

Improved Workflows for RNA Homology Search

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
eingereichte

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR rerum naturalium
(Dr. rer. nat.)

im Fachgebiet
Informatik

vorgelegt

von Master of Science Ali M. Yazbeck

geboren am 8. März 1988 in Bouday, Libanon

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Peter F. Stadler
2. Prof. Dr. Kifah R. Tout

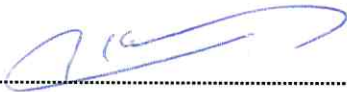
Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 17. Juni 2019 mit dem Gesamtprädikat **magna cum
laude**

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

04/07/2019

.....
(Ort, Datum)



.....
(Unterschrift)

“There are always lessons to be found in the darkest moments. It’s a moral obligation to dig deep and find that little glimmer of hope or pearl of wisdom.”

Andrea Pirlo

Abstract

Non-coding RNAs are the most abundant class of RNAs found throughout genomes. These RNAs are key players of gene regulation and thus, the function of whole organisms. Numerous methods have been developed so far for detecting novel classes of ncRNAs or finding homologs to the known ones. Because of their abundance, the sequence availability of these RNAs is rapidly increasing, as is the case for example for microRNAs. However, for classes of them, still only incomplete information is available, invertebrates 7SK snRNA for instance. Consequently, a lot of false positive outputs are produced in the former case, and more accurate annotation methods are needed for the latter cases to improve derivable knowledge. This makes the accuracy of gathering correct homologs a challenging task and it leads directly to a not less important problem, the curation of these data.

Finding solutions for the aforementioned problems is more complex than one would expect as these RNAs are characterized not only by sequences information but also structure information, in addition to distinct biological features. In this work, data curation methods and sensitive homology search are shown as complementary methods to solve these problems. A careful curation and annotation method revealed new structural information in the invertebrates 7SK snRNA, which pushes the investigation in the area forward. This has been reflected by detecting new high potential 7SK RNA genes in different invertebrates groups. Moreover, the gaps between homology search and well-curated data on the one side, and between experimental and computational outputs on the other side, are closed. These gaps were bridged by a curation method applied to the microRNA data, which was then turned into a comprehensive workflow implemented into an automated pipeline. **MIRfix** is a microRNA curation pipeline considering the detailed sequence and structure information of the metazoan microRNAs, together with biological features related to the microRNA biogenesis. Moreover, this pipeline can be integrated into existing methods and tools related to microRNA homology search and data curation. The application of this pipeline on the biggest open source microRNA database revealed its high capacity in detecting wrong annotated pre-miRNA, eventually improving alignment quality of the majority of the available data. Additionally, it was tested with artificial datasets highlighting the high accuracy in predicting the pre-miRNA components, miRNA and miRNA*.

Acknowledgements

No surprise that everyone starts with you Prof. Dr. Peter F. Stadler (P), so first I would like to say thanks to Peter for everything. Prof. Peter has been the source of inspiration for me during all my PhD period. No doubts how much anyone can learn science from him, but I also learned from him how to make a balance between being professional and friend at the same time. I won't forget our open-air office, never ever!.

I want to thank my Lebanese supervisor Prof. Dr. Kifah R. Tout for his support since the beginning of this journey, which started with him from my master's thesis. With his help, it was my first step in Bioinformatics. Of course, without him, Bioinformatics was not introduced to the Lebanese University systems.

Dr. Joerg, hard to push, easy to discuss with. Joerg was the Obi-Wan Kenobi at the end, and he pushed forward my work until it is done + proofreading. Thank a lot Joerg for being also patient when I really needed it.

Jens and Petra? yes, they are part of everything, wherever there is a technical or administrative help you will find them. My longest flight ever was with Jens, but surprisingly, I did not ask for any installation. Many thanks for Jens and Petra. The administrative competitors from the Lebanese side were Zeinab and Fatima who I would like to send many thanks for them and for their patience and understanding all time.

Of course, Bioinformatics! an environment that everyone would like to work in. Thanks for everyone, and literally for everyone in this group for being friendly and lovely and always ready to help and discuss any topic. Nancy and Ritu, thank you for being more than colleagues, thanks for being true friends.

Special thanks for Deisy and Dr. Moe Hankir for their careful proofreading and being supportive all time.

To all my friends in Lebanon and Germany. Hussein and Mohammad I would like to name you because without you guys it was not easy to handle things in Lebanon during my stay in Germany. My friends in Germany it is a really long list and it is very very special thanks, I wish I can say it face to face for each of you. If you are reading this and smiling then you are on the list... Pedro and Elke how I can forget them, the best couple I have seen in my life. One of the luckiest moments in my life was meeting them!.

Being in a warm and lovely family is really something very special and one of the things that one should really appreciate. To all my brothers and my sister many many thanks and I am lucky to be part of this family. A lot of changes, my only sister she suddenly grew up and now she is a mother of the cutest creature in the world. To my parents, no words can describe my feelings towards you, and thank you is something really very small to say. My father, he dedicated his time, life and strength for us and it is time to pay this back to him. My mother, how can I forgot your calls and support all this time, and your tasty food when I am home.

To a very special person, who in a short time became my soulmate and the person I would like to spend as much as possible of my life with her. The end? yes, I hope till the end. For old-fashion not likable traditions I will leave a space for your name and you will fill it by your hand once the traditional risks are cleared.

Bibliographic Information

This work is based on three first authorship peer reviewed published articles and one co-authorship in a book chapter (submitted):

Ali M. Yazbeck, , Kifah R. Tout, Peter F. Stadler and Jana Hertel. "*Towards a consistent, quantitative evaluation of microRNA evolution.*" *Journal of integrative bioinformatics* 14.1 (2017).

Ali M. Yazbeck, , Kifah R. Tout, and Peter F. Stadler. "*Detailed secondary structure models of invertebrate 7SK RNAs.*" *RNA biology* 15.2 (2018): 158-164.

Cristian A. Velandia-Huerto, Ali M. Yazbeck, Jana Schor, Peter F. Stadler. "*Evolution and Phylogeny of MicroRNAs Protocols, Pitfalls, and Problems.*" (Submitted 2018)

Ali M. Yazbeck, , Peter F. Stadler, Kifah R. Tout, and Jörg Fallmann. "*Automatic Curation of Large Comparative Animal MicroRNA Data Sets.*" *Bioinformatics*. 2019 Apr 16.

Contents

Selbstständigkeitserklärung	ii
Abstract	iv
Acknowledgements	v
1 Introduction	2
2 Biological and Computational background	5
2.1 Biology	5
2.1.1 Non-coding RNAs	6
2.1.2 RNA secondary structure	7
2.1.3 Homology versus similarity	8
2.1.4 Evolution	8
2.2 The role of computational biology	10
2.2.1 Alignment	10
2.2.1.1 Pairwise alignment	10
2.2.1.2 Multiple sequence alignment (MSA)	13
2.2.2 Homology search	15
2.2.2.1 Sequence-based	15
2.2.2.2 Structure-based	15
2.2.3 RNA secondary structure prediction	17
3 Careful curation for snRNA	18
3.1 Biological background	18
3.2 Introduction to the problem	19
3.3 Methods	20
3.3.1 Initial seeds and models construction	20
3.3.2 Models anatomy then merging	23
3.4 Results	25
3.4.1 Refined model of arthropod 7SK RNA	26
3.4.1.1 5' Stem	26
3.4.1.2 Extension of Stem A	27
3.4.1.3 Novel stem B in invertebrates	27
3.4.1.4 3' Stem	29
3.4.2 Invertebrates model conserves the HEXIM1 binding site	30
3.4.3 Computationally high potential 7SK RNA candidate	32
3.4.4 Sensitivity of the final proposed model	33
3.5 Conclusion	33

4	Behind the scenes of microRNA driven regulation	35
4.1	Biological background	35
4.2	Databases and problems	38
4.3	MicroRNA detection and curation approaches	40
5	Initial microRNA curation	45
5.1	Introduction	45
5.2	Methods	46
5.2.1	Data pre-processing	46
5.2.2	Initial seeds creation	46
5.2.3	Main course	47
5.3	Results and discussion	50
5.4	Conclusion	52
6	MIRfix pipeline	54
6.1	Introduction	54
6.2	Methods	56
6.2.1	Inputs and Outputs	56
6.2.2	Prediction of the mature sequences	56
6.2.3	The original precursor and its alternative	60
6.2.4	The validation of the precursor	62
6.2.5	Alignment processing	65
6.3	Results and statistics	67
6.4	Applications	70
6.4.1	Real life examples and artificial data tests	70
6.4.2	miRNA and miRNA* prediction	78
6.4.3	Covariance models	82
6.5	Conclusion	84
7	Discussion	86
A	7SK RNA project	90
B	MicroRNA project	98
	Bibliography	102

List of Figures

2.1	POL-II promoters.	6
2.2	POL-III promoters.	6
2.3	Example of structural RNA, U7 RNA.	8
2.4	Simple evolution example.	10
2.5	Smith-Waterman scoring matrix	12
2.6	Needle-Wunsch alignment scoring matrix	13
2.7	RNAfold and C o F old predictions.	17
3.1	The vertical division.	25
3.2	Phylogenetic distribution of the 7SK genes identified in Arthropoda.	26
3.3	Comparison of the 5 prime stems of 7SK RNA in Vertebrates and Hexapoda.	27
3.4	Comparison of stem B of 7SK RNA.	29
3.5	Hexapoda 7SK RNA structure.	30
3.6	7SK RNA invertebrates model.	31
3.7	General invertebrates structure.	32
3.8	Upstreams alignment.	33
4.1	MicroRNA biogenesis.	36
4.2	MicroRNA binds to mRNA.	37
4.3	Combination of microRNA to a drug in breast cancer treatment.	38
4.4	microRNA families distribution at miRBase (v. 21)	39
5.1	Entropy comparison for each group of microRNA families.	50
5.2	Distribution of miRNA homologs of miRNA families mir-3.	52
6.1	This figure shows the MIRfix general workflow as it is divided into two levels, precursors' level and alignment level. New alignments are created at the first level after processing the precursors independently of each other. The produced alignments at the previous level are processed again in the alignment level.	55
6.2	miRNA prediction method.	57
6.3	Extracting the alternative precursor.	61
6.4	Different folding states for the annotated precursors.	63
6.5	Changed precursors distribution.	69
6.6	Comparison of the secondary structures of original and corrected precursors.	71
6.7	Similar correction between MIRfix and miRBase.	72
6.8	Prediction of miRNA* follows the Dicer and AGO process.	73

6.9	Example of correction at alignment level.	74
6.10	Original alignment VS improved, after the precursor level. . .	76
6.11	Original alignment VS improved, after the alignment level. . .	77
6.12	Quality of miRNA prediction.	79
6.13	Example of miRNA prediction.	80
6.14	The second mature prediction performance in the first test set.	81
6.15	The second mature prediction performance in the second test set.	81
6.16	Comparison for the relative effective sequence number.	82
6.17	Comparison of the relative entropy of CMs.	83
6.18	Example of CM search results.	84
A.1	Human 7SK RNA.	90
A.2	Diptera model.	91
A.3	Arachnida model.	92
A.4	Metapolybia cingulata structure.	93
A.5	CMs comparison between existing CMs and the CM of our project.	94
B.1	Entropy comparison in Group 1	98
B.2	Entropy comparison in Group 2	99
B.3	Entropy comparison in Group 3	99
B.4	Entropy comparison in Group 4	100
B.5	Entropy comparison in Group 5	100
B.6	Entropy comparison in Group 6	101

List of Tables

6.1	Validation criteria for the precursors.	66
6.2	Accuracy distribution of the predicted mature sequences in a given sample.	79
A.1	The searched genomes are listed with their sources.	95
A.2	The searched genomes are listed with their sources.	96
A.3	The searched genomes are listed with their sources.	97

List of Abbreviations

A	Adenine
AGO	ArGOnaute
BLAST	Basic Local Alignment Search Tool
C	Cytosine
CM	Covariance Model
DIALIGN	DIAGONAL ALIGNment
DNA	DeoxyriboNucleic Acid
G	Guanine
International	Union of Pure and Applied Chemistry
MFE	Minimum Free Energy
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
nt	nucleotide
POL-II	POLymerase II
POL-III	POLymerase III
PSE	Proximal Sequence Element
RNA	RiboNucleic Acid
snRNA	small nuclear RNA
T	Thymine
U	Uracil

miRNA and miRNA* are not used as functional and non-functional mature sequences. It is used to differentiate between the two mature sequences of a pre-miRNA, despite the functionality.

Chapter 1

Introduction

In recent years, non-coding RNAs have gained attention, among other molecules, as key regulators of a plethora of different cellular functions and in diseases. Moreover, these RNAs have become key in understanding the evolution of the living things. After their own evolution in terms of classification, from junk to major functional players, the discovery of “functional” homologs of these RNAs has become worthwhile, although challenging task. The discovery process of these RNAs is called *Homology search*. However, the problem does not stop at the level of finding homologous, but also extends to the task of classifying novel RNAs, either by adding them into an existing family of RNAs or by creating a new class for them. In general, classes of these RNAs are defined by their sequence, structure and functional similarities. Detection of overlaps between the latter characteristics, requires highly sensitive methods and tools that can detect and predict, as accurate as possible, novel and homolog RNAs. So far, reaching maximum accuracy remains an open task, although many appreciable efforts have been conducted in this regard. Until then, data curation methods are a major part of the general homology search process. Unfortunately, sometimes old fashion curation, i.e., manual and semi-manual curation is still needed as an intermediate action in this whole process.

In this work I applied a new method related to the curation and homology search on two different ncRNAs classes, namely 7SK RNA and microRNA. The former works as a gene regulator at the transcriptional level (Ji et al., 2013), and the latter at post-transcriptional level (Cursons et al., 2018). More on the General biological and computational background is presented in [chapter 2](#). In [section 2.1](#), the biological background of non-coding RNAs and related processes is described. In the computational background ([section 2.2](#)) the common methods used in homology search in general are described, in addition to presenting the tools used in this work.

In [chapter 3](#), an iterative homology search was used including a new curation method applied during this research. 7SK RNA has been better studied in the vertebrates species than the invertebrates. However, a number of 7SK genes were identified in previous projects (Gürsoy, Koper, and Benecke, 2000; Egloff, Van Herreweghe, and Kiss, 2006; Gruber et al., 2008b; Gruber et al., 2008a; Marz et al., 2009). In contrast to the secondary structure of vertebrate 7SK RNA, the invertebrates structure had been only partially described. The previous efforts were extended by careful curation for the remaining

ambiguous parts of this structure, by dividing the general problem into sub-problems. This revealed more conservative parts in specific groups of the invertebrates species. The method described in the [chapter 3](#) revealed a novel stem in a specific invertebrate groups, similar to the conserved stem in vertebrates. Combining homology search approaches and the new careful curation method explained in the aforementioned chapter, the number of the putative 7SK genes was extended, in addition to a better description of their general structure.

Based on the aforementioned method, it was clear that the combination of homology search and curation leads to increased sensitivity compared to simple homology search. This is reflected in [chapter 3](#), in the extension of the 7SK genes, in addition to an increase of knowledge regarding their structures. Then, more comprehensive curation methods were applied to the microRNAs, as described in [chapter 4](#) and [chapter 5](#). The target was miRBase, the biggest available open source microRNA database. This database is managed by a group of researchers at *The University of Manchester*, providing organized data of pre-miRNA and mature sequences grouped into families. As a first step (see [chapter 5](#)), the aim was improving the seed alignments of the animal families of this database. The pre-miRNA sequences (precursors), are not defined by length as it varies between the families, even more, it varies between precursors of the same family. The target at this step was limited to a group of families. Those were only the families where a consensus length for a general alignment could be defined. Based on this consensus alignment, a number of sequences were modified leading to an improvement for a reasonable number of seed alignments of the target families. However, the cons of this method was the limitation to specific families and considering only a small number of problems related to the data curation and the homology search that appear for microRNA data. In addition, the false positive records included in this data were not checked. Nothing more was expected, as this was only initial work aimed to gain better quantitative and qualitative knowledge about the available microRNA data. However, this work did not only improve the seed alignments of the processed families, but also showed the feasibility of building a fully automated pipeline to improve this data and also that addressing the problems related to these annotated miRNAs can be a first step for successfully improving the homology search.

Eventually, ideas and developments covering major aspects of the here presented problems were implemented in a fully automated pipeline, **MIRfix**. In [chapter 6](#) I present a workflow that automatically curates miRNA datasets, acting on two levels, the precursors level and the alignment level. In general, **MIRfix** tries to correct false positive inputs in reference to their genomic sources, in addition to evaluating the consistency of the microRNA families by checking their alignments. The goal was to produce high quality alignments, that improve the homology search as well as producing a useful quantitative analysis of these data. This general workflow permeates useful features, e.g., predicting the miRNA and the miRNA* sequences for precursors without a

known miRNA, thereby considering important biological features. **MIRfix** was applied to all metazoan microRNA families available at miRBase release 21. As expected, the quality of the majority of the seed alignments was improved. Moreover, a number of precursors were corrected to show a better generic pre-miRNA structure. In [chapter 6](#), a qualitative and quantitative analysis of the results of the pipeline and miRBase data is shown. Additionally, a description of test results using an artificial data set, showing the accuracy levels of this pipeline, can be found there. The tests emphasize that **MIRfix** produces alignments that are comparable across families and sets the stage for improved homology search as well as quantitative analyses.

Chapter 2

Biological and Computational background

2.1 Biology

Cells are the fundamental unit of life in all living organisms, as they contain the main life components. These cells contain units of genetic information, called genes, which in turn are located on chromosomes. Chromosomes are made up of a chemical material, DesoxyriboNucleic Acid (DNA). The latter was defined in 1953 by James Watson and Francis Crick (Sussman, 2018), as a two-stranded double helix. DNA is made up of nucleotides Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). A DNA can be either coding for a protein or be non-coding. Both undergo a transcription process which leads to the production of a new molecule, RiboNucleic Acid (RNA), which is in turn either coding or non-coding. RNA is a single-stranded molecule made up of four bases A, C, and G, with T being replaced by another chemical compound Uracil (U) (Nature, 2014). Coding RNAs are known as messenger RNAs (mRNA) that are eventually translated into proteins, while non-coding ones function as RNAs.

Transcription process

All RNAs, coding as well as non-coding, are synthesized from a strand of DNA by the enzyme RNA polymerase. There are three types of RNA polymerases Pol-I, Pol-II, and Pol-III. In general, Pol-II is responsible for the synthesis of mRNAs (Soutourina, 2018). On the other hand, also some of the non-coding RNAs are synthesized by Pol-II, microRNAs for instance (Lee et al., 2004; Cai, Hagedorn, and Cullen, 2004). Other non-coding RNAs are synthesized by Pol-III like the 7SK RNA studied in this work (Krüger and Benecke, 1987; Murphy, Di Liegro, and Melli, 1987). Pol-II and Pol-III find their targets by recognizing stretches of DNA known as promoters. Pol-II promoters consist of 3 elements and a TATA box as shown in (Smale and Kadonaga, 2003) figure 2.1.

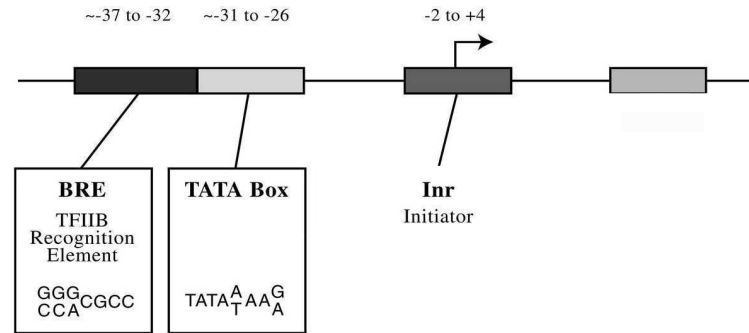


Figure 2.1: This figure shows the conserved sequences and positions POL-II promoters. BRE box is the TFIIB transcription factor binding site. TATA and Inr box act as transcription signals. Edited from (Smale and Kadonaga, 2003)

For Pol-III there are 3 different types of promoters known (Schramm and Hernandez, 2002), but here I focus on the type-3 which is related to the 7SK RNA project, as described in chapter 3. These 3 types are shown in figure 2.2.

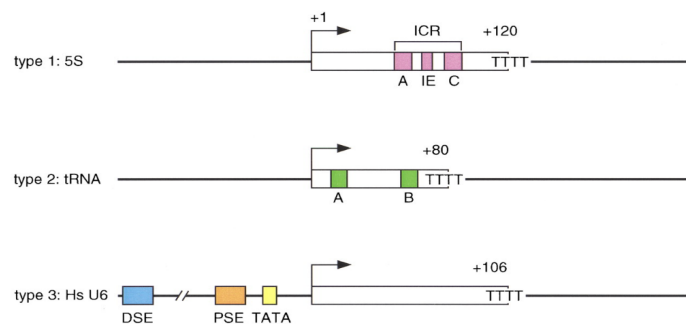


Figure 2.2: Comparing the three different types of POL-III. The first type contains the ICR (internal control region) which consists of 3 elements. A box, intermediate element (IE) and C box. Second type consists of two elements A and B boxes. Where the third type consists of 3 elements, distal sequence element (DSE) located at positions 215 to 240. Proximal sequence element (PSE) located at positions 65 to 48. A TATA box located at positions 32 to 25. Source: Edited from (Schramm and Hernandez, 2002)

2.1.1 Non-coding RNAs

The majority of RNAs in the genome are non-coding RNAs (Cheng et al., 2005; Consortium, 2004) that play important regulatory roles in a cell. There are many different types of regulatory ncRNAs (e.g., long non-coding RNAs, microRNAs (miRNAs), small nuclear RNAs (snRNAs)). snRNAs play two main regulatory roles, once at transcriptional and then at the post-transcriptional level. snRNAs that play a role at the post-transcriptional level, fold onto themselves, forming a distinct secondary structure (Corbett, 2018; Quinn and Chang, 2016). The two main post-transcriptional roles act at pre-mRNA processing, like U7 RNA (Godfrey et al., 2006) and the regulation of transcription factors like 7SK RNA (Peterlin, Brogie, and Price, 2012). Even though

microRNAs are small sequences which do not fold and form large stable secondary structures, they are processed from hairpin shaped structures called pre-miRNA.

2.1.2 RNA secondary structure

Many ncRNAs are known to form distinct secondary structures, rendering structure an important feature for detection and annotation of ncRNAs. It is known that A forms two hydrogen bond with T and C forms three hydrogen bonds with G, making a C-G basepair more stable than A-T or A-U in case of RNA. Furthermore, in RNA Uracil can pair with A and G. Moreover, these structures can form also a tertiary structure, e.g., when a single strand of RNA folds onto a duplex or interacts with other molecules of RNA. A generic hairpin loop is defined by two helices and a loop. Bulges are none base-pairing nucleotides located in a stem. These structures are important for the function of RNA as they work as binding sites to other factors or influence the direct interaction with their targets. For instance, figure 2.3 shows the function of the U7 RNA which plays the main role in the histone pre-mRNA 3' end processing. The 5' end of the U7 RNA binds to what is called *HDE* of the histone pre-miRNA (Marzluff and Duronio, 2002). At the same time, SM protein binds to a specific highly conserved motif upstream the stem-loop formed at the 3' end of the U7 RNA (Skrajna et al., 2017). Moreover, the stability of these structures is always important as it directly influences which stretches of RNA are available for interaction.

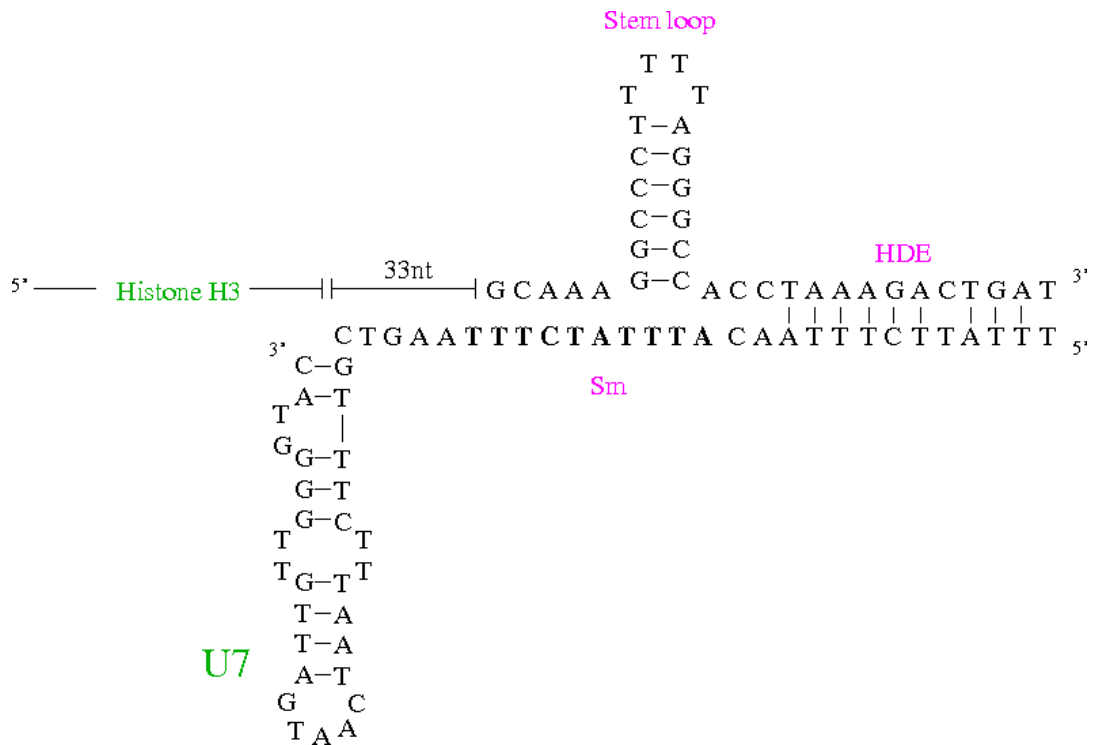


Figure 2.3: U7 RNA (bottom) binds to the histone downstream element (HDE, found the histone coding genes instead of the poly A tail in other coding genes) at the 3' end of a histone pre-miRNA. The sm region is the conserved region where the SM proteins bind. Source: (Marz and Stadler, 2011).

2.1.3 Homology versus similarity

Are the terms “similar” and “homologous” synonymous? The answer is no. In biology and computational biology, the two terms are not the same. Two DNA/RNA sequences can be similar, but they do not necessarily have to be homologs, as in the convergent evolution for instance. The similarity between two sequences is usually represented only at a single (usually the sequence) level. On the other hand, two sequences can be homologs even if they are not recognizably similar, simply because they have diverged so far, that no residual similarity is left. For example in microRNAs, two hairpin shapes not very similar at the sequence level can still share a very similar structure and a similarity in specific regions where the microRNA are located. Such two sequences are homologous even if they do not have a perfect similarity at the sequence level. In other words, similar sequences are homologous when their similarity is informative, and when this similarity is not by chance.

2.1.4 Evolution

According to Charles Darwin’s definition of evolution, all living organisms have descended from a common ancestor. Over time, all generations were subjected to changes leading to the development of novel characteristics.

During evolution, these characteristics were gained by some individuals but not others. These changes can be seen as modifications in the DNA sequences. Such modifications are called “mutation” can be a cause of internal reasons, e.g., during the DNA replication or external sources like the effects of the UV sunlight (Gruijl, Kranen, and Mullenders, 2001). Sometimes changes are only temporary, affecting only a few individuals, but sometimes such changes can become stable integrated in the genome of a species, being passed down to future generations or even leading to the evolution of a new species.

Mutation types

The mutation types mentioned here are only related to the topic at hand, which is related only to the DNA/RNA changes and not the protein products.

- Substitution, which is exchanging a nucleotide with another one. e.g., change of the sequence ACGTTGC to AGGTTGC
- Insertion, when one or more nucleotides are inserted in the sequence. e.g., change of the sequence ACGTTGC to ACCTTGTTGC
- Deletion, when one or more nucleotides are deleted from a sequence. e.g., ACGTTGC to ATGC

Suppose a mutation of DNA sequences of a given gene lead to modification of specific characteristics, is inherited by two different groups of descendants but not by any other. This means that these two groups experienced more mutation in the same DNA sequence, and this will be passed down to the next generation of each of them. This will eventually decrease the similarity between the DNA sequences of a given gene, but will still be sharing its original functionality. Considering the 3 nodes A, B, and C in figure 2.4 as; A) is the ancestor of two different species B) and C). For simplicity, suppose these sequences are two DNA sequences of a given gene. Under environmental or internal conditions, the same DNA sequence was mutated in both groups and inherited to their next generations. As shown in the figure, the sequences are not exactly the same but similar, however, they still share the same function in both species. For example, if we suppose this is the body hair gene, during evolution the modification lead to less hair in humans than the Chimpanzees. This scenario can continue over time until this function is lost or kept, gaining characteristics or retains the same characteristics. But so far, humans are still going through “machinery” evolution by using shaving machines.

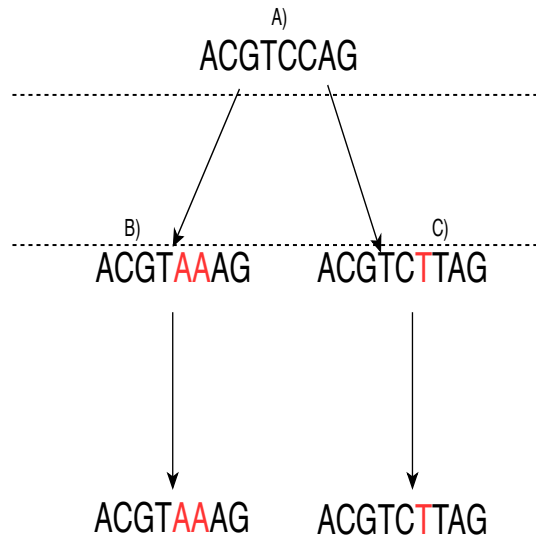


Figure 2.4: Two different species at nodes B and C of a common ancestor A. The species B) and C) inherited the same sequence from A) which was then mutated and passed on to the next generation.

2.2 The role of computational biology

In order to handle biological data and to understand the underlying relations, computer science, and mathematics have been integrated into these studies. Many tools and methods have been and are currently being developed, and until today these tools and methods are gaining more and more importance in solving problems in this context.

2.2.1 Alignment

In principle, alignment means to align a word or a sentence to another finding the most similar regions between them. In the biological context, this is applied to DNA, RNA as well as protein sequences. The main idea is to find similarity between these sequences in order to understand their evolutionary relationship and also the connection between sequence similarity and functional similarity.

2.2.1.1 Pairwise alignment

Local alignment which is a method first proposed by Temple F. Smith and Michael S. Waterman, hence known as Smith-Waterman algorithm, tries to find the best similar regions between two sequences regardless if it is from the first position to last or not. On the other hand, the global alignment which was introduced as Needleman–Wunsch algorithm by Saul B. Needleman and Christian D. Wunsch is based on aligning the whole sequences from the first position to the last one. Common use cases for both methods differ, based on

the general knowledge of the sequences that need to be aligned. In general, the global alignment is better when the sequences are almost equal in length and when we know that they have a reasonable similarity. Local alignment is used in the case when we do not see an obvious similarity for the whole, but reasonable similarity between sub-strings of the two sequences. The alignment in both methods is based on scores, defined for each state of each nucleotide in the sequence pairs. The 3 common states are match, mismatch or gap. The match state is when two nucleotides are equal, mismatch when they are not, and a gap is introduced to account for insertions and deletions (InDels). Usually, a match is given a positive score and mismatches and gaps are given negative scores.

Local alignment

Since local alignments focus on specific regions of a sequence, their score is always starting from zero, resetting negative scores from upstream sub-alignments. This guarantees that always optimal sub-alignments can be found, even if the upstream sequences do not align well. In other words, whenever a bad score results from mismatches and/or gaps, the algorithm starts again from zero allowing to aligning specific regions in the sequences. The following is an example of a local alignment with the Smith-Waterman scoring method.

Sequence x=GGACCTATGCTTGG

Sequence y=ACTAGG

$$F(i, j) = \max \left\{ \begin{array}{l} 0 \\ F(i-1, j) - G, \\ F(i, j-1) - G, \\ F(i-1, j-1) + s(x_i, y_j) \end{array} \right\}. \quad (2.1)$$

Where \mathcal{F} is the score of a given cell in the scoring table, \mathcal{G} is the gap score and s is the score of a match or a mismatch. If the match score is 1, mismatch is -1 and gap score is -2. The table will be:

<i>S</i>		A ₁	C ₂	T ₃	A ₄	G ₅	G ₆
	0	0	0	0	0	0	0
G ₁	0	0	0	0	0	1	1
G ₂	0	0	0	0	0	1	2
A ₃	0	1	0	0	1	0	0
C ₄	0	0	2	0	0	0	0
C ₅	0	0	1	1	0	0	0
T ₆	0	0	0	2	0	0	0
A ₇	0	1	0	0	3	1	0
T ₈	0	0	0	1	1	2	0
G ₉	0	0	0	0	0	2	3
C ₁₀	0	0	1	0	0	0	1
T ₁₁	0	0	0	2	0	0	0
T ₁₂	0	0	0	1	1	0	0
G ₁₃	0	0	0	0	0	2	1
G ₁₄	0	0	0	0	0	1	3

Figure 2.5: The scoring matrix of Smith-Waterman alignment for the example sequences *x* and *y*. The colored cells show the trace-back of a path of one solution of three different solutions in this matrix (The other solutions can be determined by tracing back from all other cells of score equal to 3). (Raden et al., 2018b; Raden et al., 2018a)

From the final matrix, the best alignment can be reconstructed following the path defined by the alignment, a method known as traceback. Two cells can be connected when the score of one of them is calculated from the other cell, by adding the match/mismatch score or subtracting the gap score. The cells are always calculated from the above, left or upper diagonal cell, so only these cells are checked in the traceback. The latter starts from the highest score in the table, wherever this score is, until it reaches zero, which indicates the start of a new local alignment. Moving horizontally means no gaps are introduced, and moving vertically means inserting a gap at a given position in the sequence *x* or *y*. This produces more than one possible alignment, for example, one of them is:

```

A C T A
C C T A
- * * *
```

The stars in the above alignments are matches, the dash represents non-matching nucleotides (a gap), and the result is an alignment of specific regions of the two sequences to each other without considering the up-/downstream parts of the sequences.

Global alignment

This method is used to align whole sequences together. Here, zero is only used to initialize the table at the beginning and negative values within the alignment are not reset. The general equation is:

$$S(i, j) = \max \left\{ \begin{array}{l} S(i-1, j) - G, \\ S(i, j-1) - G, \\ S(i-1, j-1) + s(x_i, y_j) \end{array} \right\}. \quad (2.2)$$

The traceback of the alignment from the table starts always from the last cell (bottom right) to the first cell (upper left), as these positions correspond to the end and the beginning of both sequences. Figure 2.6 demonstrates an example of a scoring matrix in a global alignment, for the same sequences used in the previous example of the local alignment.

