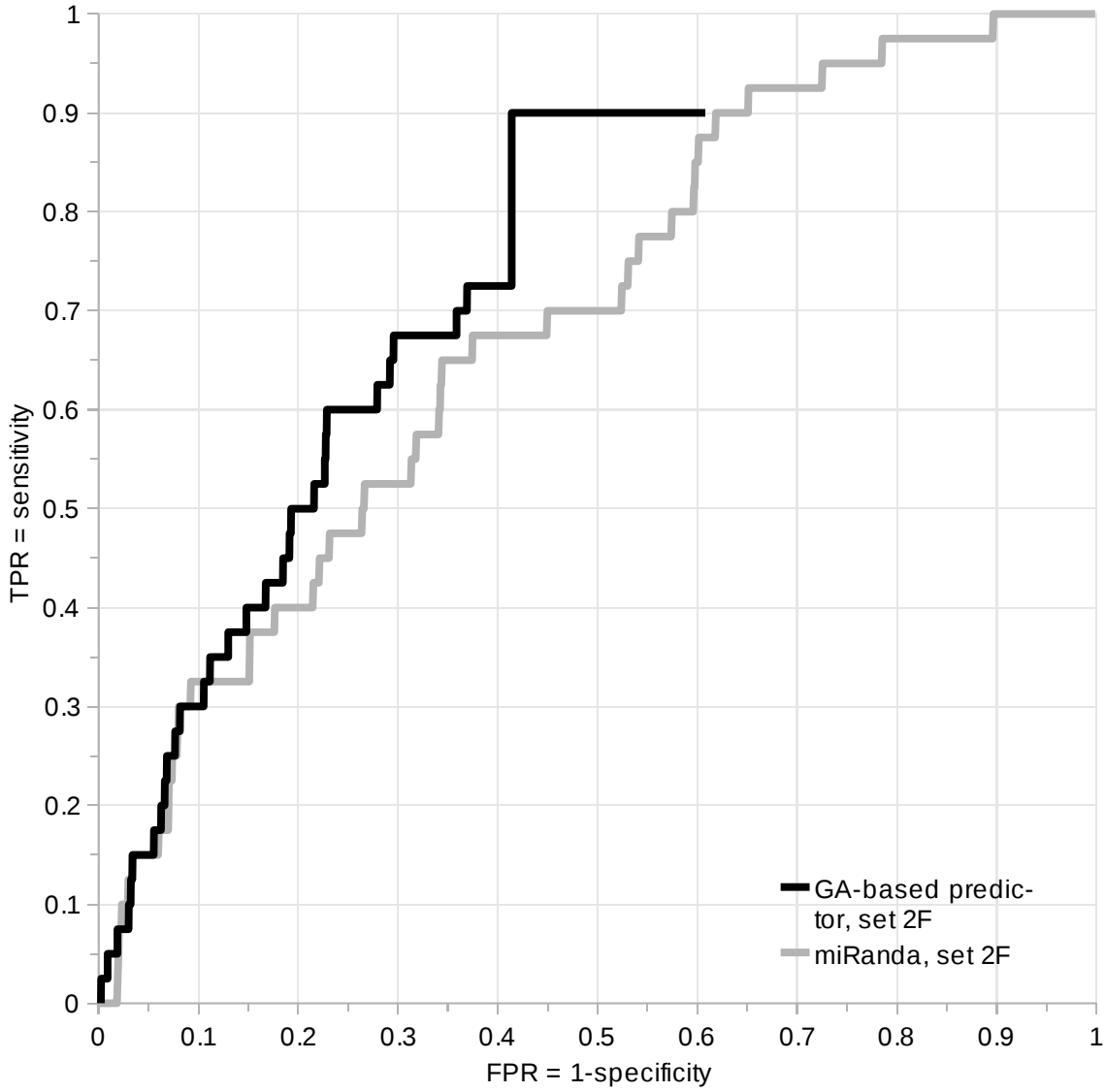

Figure 24: Receiver operating characteristic (ROC) curves for GA-based predictor and the miRanda prediction tool on set 2F.

ters are not remarkably huge though.

The miRanda prediction tool was run for the data in test set 2F. MiRanda was run using very low score and folding free energy thresholds of 0. The best score for each miRNA-mRNA pair in the set was chosen to represent the pair in question. Varying the score threshold, a ROC curve for this miRanda run was obtained. These results were compared with the performance of the GA-based predictor on the set 2F. ROC curves for the two prediction tools are shown in Figure 24.

