In-Silico-Dicer

Intrinsic and Extrinsic Prediction of mature miRNA

Bachelor-Arbeit
im Fach Bioinformatik und Genomforschung
von
Sylvia Tippmann

30. September 2005

Betreuer: Dr. Marc Rehmsmeier
Prof. Dr. Robert Giegerich

Acknowledgement

I would like to express my gratitude to Marc Rehmsmeier for being an outstanding tutor and role model as researcher. His support encouraged my scientific curiosity and made this work exciting and successful. Furthermore, my sincere thanks go to Gregor Obernosterer and Professors Dorothee Staiger and Robert Giegerich for arousing interest in the subject of RNA mediated gene regulation and providing their knowledge for questions, which emerged during the work. I am also deeply indebted to Lena Büsch for her time and effort in reviewing my thesis.

Contents

1 Introduction

At all times intense genetic efforts have been applied to comprehend development. On that account it is surprising that a relatively large class of regulatory genes has surfaced only recently: The first mircoRNA gene and its developmental role was described more than ten years ago but researchers are enlightening the broad and abundant presence of such genes only now. MicroRNAs are short RNA sequences that use antisense complementarity to repress expression of specific messenger RNAs. Studies of functional roles have shown that miRNAs are involved in complex genetic pathways regulating, among others, neuronal differentiation, stem cell division, cancer, embryogenesis and hematopoiesis. Recent reports indicate that this is likely to be only the tip of an iceberg with plenty of regulating possibilities. MiRNAs might thus be previously underestimated key participants in the field of gene expression [Pasquinelli et al., 2005]. If we could fully understand their biogenesis and function in vivo, we would possibly be able to predict a many more target genes and to annotate them and furthermore knock down unwanted, maybe pathological, ones through introduction of artificial miRNA genes. The maturation of these smallRNAs, however, appears to be of high complexity since a large number of proteins are involved and, dependent on the organism, diverse procedures seem to apply. Ever so much bioinformatical research has been done developing different approaches to reconstruct the miRNA pathway, and was partially successful already. Some programs were published to predict possible miRNA precursors from whole genomes, and there exist algorithms which match the mature miRNA with its corresponding target gene. Unfortunately, little is known about the step in between, the processing of precursors to mature and functional active miRNA, so there are only few approaches to close this prediction gap. For this reason, the question is: Is it possible to reliably predict mature miRNAs from their precursors? Within the context of my bachelor thesis I have followed this question with different general ideas and developed a tool which deploys them.

2 Spreading silence - Biological background of smallRNA mediated silencing

Small RNAs are a family of regulatory RNA fragments which are mainly located in the intronic regions of an organism's genome. These non-coding RNAs originate from double-stranded(ds)RNA through a stepwise maturation process, and commonly they reach a length of 19 to 28 nucleotides in their functional stage. In this final mode, small RNAs are able to influence gene expression on a post-transcriptional level. Target genes are silenced via specific base pairing with the complementary small regulatory RNA. This mechanism had already been recognized in some eukaryotes about several years ago: In 1998 the first observation of RNA-mediated gene silencing was made in *Caenorhabditis elegans*, where dsRNA leads to the degradation of mRNA - it was called RNA interference (RNAi) but at this time the new phenomena was not associated with small RNA fragments [Fire et al., 1998]. Further research showed that similar regulation procedures not only occur in all kinds of metazoa, but also in plants, where it is termed 'co-suppression' or 'post-transcriptional gene silencing' (PTGS), and in funghi ('quelling') [Kim, 2005a and references therein]. However, RNAi is still a common synonym for gene silencing mediated by small RNAs in general, but in fact it refers to the actual cleavage of a mRNA target gene. The emerging biotechnology, which is based on an efficient gene knock-down using small RNA molecules, is also called RNA interference. This is going to be one of the leading technologies in detecting gene functions and developing genetic therapies.
In the following Section I will review the standard of knowledge around small RNAs, their classification, biogenesis and function.

2.1 Introduction to smallRNAs – A classification

Small RNAs in their final shape and function are sometimes indistinguishable; therefore they are classified by their origin. Basically, RNA can descend either from endogenous transcripts or exogenous sources of RNA, like retro-viruses for instance.
DsRNA transcripts give rise to endogenous small RNAs through a stepwise processing by endonuclease-III-type enzymes. These RNAs can be split up into two major groups: microRNA (miRNA) and short-interfering RNA (siRNA).

MiRNAs and siRNAs are often very close in function and biochemical structure but they do arise from different kinds of precursor states. While miRNA precursors fold back on themselves to form a hairpin-like structure, siRNAs derive from a perfect long double-stranded RNA duplex. Usually only one miRNA is generated from pre-miRNA but several or many siRNAs are generated from long dsRNA. Interestingly, the first miRNA, named lin-4, was already discovered in 1993 [Lee et al., 1993]. During mutation experiments with *C.elegans*, Lee et al. observed an oppositional progression of development. While mutating lin-14 lead to a premature transition into next larval stage, a lin-4 mutation caused a rerun of the current stage. As a conclusion they assumed lin-4 to have some regulatory influence on lin-14. Similar observations have been made with let-7 miRNA and its target lin-41. At first place, this did not seem very interesting to the research community but since with the discovery of RNAi, thousands of miRNAs have been found in nematodes, vertebrates, plants and even viruses [Kim, 2005a and references therein]. A useful database where all known and experimentally evaluated miRNAs and precursors are stored is the miRNABase (http://www.sanger.ac.uk/Software/Rfam/mirna/) [Griffiths-Jones, 2004], additionally the rfam database [Ambros V. et al. 2003] classifies miRNAs into families of homologous

function. Endogenous siRNAs have been identified since 2001 in *T.brucei*, *S.pombe*, *C.elegans*, *D.melanogaster*, and *A.thaliana* [Kim, 2005a and references therein]. Among siRNAs there are three different subclasses to be distinguished: repeat-associated RNAs (rasiRNAs), endogenous trans-acting siRNAs (tasiRNAs) and small scan RNAs.
SiRNAs can be looked up in the smallRNA database [http://web.mit.edu/mmcmanus/www/siRNADB.html]
Furthermore, there exist two more kinds of small endogenous RNA classes, which are not as well examined to classify them: tiny non-coding RNA (tncRNA) and small modulatory RNA (smRNA). Neither their biogenesis nor their action mechanism has become clear up to now. On the contrary, there are exogenous dsRNAs that are also able to induce expression of small RNAs. The naturally occurring ones include virus-induced siRNAs.

Class	Subclass	mature length	Biogenesis	Mechanism
microRNA (miRNA)	N.A.	19-24 nt	2-step processing of hairpin precursors by RNase-III-type enzymes	translational repression, mRNA cleavage
Short interfering RNA (siRNA)	trans-acting siRNA	21-22 nt	cleavage of long endogenous dsRNA by Dicer	mRNA cleavage
	repeat-associated siRNA (rasiRNA	24-26 nt (plants) 24-27 nt (animals)	cleavage of long dsRNAs derived from repetitive sequences or transposons by Dicer	modificaton of histone and/or DNA
	Small scan RNA (scnRNA)	~28 nt	cleavage of long endogenous dsRNA by Dicer	Histone methylation leading to DNA elimination

Table 1: Classification of small RNAs

2.2 SmallRNA mediated silencing mechanisms

Known so far, smallRNAs guide at least 4 distinct modes of gene silencing mechanisms: (a) endonucleolytic cleavage, (b) translational repression, (c) transcriptional repression and (d) DNA elimination through histone modification. The latter two are mediated by rasiRNAs and scanRNAs: rasiRNAs recruit DNA-cytosine methyltransferase and histone modifying enzymes. Together they form the RNA-induced-initiation-of-transcriptional-silencing-complex, or RITS, which methylates the DNA, leading to silencing at transcriptional levels (transcriptional gene silencing, TGS). ScanRNAs, however, induce histone methylation that leads to complete elimination of DNA.
If the small RNA molecules interact with mRNA on a post-transcriptional level, it depends on the sequence complementarity between the regulating and the regulated RNA which kind of gene silencing follows. If the complementarity is nearly perfect, which is usually the case in plants, mRNA cleavage is induced between the 10th and 11th nucleotide measured from the 5' end of the smallRNA. While cleaved mRNA is degraded afterwards by cellular nucleases, the miRNA remains intact and can guide the recognition and cleavage of additional transcripts. Cleavage reactions are mostly guided by siRNAs, which bind to an effector complex termed RISC (RNA induced silencing complex). However, there are also examples where miRNA initiates RNAi. Actually, miRNAs are usually involved in translational repression, which occurs upon a lower sequence complementarity between the interacting molecules. Therefore, they are integrated into a different effector complex: the microribonucleinprotein, or miRNP. The miRNPs might repress translation at a step after translational initiation, in a manner that does not observably change the density of ribosomes on the transcript due to just slowing or stalling them. An alternative possibility is that translation continues at the same rate but is

non-productive because the newly synthesized polypeptide is degraded immediately. In both cases, mRNA levels are not affected, indicating that silencing occurs around translation.

An almost perfect base pairing to the target mRNA is verified at position 2-7 of the smallRNA, which seems to be important for target recognition. An important enzyme which helps forming all effector complexes is Argonaute, a large protein containing 2 domains: PAZ and PIWI. Whereas the PAZ domain, located at the center of Argonaute, interacts with the 3' overhang of the dsRNA, the PIWI domain on the c-terminus shows homology to RNase H and is thus considered to cleave the target mRNA. Additional proteins involved in silencing are less conserved between organisms, there are some dsRBDs in *Drosophila*, which appear to be involved in strand selection and RISC assembly or as co-factors to the preprocessing protein Drosha. Putative RNA helicases also function in the assembly of effector complexes and RNA-dependent RNA polymerases (RdRPs) join the smallRNA together with its target.

2.3 Focus on miRNAs - Biogenesis and function in animals and plants

A miRNA is defined as a single stranded RNA fragment of an average length of 21-22 nt, originating from endogenous hairpin-like transcripts that are processed to their final form by RNase-III-type enzymes.

MiRNA is the most investigated smallRNA species, at the moment the miRNAregistry [Ambros et al., 2003; Griffiths-Jones et al., 2004] contains 2909 entries, which means precursors and their appendant miRNAs. 31 Organisms are listed, in addition to the leadoff animals *C.elegans* and *D.melanogaster* and their related species vertebrates, especially mammalia have recently been added to the database. The higher the evolutionary level of development of an organism the more intron is contained in the pre-mRNA, suggesting that mammals show the greatest variety of regulatory smallRNAs. [Mattick JS., 2004] Furthermore, the miRNABase stores miRNAs of 7 different plants (most investigated are *A.thaliana* and *O.sativa*), and even virus miRNAs have been released [summary of release 7.0 of the miRNABase: http://microrna.sanger.ac.uk/sequences/help/summary.shtml].

2.3.1 MicroRNA biogenesis in animals and plants

The biogeneses of miRNA in animals and plants appear to be rather different, although both include homologues of the RNase-III-type enzyme Dicer. Like other RNA polymerase II transcripts, miRNA genes are capped, spliced and polyadenylated. They appear in a long primary transcript which contains the predicted miRNA precursor as part of an RNA hairpin structure. This early stage of miRNA is called the pri-miRNA [Zeng et al., 2002].

In animal cells, the pri-miRNA is excised by the nuclear RNase-III-type enzyme Drosha [Lee et al., 2005]. Drosha forms a large complex, known as the microprocessor complex, together with its essential cofactor Pasha, which is known to contain 2 dsRNA-binding domains [Denli et al. 2004]. Full particulars are not explored about recognition and cleavage of Drosha, but it has been observed to be selective for hairpins bearing a large (>10 nt) terminal loop located on a stem with partially imperfect complementarity. From the stem-loop junction, it cleaves approximately two helical turns (~22 nt) into the stem releasing a ~70 nt long hairpin precursor. Variations in the stem structure, such as bulges, and sequence around this region contribute to the fine-tuning of the actual cleavage site. Drosha has emerged as a key determinant of which part of the pri-miRNA will become the mature miRNA, because by clipping the primary transcript it generates one end of the final miRNA. The resulting pre-miRNA is exported into the cytoplasm by a nuclear export factor, called Exportin-5, which also functions as a quality control for hairpin precursors. The second step of processing occurs in the cytoplasm where Dicer trims the hairpin to a ~22 nt long RNA duplex [Lee et al., 2002]. It recognizes the characteristic 3' 2 nt overhang generated by Drosha cleavage and simply cuts off the loop, leaving an additional 3' overhang on the other side. Due to the immediately following steps, the released duplex is very short-living in vivo.

Homologues of Drosha were not found outside the animal kingdom, suggesting that the described processing does not apply for plant cells. Maturation of miRNA in plants has been primarily examined in *A.thaliana*, which shows 4 Dicer-like enzymes (DCL1-DCL4). DCL1 provides the Drosha functionality, executing the first cleavage step to yield the miRNA precursor. Experiments have shown that levels in miRNA expression are only reduced in DCL1 mutants, which indicates that this homologue of Dicer is the key protein in plant miRNA biogenesis. Therefore, the second cut is assumed to be also mediated by DCL1. In contrast to animal cells, this takes still place in the nucleus followed by an additional methylation step, which requires another protein, termed HEN1. Short RNA, present in the nucleus, may be accidentally utilized as primer leading to amplification of unwanted genes. Methylation of the last nucleotide of the miRNA prevents this by protecting the plant miRNA from polymerases adding additional nucleotides to the 3' end. The final duplex is then exported into the cytoplasm by HASTY, the plant orthologue of Exportin-5 [Chen, 2005].

Following cleavage and nucleocytoplasmic export, the miRNA pathway of animals and plants appears to be biochemically indistinguishable. One strand of the miRNA:miRNA* duplex is incorporated into RISC or a similar effector complex while the other one is degraded. Thermodynamic differences in the base-pairing stabilities of the 5' ends determine which strand is selected. Essentially this means the strand with the weaker hydrogen bonding is used in silencing [Tomari et al., 2004]. However, this selection step is not sufficiently enlighted yet and awaits further investigation.

2.3.2 MicroRNA function in animals and plants

Plant and animal miRNAs are quite different regarding their complementarity to the mRNA target. While miRNAs in animals usually mediate translational repression due to an imperfect base pairing, plant miRNAs show a high complementarity and therefore initiate target cleavage.

The most interesting question to arise from the discovery of hundreds of different miRNAs is, what are all these smallRNAs doing? For lin-4, let-7 and several other miRNAs their function and regulated targets were discovered based on in vivo experimentation even before their status as non-coding RNA genes. However, computational approaches have been developed to find the target genes of the miRNAs (see Section 3 for details). Due to the near-perfect complementarity of the miRNA this was especially successful in plants. The majority of miRNAs here have a remarkable addiction for targeting transcription gene factor families, particularly those involved in developmental patterning or cell differentiation, by mediating the degradation of key regulatory gene transcripts in specific daughter cell lineages [Review Bartel DP., 2004 and references therein]. For example, during differentiation, certain genes specifying a less differentiated state might need to be turned off. This can be achieved by repressing transcription or, to more quickly stop expression of such a gene, the differentiating cell can deploy a miRNA that cleaves that mRNA. This may explain why plant miRNAs are enriched in plant organs because most cells of plant organs are typically differentiated.