<i>D</i>		A ₁	C ₂	T ₃	A ₄	G ₅	G ₆
	0	-2	-4	-6	-8	-10	-12
G ₁	-2	-1	-3	-5	-7	-7	-9
G ₂	-4	-3	-2	-4	-6	-6	-6
A ₃	-6	-3	-4	-3	-3	-5	-7
C ₄	-8	-5	-2	-4	-4	-4	-6
C ₅	-10	-7	-4	-3	-5	-5	-5
T ₆	-12	-9	-6	-3	-4	-6	-6
A ₇	-14	-11	-8	-5	-2	-4	-6
T ₈	-16	-13	-10	-7	-4	-3	-5
G ₉	-18	-15	-12	-9	-6	-3	-2
C ₁₀	-20	-17	-14	-11	-8	-5	-4
T ₁₁	-22	-19	-16	-13	-10	-7	-6
T ₁₂	-24	-21	-18	-15	-12	-9	-8
G ₁₃	-26	-23	-20	-17	-14	-11	-8
G ₁₄	-28	-25	-22	-19	-16	-13	-10

Figure 2.6: The scoring matrix Needle-Wunsch alignment for the same sequences used above. As shown in the highlighted cells, the traceback starts from the end to the first cell. (Raden et al., 2018b; Raden et al., 2018a)

One of the results in this alignment is:

G	G	A	C	C	T	A	T	G	C	T	T	G	G
-	-	A	C	-	T	A	-	-	-	-	-	G	G
		*	*		*	*						*	*

Here we see more gaps than for the local alignment, since global alignment tries to find the best possible solution in regard to all the nucleotides in the alignment, from the first nucleotide to the last one. As we see in the table in the lower right cell, the score of this alignment was equal to -10 and thus not good.

2.2.1.2 Multiple sequence alignment (MSA)

A multiple sequence alignment (MSA), in general, is the alignment of more than two sequences to each other. i.e., Finding the most similar regions between a number of sequences. Such similar regions known as conserved regions can be revealed via MSA. The importance of MSA is in detecting the

evolutionary relationships between different sequences. These sequences can be DNA, RNA or protein sequences. Most of the time, the conserved regions of DNA, RNA or protein are related to some function or even to specific diseases. Although there are mutations without any appreciable effects, other mutations can as well lead to specific diseases. For instance, the mutation in the *PRPF31* gene leads to the retinitis pigmentosa (Xiao et al., 2017). In such cases aligning multiple sequences from normal samples to infected samples, can help to reveal the changes which lead to the disease or the disorder.

Different methods and tools have been established to perform MSAs. In my project I mainly used `clustalw` (version 2.0.12) (Thompson, Higgins, and Gibson, 1994) and the anchored aligning tool `DIALIGN` (Morgenstern et al., 2006; Morgenstern, Dress, and Werner, 1996). `T-coffee` (Karmakar et al., 2017) and `Muscle` (Edgar, 2004) are examples for other, well established MSA tools, but were not used in this work due to the speed advantage of `clustalw`.

Clustalw

`Clustalw` considers the evolutionary relationship between the sequences by producing a guide tree before aligning the actual sequences. There are 3 main steps performed by `clustalw` to find the final alignment of a given number of sequences Chenna et al., 2003. First, a pairwise alignment is applied for every set of two sequences, and a distance matrix is set up based on the resulting scores. Based on this distance matrix, a guide tree is calculated following the neighbor-joining method (Saitou and Nei, 1987). And as a final step the sequences are progressively aligned in the same order in which they are present in the guide tree.

DIALIGN

The name `DALIGN` comes from the two words `DIAGONAL` and `ALIGNMENT` because it diagonally aligns fragments of equal lengths. Two fragments are similar subsequences of two different sequences. Unlike other tools which use a substitution table, `DALIGN` depends on the local sequence similarity by comparing the P-value of the fragments after weighing all the possible fragments between the sequences. The weights are maximized at the end to find the best possible alignment. In `DALIGN 2` the weighting matrix was changed to decrease the noise produced by the short fragments (Morgenstern, 1999; Morgenstern, 2002). Moreover, `DALIGN` introduced later what is a so-called *anchored alignment* (Morgenstern et al., 2006). The idea is to give an option for the users to force aligning specific regions, making use of information available to the user but not to the program. This option was useful in my microRNA project as the alignments were anchored at the mature sequences which are parts of the pre-miRNA with a higher similarity than other parts of the precursor. Moreover, the precursors are determined in relation to the positions of the mature sequences.

2.2.2 Homology search

Homology search is the process of finding an informative similarity between sequences in different genes and genomes. In simple words, it is finding a query sequence or a similar sequence in a target genomic database. Two main categories can be introduced in this area: *sequence based search* and *structure based search*.

As the name implies, sequence-based search depends only on the sequence information, e.g., in DNA sequences only the nucleotides are considered. On the other hand, structure-based search, considers also the structural information of the sequences which fold and form secondary structures.

2.2.2.1 Sequence-based

BLAST

The most widely used sequence-based alignment tool today is BLAST (Altschul et al., 1990; Camacho et al., 2009). Blast uses a heuristic algorithm, offers different types of searches based on the type of the query and target sequences, for instance *blastx*, *blastp* and *blastn*. *blastn* is used to search nucleotide sequences and the one algorithm used in my projects. It searches the subject (target) sequences with fixed length segments from the query sequence. Each segment is a window of default size 11 for nucleotide search and 3 for protein search, moved by stepsize 1 along the query sequence. These segments are then compared to the target sequences and scored using a substitution matrix. Sequences with scores below a specific threshold are ignored in the next step. In the last step, all possible matches are extended from both directions and the extended regions are aligned to the target sequence. The extension stops when the calculated score starts to drop again or when the sequence is completely aligned.

2.2.2.2 Structure-based

Structure-based search tools evaluate structural and sequence information together. Different tools are available to find novel structure-functional ncRNAs or to detect already known ncRNAs in new genomes. Various approaches and methods were applied during this work, from using a defined descriptor, sequence alignment and consensus secondary structure to using hidden Markov models and study the covariation of a sequence alignment.

RNAmotif (Macke et al., 2001) was first introduced as *rnabob* (Eddy, 1996). This tool works with a user-defined descriptor. In the descriptor, a sequence and the corresponding structure are defined via the helix/helices and the loop(s) that make up the structure, as well as a definition of nucleotides and their variation. The tool is for detecting highly conserved motifs, allowing variation for the rest with applying specific scoring methods to rank hits and matches.

The program `Erpin` (Gautheret and Lambert, 2001) does not use descriptors. Instead, it uses an alignment of sequences and its secondary structure to infer a statistical secondary structure profile (SSP). For example in hairpins, two profiles are created, a helix profile (for the stems) and a single-strand profile for the loops. Using these profiles, it searches the target database and constructs a square matrix which is a dynamic programming matrix. In the end, it computes E-values for the matches.

The main obstacles concerning homology search in the here described and other tools which use similar approaches, are limitations of the search or sensitivity. For example, when designing a specific descriptor we are forcing the search to find very specific sequences or variation, which might lead to removing a hit with one or a few mismatches. In `Erpin` for example, it only allows gaps when using the helix profile, while they are not allowed in the single-strand profile (loop regions), which is not conform regarding means of variation and covariation that are considered in the homology search. To overcome this, *hidden Markov models* are used e.g., in `infernal` (Eddy, 2002; Nawrocki and Eddy, 2013), a tool applied in this work.

Infernal

`Infernal` is a software package which produces significant covariance models from multiple alignments with consensus secondary structure. This tool evaluates the covariation of each residue in the sequences, and it deals with all the insertions and deletions that can occur in a sequence or set of sequences. Thus, the power of this tool lies in its flexibility, considering mutation events over the distribution of nucleotides and dealing with insertions and deletions related to the consensus of the alignment. Of course, this tool works at both levels: sequence and structure. From a reasonable number of aligned sequences, the CM model can learn for each given position, if the residue or a base pairing nucleotide is highly conserved, supported by evolutionary relation or if it is a non-frequent insertion or deletion. `Infernal` consists of a collection of tools that were all used to more or less extend during this work. A CM model is built with the command `cmbuild` requiring two parameters, the alignment input file, and the output file. Before searching the genomes, the CM needs to be calibrated with `cmcalibrate`, which calculates the exponential parameters related to the E-value calculation of the model. The genomes are searched with a calibrated CM via `cmsearch`, which returns hits ranked by E-value (where possible). Moreover, with the `cmalign` command it is possible to align a number of sequences using a given CM. With `cmScan` a number of sequences can be searched against a number of covariance models. A big collection of RNA family covariance models are available at the `Rfam` database; the number of these families reached 2687 in the latest release `Rfam` 13.0 (Kalvari et al., 2017; Kalvari et al., 2018).

2.2.3 RNA secondary structure prediction

A secondary structure is not measured only in terms of the number of base pairs, but also according to its thermodynamic stability. The latter is a measure for the amount of energy needed to build/melt a double-stranded DNA or RNA. The first RNA secondary structure prediction tools started without considering the thermodynamical energy as for example (Nussinov et al., 1978) which focused on maximizing the number of base-pairing nucleotides. Later, utilizing dynamic programming and experimental data, the thermodynamics of RNA folding was introduced by Michael Zuker and Patrick Stiegler in `mfold` (Zuker and Stiegler, 1981; Zuker, 2003). The Zuker-Stiegler method was also implemented in `RNAfold` one of the components of the `ViennaRNA` package (Hofacker et al., 1994; Lorenz et al., 2011). Recently, `C o F old` (Proctor and Meyer, 2013) was developed, a tool also implementing Zuker-Stiegler thermodynamics methods and based on the `RNAfold` algorithm. `C o F old` additionally takes the kinetics of folding into account. However, compared to the implementation in the `ViennaRNA` package, there are no substantial differences in the predicted structure as shown in figure 2.7. The line containing dots and brackets represents the predicted structure of the sequence, ending with a number in the brackets. This number represents the minimum free energy (MFE) calculated for the structure. Each left bracket "(" represents a nucleotide base-pairing to its corresponding nucleotide, denoted by a right bracket ")", where The dots are non-base-pairing nucleotides.

```
GGGCUAUUAGCUCAGUUGGUUAGAGCGCACCCUGAUUAGGGUGAGGUCGCGAUUCGAAUUCAGCAUAGCCCA
(((((((..((((.....))))).((((.....))))).(((.....)))))))). (-28.90) --- RNAfold from ViennaRNA
(((((((...((((.....))))).((((.....))))).(((.....))))).))))). (-29.40) --- C o F old
```

Figure 2.7: This figure compares the RNA secondary structure predicted by `RNAfold` from `ViennaRNA` and `C o F old`. The first line is the RNA sequence input for both tools. The second and the third line are the structures calculated by `RNAfold` and `C o F old`, respectively. The numbers in brackets are the minimum free energy (MFE) (Unit: kcal/mol) calculated for each of them.

This work is using many of the `ViennaRNA` package components. The latter also makes it possible to find the consensus secondary structure of an aligned sequence using `RNAalifold`. Moreover, in both `RNAfold` and `RNAalifold` it is possible to force specific nucleotides to base pair together using the new constraint folding framework (Lorenz, Hofacker, and Stadler, 2016).

Chapter 3

Careful curation for snRNA

3.1 Biological background

The 7SK snRNA is a mammalian Pol-III transcript with a typical length of about 330nt (Krüger and Benecke, 1987; Murphy, Di Liegro, and Melli, 1987). Small nuclear RNAs, most of which are Pol-III transcripts, contain two main elements; A *PSE* and a *TATA* box (Dergai et al., 2018). In U6 snRNAs which are also Pol-III transcript, the *TATA* boxes are located around 22-23 nucleotides upstream of the initiation site, with a 13 nucleotide spacer between the *PSE* and *TAT* box. The 7SK RNA sequence is very well conserved in vertebrates (Gürsoy, Koper, and Benecke, 2000; Egloff, Van Herreweghe, and Kiss, 2006), but shows a high level of variability in invertebrates (Gruber et al., 2008b; Gruber et al., 2008a). Due to its abundance, it has been known since the 1960s. The molecule is capped at its 5' end by a highly specific methylase MePCE, also known as BCDIN3 (Jeronimo et al., 2007). Its stability is regulated by means of a highly specific interaction with LARP7 (La-related protein 7, also known as PIP7S) (He et al., 2008; Krueger et al., 2008; Markert et al., 2008; Eichhorn, Chug, and Feigon, 2016). LARP7 also plays an inhibitory role counteracting MEPCE; to this end, its xRRM domain binds to the 3' stem-loop of the 7SK RNA (Brogié and Price, 2017). The Bin3 RNA methyltransferase reinforces LARP7 binding. It adds to the stability of the interaction (Xue et al., 2009) and catalyzes the 5' methylation that is protective against degradation (Cosgrove et al., 2012).

The primary function of 7SK is to mediate an inhibitory interaction of the HEXIM1 protein with the general transcription elongation factor P-TEFb, thereby repressing transcript elongation by Pol II (Michels et al., 2004; Blazek et al., 2005; Egloff, Van Herreweghe, and Kiss, 2006; Peterlin and Price, 297-305). To activate P-TEFb it must be released from the complex with 7SK RNA (Chen et al., 2008). This process is facilitated by the PPIM1G phosphatase, which binds then to 7SK RNA and prevents re-binding of P-TEFb (Gudipaty et al., 2015).

7SK RNA suppresses the deaminase activity of APOBEC3C and sequesters this enzyme in the nucleolus (He et al., 2006). It features at least two distinctive secondary structure elements: Both HEXIM1 and P-TEFb bind specific sequence motifs at the 5'-terminal hairpin, while a 3'-terminal hairpin interacts with P-TEFb only. The high level of sequence conservation in gnathostomes contrasts with highly divergent 7SK sequences in invertebrates, however.

Still, high levels of sequence conservation have been reported for the 3'- and 5'-hairpins, suggesting that the protein interaction sites are likely homologous to the ones of vertebrates.

Two highly conserved GAUC motifs located in the upper part of the 5' hairpin form a short helix and, together with adjacent, conserved U bulges form the HEXIM1 binding site (Lebars et al., 2010; Egloff, Van Herreweghe, and Kiss, 2006). This short helix is located between the nucleotide number 40 and 60 in figure A.1, this figure represents the human 7SK RNA from (Egloff, Van Herreweghe, and Kiss, 2006). The stem-loops in order from left to right are 5' stem, stem A, stem B and 3' stem. A recent crystal structure (Martinez-Zapien et al., 2017) of the 5'-terminal hairpin shows that the major groove of the GAUC-GAUC helix is occupied by the adjacent uridines, which form non-standard base pairs with nucleotides of the helix. This structural element is also the main interaction site for PPM1G (Gudipaty et al., 2015). This motif also acts as a binding site for the HIV-1 transcriptional transactivator (Tat) protein (Muniz et al., 2010), which can cause chemical variations in some residues of the GAUC-GAUC helix (Bourbigot et al., 2016). The 7SK RNA is able to form alternative structures under the influence of different factors (Bourbigot et al., 2016). The presence of the GAUC-GAUC helix, for example, is magnesium-dependent (Brogie and Price, 2017). Finally, 7SK RNA plays a role in cancer by negatively regulating (Eilebrecht et al., 2010; Eilebrecht et al., 2011) the attenuating HMGA1 protein, which is strongly overexpressed in several cancers (Masciullo et al., 2003; Fusco and Fedele, 2007; Eilebrecht, Benecke, and Benecke, 2011).

3.2 Introduction to the problem

The main problems faced when performing homology searches to detect new genes in invertebrates 7SK snRNAs, can be summarized in two points: When the homology sequence-based search converges to a point that no more hits can be detected, or when the structural information is not sufficient to build models such as covariance models which search both, sequence and structural information. To-date 7SK RNAs have been identified in a wide range of bilaterian animals. While their sequence and structure is well conserved in vertebrates, it shows large changes in size, sequence, and secondary structure across the invertebrate 7SK genes identified so far (Gürsoy, Koper, and Benecke, 2000; Egloff, Van Herreweghe, and Kiss, 2006; Gruber et al., 2008b; Gruber et al., 2008a; Marz et al., 2009). In these respects, invertebrate 7SK behaves similarly to the vertebrate telomerase RNA component (Xie et al., 2008). It is unclear at this point whether 7SK RNA is a bilaterian innovation, or whether homologs in diploplasts or even outside the animal kingdom just have escaped detection. For example, the putative, highly derived gene reported for *Caenorhabditis* species (Marz et al., 2009) (also named T26A8.6 and CeN21-2) may in fact be a homolog of the U8 snoRNA (Hokii et al., 2010). The main purpose of this contribution is to provide an update of the sequence and secondary structure information available for invertebrates. To this end, a comprehensive homology search was conducted. Using a greatly enlarged

set of credible 7SK genes a method was applied to construct high-quality sequence alignments and detailed structural models that allowed us to trace in full detail the evolutionary history of this RNA family.

3.3 Methods

3.3.1 Initial seeds and models construction

First data origin

Candidate search began with the previously annotated invertebrates 7SK RNAs. From (Gruber et al., 2008a) the sequences of the species from the following groups were used, Apocrita: *Apis mellifera* (western/ europe honey bee) and *Nasonia vitripennis* (wasp). Coleoptera: *Tribolium castaneum* (beetle). Diptera: *Aedes aegypti* (mosquito), *Culex pipiens* (mosquito), *Anopheles gambiae* (mosquito), and *Drosophila* (Fruit flies). *Ixodes scapularis* (known as ticks) from Arachnida and *Pediculus humanus corporis* (louse) from Neoptera. From (Gruber et al., 2008b) *Lottia gigantea* (limpet) which belongs to the Gastropoda in Mollusca. From Orthogastropoda in Mollusca, *Helix pomatia* (limpet) and *Aplysia californica* (sea hare). From Annelids group, *Capitella capitata* (worm) and *Helobdella robusta* (leech). The genomic data was downloaded from NCBI databases (Wheeler et al., 2007), the antgenomes database website (<http://antgenomes.org/downloads/#category-view>) (Wurm et al., 2009), and from VectorBase database website (<https://www.vectorbase.org/downloads/>) (Giraldo-Calderón et al., 2014) at different times during 2017.

Homology search and building covariance models

The majority of the annotated invertebrates 7SK RNAs that were collected from the previous related work was from the Ecdysozoa groups, and mainly from the Endopterygota. Because of the availability of these sequences, it was more realistic to start the homology search from these sequences than starting from other sequences, for instance, the ones of the Mollusca group. The general workflow was, first run a homology sequence-based search (`blastn`) and when hits were found proceed to `cmsearch` and run the covariance model of the existing putative 7SK candidates against the genomes or the contigs of the hits that were found. When the sequences were similar to the known (previously annotated) sequences, which means they fit to the consensus secondary structure of the existing sequences as well as that they are similar at the sequence information. These sequences were added to the previous sequences and the whole set of sequences was re-aligned and new models were built. This step was repeated iteratively until the search converged and no more good hits were detected.

Endopterygota

The most relevant first hits were found using the sequences of Apocrita species as queries against the *Whole genome shotgun database* (WGS) available at NCBI. The start was with the sequences of Apocrita species to `blast` against the WGS database and 3 candidate species were detected, *Bombus impatiens* (Accession number: AEQM02001933, coordinates: 40540-40857), *Solenopsis invicta* (Accession number: AEAQ01012710, coordinates: 19471-19768) and *Pogonomyrmex barbatus* (Accession number: ADIH01016463, coordinates: 238-550). Notably, these sequences were detected from *Nasonia* sequence in particular. These 3 sequences and the *Nasonia* were then added together and aligned using `clustalw` (version 2.1) with manual editing based on the general structure provided by (Gruber et al., 2008a). For the manual curation the `emacs` text editor in `ralee` mode (Griffiths-Jones, 2005) was always used. The 3 new candidates and the *Nasonia* were then used as queries for a new `blast` search against the WGS species in NCBI. For reasonable hits (covering at least 80% of the query sequence), the related genomic sequences were downloaded and searched (`cmsearch` command) with the covariance model built for the 4 query sequences, which was the first covariance model built.

Two additional sequences were detected showing sequence and secondary structure similarity, one from *Linepithema humile* (Accession number: ADOQ01006353, coordinates: 44230-44538) and one from *Atta cephalotes* (Accession number: ADTU01023827, coordinates: 11509-11816). From these sequences, the 5 new candidates were annotated and added to the 3 known sequences *Apis mellifera*, *Nasonia vitripennis* and *Tribolium*, and a new alignment was calculated, followed by building a covariance model of these sequences after the alignment was manually refined. The same step was repeated (`blastn-cmsearch`) and a new candidate was detected, the *Acromyrmex echinator* (Accession number: AEVX01005901, coordinates: 226775-226466). In the end 14 more sequences having sequence and structure conservation could be identified. For a full list of searched genomes, stating where significant hits were found and where not, please refer to tables A.1, A.2 and A.3.

This step was important since new candidates in new groups (Orussoidea, Tenthredinoidea) and also one candidate outside of the Endopterygota group (*locusta migratoria* from the Orthoptera group) could be identified.

All the previously mentioned sequences (candidates and known ones) were added together into one alignment as in the Endopterygotas sequences, and a "`blastn-cmsearch` search was applied to the nucleotide collection and WGS databases available at NCBI. Nine more sequences were detected this way, increasing the total number of the putative 7SK RNAs in Endopterygota into 30 sequences. At this step, a highly likely 7SK RNA candidate was found in the ant *Metapolybia cingulata* (Accession number: KM421505.1, coordinates: 996-673) from the nucleotide collection database, see 3.4.3.

Diptera

The Diptera contained 3 non-Drosophila sequences and 12 from Drosophilas. The main problem here was the big difference in length between the Drosophila species and the rest. The alignment of all known Diptera sequences did not lead to any clear structural annotation, eventually producing bad alignments. To overcome this, sequences were divided vertically in different ways based on the best-aligned parts and knowledge about the known stems (5' stem, stem A and 3' stem) which lead to 7 sub-alignments. The sub-alignments were glued together to produce better alignments. A covariance model was built for these sequences with the new alignment, and the procedure was applied again, blasted against different databases, the nucleotide collection database and WGS, this time including the RefSeq Genome database with the expectation threshold $E=100$. 7 hits were detected with `blastn`, and 6 of them gave reasonable hits after searching their genomes with the initial Diptera covariance model. After manual filtering, only 3 sequences were left at this level and two were applied later to the Arthropoda model. The 3 candidates at this level were: *Mochlonyx cinctipes* (Accession number: JXPH01043687, coordinates: 608-303), *Ephydra gracilis* (Accession number: JXPQ01011706, coordinates: 27681-27235) and *Scaptodrosophila lebanonensis* (Accession number: 35006-35493).

Towards a full Arthropoda model

The previous steps and the newly detected candidates showed weakest conservation between the species in Endopterygota. This led to the expectation that this conservation could be retained to appear at higher levels of the phylogenetic tree. This conservation was expected at Arthropoda as it is one of the biggest groups which descend directly from Ecdysozoa and the aforementioned models (Endopterygota and Diptera) both belong to this group.

The first challenge was merging the different Diptera species (Drosophila and non-Drosophila), and the second one was, merging Diptera sequences to the other Endopterygota as the main step to finding general Arthropoda model with a reasonable sequence and structural conservation. The scanning feature (`cmscan`) from `infernal` was used to merge the two models and the Diptera sequences were scanned with the Endopterygota model, then the Endopterygota with the Diptera model. The latter gave better structural alignment, so the Endopterygota sequences were merged to the Diptera model. Additionally, two candidates (*Phortica variegata* and *Glossina morsitans*) were added from Diptera which were detected in the previous step but not added because of lower similarity to the model than other sequences, and the *Ixodes scapularis* from Arachnida.

The result was used as a first Arthropoda model, to find more candidates in different Arthropoda groups. Indeed, a sequence was detected in Coleoptera (*Anoplophora glabripennis*) and another candidate was removed later from

the model (*Priacma serrata*). Three more candidates were also detected and added to this model initially and used in the search, but two of them were removed later from the final model when the general consensus structure of the Arthropoda was defined. The sequences which were removed at the end are *Trichuris suis* and *Ladona fulva* from Nematoda and Odonata, respectively. The sequence which was left in the final model was from the Ephemeroptera group. These sequences were removed because they either did not show any of the novel discovered stem B or extensions in the stem A and moreover, the Nematoda candidate was different even at the base stem of the general structure.

The first model used here was based on the sequences from the work of (Gruber et al., 2008b) containing *Capitella capita*, *Lottia gigantea*, *Helix pomatia*, *Helobdella robusta* and *Aplysia californica*. After 3 iterations of `blastn-cmsearch`, 3 more candidates were detected in the Mollusca group, *Biomphalaria glabrata*, *Lingula anatina* and *Crassostrea gigas*.

Pol III promoter in the candidates

As mentioned in the introduction, 7SK genes are Pol-III transcripts, suggesting that the upstream region of the 7SK RNA must contain the characteristics of the Pol-III promoter. The U6 RNAs are also Pol-III transcripts (Dergai et al., 2018) so the promoters from (Hernandez Jr, Valafar, and Stumph, 2006) were used as a reference to investigate the upstream of the candidates for these promoters. The 100 nucleotides upstream of the candidate sequences were extracted and searched these sequences for a Pol-III promoter similar to that of U6 promoters.

3.3.2 Models anatomy then merging

Besides using existing bioinformatics tools, manual curation was necessary in order to reach the best possible structural annotation for the groups of the invertebrates species and generic invertebrates model. To this end, the alignments were divided horizontally and vertically (Figure 3.1 shows the vertical split). The horizontal division takes each group alone, hence, working with each group separately. The groups are Apocrita, Diptera, Coleoptera and Tenthredinoidea. For each of these 4 groups, a vertical division was applied with a slight difference in the Diptera sequences.

Vertical division in all groups

The 5' stem and the 3' stem are almost perfectly conserved between all species, but the part in between was adding noise to the alignments. This part, from now on referred to as "noisy" part, includes the stem A as well. For each of the groups, this part was cut out the alignment, and re-aligned alone using `clustalw` (v 2.1) in addition to calculating the secondary structure of these alignments. These *noisy* parts were then treated as independent models and a

covariance model was created for each of them.

Diptera sequences went through more anatomical processing because of the diversity in these sequences, especially in the region between the 5' and 3' end which I also refer to as *noisy*. A pairwise alignment was performed between the *noisy* regions of each two sequences. In other words, for each of the two sequences, their regions were aligned independently of the other sequences. This step was applied in order to group the most similar sequences together. This process ended with dividing the Diptera sequences into two groups: The first group contained *Drosophila*, *Scaptodrosophila lebanonensis*, and *Ephydra gracilis*. The second group contained *Aedes aegypti*, *Mochlonyx cinctipes*, *Culex pipiens*, and *Anopheles gambiae*. For each of these groups, these noisy parts were aligned together (multiple alignment), and built covariance models.

Merging the groups

The *noisy* part was the major obstacle and this strategy was applied to overcome this problem and find the best possible alignment for all sequences together. In this step, the goal matching the sequences and their alignment in the best possible way regarding their differing regions. The vertical division ended with 4 different covariance models built for the *noisy* parts, one for each of the processed group, except for *Drosophila*, which ended up with two different models. A listing of the models follows:

- AN: Apocrita noisy part model
- TN: Tenthredinoidea noisy part model
- CN: Coleoptera noisy part model
- DDN: Diptera (with *Drosophila*) noisy part model
- DN: Diptera non-*Drosophila* noisy part model

Using `cmscan` by `infernai`, each group was scanned by the covariance models of all of the other groups. For example, AN sequences (noisy part only) scanned simultaneously by all other models, then TN scanned simultaneously by all other models (including AN). This step was applied for all models, and the best hits were between Apocrita and Tenthredinoidea.

The next step was merging these sequences together to find the consensus structure of this part. Using the CMALIGN feature from `infernai`, The sequences of the other groups were simultaneously merged (aligned) to the model of Apocrita and Tenthredinoidea. At this step, the stem A and its extension was defined clearly, and the problem was minimized to a specific region between the extended stem A and 3' end.

The exactly same steps were applied again but for the reduced *noisy* part of these sequences. At this step the noisy parts alignment showed obvious

improvement, so all parts of all species from all groups were glued again defining the final alignment and consensus structure. Of course, manual curation was also a prerequisite at this level. Using this final model, the new genomes were searched and new candidates were found, 3 species from Tenthredinoidea, 2 from Coleoptera and one from each of Orthoptera, Dictyoptera and Cephoidea.

For a general invertebrate model, the Lophotrochozoa sequences were aligned to the final model described in the previous paragraph.

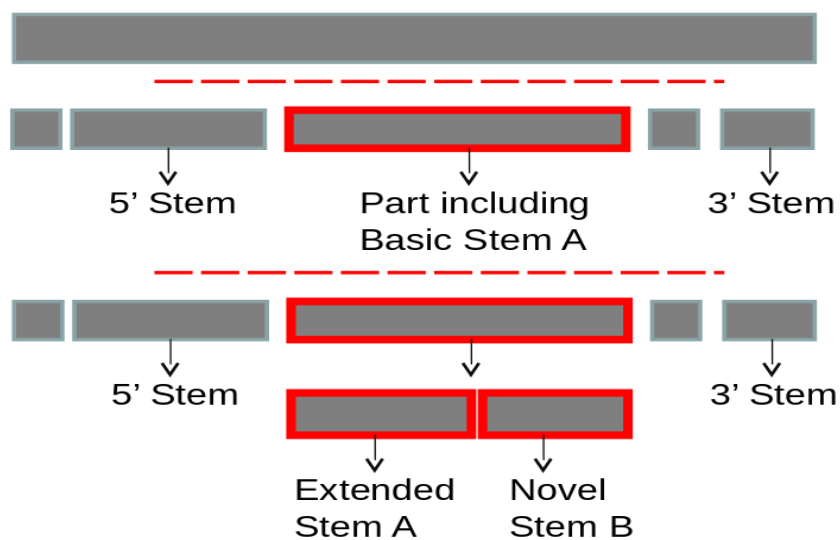


Figure 3.1: This figure shows how sequences were split vertically. First, the regions between 5' stem and 3' stem were cut out of the sequences and aligned together. Second, the same regions were split again, at the stem defined in the previous step (extended stem A), then the remaining parts were aligned alone together (Novel stem B).

3.4 Results

In total, 48 candidates for 7SK gene in invertebrate genomes, were identified. The phylogenetic distribution of the sequences is summarized in figure 3.2. No ecdysozoan candidates were found outside Arthropoda. However, the phylogenetic scope within the group was substantially extended by identifying not only many additional sequences from diverse Hexapoda, but also new genes from the subphyla Crustacea and Chelicerata. Also, the much smaller collection of known and predicted lophotrochozoan 7SK RNA genes were reanalyzed.

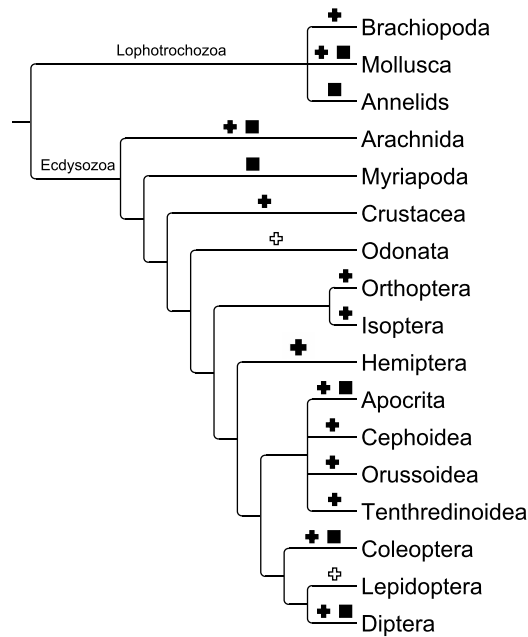


Figure 3.2: Phylogenetic distribution of the 7SK genes identified in Arthropoda. Black crosses mark clades for which 7SK RNA are newly reported, groups for which homologs were reported in earlier studies are marked by black squares. White crosses indicates clades where homology searches remained unsuccessful. For simplicity only a single branch for Crustacea is shown.

3.4.1 Refined model of arthropod 7SK RNA

3.4.1.1 5' Stem

The structure of invertebrate 7SK RNAs is best discussed by first using the Hexapoda consensus, figure 3.5, as a paradigmatic example. Overall, the molecule is well-conserved in both sequence and structure. The 5' stem is highly conserved and very similar to the corresponding well-characterized structure of its vertebrate homolog (Egloff, Van Herreweghe, and Kiss, 2006; Uchikawa et al., 2015), see figure 3.3. The model, in particular, conserves the P-TEFb binding motif including the U-bulges. Moreover, the nucleotides around this binding site are also conserved between Hexapoda and vertebrates with a notable difference, the A bulge at the top of the Hexapoda P-TEFB binding is conserved as a U bulge in vertebrates. The 5' stem is also well conserved in flies (*Drosophila*), as demonstrated by a comparison of the model proposed here, with the structure reported by (Cosgrove et al., 2012) for the 7SK RNA of *Drosophila melanogaster*. Both structures feature the same P-TEFb binding motif as well as a second identical helix.

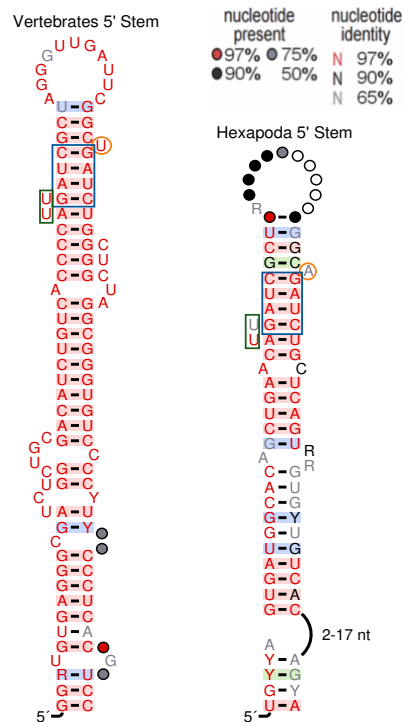


Figure 3.3: Comparison of the 5 prime stems 7SK RNA in Vertebrates and Hexapoda. Nucleotides marked by a letter are present in at least 65%, 90% or 97% of the sequences depending on the color as indicated in the legend. Circled positions without a letter are more variable and are present in a fraction of the sequences as indicated in the legend. Base pairs shaded in red show no variation, green indicates covariation and blue refers to compatible mutation. The boxed nucleotides are crucial for P-TEFB binding.

3.4.1.2 Extension of Stem A

Stem A was clearly defined previously with 5 to 6 base-pairing nucleotide and a bulge nucleotide at the 5' arm of the stem (Gruber et al., 2008b). These results strongly suggest that the conserved *Stem A* structure is more extensive than previously reported. In particular, there are at least two to 3 additional base pairs. Key features of this structure, in particular, these additional pairs at the base of the stem, are clearly conserved between vertebrates and invertebrates. See the *Stem A* marked region in 3.6 and the second stem in the figure A.1.

3.4.1.3 Novel stem B in invertebrates

Interestingly, the vertical division of the sequences explained before, revealed additional stem in the invertebrates species which was not recognized as a conserved element in earlier studies (Gruber et al., 2008b; Gruber et al., 2008a), which in particular focussed on drosophilas. A similar structure is part of the vertebrate 7SK RNA secondary structure model (Egloff, Van Herreweghe, and Kiss, 2006; Uchikawa et al., 2015). Because of this similarity, this novel stem is

confidently named as *Stem B* now.

As shown in figure 3.4, this stem features a highly conserved loop as well as a quite well-conserved sequence over much of its stem. Nevertheless, the Coleoptera species were a special case in Hexapoda species as they did not share the whole stem components with the others, as it shared only part of the stem (yellow boxed regions) in addition to the highly conserved loop and few nucleotides located directly under the loop (green boxed regions).