The results in Figure 24 show that the GA-based target predictor outperforms the miRanda tool on a very large range of false positive rates. The seed pair step of

the GA-based predictor cuts out the last 10% of true positives, as can be seen from the table in Figure 21, so there is no further improvement in sensitivity after the false positive rate of 0.414. On false positive rates lower than 0.1, the two tools performed equally well.

Using the folding free energy and alignment score thresholds of 0, the miRanda run took approximately 5 days. This sort of run time was felt to be a bit over the line, so the further comparisons were done with miRanda runs on the sets 1–5, which have significantly smaller number of miRNA-mRNA pairs to be studied.

The ROC curves of the GA-based predictor and the miRanda tool on sets 1–3 are shown in Figure 25 and the ROC curves on sets 4–5 are shown in Figure 26. The figures show that the GA-based target predictor outperforms miRanda on these smaller data sets. With false positive rates above 0.13, the ROC curve of the GA-based predictor is above the ROC curve of miRanda in all of the cases.

When comparing the performances of the two tools on set 2 to the performances on set 2F, one can note that sensitivity values on similar false positive rates are much lower on set 2. It is likely that a random sequence in a 3' UTR of a known miRNA target has more in common with the predicted binding sites in known miRNA-mRNA interactions, than a random sequence in a random 3' UTR. Because of this, the 3' UTRs of verified interactions provide a more challenging ground for the classification problem, than the whole collection of human 3' UTRs. As can be seen in Figures 25 and 26, the performance of miRanda predictor on sets 1–5 is actually very close to a random classifier. Considering the fact that miRanda is a very popular and established miRNA target prediction tool, it is easy to realize the level of the challenge in this classification problem with these kind of data sets.