As in plants, the predicted targets in animals are significantly high in genes with known or suspected roles in transcriptional regulation, suggesting that the described model could also be operating in animal tissues. A popular example is lin-4 and let-7, the first miRNAs discovered in *C.elegans* which act as posttranscriptional repressors of their target genes involved in regulating developmental timing. Nonetheless, this enrichment for transcriptional regulators is much smaller in mammals, and functions of target genes represent a surprisingly broad diversity. Mir-181 in mammals, for instance, is involved in the control of hematopoiesis. Recent researches show the regulatory relevance of miRNAs in stem cell division, cancer and other human diseases [Ambros, 2004].

Like in the latter example, one single target can cooperatively be controlled by several different miRNAs and, the other way round, a single miRNA species may target different mRNAs.

3 Understanding complexity - Bioinformatical approaches

Computational and system biologists will have to deal with the prospect that a substantial fraction of all animal mRNAs could have their precise level of expression defined by miRNA regulation. To the extend, that the miRNAs direct translational repression rather than mRNA cleavage, this regulation will be invisible to one of the most powerful tools of biologists, microarray analysis of mRNA levels. In this field, bioinformatics is becoming downright essential. In this Section I will present bioinformatical approaches developed over the last years, which are deployed to reconstruct the miRNA pathway.

3.1 Three steps to reconstruct the miRNA pathway

In order to fully understand the miRNA pathway, we have to consider three main steps: First, on the basis of a genomic sequence, miRNA precursor sequences, which are generated from the primary transcript in vivo, can be predicted from secondary structure characteristics. Second, we have to find out the rules defining the following cleavage steps mediated by cellular RNases to yield potential mature miRNAs. Third, with these sequences it will be necessary to find the genomic targets regulated by miRNA mediated silencing mechanisms. If we were able to reconstruct all these three steps completely, it would offer the possibility to find out all about the amount of miRNAs and assign functions to each of them if the targeting gene is annotated. Conversely, it might open a new way for gene annotation by associating the repressing function of known miRNA with a new, yet un-annotated gene or gene cluster.

3.2 Precursor Prediction

The first step in reproducing the miRNA biogenesis is the prediction of possible precursors in an organism's genome. There are several algorithms and a series of programs developed to answer the question of how many miRNA genes are encoded in animals and plants (Table 2). Through simple sequence homology search using BLASTN, homologues or orthologues of miRNAs, isolated by cloning, have been identified [Pasquinelli et al., 2000].

The most noticeable property of pre-miRNA is its secondary structure: a hairpin, defined as a single stranded RNA folding back on itself. Alternatively to homology search, one can therefore use RNA folding algorithms such as Mfold [Zuker, 2003], RNAfold [Vienna RNA package, Schuster et al., 1994] or RNALfold [Hofacker et al., 2004] to predict hairpin-like structures. The idea behind this is to minimize the energy of the given RNA sequence as it is thought that the one with the lowest free energy is the most stable one. For example the MiRseeker procedure examines the folding of RNA sequences conserved between two *D.melanogaster* species using Mfold [Lai et al., 2003]. This algorithm uses predictions of stem-loop structure formation as key criteria. It also takes into account the nucleotide divergence of miRNA candidates, as the authors detected less selective pressure in the loop sequence of orthologous precursors. MiRscan is another computational approach that has been applied to the genomes of *C.elegans* and humans [Lim et al., 2003]. It is similar to MiRseeker, but uses RNAfold as secondary structure prediction method. Recently, a particular phylogenetic approach has been published by Berezikov et al. [2005], which has been used to identify novel human miRNA precursors. This one, in contrast to former algorithms, is based on the noticeable conservation of precursor sequences and secondary structures among species.

Program	URL	Species	References
MiRseeker	http://genes.mit.edu/mirscan	*D.me*	Lai et al., 2003
MiRscan		*C.el/H.sa*	Lim et al., 2003
Phylogenetic shadowing	-	*H.sa*	Berezikov et al., 2005

Table 2: miRNA precursor prediction programs

3.3 Target Prediction

The major challenge in determining miRNA functions is to identify their regulatory targets. It is much easier to find them in plants due to a high complementarity of the miRNA/mRNA duplex. In a systematic search for the targets of 13 *Arabidopsis* miRNA families in 2002, for instance, 49 unique targets were found simply by looking for transcripts with near-perfect complementarity to the miRNA [Matthew, Rhoades et al. 2002]. Confidence in many of these predictions was backed by the observation that the complementarity is conserved among flowering plants, and the majority of the 49 targets has since been confirmed experimentally.

Hundreds of animal miRNAs have also been identified, but only a few of their targets are known. Prediction of mRNA/miRNA duplexes is especially challenging in animals, since the interaction, in contrast to plant cells, usually occurs via incomplete base pairing. The rules that govern such interactions are incompletely defined, but as mentioned before, the near-perfect complementarity at the 5' end of the leading miRNA is observed throughout the animal kingdom and therefore holds as a proper starting point for computational approaches. Especially residues 2-8 of invertebrate miRNAs pair perfectly to 3' UTR of the targeting mRNA and are perfectly conserved in othologous transcripts of other metazoan species. A contiguous helix of at least 7 basepairs is nearly always seen in this region. Based on this characteristic trait, a number of algorithms has been developed to predict animal miRNA targets. The method of Stark et al. for the prediction of *Drosophila* target genes provides a list of candidate targets which has to be combined with additional biological criteria, including functional relationships shared among them [see Review Lewis et al. 2003 and Refs therein] This is not an accurate solution in cases where predicted targets do not show clear functional relatedness. Subsequent approaches additionally used RNA folding algorithms (Mfold [Zuker et al. 2003], RNAfold [Hofacker et al. 2003]) to estimate free energy to each miRNA/target site interaction to make sure the predicted duplex has a relatively high thermodynamic stability. Furthermore, programs like 'TargetScan' [Lewis et al. 2003] and 'findMiRNA' [Sundaresan et al. 2005] in addition require candidates to have multiple binding sites in the target 3' UTR and evolutionary conservation of the target between species. 'RNAhybrid', a target prediction tool developed at Bielefeld University [Rehmsmeier et al. 2004], includes also miRNA specific extreme value distribution parameters. Most experimentally detected targets have been confirmed with these approaches and many new ones have been predicted in *Drosophila* and mammals. However, this step, authentic prediction of miRNA targets, appears to be a difficult mission in the future, particularly in animals.

Program	URL	Species	References
TargetScan/ TargetScanS	http://genes.mit.edu/targetscan	Vertebrates	Lewis et al., 2003 Pfeffer et al., 2004
miRanda	http://www.microrna.org	*D.me/H.sa*	Kiriakidou et al., 2004
miRNA-target prediction	http://www.russel.embl.de/miRNAs	*D.me*	Stark et al., 2003 John et al.,2004
RNAhybrid	http://bibiserv.techfak.uni-bielefeld.de	*D.me*	Rehmsmeier et al., 2004

Table 3: miRNA target prediction programs

3.4 The missing link: Prediction of mature miRNA

The chain of reconstructing the miRNA pathway is still incomplete since target prediction needs a mature miRNA to find regulated transcripts and precursor prediction only provides a ~70 nt long immature pre-miRNA. To link these two procedures, a method to predict ~22 nt mature miRNA from its precursor should be developed. Surprisingly, less efforts have been undertaken to close this intermediate step, maybe because this certainly is a serious enterprise. No program executing this step was published as of April 2005, at the time of beginning my Bachelor Thesis, so I started to develop different ideas to overcome this challenge.
For the sake of completeness I want to mention that another computational approach dealing with the same problem (immediately following a precursor prediction), especially in animals, was published in June by Wang et al. [2005]. The authors developed an algorithm, called miRAlign, to find new miRNAs using both sequence information and structural characteristics of already known miRNAs. Given a ~70 nt candidate precursor sequence, miRAlign scores this possible precursor and predicts its mature form in 5 steps: (1) Secondary structure of both strands of the candidate precursor stem are predicted by RNAfold, all sequences with a MFE< -20 kcal/mol are aligned in a pairwise fashion to all known ~22 nt miRNAs in their training set by CLUSTALW. Only those pairs exceeding a defined minimum threshold are kept for further analysis. For each of those potential homolog pairs, the ~22 nt subsequence on the candidate that aligns to the known miRNA is regarded as the potential mature miRNA. The last predefinition is remarkable because it implies that the position of the mature miRNA in the comparing precursor from the training set has to be already defined. (3) The position of the mature miRNA on the candidate precursor is than predicted by considering 3 conditions: (a) the ~22 nt potential miRNA should not locate on the terminal loop of the hairpin, (b) it should locate on the same stem side of the hairpin as its known homologues and should (c) not differ too much in position relative to them. The resulting position difference between known and candidate mature miRNA is than calculated, all candidates with a difference lower than a defined cut-offs are selected for the next step: (4) RNA secondary structure alignment. RNAforester is deployed to calculate a normalized similarity score, which ranges from 0 to 100 after normalization by the self-alignment score of the known precursor. (5) Finally, a total similarity score is assigned to the candidate sequence as the maximum of the similarity scores of all homologues to the candidate precursor. The higher this total score the more likely is the authenticity of the candidate. While the original intention of this algorithm is the precursor prediction, it includes approaches of mature miRNA prediction as well. However, this algorithm supposes the knowledge of the position of mature miRNA in the related precursors. This can not be assumed for all newly predicted precursors. We will have a closer look on possibly better solutions while trying to develop an approach that deals with the prediction of mature miRNA from given precursors exclusively.

4 Trying to bridge the gap – ‘In-Silico-Dicer’

4.1 The Concept: Two general approaches

The central question of this work is: Can we predict the mature miRNA from a known precursor? I suggested ‘Yes, we can’, thus, the main challenge is to develop reasonable algorithms to prove this thesis. More formal, I define the task as follows: Given a validated miRNA precursor sequence *pcS*: Where is the relative position of the mature miRNA, *mmP*, relative to *pcS*?

Imprinciple, I took two main ideas into consideration: Intrinsic and Extrinsic prediction of mature miRNA. Extrinsic prediction seizes a method, intentionally published to predict miRNA precursor sequences on the basis of their conservation between species [Berezikov, 2005]. For this approach I needed to obtain external information, more precisely related miRNAs, to compare to. Therefore, this approach is termed ‘Extrinsic’. Intrinsic uses only internal information the given precursor sequence provides. The idea was to simulate the cleavage steps mediated by the RNase-III-type enzymes Drosha and Dicer, as mentioned in Section 2. Due to the fact that these enzymes have nothing else than the precursor’s sequence and secondary structure to cut properly, I suggested that prediction might be possible with this information only. I liked the vision to reconstruct the processing enzymes recognition and cleavage and therefore named the tool to develop ‘In-Silico-Dicer’.

4.1.1 The intrinsic approach and its extension

The intrinsic approach to predict mature miRNA from its precursor does not require external information. As input it uses only the nucleotide sequence of the precursor. Based on this sequence, a secondary structure prediction method, e.g. RNAfold [Hofacker et al, 2004] or Mfold [Zuker et al., 2003], is executed to yield the folded miRNA precursor, which should form a hairpin structure. Now we assume every substring of the precursor to be a mature miRNA candidate, the length of the subsequence should therefore range between 19-24 nt (for predicting animal miRNAs a length of ~21 nt is suggested). Several characteristics are known about mature miRNAs:

(1) they do not locate on the terminal loop, but at one arm of the hairpin stem,
(2) mostly, they do not locate in a region with too many bulges or internal loops and
(3) they are often observed to start with the nucleotide ‘U’.

Although, these conditions are not very precise, it might be possible to define a formula that, on the basis of sequence and secondary structure of every candidate, can be applied to assign each window of the specified length a probability score to be a mature miRNA. The candidate with the highest probability is proposed as mature miRNA.

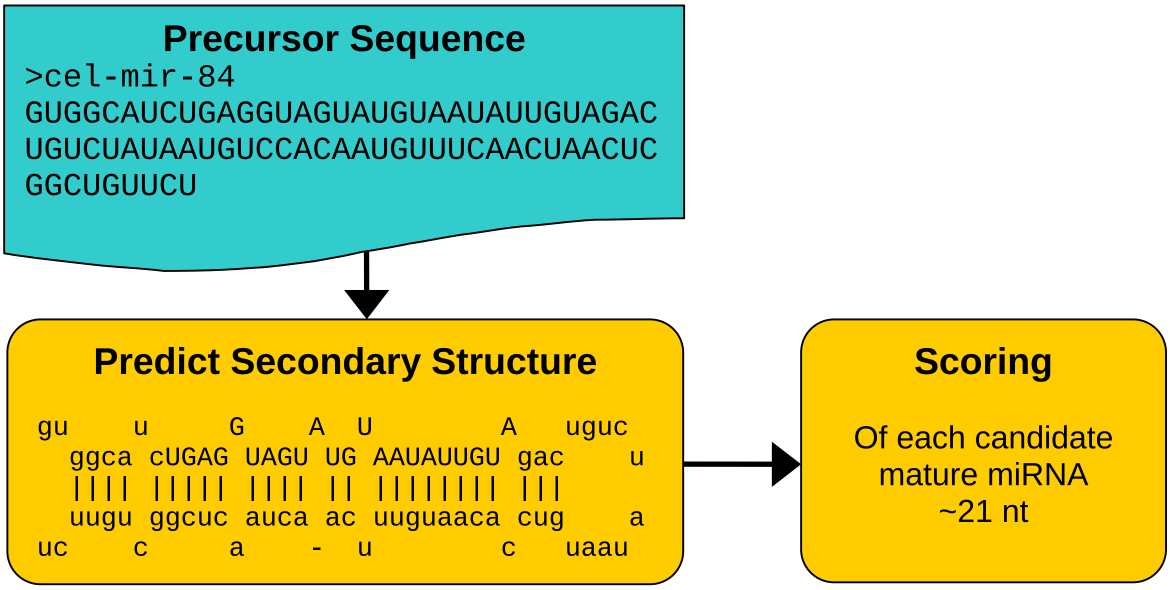

Precursor Sequence
>cel-mir-84
GUGGCAUCUGAGGUAGUAUGUAAUAUUGUAGAC
UGUCUAUAAUGUCCACAAUGUUUCAACUAACUC
GGCUGUUCU
Predict Secondary Structure
gu   u   G   A U   A   uguc
  ggca cUGAG UAGU UG AAUAUUGU gac   u
  |||| ||||| |||| || |||||||| |||
  uugu ggcuc auca ac uuguaaca cug   a
uc   c   a   - u   c   uaau
Scoring
Of each candidate
mature miRNA
~21 nt

Figure 1: Procedure of the Intrinsic mature miRNA prediction method. The input precursor sequence (blue) is processed in 2 action steps (orange)

The intrinsic prediction works also for several precursors at the same time. When applying this method to some closely related precursors I observed a quite similar position of predicted mature miRNA (data not shown). Due to this observation I thought of a possible improvement of the algorithm that may fine-tune the prediction by aligning the input precursor to closely related miRNA precursors (e.g. from the same miRNA family). This set of closely related comparison precursors can be derived from a database of miRNA precursors, e.g. the miRNABase, by searching similar precursors by sequence. For this task, the Smith-Waterman Algorithm [Smith and Waterman, 1981] for pairwise local alignment is deployed. The algorithm generates a similarity matrix and returns – in this application – the highest value from this matrix, which represents the similarity score. All precursors which, aligned to the input precursor, return a similarity score higher than a user-defined cut-off are considered to be related precursors to the given input sequence (see 'Implementation' for details). Thereafter, a multiple alignment is performed on the set of related precursor sequences and than matched together with the scores calculated for each of the precursor sequences (see 'Implementation' for details on this matching step). As a result, we have a multiple alignment with a score assigned to each single character, which represents the possible mature miRNA sequence surrounding this nucleotide. In the final step, a column-wise average score is calculated and, as in the simple intrinsic approach, the position identified by the highest score is considered as position for the mature miRNA. One has to pay attention to the fact that this predicted position is relative to the alignment, not to the precursor sequences itself. As a consequence all gaps of the analyzed precursor have to be removed to yield the absolute miRNA position.