Diptera

On the other hand, *Drosophilas* have insertions in this region that effectively made it impossible to construct reliable alignments of this region with other arthropods. Hence *drosophilid* sequences were excluded from the construction of consensus models for the region between *Stem A* and the 3' hairpin. While the 7SK RNA of other Diptera species (Non- *Drosophila*) is within the range of other arthropods, their sequences are highly divergent in this region and did not show any credible alignment with other arthropods. Hence separate alignments and structure models have been constructed for *Drosophilas* and the remaining Diptera, respectively. The Dipteran (non-*Drosophila*) model, A.2 shares at least an overall structural similarity as well as the P-TEFb binding sites in both the 5' and 3' stems with the other arthropods. In these regions, the alignment clearly shows the sequence homology. The structures in the *Stem A* and *Stem B* regions, on the other hand, do not seem to conform to the arthropod consensus.

Arachnida

The novel *Stem B* did not show any presence in the Arachnida model, even as dissimilar stem, figure A.3. Arachnida model which was built of nicer sequences, completely misses this region and shows less conservation in the 5' stem region, although the long-range interaction does appear to be well preserved. Where there is also a clearly recognizable *Stem A* and a 3' hairpin.

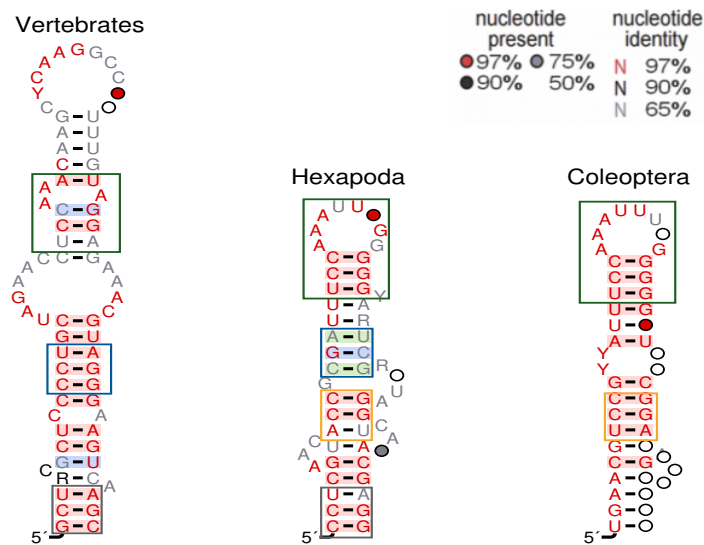


Figure 3.4: Comparison of stem B 7SK RNA in vertebrates, Hexapoda without Coleoptera, and Coleoptera alone, respectively. Nucleotides marked by a letter are present in at least 65%, 90% or 97% of the sequences depending on the color as indicated in the legend. Circled positions without a letter are more variable and are present in a fraction of the sequences as indicated in the legend. Base pairs shaded in red show no variation, green indicates covariation and blue refers to compatible mutation. The boxes of same colors represent the conserved regions between models.

3.4.1.4 3' Stem

In the final models, the 3' stem conformed with the 3' stem annotated by (Gruber et al., 2008b) as well as showing high similarity to the same part in the vertebrates model. This part starting with GGC-CCG has been experimentally proven by (Egloff, Van Herreweghe, and Kiss, 2006) to be a crucial part for P-TEFB binding, as without this short stem-loop P-TEFB is abolished. Back to figure A.1, removing the part labeled as d6 completely abolished the P-TEFB binding process. These triple base-pairings at the base of the 3' stem are conserved in all presented models, with a very small decrease in the conservation for the rest of the nucleotides under the loop.

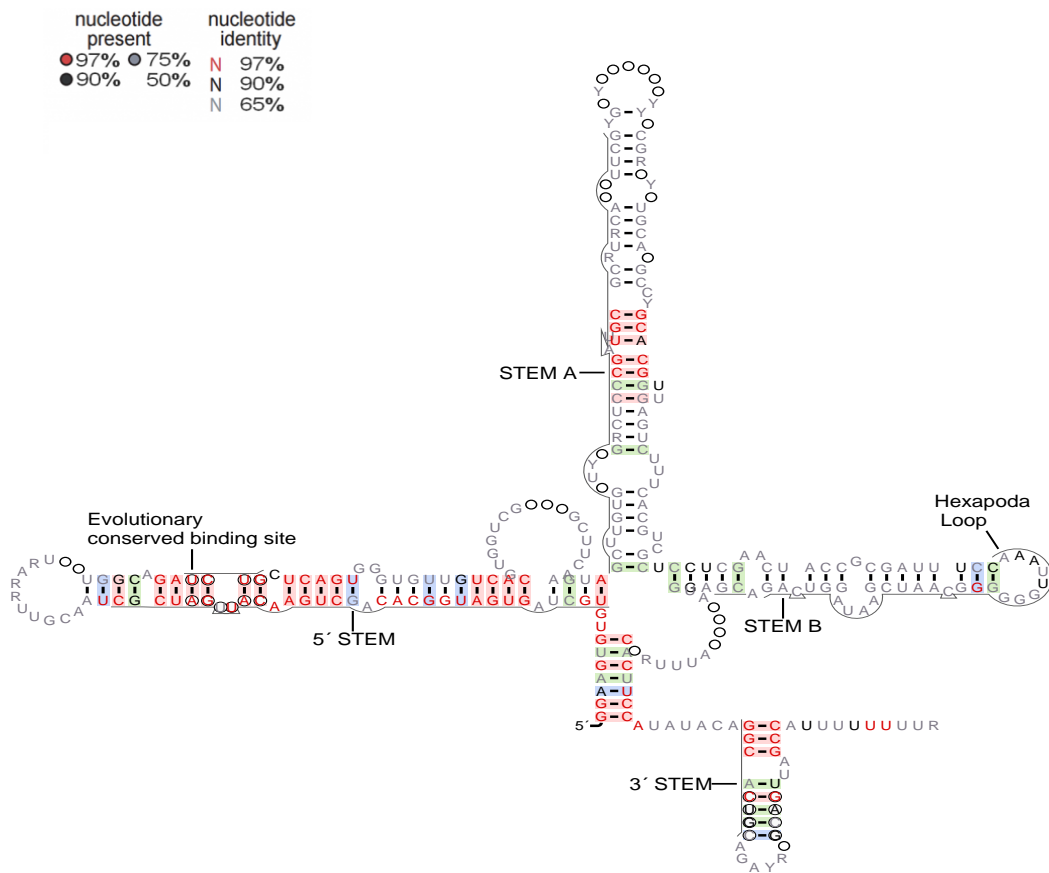


Figure 3.5: Hexapoda 7SK RNA structure drawn with R2R (Weinberg and Breaker, 2011). Nucleotides marked by a letter are present in at least 65%, 90% or 97% of the sequences depending on the color as indicated in the legend. Circled positions without a letter are more variable and are present in a fraction of the sequences as indicated in the legend. Base pairs shaded in red show no variation, green indicates covariation and blue refers to compatible mutation. In addition, the circled nucleotides at the 5' stem represent the P-TEFb and HEXIM1 binding sites and the target site of PPIM1G. The circled nucleotides at the 3' stem, are nucleotides crucial for P-TEFb binding.

3.4.2 Invertebrates model conserves the HEXIM1 binding site

Superpositions of the structural models for individual phylogenetic groups (see Figure 3.6) show that there is significant variation also at the level of secondary structures. On the other hand, there is also a clearly identifiable core structure common to the invertebrate structures. This core in particular covers the top 11 base pairs of the 5' stem including the P-TEFb binding site as well as the 3' hairpin. At least six base pairs of *Stem A* are also essentially immutable.

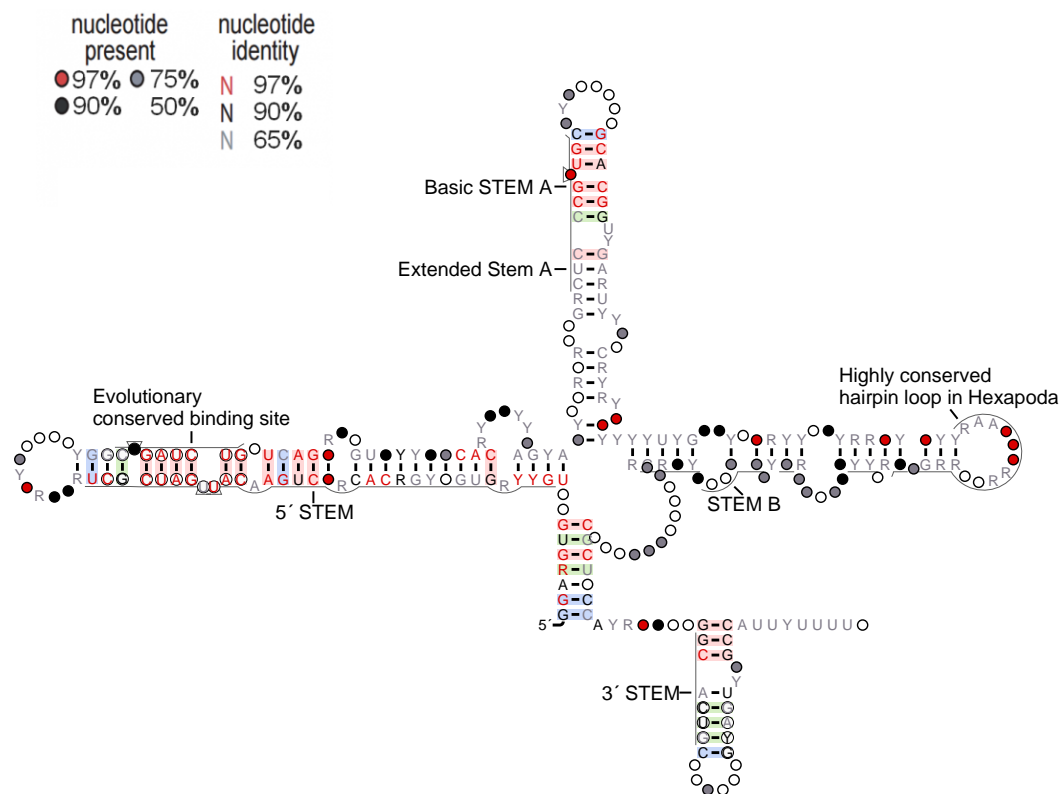


Figure 3.6: General 7SK RNA invertebrates model. Nucleotides marked by a letter are present in at least 65%, 90% or 97% of the sequences depending on the color as indicated in the legend. Circled positions without a letter are more variable and are present in a fraction of the sequences as indicated in the legend. Base pairs shaded in red show no variation, green indicates covariation and blue refers to compatible mutation. In addition, the circled nucleotides at the 5' stem represent the P-TEFB and HEXIM1 binding sites and the target site of PPIM1G. The circled nucleotides at the 3' stem, are nucleotides crucial for P-TEFB binding.

Stem B can be ascertained only in Hexapoda. At present, the data remain very sparse for Lophotrochozoa, where 7SK sequences so far could be identified only for a few representatives of the Mollusca. In isolation, these data were insufficient to infer a *Stem B*-like structure. In particular, the relevant sequence regions were too dissimilar to the *Stem B* sequences of Hexapoda for a credible sequence alignment. Similarly, no evidence shows that the *Stem B* in Hexapoda (and possibly lophotrochozoan) is homologous to a corresponding structure in vertebrates.

As expected, the HEXIM1 binding motif GAUC–GAUC (Lebars et al., 2010) is perfectly conserved across the invertebrates, as noted already in (Gruber et al., 2008a; Gruber et al., 2008b). Consistent with the results of (Martinez-Zapien et al., 2017), furthermore, the adjacent uracil residues are also highly conserved, with the notable exception of the substitution of adenine. While the consensus motif for vertebrates reads UuGAUC–uGAUC, invertebrates instead

feature UUGAUC–AGAUC as their consensus HEXIM binding motif.

The 7SK RNA appears in an open and a closed conformation. In the closed structure, a helix is formed between the 5' end of the molecule and a complementary region just upstream of the 3' stem (Martinez-Zapien et al., 2017). Mutations in this stem affect LARP7 binding (Uchikawa et al., 2015). The alternative, open conformation has been described e.g. in (Egloff, Van Herreweghe, and Kiss, 2006) and (Wassarman and Steitz, 1991). In this work the structure described as a closed structure, summarized in figure 3.7.

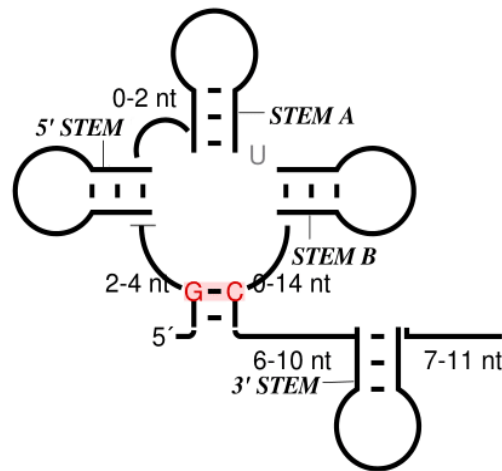


Figure 3.7: The general core structure of invertebrates 7SK RNA.

3.4.3 Computationally high potential 7SK RNA candidate

After checking the upstream of all candidates, interestingly, *Metapolybia cingulata* upstream contained the Pol-III promoter. As shown in figure 3.8 the TATA box located 23 nucleotides upstream the 7SK RNA sequence as it clearly shows high similarity to the upstream regions of different U6 RNA sequences. Moreover, the secondary structure of this candidate contains all the binding sites described above, and it follows the invertebrate snRNA secondary structure proposed in this work. The secondary structure is shown in the figure A.4.

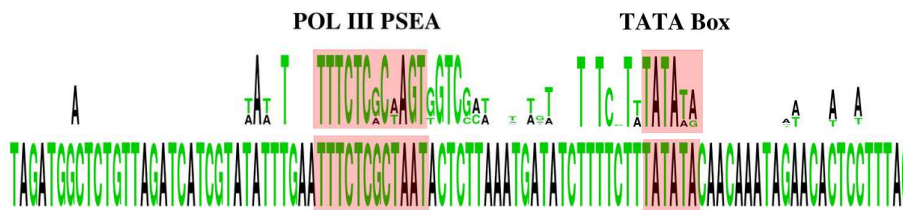


Figure 3.8: Alignment of the region upstream of the 7SK RNA candidate from the genome of the ant *Metapolybia cingulata* (below) and a sequence logo (above) constructed from the upstreams regions of the three genomic copies of the U6 snRNA, the single U6atac snRNA, and the 7SK RNA from honey bee (*Apis mellifera*). The PSE and TATA elements are highlighted. The sequence logo was generated with WebLogo (Crooks et al., 2004).

3.4.4 Sensitivity of the final proposed model

To test the main goal which is improving the sensitivity of detecting new candidates, the final models of this work and the ones of the Rfam database were compared. The comparison was on two levels, the genomic sources of the candidates and the candidates themselves. Not surprisingly, the Arthropoda model RF1052 easily detected the Hexapoda sequences in the genomes and in direct search against the candidate sequences, but with a better E-value for the proposed model. For Lophotrochozoa and Arachnida it was always better to search with the model of this work, since RF1052 did not detect all candidates used as targets. On the other hand, the Rfam 7SK model RF00100 only competed in finding the Lophotrochozoa sequences. Figure A.5 shows all the E-values of these comparisons. The green E-values represents almost perfect hits; the yellow reasonable to good hits and the rest are values of bad or no hits.

3.5 Conclusion

In this work, the models of the 7SK RNA throughout the bilaterian animals were significantly extended and refined. Based on carefully curated results of iterative homology searches, the detailed model of the secondary structure and the structural variation of 7SK throughout the Ecdysozoa were constructed, but to a lesser extent in the Lophotrochozoa. This analysis done in this work, confirms in particular, the evolutionarily well-conserved sequence and secondary structure elements at both the 5'- and 3'-ends of the molecule that are associated with P-TEFb and HEXIM binding. In addition, The *Stem A* was also found to be a universally conserved feature of this snRNA. In contrast,

an additional structural domain was established, *Stem B*, located between *Stem A* and the 3' hairpin, that is conserved in most Hexapoda and possibly across some other major invertebrate lineages. Despite the presence of well-conserved domains and a common overall organization, we observe that the structure of the molecule was subject to substantial changes throughout animal evolution, with highly derived homologs in Diptera and in particular in fruitflies. The homology search step reconfirmed earlier observations (Gruber et al., 2008b; Gruber et al., 2008a; Marz et al., 2009) regarding the difficulty of finding 7SK genes. The `blastn` based searches almost exclusively recognize the 5' hairpin region as soon as phylogenetic distances reach (sub)phylum level. As expected, `infernald` adds at least a moderate gain in sensitivity (Menzel, Gorodkin, and Stadler, 2009). Once identified, it is surprisingly easy to recognize *bona fide* 7SK RNA sequences and to separate them from false positive candidates, owing to the total length of the molecule and its multiple conserved domains.

Chapter 4

Behind the scenes of microRNA driven regulation

4.1 Biological background

MicroRNAs (miRNAs) are small non-coding RNAs produced from endogenous transcripts in humans, animals, plants, and viruses (Gebert and MacRae, 2018; Vidigal and Ventura, 2015; Iwakawa and Tomari, 2015). The majority of these miRNAs are located within introns, however, a reasonable number of them are generated from exonic regions (Rodriguez et al., 2004). miRNAs are ~ 22 nucleotides cleaved from the stems of the pre-miRNAs hairpin (precursor) that have lengths ranging from 60 to 80 nucleotides in animals and more variable in plants (Kim, 2005). The precursors in its turn, are produced by the polycistronic primary transcripts, called pri-miRNAs (Bologna, Schapire, and Palatnik, 2013; Lee et al., 2002). The generic structure of a pri-miRNA conserves a capped 5' end and poly(A) tail at the 3' end (Cai, Hagedorn, and Cullen, 2004).

MicroRNA genes are transcribed by polII, producing the primary miRNA (pri-miRNA) (Lee et al., 2004). pri-miRNAs are processed by the microprocessor complex that releases the pre-miRNA hairpin-loop (precursor) (Lee et al., 2002; Denli et al., 2004). The microprocessor complex is formed by the interaction between Drosha and other nuclear proteins. For instance, this complex in humans is composed of Drosha and DiGeorge syndrome critical region gene 8 protein (DGCR8), wherein *D. melanogaster* and *C. elegans* this interacting protein is known as Pasha (Han et al., 2004; Denli et al., 2004; Kim, 2005). The processing of pre-miRNAs occurs in the cytoplasm after they have been exported from the nucleus by Exportin-5 (Exp5), which binds directly to the pre-miRNA in a RanGTP-dependent manner (Lund et al., 2004; Kim, 2004). Moreover, Exp-5 has a major role in protecting the pre-miRNA from degradation (Zeng and Cullen, 2004). In this context, the pre-miRNA length has a direct influence on its binding capabilities to Exp5. pre-miRNA with length less than 16 nucleotides can abolish pre-miRNA-Exp5 binding, as well as there is an effect of bulges and loop size specific cases (Zeng and Cullen, 2004). The successfully exported pre-miRNAs are processed with the Dicer protein. Dicer cleavage extracts usually imperfect duplex from the stem that features a 2 nt overhang and the 3' ends (Kim, Kim, and Kim, 2016;

Yu et al., 2016). When Argonaute (AGO) loads this duplex the two strands are separated, and the one with the less stable base-pairing at its 5' terminal nucleotides, the miRNA, is bound to AGO. The strand with the more stable base pairing at the 5' terminal nucleotides, the miRNA*, is usually excluded. Although both miRNA and miRNA* are often visible in small-RNA-seq data and in some cases both miRNA and miRNA* are functional, there is no data on the miRNA* for many miRNA families. Figure 4.1 summarizes the process of microRNA synthesis, showing the 2 overhanging nucleotides remaining after Dicer processing.

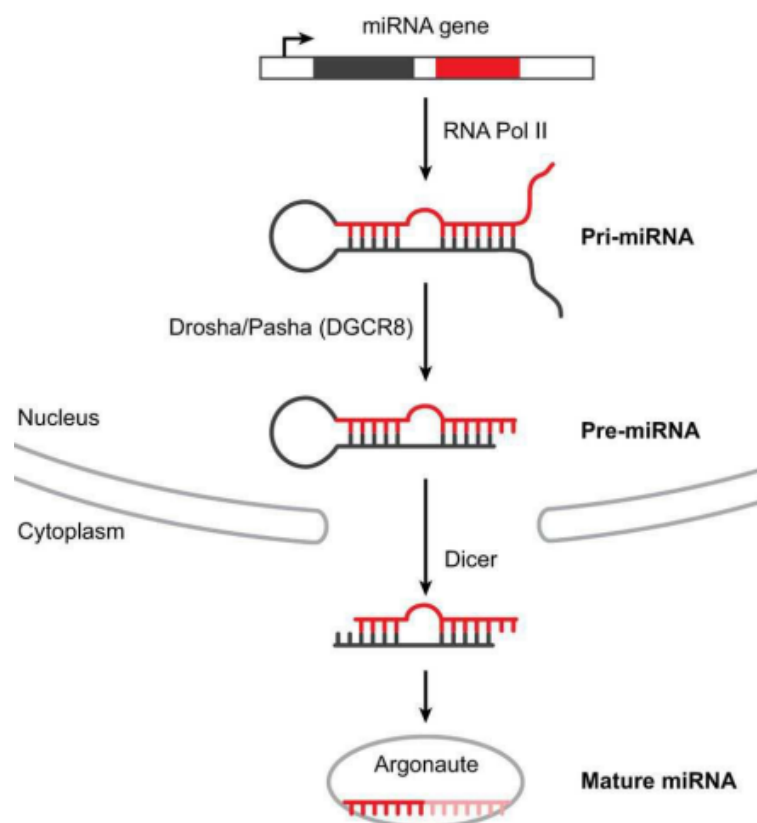


Figure 4.1: Synthesis of a mature microRNA, from the transcription process to the cleavage process. Figure is adopted from (Alberti and Cochella, 2017).

These small RNAs play different roles in gene regulation (He and Hannon, 2004). Gene negative regulation is one of the main and best understood roles. A miRNA represses or downregulates the translation and the expression of its targets (Olsen and Ambros, 1999; Reinhart et al., 2000; Meister, 2013). This starts with miRNA-target interaction when the miRNA binds its target usually along the 3' UTR region in humans (Bertoli, Cava, and Castiglioni, 2015) (figure 4.2).

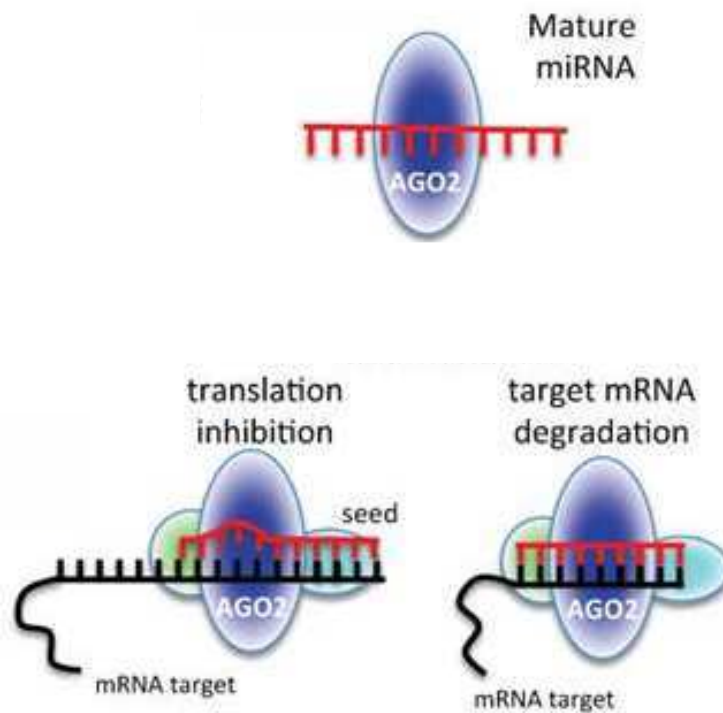


Figure 4.2: This figure shows a miRNA target interaction. Based on the position and level of complementarity between miRNA and mRNA, the mRNA translation can be inhibited or interaction will lead directly to the degradation of the mRNA. This figure is edited from (Bertoli, Cava, and Castiglioni, 2015).

The main difference between miRNA driven regulation in plants and humans is that in plants the miRNA binds to transcribed regions with almost perfect complementarity (Voinnet, 2009). Nevertheless, there are always some exceptions, as e.g. mentioned in (Rhoades et al., 2002). But not regulation by miRNAs, also regulation of miRNA levels play a role in various human diseases. For instance, in chronic lymphocytic leukemia, miR15 and miR16 are downregulated (Calin et al., 2002) as well as miR9 and miR153 in Alzheimer's disease (Femminella, Ferrara, and Rengo, 2015; Goodall et al., 2013) while in contrast, for example in lung cancer, miRNAs are overexpressed (Liu et al., 2012). Understanding the role of these small RNAs in diseases is essential since it has been proven that these microRNAs can be also targeted or hitchhiked for the treatment process of many incurable diseases. For instance, in (Kim et al., 2011) it was shown that combining miR-145 with the Fluorouracil (5FU) drug has improved the treatment of breast cancer. Figure 4.3 illustrates an experimental process of combining microRNA with a drug can lead to apoptosis of cancer cells.

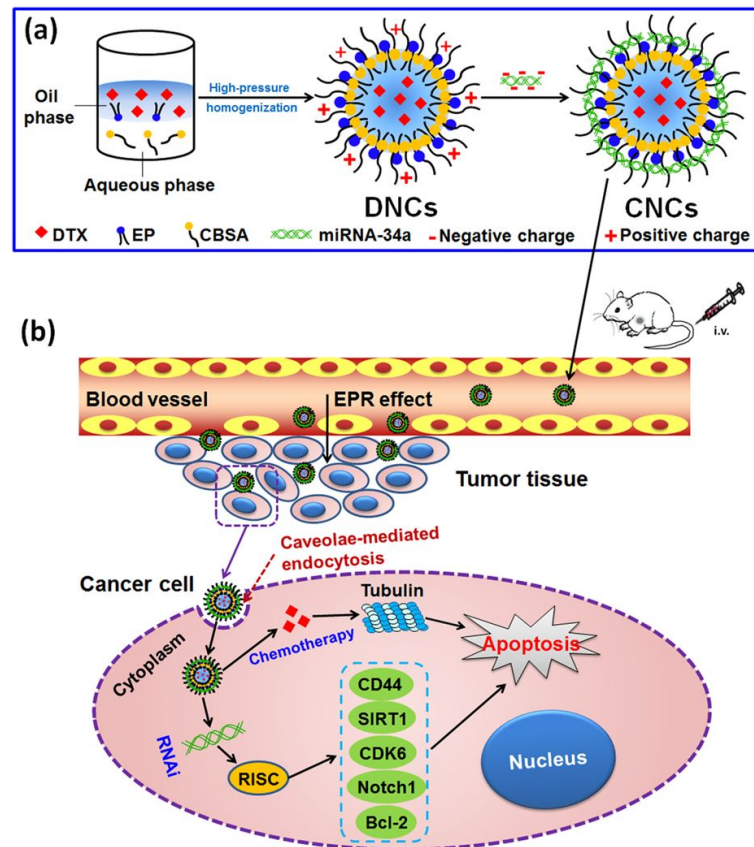


Figure 4.3: The full process of combining the drug docetaxel (DTX), which is commonly used for breast cancer chemotherapy, to miR-34a. (a) shows the process of combining DTX to miRNA-34a using a nanocarrier. (b) the process of penetrating a cancer cell with the nanocarrier, eventually killing the cancer cell. Source: (Zhang et al., 2017).

4.2 Databases and problems

Different miRNA databases are available as open source resources, e.g., miRNAMap, Rfam (Griffiths-Jones et al., 2003; Kalvari et al., 2017) or miRBase (Griffiths-Jones, 2004; Kozomara and Griffiths-Jones, 2014). The latter database is the main source of miRNA annotation for many other databases, and naturally, it contains more miRNA and pre-miRNA sequences than any other database. Moreover, this database offers organized files to download which also contain useful details concerning the sequences. For instance, the *miRNA.dat* file, which contains all the details about a given precursor and its mature sequences. This includes the mapping of the precursor to their related mature sequence, positions of the mature sequences in the precursor, the origin of the record and more. The latest release of miRBase is version 22, but since some critical files were not available when conducting this work, the here presented methods were applied to release 21. This registry is based on microRNAs, published in peer-reviewed journals, and aims at organising this data and provide a consistent nomenclature for families and precursors,

in addition to providing alignments and thus conservation information for annotated precursors. The used release contains 21263 pre-miRNAs in animal species, 14712 are classified into families summing up to 1415 metazoan families. Biologically, each precursor contains two mature sequences (miRNA and miRNA*). However, 9258 precursors are annotated only with one mature sequence. Figure 4.4 shows the distribution of families according to the number of annotated mature sequences in the precursors.

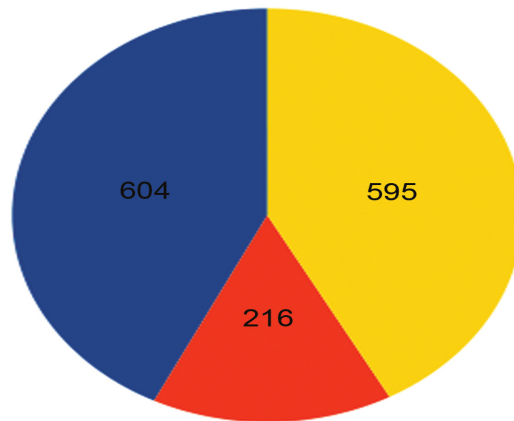


Figure 4.4: Distribution of mature microRNA sequence entries for miRBase (v. 21) families. The majority of families reports a miRNA and miRNA* for all (red) or at least some (blue) of the family members. Families in which only a single mature product is reported for every member are shown in yellow.

The main problems in annotating this data are related to the precursors (sequence and structure information) and to the classification of precursors to their correct families. Regarding precursors with one annotated mature sequence, it is possible that the mature sequence is oriented in the wrong orientation. In other words, the mature sequence can be annotated on the wrong arm, on the 5' end instead of 3' end, for instance. This error will have a lot of consequences for downstream analysis, starting from annotation of the precursor itself and ending with decreasing the quality of the homology search by suboptimal alignments. Clearly, in the mis-orientation case, the precursors will be incorrect as only part of them is correct. For example, when the mature sequence is annotated at the 5' arm and it should be at the 3' arm, the precursor is extended from the 3' instead of the 5' end. This problem can already be seen at a sequence-based search, as using a wrong sequence for a BLASTn search, for instance, will detect a wrong homolog sequence, leading to false positive results. Furthermore, increasing the number of false positives in a family of precursors will also lead to a bad alignment or a wrong one. The previous two problems also have a harmful effect on the homology search, e.g., for the construction of CMs for example. Another problem that affects downstream analysis, is the length of the sequences. Even when the positions of the mature sequence(s) are correctly annotated in their correct

arms, the longer than correct sequences become a problem when predicting the secondary structure. The generic structure of a pre-miRNA is a single hairpin, which consists only of one stem and one loop. When the sequence is too long, structure prediction can lead to the computation of an additional hairpin structure at one of the ends of the sequences or even both ends. This directly affects the computation of consensus secondary structures and as a consequence any kind of homology search. One of the remaining challenges is the assignment of the pre-miRNA to their correct families.

4.3 MicroRNA detection and curation approaches

As diverse as the approached for homology search and data curation are the tools used for microRNA detection and curation. They mainly range from tools studying microRNAs based on their conservation across species to tools curating the outputs of homology search methods and biological experiments. These microRNA homology search and curation tools were reviewed and classified into at least two main different perspectives. In (B Malas and Ravasi, 2012) they were classified into comparative and non-comparative tools, where the comparative tools are based on the genomic conservation, and the non-comparative tools focus more on the sequence information and structures. Another classification in (Li et al., 2010) classified these approaches into three main categories, sequence or structure conservation-based, machine learning-based and the experimental based. The first tools rely only on the genomic and sequences information, without any help of the machine-learning methods or methods like hidden Markov or covariance models. *miRseeker* (Lai et al., 2003) was one of the first computational approaches to identify new pre-miRNAs homologs from known *Drosophila* pre-miRNAs. Structural information of known pre-miRNAs was used to figure out main features that describe these precursors, and these characteristics were used for filtering at each step of the workflow. A first filter was detecting sequences in two *Drosophila* species, which are similar to the known set of used pre-miRNAs and eliminate those annotated in exons, transposable elements, snRNA, snoRNA, tRNA and rRNA genes. The second filtering step regards the structure of the sequences after folding, and removing sequences of less than 23 base-pairs in a stem and sequences of minimum free energy greater than -23 kcal/mol. The final step was comparing the structures of final candidates to the different scenarios which appeared in the known set of the pre-miRNA. The advantages of this approach can detect true homologs with less false positives, and the candidates can be classified among other non-coding RNAs. However, exploring candidates not closely related in the phylogenetic distribution of the species is still a big challenge that is not covered in this approach, in addition to the limitation of search to a specific family and species.

Similarly to *miRseeker* approach, but it extends the strategy to reach more

distant species, *miRAlign* (Wang et al., 2005) is also a computational approach for detecting microRNA. It is based on sequence and structure alignments, with more focus on the mature sequences of the known set of miRNAs. The genomes are first searched by *Blast* using the known set of miRNAs, then 70 flanking nucleotides around the potential hits were cut from the genome, and searched with 100 nucleotide long windows. The sequences which are not overlapping with repeat sequences were considered as potential candidates to be processed. This is an example of an approach that can lead to mature sequences shifted to the wrong end of the annotated precursors. The secondary structure of the potential sequences is calculated using *RNAfold* and the sequences with MFE score less than -20 kcal/mol are kept and aligned to the known miRNA sequences, keeping only sequences of similarity score greater than 70 (or user-defined). The position of the mature sequence is then checked with more constraints, the mature sequence should not be located in the loop, should be in the same arm (5' or 3' stem) and the distance from the loop should also be similar according to their defined distance calculation. In the end, the secondary structure of the remaining candidates is compared to the ones of their homologs in the known set of microRNAs. Furthermore, this approach is extended to be able to search any available genome. However, it is also restricted to known homologs, especially regarding the positioning of the mature sequence. The main issue here is that the miRNA can be wrongly annotated or very similar mature sequences can appear in different stems but only for different families.

The two described approaches focus on genomes and detecting homologs as a starting point for their workflows. Selection of candidates rely on either homolog miRNAs or pre-miRNA structures. *RNAmicro* (Hertel and Stadler, 2006) focuses on the whole alignment of sequences and their consensus structure without referring directly to the genomes or comparing the sequences to their homologs. Instead, it uses support vector machines to distinguish microRNA sequences from other non-coding RNA sequences in a given alignment. *RNAmicro* is pre-processing the alignments and then designs specific descriptors. At the pre-processing level, a window of defined size is moved in one nucleotide steps, computing its consensus sequence and structure. Windows with stems of less than 10 base-pairing nucleotides or constituting of two or more hairpins with at least 5 base-pairing nucleotides each are removed. Then descriptors are computed for the rest. These descriptors are calculated to be used in the SVM classification. The calculation is based on the length of the stem and the loop region in addition to the sequence composition in terms of G-C content; the secondary structure i.e., the thermodynamic stability, as it is known that microRNAs are more stable than other non-coding RNAs; the structural conservation of the sequences and the energy of the consensus secondary structure; and finally, the position of the mature miRNA sequence (i.e., same orientation which can be either at 5' end or 3' end, another part of the stem, or it can be located in the loop region). Then, based on positive and negative training sets, the SVMs are trained to classify microRNA sequences and others in a given alignment.