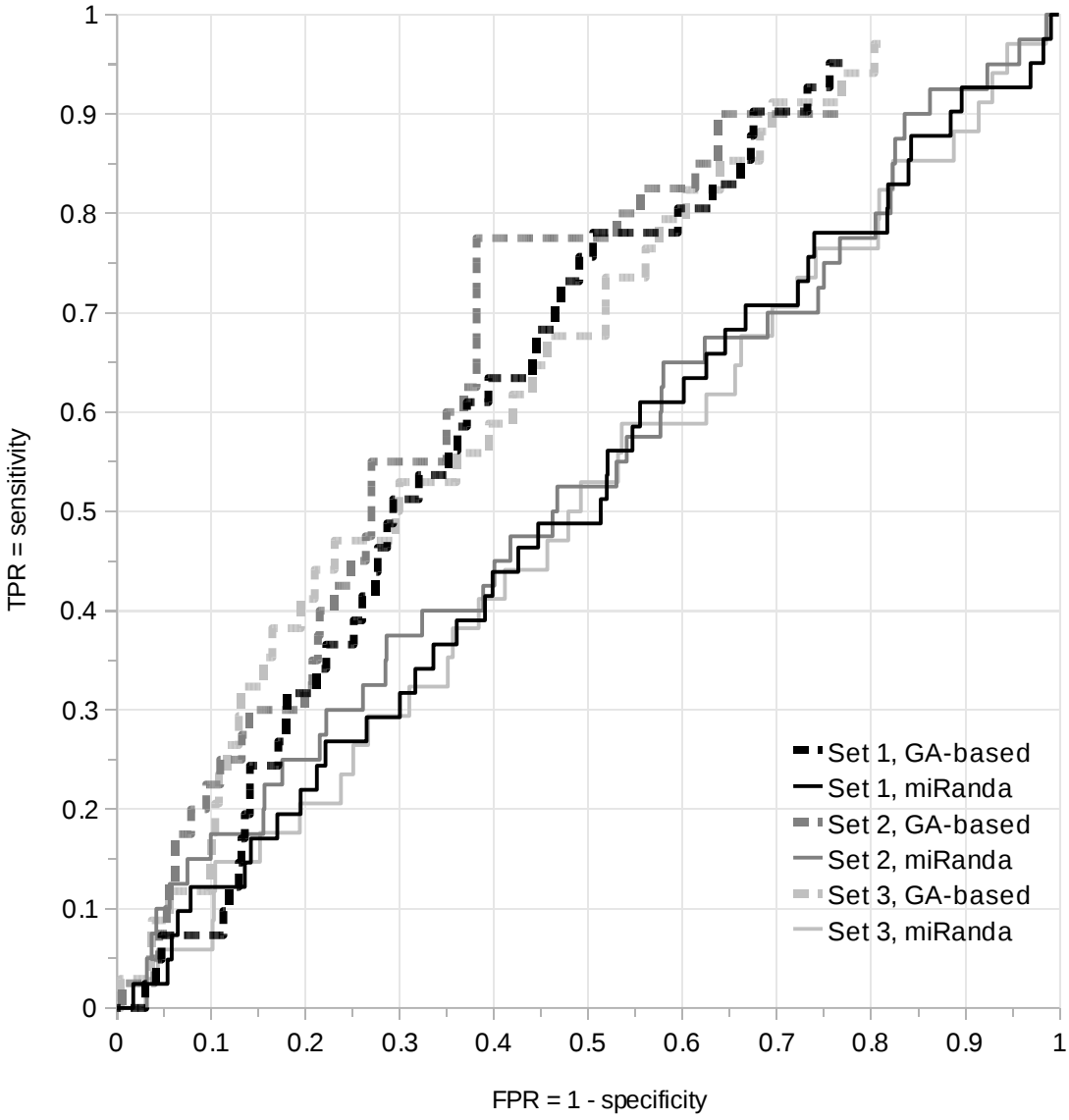

Figure 25: Receiver operating characteristic (ROC) curves for GA-based predictor and the miRanda prediction tool on sets 1, 2 and 3.