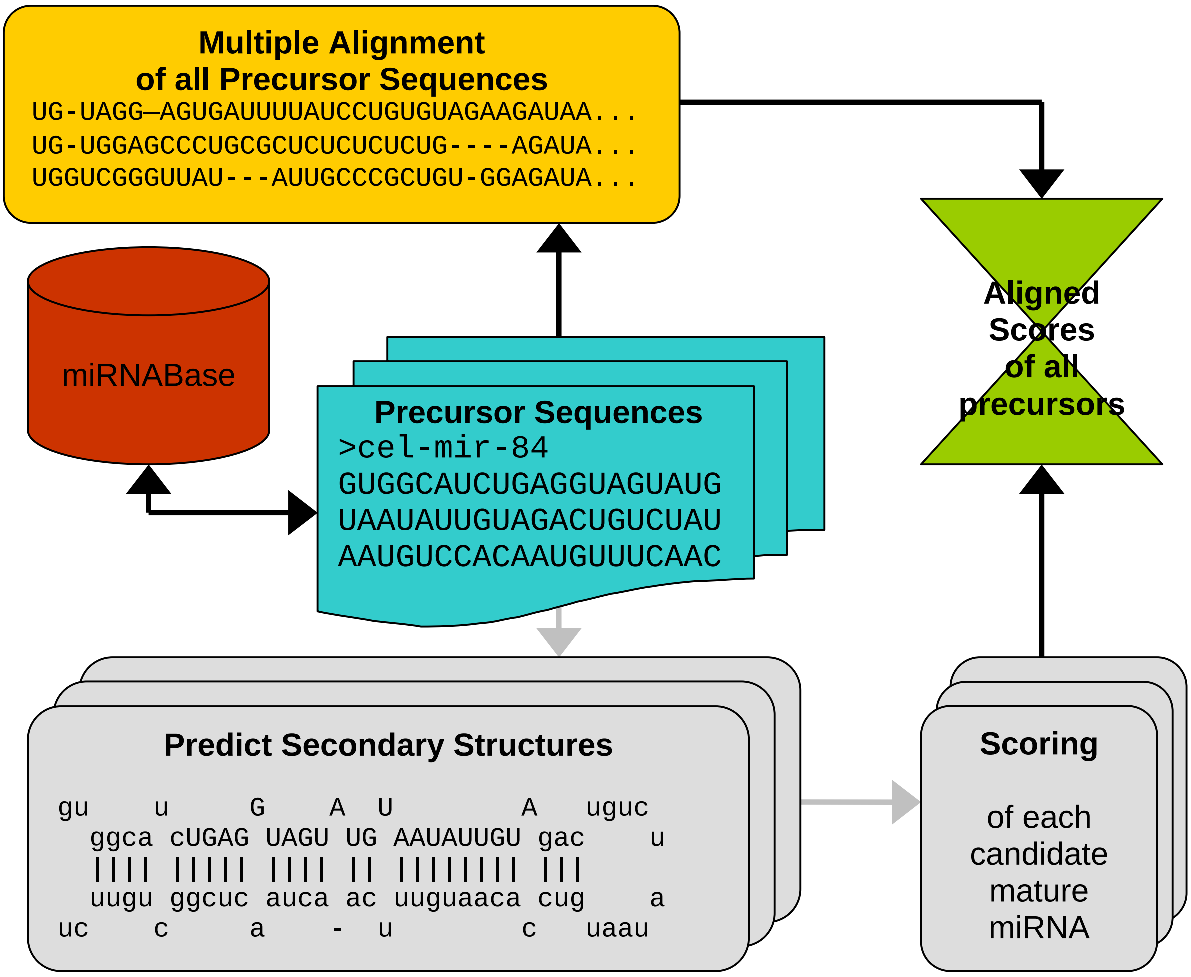

Multiple Alignment
of all Precursor Sequences
UG-UAGG–AGUGAUUUUAUCCUGUGUAGAAGAUAA...
UG-UGGAGCCCUGCGCUCUCUCUCUG----AGAUA...
UGGUCGGGUUAU---AUUGCCCGCUGU-GGAGAUA...
miRNABase
Precursor Sequences
>cel-mir-84
GUGGCAUCUGAGGUAGUAG
UAAUAUUGUAGACUGUCUAU
AAUGUCCACAAUGUUUCAAC
Aligned Scores of all precursors
Predict Secondary Structures
gu u G A U A uguc
ggca cUGAG UAGU UG AAUAUUGU gac u
|||| ||||| |||| || |||||||| |||
uugu ggcuc auca ac uuguaaca cug a
uc c a - u c uaau
Scoring
of each candidate mature miRNA

Figure 2: Extended Intrinsic Prediction of mature miRNA. Grey areas are similar to the simple intrinsic prediction, except, that they are multiple executed. The input may be one or more precursor sequences. (blue) With this input related precursor sequences are searched within a miRNA-DB (red). The returned set of sequences is multiple aligned (orange) and thereafter matched together with the scores of each precursor.

4.1.2 Extrinsic Approach

The general idea behind this approach is the assumption that miRNAs appear to have important, maybe even essential, regulatory roles throughout all explored organisms. Therefore, miRNA genes should be conserved among genomes, especially among closely related species. Berezikov et al. [2005] already applied this idea to the prediction of precursor sequences, many of which were confirmed experimentally later on. Precursor sequences, however, may vary in their nucleotide sequence as long as they are able to form a stable hairpin secondary structure. In contrast, the sequences of the mature miRNA as the actual regulatory core sequences of the pre-miRNA must not differ too much in its primary sequence to ensure complementarity to the target gene. For this reason, one can assume a noticeably higher conservation of the mature sequence among species relatively to its known precursor. So, for this prediction approach, a set of related precursor sequences is unconditionally needed because all following actions are based on this set. It is searched within the miRNABase as described in Section 4.1.1. Just as well, a multiple alignment is created on the derived most similar sequences. Now, based on this alignment, a conservation value is calculated column-wise (see ‘Implementation’ for details) and thereafter these conservation values are summed up for all subsequences of length 21 (nt). The higher the conservation sum, the higher the probability for the mature miRNA to locate in this window. Again, this position has to be seen in relation to the alignment.

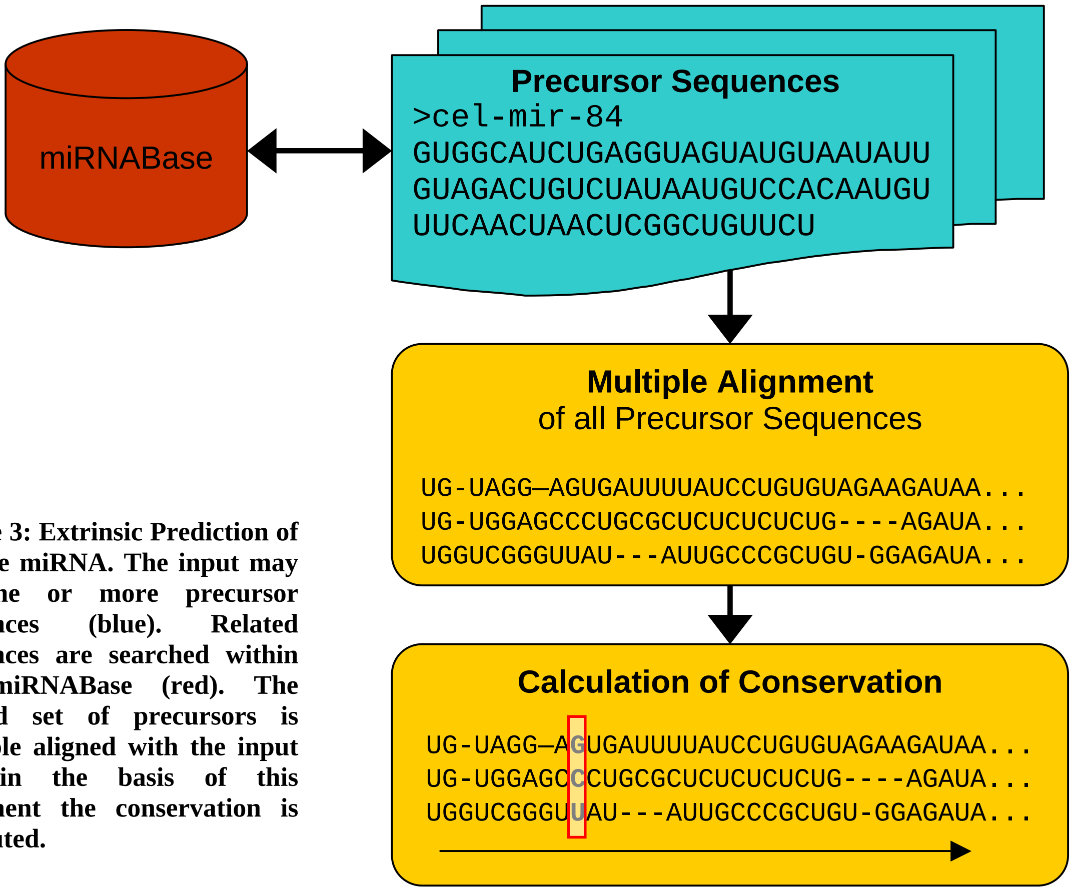

miRNABase
Precursor Sequences
>cel-mir-84
GUGGCAUCUGAGGUAGUAUGUAAUAUU
GUAGACUGUCUAUAAUGUCCACAAUGU
UUCAACUAACUCGGCUGUUCU
Multiple Alignment
of all Precursor Sequences
UG-UAGG–AGUGAUUUUAUCCUGUGUAGAAGAUAA...
UG-UGGAGCCCUGCGCUCUCUCUCUG----AGAUA...
UGGUCGGGUUAU---AUUGCCCGCUGU-GGAGAUA...
Calculation of Conservation
UG-UAGG–AGUGAUUUUAUCCUGUGUAGAAGAUAA...
UG-UGGAGCCCUGCGCUCUCUCUCUG----AGAUA...
UGGUCGGGUUAU---AUUGCCCGCUGU-GGAGAUA...

Figure 3: Extrinsic Prediction of mature miRNA. The input may be one or more precursor sequences (blue). Related sequences are searched within the miRNABase (red). The yielded set of precursors is multiple aligned with the input and in the basis of this alignment the conservation is computed.

4.2 Design and Implementation

The Tool ‘In-Silico-Dicer’ has been developed to predict mature miRNA from a given miRNA precursor. For every input pre-miRNA it makes a suggestion for the most probable position of the mature sequence within this precursor. Additionally, it provides the professional user with several graphical representations of the output, which may be helpful to evaluate the predicted position in a biological context.

‘In-Silico-Dicer’ has been implemented in the JAVA programming language (JDK 1.5.0.), and consists of several packages containing too many classes to describe them all in detail. For this reason, I refrain from giving a complete UML description and will primarily describe the most important classes and interactions among them in the following Section.

4.2.1 Overview

A graphical overview of the structure and design of ‘In-Silico-Dicer’ is shown in Figure 4. The following description will refer to this illustration.

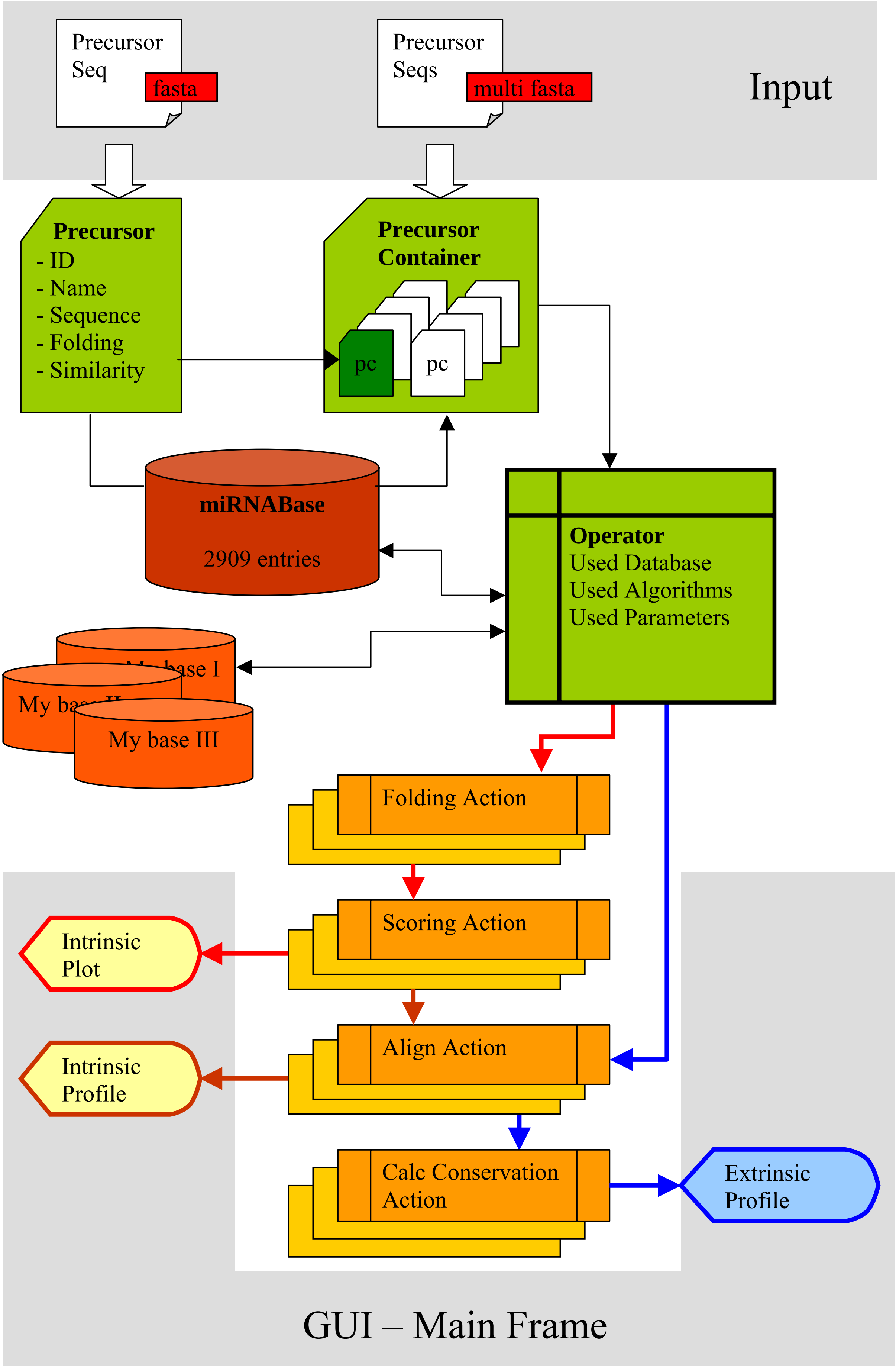


Figure 4: Overview over the program structure of 'In-Silico-Dicer'. See Section 4.2. for details.

