MiRNA and co: Methodologically exploring the world of small RNAs

Susan Higashi

N^o d'ordre: 252-2014 Année 2014

Thèse

Présentée

devant l'Université Claude Bernard - Lyon 1

pour l'obtention

du Diplôme de Doctorat

(arrêté du 7 août 2006)

et soutenue publiquement le
26 Novembre 2014

par

Susan Higashi

MiRNA and co: Methodologically exploring the world of small RNAs

---

Directeur de thèse: Marie-France Sagot
Co-Directeur: Christian Gautier
Co-Encadrant: Stefano Colella

Jury:	Hubert Charles,	Examinateur
	Christine Gaspin,	Examinateur
	Hervé Seitz,	Rapporteur
	Peter Stadler,	Rapporteur
	Hélenè Touzet,	Rapporteur

UNIVERSITÉ CLAUDE BERNARD-LYON 1

Président de l'Université	**M. François-Noël GILLY**
Vice-président du Conseil d'Administration	M. le Professeur Hamda BEN HADID
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur Philippe LALLE
Vice-président du Conseil Scientifique	M. le Professeur Germain GILLET
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est - Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maëutique Lyon Sud - Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme. la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. FLEURY
Département Chimie Biochimie	Directeur : Mme Caroline FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. Georges TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. Jean-Claude PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. P. FOURNIER
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Acknowledgements

I foremost offer my sincerest gratitude to my advisors Marie, Christian, and Christine, whose expertise and constant support, always provided with kindness and patience, were crucial to this thesis. A special thanks to Christine, who was not an official advisor, however, was always present during this three years. I also thank my co-encadrant Stefano for the biological support always provided with enthusiasm.

Many thanks to our collaborators: Ana Tereza Ribeiro de Vasconcelos, Nuno Mendes, Olivier Rue, Hubert Charles, Federica Calevro, Karen Gaget, Gabrielle Duport, Susana Vinga, Nadia Pisanti, Roberto Grossi, Caio Padoan de Sá Godinho. It has been extremely rewarding to work and learn with all of them.

A very special thanks to all my colleagues of Erable (former Bamboo) team for the friendly atmosphere, which was essential to the progress of this thesis.

Finally, my deepest gratitude to my family Alice, Nelson, and Yudji, without whom none of this would be possible.

Contents

Introduction and Motivation **11**

1 Background **1**
1.1 Biological background . . . 2
1.1.1 MicroRNA definition, history, and landscape . . . 2
1.1.2 MicroRNA biogenesis . . . 5
1.1.3 RNA-induced silencing complex . . . 6
1.1.4 Plant microRNAs . . . 7
1.2 Methodological background . . . 8
1.2.1 miRBase: a reference for microRNA studies . . . 8
1.2.2 Computational methods for microRNA identification . . . 9
1.2.3 Experimental methods for microRNA detection and quantification . . . 18
1.2.4 Computational methods for target prediction . . . 22
1.2.5 Experimental methods for microRNA target identification . . . 23

2 MIRINHO: Efficient precursor miRNA predictor **25**
2.1 Introduction . . . 25
2.2 Material and methods . . . 26
2.2.1 Algorithm . . . 26
2.2.2 Dataset . . . 27
2.2.3 Compared methods . . . 29
2.2.4 Measuring sensitivity and precision . . . 30
2.3 Results and discussion . . . 30
2.3.1 Regression analysis of the free energies . . . 30
2.3.2 Time efficiency . . . 30
2.3.3 miRNA hairpin structure prediction in sRNA-seq of plant . . . 33
2.3.4 Sensitivity and precision . . . 34
2.4 Conclusion . . . 40

3 MicroRNA expression profile during embryonic development in *A. pisum*: combining deep sequencing data and MIRINHO to identify miRNAs **43**
3.1 Introduction . . . 44
3.2 Material and methods . . . 44
3.2.1 Aphid rearing and embryo isolation . . . 44
3.2.2 RNA extraction . . . 46
3.2.3 Next-generation Illumina Sequencing . . . 46
3.2.4 Treatment of the small RNA sequencing data . . . 46
3.2.5 MicroRNA expression profile . . . 49

3.2.6 Target prediction . . . 49
3.3 Results and discussion . . . 50
3.3.1 Statistical summary of the sequenced reads . . . 50
3.3.2 MicroRNAs Expressed in *Acyrthosiphon pisum* . . . 53
3.3.3 MicroRNA gene expression profile . . . 63
3.3.4 MicroRNA target prediction . . . 63
3.4 Conclusion . . . 67