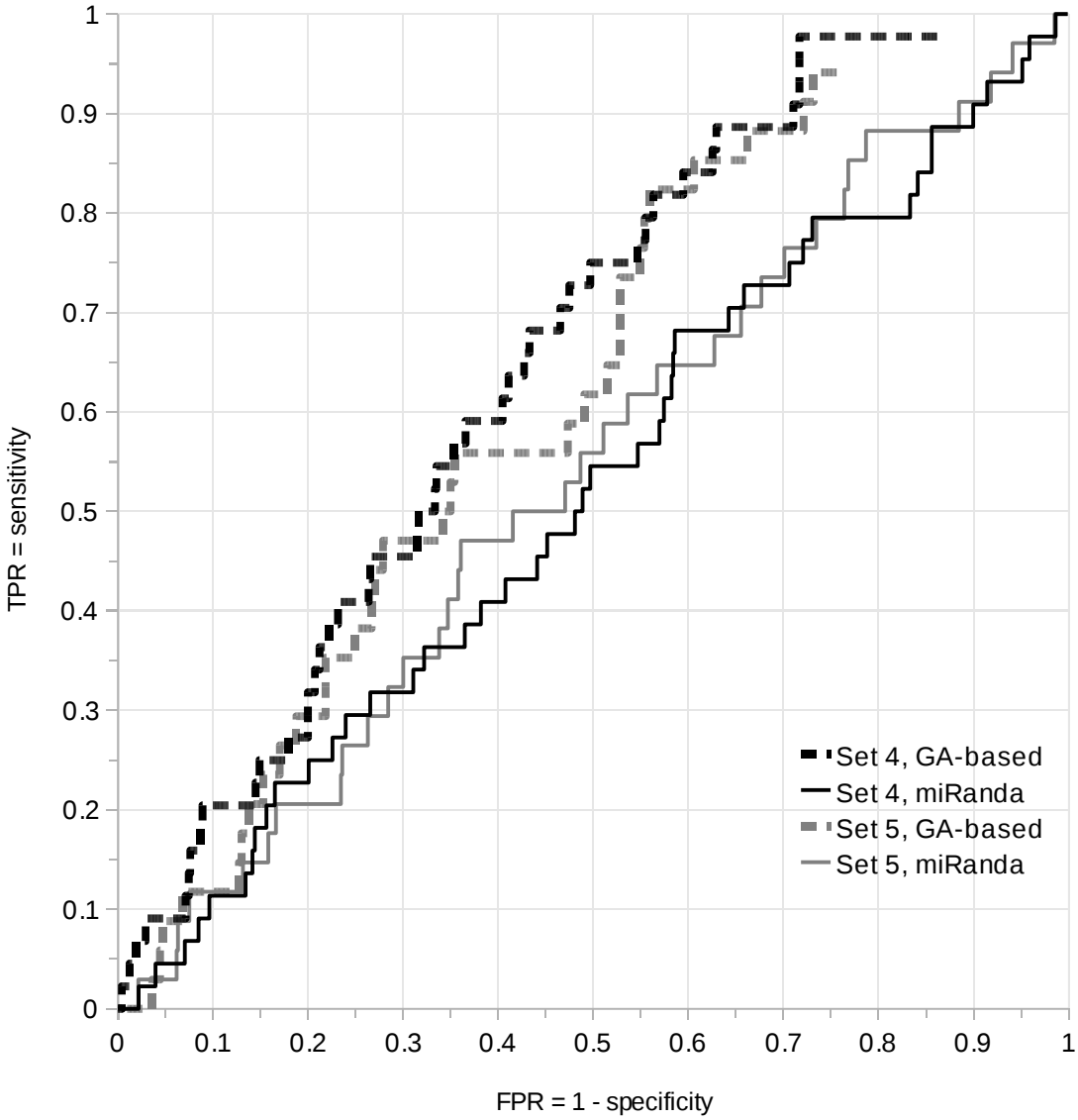

Figure 26: Receiver operating characteristic (ROC) curves for GA-based predictor and the miRanda prediction tool on sets 4 and 5.

5 Conclusions

5.1 Summary

The objective of this thesis was to provide an insight to the field of miRNA target prediction in animals. This thesis introduced a new GA-based miRNA target prediction tool. Additionally, total of seven existing miRNA target prediction tools were presented. Four of these, the TargetScan [32], miRanda [22], RNAhybrid [42] and Sfold [33], were traditional target prediction tools. Three of the tools, namely TargetBoost [44], PicTar [29] and miTarget [27], were machine learning tools.

Necessary biological background information was presented, to describe the biological frame of the target prediction problem. Examples of dynamic programming and genetic algorithms were also introduced, in order to help the reader in understanding the workings of various methods used by the various prediction tools.

It was shown that the GA-based target predictor outperformed the miRanda tool in the classification of potential miRNA-mRNA interactions, resulting in higher sensitivity values with identical specificity values. It was additionally shown, that the affine gap alignment parameters produced by the GA resulted in better performance than the hand tuned parameters used by miRanda target prediction tool.

5.2 Discussion and future work

There are still open questions about the miRNA binding to mRNA as a biological phenomenon. This causes problems especially for traditional target prediction tools, which try to mimic the biological workings of the binding as well as possible. For machine learning prediction tools, the relatively small number of verified miRNA-mRNA interactions poses the main problem. This is especially true for verified negative data. Machine learning methods commonly use unknown miRNA-mRNA pairs or predicted binding of shuffled miRNAs to cover up the lack of adequate negative training data.

Improvements in the level of biological knowledge of the binding will have a positive effect on the performance of prediction tools. New discoveries of the binding mechanics can be put in to use by mimicking them with new kind of methods. This is true for both traditional and machine learning tools. As more verified miRNA-mRNA interactions are discovered, even the existing machine learning tools will most probably start producing more reliable and accurate classificators.

There are numerous improvements that can be done for the GA-based predictor presented in Section 4, in both the seed pair step and the alignment parameter optimization step. Some ideas for future work, which emerged during the process of working with the subject, are presented here.

Currently, the GA-based tool is not using any of the common prefilterative steps used by some other algorithms, such as requiring certain level of orthology for a potential target site. The influence such filters would have on the ROC curves of the final classification could be studied.

The interaction of multiple miRNA target sites in a single 3' UTR could be taken into account in the predictions. Examples of such target site interactions are introduced in the paper [10].

The addition of new parameters and modification of the current ranges could be further studied. Especially the hypothesis about different types of miRNA-mRNA duplexes could be taken into account. This could be done by allowing multiple weighting schemes (3', middle and 5' positional multipliers) in the affine gap alignment parameter optimization, selecting the weighting scheme resulting in the highest alignment score. Additionally, the alignment score could be normalized by the length of the alignment and using the average result for a random sequence of the same length. This would allow the user to compare potential targets from two different runs, using two different sets of alignment parameters.

The fitness function of the genetic algorithm could be modified. One idea would be to use weighting, which relatively increases the importance of correct predictions on low sensitivity and false positive rates. This is the area where the performance of the GA-based predictor is worst, when comparing to miRanda.

Currently the GA is terminated manually, which is not very sophisticated or convenient, especially for an outside user. Possible termination conditions, which would not cause too early termination, could be studied further.

The processor and memory consumption of the tool in various phases of the execution could be further analyzed in order to make the tool more computationally efficient.