Input

As the main goal is to predict the mature miRNA from its precursor, valid inputs to the program are a FASTA file (see Appendix) of the given precursor as well as multi-FASTA files to process more than one precursor. One can run predictions on every input precursor or just use the set for quick comparison with the focused precursor sequence.

The Central Objects

All classes which hold the main information, needed to process the input precursor, are collected in the Operating package (objects shaded in green). First I will just concentrate on the Classes Precursor and PrecursorContainer. From the input precursor file a Precursor-Object is created which holds the following properties:

- Name - the whole description part of the FASTA file
- ID - refers to the first part of this description which is an unique key defined in the miRNAregistry annotation system [Griffith-Jones, 2004]
- Sequence – rest of the FASTA file
- Folding - represents the secondary structure of the precursor, initially set to null, not assigned until the precursor traverses further processing
- Similarity – refers to a pairwise alignment with another precursor, initially set to the maximum value (derived from score setting of the used alignment algorithm) because aligned to itself the precursor should have full similarity

If the input consists of more than one precursor sequence, a collection of Precursor objects is created, the PrecursorContainer. The PrecursorContainer object is an extension of the Vector Class that holds only Precursor objects. This class implements some useful methods to search for a specific Precursor object by:

- ID of the Precursor

  ```
  public Precursor
  getPrecursorByID(String theID);
  ```

  Fast search method due to the use of Hashtable.

- Name of the Precursor

  ```
  public PrecursorContainer
  getPrecursorByName(String theName);
  ```

  If it is already known to which family the precursor belongs, it may be useful to search related family members by name

- Sequence of the Precursor

  ```
  public PrecursorContainer
  getPrecursorBySequence(String theSequence);
  ```

  This search method is mainly used to find related Precursors by similar sequences. It implements the Smith-Waterman Algorithm [Smith and Waterman, 1981] and returns all precursors with a similarity to the input higher than a defined cut-off value. The returned Precursors are subsequently sorted by similarity.

The miRNAregistry – the central database

The miRNAregistry is actually a permanently stored object of the type PrecursorContainer. It has been initially created from the multi-FASTA file of all precursors listed in the miRNAregistry. From version 7.0 of the database, 2909 precursors were available [Griffith-Jones, 2004]. Once read into a PrecursorContainer object, one can apply the search methods mentioned before to yield comparable precursor sequences from the miRNAregistry. Additionally, the program offers the possibility to create own databases from multi-Fasta files similar to the 'hairpin.fa' from the miRNAregistry, e.g. one might create a database which contains only plant precursors because comparisons between plant and animal precursors are not reasonable to predict new mature miRNA due to the fact that maturation differs between the kingdoms. During the use of the application, the user can switch between databases. These self-defined databases persist, so they can be used during another program run again.

The Action Steps

The core of the program to predict the mature miRNA position is realized by a chain of actions (actions shaded in orange) on the Precursor or the PrecursorContainer-Object. The Action-Chains of the general prediction approaches, explained in Section 4.1., overlap at some point, so there is no need to implement a similar step twice. This concept makes the program more flexible, if one develops another general approach, the order might be simply rearranged or an action can be added to the chain. Action-Chains for the proposed approaches are as follows:

- Intrinsic Prediction (simple)
  [Folding] → [Scoring]
- Intrinsic Prediction (extended)
  [Folding] → [Scoring]→ [Multiple Alignment]
- Extrinsic Prediction
  [Multiple Alignment] → [Calculation of Conservation]

In the end of each action chain for every Precursor/PrecursorContainer several score/conservation values are returned. These can be output to the user as raw data or in the form of a plot (see 'Graphical User Interface').
To achieve a maximum of flexibility of the program, the algorithms, which are applied in each action step, should be changeable. While there are already several existing standard algorithms for secondary structure prediction (Folding) and multiple alignment of sequences, there are rules neither for calculating the conservation between sequences nor for scoring precursor sub-sequences regarding their probability to be the mature miRNA. For this reason, all action steps in the action package have been implemented through an interface which defines parameters and returns values for each of the actions. For every action step interface, I have implemented exactly one concrete class (see 'Used Algorithms' for details). This concept makes it easy for future developers. They only have to include alternative algorithms in order to improve the mature miRNA prediction; they only have to develop a class which implements the respective Interface. All available algorithms choices are displayed automatically and the user may choose which one to apply for his purpose.

The Operator Concept

In order to maintain the control over all this flexibility, a central object that holds all settings is necessary. The application therefore deploys the Operator, which is also part of the Operating package (shaded in green). The Operator holds all variables and provides Getter and Setter Methods to change and retrieve them from other classes. Most important variables and initial values are listed in Table 4.

Variable	Description	Initial Value
Precursor Database	current precursor DB to search in for comparable precursors	miRNAregistry
FoldingAction	the current algorithm to predict secondary structure	RNAstructure
ScoreAction	the current algorithm set to score the precursor	Score_ Gaasterland
MultipleAlignAction	the current multiple alignment method	ClustalW
CalcConservationAction	the current algorithm to calculate the conservation between several precursors	CalcCons_ Entropy
mmRNA length	length of the mature miRNA to predict	21
MATCHscore, INDELscore, MISMATCHscore	score settings for the Smith-Waterman algorithm to search for similar precursors by sequence	2 -1 -2

Table 4: Some Variables and its initial values holded by the Operator

Additionally, the Operator provides the methods to actually predict the position of mature miRNA within its given precursor. These methods need the scores derived from the action steps as parameters:

```
     public int predictMatureMiRNAin(Vector dataVec);
     public int predictMatureMiRNAex(Vector dataVec);
```

These methods however, return the position of the mature miRNA relative to alignment. To get the absolute position of the mature miRNA, one further step is required:

```
     public int[] findStartInSinglePC
                             (int startINalign, Vector align);
```

The Operator is also in charge of creating new databases from FASTA files and reading the selected ones to provide their content to classes working on the current database. In this application, only one project, which means one miRNA prediction, at a time is possible. To ensure that there is always only one definite value selected for every setting, the Operator is implemented as a Singleton.

4.2.2 Used Action Algorithms

All Actions in the Action Package operating on a Precursor or PrecursorContainer are defined through interfaces which await concrete implementation. I implemented one single class for every interface, which at the moment leaves no choice of algorithms to the user. It will be challenging to develop alternative classes and methods for every each interface within a biological context to improve the prediction of mature miRNA.

IFolding Action

For secondary structure prediction of a sequence, I deployed RNAstructure, which is a Windows implementation of the Zuker Algorithm [Zuker et al., 2004] based on free energy minimization. ‘In-Silico-Dicer’ calls the program by creating a new process.

```
      Process p = Runtime.getRuntime().exec(RNAfold);
```

This algorithm returns a String representing the secondary structure of the sequence. The String is given in ‘Vienna’ notation, where brackets represent paired nucleotides and dots represent unpaired ones e.g.,

```
      Sequence: CUACUCUGUCAUGUAUAACUAAAUUUGAUUGACACUUCUGUGAGUA
      Folding:  .(((((((((((..................)))))).......)))))
```

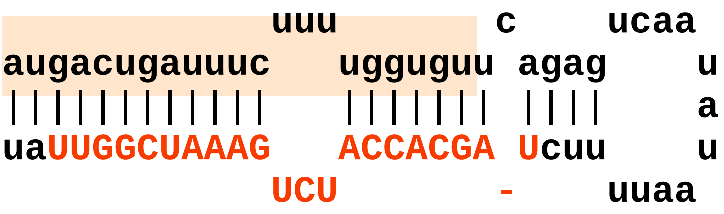
IScoring Action

To score a precursor in the context of this application means to assign a score to every potential mature miRNA sequence within the precursor sequence which represents the probability of this sequence to be the mature miRNA. Basically, a potential mature sequence can be every valid subsequence of the precursor. It can be excluded that the miRNA locates on the stem-loop, but for the following algorithm all possibilities are considered. The IScoringAction Interface is designed to reproduce recognition and cleavage on the precursor mediated by RNase-III-type enzymes (see Section 2); therefore it is only provided with the same input as these enzymes are: (1) nucleotide sequence of the precursor and (2)its secondary structure. So, to yield a useful result, the Interface requires (a PrecursorContainer holding) Precursors with an assigned Folding property, derived through a previous run of IFoldingAction. So far, no validated reliable method to score precursors has been published perhaps due to insufficient biological knowledge about the complex processing of miRNA. However, for *Arabidopsis thaliana*, an algorithm to predict precursor miRNAs and their mature sequences at once has recently been developed by a New Yorker research group [Wang et al., 2004]. Their prediction covered 63% of known *Arabidopsis* miRNAs and identified 83 new ones, which have either been validated through northern blotting, massively parallel signature sequencing (MPSS) or microarray analysis. Additionally, for the recently predicted miRNAs, putative targets, which were functionally conserved between *A.thaliana* and *O.sativa*, have been identified. Because miRNA biogenesis is a stepwise process in Arabidopsis as well as in animals [Chen, 2005], the scoring method may be applicable also to animals. Wang et al. considered both sequence and secondary structure to develop a formula, which is applied to every potential mature sequence (figure 5).

```
miRNAscore =
  (number of mismatches)
+ (2*number of nucleotides in terminal mismatches)
+ (number of nucleotides in internal bulges/number of internal bulges)
+ (1) if the miRNA sequence does not start with U
```

One has to be attentive because in this formula the score is calculated using penalties for properties which are suggested not to apply to mature miRNA sequences.

Figure 5: The Scoring formula is applied to every subsequence of the precursor


uuu c ucaa
augacugauuuc uggugu agag u
|||||||||||| ||||||| |||| a
uaUUGGCUAAAG ACCACGA Ucuu u
UCU - uuaa

Above all, this Action step may be subject to further development because more details of how the processing enzymes recognize and cleave the miRNA precursor await closer investigation. In collaboration with biologists an optimal scoring, that nearly perfectly reconstructs the selection of mature miRNA sequences mediated by cleavage enzymes, might be found.

IMultipleAlign Action

For aligning multiple precursor sequences I used CLUSTALW [Chenna et al., 2003], a downloadable tool provided by the EBI (European Bioinformatics Institute). Basically CLUSTALW follows these 3 main steps:

1. Determine all pairwise alignments between sequences and the degree of similarity between them.
    - 1.1. Using the pairwise alignment, compute a distance between the sequences. Commonly this distance is calculated by looking at the non-gapped positions and counting the number of mismatches between the two sequences. Then divide this value by the number of non-gapped pairs.
    - 1.2. Transfer the distance values into a matrix representing all possible pairs of sequences.
2. Construct a similarity tree based on the matrix from 1.2 and Neighbour-Joining.
3. Combine the alignments from 1 in the order specified in 2 using the rule “once a gap always a gap”

‘In-Silico-Dicer’ calls the program by creating a new process:

```
Process p = Runtime.getRuntime().exec(clustalw
            IOfiles/SimilarPrecursors.txt
            /output=fasta
            /outorder=input);
```

As specified in the settings, clustalW takes a multi-FASTA file of the precursors to be aligned as input and outputs a multi-FASTA file of the alignment (See ‘Appendix’ for details on output format). To ensure correct processing of further Action Steps, the aligned precursors in the multi-FASTA file should be ordered the same way they are input.

ICalcConservation Action

For calculating the conservation between aligned precursor sequences I developed an algorithm using free entropy H defined by the following formula:

$$H = -\sum_{c=1}^{|\Lambda|} p_c \log_2 p_c$$

where Λ={A, C, G, U}, the Alphabet of nucleotides and c one character from this alphabet. This formula is applied to every column of the alignment obtaining a rate of ‘disarrangement’, which is indirectly proportional to the conservation. The maximal ‘disarrangement’ is achieved when the number of nucleotides is uniformly distributed, e.g. when the alignment column consist of exactly one A, C, G and U each. Due to the known size of the alphabet, one can easily get the maximum value for the entropy of one column by applying 4 to the formula.

$$H_{\max} = -\left[4 * \left(\frac{x}{4}\right) \log_2 \frac{x}{4}\right] = 2$$

Consequently, the Conservation C of an alignment column is calculated by subtracting the entropy from the maximal entropy=