4 Prediction of non-coding RNAs and targets in *Mycoplasma hyopneumoniae* 69
4.1 Introduction . . . 70
4.2 Material and methods . . . 70
4.2.1 Prediction of non-coding RNAs . . . 70
4.2.2 Alvinho: An algorithm for the prediction of non-coding RNA targets 74
4.2.3 Conservation analysis . . . 75
4.3 Results and discussion . . . 77
4.3.1 Identified ncRNA candidates . . . 77
4.3.2 Predicted non-coding RNA targets . . . 77
4.3.3 Conserved ncRNAs . . . 83
4.4 Conclusion . . . 83

5 Cluster analysis of structured motifs 85
5.1 Introduction . . . 85
5.2 Materials and methods . . . 86
5.2.1 A brief reminder on Smile . . . 86
5.2.2 Unweighted Pair Group Method with Arithmetic Mean . . . 89
5.3 Initial results and discussion . . . 90
5.4 Conclusion . . . 90

Conclusion and Perspectives 93

Bibliography 95

Abstract

The main contribution of this thesis is the development of a reliable, robust, and much faster method for the prediction of pre-miRNAs. With this method, we aimed mainly at two goals: efficiency and flexibility. Efficiency was made possible by means of a *quadratic* algorithm. Since the majority of the predictors use a *cubic* algorithm to verify the pre-miRNA hairpin structure, they may take too long when the input is large. Flexibility relies on two aspects, the input type and the organism clade. MIRINHO can receive as input both a genome sequence and small RNA sequencing (sRNA-seq) data of both animal and plant species. To change from one clade to another, it suffices to change the lengths of the stem-arms and of the terminal loop. Concerning the prediction of plant miRNAs, because their pre-miRNAs are longer, the methods for extracting the hairpin secondary structure are not as accurate as for shorter sequences. With MIRINHO, we also addressed this problem, which enabled to provide pre-miRNA secondary structures more similar to the ones in MIRBASE than the other available methods.

MIRINHO served as the basis to two other issues we addressed. The first issue led to the treatment and analysis of sRNA-seq data of *Acyrthosiphon pisum*, the pea aphid. The goal was to identify the miRNAs that are expressed during the four developmental stages of this species, allowing further biological conclusions concerning the regulatory system of such an organism. For this analysis, we developed a whole pipeline, called MIRINHOPIPE, at the end of which MIRINHO was aggregated.

We then moved on to the second issue, that involved problems related to the prediction and analysis of non-coding RNAs (ncRNAs) in the bacterium *Mycoplasma hyopneumoniae*. A method, called ALVINHO, was thus developed for the prediction of targets in this bacterium, together with a pipeline for the segmentation of a numerical sequence and detection of conservation among ncRNA sequences using a k-partite graph.

We finally addressed a problem related to motifs, that is to patterns, that may be composed of one or more parts, that appear conserved in a set of sequences and may correspond to functional elements. This had already been addressed in a robust method called SMILE. However, depending on the input parameters, the output may be too large to be tractable, as was realized in other works of the team. We then presented some clustering solutions to group the motifs that may correspond to a same biological element, and thus to better distinguish the biologically significant ones from noise that may be present in what often are large outputs from many motif extraction algorithms.

Introduction

This thesis mainly addresses methodological problems related to the prediction of small regulatory RNAs, specially microRNAs (miRNAs). The first topic involved the elaboration of a robust and efficient method, called MIRINHO, for the prediction of pre-miRNAs in both genomic and small RNA sequencing (sRNA-seq) data of both animal and plant species. The second topic led to the development of a pipeline, called MIRINHOPIPE, for the treatment of small RNA sequencing data. It was specially implemented to identify the expressed miRNAs of *Acyrthosiphon pisum*, the pea aphid. We then moved on to solve a few problems related to the prediction and analysis of non-coding RNAs (ncRNAs) in the bacterium *Mycoplasma hyopneumoniae*. A method, called ALVINHO, was thus developed for the prediction of targets in this bacterium, together with a pipeline for the segmentation of a numerical sequence and detection of conservation among ncRNA sequences using a k-partite graph. We finally addressed a problem related to motifs, that is to patterns, that may be composed of one or more parts, that appear conserved in a set of sequences and may thus correspond to functional elements such as DNA binding sites or miRNA families (i.e., all the isoforms of a same miRNA). We presented some clustering solutions to group the motifs that may correspond to a same such biological element, and thus to better distinguish the biologically significant ones from noise that may be present in what often are large outputs from many motif extraction algorithms.

All the methodological and biological concepts required to understand the previous topics are presented in **Chapter 1**. We now provide a brief introduction to each of these topics.