Some experimenting with using shuffled miRNAs to get a measure for signal to noise ratio could be done. The methods the authors of TargetScan [32] used could be mimicked when producing these shuffled miRNAs.

Additional comparisons to other miRNA target prediction tools, especially machine learning tools, could be done. If the run times of the tools cause problems for plotting ROC curves with sensible amount data, smaller level comparisons of predictions could be done instead.

References

[1] S. F. Altschul and B. W. Erickson. Optimal sequence alignment using affine gap costs. *Bulletin of Mathematical Biology*, 48:603–616, 1986.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, October 1990.

[3] D. Betel, M. Wilson, A. Gabow, D. S. Marks, and C. Sander. The microrna.org resource: targets and expression. *Nucleic Acids Research*, 36(Database Issue):D149–53, 2008.

[4] E. Birney, T. D. Andrews, P. Bevan, M. Caccamo, Y. Chen, L. Clarke, G. Coates, J. Cuff, V. Curwen, and T. Cutts. An overview of ensembl. *Genome research*, 14(5):925–928, 2004.

[5] J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen. Principles of microrna–target recognition. *PLoS Biology*, 3(3): e85, 2005.

[6] A. Brindle. Genetic algorithms for function optimization. *(Doctoral dissertation and Technical Report TR81-2). Edmonton: University of Alberta, Department of Computer Science.*, 1981.

[7] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[8] F. H. C. Crick. Codon-anticodon pairing: The wobble hypothesis. *Journal of Molecular Biology*, 19:548–555, 1966.

[9] B. R. Cullen and H. Hughes. Transcription and processing of human microrna precursors. *Molecular Cell*, 16:861–865, 2004.

[10] D. Didiano and O. Hobert. Molecular architecture of a mirna-regulated 3' utr. *RNA*, 14:1297–1317, 2008.

[11] J. G. Doench and P. A. Sharp. Sirnas can function as mirnas. *Genes & Development*, 17:428–442, 2003.

[12] J. Dostie, Z. Mourelatos, M. Yang, A. Sharma, and G. Dreyfuss. Numerous micrornps in neuronal cells containing novel microrna. *RNA*, 9:180–186, 2003.

[13] S. R. Eddy. What is dynamic programming? *Nature biotechnology*, 2004.

[14] A. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. Marks. Microrna targets in drosophila. *Genome Biology*, 4, 2003.

[15] A. Eulalio, E. Huntzinger, T. Nishihara, J. Rehwinkel, M. Fauser, and E. Izaurralde. Deadenylation is a widespread effect of mirna regulation. *RNA*, 15:21–32, 2009.

[16] F. Gray. Pulse code communication, 1947.

[17] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright. mirbase: tools for microrna genomics. *Nucleic Acids Research*, 36:D154–D158, 2008. [online] `http://www.mirbase.org/`.

[18] J. Han, Y. Lee, K.-H. Yeom, J.-W. Nam, I. Heo, J.-K. Rhee, S. Y. Sohn, Y. Cho, B.-T. Zhang, and V. N. Kim. Molecular basis for the recognition of primary micrornas by the drosha-dgcr8 complex. *Cell*, 125:887–901, 2006.

[19] P. G. Hawkins and K. V. Morris. Rna and transcriptional modulation of gene expression. *Cell Cycle*, 7(5):602–607.

[20] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tracker, and P. Schuster. Fast folding and comparison of rna secondary structures. *Monatshefte für Chemie*, 125:167–188, 1994.

[21] G. Hutvágner and P. D. Zamore. A microrna in a multiple-turnover rnai enzyme complex. *Science*, 297(2056), 2002.

[22] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks. Human microrna targets. *PLoS Biology*, 2(11), 2004.

[23] M. W. Jones-Rhoades and D. P. Bartel. Computational identification of plant microrna and their targets, including a stress-induced mirna. *Molecular Cell*, 14:787–799, 2004.

[24] A. K. D. Jong. An analysis of the behavior of a class of genetic adaptive systems. *Ph. D. Dissertation, University of Michigan*, 1975.

[25] K. Karhu, S. Khuri, J. Mäkinen, and J. Tarhio. Identifying human mirna targets with a genetic algorithm. *To appear in Proc. ISB 2010, International Symposium on Biocomputing*, 2010.

[26] H. Kawasaki and K. Taira. Microrna-196 inhibits hoxb8 expression in myeloid differentiation of hl60 cells. *Nucleic Acids Symposium Series*, 48:211–212, 2004.