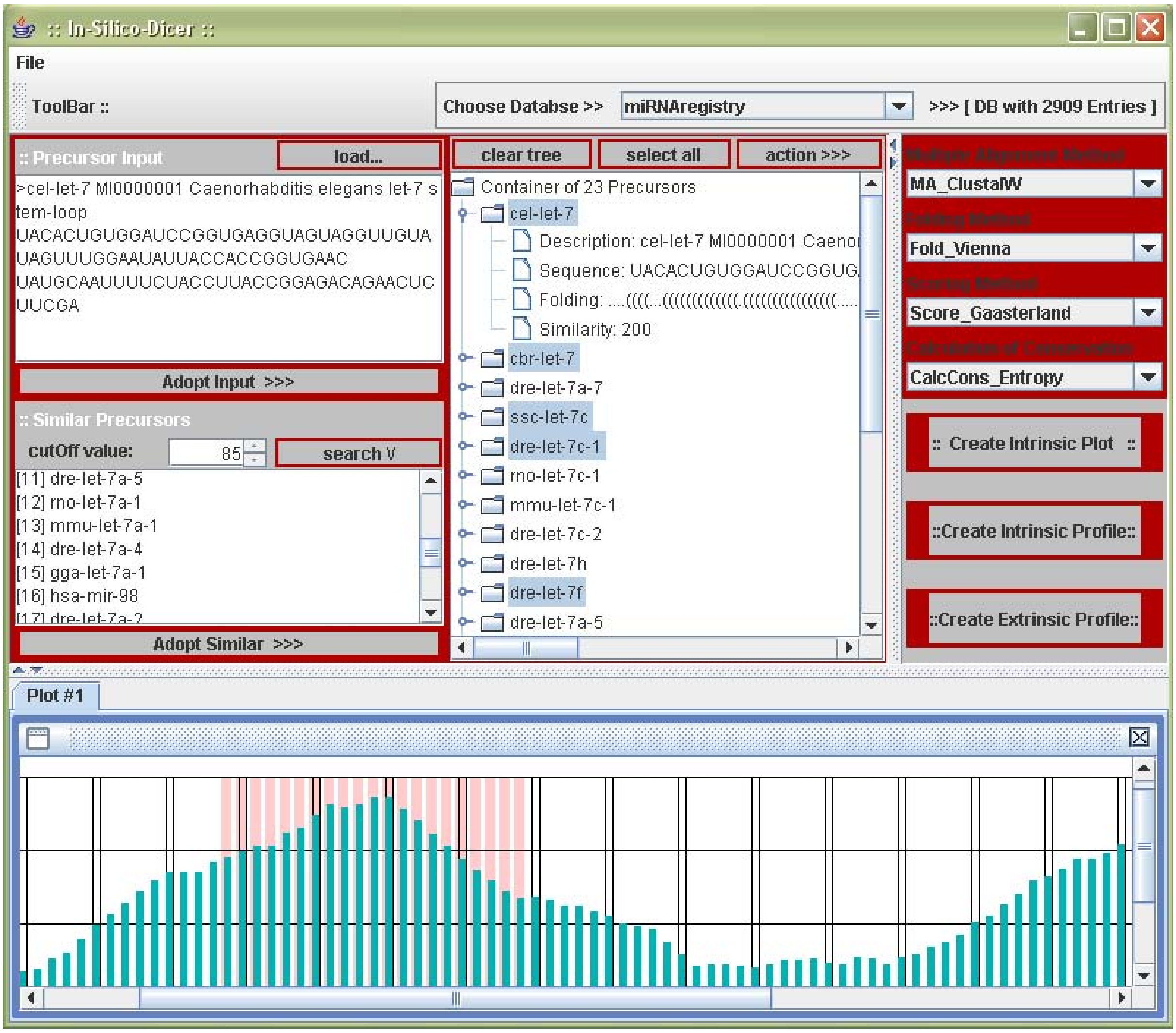
$$C = H_{\max} - H = 2.0 - H$$

Gaps do not contribute to a good conservation in a biological context. However, the given formula will return minimal entropy and a conservation of two given an alignment column consisting only of gaps but one single character. To overcome this problem a gap count is introduced. Besides the number of nucleotides per column, the gaps are counted and afterwards this value is divided by four and assigned to every each nucleotide count. So, in the case where the alignment column consists of hardly any nucleotide, every nucleotide count will be assigned ¼ of the total gap count which leads to a nearly uniform distribution, a high entropy and consequently a very low conservation.

4.2.3 Graphical User Interface

The Graphical User Interface of ‘In-Silico-Dicer’ has been designed to provide the user with full functionality and easy navigation through all settings, program steps and output.

In Figure 6 an overview of the programs interface is shown. I will explain functions of all parts of the interface by an exemplary virtual run through the application. While working with ‘In-Silico-Dicer’ the user might want to use different databases to work on: The repertoire of databases is displayed in a drop down menu located on the toolbar on top of the frame. The number of precursors stored in the current database is shown alongside this drop down menu so that the user can always estimate which data he is working on and how long a search for similar precursors within the database may take. It is possible to add new self-defined databases during the application run by reading in a multi-FASTA file. The user can retrieve this option over the menu item ‘File → create new database…’. The new database is automatically added to the drop down menu as the currently selected database. The input, which may be one or more precursors, is realized by an editable text area in the upper left of the frame. The user has two options to input his data: (1) pasting the precursor in FASTA format or (2) directly read in a FASTA/multi-FASTA file to be displayed in the input area. One can reprocess the input manually after loading a file. By clicking the ‘Adopt Input’-Button, all precursors are adopted into the tree on the right side. In this transferring step, all input sequences are folded. By double clicking on a precursor in the tree, which is represented by its unique ID, all features of the precursor, Description, Sequence, Folding and Similarity, are displayed. As mentioned before, the similarity of the input precursor is initially set to the maximum because the precursor equals itself. Precursors that are similar in sequence and therefore are considered to be related to the input, can be found by browsing the selected database. By means of a continuous cut-off value, the user is able to regulate the distance of relation and accordingly the number of precursors returned as similar precursors. These precursors are listed serially numbered in the lower left text area (which is non-editable), sorted by their similarity to the input. Also these search results can now be adopted into the tree, where they are displayed in a similarity order. With the tree the user keeps an overview over his data, he may look for details on specific precursors, add new ones and afterwards select only the favoured precursors to work with further on. By clicking ‘Action’, an additional panel is displayed which shows all plotting options. On this panel, the desired Action Classes which will be used for the prediction process may be selected through drop down menus. Below these selection boxes, one can find the actual action buttons. Each button will call another prediction approach as described in Section 4.1. The particular action steps are executed on the tree-selected precursors only. The computed data is then graphically displayed to the user through plotting. Extrinsic and intrinsic prediction approaches provide diverse kinds of plots: While conservation (extrinsic) is illustrated by a bar plot, scores (intrinsic) are plotted in lines. Examples for intrinsic and extrinsic plotting are shown in Figures 6-8. The plots are arranged within the frame with the help of a tabbed pane. For every action executed on the selected data, a new tab is created in which the plot is displayed. Additionally, an info frame pops up, providing the professional user with the raw data of the plot.



Choose Databse >> miRNAregistry >>> [DB with 2909 Entries]

>cel-let-7 MI0000001 Caenorhabditis elegans let-7 s
tem-loop
UACACUGUGGAUCCGGUGAGGUAGUAGGUUGUA
UAGUUUGGAAUAUUACCACCGGUGAAC
UAUGCAAUUUUCUACCUUACCGGAGACAGAACUC
UUCGA

cutOff value: 85

[1 1] dre-let-7a-5
[1 2] rno-let-7a-1
[1 3] mmu-let-7a-1
[1 4] dre-let-7a-4
[1 5] gga-let-7a-1
[1 6] hsa-mir-98
[1 7] dre-let-7a-2

Container of 23 Precursors
- cel-let-7
  - Description: cel-let-7 MI0000001 Caenoı
  - Sequence: UACACUGUGGAUCCGGUG.
  - Folding: ....((((...(((((((((((.((((((((((((((.....
  - Similarity: 200
- cbr-let-7
- dre-let-7a-7
- ssc-let-7c
- dre-let-7c-1
- rno-let-7c-1
- mmu-let-7c-1
- dre-let-7c-2
- dre-let-7h
- dre-let-7f
- dre-let-7a-5

Multiple Alignment Method: MA_ClustalW
Folding Method: Fold_Vienna
Scoring Method: Score_Gaasterland
Calculation of Conservation: CalcCons_Entropy

Figure 6: Overview of 'In-Silico-Dicer'`s User Interface

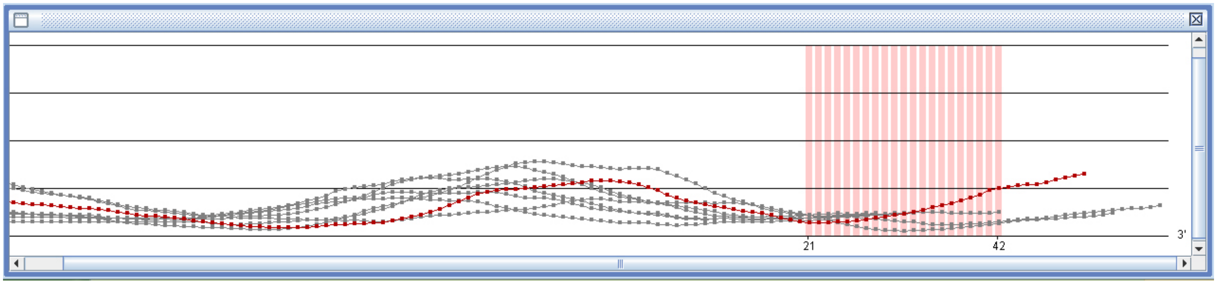


Figure 7: Intrinsic Plot. The precursors are separately scored and plotted unaligned. On the x-axis the precursor position can be read off, the score is mapped to the y-axis. Of the selected precursors from the tree, the topmost (usually, the input precursor) is highlighted in red. Position of predicted mature miRNA for the first selected precursor is displayed and shaded in salmon. The position can additionally be read off from the values beneath the shaded area. In this plot, 2 local minima can be observed. Due to the scoring method it holds: the lower the score, the higher the probability to be the position of mature miRNA. These two minima coincide with the complementary stem sequences which are paired. The area with relatively high scores between them refers to the stem loop which should have a very low probability to hold the mature miRNA.

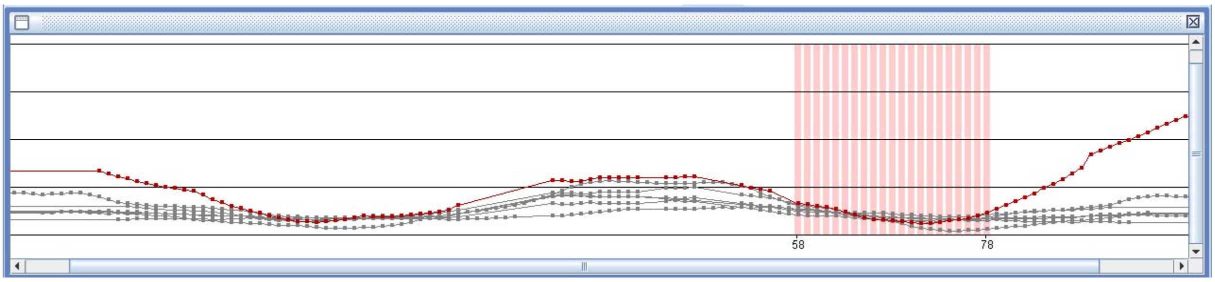


Figure 8: Intrinsic Profile. The precursors are scored separately and thereafter aligned. Gaps have no score, so they do not have a score data point. One can see in the plot where each precursor is gapped. Again, the position of mature miRNA is calculated for the topmost, red, precursor. The salmon shading shows the position within the plot, the values displayed beneath, specify the position relative to the precursor.

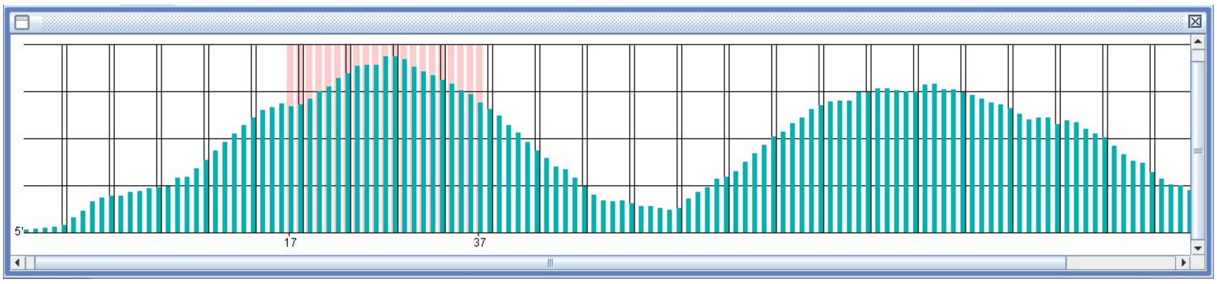
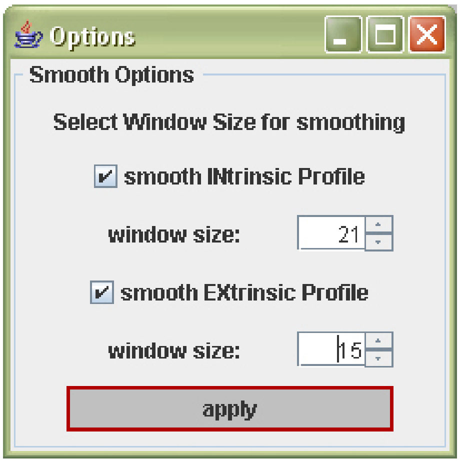


Figure 9: Extrinsic Profile. This plot displays the conservation profile of several precursors. Alignment position is displayed on the x-axis, conservation on the y-axis. The higher the bars, the higher the conservation at this position. The two 'hills' are the complementary stem regions, which are highly conserved, whereas the terminal loop between them differs among precursors. The area shaded in salmon is the most conserved one and represents the position where the mature miRNA is predicted. The shaded bars are located at the position of mature miRNA relative to the alignment, while the values displayed beneath specify the position relative to the considered precursor.

An additional plot feature of the application is the ‘smooth option’ which, however does not interfere with the prediction of mature miRNA. This feature has been implemented to display the plots in a user-friendly way. The grade of smoothing is selectable to the user by the following dialog accessible via ‘File→settings…’ (Figure 10)



☑ smooth INtrinsic Profile

window size: 21

☑ smooth EXtrinsic Profile

window size: 15

Figure 10: Smooth option dialog. The user can select whether he wants the plot to be smoothed or not by checking the accordant box. Grade of selection is managed by changing the window size over a user-friendly spinner. The window size refers to the length of the subsequence for which an average value is calculated and plotted instead of the original data point. The higher the selected window size, the smoother the plot. Initially, the smooth window size is set to 21. However, the raw data which is used to predict the position of mature miRNA is not affected by changing this setting. It is only for display purpose.

4.3 Evaluation of ‘In-Silico-Dicer’

In Sections 4.1 and 4.2 I gave an overview of ‘In-Silico-Dicer’, the program structure, the design and implementation details and concrete used algorithms. I showed that the application is highly flexible to developers who want to implement new classes to predict mature miRNA maybe with improved algorithms. ‘In-Silico-Dicer’ provides a scheme for the three approaches presented in Section 4.1,

- Intrinsic approach
- Intrinsic approach with alignment
- Extrinsic approach.

In this Section I want to assess whether these approaches produce reliable results. To examine this question, one has to apply validated miRNA precursors, where the position of mature miRNA is already known. After the program run, one can compare the predicted results with the genuine ones, which are experimentally verified.

4.3.1 Test sets and comparison details

The miRNA registry

The miRNAregistry is the most relevant database containing miRNA precursors from about 35 different organisms including plants, animals and viruses. The precursors are stored together with their corresponding authors, source organism, information on genome location, folding and the position of mature miRNA. Most of these positions have been discovered experimentally or at least verified, but for some recently added entries the mature sequence have been predicted by regarding homologues in related organisms [Berzikov et al., 2005]. However, a great part of these sequences has been validated afterwards. Thus, the data of the miRNA registry can be applied as confirmed comparison material for evaluation purposes. On the ftp server, there are several files of the databases provided for download. The ‘miRNA.dat’ file contains all relevant information that is necessary for evaluation of a ‘mature miRNA prediction program’:

- The precursor’s primary sequence
- The position of mature miRNA within the precursor sequence

I stored all the mature positions together with the unique precursor ID in a hashtable, so the predicted positions can be compared immediately to the genuine ones during the evaluation process.