Given the importance and ubiquity of miRNAs in a wide range of biological processes and diseases, a plethora of methods for the prediction of miRNAs were developed. Despite all the effort put in developing them, there remained a number of issues that needed to be addressed:

1. the vast majority of the existing softwares rely on a folding algorithm of cubic time complexity to predict the characteristic hairpin structure of a pre-miRNA: this is suitable when the input is small enough, but it can become impracticable when the size of the input increases;

2. for longer pre-miRNAs (such as in plant), such folding methods moreover can produce hairpin structures different from the ones provided in MIRBASE (Kozomara et Griffiths-Jones, 2011), as a consequence the miRNA may be located in a different place than a stem-arm;

3. together with folding, most methods then rely on further information that must be learned from previously validated miRNAs of closely related genomes (at a minimum within the same clade, plant or animal) for the final prediction of new miRNAs in order either to set the parameters of the model or to restrict the search to a limited space.

MIRINHO was therefore developed to address all three issues. The search for pre-miRNAs is concentrated on regions with the same length as the two stem-arms separated by the length of the terminal loop. The direct application to *sRNA-seq* data guarantees a better quality in the prediction of the pre-miRNA structures. A *quadratic* time complexity algorithm improves the practical efficiency of the free energy computation. As neither of the two attributes used (length of stem-arm and terminal loop) are species-specific within the animal or the plant kingdom (they differ only between these two kingdoms), the method can easily be applied for predicting pre-miRNAs in either clade. Importantly, while the method we provide is thus much simpler, faster, and general to use, we also show for tested examples that it has a sensitivity and precision as good as other methods, in some cases even better. Moreover, we show that the secondary structures predicted by MIRINHO are much closer to the ones available in MIRBASE than for the other compared methods. MIRINHO is described in detail in **Chapter 2** which is strongly based on our paper Higashi *et al.* (ress).

Still concerning the identification of miRNAs, however from another perspective, we treated and analysed the small RNA sequencing (sRNA-seq) data of *Acyrthosiphon pisum*, the pea aphid. The unique feeding habit of aphids combined with their ability to rapidly reproduce makes of them one of the most damaging pests of crops with economical importance worldwide. Considering their impact on agriculture and the role miRNAs play in gene regulation, it is imperative to better characterise and understand the function of these miRNAs. One first effort has already been made by Legeai *et al.* (2010a) in *A. pisum*, a laboratory model for the study of these pests whose genome was sequenced. It is worth noting that in Legeai's work, the miRNAs of parthenogenic females were sequenced and analysed, while we focus on the miRNAs expressed in three embryonic developmental and one larval stages. Furthermore, the potential mRNA targets of the detected miRNAs were identified by the overlapped predictions of two methods (PITA and MIRANDA), and correlated with the gene expression profile of the pea aphid.

To treat the data in order to guarantee a more accurate set of reads, as well as to detect the expressed miRNAs, three approaches were used: (i) MIRINHOPIPE, specially developed for this analysis; (ii) SRNA-PLAN, a pipeline designed for the annotation of small RNAs; and (iii) MIRDEEP, a classical method for the discovery of miRNAs from deep sequencing data (Friedländer *et al.*, 2008). The detected miRNAs were submitted to the prediction of mRNA targets. Together with such predictions, the gene expression profile of *A. pisum* was analysed and compared to the miRNA expression profile, leading to very interesting results. All the methodology, results and discussion are presented in **Chapter 3** which is strongly based on our paper Higashi *et al.* (tion).

Besides miRNAs, another small regulatory molecule was investigated. This was non-coding RNAs in *Mycoplasma hyopneumoniae*. The bacterium *M. hyopneumoniae* strain 7448 is a pathogenic and obligate parasite of porcine respiratory systems. It lives adhered to the epithelium of its host respiratory tract, and together with other bacteria and viruses, it is considered one of the ethiologic agents of swine enzootic pneumonia. The disease can cause a decrease in the productivity of these animals, sometimes resulting in their death (BYRT *et al.*, 1985; DeBey et Ross, 1994; Brockmeier *et al.*, 2002). Although some effort has already being put in understanding the infection process, the specific mechanisms relating the bacterium to the disease remain unknown (Gardner et Minion, 2010; Hsu et Minion, 1998; Nicolás *et al.*, 2007; Siqueira *et al.*, 2013).

M. hyopneumoniae 7448 has only one known transcription factor (TF) and a complex gene expression pattern. The incomparability between the number of regulatory elements and the complexity of the gene expression of the bacterium, together with increasing evidences that

ncRNAs are involved in this phenomenon, strongly encourage the search for ncRNAs in the genome of *M. hyopneumoniae* 7448. After predicting the regions with a potential to harbour ncRNA genes, additional analyses were performed in an attempt to provide more evidences to carry on with experimental validation of the ncRNAs.