Later on, methods based on the outputs of the deep sequencing experiments appeared. The *miRDeep* (Friedländer et al., 2008) approach is using the read-out of deep sequencing and mapping profiles to the genomes to detect novel microRNAs. The location of the reads and the pre-miRNA structural information are key features used to find novel microRNAs. Reads are initially filtered, removing all reads that don't seem promising for the next steps, and keeping the reads which show high nucleotide conservation with a minimum length of 18 nucleotides. Remaining reads are used next to find putative hairpins (precursors). Every two reads, located within 30 nucleotides or less, are considered to be potential miRNA and miRNA*, considering 25 flanking nucleotides around the reads (putative miRNAs). Finally, sequences which sum up into 100 nucleotides length are discarded. For reads with a distance more than 30 nucleotides from each other, two precursors are considered, one having the reads at the 3' end and the other one with the same reads at the 5' end. The length of these latter precursors is fixed at 110 nucleotides. These extracted sequences are filtered, and the ones that don't look like miRNA precursors are removed from further study. Filtering is based on the following points: The mature sequence position is the position where the most potential miRNA are aligned to. The miRNA* is defined as the sequence base-pairing to the mature miRNA with 2 overhanging nucleotides at the 3' end, and the loop is the region between the candidate miRNA and miRNA*. The putative sequence should fold into only one hairpin loop with a minimum of 14 base pairing nucleotides between potential mature miRNA and miRNA* sequences. The putative mature sequence has to be consistent with Dicer processing, i.e., the potential sequence has 2 overhanging nucleotides if it is at the 5' end, or up to 5 nucleotides when it is at the 3' end. If more than 10% of the reads are inconsistent, the precursor is discarded. More steps in *miRDeep*, after discarding the precursors which according to their used criteria don't show a clear pre-miRNA shape, they designed a probabilistic scoring method for further filtering. This *miRDeep* core algorithm relies on the 3 main points: Presence of one or more reads correspond to a miRNA* sequence with overhanging nucleotides at the 3' end, the stability of the structure and the 5' end of the predicted mature sequence should be similar to the 5' end of the known mature sequences.

More approaches started from the output of the deep sequencing and used the output reads to predict novel microRNAs based on the known ones, this time utilizing machine learning. *miRanalyzer* (Hackenberg et al., 2009) combines reads, known microRNAs and machine learning to detect novel microRNAs. *miRanalyzer* maps reads to known miRNAs in *miRBase* and then to a database of transcribed sequence. Sequences which are already annotated as miRNA or detected as putative degradation products or other small non-coding RNAs are optionally removed. In the alignment step reads are mapped to known microRNAs from *miRBase* and a theoretical miRNA* library is predicted for *miRBase* precursors with only one annotated mature sequences. To annotate these miRNA* sequences, the original precursor are folded, based

on the coordinates of the corresponding mature sequence and considering the 2 nucleotides overhang, with the predicted miRNA* base-pairing to the annotated mature sequence. The remaining reads are then aligned to the genome and sequences which overlap in the genome are clustered. Clusters which have mature sequences with more than 25 nucleotides are removed. For each putative mature sequence in the clusters, two different putative precursors are checked, since the mature sequence can be at the 5' end or the 3' end. One sequence with 60 nucleotides upstream and 10 nucleotides downstream of the mature sequence and vice versa for the other sequence are chosen. These sequences are folded using `RNAfold` and the sequence with a better score is retained in addition to removing the sequences that didn't fold at all. The precursors with miRNA-miRNA* base-pairing with less than 14 nucleotides are also removed. At the end, the precursors were defined by merging the clusters in a way forming a hairpin shape, i.e., clusters with putative miRNA, miRNA*, loop, etc.

More approaches apply SVMs, but also include more sequence and structure characteristics to train these models. In `miRPara` (Wu et al., 2011) all the characteristics of all the sequences involved in the microRNA biogenesis are included. Hence, including the characteristics of pri-miRNAs in addition to pre-miRNAs and miRNA characteristics that were used in different ways in the previous approaches. In this approach, the physical characteristics of the aforementioned sequences are defined to use as input parameters for SVM training. 3 different SVMs are used. Two trained only with mammalian and plant sequences of `miRBase` and a third one with all the `miRBase` sequences. Different features or parameters were grouped into different sets, descriptive, sizes, stability, sequences information and structure. The last filter of these parameters is rested on 5 main characteristics: Length (size of miRNA-miRNA* base-pairing region, G-C content (in miRNA and pre-miRNA), nucleotide content (miRNA and pre-miRNA), internal loops (miRNA and the top stem, the top stem is the stem between the miRNA-miRNA* stem and the loop) and unpaired bases rate (in pre-miRNA, the lower stem which is the stem before the mature sequences stem and the top stem).

The new sequencing methods helped also in developing new strategies and approaches for microRNA studies. Degradome sequencing data were recently integrated into these studies (Yu et al., 2018). Degradome sequencing datasets represent the cleavage sites of miRNAs. `miRNA Digger` (Yu et al., 2016) uses this data as an initial step to define the different loci in the genomes which could contain potential pre-miRNA sequences. The target genomic sequences are first scanned with the degradome data sets to define potential cleavage sites, and then it is assumed that Dicer processing happened at this site to cleave out the miRNA sequence. According to this assumption, and to the fact that a miRNA sequence can be located at the 5' or 3' end of the pre-miRNA, two candidate precursors are considered. The length of each candidate is a defined length based on `miRBase` data plus 30 nucleotides (a bit more than the average length of the miRNA in general). The defined

length is the length of the longest pre-miRNA sequences in the miRBase data in a given kingdom of species (plant or animal). The first candidate has the defined length upstream of the cleavage site and 30 nucleotides downstream of the cleavage site, representing a putative miRNA at the 3' arm. The second candidate has the defined length downstream of the cleavage site and 30 nucleotides upstream of the cleavage site, and in this case, the miRNA is presented at the 5' arm. These two candidate sequences are compared by their secondary structure calculated with `RNAfold`. These candidate precursors are retained or removed at this level according to two main filters, the MFE score, and the miRNA-miRNA* complementarity. When the sequence is folded, then there are two possibilities for the miRNA sequence either it is 3 nucleotides upstream of the cleavage site and 27 nucleotides downstream or vice versa. This step is applied because the real 5' and 3' ends of the candidate sequence are unknown. Then the sequences having better MFE scores and potential miRNA-miRNA* complementarity of more than 70%, are retained. The remaining potential candidates are then scanned with high throughput sequencing data to create corresponding miRNA and miRNA* clusters. The candidates which have both, miRNA and miRNA* clusters are retained, otherwise removed. The most abundant miRNA sequence in the miRNA cluster of a given pre-miRNA candidate is considered as the potential candidate, and its position is considered as the miRNA position. Then the miRNA* is supposed to be the sequence base-pairing to the potential miRNA considering the 2 overhanging nucleotides. Then the miRNA* is extracted from the miRNA* cluster. In case the exact miRNA* sequence is not found in the cluster, then an isomiR* is considered. An isomiR is a variant of a miRNA modified during dicer processing (Nielsen, Goodall, and Bracken, 2012; Lee and Doudna, 2012).

Chapter 5

Initial microRNA curation

5.1 Introduction

The main idea at this stage was to start a careful curation for the microRNA precursors, accompanied with quantitative analysis of the annotated data. A strategy was applied to the `miRBase` families release 21 since this work was done before the new release 22. The goal was to improve available alignments, by decreasing the noise in the latter. To this end, a somewhat limited strategy was applied covering two of the problems discussed before. To improve these alignments, the focus was mainly on the pre-miRNA sequences that are longer than the consensus length of their given family. Additionally, precursors that are not removed were fit to the consensus model of their corresponding families.

The strategy was applied specifically to families belonging to two types of precursors: precursors with one annotated mature sequence and precursors with two annotated mature sequences. The precursors with two annotated mature sequences contain miRNA and miRNA* sequences, while the others have only the miRNA sequence annotated and it can be at either located on the 5' arm or 3' arm. The `infernal` tool was used to build the related covariance models and process the targeted families. Based on these models, the sequences that did not show reasonable similarity to the consensus alignment of their families were removed. Otherwise, the sequences are added back to their families, correcting the ends of the long sequences, to fit the consensus alignment.

In this section, I present a first step towards a complete automated pipeline for the curation of available microRNAs, and as advanced application, understanding the evolution of these microRNAs. The seed alignment was improved for most of the processed families by editing or removing noisy sequences. The quality of these alignments was measured as Shannon entropy. The processed sets of metazoan precursors in this work showed us how feasible it is to build such pipelines, which can consequently improve the quality of the strategies and tools that aim to detect new miRNA candidates. A quantitative analysis is also part of this chapter, demonstrating the connection between both computational and biological approaches.

5.2 Methods

5.2.1 Data pre-processing

Precursors were initially separated into metazoan and non-metazoan, and then the corresponding family files were created. The family file is a fasta file containing all the precursors which belong to a given family. The assignment of the precursors to their families was based on the *miRNA.dat* data file provided by miRBase.

In order to sort out the families having both types of annotated precursors, precursors with two annotated mature sequences and precursors with one annotated mature sequence, a mapping file is required. This mapping file contains all the essential information about each precursor. Each line corresponds to a precursor's record. The records contain: the corresponding family ID (which is also the family's file name), the corresponding family name, precursor name, precursor ID and the rest of the line is information about the related annotated mature sequence(s). The last information is the name of the annotated mature sequence(s), the related coordinates of the annotated mature sequence(s) in the precursor and the name(s) of the annotated mature sequence(s). The mapping file was then scanned, and for each family with at least two precursors with two annotated mature sequences and one precursor with one annotated mature sequence, a fasta file was created and named by the ID of the family. Generation of the mapping files was based on the two files provided by miRBase *miRNA.dat* and *miFam.dat*, which are available at <http://mirbase.org/pub/mirbase/21/>.

5.2.2 Initial seeds creation

Creating the initial alignments and Covariance models

For each of the targeted families, two different fasta files were created. One containing only the precursors of the two annotated mature sequences, and the other containing the precursors of one annotated mature sequence. The files with two annotated mature sequences were processed alone. First, the upstream and the downstream nucleotides of the miRNA and miRNA* sequences of each precursor were cut to contain 10 flanking nucleotides, when these flankings are more than 10. In case the upstream or the downstream nucleotides around the miRNA or miRNA* are equal or less than 10 nothing was changed.

The processed sequences (with two annotated matures) are then aligned together using `clustalw` version 2.0.12. And then stockholm files were created for these alignments, calculating the consensus secondary structure of each alignment with `RNAalifold`. The stockholm files were then used to build covariance models of these families, and for each family, a covariance model was created using the `cmbuild` command of the `infernal` suite.

Scanning the one mature sequences sub-files

At this stage, covariance models are ready for precursors containing two annotated mature sequences of the targeted families, and the sequences with one annotated mature of the same families in separate fasta files. For each family, the fasta files containing one annotated mature sequence were scanned with the `cmscan` command of the `infernal` tool. Thus, the covariance models of the two annotated mature sequences were used as target CM database and the sequences of the one annotated mature were used as query sequences.

5.2.3 Main course

Hits Evaluation and new alignments

The default inclusion threshold of the `infernal` tool is 0.01, that means for every 100 scans of a given model against different query sequences, we expect on an average one false positive hit (Nawrocki and Eddy, 2016). In this work, a more restricted inclusion threshold was used, equal to 0.0001. And this was chosen to ensure the conservation of both sequence and secondary structure of the precursors of the same family. The sequences with E-values greater than or equal to 0.0001 were totally excluded from the new alignment. The sequences that showed E-value less than 0.0001 were modified and added back to the sequences containing two annotated matures. Interestingly, for some query sequences, their reverse complement was detected as a good hit. This is not wrong biologically as it was clearly shown for the miRNAs *iab-4* and *iab-8* that each of them is produced from a different strand (Tyler et al., 2008; Hui et al., 2013). This was shown also for *mir-3120* and *miR-214* by (Scott et al., 2012). Because this is considered biologically correct, these reverse complement hits were also considered and added back to the alignment.

Sequences were then modified based on the `cmscan` result, i.e, the start and the end of the sequences were defined according to hits in the `cmscan` result file. The acceptable and modified sequences are added back to the two mature sequences, which were in turn already processed before building the models. All these sequences were then re-aligned with `clustalw`.

The decision of using a sequence-based alignment tool was based on the comparison of the entropies of `clustalw` to the structural- based alignment tool `mlocarna`. Entropies were in most cases better for the the alignments produced with `clustalw`, especially for big alignments. However, `mlocarna` alignments showed better entropies in small families with a small difference from the sequence-based tool `clustalw`. For example, the entropy score of the *MIPF0000160* family was 63.604 using *Clustlaw* and 61.426 using `mlocarna`. *MIPF0000160* is the *mir-52* family with only 7 annotated precursors. On the other hand, the entropy of the *MIPF0000033* alignment which contains *mir-10* with 415 annotated precursors, is 161.138 for `clustalw` and

229.30 for `mlocarna`.

This decision was supported by the results of the *weighting* strategy that I developed, which is explained in the next paragraph. Basically, this strategy was applied to the small family alignments for two reasons. First, this strategy is applicable only when the alignment can produce a consensus structure (using `RNAalifold`) and second, the entropy difference between `clustalw` and `mlocarna` in the big families is much bigger than those for small families. Thus, it became obvious to favorite `clustalw` over `mlocarna` as E-values in the small families were much closer. Again, the *weighting* score of sequence-based alignments was better than those of structure-based alignments.

Weighting Strategy

This strategy works on ranking the sequences of an alignment based on both, structure and sequence information. It calculates the weight of the whole alignment. When it is applied, the consensus sequence and the consensus secondary structure of the alignment are considered. The sequences are ranked evaluating their nucleotide and structure information in reference to the calculated consensus. The final calculated weight of the alignment is the total weight of its sequences. The weight of each sequence is the sum of the frequency of each nucleotide multiplied by its own calculated factor. The following equation demonstrates the calculation of the weight of a given sequence:

$$SW = \sum_{i=0}^n f(i) * factor(i) \quad (5.1)$$

Where SW is the sequence weight. $f(i)$ is the frequency of the nucleotide at position i . The frequency is, how frequent this nucleotide appears at the same position in all sequences. $factor(i)$, is the factor calculated (below) for a nucleotide at position i in the sequence.

In order to calculate the factor of each nucleotide, the nucleotides are classified into 3 categories:

1. In the stem (base-pairing or not)
2. Outside the stem (but not in the loop)
3. Nucleotides present in the loop

The nucleotides of the first category are always given a factor equal to 1. For the nucleotides of category 2, the nucleotides are given factor α , α' or 1. The nucleotides with $f(i)$ greater than or equal to 0.8 are given factor 1, with $f(i)$ between 0.5 (included) and 0.8 (excluded) are given factor α and for nucleotides of $f(i)$ less than 0.5 factor α' .

α and α' are calculated as follows:

$$\alpha = X/N \quad (5.2)$$

and

$$\alpha' = LX/N \quad (5.3)$$

Where X is the total number of the non-pairing nucleotides in the sequence, LX is the number of the non-pairing nucleotides having frequency ($f(i)$) less than 0.5 and N is the sequence length.

For the nucleotides in the loop (3rd category), they are either in *non-informative* loop or *informative* loop. The loop is considered as *informative* if the calculated *average* is less than or equal to 0.125, otherwise it is not. The 0.125 threshold was used based on observations of the 7skRNA and microRNA data. The *average* is calculated as follows:

$$average = L/NS \quad (5.4)$$

Where L is the number of nucleotides in the loop and NS is the length of the stem.

All loop nucleotides are given a factor equal to 1 when the loop is considered *informative*. Nucleotides in *non-informative* loops are given a factor equal to β calculated as follows:

$$\beta = |\log(L/NS)| \quad (5.5)$$

L and NS same as in equation 5.4 The absolute and the log here are used to describe that a decrease in the size of the loop makes the loop more *informative*.

Alignments measurement using Shannon Entropy

Shannon entropy is a probabilistic information entropy with general form

$$E(\mathcal{W}) = - \sum_{i=1}^n p(x_i) \log_2 p(x_i) \quad (5.6)$$

where \mathcal{W} is a given word, x is a given character in the word and i is a given position in the word. This Entropy increases as the characters are more diverse, and decreases until zero if the characters are similar. For example, the substring "aaa" has a Shannon entropy *zero* and the entropy of the substring "aab" is equal to 0.9183.

To compare the alignments of each of the processed families before and after processing, these alignments were measured using Shannon entropy. The Entropy was calculated column-wise, so the consensus alignment is divided into sites. Each site represents a specific column in the whole alignment. The Shannon entropy is then calculated for each column, and after calculating the entropy of each column, the entropies are summed up together to give the final entropy of the alignment.

$$E(\mathcal{A}) = \sum_{c=1}^l \left(- \sum_{t=1}^d p(s_p) \log_2 p(s_p) \right) \quad (5.7)$$

Here \mathcal{A} is the input alignment, c is a given column in the alignment, t is a given position in each column and s is an IUPAC nucleotide code or a gap character '-' presented at each position in the column (site).

5.3 Results and discussion

Out of the 604 families showed in figure 4.4, 376 families were processed since the rest (237 families) have only one precursor with two annotated mature sequence. However, these 376 families contain ~72% of all metazoan families in MIRbase release 21. As mentioned before these families vary in size (number of precursors), and to make fair entropy comparisons, these families were grouped into 6 groups according to their size. Then the average old entropy and average new entropy of each group were compared and the results are shown in figure 5.1.

The distribution of the groups according to the number of precursors was as follows:

- Group 1: between 3 and 10 (included) precursors
- Group 2: between 11 and 20 precursors
- Group 3: between 21 and 40 precursors
- Group 4: between 41 and 100 precursors
- Group 5: between 101 and 200 precursors
- Group 6: more than 200 precursors

See also figures B.1, B.2, B.3, B.4, B.5 and B.6 for the entropy comparison of the families of each group.

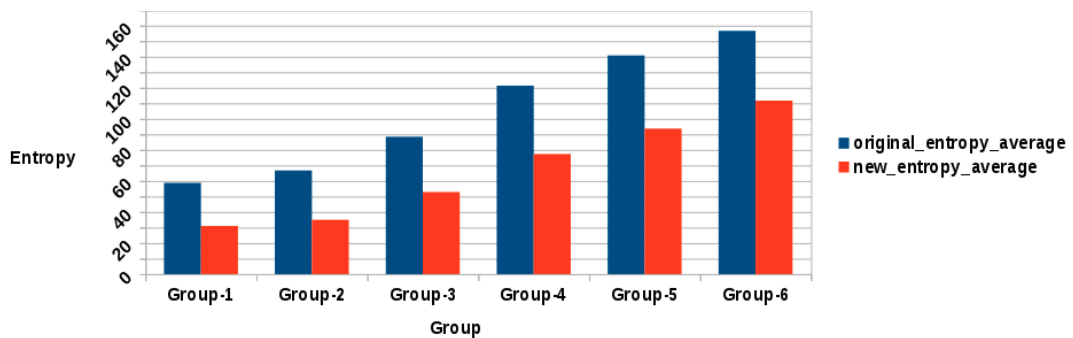


Figure 5.1: This figure shows the average entropy of each group, comparing the old entropy to the new. As mentioned before, the entropy decreases as the quality of the alignments improve.

Interestingly, this method improved all the families with a number of precursors greater than 40, i.e., Groups 4, 5 and 6. However, a number of families were not improved in the first 3 groups, resulting in improving a total of $\sim 0.95\%$ of the 367 processed families. For a deeper understanding, the number of the precursors that were totally removed was also checked. The highest number of removed precursors was from group 3 with 30 removed precursors, and the lowest was from the first group with only 8 precursors removed. Importantly, the number of removed precursors was not proportional to the original number of precursors. For instance, the original number of precursors in groups 5 and 6 was 2504 and 1726, respectively, and the number of the removed precursors was 11 for both. On the other hand, the groups 2 and 3 with a lower number of precursors, 1321 and 1688 respectively, were processed with 19 and 30 precursors removed. This is a clear evidence that quality of alignments doesn't depend only on the number of precursors, but it is also the quality of the precursors itself and how miRNAs are conserved in a given family.

Precursors tend to be more conserved in the big families since no sequences had to be totally removed. However, the average improvement of alignments was proportional to the sizes of the groups. This relation is understandable and can be explained as fewer sequences make less noise in the alignments. Based on these two observations, the relation between the family sizes and the number of precursors removed on the one side, and the relation between the sizes of the families and the average improvement of entropies on the other side, one can ask two questions. First, how should miRNAs families be classified (biologically, sequence conservation or both?) and their evolutionary conservation is also another question in this context. Second, assuming miRNAs are well classified, then what are the problems on the precursors level leading to bad alignments or low-quality alignments?

The starting point for this work were the computational problems that have a big impact on the analysis of microRNAs. Here a small test is done to see how one can, starting with the computational part, push the investigation into microRNAs forward. The microRNAs used in this study were *mir-3* and *mir-309* of the miRBase family MIPF0000140. All the sequences in this file belong to the Diptera group. These sequences were then used as query sequences in a `blastn` search with the default parameters against the nucleotide collection and the whole genome shotgun databases. Using this homology search, no significant hits were detected out of the Diptera group. Same sequences were used to build a covariance model with `infern` for both alignments, original and processed. After scanning Hexapoda genomes, specifically the *Apis mellifera* and *Nasonia vitripennis* species, a *mir-309* microRNA homologs could be detected. Both models, original and processed could be used to detect the same sequences. However, the E-value of the hit in *Apis mellifera* was better in the processed model than the original one ($7.3e-05$ in the processed and $2.5e-05$ in the original).

To bring this into an evolutionary context, the new hits were aligned to the processed alignment using `cmalign` and the evolution of these sequences was then depicted by the tool `ePoPe` (Hertel and Stadler, 2015). `ePoPe` implements a Dollo parsimony approach that determines the most likely point of origin (as the lineage leading up to the last common ancestor of all observed paralogs) as well as the most parsimonious scenario of duplications and losses in each family. As expected, the computational approaches were able to positively contribute in the evolutionary study, as the origin of the *mir-309* was moved from Diptera level to at least *Endopterygota* level, see figure 5.2.

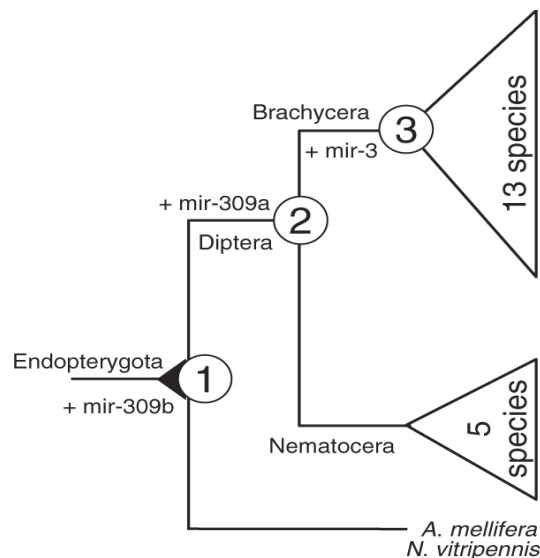


Figure 5.2: Distribution of miRNA homologs of miRNA families mir-3.

5.4 Conclusion

In this part of my work, the focus was only on families with two forms of precursors, with one or with two annotated mature sequences. With the help of `infernal` the quality of the alignments in a number of families could be improved. However, although some problems could be overcome, a lot still remain. Even though the main goal of this work was accomplished, it was also clear that trying to improve the alignments from the available seeds is not enough, as the built consensus models were based on data that might contain false positive information (i.e. wrong oriented precursors). This issue can be covered only by checking each precursor of a given family alone, regardless of the other sequences in the alignment. The next chapter will introduce the reader to a pipeline that addresses this problem. From this work quantitative and qualitative knowledge about the data in `miRBase` was derived, as well as about pre-miRNA and miRNA annotation in general. One of the facts discovered is that the last common ancestor of a `miRBase` family is frequently older than what would be inferred from raw `miRBase` data. At the end and as expected, this preliminary work provided evidence that a fully automatized

pipeline to process miRBase alignments into a complete and consistent set of miRNA sequences is feasible.

Chapter 6

MIRfix pipeline

6.1 Introduction

Based on the structure features, sequences information and alignments, the goal is to correct wrong sequences and decrease the noise in the data. The **MIRfix** pipeline simultaneously works at two levels. The first level is called the precursors level, as it focuses on the precursor itself despite the consensus alignment. The second level is the level where the consensus alignment of a group of precursors or microRNA family is considered. Sequence-based improvements by evaluating each precursor alone was achieved by removing wrong sequences or replacing them with corrected ones according to defined criteria, applying the aforementioned tools. The focus here lies on the sequence information of the precursor and its corresponding mature sequences as well as the structural characteristics of the precursors. Alignment-based improvements were implemented by trying to fix the ends of the alignments and correctly aligning mature regions of precursors with a minimal number of removed sequences. This is done based on an automated approach for detection of misaligned sequences and consequently, changing parts of the sequence if needed. Figure 6.1 shows the general workflow of **MIRfix**.

At the precursors level, the input sequences can be divided into three main categories: Precursors with only one mature sequence, precursors without a given or annotated mature sequence and precursors with two annotated mature sequences. At this level, only the first and the second categories are processed. Precursors without an annotated mature sequence are processed alone and mature sequences are predicted in reference to the mature sequences of the other precursors of corresponding family. Each precursor of one annotated mature sequence or newly predicted mature sequence, is compared to a modified version. In the modified sequence, the orientation of the mature sequence is changed, i.e., if the mature sequence is at the 3' end of the original precursor, then the modified version contains the mature sequence at the 5' end, and vice versa. Consequently, also part of the original sequence will be changed. According to certain criteria explained in the next section, the original precursor will be kept or replaced by its modified version. It is also possible that neither the original or the modified sequence are added back to the alignment. After the comparison, the second mature sequence (miRNA*) is predicted, so all the "non-removed" precursors of the first and

second categories will have two annotated mature sequences. Of course, precursors of the second category (lacking any originally annotated mature sequence) where no mature sequence could be predicted, are not processed further and they are discarded in the next steps and removed from the final alignment as well as those precursors with none of their two versions accepted.

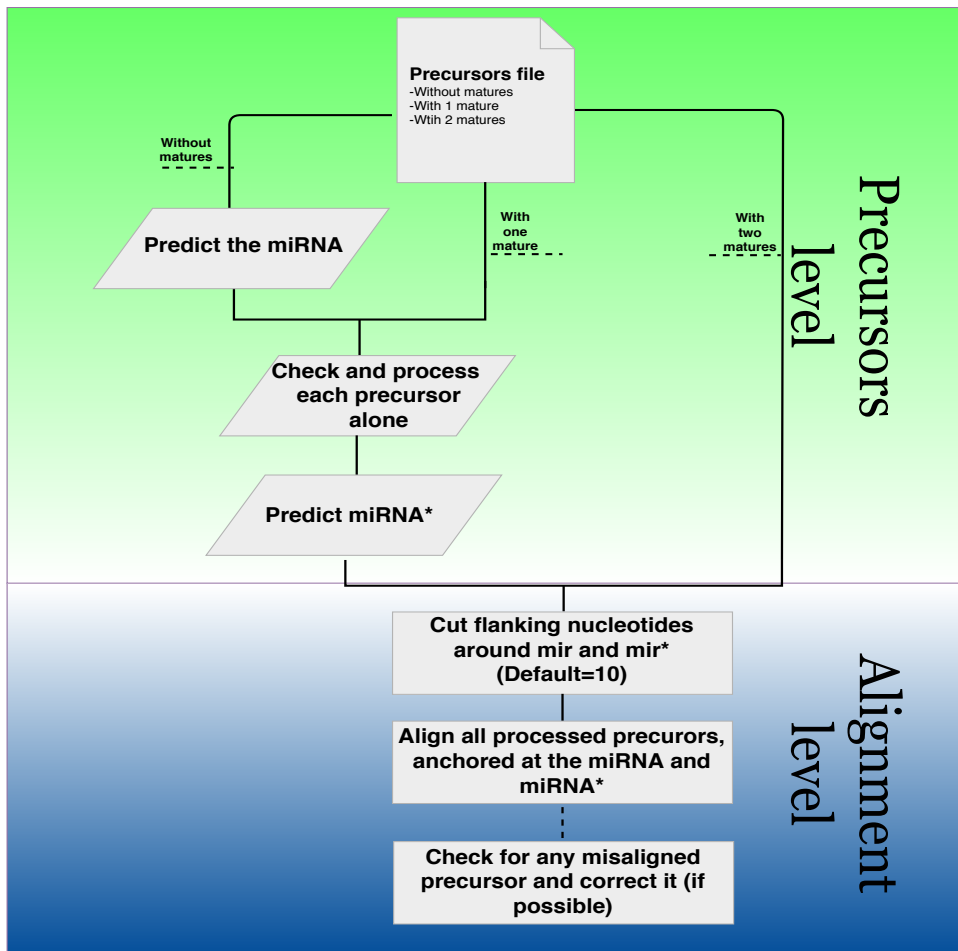


Figure 6.1: This figure shows the **MIRfix** general workflow as it is divided into two levels, precursors' level and alignment level. New alignments are created at the first level after processing the precursors independently of each other. The produced alignments at the previous level are processed again in the alignment level.

Next is the alignment level, which considers all the successfully processed precursors of the first level, in addition to the precursors with two annotated mature sequences (third category). As a first step at this level, all precursors are fixed into a defined number of flanking nucleotides upstream of the mature sequence of the 5' arm and downstream the mature sequence of the 3' arm. The number of these flanking nucleotides is by default 10, or can be a user-defined number between 10 and 50. All the precursors are then considered together by processing the consensus alignment of a given group of microRNA precursors or a family of microRNA precursors. The aim is

to find the best possible consensus alignment by detecting any misaligned precursors. A misaligned precursor, is any sequence that doesn't fit to the consensus alignment. In this case the misaligned sequence is replaced, either by the modified version if the original sequence was kept at the first level of the workflow, or by the original sequence if it was already changed to the modified version at the first level. The alignment at this level is always anchored at the mature sequence of the precursors using `DIALIGN`.

6.2 Methods

6.2.1 Inputs and Outputs

This pipeline requires 4 different input files.

Precursors file: A fasta file contains precursors of the same family or group.

Mature sequences file: A fasta file contains all mature sequences related to the precursors' file(s).

Mapping file: A *txt* file mapping all the precursors, mature sequences and the file or family names.

Genomes file: A *txt* file contains all the genomes of the precursors

Two directories:

The output directory: A location specified by the user to write the outputs in.

The files directory: The location of the precursors' file or families, as they should be stored in the same location.

A number of output files are produced, but here I only mention the summary file. It is a file summarizing all the results of the run, for instance, the number of changed sequences, the removed sequences, the old and new alignment entropies and some other statistics. For the rest of the output files see appendix B.

6.2.2 Prediction of the mature sequences

miRNA prediction

As the mature sequences are homologs, conserved with high similarity, we do expect that the mature sequence of a given precursor is located in a region homolog to a known miRNA(s). Moreover, this possibility can increase when the precursors belong to the same family. From that perspective, the known mature sequences are used as query sequences to predict the mature sequence of the precursors without an annotated mature sequence. Using `clustalw`, for each given precursor without an annotated mature sequence, a pairwise alignment is predicted where every given mature sequence belongs to the same family. The best aligned mature sequence is selected and used as a reference to predict the mature sequence of the subject precursor. The selection of the reference sequence is based on the alignment score calculated by `clustalw`. Obviously, the mature sequence with the highest alignment score always is chosen but only when it is equal or over the threshold. This threshold is equal to 21, and this threshold is chosen based on the observation

of the miRBase data. This threshold was defined after aligning all mature sequences of the biggest miRBase family *MIPF0000001*, and the minimum score was 18.

After selecting the reference mature sequence, a window of corresponding size is moved all along the subjected precursor. For each window, the exact matches between the selected miRNA and the given region of the precursor, are counted, in addition to counting the number of the nucleotides (non-gapped positions) of the precursor's region only. Always, the window with the highest number of matches is considered as the region where the new miRNA is located. Although, more than one window with the same number of matches can be found. If this is the case, then the counted number of nucleotides (non-gapped positions) is used to favorite one of the windows of equal exact matches, over the other(s).

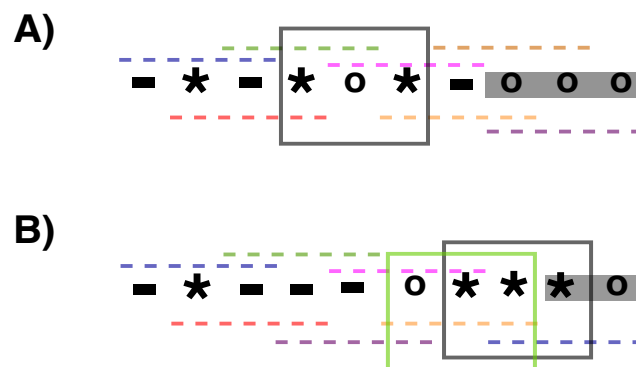


Figure 6.2: A) and B) are two different cases in miRNA prediction. The dotted lines are all the windows searched for the miRNA. The boxed windows are the windows were chosen as miRNA window. The grey regions are the loop regions. Stars are exact matches, circles are non-matching nucleotides and dashes are gaps. In B), the best box is shifted back one nucleotide to avoid the loop.

What if part of the best window was located in the loop of the precursor as shown in figure 6.2-B ?. This case is considered as follows, first, the perfect position of the mature microRNAs is exactly in one of the stems and second, it is more practical when predicting the miRNA*. For this, the window is shifted back by the number of nucleotides located in the loop when the new miRNA is located at the 5' arm, and shifted backward when the new miRNA is located at the 3' arm of the precursor.

Finally, when the new miRNA is predicted, it is added to the input files *mature* and *mapping*. The newly added records are mapped following the input mapping file, which makes it possible to re-use the new records easily. In addition to the main goal, the advantage of this strategy is that it can be an indicator that a given sequence is not a real precursor or less likely it is a precursor that doesn't belong to any of the families of the query mature

sequences. This can be deduced when none of the mature sequences has an alignment score equal or more than a defined threshold.

Predicting of the second mature sequence (miRNA*)

The proper secondary structure of every pre-miRNA is a hairpin shape, producing two almost perfect complementary base-pairing mature sequences and each of them in a different arm (on the 5' arm and 3' arm). This step is applied for all precursors with only one mature sequence, (originally annotated or predicted mature sequence). The miRNA is considered as a fixed window of nucleotides on one of the two arms of the precursor. Theoretically, a complementary window at the opposite arm should be representing the corresponding miRNA*, setting its size equivalent to the size of the miRNA window.

In case miRNA is at the 5' arm

The miRNA start position:*

If the miRNA window ends with a base-pairing nucleotide, then the miRNA* window (at the 3' arm) starts at position 2 nt downstream the nucleotide base-pairing to the last nucleotide of the miRNA window. If the miRNA ends with a non-base-pairing nucleotide, miRNA* window starts at the position m and this position is calculated as

$$m = F - x + 2 \quad (6.1)$$

where:

F : is the position of the first nucleotide at the 3' arm, base pairing to the last nucleotide in the miRNA window.

x : is the difference between the position of the last nucleotide (non-base-pairing) and the position of the last base-pairing nucleotide, in the miRNA window.