Comparison details

When evaluating a particular method, one has to define a precise measure of quality of the results. In case of ‘In-Silico-Dicer’ there is the predicted miRNA position within the given precursor and the genuine position of the mature miRNA, which is derived from the validated data from the miRNAregistry. I defined the ‘overlap’ of these two miRNA positions as:

$$\frac{END\min - START\max}{overlap[\%]} = \frac{END\max - START\min}{100\%}$$

$$overlap[\%] = \frac{(END\min - START\max)*100\%}{(END\max - START\min)}$$

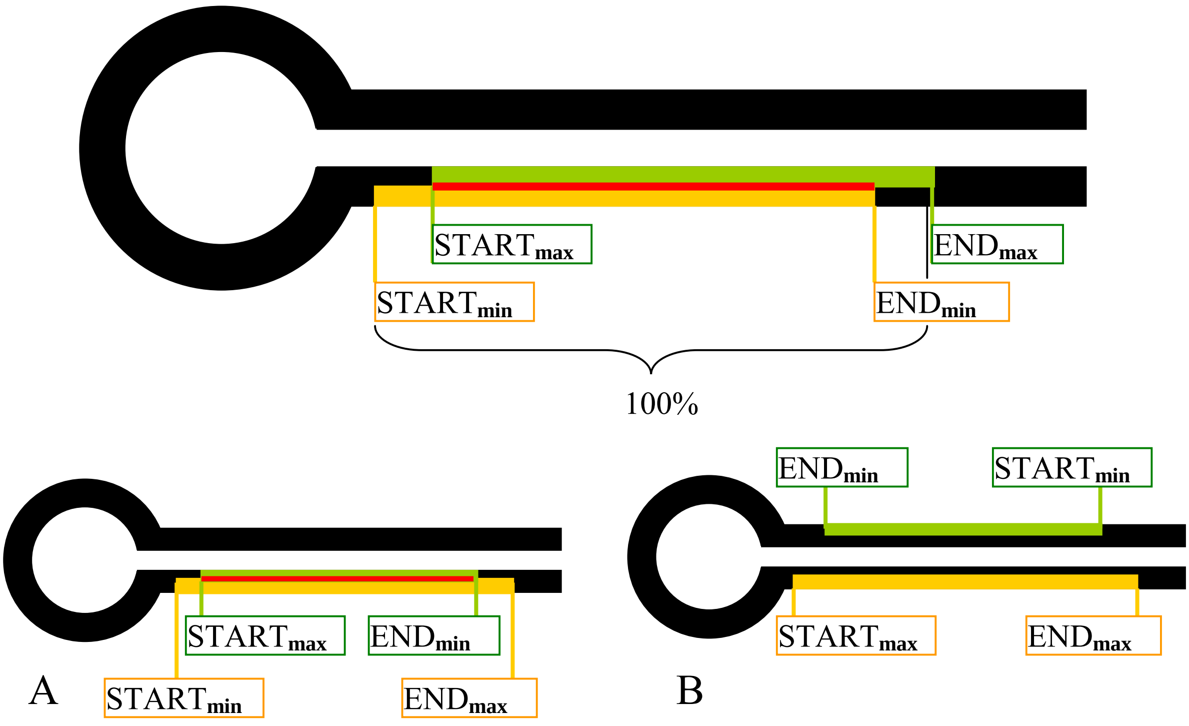


Figure 11: Illustration of the evaluation principle: Given a miRNA precursor (black, hairpin shaped), predict the position of putative miRNA (yellow sequence). Get the genuine position of mature miRNA within this precursor (green sequence) from the miRNAregistry. Compute the overlap, where overlap is defined as the length of the sequence the mature miRNA shares (red sequence).

The calculation of the overlap also works for all possible cases: (A) the predicted and the genuine sequence are not of the same length and (B) the predicted sequence is not are the same stem site, so there is no overlap at all. In case (B) the overlap will be negative.

4.3.2 Results

At the moment there are 35 miRNA precursor families known throughout the animal kingdom. In order to cover a broad spectrum of miRNAs and to get a representative result, I picked out one single precursor (always the first one) of each miRNA family. To evaluate the large amount of data automatically, I implemented an additional class that starts the application without the Graphical User Interface. All approaches are applied to the input data with their default settings for all implemented algorithms. The predicted and genuine positions of mature miRNA are output on the console in tabular form. On the basis of theses tables, the overlap was computed.

Intrinsic Plot

The run of the simple intrinsic approach, without aligning the particular precursor to any other, provides only partially satisfying results: the average overlap of all 35 miRNA predictions is 4%. This result includes all 19 putative miRNA positions which were predicted on the wrong stem side of the precursor, resulting in a negative overlap. Consequently, 16 out of the 35 examples have been predicted on the correct strand. In the bar plot in figure 12 the family representatives are sorted according to their overlap with the genuine mature position. One may speculate, that the prediction is better in several precursor families because these families follow a typical secondary structure scheme, which was the basis of the applied scoring algorithm. Maybe, the secondary structure of precursors is more diverse than expected. For those representatives, predicted on the correct strand, the overlap is never less than 11%. 34% of the evaluated data has an overlap above 50%, and even 22% reach an overlap of nearly 80%, which means a deviation from the genuine mature miRNA position of less than 2 nt. Only one of the test precursors, namely dme-mir-10, yields a 100% correct prediction. One can read off the amount of prediction success in percent of the evaluated data for all overlap values from the line plot in figure 13. In this diagram, a plateau between overlaps -25% and 15% attracts attention. This percentage derivation corresponds to an absolute derivation of about 16-34 nt. This is approximately the distance where the stem loop is expected. So, from this illustration, one can assume, that the simple intrinsic approach at least did not predict any mature miRNA on the terminal loop.
A very relevant fact to keep in mind for this approach is that the scoring function is just one step in a complete system of miRNA prediction which has been isolated to fit in this application. Wang et al. [2005] primarily executed several filter methods on their data set to reduce the amount of potential mature miRNAs. I ignored all these filters to simplify the procedure to its essence. Moreover, the authors developed their algorithm for the prediction of *A.thaliana* miRNAs, which may be very organism specific. In a review paper [Chen, 2005] it is stated that miRNA biogenesis is quite similar in *A.thaliana* and most animals, so there was the assumption, the method of Wang et al. may be also applicable also to the selected test set of metazoa. Interestingly, the prediction method was even less successful when evaluating with a set of *A.thaliana* precursors (data not shown) but this result may be due to the missing filter steps.
As a conclusion to this method, one can say that it is not too bad, considering the fact that this approach requires no additional information. Only by applying a scoring formula to the nucleotide and secondary structure sequence, the prediction of ‘In-Silico-Dicer’ reaches a significantly higher credibility than randomly selected mature miRNA positions. The 54% of totally wrongly predicted positions may lead to the assumption that many more diverse structural features of the precursor must be taken into account. Maybe, with increasing knowledge on precursor processing, one can develop a more complex scoring formula which includes these features.

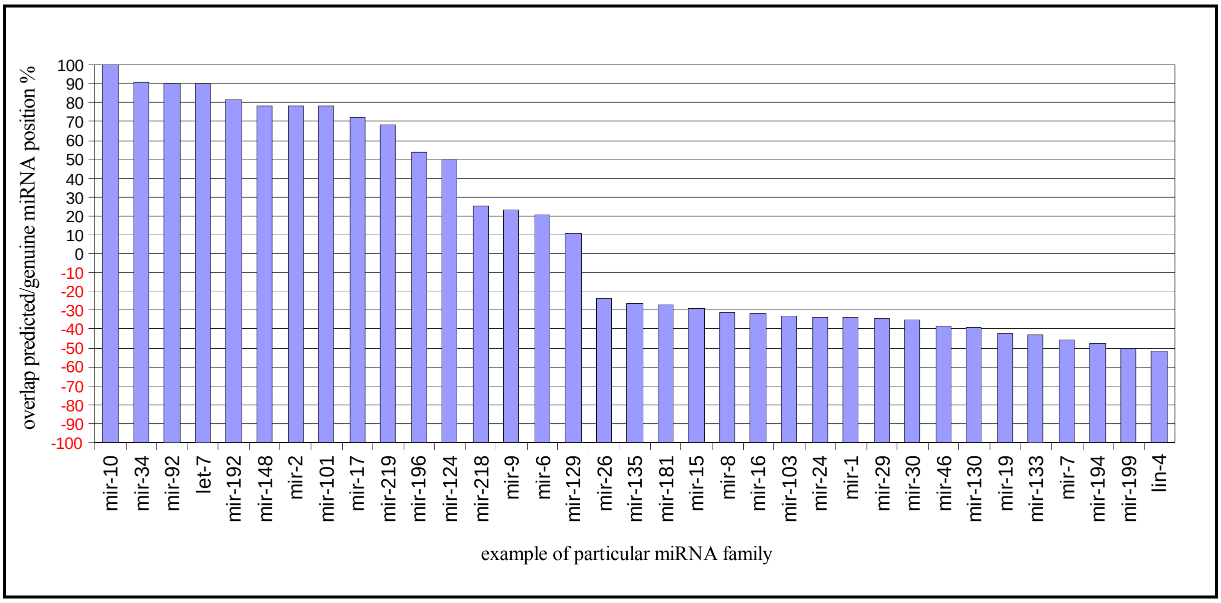


Figure 12: Intrinsic Approach (simple). The precursor families are ranked according to their reached overlap. 16 out of 35 selected examples (>45%) have been predicted on the correct stem side (positive overlap). 8 out of them (>22%) reached an overlap of nearly 80% which means < 2 nt deviation from the genuine position.

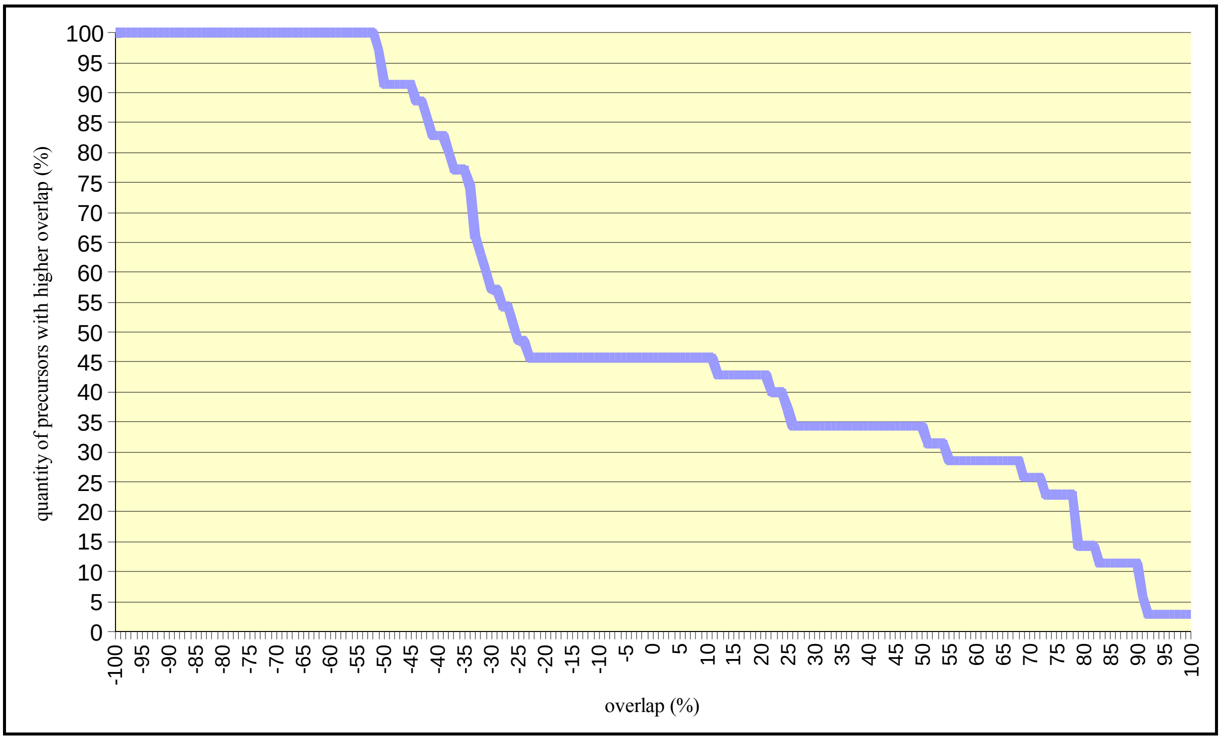


Figure 13: Intrinsic Approach (simple). The amount of successfully predicted miRNAs (in %) is mapped against the reached overlap. This means 100% of the evaluated miRNAs reached an overlap with their genuine position higher than -55%, 45% reached an overlap of 10% and only 1 miRNA (~2.9%) reached the overlap maximum of 100%.

Profiles – Intrinsic & Extrinsic

In order to yield comparable results, I used the same test set of representative precursors of every metazoan precursor family to evaluate the two similarity-based approaches: Intrinsic and Extrinsic Profiling. In these cases, the selection of comparable precursors is a variable which is considered to have a great influence on the results. The amount and grade of relation of these precursors can be regulated by varying the cut-off value for searching similar sequences. As a consequence, the evaluation of the two profiling approaches has to be performed referring to the cut-off value. Before evaluating, I tried some cut-off values and determined that there is a relatively small interval in the cut-off range where the amount of close related precursors suddenly alters greatly. This interval is between cut-offs 85 and 120, cut-offs lower than 85 return almost the whole database as similar precursors, while a cut-off higher than 120 usually returns the given precursor itself. So, consequently I restricted the evaluation to this interval. I chose a step-size of 5 to yield relatively exact and differentiated values.

Intrinsic profile

In the line plot in Figure 14, the resulting percentage of the whole test set of precursors (35 precursors) are mapped. The precursor reached an overlap higher than the according value on the x-axis. Every data-sequence (one line) refers to a specific cut-off value between 85 and 120. Additionally I added the success curve of the intrinsic plot from Figure 13 for comparison purposes. This data-sequence is mostly lower than all other lines referring to the extended intrinsic prediction approach, which means the average rate of overlap is significantly higher in the proposed extension of Wang's algorithm. Due to an alignment with similar precursor sequences, an average advancement in overlap to the genuine mature miRNA position of approximately 10% has been achieved. However, it is remarkable, that in the runs with minimal cut-off values (85, 90) no miRNA has been predicted with 100% overlap. With theses cut-off values up to 98 similar precursors were found throughout the database (data not shown), which may include also distantly related ones. One can assume from this result that an increasing amount of comparison precursors, yielded by a low cut-off, not always causes increasing prediction reliability. On the contrary, this may lead to an alignment with incomparable precursors, which consequently tampers the prediction. From this perception, one can learn the importance of the optimal cut-off adjustment. The arising question is: Is it possible to determine a universal cut-off optimum for all precursors? To answer this question, I compared the reached overlap precursor-family-wise for all cut-off steps. The bar plot in Figure 15 shows the stacked overlaps for each miRNA family (its representative precursor). From this diagram one can spot that in the minority of examples the overlap differs significantly according to the chosen cut-off value. Especially, there are very few predicted miRNAs whose overlap with their genuine position is positive for several cut-offs and negative for the rest. In the majority of test precursors, the prediction (which means the % overlap) is either solely positive or solely negative. Furthermore, in most of these cases, the prediction success in one miRNA family remains the same for all selected cut-off values. As a conclusion to this diagram one can assume, that there is no comprehensive cut-off value for all miRNA families to be found. But this plot confirms the supposition that there are some families whose predicted mature miRNA position appears to be more reliable than others. Again mir-10, mir-192, mir-34, mir-9, mir-92 are ahead in this statistic, like in the intrinsic (single) approach, which is not surprising due to the fact, that this analysis refers to an extension of the single approach.