The first problem that concerned us was related to the output of the pipeline for the prediction of ncRNAs: such pipeline was generating one single assembled ncRNA sequence where two or more different ncRNA candidates were in fact present. We solved this by applying a segmentation algorithm on these outputs. To then provide stronger evidence that the candidates were indeed functional, we performed the prediction of the ncRNA targets with a method, called ALVINHO, that was specially developed for this purpose. Finally, to verify if conservation could play any role in the functionality of ncRNAs, the identity of intergenic regions was assessed between closely-related *Mycoplasma* species by means of a k-partite graph. Genomic motifs surrounding the ncRNA, such as promoters and terminators, were also verified to reinforce the functional evidence of the ncRNA candidates. All the three steps of the pipeline are available in the form of a script or a C++ implementation. All the details concerning the methods developed for the analysis of the ncRNAs are presented in **Chapter 4** that is based on the paper Godinho *et al.* (tion).

We then looked at a problem related to structured motifs, which corresponds to a possibly complex pattern that is conserved in a sequence or a set of sequences. This is an issue that may seem unrelated to the study of miRNAs but the two may however appear combined in some studies. For instance, the motifs associated to the miRNAs that are exported from a human tissue might enable to understand what distinguishes such miRNAs from those that are not exported.

The problem of finding structured motifs was first addressed by Marsan et Sagot (2000) and implemented as a software called SMILE (Structured Motifs Inference and Evaluation). Depending on the parameters given to SMILE, the algorithm can generate a large output that may contain redundant information. This will happen in particular when the characteristics of the motifs are not precisely known, thus requiring that more permissive parameters are adopted in an attempt to recover them. We therefore present some clustering solutions to group together motifs that may correspond to a same biological "object", and to better identify the noise that may be present in such large outputs.

Efficiently extracting consensus sites in a set of sequences is an essential approach to identify functional elements in a genome. Examples of such elements are DNA binding sites and miRNA families (i.e., a consensus that represents all the miRNA isoforms). There are two main problems related to this identification. One is the prediction of the location of the element site, and the second is the extraction of the consensus. The algorithm SMILE (Marsan et Sagot, 2000) addresses both problems: extracting and locating consensus motifs in a set of sequences. To solve this problem, SMILE implements an exact algorithm for finding motifs in a set of sequences. A suffix tree is used to represent the input sequences, which together with the strategies implemented in the algorithm, result in an efficient method for the extraction of motifs. SMILE requires a number of parameters, such as the number p of boxes a (structured) motif may have, the minimum number of substitutions e (one per box) between the motif and its occurrence, and the minimum number of times q (stands for quorum) the motif has to appear among the sequences.

Depending on the values of these parameters, the size of the output generated by SMILE may be very large, containing redundant motifs. For example, the larger is the number e or the smaller the quorum q, the larger will be the output. In an attempt to organise such output eliminating the redundancy, we implemented an UPGMA (Unweighted Pair Group Method

with Arithmetic Mean) algorithm to cluster the similar motifs according to the positions where they appear. The implementation of this algorithm was performed during the internship of Thomas Balezeau, an undergraduate student in information technology whom I co-advised together with Marie-France Sagot. Another approach that has been explored, but not yet implemented, is the use of hashing for list intersection as an estimator to find redundant motifs. All the details concerning structured motifs and the clustering approaches are presented in **Chapter 5**.

Chapter 1

Background

Contents

1.1 Biological background 2
1.1.1 MicroRNA definition, history, and landscape 2
1.1.2 MicroRNA biogenesis 5
1.1.3 RNA-induced silencing complex 6
1.1.4 Plant microRNAs 7
1.2 Methodological background 8
1.2.1 miRBase: a reference for microRNA studies 8
1.2.2 Computational methods for microRNA identification 9
1.2.3 Experimental methods for microRNA detection and quantification . 18
1.2.4 Computational methods for target prediction 22
1.2.5 Experimental methods for microRNA target identification 23

In this chapter, we present the biological and computational backgrounds required to the comprehension of this thesis. It is certainly not possible to cover all the details about the concerned topics to provide a self-contained thesis; we therefore provide only the concepts that we find crucial for both the computational and biological sides. The chapter is divided in two sections: Section 1.1 presents the biological concepts and Section 1.2 covers the methodological concepts (computational and experimental).

The purpose of Section 1.1 is to present microRNAs (miRNA) and the involved machinery. We thus begin by defining a miRNA and by providing its historical background and current landscape to place miRNAs in a small regulatory context; from this exposition, one should be convinced of the importance of miRNAs in the different biological processes in which they are involved. We then present the miRNA biogenesis process, an important issue since any computational modelling of a miRNA is strongly based on this process. We thus address the RNA-induced silencing complex (RISC) that is responsible for the functional regulatory interaction between a miRNA and its target messenger RNA (mRNA).