The miRNA end position:*

If the miRNA window starts with a base-pairing nucleotide, then the miRNA* window (at the 3' arm) ends at position 2nt downstream of the nucleotide base-pairing to the first nucleotide of the miRNA window. If the miRNA starts with a non base-pairing nucleotide, miRNA* window ends at the position m' and this position is calculated as

$$m' = F' + x' + 2 \quad (6.2)$$

where:

F' : is the position of the last nucleotide at the 3' arm, base pairing to the first nucleotide in the miRNA window.

x' : is the difference between the position of the first nucleotide (non-base-pairing) and the position of the first base-pairing nucleotide, in the miRNA window.

In case miRNA is at the 3' arm

The miRNA start position:*

If the miRNA window ends with a base-pairing nucleotide, then the miRNA* window (at the 3' arm) starts at position 2 nt downstream the nucleotide base-pairing to the last nucleotide of the miRNA window. If the miRNA ends with a non-base-pairing nucleotide, miRNA* window starts at the position p and this position is calculated as

$$p = L - y + 2 \quad (6.3)$$

where:

L : is the position of the first nucleotide at the 5' arm, base pairing to the last nucleotide in the miRNA window.

y : is the difference between the position of the last nucleotide (non-base-pairing) and the position of the last base-pairing nucleotide, in the miRNA window.

The miRNA end position:*

If the miRNA window starts with a base-pairing nucleotide, then the miRNA* window (at the 3' arm) ends at position 2nt downstream of the nucleotide base-pairing to the first nucleotide of the miRNA window. If the miRNA starts with a non base-pairing nucleotide, miRNA* window ends at the position p' and this position is calculated as

$$p' = L' + y' + 2 \quad (6.4)$$

where:

L' : is the position of the last nucleotide at the 5' arm, base pairing to the first nucleotide in the miRNA window.

y' : is the difference between the position of the first nucleotide (non-base-pairing) and the position of the first base-pairing nucleotide, in the miRNA window.

Considering the AGO process

In addition to Dicer processing, also AGO processing was considered (explained in [chapter 4](#)). The miRNA and miRNA* 5' ends are compared by means of stability, by comparing the nucleotides base-pairing at both ends. The 5' end of the miRNA* is always checked if it is A-U base-pairing or a non-base-pairing nucleotide. If this is the case, then the predicted miRNA*

is trimmed recursively by one nucleotide until a G-C base-pairing is found. However, this trimming reaches a cutoff equal to 3 nt, to keep as much as possible the characteristics of the Dicer process.

Finally, these predicted miRNAs* are added to a separate file recording the mapping between miRNA* IDs and corresponding precursors. A fasta file containing all the miRNA* sequences is created and added to the output folder of each family or input file.

6.2.3 The original precursor and its alternative

This step is one of the major steps in **MIRfix**, it is even the most critical step in the whole workflow. Here the precursor itself is studied ignoring all other precursors and the consensus alignment. Each precursor is studied alone considering its presence in the source genome. Then, changes are applied to the *original precursor*. For precursors with one annotated mature, and *alternative* version of the former is considered as well.

Changes happen in the number of the flanking nucleotides upstream of the mature sequence when it is at the 5 prime end of the original precursor, or downstream of the mature sequence when it is at the 3 prime end. The number of the flanking nucleotides can range from zero to ten nucleotides. The number of nucleotides here is limited to a maximum of 10. When the number of the flanking nucleotides is greater than 10 they are cut. For precursors having less than 10 flanking nucleotides, the original nucleotides are not changed.

Alternative precursors are derived from the original ones, retaining part of the sequence without any changes. Nevertheless, a notable part of this sequence is changed, as well as the arm's relative position of the mature sequence. The orientation of the precursor and the mature sequence is changed from 5' to 3' or vice-versa. If the mature sequence is for example located at the 5' end of the original precursor, it will appear at the 3' end of the alternative version. In principle, the precursors that undergo this process, have to be found in a source genome fasta file, uploaded into the pipeline. Once the original precursor is found in the source file, a number of nucleotides equal to 250 nucleotides are added to both sides of the precursor to avoid any error when working on very long sequence(s).

By default, 10 flanking nucleotides upstream or downstream of the mature sequence. In both orientations (miRNA at 5' or 3' end), the number of flanking nucleotides left to the further steps of the workflow is equal to 10. For a precursor with a mature sequence at 5' end, the number of the nucleotides upstream of the miRNA is checked, if the number of the nucleotides in the original sequence is greater than 10 nucleotides, then the sequence is cut until only 10 flanking nucleotides left. On the other hand, if the number of the nucleotides is less than 10 nucleotides, these flanking are extended into 10 from the upstream 250 nucleotides that were extracted from the related genome. For a precursor with a mature sequence at the 3' end, the same procedure is applied with the main difference, that extending or cutting the flankings, is applied at the downstream region of the 3' end mature sequence. 10 nucleotides is the

default number used in the pipeline, and this number can be defined by the user without exceeding 50 nucleotides.

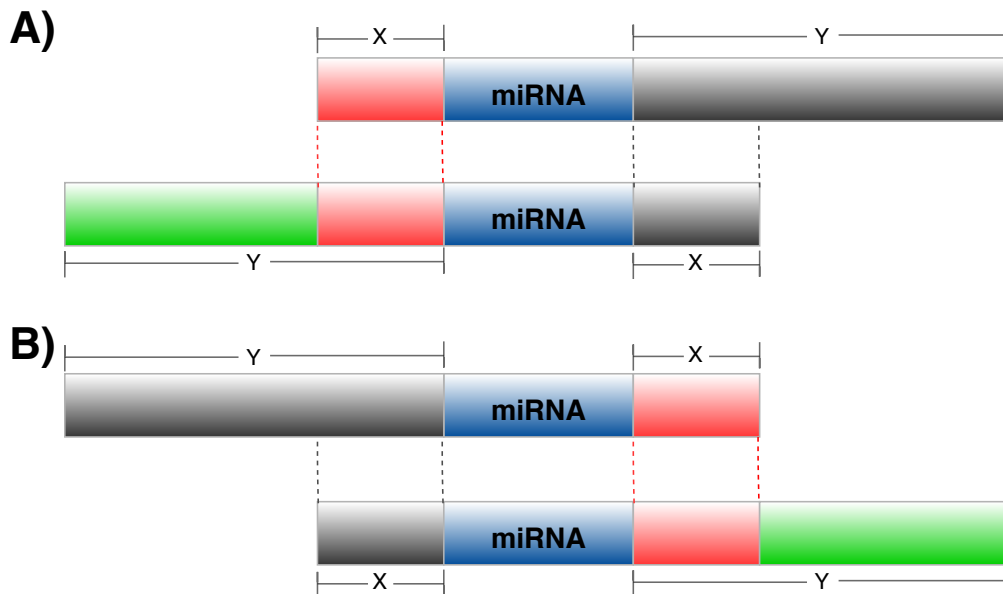


Figure 6.3: A) and B) Shows how the original precursor (above) is changed to the new precursor (below), X and Y are fixed lengths. The regions with the same colors are regions with the same nucleotides, i.e., just repositioned sub-sequences. The green part is the new extended part of the sequence from its genomic source. A) represents a precursor with a miRNA at the 5' prime end changed to be at the 3' prime end in the new sequence. B) represents a precursor with miRNA at the 3' prime end changed to be at the 5' prime end in the new sequence.

Defining the modified version, i.e., the alternative candidate sequence. The alternative sequence has the same length as the original sequence, but the position of the mature sequence is changed. When the mature sequence of the original precursor is at the 5' end, the sequence is divided into three regions denoted as X , Y and m . X is the number of nucleotides in the upstream region of the mature sequence, Y is the number of nucleotides in the downstream region of the mature sequence and m is the mature sequence itself. To change the original sequence into new modified version, first, the upstream region X is extended by adding more nucleotides upstream, until X is equal to Y . Second, the mature sequence is not modified so the length m won't change. Third, Y is cut from the end so that Y is equal the original length of X . As an example and as the default the number of the flanking nucleotides equal to 10, at the end of the above process (in figure 6.3) the length of Y is equal to 10. Although, this number can be any another number between 0 and 50 depending on the user's choice. The mature sequence is shifted from the 5' end in the original sequence to the 3' end in the alternative version. The nucleotides of the X region of the alternative sequence come from the Y region of the original precursor with a length between 0 and 50 (10 by default). The same when

the mature sequence of the original precursor is located at the 3' end of the sequence, but with X as a number of nucleotides downstream the mature sequence and Y as the number of nucleotides upstream the mature sequence. So, the mature sequence will be shifted from the 3' end of the original sequence to the 5' end of the alternative version. N.B. the extended nucleotides at the X region always come from the flanking 250 nucleotides extracted already from the source genome. Figure 6.3 demonstrates both, alternative 5' oriented precursor and alternative 3' oriented precursors.

6.2.4 The validation of the precursor

The decision of choosing one precursor over the other is based on sequence-based information and the secondary structure information. If possible, the original precursor is always preferred. The comparison points fall into 4 main criteria:

1. Stability of the secondary structure
2. Mature sequence position
3. Number of hairpins in the secondary structure
4. Number of nucleotides of a given hairpin

The stability of the pre-miRNA is necessary for the exportation process from the nucleus and this was represented by the MFE (minimum free energy) calculated with `RNAfold` from the `ViennaRNA` package. The more stable a secondary structure is the lower the MFE becomes. The MFE is used as the first and main factor when comparing the candidates, preferring the most stable one. The other factors are discussed in the following.

The position of the mature sequence can explain a lot about the precursor, since it is known to be located exactly in one of the arms of the hairpin. To this end, the position of the mature sequence for each given precursor is considered, with all its possible positions. There are 3 possibilities here, (a) The mature sequence is located exactly in one of the two arms, (b) Part of the mature sequence is located in the loop and (c) All the mature sequence is located inside the loop. In this pipeline, (a) is always accepted as it is the perfect possibility, sometimes (b) is accepted with some exceptions (explained in 6.2.4), where (c) is always considered as a wrong position.

One or more hairpins can define the secondary structure of a precursor, where the perfect shape is when the pre-miRNA sequence folds into one clear hairpin. The status of the precursor is considered carefully as the precursor can fold into more than one hairpin. This can be caused by unreasonable elongation of the precursor, thus, it might then contain more than one hairpin, but only one hairpin holds the miRNA and miRNA*. Figure 6.4 shows the different scenarios of foldings. Figure 6.4-A shows structurally perfect hairpin shape but not necessarily a true pre-miRNA because it depends on the position of the

mature sequence. In the used classification, a precursor with a similar shape is considered as a wrong precursor and excluded from the study if the mature sequence was covering all of the loop (the red region) without any nucleotide being present in any of the arms (black nucleotides). Figures 6.4-B and 6.4-C are secondary structures with more than one hairpin, and these structures are treated as special cases. Such structures will be processed in a way that each hairpin in the structure will be considered separately. In figure 6.4-B, the one structure gave two hairpins (black precursor and green precursor) and each of them is treated separately in the same way explained above for figure 6.4-A with an additional constraint for both hairpins together. In this case, it is likely that the mature sequence is shared between the two hairpins, i.e., part of the mature sequence is located in the black hairpin and the other part in the green hairpin. In this situation, the sequence is also considered as wrong precursor and excluded and removed from the data. Structure of figure 6.4-C is processed in the same way as in figure 6.4-B with only one difference, the precursor is removed in case the mature sequence is located in any of the red arms without any further checking.

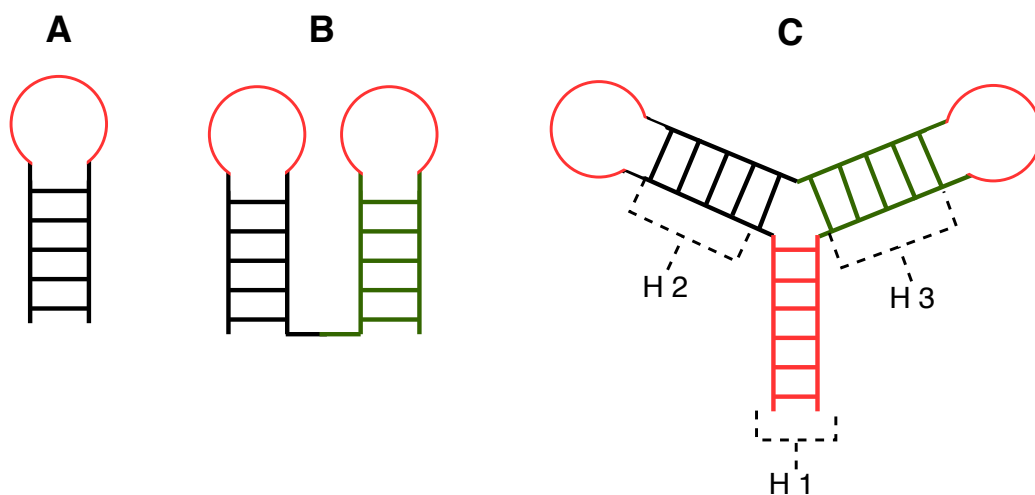


Figure 6.4: Different folding states for the annotated precursors. In all cases, the red region is a forbidden region for the mature sequence(s). In B and C, it is also not allowed to be shared between the black and green precursors.

The number of the base-pairing nucleotides is counted for each valid hairpin. In typical hairpin structures where the sequence folds into one hairpin, all the base-pairing nucleotides of the structure are counted. For the sequences folding into more than one hairpin the counting is a bit different, where for each valid hairpin the number of base pairing nucleotides is counted separately. For example in figure 6.4-C h2 and h3 are valid regions, while h1 is an invalid one. If the mature sequence is h2 for instance, then the base pairing nucleotides from the h2 region are counted. In the next explanation I will refer to these counts as *Ncounts* and *Ocounts*, the counts in the alternative sequence and the original sequence, respectively.

Comparative method for the original precursor and its alternative

For the three main cases already discussed, specific terms will be used in the following to make it easier in the explanation of the comparison constraints. The term *broken* is used when the sequence is excluded and removed, so it is used when the mature sequence is located in the loop (whole loop), shared between 2 hairpins or in both arms as shown in figure 6.4-C. The term *bad position* is used when the mature sequence is partially located in the loop. For cases where the mature sequence is located exactly in one of the arms, the term will be *True position*.

In the comparison, if the sequence is *broken* then it is directly excluded from any more comparison and the other precursor is checked if it is valid or not. When both the original precursor and the alternative precursor are *broken* then no further comparison or checking is conducted. When both, alternative and original, are not broken are then they are compared together, even if one of them or both have the *bad position* state. Once both precursors are valid, the comparison starts from the MFE score and in the following the explanation will be separated into two main paragraphs, the first when the original precursor has better MFE which is smaller than the MFE of the alternative precursor that means it is more stable. The second paragraph, is the opposite case, the MFE of the alternative sequence is better than the one of the original.

The MFE of the original is better... when the original sequence is in the *true position* state, and it has only one hairpin (i.e. one part) and the alternative sequence is in any another state, the original precursor is always chosen. Also when the original sequence is in the *bad position* state and one part, it always was chosen and reported in the summary file as *bad positioned* for all the states of the alternative sequence, but with only one exceptional case of the alternative sequence. This case happens when the alternative sequence has only one part and *true position* state and an MFE score less than -10.0 kcal/mol and the number of base pairing nucleotides of the alternative sequence is greater than the one of the original precursor. These constraints are added because the priority is always for keeping the original sequence and changing it only when the alternative is almost perfect. However in this case, if the alternative sequence status isn't conform with the additional constraints, the original sequence is kept and it is reported in the summary file as a *bad positioned sequence*. Continuing with the original precursors of one hairpin, for which they are in the *broken* state, they are either totally removed (i.e. not replaced) or replaced with the alternative sequence. In two cases the precursor is totally removed and not replaced with the alternative sequence, it's when the alternative sequence is also in *broken* state. In the other states, it is possible to replace the original by the alternative under one additional condition, that the alternative sequence should have MFE score less than or equal to -10.00 kcal/mol.

The MFE of the alternative is better... The original precursor is always replaced by the alternative sequence when the alternative is in a *true position* and the secondary structure of the alternative sequence is built up of only one hairpin.

That is the only case when the alternative sequence is definitely chosen over the original one. In all other cases, either the original is kept, both removed or the alternative sequence is chosen after adding additional constraints for the comparison between the original and the alternative sequence.

To avoid the repetition in the text, table 6.1 summarizes all the possible cases covered by the workflow. The table is divided into two main parts, the first one is when the original sequence has better MFE score, and second part when the alternative MFE score is better. Each part in turn is also split into two sub-parts. The first part includes the comparisons when the original sequence has better MFE score than the alternative version. One of the two sub-parts refers to when the original sequence folds into one hairpin and the other one is when the sequence folds into more than one hairpin, named respectively, parts=1 and parts > 1

Upstream and downstream of the mature sequences

In between processing the precursors and the final alignment, all precursors are fixed at the end, and this step includes all the valid precursors. Of course, at this step all precursors are with two defined mature sequences. For the consistency between sequences and improving the final alignments, the end of the sequences is defined. For all precursors, the number of the flanking nucleotides is modified to 10. For sequences with more than 10 flanking nucleotides at any of the two mature sequences upstream or downstream, the number is decreased to 10. For sequences with less than 10 flanking nucleotides the number is extended by retrieving the nucleotides from their given source genomes. Thereby 10 flanking nucleotides is the default and can be any other user-defined number between 0 and 50 nucleotides.

6.2.5 Alignment processing

Primary alignment

After all processes at the precursors level are finished, a primary alignment is constructed for each group of precursors or family. Here an anchored alignment is applied using `DIALIGN`, and this type of alignment is applicable at this stage because all precursors will have the miRNA and miRNA* defined. Thus, the alignment is anchored at the miRNA and miRNA* which is an important step, in order to detect any misaligned sequence as explained in the following, as well as improving the alignment quality.

Final alignment and misaligned sequences

First, let's define the misaligned sequences. As the mature sequences are very similar especially in the precursors that belong to the same family or group, one expects to see both ends and the mature sequences of any sequence are consistent within the consensus alignment. In some cases, the mature sequences are aligned to the opposite end of the alignment. For instance, if the

Original score is better than the new score (original MFE < new MFE)							
Original/New	Parts=1			Parts>1			
Parts = 1	FF	Original	Original	Original	Original	Original	Original
	FT	if: new score <= -10.0 and Ncounts > Ocounts =>new else: Original (BP)	Original (BP)	Original (BP)	Original (BP)	Original (BP)	Original (BP)
	TF & TT	if: new score <= -10.0 =>new else: Both excluded	if: new score <= -10.0 =>new else: Both excluded	Both Excluded	if: new score <= -10.0 =>new else: Both excluded	if: new score <= -10.0 =>new else: Both excluded	Both Excluded
Parts > 1	FF	if: Ncounts > Ocounts =>new else: Original	Original	Original	if: Ncounts > Ocounts =>new else: Original	Original	Original
	FT	if: new score <= -10.0 and Ncounts > Ocounts =>new else: Original (BP)	Original (BP)	Original (BP)	Original (BP)	Original (BP)	Original (BP)
	TF & TT	if: new score <= -10.0 =>new else: Both excluded	if: new score <= -10.0 =>new else: Both excluded	Both Excluded	if: new score <= -10.0 =>new else: Both excluded	if: new score <= -10.0 =>new else: Both excluded	Both Excluded
New score is better than the original score (new MFE < original MFE)							
Parts = 1	FF	New	New	New	New	New	New
	FT	Original	Original (BP)	if: new score <= -10.0 =>new else: Both excluded	Original	Original (BP)	if: new score <= -10.0 =>new else: Both excluded
	TF & TT	Original	Original (BP)	Both Excluded	Original	Original (BP)	Both Excluded
Parts > 1	FF	Original	if: new score <= -10.0 =>new else: Original (BP)	if: new score <= -10.0 =>new else: Both Excluded	if: new score <= -10.0 and Ncounts > Ocounts =>new else: Original	if: new score <= -10.0 and Ncounts > Ocounts =>new else: Original (BP)	if: new score <= -10.0 =>new else: Both Excluded
	FT	Original	Original (BP)	if: new score <= -10.0 =>new else: Both Excluded	Original	Original (BP)	if: new score <= -10.0 =>new else: Both Excluded
	TF & TT	Original	Original (BP)	Both Excluded	Original	Original (BP)	Both Excluded

Table 6.1: This table shows the comparisons between the original sequence and the new sequence when the original MFE is better than the new MFE and vice versa. The *F* and *T* alphabets are *False* and *True*. The first character always refers to Broken or not (*T* or *F*) and the second refers to In loop or not (*T* or *F*). For example: if not broken and In loop then it is *FT* where the not broken and not in loop is *FF*. *BP* is the abbreviation for “Bad Positioned” mature sequence.

consensus alignment with a given mature sequence at its 3' end and the same mature sequence of a given precursor is located at the 5' end of the precursor, then this precursor is misaligned because it is shifted to the opposite end of the consensus alignment. For example, the precursor *ptr-mir-453* in figure 6.10.

This step is important as it is double checking the precursors after they are processed separately. One of the possible scenarios that could happen, is that the original precursor and the alternative version are real competitors and in cases where the original precursor is not replaced, but the alternative precursor is the correct one, the opposite scenario is also possible. In such cases, the anchored alignment can judge on that by detecting the misaligned precursor. When the misaligned precursor is detected it is replaced then by the alternative sequence if the misaligned is the original precursor. If the misaligned sequence was the alternative version, then the original precursor is added back to the alignment. Albeit, it is possible to leave the detected misaligned sequence when the other version is not a potential candidate.

Detection of the shifted sequences

As shown in figure 6.10, when a sequence or more are misaligned in the consensus alignment, the alignment looks like two blocks. One of the 2 blocks contains the majority of the sequences or the aligned nucleotides, and the other one contains the majority of gapped positions and few nucleotides and sequences. When an alignment is divided in such a way, the sequence(s) in a region with a majority of gaps is a misaligned sequence. this was implemented in a computational method to detect the misaligned sequences. The alignment is always divided into two halves, and for each half, the average number of nucleotides and the average number of gaps is counted. Then for each sequence, the number of the nucleotides is counted in each of the halves. If the number of nucleotides in a given a half of a sequence much greater than the average and, the number in the second half is less than half of the average of the second half, or zero. The following equations show when a sequence is calculated as a misaligned sequence. Denote by n_1 and n_2 the number of nucleotides of the query sequence in each part, and by N_1 and N_2 the corresponding average number of nucleotides in a row. The sequence is considered as misaligned whenever

$$\begin{aligned} n_1 > \min(1.5N_1, 0.7(N_1 + N_2)) \text{ and } n_2 < 0.5N_2, \text{ or} \\ n_2 > \min(1.5N_2, 0.7(N_1 + N_2)) \text{ and } n_1 < 0.5N_1. \end{aligned} \quad (6.5)$$

6.3 Results and statistics

This pipeline was applied to the miRBase data as it is the biggest available registry for the annotated pre-miRNA (precursors) and their related miRNAs. **MIRfix** was applied for all families, metazoan and non-metazoan. However, in this first version of this pipeline, the focus on the metazoan families in

principle.

Out of the total 1415 metazoan families, the Shannon entropy of 1104 families was decreased showing improvement in the quality of the alignments of these families, which is around 78 % of total metazoan families. For the rest, (311 families) the entropy was either the same or increased. The negative change can be explained by the means of changing a precursor or because of using a different aligning method (anchored alignment) that forces specific regions (mature sequences) to align together. On the other hand, the improvement was either due to the replacement of the original sequences with the new alternative precursor, which shifts the annotated mature sequence from one end to another, or due to the modification of the length of sequences.

The total number of the original precursors which had high potential alternatives at precursors level was 259 out of a total number of metazoan precursors equal to 14712, these precursors were detected in 152 families out of the 1415 metazoan families. However, as this pipeline works at two levels, some of these replaced precursors at the precursor level, were changed back to the original precursors because they did not fit to the consensus alignment of their families. Of course, these precursors were changed back to the original sequences, only when the original sequence is a potential precursor and not classified at precursor level as *broken* (to remove). Figure 6.5 show the distribution of these 259 precursors. Not only the changed sequences, but also the original sequences are checked at the alignment level. For those sequences which were not changed at the precursor level and did not fit to the final alignment, were replaced by their potential alternatives but only when it exists. The number of these sequences was 149 precursors in 24 families, summing up the final precursors replaced by their alternatives to 312 precursors. The maximum number of original sequences replaced by alternative in one family was 8 sequences in the improved sequences and 5 in the non-improved families.

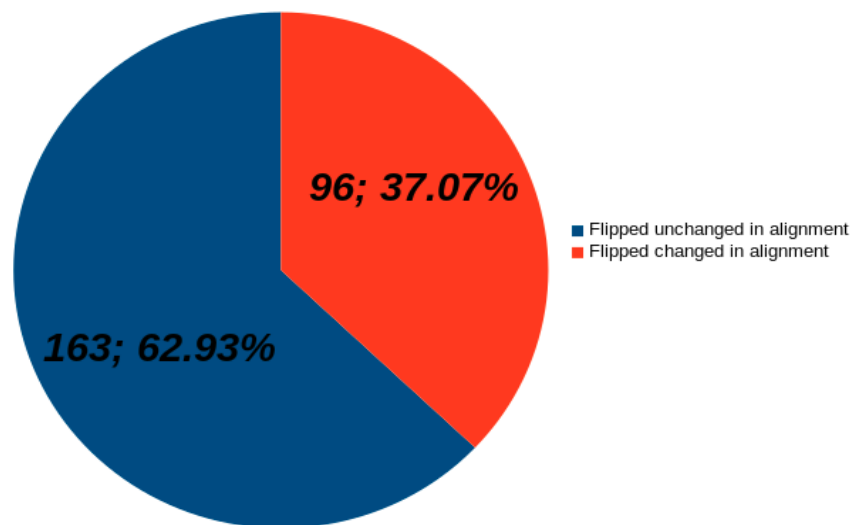


Figure 6.5: This figure shows the number of precursors which were changed to their alternatives (alternative). The unchanged, are the precursors which did not change at the alignment level. The changed are the precursors that are changed at the alignment level.

For the precursors which were not replaced by an alternative, 247 precursors were totally removed as neither original nor the alternative showed an acceptable pre-miRNA structure according to the strategy explained before. For the sequences which were retained (were not changed or removed), the modification of these sequences was checked in the sense of trimming and extension, in relative to the default 10 flanking nucleotides upstream and downstream the mature sequences. Total number of these sequences was 14094 precursors, 6419 were extended with an average of ~ 18 nucleotide extended per sequence, and 7675 were trimmed with a trimming average ~ 15 nucleotide per sequence.

Detailed numbers

Here is a discussion of more detailed numbers to show the feasibility of the methods of replacing the original sequences with their alternatives and checking the misaligned sequences. To this end, a comparison between the numbers of the improved and non-improved families is done. Concerning the replaced sequences at precursors' level, the ratio of the changed sequences in the non-improved families is ~ 1.4 per family and ~ 1.9 per family in the improved families. These numbers show that the method used to replace the sequences is not negatively affecting the alignments, but positively. Furthermore, the average of the removed sequences out of total removed sequences was checked, only for the families that had replaced sequences. The average of removed sequences in the non-improved families was ~ 0.73 and ~ 0.34 in the improved families. This shows that the main reason of improving the families was not the removal of the families, but the shifting of the mature from one end to another, which is the replacement of the original precursor with its alternative. Moreover, The percentage of the alternative sequences

which was changed back to the original when detected as misaligned at the alignment level, was calculated as well. This percentage in the non-improved families was ~ 0.16 and ~ 0.49 in the improved ones. Latter numbers suggest that checking even the alternatives at the alignment level is important, and the pipeline can also clean any possible noise produced by the correction at the precursors' level. The average in the non-improved families is low because the original sequences do not show a potential pre-miRNA structure.

6.4 Applications

6.4.1 Real life examples and artificial data tests

In this section are the real-life outputs of the pipeline, after its application on the metazoan miRNA families, in addition to an artificial data was used to test the performance of the pipeline.

In the following are different examples showing the effectiveness of the pipeline at precursors' level, alignment level and in the combination of both levels. The examples shown here are from different families.

Correction at precursors level

From the miRBase family *MIPF0000018* which is the *mir-154* microRNA family, here are two different examples of corrected precursors at precursors' level, alignment level and the combination of both. This family is part of the imprinted DLK1-DIO3 "mega-cluster" and has important functions in embryonic development and the epithelial to mesenchymal transition (Benetatos, Vartholomatos, and Hatzimichael, 2014; Dill and Naya, 2018).

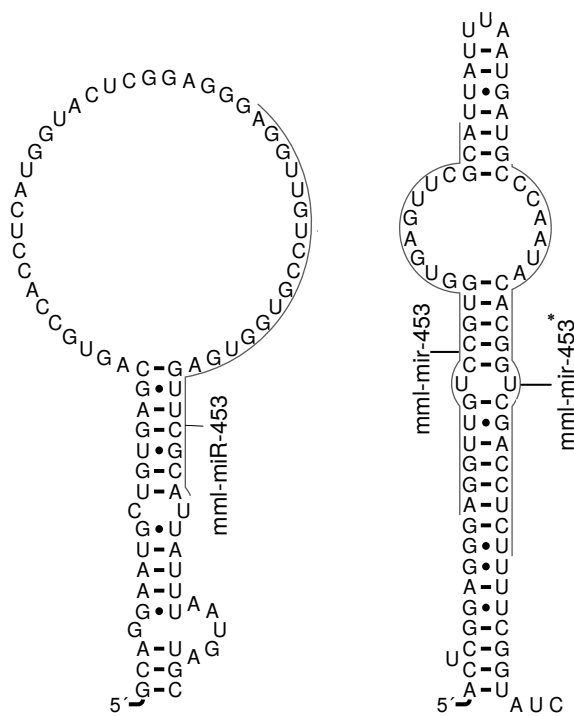


Figure 6.6: Comparison of the secondary structure and miRNA/miRNA* positions of mml-miR-453 as annotated in miRBase (l.h.s.) and the shifted alternative identified by **MIRfix**.

Figure 6.6 shows how the precursor of ID *MI0007754* (mml-mir-453) from the macaque (scientific name: *Macaca mulatta*) is corrected. Originally, the mature sequence of this precursor was annotated to the 3' stem of the precursor. After processing this family in **MIRfix**, this precursor was replaced by an alternative one, suggesting that the mature sequence should be at the 5' stem. In this case, the original sequence should be shortened from the 5' end and extended from the 3' end. The pipeline corrected it in this way by trimming the 5' end into 10 nucleotides upstream the mature sequence and the 3' end was then extended based on the source genome of this sequences. As shown in the figure, the general structure of the hairpin is totally improved comparing to the original structure that included the biggest part of the mature sequence in the loop. The corrected precursor not only looks like the correct generic structure of a pre-miRNA, but also it conserves the 2 nucleotides overhang at the 5' ends of the mature sequences. Interestingly, this precursor was corrected in by miRBase community exactly as this pipeline, did and the record was merged to another one, see figure 6.7.

Stem-loop sequence mml-mir-323b	
Accession	MI0023688 (change log)
Description	Macaca mulatta miR-323b stem-loop
Gene family	MIPF0000018; mir-154
Stem-loop	<pre> -- u gugaguuc 5' agguug ccguq gcauuauu u 3' uccagc ggcac --auaacc uc u </pre>

Figure 6.7: The correction of the precursor *MI0007754* in the release 22, confirms with **MIRfix** result. The miRNA sequence is corrected to be at the 5' arm instead of the 3' arm. The precursor predicted by **MIRfix** is shown in figure 6.6 (NB: The record was renamed and the accession number was changed in the new release).

There are cases when the 2 overhanging nucleotides are increased, and this is because of comparison of the 5' ends of the miRNA and miRNA*, following the AGO process. Figure 6.8 shows the precursor *MI0020652* (ggo-mir-487b) from the Gorilla animal, as an example when the 2 overhanging nucleotides increased to 3. These nucleotides were 3 in this case (the last hairpin to the right) because the miRNA* star was trimmed from its 5' end to make more stable than the one of the miRNA, which make it biologically correct. The same sequence was tested again without considering the stability problem, and the miRNA* was correctly predicted typically with 2 overhanging nucleotides (the hairpin in the middle). The labeled nucleotide number 10, is the only difference between the two alternatives. This nucleotide was trimmed from the miRNA* prediction when the stability problem is considered, and it is the final **MIRfix** output.

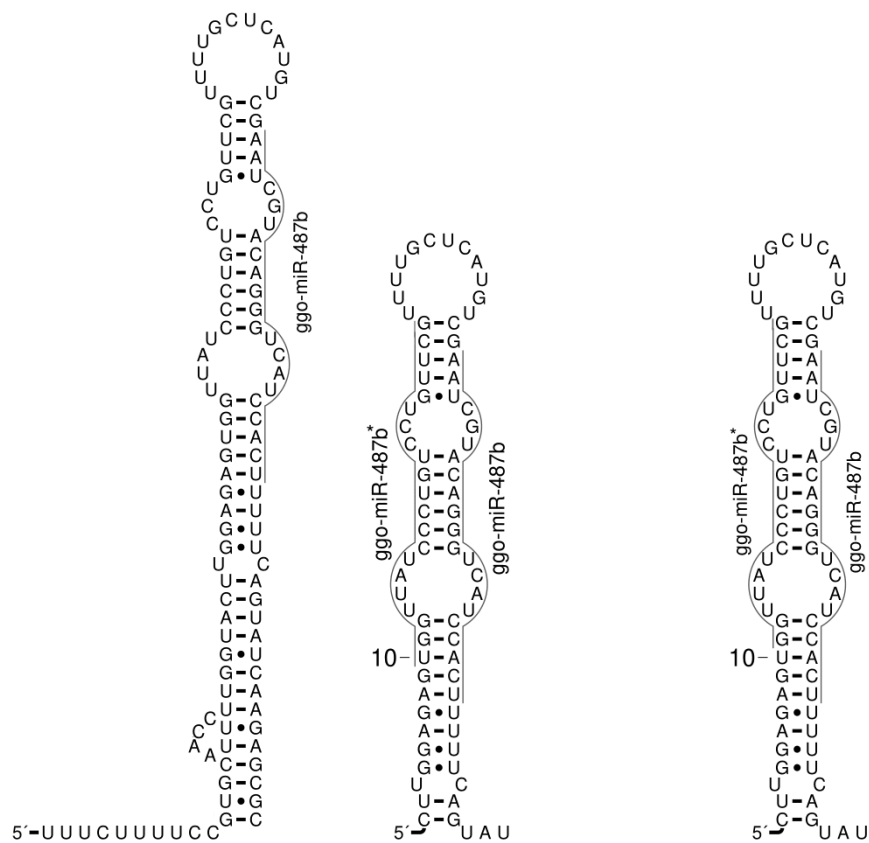


Figure 6.8: The precursor *ggo-mir-487b* was corrected by changing the orientation of the mature sequence from 3' to 5' end. The first hairpin from the left end shows how the precursor was annotated at miRBase. The hairpin in the middle is the alternative precursor without considering the stability of miRNA and miRNA*. The right-hand side precursor is the alternative with the consideration of miRNA 5' end stability.