Plots for Extended Intrinsic Approach

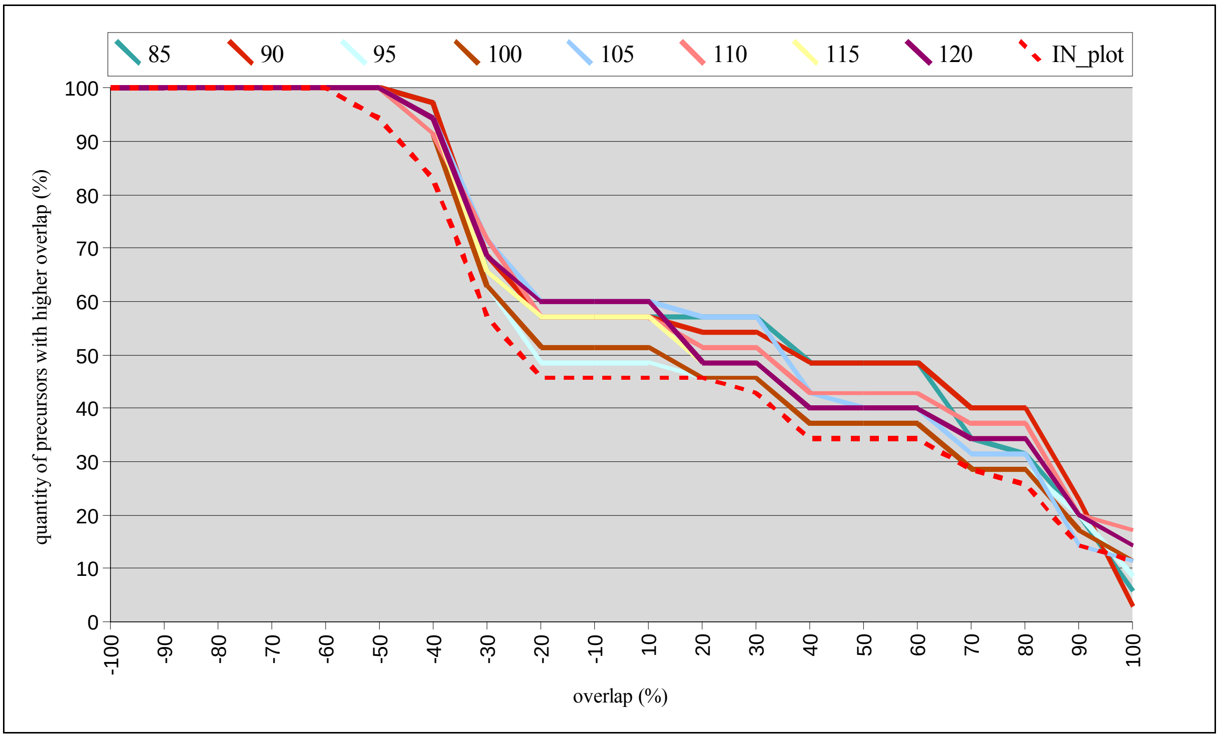


Figure 14: Extended Intrinsic Approach. The particular scored precursor has been aligned to the similar precursors derived according to the cut-off value. The amount of successfully predicted miRNAs (in %) is mapped against the reached overlap. The lines which refer to the different cut-offs are clustered in the plot. The red dotted data-sequence is the curve derived from the intrinsic simple prediction method.

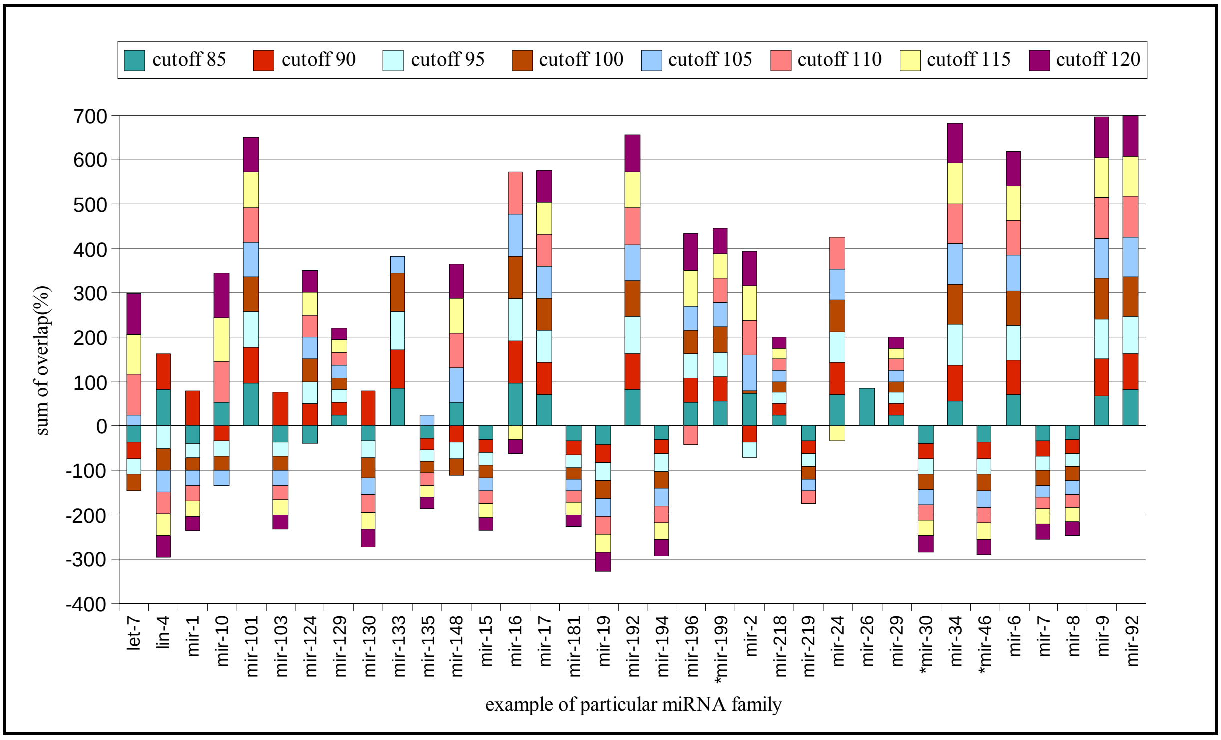


Figure 15: Extended Intrinsic Approach. Overlaps derived from different cut-offs are shown in a stacked way for each single precursor family. As the overlap interval 200 (-100 % to 100%) the upper and lower bounds in this diagram are theoretically 800 and -800 because of the 8 cut-offs.

Plots for Extrinsic Approach

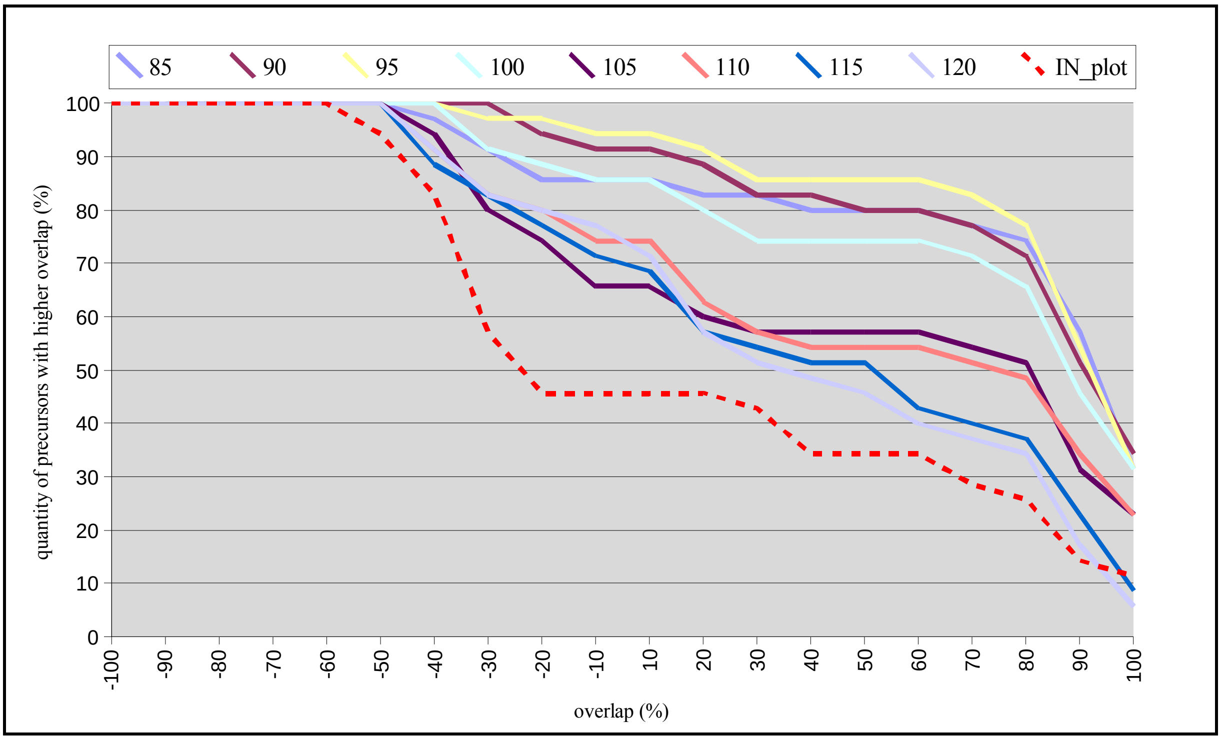


Figure 16: Extrinsic Approach. The particular precursor has been aligned to the similar precursors derived according to the cut-off value, than the prediction was made based on the conservation. The amount of successfully predicted miRNAs (in %) is mapped against the reached overlap. The red dotted data-sequence is the curve derived from intrinsic simple prediction method.

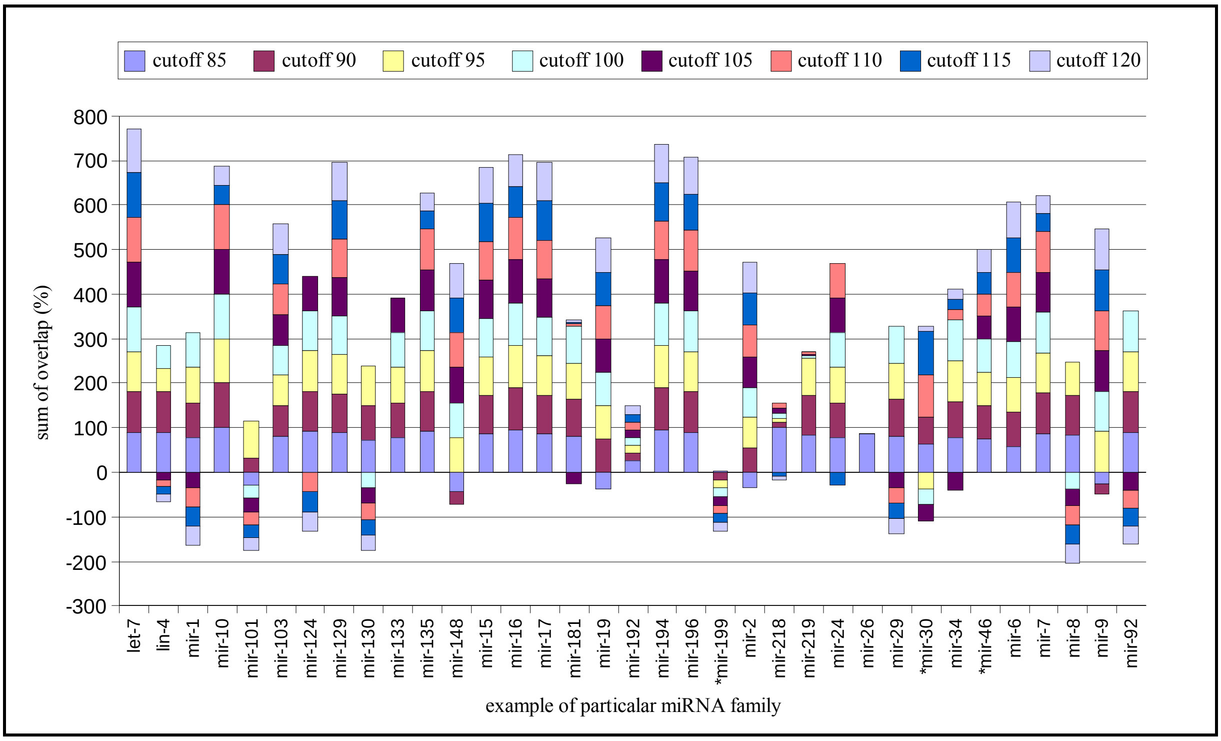


Figure 17: Extrinsic Profile. Overlaps derived from different cut-offs are shown in a stacked way for each single precursor family. As the overlap interval is 200 (-100 % to 100%) the upper and lower bounds in this diagram are theoretically 800 and -800 because of the 8 cut-offs.

Extrinsic Profile

To evaluate the third prediction method I applied the same test set of representative precursors again. The extrinsic miRNA prediction approach described in Section 4.1.2 was outstandingly successful on the chosen data. In Figure 16 a line plot, similar to Figure 14, is shown. The x-axis represents the degree of overlap in %. On the y-axis, the percentage of miRNA precursors that reached a certain degree of overlap is registered. Like in Figure 14 I added the overlap curve of the intrinsic (simple) prediction method; the advancement in prediction reliability is remarkable: regarding all cut-offs, at least 50% of the miRNAs have been predicted with an overlap of 30%. Overlaps of 80% were reached by nearly 80% of the predicted miRNAs using the lower cut-off values (85, 90, 95, 100) while only about 50% of the miRNAs reached this overlap level with cut-offs higher than 105. This noticeable difference shows, that the cut-off and consequently the amount of comparison-precursors in the extrinsic approach has a different impact on the prediction result than in the intrinsic approach. The extended intrinsic method can return a misleading prediction on a too low cut-off causing many similar precursors. On the contrary, the extrinsic method appears to yield better results when applying these low cut-off values and yields decreasing overlaps with increasing cut-offs. This may be due to the fact that the extrinsic prediction method is mainly based on the conservation which depends only on the alignment, whereas the intrinsic method basically depends on the scoring, even if it is aligned afterwards. When there are many comparison precursors, the conservation can be calculated more reliably. This does not imply that an alignment to nearly all precursors of the whole database will return the best result, from Figure 16 one could assume a cut-off around 95 may be the best choice for this dataset. Again, I questioned if there exists one ultimate cut-off for the extrinsic prediction of mature miRNA in general, which may be possible regarding the success curves in Figure 16. From the plot in Figure 17, however, this appears unlikely. In most cases, this plot shows the same features as the comparable intrinsic one (Figure 15): the percent overlap per precursor is either solely positive or solely negative and mostly the overlap values are similar within one example – there are some exceptions like mir-1, mir-8 and mir-101 for instance. Regarding this illustration one can again obtain the high prediction reliability of this approach: some miRNAs, the miRNAs which appear to be easily predictable with the intrinsic approach, show relatively good results also here and additionally there are several miRNAs whose prediction success increased remarkably. I observed that these precursors are mostly assigned to have many relatives in other organisms: one exemplary miRNA family here is let-7, whose prediction success more than doubled in comparison to the extended intrinsic approach.