The main goal of Section 1.2 is to introduce the current computational and experimental methods used to detect miRNAs and targets. We first present how the two problems, prediction of miRNAs and prediction of targets, are computationally addressed. We then introduce the experimental methods used to detect miRNAs and targets, an important aspect that may complement and validate the results obtained by the computational methods.

1.1 Biological background

1.1.1 MicroRNA definition, history, and landscape

A miRNA is a small non-coding regulatory molecule, present in animals, plants, and in a few viruses. It is responsible for the post-transcriptional regulation of gene expression via complementarity base-pairing with the target mRNA; frequently the result of the regulation is the silencing of the target, however, there are fewer cases in which the expression is enhanced. These transcripts of ∼22nt are derived from a precursor-miRNA (pre-miRNA) with a specific hairpin (stem-loop) structure, with small internal loops and bulges, and are located in the stem of the hairpin. A "classical" miRNA would meet all the previous features. Although in practice variations are obviously possible, the minimum requirement to classify a sequence as a miRNA is its length (∼22nt) and the presence of a hairpin loop (Berezikov *et al.*, 2006; Chen et Rajewsky, 2007; He et Hannon, 2004).

These molecules are believed to be involved in the regulation of several basic pathways, such as in the transition of developmental stages in nematodes (*lin-4* and *let-7*) (Reinhart *et al.*, 2000; Wightman *et al.*, 1993; Lee *et al.*, 1993), cell proliferation and apoptosis (miRNA *Bantam*) (Brennecke *et al.*, 2003), regulation of fat metabolism (*miR-14*) (Xu *et al.*, 2003), etc. Furthermore, they also known to play a role in diseases such as autoimmune and neurodegenerative diseases, and in cancer (Almeida *et al.*, 2011).

The first investigations concerning RNA interference (RNAi), which is a process of inhibiting gene expression (i.e., of gene silencing), started in 1990 with the efforts of two teams (Napoli *et al.*, 1990; Van der Krol *et al.*, 1990). The authors used a transgene in an attempt to over-express an enzyme related to the violet color of petunias. In the end, instead of obtaining darker violet petunias as it was expected, they observed white ones. They thus raised the hypothesis that the endogenous and transgenic genes were co-suppressed.

Three years later, the first miRNA was identified in the nematode *Caenorhabditis elegans* by Lee *et al.* (1993) and Wightman *et al.* (1993). The authors cloned the gene *lin-4* and discovered that it did not encode a protein but instead a small RNA of 21nt. They observed that, by partial complementarity between the miRNA and the 3′UTR of the *lin-14* mRNA, the translation of the protein LIN-14 was being repressed.

In 1998, the classically established flow of information inside a cell (the so-called central dogma) became more complex as the pathway of RNAi was first described by Fire *et al.* (1998), with RNAs regulating other RNAs, instead of only producing proteins. The authors discovered that, instead of a single strand RNA (ssRNA), the trigger for the gene silencing in *Caenorhabditis elegans* was a double strand RNA (dsRNA) (Sen et Blau, 2006). This work introduced a new concept for the gene silencing pathway, clarifying the results of previous works (Napoli *et al.*, 1990; Van der Krol *et al.*, 1990; Guo et Kemphues, 1995), and maybe becoming one of the best known pathways for RNA silencing, since it is possible to repress the expression of a wide range of genes with just partial sequence complementarity.

In 2000, a second miRNA, namely *let-7*, was discovered in *Caenorhabditis elegans* by Reinhart *et al.* (2000). The authors noticed that the repression of *let-7* caused the reappearance of larval characteristics during the adult stage, while the over-expression of *let-7* caused the early expression of adult characteristics. The authors then concluded that miRNA *let-7* was controlling the transition of developmental stages in *Caenorhabditis elegans*. A timeline with the major discoveries in gene silencing can be found in Figure 1.1.

Gradually other types of small non-coding RNAs were being discovered: Piwi-interacting RNAs (piRNA) (Siomi *et al.*, 2011), transcription initiation RNAs (tiRNA) (Taft *et al.*, 2009),