Correction at the alignment level

The correction of the precursors was not limited only to the structural information of the precursor. So, the precursors were also checked in the final alignment of the sequence, as we expect to see a high quality of alignment for precursors belong to the same family according to the similarities between the mature sequences which should be also expressed in the containing precursors with a reasonable similarity. Figure 6.9 shows the precursor *MI0009839* (*bta-mir-453*) from the colloquially cows (scientific name: *Bos taurus*) that was corrected at the precursor level. The left hairpin is the original hairpin annotated in the miRBase. The general structure of the precursor and the position of the mature sequences together, form almost typical pre-miRNA structure. In comparison to the alternative (right hairpin), there is no reason to change or replace this precursor. Albeit, this precursor did not fit to the final alignment of the family and it was shifted to the left of the consensus

alignment figure 6.10-A. In the same time, the alternative was a potential candidate so it was at the end replaced by its alternative precursor.

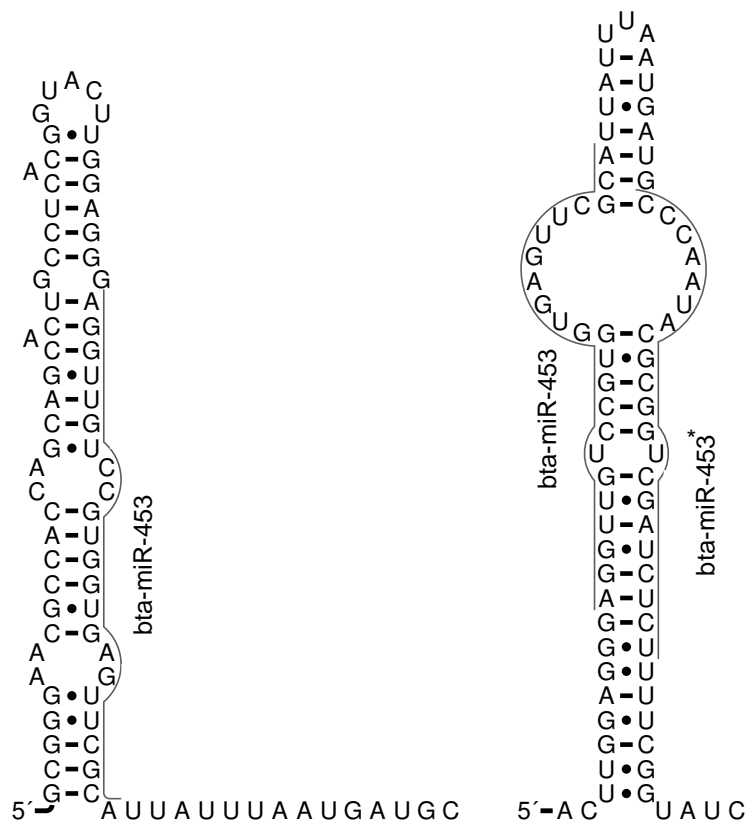


Figure 6.9: To the left, the original hairpin of the precursor bta-mir-453. To the right its alternative including a miRNA*.

Re-correcting the alternatives

In 96 cases the changed precursors at the alignment level were changed back to the original because the alternative did not fit to the final alignment. In this family (MIPF0000018), 3 precursors were changed to their alternatives according to the structural comparison, and at the end were changed back to the original precursors based on the alignment result. Figure 6.11-A is the alignment result after processing the family in the precursor's level only. The three precursors that are shifted to the right end of the alignment, are the alternatives of the original precursors. These precursors were detected again at the alignment level as misaligned sequences and thus were changed back to their original state, as the original sequences are potential pre-miRNA sequences based on their structural features. Figure 6.11-B demonstrates how these precursors fit to the final alignment.

In the end, all the precursors were aligned together without any misaligned sequence. With a 5 original precursors replaced with their alternatives, one of

them was changed at alignment level, and the rest based on their structural features. This final alignment is partially shown in figure [6.10-B](#).

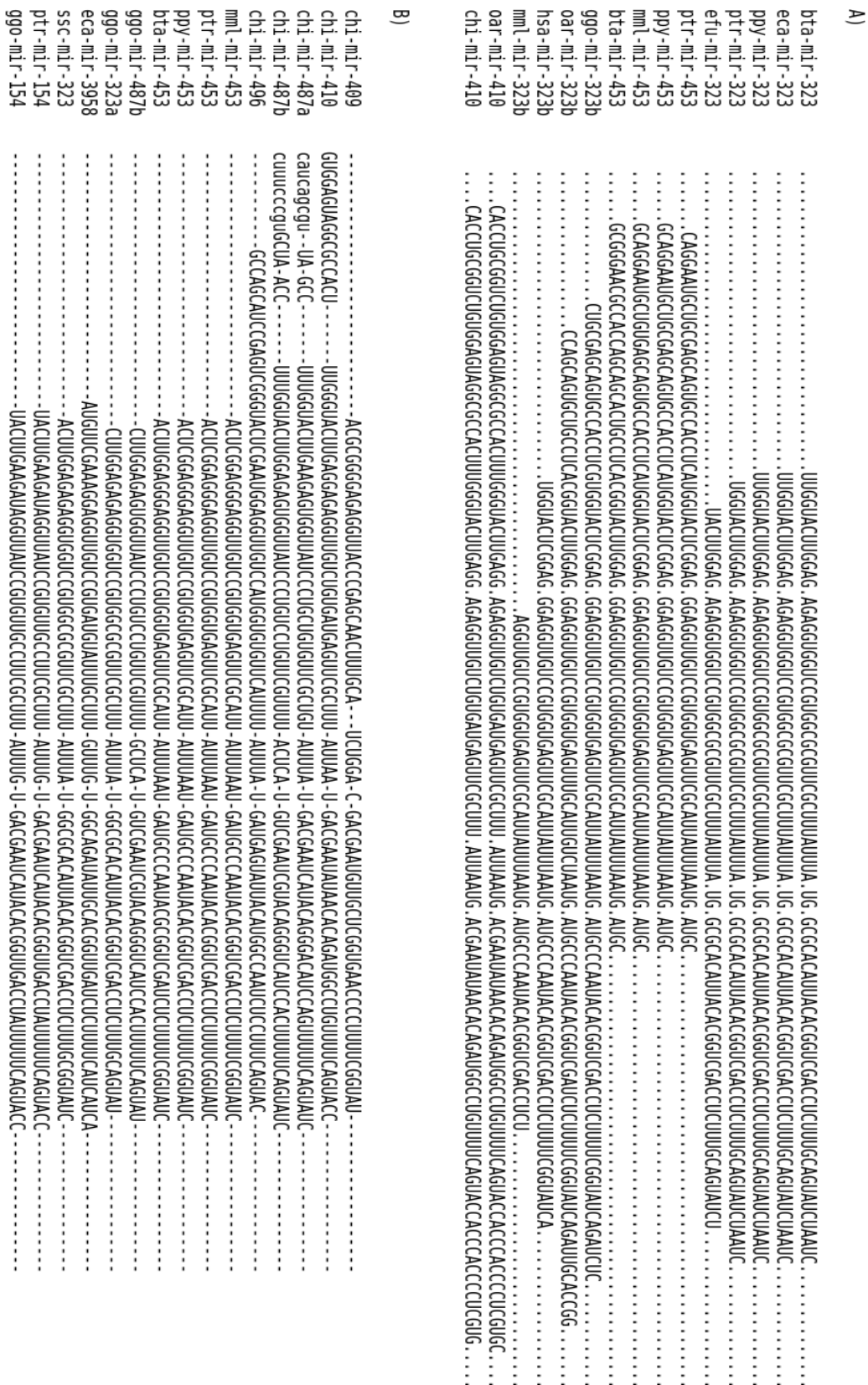


Figure 6.10: Part of the alignment of MIPF000018 family. A) is the original alignment. B) is the final alignment. The four sequences of IDs ptr-mir-453, ppy-mir-453, mml-mir-453 and bta-mir-453 were misaligned and corrected to fit the alignment (in B)).

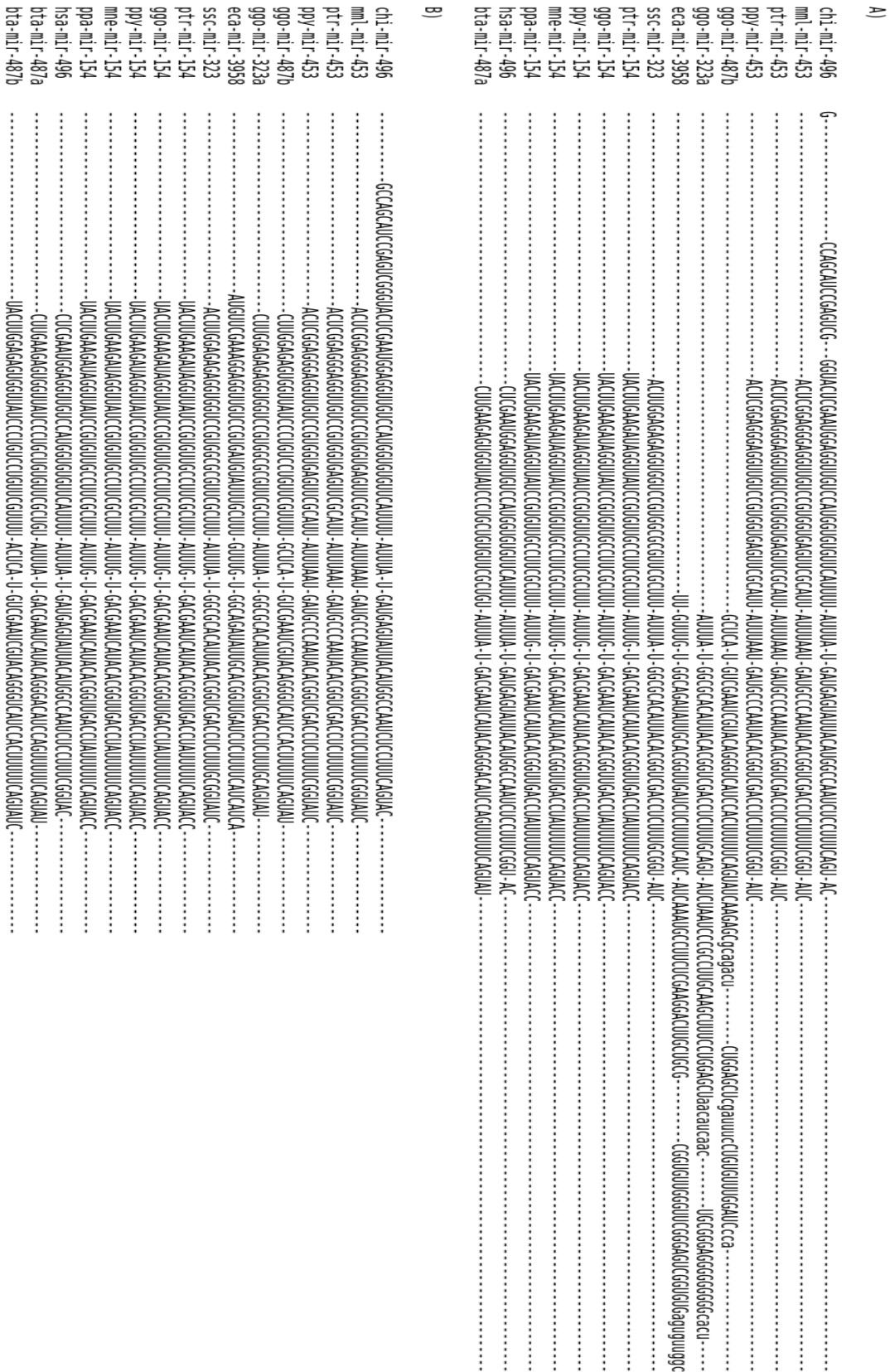


Figure 6.11: Part of the alignment of MIPF000018 family. A) Alignment after the precursor level. B) is the final alignment. The Three sequences of IDs ggo-mir-487b, ggo-mir-323a and eca-mir-3958 were changed at the precursor level but the alternatives did not fit the new alignment and were detected as misaligned in the alignment level. These precursors were changed back to the original sequences to fit the final alignment produced by the pipeline.

As further checking of the consistency between the two levels (precursor level and alignment level), the family *MIPF0000005* was processed under different conditions. Two precursors of this family were changed at the alignment level, *MI0019479* and *MI0019480*. The parameters of the *RNAfold* were changed to allow lonely base-pairing nucleotides. This was done to test if the pipeline can detect the wrong sequence independently at both levels. This time, the precursor *MI0019479* was detected then as a wrong precursor at precursors' level and it was replaced by its alternative. This test proved that those wrong precursors can be detected independently at both levels which then increases the probability of detecting such cases.

6.4.2 miRNA and miRNA* prediction

This is also one of the main goals of this pipeline, to predict the mature sequence(s) for a given precursor which might not have annotated mature sequence or only one. The first case does not exist in *miRBase* since all precursors are annotated with at least one mature sequence. So a modified data from *miRBase* was used as test data set for the pipeline.

miRNA prediction

This test was divided into two main tests: in the first the mature sequences and precursor were from the same family, of the precursors without annotated mature sequence. In the second, the sequences were from different families to predict the mature sequence for a given precursor.

In the first test, the *miRBase* family *MIPF0000033* was used. It is the biggest family in *miRBase* release 21 of 415 sequences. This family consists of 159 precursors of two annotated mature sequences and 256 precursors of one annotated mature sequence. This family was chosen because of the diversity in the mature and precursor sequences, but same time, high similarity still can be there between a group of sequences in the same family. Randomly, the annotated mature sequences were removed for 200 precursors of the 256 precursors which had only one annotated mature sequence. So the run was for 415 precursors of the same family, 215 with at least one annotated mature sequence and 200 without any annotated mature sequence. In the same way, 100 samples were created and tested. For the outputs of all 100 samples together, the distances between the predicted mature sequences and their correspondings annotated at *miRBase* were calculated using Levenshtein distance. The Levenshtein distance is the edit distance shown in the equation 2.2 in the introduction, but instead, adding score one for mismatches and indels, and zero for matches.

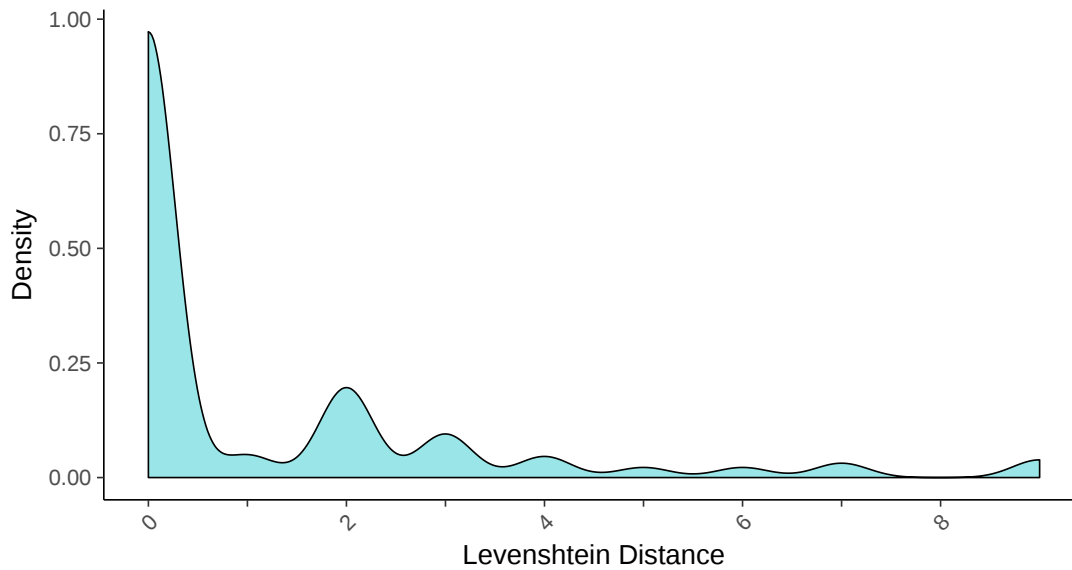


Figure 6.12: Density plot for 100 runs comparing Levenshtein distances between original and predicted mature sequences for 200 randomly selected precursors with one annotated mature sequence of MIPF0000033. The peak at distance 0 shows that for most of the cases the predicted sequence resembles the original.

In one of the samples in particular, out of the **200** precursors the pipeline predicted perfectly **111** mature sequences as they are annotated in *miRBase*. The table 6.2 shows how the number was produced in this specific sample.

Perfectly predicted	mispredicted		Mature of the opposite arm
	modified from one end	modified from both ends	
111	71	4	14

Table 6.2: This table shows the distribution of predicted mature sequence in a sample of 200 precursors. Perfectly predicted, mature sequences exactly as annotated at *miRBase*. Mispredicted had few nucleotides changes at one of the ends or both. last column to the right, are the mature sequences predicted at the opposite arm of the annotated mature sequence.

In the second test, the accuracy of the pipeline in predicting a mature sequence of a precursor was checked, in reference to precursors and mature sequences from different families, i.e., having less similarity. To this end, 5 precursors were used from 4 different families and with removing the annotated mature sequences for one of them. The 5 precursors were:

- precursor MI0019480 (ola-mir-30c *Oryzias latipes*), belongs to MIPF0000005 family (mir-30)
- precursors MI0013837 (tgu-mir-18b *Taeniopygia guttata*) and MI0004900(xtr-mir-93a *Xenopus tropicalis*), belong MIPF0000001 family (mir-17)
- precursor MI0014113 (oar-let-7b *Ovis aries*), belongs MIPF0000002 family (let-7)

and the tested precursor was *MI0014874* from *MIPF0000159* family (mir-296). Figure 6.13 shows the alignment of the predicted mature sequence to the annotated ones in miRBase. The first mature was predicted with 86% similarity and the second one was annotated with 80% similarity. These results show the ability of the pipeline to predict the mature sequences of a precursor with a reasonable similarity despite if the sequences belong to the same family or not.

```

Predicted mir : CAGAGGGUUGGGUGGAGGCUC
Annotated mir*: GAGGGUUGGGUGGAGGCUCUC
-----
Predicted mir*: CCCCCCUCAAUCCUGUUGUG
Annotated mir :AGGGCCCCCUCAAUCCUGU

```

Figure 6.13: Aligning the predicted mature sequences of the precursors *MI0014874* to its correspondings at miRBase. The first predicted miRNA sequence overlapped with the miRNA* annotated at miRBase, and the mature sequence predicted as miRNA*, overlapped with the annotated miRNA. The miRNA/miRNA* are only used to differentiate between the the two precursors and not referring to the functionality. The reference sequences were taken from different families which affect the accuracy of the prediction. See figure 6.12 for the cases when the sequences belong to the same family.

miRNA* prediction

The prediction of the second mature sequence was also tested using an artificial dataset from miRBase. In this test, only the precursors with two annotated mature sequences were used, since a real information about both annotated mature sequences for each precursor is needed. This test was done in two steps, in each step one of the mature sequences annotated in the sample precursors was removed. Hence, the mature sequences annotate at the 5' end was removed first from the precursors and then these sequences were processed in the pipeline. The same step was done with removing this time the sequences at the 3' end of the precursors. In this test 1402 precursors were used, by taking two precursors, of two annotated mature sequences from each family. For each set, the predicted mature sequences were compared to their corresponding annotated ones at miRBase. In figures 6.14 and 6.15 are the comparisons between the predicted position of the miRNA* sequences and the annotated ones. The margin of 2 to 3 nucleotides is always acceptable since the miRNA* in miRBase some times they lack the 2 overhanging nucleotides.

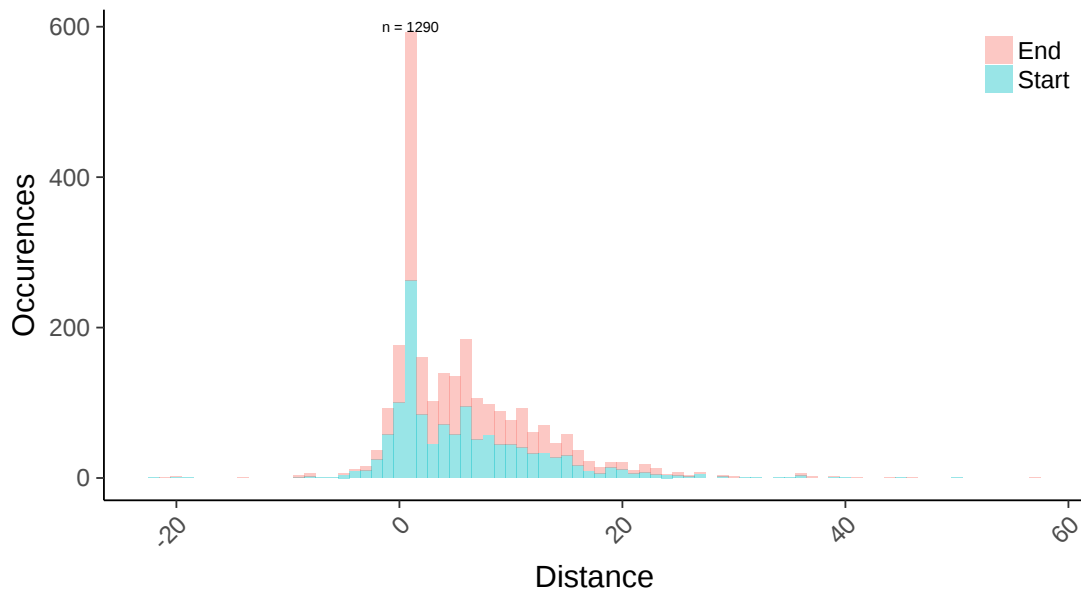


Figure 6.14: Distances of start to start and end to end of original to predicted miRNA* positions. The peak around 0 shows that most predictions overlap directly with the original annotation.

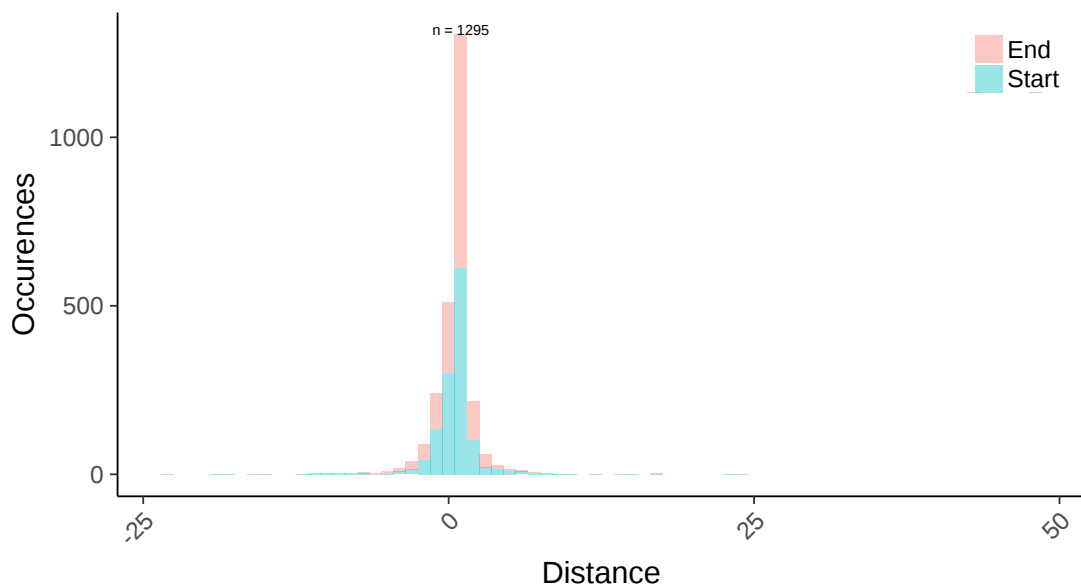


Figure 6.15: Distances of start to start and end to end of original to predicted miRNA* positions. The peak around 0 shows that most predictions overlap directly with the original annotation.

Moreover, one of the 2 annotated matures was removed randomly from all the precursors having two annotated mature sequences, which is in total 5329 in 1392 families. The predicted sequences were aligned using `clustalw`. According to the alignment scores, 3560 (66.8%) miRNA* sequences were recovered perfectly. In 1655 (31.0%) a slight variation is obtained in position or length of the predicted miRNA. Major deviations were observed in less

than 2% of the tests.

In addition, all precursors in which one mature sequence was annotated in miRBase v.21 while both a miRNA and a miRNA* are reported in miRBase v.22 were considered. There are 43 cases of this type, of which 29 are predicted exactly and another 9 to a good approximation. Only 5 miRNAs exhibit major discrepancies.

6.4.3 Covariance models

In the end, it was meaningful not only to check the sequences and alignments information of the families, but also the covariance models.

Covariance model measurements

Covariance models were built with `cmbuild` and with `cmstat` the related information was extracted. Among the information that `cmstat`, the main categories of these CMS which affect the homology search was compared, which are the relative CM entropy (the information content in bits) and effective total sequence number. Figures 6.16 and 6.17 shows the comparison of these two measures normalized by the number of sequences in the alignment.

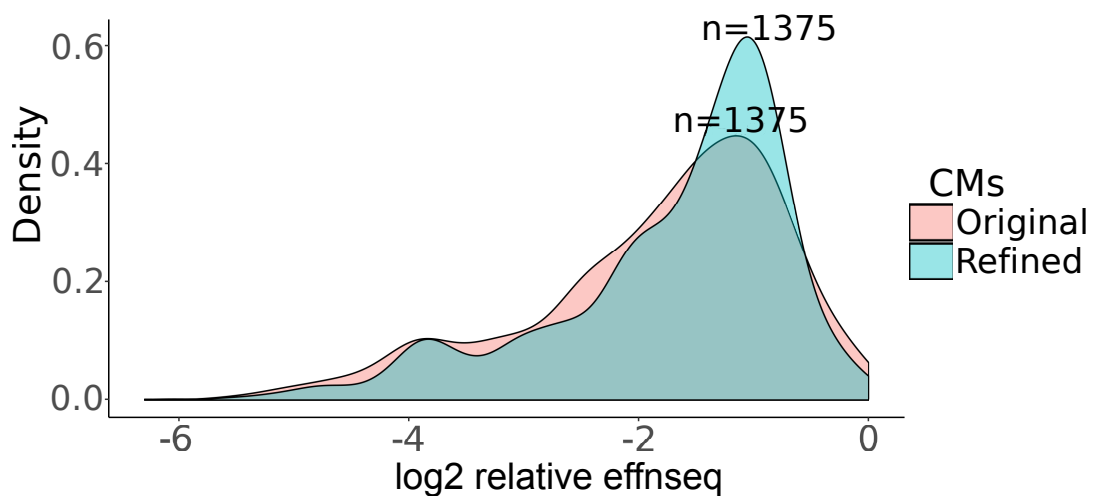


Figure 6.16: CMs built from **MIRfix** alignments have a higher relative effective sequence number than the original, thus more sequences per alignment that add to a CMs entropy.

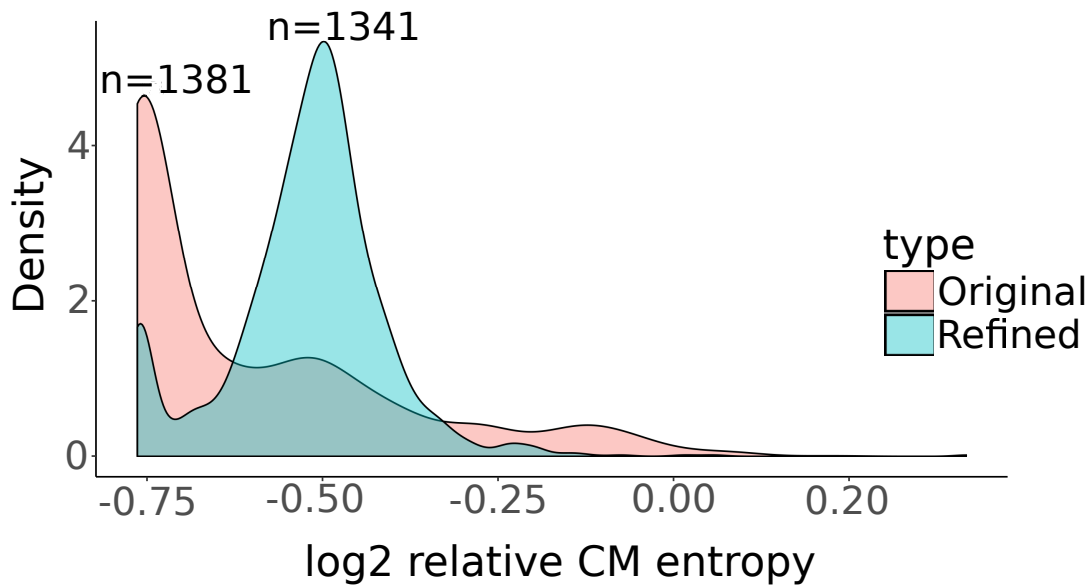


Figure 6.17: Relative entropy of CMs built from **MIRfix** refined alignments. Compared to CMs from original alignments, the refined CMs have higher entropy, thus contain more information.

Homology search improvement

In the previous section, both the information content (entropy) and the effective total sequence number were improved after processing the families in the pipeline. Theoretically, this improvement should be reflected in the homology searches using the new CMs. This was practically checked using the improved family *MIPF0000005* (mir-30), and mir-30 plays important role with other factors as a diagnostic biomarker and therapeutic target for cancer (Su et al., 2018). As a final result of processing this family with **MIRfix**, two precursors were corrected *MI0019480* and *MI0019479*. After running both, original CM and the new CM against the genome of the *MI0019479*, which is a precursor from the Japanese rice fish. Interestingly, both models (the original and the refined) detected a hit in the genome similar to the alternative precursor corrected by the pipeline. In figure 6.18 is the alignment of the four sequences:

- Sequence A), the alternative of the original sequence (B) corrected with the pipeline
- Sequence B), the original sequence as it is presented in miRBase release 21
- Sequence C), the sequence hit detected with the original CM
- Sequence D), the sequence hit detected with the refined CM

As this alignment shows, the pipeline's output sequence is very similar to the sequences detected with both CMs with just a few nucleotides added at ends,

which does not affect the main structure of the sequence and conserves the correct mature sequences. Moreover, the E-value of the hits was greater in the refined model **4.1e-14** than the original CM **1.3e-12**. This emphasizes that the correction of the sequences according to the method used in this pipeline, is definitely increasing the accuracy of the CMs in detecting homologs.

```

A)          GAUGCUGUAAACAUCUUGACUGGAAGCUGGGAUUUUGUCAGUGUGGGCUUUCAGUCGGAUGUUUGCAUCAUCUUAU
B)                                UGGAAGCUGGGAUUUUGUCAGUGUGGGCUUUCAGUCGGAUGUUUGCAUCAUCUUAUUGCCACAAACCACCAACAGAAACCCUUAUUGCCAACAUUGAUUAUUUG
C) GGCAAUGUCGAGAUUCUGUAAACAUCUUGACUGGAAGCUGGgauuuUGUCAGUGUGGGCUUUCAGUCGGAUGUUUGCAUCAUCUUAUUGCC
D)          GUCGAGAUUCUGUAAACAUCUUGACUGGAAGCUGGGAUUUUgUcAGUGUGGGCUUUCAGUCGGAUGUUUGCAUCAUCUUAUUGC

```

Figure 6.18: Alignment of the CMs hits. From A) to D), the pipeline’s output, the original sequence annotated at miRBase, the sequence detected with the original CM and the sequence detected with the refined CM. The results of both CMs confirms with the correction made by **MIRfix**.

6.5 Conclusion

In this chapter I described the pipeline that covers the main problems introduced previously. This work was an extension to the previous chapter, as an automated curation methods was implemented which lead to an improvement in the majority of the seed alignments of microRNA families found at the miRBase database. In these methods two levels were considered, sequence and alignment level. At the sequence level, the precursors were treated independently of each other and of the alignment. The precursors were checked based on their conservation of the typical generic features of a pre-miRNA, then their consistency with the consensus alignment. These methods included also predicting the miRNA* sequence and miRNA sequences for the processed precursors without any annotate mature sequence, by the mean of base-pairing properties between the miRNA and miRNA* sequences and the homology search, respectively. The prediction of the miRNA* considered carefully the biological features like the 2 overhanging nucleotides and the stability at the 5 prime ends of the mature sequences.

As an application, the pipeline was applied to the microRNA families of the miRBase release 21. Of the metazoan families, ~78% of the families could be improved in their seed alignments, in addition to correcting a number of precursors and showing theoretically that they are annotated wrong. The pipeline was also tested using artificial data sets which showed a good level of accuracy in predicting the mature sequences for the precursors. Because the main issue in all this work is improving the homology search at the end, the covariance models were also compared (CMs of the original and CMs of the refined alignments). The results showed that this careful automated curation implemented in **MIRfix** contributes positively to microRNA homology search the search for true homologs.

Comparing **MIRfix** to the other methods explained in section 4.3, e.g., miRseeker. For instance, the number of base-pairing nucleotides is only counted when

comparing two versions of a given precursor in the latter. Additionally, in *miRseeker*, the structures with bulges or internal loops are penalized in the score with increasing the score of the sequences of continuous stems. This was covered in **MIRfix** by separating the stem-loops and considering each independently when the sequences show multi-hairpin shapes. These differences make the method more flexible by focusing only on the main parts, and not penalizing the whole sequence, while instead, the parts which describe a putative precursor can be cut out of the whole sequence, i.e., cutting the sequence at the correct ends. In contrast to *RNAmicro*, which checks the precursors according to pre-miRNA characteristics and relies also on information extracted from the given alignment, **MIRfix** checks the precursors as an initial step independently from the alignment or other precursors. Instead, it uses the generic biological characteristics to evaluate the precursors at the first level, before considering the alignment. Nevertheless, *RNAmicro* showed an ability to distinguish microRNA from others in an alignment. In the method described in this chapter, the lengths of the precursors are not pre-defined when comparing two precursors as in *miRDeep* and *miRanalyzer*. Instead, the miRNA* is predicted first, then the ends of the precursors are defined by the positions of the miRNA and miRNA*. By this, the possibility of trimming totally or partially the miRNA* when pre-defining the length of the precursor is avoided. Moreover and explicitly, **MIRfix** considers AGO processing while predicting the miRNA*.

Chapter 7

Discussion

In this dissertation, different methods and workflows regarding homology search and data curation were introduced and discussed. The described workflows have highlighted the correlation between improved data curation and improving the quality and sensitivity of the homology search methods. It becomes evident, that curation is a substantial step during homology search for non-coding RNAs. Even considering the high quality and the sensitivity that current methods and tools provide, none of them so far found a complete solution for the homology search problem. This limitation is due to extensive variation between the ncRNA sequences, which is even found in the same classes of RNA, like the variation shown for 7SK RNAs of vertebrates and invertebrates. In this context, and to break those limitations, the main problem was divided into sub-problems. Accordingly, the previous knowledge about the invertebrates 7SK snRNA was confirmed. Furthermore, the new results provide a more detailed structure, extending the known stems and annotating a novel one, stem A and stem B, respectively. A remaining question about 7SK snRNA evolution is whether 7SK is a true innovation that appeared relatively late in animal evolution, or whether it has diverged far from its distant ancestors. With currently available data and homology search tools, none of these theories can be confirmed. However, this detailed analysis of the invertebrates structure can support future researches trying to explore invertebrates genes and detect more novel 7SK snRNA genes. It has been shown, that the proposed models could already help to detected new 7sk RNA genes in new groups of invertebrates species. These results conform to assumptions that building informative models with better structural information will significantly increase the quality of the homology search. The here presented workflow can be implemented in an automated pipeline applicable to other non-coding RNAs of homology search problems, similar to those of 7SK snRNA, for instance, U7 RNA and telomerase RNA.