5 Discussion

In this thesis, I presented some ideas to solve the problem of mature miRNA prediction and implemented an application, which deploys these ideas. I evaluated the applied approaches and compared the results.

The intrinsic approach, which scores the precursor by means of primary and secondary structure was the most ambitious attempt because we do not have any more information than the precursor provides. The idea was to reconstruct the recognition and the cleavage process mediated by the RNase-III-type processing enzymes. Due to a lack of biological experience I build in a piece of an already existing algorithm, which however is very specific for *Arabidopsis*. All in all, this pure approach has not been truly successful when applying it to the chosen test set of animal precursors, but nevertheless, there were some good predictions. These lead to the conclusion that the prediction success is probably just a question of how much we known about the precursor's properties and consequently how precise the scoring formula can be defined. Maybe, in the future, one can develop the ultimate scoring algorithm applicable to miRNAs of all families. In order to improve the intrinsic method despite the existing scoring formula, I extended this approach by using related precursor sequences. Due to this additional alignment step, a noticeable prediction advance could be observed; the average overlap result was at least 10% higher than before. Dependent on the amount of precursors to compare, even better results can be achieved. I assumed the cut-off for searching similar sequences has a great influence on the prediction and therefore I evaluated with several adjustments in a relevant interval. High cut-off values, which yield a little set of mostly closely related precursors, appear to be favourable here. The reason may be that the results of the scoring are blurred by aligning the precursor to others. Although the extended intrinsic strategy has not been very reliable in all cases, one can obtain some useful information from the evaluation for developing a similar approach.

Furthermore, I considered a totally different idea to overcome the challenge to predict mature miRNA: the extrinsic approach. This one is based on a multiple alignment with related precursors and regards the conservation among them. In comparison to the previous approaches its prediction was significantly more reliable. Especially good results have been achieved with lower cut-off values; which indicates that a lot of comparison precursors contribute to the prediction credibleness in this case. On account of this oppositional outcome compared to the intrinsic strategy, it might be a suggestive consideration to somehow combine these two approaches. Probably, this will compensate the particular disadvantages of intrinsic and extrinsic miRNA prediction.

Due to the highly flexible implementation, 'In-Silico-Dicer' is open to integrate new algorithms which may increase the reliability of the prediction. There are already some new ideas to improve the application: one could optimize the calculation of conservation by regarding the perfect complementary on the 5' end of the mature miRNA to the target. Maybe, the sequence between nt 2-7 appears to be even higher conserved than the rest of the miRNA sequence is. By improving the algorithms, the application also becomes more effective and I am also going to make it applicable for high-throughput computation. Certainly, the development phase of 'In-Silico-Dicer' is not completed yet.

6 Conclusion & Future Prospects

The astonishing diversity of regulatory pathways directed by small RNAs has been discovered through a combination of genetic and biochemical approaches. However, to completely understand the whole dimension of this complex system, we have to apply bioinformatical methods, too. Recent researches, which combine these two approaches, have shown that the total number of miRNAs in humans is much larger than previously expected [Bentwich et al., 2005]. With their bioinformatic-experimental prediction method the authors nearly doubled the current number of sequenced human miRNAs. However, to confirm new miRNAs, which appear to have no related ones in other primates for instance, they had to deploy expendable microarray analyses and sequence-directed cloning. If we were able to develop a reliable fully computational prediction method for miRNAs and their targets, the enlightenment of RNA mediated gene regulation would be considerably more seizable. Several uncertainties associated with miRNA predictions include: (1) the orientation of the transcript (plus or minus strand) for a genomic location encoding hairpin sequence, (2) the position of the processing sites within the hairpin structure and (3) the determination of which of the paired segments of the hairpin will constitute the mature miRNA [Aravin et al., 2005]. There already exist several programs which yield very satisfying results in genome-wide precursor prediction and searching for possible miRNA targets but relatively less effort has been spend to predict the processing from pri/pre-miRNA to the mature sequence. Within the time of my bachelor thesis I tried to develop approaches to overcome the missing link in this chain of computationally reconstructing the miRNA pathway. Especially regarding the proposed intrinsic approach one can conclude that the processing by several enzymes requires further investigation; to transcribe this special step in a reliable algorithm will be a great challenge in the future.

7 References

[1] Ambros V., Bartel B., Bartel DP., Burge CB., Carrington JC., Chen X., Dreyfuss G., Eddy SR., Griffiths-Jones S., Marshall M., Matzke M., Ruvkun G., Tuschl T.(2003), A uniform system for microRNA annotation. RNA 2003 Mar 9(3):277-9

[2] Ambros V.(2004); The functions of animal microRNAs. Nature 2004 Sep 16;431(7006):350-5

[3] Aravin A., Tuschl T. (2005), Identification and characterization of small RNAs involved in RNA silencing. FEBS Lett. 2005 Sep 6; [Epub ahead of print]

[4] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004 Jan 23;116(2):281-97.

[5] Bentwich I., Avniel A., Karov Y., Aharonov R., Gilad S., Barad O., Barzilai A., Einat P., Einav U., Meiri E., Sharon E., Spector Y., Bentwich Z. Identification of hundreds of conserved and nonconserved human microRNAs. Nat. Gent. 2005 Jul;37(7):766-70. Epub 2005 Jun 19.

[6] Berezikov E., Guryev V., van de Belt J., Wienholds E., Plsterk RH., Cuppen E.(2005), Phylogenetic shadowing and computational identification of human microRNA genes. Cell. 2005 Jan 14;120(1):21-4

[7] Burgler C., Macdonald PM. (2005), Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method. BMC Genomics. 2005 Jun 8;6(1):88.

[8] Chen X.(2005); microRNA biogenesis and function in plants. FEBS Lett. 2005 Sep 3; [Epub ahead of print]

[9] Chenna R., Sugawara H., Koike T., Lopez R., Gibson TJ., Higgins DG., Thompson JD. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 2003 Jul 1;31(13):3497-500

[10] Dykxhoorn DM., Novina CD., Sharp PA.(2003), Killing the messenger: short RNAs that silence gene expression. Nat Rev Mol Cell Biol. 2003 Jun;4(6):457-67

[11] Filipowicz W., Jaskiewicz L., Kolb FA., Pillai RS.(2005), Post-transcriptional gene silencing by siRNAs and miRNAs. Curr Opin Struct Biol. 2005 Jun; 15(3):331-41

[12] Griffiths-Jones S.(2004), The mircoRNA Registry. Nucleic Acids Res. 2004 Jan 1; 32(Database issue):D109-11

[13] Hofacker IL., Priwitzer B., Stadler PF., Prediction of locally stable RNA secondary structures for genome wide-surveys surveys. Bioinformatics 2004;20, 186-190

[14] John B., Enright AJ., Aravin A., Tuschl T., Sander C., Marks DS. Human MicroRNA targets. PloS Biol. 2004 Nov;2(11):e363. Epub 2004 Oct 5.

[15] Kim VN. (2005a), SmallRNAs: Classification, Biogenesis, and Function. Mol. Cells 2005 Feb 19(1):1-15.

[16] Kim VN. (2005b), MicroRNA biogenesis: coordinated cropping and dicing. Nature Rev Mol Cell Biol 2005 May; 6(5): 376-85

[17] Kiriakidou M., Nelson PT., Kouranov A., Fitziev P., Bouyioukos C., Mourelatos Z., Hatzigeorgiou A. A combined computational-experimental approach predicts human microRNA targets. Genes Dev 2004 May 15; 18(10),1165-1178. Epub 2004 May 6

[18] Lai EC., Tomancak P., Williams RW., Rubin GM. Computational identification of Drosophila microRNA genes. Genome Biol. 2003;4(7):R42. Epub 2003 Jun 30.

[19] Lee Y., Jeon K., Lee JT., Kim S., Kim VN. (2002), MicroRNA maturation: stepwise processing and subcellular localization. EMBO J. 2002 Sep 2; 21(17):4663-70

[20] Lee Y., Ahn C., Han J., Choi H., Kim J., Yim J., Lee J., Provost P., Radmark O., Kim S., Kim VN. (2005), The nuclear Rnase II Drosha initiates microRNA processing. Nature 2003 Sep 25; 425(6956):415-9.

[21] Legendre M., Lambert A., Gautheret D. Profile-based detection of microRNA precursors in animal genomes. Bioinformatics 2005 Apr 1;21(7):841-5 Epub 2004 Oct 27

[22] Lewis BP., Shih IH., Jones-Rhoades MW., Bartel DP., Burge CB. (2003) Prediction of mammalian microRNA targets. Cell 2003 Dec 26;115(7):787-98

[23] Lim LP., Lau NC., Weinstein EG., Abdelhakim A., Yekta S., Rhoades MW., Burge CB., Bartel DP. The microRNAs of Caenorhabditis elegans. Genes Dev. 2003 Apr 15;17(8):991-1008. Epub 2003 Apr 2.

[24] Meister G., Tuschl T.(2004), Mechanisms of gene silencing by double-stranded RNA. Nature 2004 Sep 16;431(7006):343-9

[25] Pasquinelli AE., Hunter S., Bracht J. (2005), MicroRNAs: a developing story. Curr Opin Genet Dev. 2005 Apr; 15(2):200-5

[26] Rehmsmeier M., Steffen P., Hochsmann M., Giegerich R.(2004), Fast and effective prediction of microRNA/target duplexes. RNA 2004 Oct; 10(10):1507-17

[27] Rhoades MW., Reinhard BJ., Lim LP., Burge CB.; Bartel B.; Bartel DP. Prediction of plant microRNA targets. Cell 2002 Aug 23;110(4):513-20

[28] Smith TF., Waterman MS., Identification of Common Molecular Subsequences. J Mol Biol. 1981 Mar 25;147(1):195-7.

[29] Stark A., Brennecke J., Russell RB., Cohen SM. Identification of Drosophila microRNA targets. PloS Biol. 2003 Dec;1(3):E60. Epub 2003 Oct 13.

[30] Tomari Y., Matranga C., Haley B., Martinez N., Zamore PD. A protein sensor for siRNA asymmetry. Science 2004 Nov 19;306(5700):1377-80

[31] Vermeulen A., Behlen L., Reynolds A., Wolfson A., Marshall Ws., Karpilow J., Khvorova A.(2005), The contibutions of dsRNA structure to Dicer precifity and efficiency. RNA 2005 May; 11(5):674-82. Epub 2005 Apr 5.

[32] Wang X.; Zhang J., Li F., Gu J., He T., Zhang X., Li Y.(2005), MicroRNA identification based on sequence and structure alignment. Bioinformatics. 2005 Sep 15; 21(18):3610-4. Epub 2005 Jun 30

[33] Wang XJ., Reyes JL., Chua NH.; Gaasterland T.(2004), Prediction and identification of Arabidopsis thaliana microRNAs and their targets. Genome Biol. 2004; 5(9):R65. Epub 2004 Aug 31.

[34] Zeng Y.; Yi R., Cullen BR. (2005), Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. EMBO J. 2005 Jan 12; 24(1):138-48. Epub 2004 Nov 25.

[35] Zuker M., Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003 Jul 1;31(13):3406-15.

8 Appendix

FASTA Format

- This format contains a one line header followed by lines of sequence data.
- Sequences in fasta formatted files are preceded by a line starting with a “>” symbol.
- The first word on this line is the name of the sequence. The rest of the line is a description of the sequence

Term	Entry Name (ID)	Accession	Organism	Family	Sequence Type
e.g.	cel-let-7	MI0000001	Caenorhabditis elegans	let-7	stem-loop

- The remaining lines contain the sequence itself.
- Blank lines in a FASTA-file are ignored, and so are spaces or other gap symbols(dashes, underscores, periods) in s sequence.
- FASTA-files containing multiple sequences are just the same, with one sequence after another. This format is called multi-FASTA

```
>cel-let-7 MI0000001 Caenorhabditis elegans let-7 stem-loop
UACACUGUGGAUCCGGUGAGGUAGUAGGUUGUAUAGUUUGGAAUAUUACCACCGGUGAAC
UAUGCAAUUUUCUACCUUACCGGAGACAGAACUCUUCGA
```

ClustalW, used Output Format

For the given multi-FASTA input:

```
>ssc-let-7i MI0002447 Sus scrofa let-7i stem-loop
CUGGCUGAGGUAGUAGUUUGUGCUGUUGGUCGGGUUGUGACAUUGCCCGCUGUGGAGAUAACUGCGCAAGCUACU
GCCUUGCUAG
>rno-let-7i MI0000835 Rattus norvegicus let-7i stem-loop
CUGGCUGAGGUAGUAGUUUGUGCUGUUGGUCGGGUUGUGACAUUGCCCGCUGUGGAGAUAACUGCGCAAGCUACU
GCCUUGCUAG
>mmu-let-7i MI0000138 Mus musculus let-7i stem-loop
CUGGCUGAGGUAGUAGUUUGUGCUGUUGGUCGGGUUGUGACAUUGCCCGCUGUGGAGAUAACUGCGCAAGCUACU
GCCUUGCUAG
```

The multi-FASTA output format looks like that:

```
>ssc-let-7i
----------------------------------------------------CUGGCUGAG
GUAGUAGUUUGUGCUGUUGGUC-GGGUU--GUGACAUUG--CCCGCUGUGGAG--AUAAC
UGCGCAAGCUACUGCCUUGCUAG----------------------
>rno-let-7i
----------------------------------------------------CUGGCUGAG
GUAGUAGUUUGUGCUGUUGGUC-GGGUU--GUGACAUUG--CCCGCUGUGGAG--AUAAC
UGCGCAAGCUACUGCCUUGCUAG----------------------
>mmu-let-7i
----------------------------------------------------CUGGCUGAG
GUAGUAGUUUGUGCUGUUGGUC-GGGUU--GUGACAUUG--CCCGCUGUGGAG--AUAAC
UGCGCAAGCUACUGCCUUGCUAG----------------------
```