1960s July 1969 Britten and Davidson propose that RNA regulates eukaryotic gene expression	**1970s** October 1972 Human cells are shown to contain nuclear double-stranded RNA	**1990s** April 1990 Cosuppression discovered in plants December 1993 The first microRNA, lin-4, is discovered February 1994 RNA found to direct DNA methylation of plant viroids May 1994 Calgene's "antisense" Flavr Savr tomato approved for sale by the FDA	May 1995 Both sense and antisense RNA found to inhibit gene expression in C. elegans June 1997 An Argonaute protein, Piwi, is linked to stem cell maintenance February 1998 Double-stranded RNA is discovered to be the trigger of RNA interference (RNAi)	October 1998 Plant viruses shown to encode RNA silencing suppressors October 1999 Argonaute proteins found to be required for RNAi October 1999 - March 2000 Small interfering RNAs (siRNAs) discovered as guides for RNA silencing
2000s October 2000 Double-stranded RNA shown to direct DNA methylation January 2001 Dicer shown to make siRNAs May 2001 RNAi discovered in human cells July 2001 Dicer found to make microRNAs (miRNAs)	October 2001 miRNAs are established as a large class of gene regulators July 2002 Plant miRNAs are discovered July 2002 siRNAs are revealed as triggers of RNAi in mice	September 2002 Small RNAs guide the production of heterochromatin at centromeres November 2002 miRNAs implicated in cancer September 2003 It is clear that miRNA maturation begins in the nucleus	November 2003 Dicer shown to be required for mouse embryogenesis, and perhaps for stem cell production March 2004 Human genome-wide RNAi libraries become available April 2004 Animal viruses found to encode miRNAs	August 2004 First "investigational new drug" application filed for a therapeutic siRNA September 2004 Argonaute is revealed as the RNAi endonuclease, "Slicer" June 2005 miRNAs shown to act as oncogenes July 2005 Primate-specific miRNAs identified
March 2006 miRNAs hsa-mir-155 and hsa-let-7a-2 associated to lung cancer June 2006 Epigenetic regulation of miRNAs	June 2007 miRNA target can also occur in 5'-UTR September 2007 miRNAs can regulate ncRNAs from the category of long ultraconserved genes (UCGs) December 2007 miRNAs can up-regulate mRNA expression and initiate the translation of proteins	February 2008 miRNA (miR-373) targets promoter sequences and induces gene expression October 2008 Functional single nucleotide polymorphism (SNP) in the miRNA seed region; miRNA binding sites located within coding sequence June 2009 proof of concept of miRNA delivery as cancer therapy	**2010s** March 2010 miRNA as molecular decoys August 2010 miRNAs predominantly cause mRNA destabilization September 2010 Overexpression of a single miRNA is sufficient to cause cancer	August 2011 Competing endogenous RNA (ceRNA) communicate with and regulate other RNA transcripts by competing for shared miRNAs

Figure 1.1: *Timeline of the main discoveries in gene silencing (Zamore et Haley, 2005; Kunej et al., 2012).*

nucleolar RNAs (snoRNA) (Filipowicz et Pogačić, 2002; Dieci *et al.*, 2009), and other miRNAs (Winter et Diederichs, 2011; Siomi et Siomi, 2010; de Planell-Saguer et Rodicio, 2011) have also been identified. The full landscape of such small regulatory RNAs is presented in Figure 1.2.