Easy to find homologs are likely false positives, this is definitely true for microRNA homology search. Homologs of these 22 nucleotide small sequences are easy to be found in various genomes, which produces a massive amount of potential homologs that need to be filtered. Computationally, these miRNAs are filtered based on the pre-miRNA, as these mature sequences are located on both arms of the hairpin shape of the pre-miRNA. Many state-of-the-art microRNA investigation tools exist, either directly based on sequence and structural conservation, or implementing them as features for machine

learning approaches. All of them are based on knowledge about functional and structural features of these RNAs which are then implemented in the homology search. For instance, consideration of DICER cleavage sites. Or, using datasets to train SVM machines which will be based on the amount of information that the datasets include. In all cases, the quality and quantity of information provided from those data sets is a bottleneck. Another concern is the sensitivity and accuracy of the available tools and pipelines, especially those based on sequence and structure information. This highlights the importance of data curation as an essential part of the whole homology search process. The here presented work includes two related parts about data curation and homology search in microRNAs. The first presents an initial step towards a complete coverage of this problem. In the initial curation, and relying in big parts on existing work in regard to covariance models (*infernal*), the curation problems were partially addressed. As for example in *miRBase* data, some precursors are annotated with one mature sequence and others are annotated with both mature sequences, leading to inconsistency in the annotated data which subsequently affects the quality of the alignments. Families consisting of both types of precursors were initially processed. Each of these families was divided into two groups, one containing the precursors with the one mature sequences, and the other containing the precursors with both annotated mature sequences. Based on the consensus alignment of the latter, the former group was processed. This was done by building covariance models for the group of precursors with two annotated mature sequences and scanning the other sequences of the same family. This process defined the ends of the precursors with one mature sequence, making all the sequences in a given family consistent. Moreover, sequences that didn't show any consistency with other family members were automatically detected in the scan and removed. This leads to an improvement in the alignment of the majority of the processed metazoan families, with an average improvement in Shannon entropy equal to 33%.

In the second part, a more comprehensive workflow was developed to solve the aforementioned problems. A method which focuses on the details of the mature and the pre-miRNA characteristics, considering the consistency of the sequences and the alignment at the same time. This method was implemented in **MIRfix**, a pipeline which automatically curates precursor input sequences based on structural features and seed alignments, in reference to their genomes. In addition to techniques similar to ones in already established tools, new and improved techniques were developed. The new and improved methods used in this pipeline are more flexible than any of the other existing tools, without discarding the main features of the pre-miRNA structure and miRNA processing. For instance, long sequences, which fold into more than one hairpin were considered carefully, and the putative hairpins were extracted from those sequences, without any limitation to the length of the precursors. After applying this pipeline to *miRBase* data (v. 21), these detailed considerations made it possible to correct a considerable number of the annotated pre-miRNA. Furthermore, some of them were corrected in the

new version of the miRBase (v .22) in the same way than by the pipeline. This highlights the significance of **MIRfix**, which could also be used as an automatic data curator for the next versions of miRBase. Moreover, the biological features of miRNA processing are considered, e.g., Dicer processing and the stability differences at 5' ends of the mature sequences needed for the mature sequence cleavage process. Results of a real-life application of **MIRfix** (on miRBase) were also presented (see chapter 6), showing high potential for future applications by covering the major problems discussed throughout this thesis. For example, correcting the precursors and improving the consistency between the precursors and their mature sequences in the metazoan microRNA families. This led to an improvement in the alignment of 78% of the metazoan families available at miRBase. Furthermore, this pipeline showed notable accuracy and sensitivity in predicting the mature sequences when it was tested with an artificial data set. The mature sequences predicted by **MIRfix** perfectly overlapped with the experimentally annotated sequences. It also showed high accuracy in predicting the miRNA sequence of the precursors with one missing mature sequence, not only for miRNA, but also for miRNA* sequences. Moreover, the predicted miRNA* sequences fit exactly to DICER and AGO processing features.

Back to the main point, and how curation affects the sensitivity of the homology search, it could be shown that the curation method presented in this work has a direct positive effect on homology search results, as was proven by tests with newly build CMs. A key feature of **MIRfix** is that it fits to most of the existing microRNA homology search tools. It can be included at the pre-processing, intermediate or post-processing levels. This comes from the accurate assessment that this pipeline provides through a detailed and carefully automated annotation method of the pre-miRNA independently of the other precursors, considering consensus information at the final stage. For instance, it can be combined with RNAmicro and used to harness comparative information to distinguish *bona fide* miRNAs from other small RNAs. However, this pipeline covers only the metazoan microRNAs that are produced by Drosha and Dicer but not other types like mirtrons, which are not processed by Drosha (Berezikov et al., 2007). Even though a major step in the right direction, **MIRfix** is still not covering all known problems, and can still be enhanced in terms of sensitivity. In many cases, e.g., in miRBase, precursors and mature sequences could not be assigned to families. Including a sophisticated family classification method is one of the features that could be implemented into **MIRfix** in the future to provide an almost complete solution for such issues.

In the end, it is clear that sequence-based and structural-based homology search results started to converge, and the existing tools and methods have reached almost the peak in detecting novel classes of ncRNAs or homologs to the existing ones. Of the main problems that can be deduced from the work presented in this thesis, are the unavailability of the genomes or missing

structural and sequence information about the homologs in the newly available genomes e.g., phylogenetically unrelated species. The latter problem is deduced from the 7SK RNA presented above, as the available sequence and structural information were not enough to detect new 7SK genes in new species in farther groups, until a careful annotation has been applied. More complex homology search problems exist, like the one we have seen in microRNAs. As even with almost full knowledge about these small 22 nt sequences and their pre-miRNAs structures, still, a lot of false positives are produced. To this end, the future homology search approaches can make a big step by considering more the biological functions, as far as these features are applicable and can be implemented computationally. In addition, a new fashion homology search can go further towards including the targets of the ncRNA as part of the search as it could reduce the false outputs of the search. For instance, this can be applicable on the U7 RNA which binds to the histone pre-mRNA, by studying the interaction between the U7 RNA candidates and the existing data of histone pre-mRNAs, as a filtering step after finding homologs based on the sequence and structural information. Thus, the future homology search can be significantly pushed forward by applying the aforementioned strategies as an “out of box” approaches. However it looks promising, the main question can be asked how to find the balance between the structural and sequence information, and including the biological functions as described before, when using it as criteria in such methods. Further questions can be answered only by more sensitive homology search approaches, does the combination of the evolutionary information about two different classes of ncRNAs help in finding new classes of them?. In other words, if a specific classes of ncRNAs in specific species share the same pattern of the evolution, can their pattern be used as a query to search for sequences of the same pattern, to be detected as a novel class of ncRNAs?

Appendix A

7SK RNA project

In this Appendix you can find more invertebrates models represented by their consensus secondary structures, as well as other supplementary data. The caption of the detailed models follows the caption of 3.5.

Specific detailed models

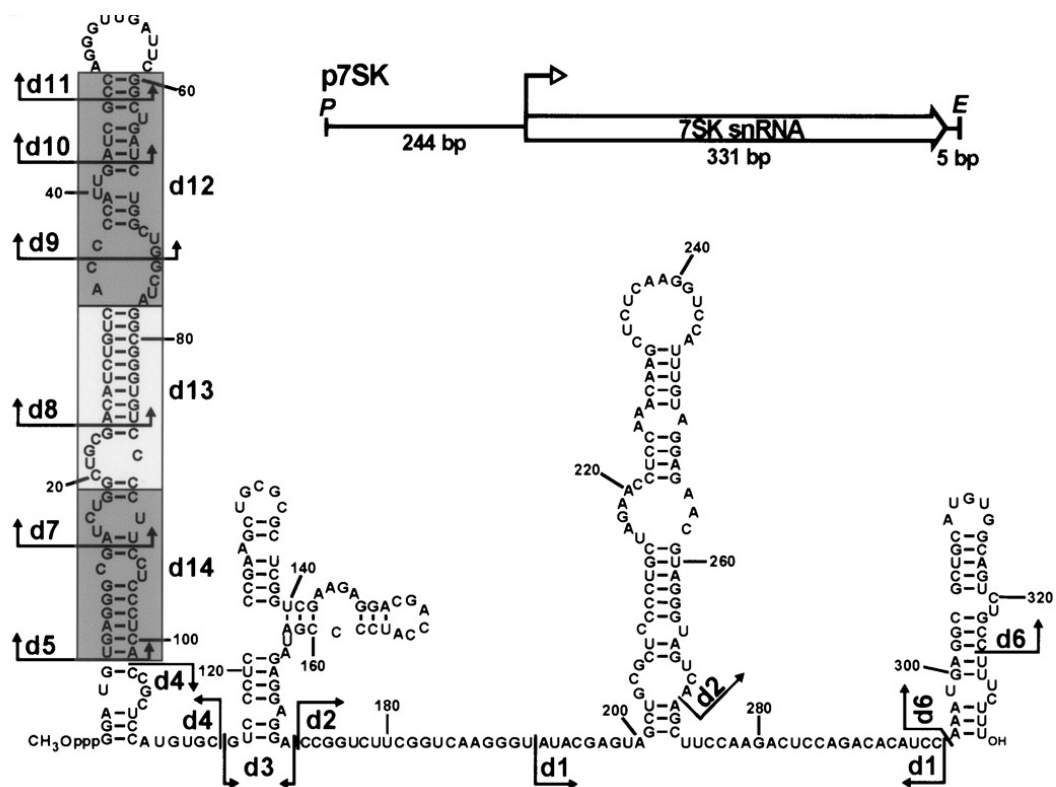


Figure A.1: Human 7SK RNA. The parts d7,d8,d9,d10,d12 and d13 are crucial parts for the P-TEFB binding. The parts d5, d12, and d13 are crucial HEXIM1 binding Egloff, Van Herreweghe, and Kiss, 2006.

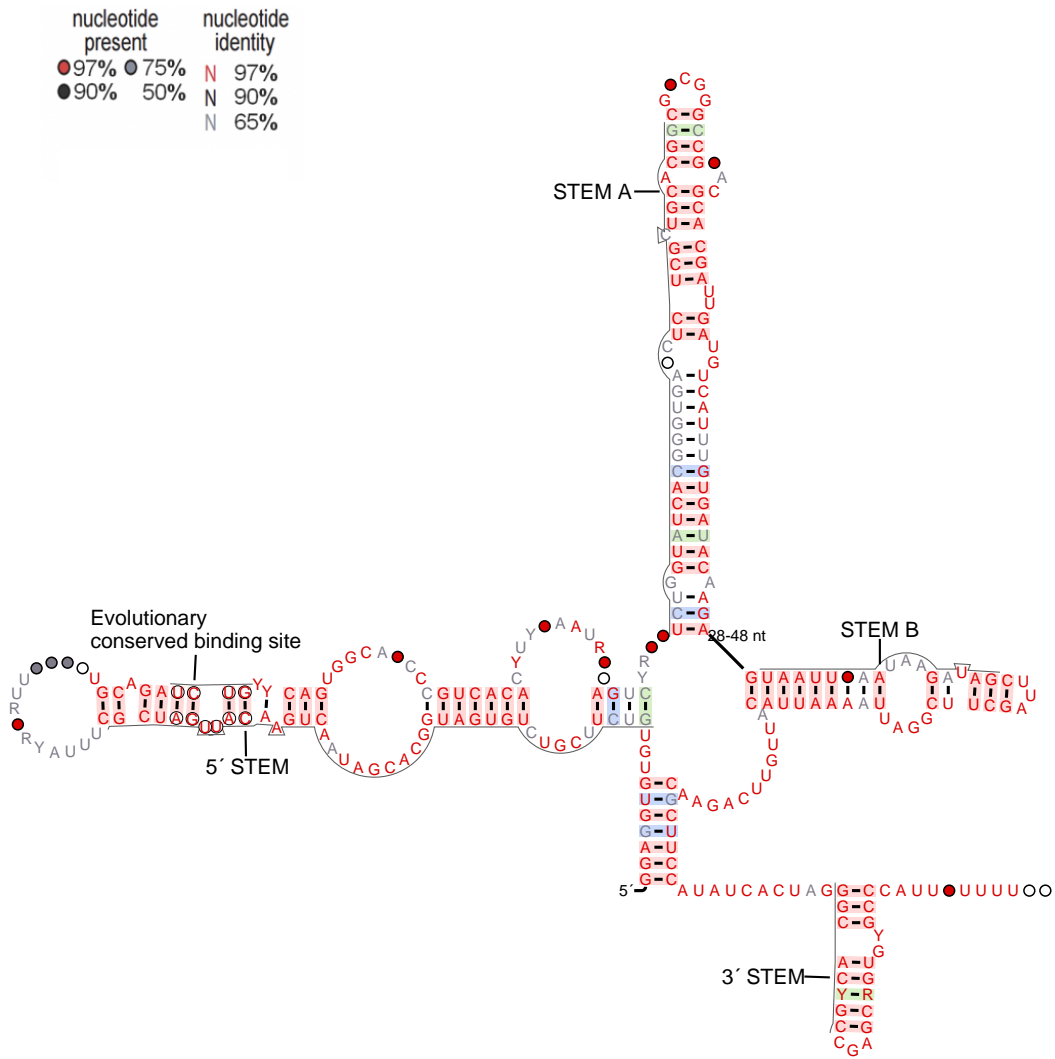


Figure A.2: Diptera model, without the drosophila species.

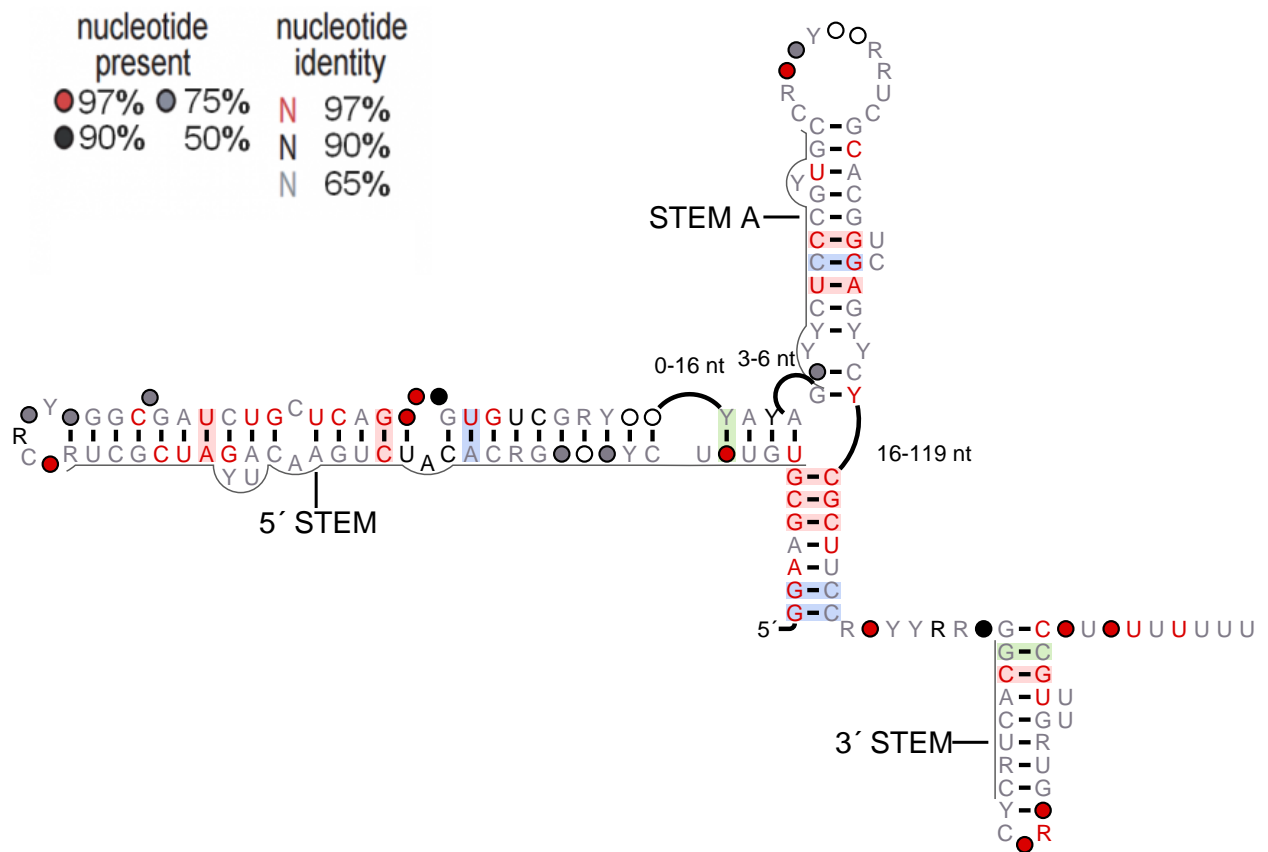


Figure A.3: Arachnida model. Unlike other invertebrates species, this group did not seem to have stem B, at least from the sequences available so far.

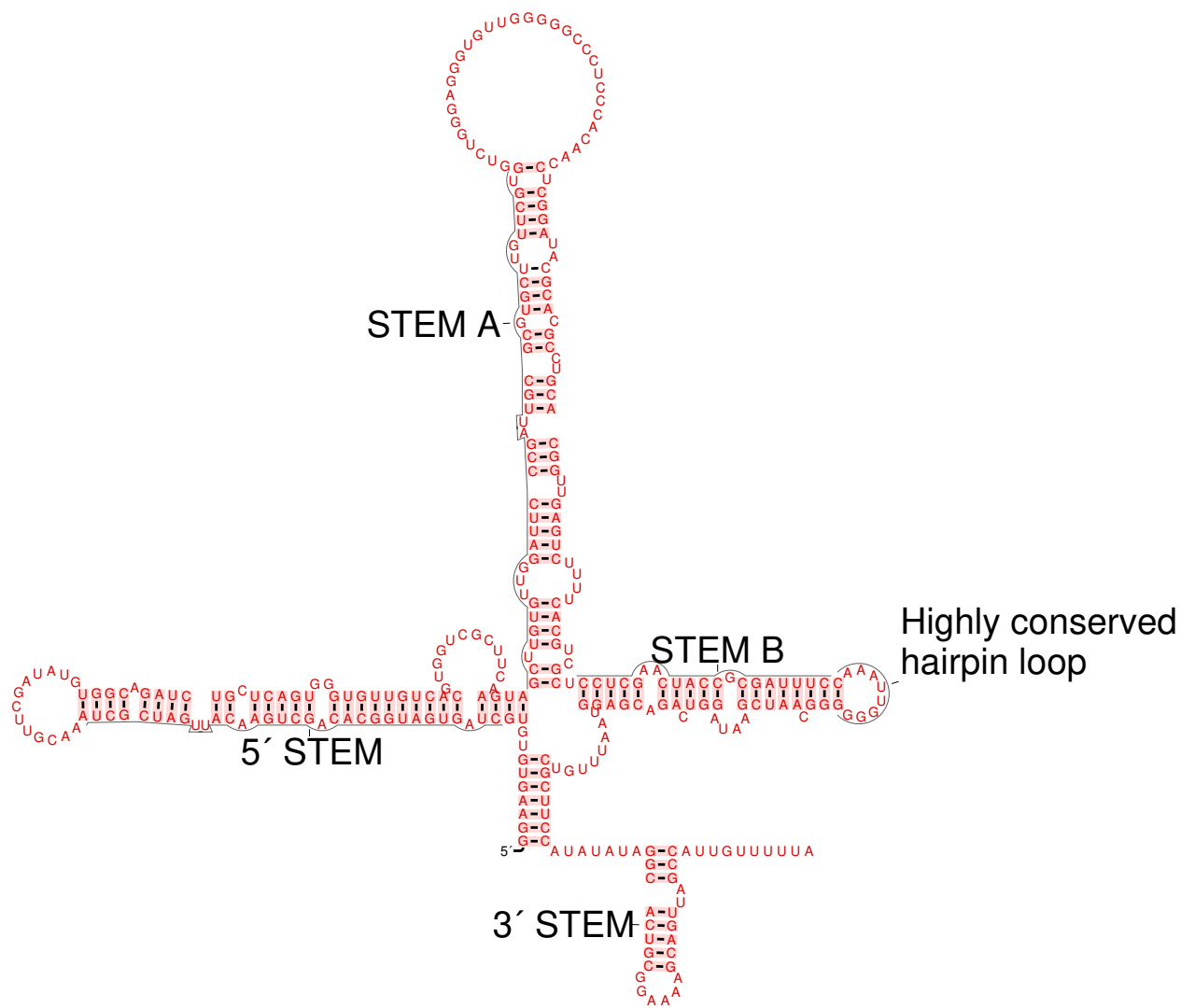


Figure A.4: *Metapolybia cingulata* structure contains all the conserved parts of a typical 7SK RNA. Additionally, it shows clearly the Hexapoda conserved novel stem B and its loop.

Detailed data related to the investigated species

SPECIES	RF00100		RF01052		this work
	Genomes	Candidates	Genomes	Candidates	
Acordulecera pellucida	0.000009	0.000092	8E-036	8.2E-035	1.30E-070
Acromy	7	0.00033	1.30E-028	6.3E-033	2.20E-061
Andrena asteris	5.60E-006	0.00011	3.40E-039	6.50E-038	2.20E-071
Anoplophora glabripennis	0.0045	2.10E-007	1.40E-030	3.80E-035	6.90E-043
Aporus niger	4.20E-006	5.40E-005	1.60E-039	2.00E-038	6.50E-074
Athalia rosae	2.60E-002	2.30E-006	1.90E-031	1.70E-035	1.00E-068
Bombus impatiens	0.066	7.70E-005	1.10E-033	6.00E-038	3.00E-067
Bombus pensylvanicus	2.70E-006	3.60E-005	4.60E-040	9.70E-039	5.70E-073
Cephus cinctus	0.38	3.30E-005	2.70E-029	2.40E-033	2.50E-067
Cerapachys biroi	2.4	0.00016	7.10E-034	4.70E-038	7.60E-065
Ceratosolen solmsi marchali	1.6	7.90E-005	9.40E-036	4.70E-040	3.60E-068
Chalybion californicum	1.80E-006	1.80E-005	2.60E-039	2.70E-038	1.30E-072
Cotesia vestalis	NO	0.00059	2.10E-031	1.50E-035	1.50E-058
Dendroctonus ponderosae	7.00E-005	4.80E-009	9.00E-020	6.20E-024	1.40E-035
Ephydra gracilis	NO	NO	8.30E-064	1.10E-067	8.70E-015
Fopius arisanus	0.61	6.00E-005	3.30E-034	3.30E-038	8.70E-069
Gerris buenoi	0.00011	4.10E-009	5.40E-020	2.00E-024	3.60E-037
Homalodisca vitripennis	NO	0.0026	5.80E-021	1.60E-025	1.20E-044
Linepithema humile	0.0033	2.10E-007	8.60E-034	5.60E-038	1.80E-063
Locusta migratoria	1.9	2.80E-010	1.90E-023	7.50E-028	2.50E-049
Megachile rotundata	0.18	9.10E-006	2.70E-034	1.40E-038	6.40E-069
Metapolybia cingulata	0.0014	0.012	1.40E-035	1.20E-034	4.90E-073
Microplitis demolitor	4.1	0.00028	2.80E-035	1.90E-039	6.50E-069
Mochlonyx cinctipes	NO	0.0026	9.70E-052	2.00E-055	3.10E-024
Monomorium pharaonis	8.3	0.0025	1.30E-031	6.80E-036	2.60E-057
Nematus tibialis	6.40E-007	5.50E-006	2.70E-036	4.60E-034	8.60E-074
Neodiprion lecontei	4.7	0.00027	7.00E-030	4.00E-034	5.30E-068
Nicrophorus vespilloides	0.0054	3.80E-007	3.50E-031	2.50E-035	8.00E-042
Orthogonalys pulchella	9.70E-006	7.20E-005	3.50E-037	6.40E-036	6.40E-073
Pogonomyrmex barbatus	9.7	0.0035	6.30E-030	4.00E-034	6.00E-061
Sapyga pumila	0.0014	0.012	1.80E-039	2.40E-038	9.70E-074
Scolia verticalis	2.40E-008	1.30E-007	3.90E-039	2.00E-038	8.40E-075
Sericomyrmex harekulli	2.80E-005	0.00033	1.20E-034	1.40E-033	8.40E-066
Solenopsis invicta	NO	0.0025	1.60E-027	1.20E-031	9.60E-060
Taxonus pallidicornis	1.20E-005	0.0001	1.40E-035	1.20E-034	1.40E-072
Trichomalopsis sarcophagae	18.4	1.50E-005	4.00E-037	4.50E-013	1.00E-066
Wasmannia auropunctata	4.4	0.017	1.90E-029	9.20E-034	3.60E-058
Zootermopsis nevadensis	0.0012	3.70E-008	2.60E-028	7.70E-033	1.70E-055
Lophotrochozoa					
Biomphalaria glabrata	9.10E-018	4.50E-021	0.0099	0.04	5.00E-022
Crassostrea gigas	2.70E-011	7.60E-016	7.20E-006	5.00E-009	3.90E-022
Lingula anatina	6.80E-015	2.20E-019	6.30E-011	1.00E-012	1.10E-026
ARACHNIDA					
Rhipicephalus microplus	NO	0.01	0.00096	0.0043	4.10E-008
Centruroides sculpturatus	0.0046	1.10E-007	0.00035	1.70E-015	5.80E-025
Mesobuthus martensii	0.11	3.70E-007	1.30E-013	4.40E-018	8.10E-026
Hypochthonius rufulus	NO	0.052	3.40E-014	1.10E-015	2.90E-023
Loxosceles reclusa	NO	NO	0.022	NO	1.5
Stegodyphus mimosarum	NO	NO	0.08	NO	3.10E-006
Steganacarus magnus	NO	0.0015	0.015	2.20E-007	1.60E-014

Figure A.5: CM hit scores, compared between RF00100, RF01052 and the general model we proposed. The green shaded number are perfect or almost perfect hits, the yellow shaded are acceptable hits and the non-shaded are bad hits.

Group	No significant hits	Significant hits
Apocrita	Camponuts floridanus : cflo_v3.3.fa antgenomes.org	Bombus pensylvanicusgi: gi 692921418 gb KM417872.1 NCBI
	Cardiocondyla: Cobs_1.4_OGS_CDS.fa.gz and Cobs1.4.scf.fa.gz antgenomes.org	Bombus impatiens: ENA AEQM02001933 AEQM02001933.1 NCBI
	Harpegnothos Saltator: hsal_v3.3.fa.zip antgenomes.org	Pogonomyrmex barbatus: ENA ADIH01016463 ADIH01016463.1 NCBI
	Atta cephalotes: acep_scaffolds.fasta.gz antgenomes.org	Acromyrmex echinaior: gi 328898774 gb AEVX01005901.1 NCBI
		Megachile rotundata: gi 340013494 gb AFJA01011723.1 NCBI
		Microplitis demolitor: gi 585252028 gb AZMT01026853.1 NCBI
		Linepithema humile: ENA ADOQ01006353 ADOQ01006353.1 NCBI
		Solenopsis invicta: ENA AEAQ01012710 AEAQ01012710.1 NCBI
		Fopius arisanus: gi 735031049 gb JRKH01004282.1 NCBI
		Cerapachys biroi: gi 602381356 gb JASI01008798.1 NCBI
		Wasmannia auropunctata: gi 753959420 dbj BBSV01055078.1 NCBI
		Ceratosolen solmsi: gi 563376351 gb ATAC01000351.1 NCBI
		Monomorium pharaonis: gi 812160015 dbj BBSX02021467.1 NCBI
		Cotesia vestalis: gi 770384568 gb JZSA01001595.1 NCBI
		Sapyga pumila: gi 692938209 gb KM426205.1 NCBI
		Scolia verticalis: gi 692938673 gb KM426669.1 NCBI
	Orthogonalys pulchella: gi 692935924 gb KM424435.1 NCBI	
	Sericomyrmex harekulli: gi 692939233 gb KM427229.1 NCBI	
	Aporus niger: gi 692919716 gb KM417002.1 NCBI	
	Metapolybia cingulata: gi 692928781 gb KM421505.1 NCBI	
	Andrena asteris: gi 692910530 gb KM412452.1 NCBI	
	Chalybion californicum: gi 692923094 gb KM418716.1 NCBI	
	Trichomalopsis sarcophagae: GenBank: NNAY01000270.1 NCBI	

Table A.1: The searched genomes are listed with their sources.

Group	No significant hits	Significant hits
Orussoidea		Orussus abietinus: gi 604465590 gb AZGP01002431.1 NCBI
Tenthredinoidea		Neodiprion lecontei: gi 914280525 gb LGIB01000732.1 NCBI
		Nematus tibialis: KM423963.1 NCBI
		Taxonus pallidicornisg: gi 692945457 gb KM433453.1 NCBI
		Athalia rosae: gi 459169241 gb AOFN01002401.1 NCBI
		Acordulecera pellucida: gi 692909736 gb KM412053.1
Coleoptera		Anoplophora glabripennis: gi 497030437 gb AQHT01007481.1 NCBI
		Nicrophorus vespilloides: gi 941342577 gb LJCH01003818.1 NCBI
		Dendroctonus ponderosae: gi 462276725 gb APGK01059300.1 NCBI
Orthoptera	Acheta domesticus: gi 550455633 ref NC_022564.1 gi 481852725 ref NC_021074.1 gi 23334625 ref NC_004290.1 All from NCBI	Locusta migratoria: gb AVCP011138713.1 NCBI
	Melanoplus saguinipes: gi 4049647 gb AF063866.1 NCBI	
	Gryllus bomaculatus: gi 134303398 ref NC_009240.1 NCBI	
Chelicerata/Arachnida		Rhipicephalus microplus: Sequence ID: LYUQ01249707.1 NCBI
		Centruroides sculpturatus: Sequence ID: AXZI01030636.1 NCBI
		Mesobuthus martensii: GenBank: AYEL01061488.1/ NCBI
		Acanthoscurria geniculata: Sequence ID: AZMS0100928822.1 NCBI
		Hypochthonius rufulus: GenBank: LBFL01042102.1 NCBI
		Loxosceles reclusa: Sequence ID: JJRW010450415.1 NCBI
		Stegodyphus mimosarum: GenBank: AZAQ01076786.1 NCBI
		Steganacarus magnus: GenBank: LBFN01095989.1 NCBIAADK01.1.fsa_nt.gz BAAB01.1.fsa_nt.gz BABH01.1.fsa_nt.gz and BABU01.1.fsa_nt.gz all from NCBI

Table A.2: The searched genomes are listed with their sources.

Group	No significant hits	Significant hits
Lepidoptera	Bombyx mori AADK01.1.fsa_nt.gz BAAB01.1.fsa_nt.gz BABH01.1.fsa_nt.gz and BABU01.1.fsa_nt.gz all from NCBI	
	Chilo suppressalis ANCD01.1.fsa_nt.gz NCBI	
	Melitea cinixia APLT01.1.fsa_nt.gz NCBI	
	Papilio glaucus JWHW01.1.fsa_nt.gz NCBI	
	Heliconius melpomene CAEZ01.1.fsa_nt.gz CAFA01.1.fsa_nt.gz NCBI	
Odonata		Nocardiopsis halotolerans: ANAX01.1.fa NCBI
Chelicerata/Merostomata		Limulus polyphemus: GenBank: AZTN01102591.1 NCBI
Hemiptera		Rhodnius prolixus: RhodniusprolixusCDC_SCAFFOLDS_RproC1.fa Vectorbase.org
Dictyoptera/Isoptera		Zootermopsis nevadensis: gi 639534756 gb AUST01000745.1 NCBI
Cephoidea		Cephus: gi 452905040 gb AMWH01004575.1 NCBI
Mollusca	Mytilus galloprovincialis: APJB01.1.fsa_nt APJB01.2.fsa_nt APJB01.3.fsa_nt APJB01.4.fsa_nt APJB01.5.fsa_nt NCBI	Biomphalaria glabrata: gi 530206112 gb APKA01045576.1 NCBI
		Crassostrea gigas: gi 404587963 gb AFTI01030118.1 NCBI
		Lingula anatina: gi 848959377 gb LFEI01000133.1 NCBI
Diptera	Glossina morsitans: JXPS01.1.fa NCBI	Ephydra gracilis: gi 824207163 gb JXPQ01011706.1 NCBI
	Glossina austeni: GlossinaausteniTTRI_SCAFFOLDS_GausT1.fa Vectorbase.org and JMRR01.1.fsa_nt NCBI	Mochlonyx cinctipes: gi 824463314 gb JXPH01043687.1 NCBI
	Phortica variegata: JXPM01.1.fsa_nt NCBI	

Table A.3: The searched genomes are listed with their sources.

Appendix B

MicroRNA project

Availability of MIRfix

MIRfix is available at: <https://github.com/Bierinformatik/MIRfix.git>.

Also there you can find the documentation needed.

Entropies of families of the initial curation

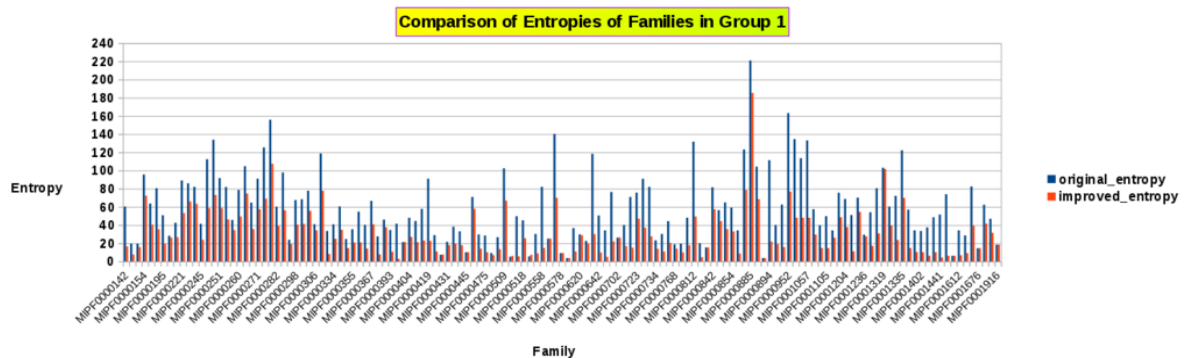


Figure B.1: Comparison between old and new entropy in the initial curation (Group1). The entropy decreases as the alignment quality improves.

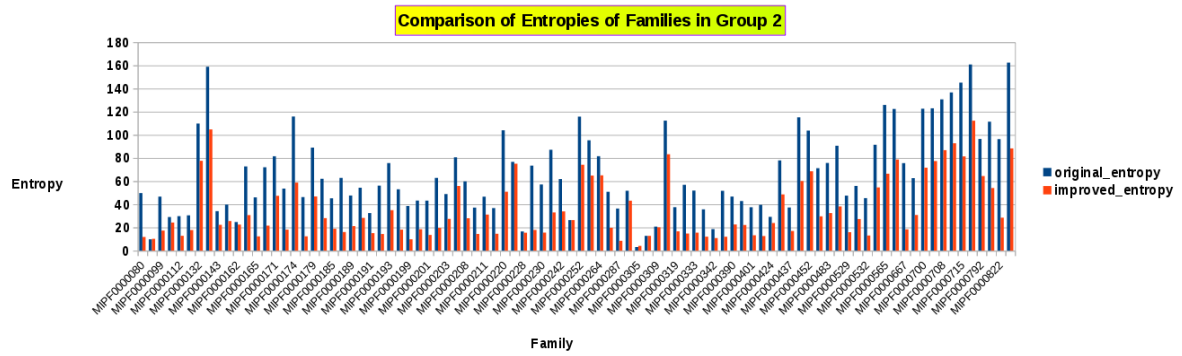


Figure B.2: Comparison between old and new entropy in the initial curation (Group2). The entropy decreases as the alignment quality improves.