[27] S.-K. Kim, J.-W. Nam, J.-K. Rhee, W.-J. Lee, and B.-T. Zhang. mitarget: microrna target gene prediction using a support vector machine. *BMC Bioinformatics*, 7:411, 2006.

[28] M. Kiriakidou, P. T. Nelson, A. Kouranov, P. Fitziev, C. Bouyioukos, Z. Mourelatos, and A. Hatzigeorgiou. A combined computational-experimental approach predicts human microrna targets. *Genes & Development*, 2004.

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

[30] Y. Kurihara and Y. Watanabe. Arabidopsis micro-rna biogenesis through dicer-like 1 protein functions. *Proc. Natl. Acad. Sci. U.S.A.*, 101(34):12753–12758, 2004.

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

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

[33] D. Long, C. Y. Chan, and Y. Ding. Analysis of microrna-target interactions by a target structure based hybridization model. *Pacific Symposium on Biocomputing*, 13:64–74, 2008.

[34] B. Mallick, Z. Ghosh, and J. Chakrabarti. Microrna switches in trypanosoma brucei. *Biochemical and Biophysical Research Communications*, 372:459–463, 2008.

[35] M. Mitchell. *An introduction to genetic algorithms.* MIT Press, 1st edition, 1998.

[36] P. T. Nelson, W.-X. Wang, and B. W. Rajeew. Microrna (mirnas) in neurodegenerative diseases. *Brain Pathology*, 18:130–138, 2008.

[37] R. Nussinov, G. Piecznik, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 25:68–82, 1978.

[38] G. L. Papadopoulos, M. Reczko, V. A. Simossis, P. Sethupathy, and A. G. Hatzigeorgiou. The database of experimentally supported targets: a functional update of tarbase. *Nucleic Acids Research*, 37:D155–D158, 2009.

[39] R. S. Pillai. Microrna function: multiple mechanisms for a tiny rna? *RNA*, 11:1753–1761, 2005.

[40] N. Rajewsky and N. D. Socci. Computational identification of microrna targets. *Dev. Biol.*, 267:529–535, 2004.

[41] N. Rajewsky, M. Vergassola, U. Gaul, and E. D. Siggia. Computational detection of genomic cis-regulatory modules applied to body patterning in the early drosophila embryo. *BMC Bioinformatics*, 3(30), 2002.

[42] M. Rehmsmeier, P. Steffen, M., and H. et al. Fast and effective prediction of microrna/target duplexes. *RNA*, 10:1507–1517, 2004.

[43] G. Ruvkun. Molecular biology: Glimpses of a tiny rna world. *Science*, 294(5543):797–799, 2001.

[44] O. Sætrom, O. Snøve, and P. Sætrom. Weighted sequence motifs as an improved seeding step in microrna target prediction algorithms. *RNA*, 11:995–1003, 2005.

[45] P. Sætrom. Predicting the efficacy of short oligonucleotides in antisense and rnai experiments with boosted genetic programming. *Bioinformatics Advance Access*, 2004.

[46] A. A. Saraiya and C. C. Wang. snorna, a novel precursor of microrna in giardia lamblia. *PLoS Pathog.*, 4(11), 2008.

[47] D. S. Schwarz and P. D. Zamore. Why do mirnas live in the mirnp? *Genes & Development*, 16:1025–1031, 2002.

[48] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[49] A. Stark, N. Bushati, and C. H. J. et al. A single hox locus in drosophila produces functional micrornas from opposite dna strands. *Genes & Development*, 22:8–13, 2008.

[50] K. Ukumura, J. W. Hagen, H. Duan, D. M. Tyler, and E. C. Lai. The mirtron pathway generates microrna-class regulatory rnas in drosophila. *Cell*, 2007.

[51] A. E. Williams. Functional aspects of animal micrornas. *Cellular and molecular life sciences*, 65(4):545–562, 2008.

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

[53] Y. Zhang and J. Skolnick. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophysical Journal*, 87(4):2647–2655, 2004.

[54] H. Zhou, S. B. Pandit, S. Y. Lee, J. Borreguero, H. Chen, L. Wroblewska, and J. Skolnick. Analysis of tasser-based casp7 protein structure prediction results. *Proteins: Structure, Function, and Bioinformatics*, 69:90–97, 2007.

[55] M. Zuker and P. Stiegler. Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.