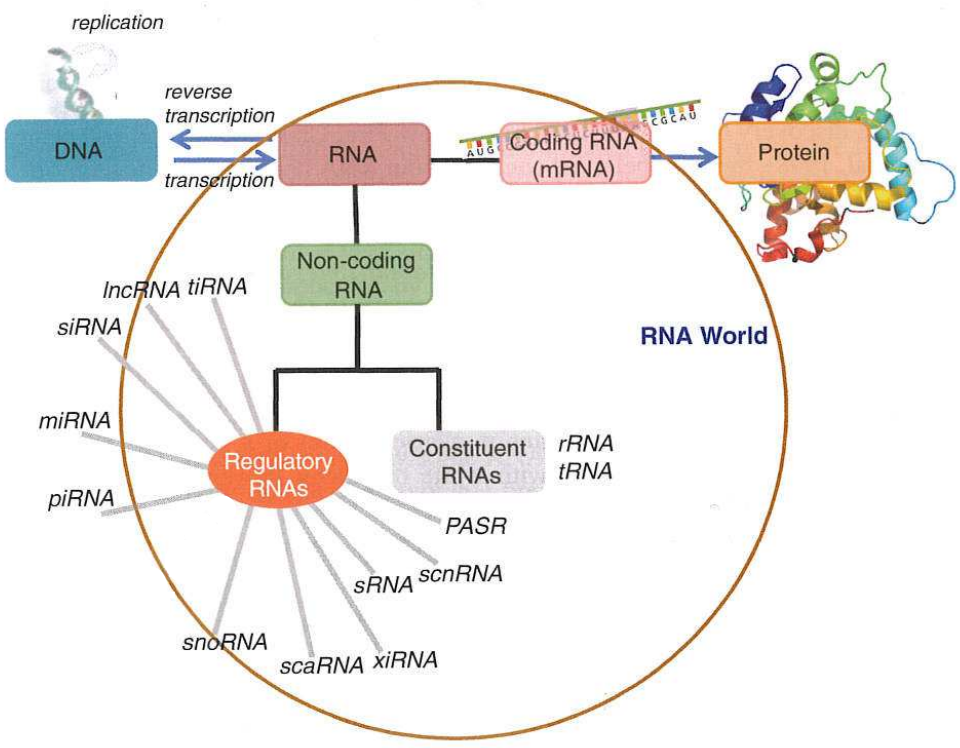


Figure 1.2: *RNA landscape and the different types of small non-coding RNAs: transcription initiation RNA (tiRNA) (Taft* et al., *2009), long non-coding RNA (lncRNA) (Geisler et Coller, 2013; Batista et Chang, 2013), small interfering RNA (siRNA) (Castel et Martienssen, 2013; Davidson et McCray, 2011), Piwi-interacting RNA (piRNA) (Siomi* et al., *2011), small nucleolar RNA (snoRNA) (Filipowicz et Pogačić, 2002; Dieci* et al., *2009), small Cajal body-specific RNA (scaRNA) (Darzacq* et al., *2002), X-inactivation RNA (xiRNA) (Ogawa* et al., *2008), small RNA (sRNA) (Gottesman et Storz, 2011), small-scan RNA (scnRNA) (Kim, 2005), promoter-associated small RNA (PASR) (Kapranov* et al., *2007) (image modified from Ghosh et Mallick (2012)).*

1.1.2 MicroRNA biogenesis

Concerning the transcription of miRNA genes, these molecules can arise either from intergenic regions or from introns of spliced genes. They are either transcribed as independent units or in clusters of miRNAs by means of a polycistronic transcript. Figure 1.3 will serve as a support for all the explanation given in what follows.

The transcription of miRNA genes is mainly performed by the RNA polymerase II (Pol II). The Pol II begins the transcription in the nucleus by binding to the promoter, and its first product is a longer transcript (between 500bp and 10Kbp) called primary-miRNA (pri-miRNA), which is capped at the $5'$ end and polyadenylated at the $3'$ end. Still in the