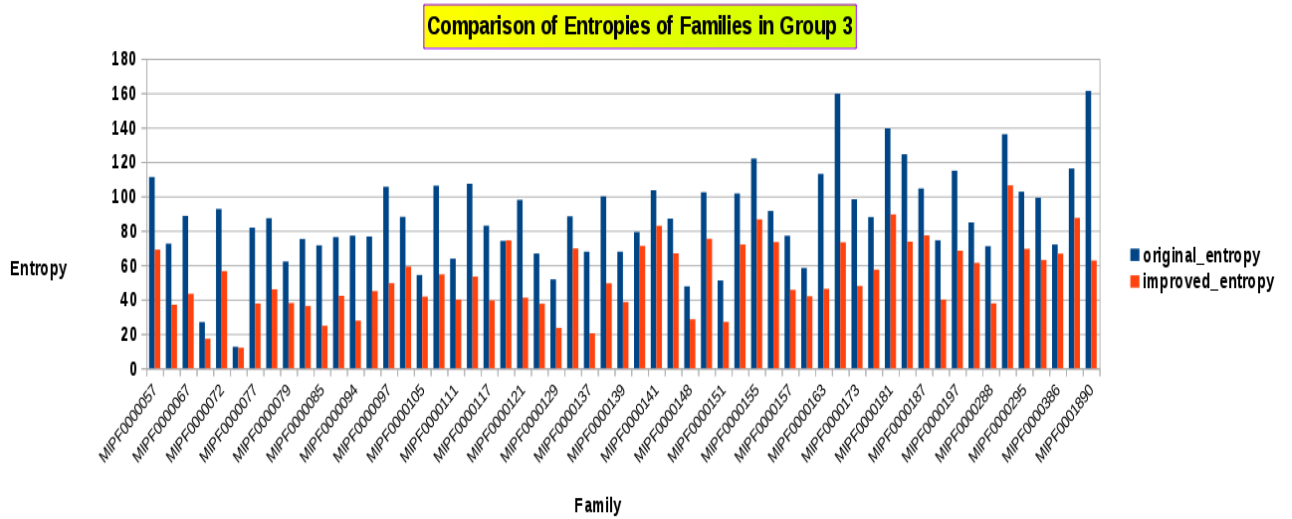


Figure B.3: Comparison between old and new entropy in the initial curation (Group3). The entropy decreases as the alignment quality improves.

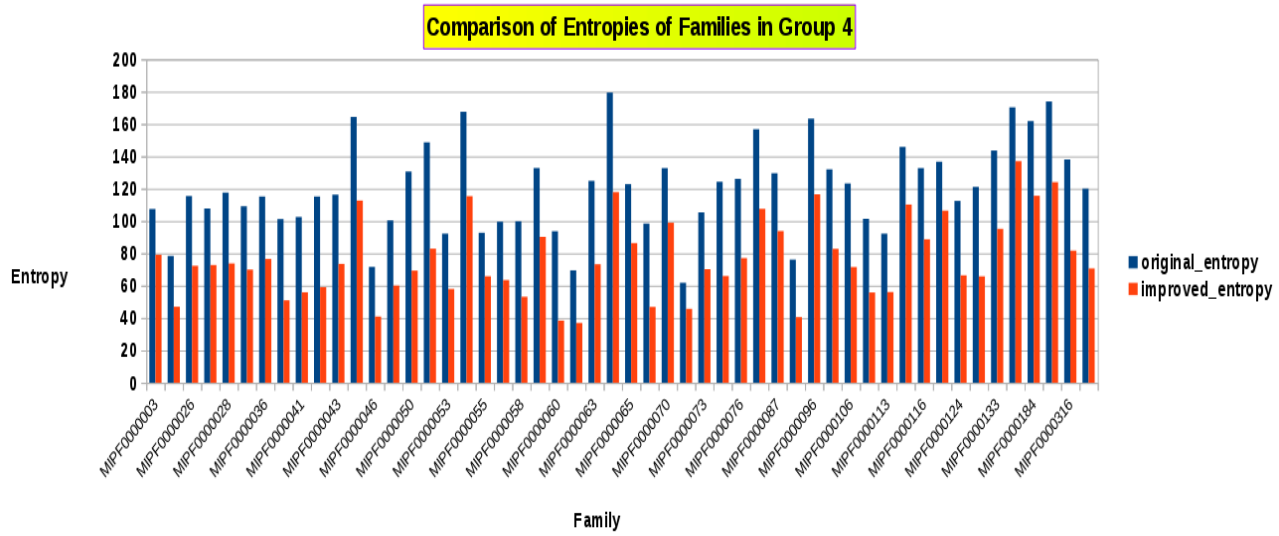


Figure B.4: Comparison between old and new entropy in the initial curation (Group4). The entropy decreases as the alignment quality improves.

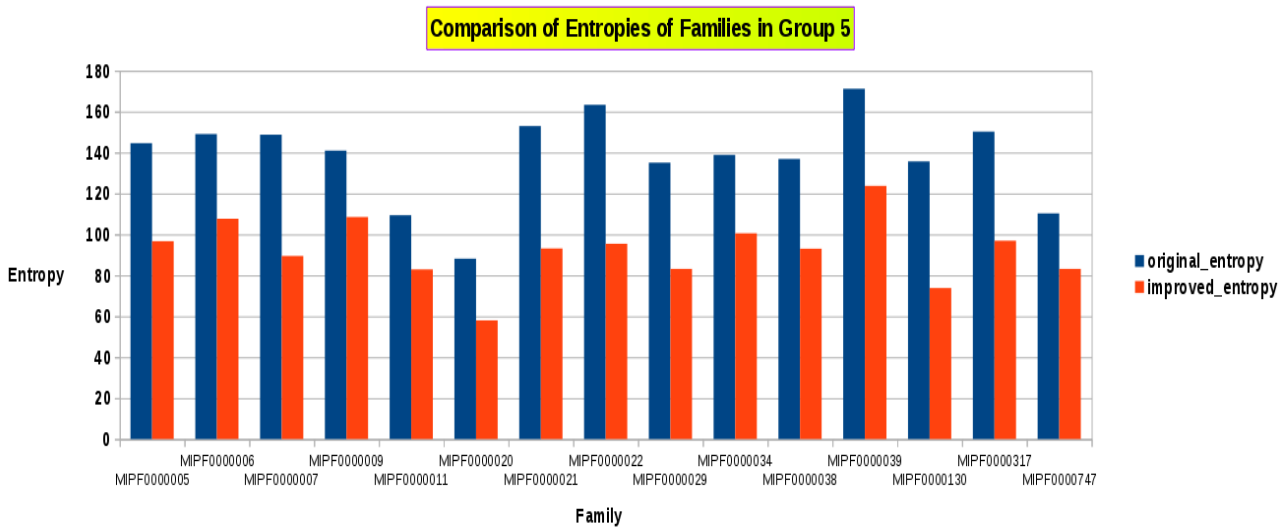


Figure B.5: Comparison between old and new entropy in the initial curation (Group5). The entropy decreases as the alignment quality improves.

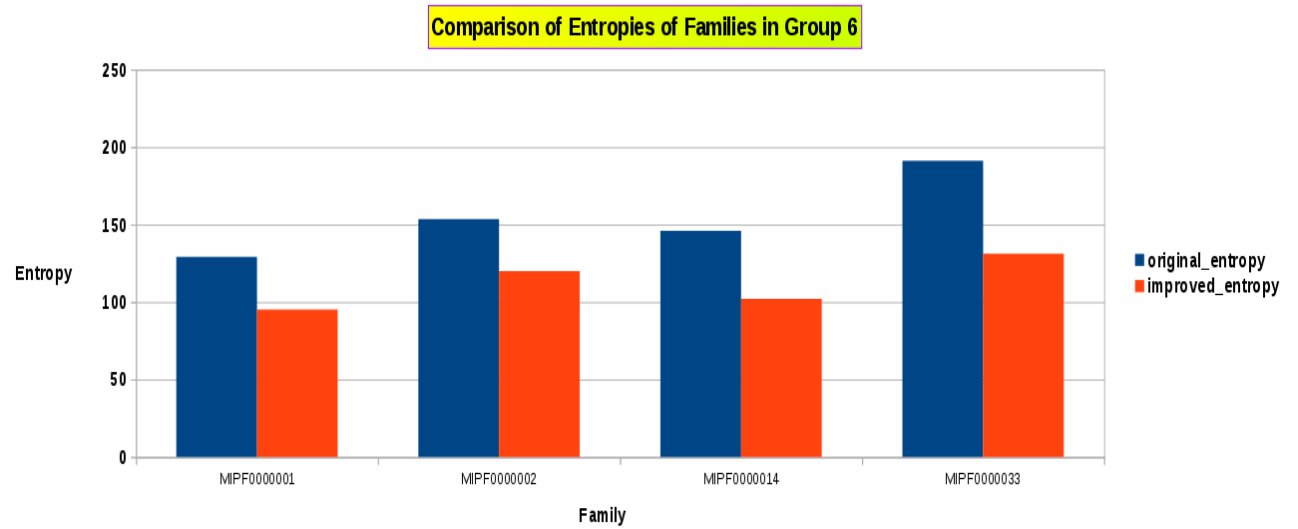


Figure B.6: Comparison between old and new entropy in the initial curation (Group6). The entropy decreases as the alignment quality improves.

Bibliography

- Alberti, C. and L. Cochella (2017). "A framework for understanding the roles of miRNAs in animal development". In: *Development* 144.14, pp. 2548–2559.
- Altschul, S. F. et al. (1990). "Basic local alignment search tool". In: *J Mol Biol* 215, pp. 403–410.
- B Malas, T. and T. Ravasi (2012). "Computational tools for genome-wide miRNA prediction and study". In: *The Open Biology Journal* 5.1.
- Benetatos, L., G. Vartholomatos, and E. Hatzimichael (2014). "DLK1-DIO3 imprinted cluster in induced pluripotency: landscape in the mist". In: *Cellular and molecular life sciences* 71.22, pp. 4421–4430.
- Berezikov, E et al. (2007). "Mammalian mirtron genes". In: *Mol Cell* 28, pp. 328–336.
- Bertoli, G., C. Cava, and I. Castiglioni (2015). "MicroRNAs: new biomarkers for diagnosis, prognosis, therapy prediction and therapeutic tools for breast cancer". In: *Theranostics* 5.10, p. 1122.
- Blazek, D et al. (2005). "Oligomerization of HEXIM1 via 7SK snRNA and coiled-coil region directs the inhibition of P-TEFb". In: *Nucleic Acids Res.* 33, pp. 7000–7010.
- Bologna, N. G., A. L. Schapire, and J. F. Palatnik (2013). "Processing of plant microRNA precursors". In: *Briefings in Functional Genomics* 12.1, pp. 37–45. DOI: [10.1093/bfgp/els050](https://doi.org/10.1093/bfgp/els050). eprint: [/oup/backfile/content_public/journal/bfg/12/1/10.1093/bfgp/els050/2/els050.pdf](http://oup/backfile/content_public/journal/bfg/12/1/10.1093/bfgp/els050/2/els050.pdf). URL: <http://dx.doi.org/10.1093/bfgp/els050>.
- Bourbigot, S et al. (2016). "Solution structure of the 5'-terminal hairpin of the 7SK small nuclear RNA". In: *RNA* 22, pp. 1844–1858.
- Brogie, J. E. and D. H. Price (2017). "Reconstitution of a functional 7SK snRNP". In: *Nucleic Acids Res.* 45, pp. 6864–6880. DOI: [10.1093/nar/gkx262](https://doi.org/10.1093/nar/gkx262).
- Cai, X., C. H. Hagedorn, and B. R. Cullen (2004). "Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs". In: *Rna* 10.12, pp. 1957–1966.
- Calin, G. A. et al. (2002). "Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia". In: *Proceedings of the National Academy of Sciences* 99.24, pp. 15524–15529.
- Camacho, C. et al. (2009). "BLAST+: architecture and applications". In: *BMC bioinformatics* 10.1, p. 421.
- Chen, R. et al. (2008). "PP2B and PP1 α cooperatively disrupt 7SK snRNP to release P-TEFb for transcription in response to Ca²⁺ signaling". In: *Genes & Development* 22.10, pp. 1356–1368.

- Cheng, J. et al. (2005). "Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution". In: *Science* 308.5725, pp. 1149–1154.
- Chenna, R. et al. (2003). "Multiple sequence alignment with the Clustal series of programs". In: *Nucleic acids research* 31.13, pp. 3497–3500.
- Consortium, E. P. et al. (2004). "The ENCODE (ENCyclopedia of DNA elements) project". In: *Science* 306.5696, pp. 636–640.
- Corbett, A. H. (2018). "Post-transcriptional regulation of gene expression and human disease". In: *Current opinion in cell biology* 52, pp. 96–104.
- Cosgrove, M. S. et al. (2012). "The Bin3 RNA methyltransferase targets 7SK RNA to control transcription and translation". In: *Wiley Interdiscip Rev RNA* 3, pp. 633–647. DOI: [10.1002/wrna.1123](https://doi.org/10.1002/wrna.1123).
- Crooks, G. E. et al. (2004). "WebLogo: a sequence logo generator". In: *Genome Res.* 14.6, pp. 1188–1190.
- Cursons, J. et al. (2018). "Combinatorial targeting by microRNAs co-ordinates post-transcriptional control of EMT". In: *Cell systems* 7.1, pp. 77–91.
- Denli, A. M. et al. (2004). "Processing of primary microRNAs by the Microprocessor complex". In: *Nature* 432.7014, p. 231.
- Dergai, O. et al. (2018). "Mechanism of selective recruitment of RNA polymerases II and III to snRNA gene promoters". In: *Genes & development*.
- Dill, T. and F. Naya (2018). "A hearty dose of noncoding RNAs: the imprinted DLK1-DIO3 Locus in cardiac development and disease". In: *Journal of cardiovascular development and disease* 5.3, p. 37.
- Eddy, S. R. (2002). "A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure". In: *BMC bioinformatics* 3.1, p. 18.
- Eddy, S. (1996). "RNABOB: a program to search for RNA secondary structure motifs in sequence databases". In: *Manual Google Scholar*.
- Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput". In: *Nucleic acids research* 32.5, pp. 1792–1797.
- Egloff, S., E. Van Herreweghe, and T. Kiss (2006). "Regulation of polymerase II transcription by 7SK snRNA: two distinct RNA elements direct P-TEFb and HEXIM1 binding". In: *Mol. Cell. Biol.* 26, pp. 630–642.
- Eichhorn, C. D., R Chug, and J Feigon (2016). "hLARP7 C-terminal domain contains an xRRM that binds the 3' hairpin of 7SK RNA". In: *Nucleic Acids Res.* 44, pp. 9977–9989.
- Eilebrecht, S, B. J. Benecke, and A Benecke (2011). "7SK snRNA-mediated, gene-specific cooperativity of HMGA1 and P-TEFb". In: *RNA Biol* 8, pp. 1084–1093. DOI: [10.4161/rna.8.6.17015](https://doi.org/10.4161/rna.8.6.17015).
- Eilebrecht, S et al. (2011). "HMGA1-dependent and independent 7SK RNA gene regulatory activity". In: *RNA Biol.* 8, pp. 143–157.
- Eilebrecht, S. et al. (2010). "7SK small nuclear RNA directly affects HMGA1 function in transcription regulation". In: *Nucleic acids Res* 39.6, pp. 2057–2072.

- Femminella, G. D., N. Ferrara, and G. Rengo (2015). "The emerging role of microRNAs in Alzheimer's disease". In: *Frontiers in physiology* 6, p. 40.
- Friedländer, M. R. et al. (2008). "Discovering microRNAs from deep sequencing data using miRDeep". In: *Nature biotechnology* 26.4, p. 407.
- Fusco, A. and M. Fedele (2007). "Roles of HMGA proteins in cancer". In: *Nature reviews. Cancer* 7.12, p. 899.
- Gautheret, D. and A. Lambert (2001). "Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles¹". In: *Journal of molecular biology* 313.5, pp. 1003–1011.
- Gebert, L. F. R. and I. J. MacRae (2018). "Regulation of microRNA function in animals". In: *Nat. Rev. Mol. Cell Biol.*
- Giraldo-Calderón, G. I. et al. (2014). "VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases". In: *Nucleic Acids Res.* 43.D1, pp. D707–D713.
- Godfrey, A. C. et al. (2006). "U7 snRNA mutations in *Drosophila* block histone pre-mRNA processing and disrupt oogenesis". In: *Rna* 12.3, pp. 396–409.
- Goodall, E. F. et al. (2013). "Neuronal dark matter: the emerging role of microRNAs in neurodegeneration". In: *Frontiers in cellular neuroscience* 7, p. 178.
- Griffiths-Jones, S (2004). "The microRNA Registry". In: *Nucleic Acids Res.* 32, pp. D109–D111.
- Griffiths-Jones, S. (2005). "RALEE—RNA ALignment editor in Emacs". In: *Bioinformatics* 21, pp. 257–259.
- Griffiths-Jones, S. et al. (2003). "Rfam: an RNA family database". In: *Nucleic Acids Res.* 31, pp. 439–441.
- Gruber, A. et al. (2008a). "Arthropod 7SK RNA". In: *Mol. Biol. Evol.* 1923-1930, p. 25.
- Gruber, A. R. et al. (2008b). "Invertebrate 7SK snRNAs". In: *J. Mol. Evol.* 107-115, p. 66.
- Grujil, F. R. de, H. J. van Kranen, and L. H. Mullenders (2001). "UV-induced DNA damage, repair, mutations and oncogenic pathways in skin cancer". In: *Journal of Photochemistry and Photobiology B: Biology* 63.1-3, pp. 19–27.
- Gudipaty, S. A. et al. (2015). "PPM1G Binds 7SK RNA and Hexim1 To Block P-TEFb Assembly into the 7SK snRNP and Sustain Transcription Elongation". In: *Mol Cell Biol* 35, pp. 3810–3828. DOI: [10.1128/MCB.00226-15](https://doi.org/10.1128/MCB.00226-15).
- Gürsoy, H.-C., D. Koper, and B.-J. Benecke (2000). "The vertebrate 7S K RNA separates hagfish (*Myxine glutinosa*) and lamprey (*Lampetra fluviatilis*)". In: *J. Mol. Evol.* 50, pp. 456–464.
- Hackenbarg, M. et al. (2009). "miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments". In: *Nucleic acids research* 37.suppl_2, W68–W76.
- Han, J. et al. (2004). "The Drosha-DGCR8 complex in primary microRNA processing". In: *Genes & development* 18.24, pp. 3016–3027.

- He, L. and G. J. Hannon (2004). "MicroRNAs: small RNAs with a big role in gene regulation". In: *Nature Reviews Genetics* 5.7, p. 522.
- He, N et al. (2008). "A La-related protein modulates 7SK snRNP integrity to suppress P-TEFb-dependent transcriptional elongation and tumorigenesis". In: *Mol Cell* 29.5, pp. 588–599. DOI: [10.1016/j.molcel.2008.01.003](https://doi.org/10.1016/j.molcel.2008.01.003).
- He, W. J. et al. (2006). "Regulation of two key nuclear enzymatic activities by the 7SK small nuclear RNA". In: *Cold Spring Harb Symp Quant Biol.* 71, pp. 301–311.
- Hernandez Jr, G., F. Valafar, and W. E. Stumph (2006). "Insect small nuclear RNA gene promoters evolve rapidly yet retain conserved features involved in determining promoter activity and RNA polymerase specificity". In: *Nucleic acids research* 35.1, pp. 21–34.
- Hertel, J. and P. F. Stadler (2006). "Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data". In: *Bioinformatics* 22.14, e197–e202.
- Hertel, J. and P. F. Stadler (2015). "The Expansion of Animal MicroRNA Families Revisited". In: *Life* 5, pp. 905–920. DOI: [10.3390/life5010905](https://doi.org/10.3390/life5010905).
- Hofacker, I. L. et al. (1994). "Fast Folding and Comparison of RNA Secondary Structures". In: *Monatsh. Chem.* 125, pp. 167–188.
- Hokii, Y. et al. (2010). "A small nucleolar RNA functions in rRNA processing in *Caenorhabditis elegans*". In: *Nucleic Acids Res.* 38, pp. 5909–5918. DOI: [10.1093/nar/gkq335](https://doi.org/10.1093/nar/gkq335).
- Hui, J. H. et al. (2013). "Structure, evolution and function of the bi-directionally transcribed iab-4/iab-8 microRNA locus in arthropods". In: *Nucleic acids research* 41.5, pp. 3352–3361.
- Iwakawa, H. O. and Y. Tomari (2015). "The Functions of MicroRNAs: mRNA Decay and Translational Repression". In: *Trends Cell Biol.* 25.11, pp. 651–665.
- Jeronimo, C et al. (2007). "Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme". In: *Mol Cell* 27, pp. 262–274. DOI: [10.1016/j.molcel.2007.06.027](https://doi.org/10.1016/j.molcel.2007.06.027).
- Ji, X. et al. (2013). "SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase". In: *Cell* 153.4, pp. 855–868.
- Kalvari, I. et al. (2017). "Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families". In: *Nucleic acids research* 46.D1, pp. D335–D342.
- Kalvari, I. et al. (2018). "Non-Coding RNA Analysis Using the Rfam Database". In: *Current protocols in bioinformatics*, e51.
- Karmakar, R et al. (2017). "A Comparative Study of Multiple Sequence Alignments". In:
- Kim, S.-J. et al. (2011). "Development of microRNA-145 for therapeutic application in breast cancer". In: *Journal of controlled release* 155.3, pp. 427–434.
- Kim, V. N. (2004). "MicroRNA precursors in motion: exportin-5 mediates their nuclear export". In: *Trends in cell biology* 14.4, pp. 156–159.

- Kim, V. N. (2005). "MicroRNA biogenesis: coordinated cropping and dicing". In: *Nature reviews Molecular cell biology* 6.5, p. 376.
- Kim, Y. K., B Kim, and V. N. Kim (2016). "Re-evaluation of the roles of DROSHA, Exportin 5, and DICER in microRNA biogenesis". In: *Proc Natl Acad Sci U S A* 113, E1881–E1889. DOI: [10.1073/pnas.1602532113](https://doi.org/10.1073/pnas.1602532113).
- Kozomara, A and S Griffiths-Jones (2014). "miRBase: annotating high confidence microRNAs using deep sequencing data". In: *Nucleic Acids Res.* 42, pp. D68–D73.
- Krueger, B. J. et al. (2008). "LARP7 is a stable component of the 7SK snRNP while P-TEFb, HEXIM1 and hnRNP A1 are reversibly associated". In: *Nucleic Acids Res* 36.7, pp. 2219–2229. DOI: [10.1093/nar/gkn061](https://doi.org/10.1093/nar/gkn061).
- Krüger, W. and B. J. Benecke (1987). "Structural and functional analysis of a human 7 S K RNA gene". In: *J. Mol. Biol.* 195, pp. 31–41.
- Lai, E. C. et al. (2003). "Computational identification of Drosophila microRNA genes". In: *Genome biology* 4.7, R42.
- Lebars, I. et al. (2010). "HEXIM1 targets a repeated GAUC motif in the riboregulator of transcription 7SK and promotes base pair rearrangements". In: *Nucleic Acids Res.* 38.21, pp. 7749–7763.
- Lee, H. Y. and J. A. Doudna (2012). "TRBP alters human precursor microRNA processing in vitro". In: *Rna* 18.11.
- Lee, Y. et al. (2002). "MicroRNA maturation: stepwise processing and subcellular localization". In: *The EMBO journal* 21.17, pp. 4663–4670.
- Lee, Y. et al. (2004). "MicroRNA genes are transcribed by RNA polymerase II". In: *The EMBO journal* 23.20, pp. 4051–4060.
- Li, L. et al. (2010). "Computational approaches for microRNA studies: a review". In: *Mammalian Genome* 21.1-2, pp. 1–12.
- Liu, B. et al. (2012). "MiR-26a enhances metastasis potential of lung cancer cells via AKT pathway by targeting PTEN". In: *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1822.11, pp. 1692–1704.
- Lorenz, R., I. L. Hofacker, and P. F. Stadler (2016). "RNA folding with hard and soft constraints". In: *Algorithms for Molecular Biology* 11.1, p. 8.
- Lorenz, R. et al. (2011). "ViennaRNA Package 2.0". In: *Alg. Mol. Biol.* 6, p. 26.
- Lund, E. et al. (2004). "Nuclear export of microRNA precursors". In: *Science* 303.5654, pp. 95–98.
- Macke, T. J. et al. (2001). "RNAMotif, an RNA secondary structure definition and search algorithm". In: *Nucleic acids research* 29.22, pp. 4724–4735.
- Markert, A. et al. (2008). "The La-related protein LARP7 is a component of the 7SK ribonucleoprotein and affects transcription of cellular and viral polymerase II genes". In: *EMBO Rep.* 9, pp. 569–575.
- Martinez-Zapien, D et al. (2017). "The crystal structure of the 5' functional domain of the transcription riboregulator 7SK". In: *Nucleic Acids Res* 45, pp. 3568–3579. DOI: [10.1093/nar/gkw1351](https://doi.org/10.1093/nar/gkw1351).

- Marz, M. and P. F. Stadler (2011). "RNA interactions." In: *Advances in experimental medicine and biology* 722, pp. 20–38.
- Marz, M. et al. (2009). "Evolution of 7SK RNA and its Protein Partners in Metazoa". In: *Mol. Biol. Evol.* 26, pp. 2821–2830.
- Marzluff, W. F. and R. J. Duronio (2002). "Histone mRNA expression: multiple levels of cell cycle regulation and important developmental consequences". In: *Current opinion in cell biology* 14.6, pp. 692–699.
- Masciullo, V. et al. (2003). "HMGA1 protein over-expression is a frequent feature of epithelial ovarian carcinomas". In: *Carcinogenesis* 24.7, pp. 1191–1198.
- Meister, G. (2013). "Argonaute proteins: functional insights and emerging roles". In: *Nature Reviews Genetics* 14.7, p. 447.
- Menzel, P., J. Gorodkin, and P. F. Stadler (2009). "The Tedious Task of Finding Homologous Non-coding RNA Genes". In: *RNA* 15, pp. 2075–2082.
- Michels, A. A. et al. (2004). "Binding of the 7SK snRNA turns the HEXIM1 protein into a P-TEFb (CDK9/cyclin T) inhibitor". In: *EMBO J.* 23, pp. 2608–2619.
- Morgenstern, B. (1999). "DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment." In: *Bioinformatics (Oxford, England)* 15.3, pp. 211–218.
- (2002). "A simple and space-efficient fragment-chaining algorithm for alignment of DNA and protein sequences". In: *Applied Mathematics Letters* 15.1, pp. 11–16.
- Morgenstern, B., A. Dress, and T. Werner (1996). "Multiple DNA and protein sequence alignment based on segment-to-segment comparison". In: *Proceedings of the National Academy of Sciences* 93.22, pp. 12098–12103.
- Morgenstern, B. et al. (2006). "Multiple sequence alignment with user-defined anchor points". In: *Alg. Mol. Biol.* 1, p. 6.
- Muniz, L. et al. (2010). "Controlling cellular P-TEFb activity by the HIV-1 transcriptional transactivator Tat". In: *PLoS pathogens* 6.10, e1001152.
- Murphy, S, C Di Liegro, and M Melli (1987). "The *in vitro* transcription of the 7SK RNA gene by RNA polymerase III is dependent only on the presence of an upstream promoter". In: *Cell* 51, pp. 81–87.
- Nature (2014). *Scitable, by nature education*. URL: <https://www.nature.com/scitable/glossary> (visited on 11/01/2018).
- Nawrocki, E. and S. Eddy (2016). *INFERNAL User's Guide*. English. Version Version 1.1.2. EddyRivasLab. 116 pp. July 1, 2016.
- Nawrocki, E. P. and S. R. Eddy (2013). "Infernal 1.1: 100-fold faster RNA homology searches". In: *Bioinformatics* 29.22, pp. 2933–2935.
- Neilsen, C. T., G. J. Goodall, and C. P. Bracken (2012). "IsomiRs—the overlooked repertoire in the dynamic microRNAome". In: *Trends in Genetics* 28.11, pp. 544–549.
- Nussinov, R. et al. (1978). "Algorithms for loop matchings". In: *SIAM Journal on Applied mathematics* 35.1, pp. 68–82.

- Olsen, P. H. and V. Ambros (1999). "The lin-4 regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation". In: *Developmental biology* 216.2, pp. 671–680.
- Peterlin, B. M. and D. H. Price (297-305). "Controlling the elongation phase of transcription with P-TEFb". In: *Mol. Cell.* 2006, p. 23.
- Peterlin, B. M., J. E. Brogie, and D. H. Price (2012). "7SK snRNA: a noncoding RNA that plays a major role in regulating eukaryotic transcription". In: *Wiley Interdisciplinary Reviews: RNA* 3.1, pp. 92–103.
- Proctor, J. R. and I. M. Meyer (2013). "C o F old: an RNA secondary structure prediction method that takes co-transcriptional folding into account". In: *Nucleic acids research* 41.9, e102–e102.
- Quinn, J. J. and H. Y. Chang (2016). "Unique features of long non-coding RNA biogenesis and function". In: *Nature Reviews Genetics* 17.1, p. 47.
- Raden, M. et al. (2018a). "Freiburg RNA tools: a central online resource for RNA-focused research and teaching". In: *Nucleic acids research*.
- Raden, M. et al. (2018b). "Interactive implementations of thermodynamics-based RNA structure and RNA–RNA interaction prediction approaches for example-driven teaching". In: *PLoS computational biology* 14.8, e1006341.
- Reinhart, B. J. et al. (2000). "The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*". In: *nature* 403.6772, p. 901.
- Rhoades, M. W. et al. (2002). "Prediction of plant microRNA targets". In: *cell* 110.4, pp. 513–520.
- Rodriguez, A. et al. (2004). "Identification of mammalian microRNA host genes and transcription units". In: *Genome research* 14.10a, pp. 1902–1910.
- Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." In: *Molecular biology and evolution* 4.4, pp. 406–425.
- Schramm, L. and N. Hernandez (2002). "Recruitment of RNA polymerase III to its target promoters". In: *Genes & development* 16.20, pp. 2593–2620.
- Scott, H. et al. (2012). "MiR-3120 is a mirror microRNA that targets heat shock cognate protein 70 and auxilin messenger RNAs and regulates clathrin vesicle uncoating". In: *Journal of Biological Chemistry* 287.18, pp. 14726–14733.
- Skrajna, A. et al. (2017). "U7 snRNP is recruited to histone pre-mRNA in a FLASH-dependent manner by two separate regions of the stem-loop binding protein". In: *RNA* 23.6, pp. 938–951.
- Smale, S. T. and J. T. Kadonaga (2003). "The RNA polymerase II core promoter". In: *Annual review of biochemistry* 72.1, pp. 449–479.
- Soutourina, J. (2018). "Transcription regulation by the Mediator complex". In: *Nature Reviews Molecular Cell Biology* 19.4, p. 262.
- Su, W. et al. (2018). "miR-30 disrupts senescence and promotes cancer by targeting both p16INK4A and DNA damage pathways." In: *Oncogene*.

- Sussman, I. (2018). "65 YEARS OF THE DOUBLE HELIX: Could Watson and Crick have envisioned the true impact of their discovery?" In: *Endocrine-related cancer* 25.8, E9–E11.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". In: *Nucleic acids Res.* 22.22, pp. 4673–4680.
- Tyler, D. M. et al. (2008). "Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci". In: *Genes & development* 22.1, pp. 26–36.
- Uchikawa, E. et al. (2015). "Structural insight into the mechanism of stabilization of the 7SK small nuclear RNA by LARP7". In: *Nucleic Acids Res* 43, pp. 3373–3388. DOI: [10.1093/nar/gkv173](https://doi.org/10.1093/nar/gkv173).
- Vidigal, J. A. and A. Ventura (2015). "The biological functions of miRNAs: lessons from in vivo studies". In: *Trends Cell Biol.* 25.3, pp. 137–147.
- Voinnet, O. (2009). "Origin, biogenesis, and activity of plant microRNAs". In: *Cell* 136.4, pp. 669–687.
- Wang, X. et al. (2005). "MicroRNA identification based on sequence and structure alignment". In: *Bioinformatics* 21.18, pp. 3610–3614.
- Wassarman, D. A. and J. A. Steitz (1991). "Structural analyses of the 7SK ribonucleoprotein (RNP), the most abundant human small RNP of unknown function." In: *Molecular Cellular Biol* 11.7, pp. 3432–3445.
- Weinberg, Z. and R. R. Breaker (2011). "R2R-software to speed the depiction of aesthetic consensus RNA secondary structures". In: *BMC bioinformatics* 12.1, p. 3.
- Wheeler, D. L. et al. (2007). "Database resources of the national center for biotechnology information". In: *Nucleic Acids Res.* 36.suppl_1, pp. D13–D21.
- Wu, Y. et al. (2011). "MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences". In: *BMC bioinformatics* 12.1, p. 107.
- Wurm, Y. et al. (2009). "Fourmidable: a database for ant genomics". In: *BMC genomics* 10, p. 5.
- Xiao, X. et al. (2017). "Novel Mutations in PRPF31 Causing Retinitis Pigmentosa Identified Using Whole-Exome Sequencing". In: *Investigative ophthalmology & visual science* 58.14, pp. 6342–6350.
- Xie, M. et al. (2008). "Size Variation and Structural Conservation of Vertebrate Telomerase RNA". In: *J. Biol. Chem.* 283, pp. 2049–2059.
- Xue, Y. et al. (2009). "A capping-independent function of MePCE in stabilizing 7SK snRNA and facilitating the assembly of 7SK snRNP". In: *Nucleic acids research* 38.2, pp. 360–369.
- Yu, D. et al. (2018). "Tracking microRNA processing signals by degradome sequencing data analysis". In: *Frontiers in Genetics* 9, p. 546.
- Yu, L. et al. (2016). "miRNA Digger: a comprehensive pipeline for genome-wide novel miRNA mining". In: *Sci. Rep.* 6, p. 18901.

- Zeng, Y. and B. R. Cullen (2004). "Structural requirements for pre-microRNA binding and nuclear export by Exportin 5". In: *Nucleic acids research* 32.16, pp. 4776–4785.
- Zhang, L. et al. (2017). "Cytosolic co-delivery of miRNA-34a and docetaxel with core-shell nanocarriers via caveolae-mediated pathway for the treatment of metastatic breast cancer". In: *Scientific reports* 7, p. 46186.
- Zuker, M. (2003). "Mfold web server for nucleic acid folding and hybridization prediction". In: *Nucleic acids research* 31.13, pp. 3406–3415.
- Zuker, M. and P. Stiegler (1981). "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information". In: *Nucleic acids research* 9.1, pp. 133–148.