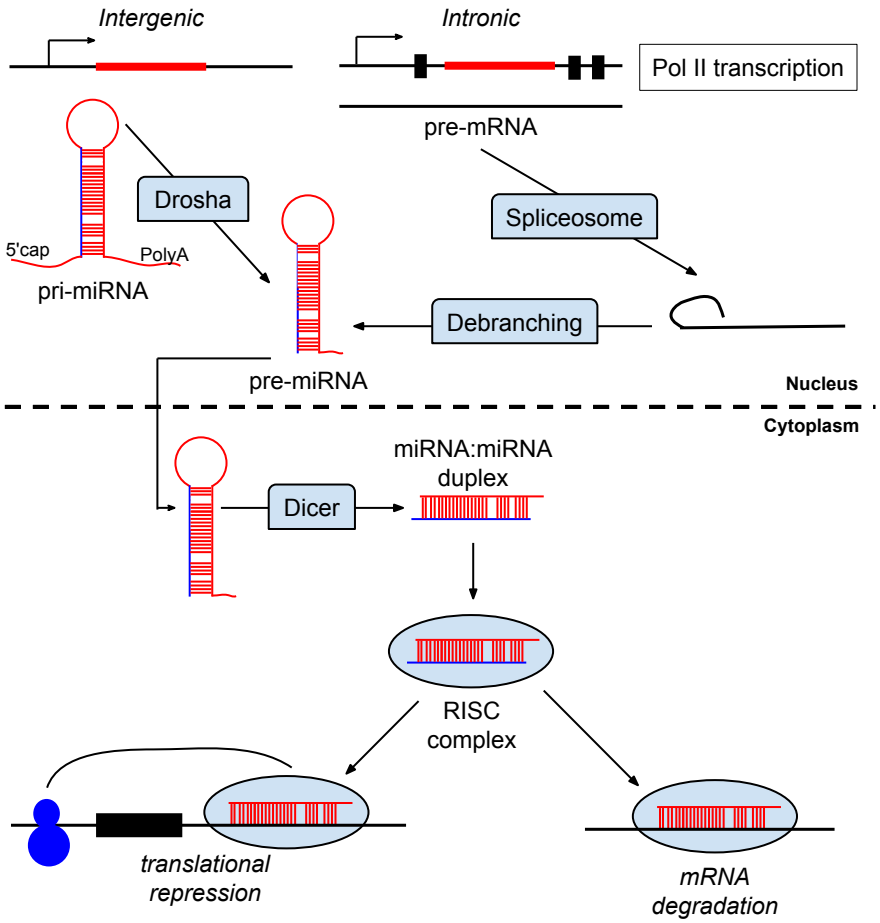


Figure 1.3: *Animal miRNA biogenesis: A miRNA can be located either in an intergenic region or in an intronic region of a protein encoding gene. In the first canonical pathway (left side of the figure), the miRNA is transcribed from its own gene into a pri-miRNA, which will then be processed by Drosha into a pre-miRNA. In the second non-canonical pathway (right side of the figure), the host gene is transcribed, spliced and the miRNA comes from an intron—in this case the miRNA is called mirtron. After debranching, the sequence folds itself into a pre-miRNA, and does not require processing by Drosha. The pre-miRNA is exported by Exportin-5 to the cytoplasm, where it is further processed by the enzyme Dicer into a duplex miRNA:miRNA*; usually only one of the strands is loaded into the RISC complex and the other miRNA* is degraded.*

nucleus, the pri-miRNA is processed by a microprocessor (composed of the Drosha RNase III enzyme and its cofactor DiGeorge syndrome critical region gene 8 (DGCR8)) into a ∼80nt stem-loop precursor miRNA (pre-miRNA) with a 2 nucleotides overhang at the 3′ end. The pre-miRNA is then exported by the nucleocytoplasmic shuttler Exportin-5 to the cytoplasm, where another RNase III enzyme called Dicer and its cofactor transactivating response RNA-binding protein (TRBP) recognise the 2 nucleotides overhang left by Drosha and cleaves the terminal loop. The result is a short imperfect miRNA:miRNA* duplex of length ∼22nt that is unwounded, producing one functional strand (mature miRNA) and another non-functional miRNA* (miRNA star) that is usually degraded, although sometimes it can be functional too (Petersen *et al.*, 2006; Yang *et al.*, 2011; Okamura *et al.*, 2008). The duplex configuration is known to stabilise the miRNA by protecting it from RNases degrading single-stranded miRNAs. Recently, Winter *et al.* (2013) provided a first evidence that single-stranded loop regions may give origin to functional regulatory miRNAs, which the authors call loop-miRNAs.

During the splicing of other genes, miRNAs can also arise from introns. After splicing, the intronic region folds into a pre-miRNA stem loop and it is then submitted to the same canonical biogenesis pathway. This kind of miRNA is called mirtron and is independent of the activity of Drosha. After all the processing, the mature miRNA is incorporated into a complex called RNA-induced silencing complex (RISC) to be further driven for target regulation, as it is detailed in the next section (Okamura *et al.*, 2007; Meister, 2013; Gommans et Berezikov, 2012).

1.1.3 RNA-induced silencing complex

Once the miRNA is assembled into RISC, it anneals to its mRNA target for regulation. The RISC complex is comprised of several proteins, among them one is well known: the Argonaute (AGO) protein that is the component that links the miRNA with the complex by means of two domains, PAZ and PIWI, responsible for miRNA recognition. The RISC assembly may be divided into at least two successive steps: RISC-loading, in which miRNA duplexes are inserted into the AGO proteins; and (ii) strand dissociation, in which the two miRNA strands are separated within the AGO protein. During assembly, the AGO proteins suffer conformation changes, made by chaperones, to allow for the incorporation of the miRNA duplex. Once the duplex is incorporated, AGO releases the tension to recover its original conformation unwinding the duplex and discarding the non-functional miRNA strand called "passenger strand". The remaining functional mature miRNA (or "guide strand") is usually the one with the less stable 5′ end. This mechanism of RISC assembly is mostly studied in *Drosophila* using AGO2-RISC as a model system. Although the exact molecular composition of RISC is unknown, a sufficient requirement for target regulation is the Argonaute protein (Kawamata et Tomari, 2010; Meister, 2013; Scott, 2012).

Usually, the effect of RISC, miRNA and target interaction is down-regulation, either through the cleavage of the mRNA target or by repression of the translation. To cleave the target, at least two requirements are necessary: an Argonaute with catalytic activity (in humans only AGO2 has this characteristic), and a near-perfect complementarity between the guide miRNA and its target. Different from the cleavage, near-perfect complementarity is not required to repress translation. Instead, only a smaller region of 6nt, that is called *seed*, requires perfect complementarity. It is usually located at the 5′ end of the miRNA (positions 2 to 8) and is known to be more frequent in animals (Zheng et Zhang, 2010). When the pairing at the 5′ end is insufficient, stronger pairing at the 3′ end compensates for it (Brennecke *et al.*, 2005). Concerning the mRNA target, the interaction is usually located in the 3′ untranslated

region (3UTR) (Pratt et MacRae, 2009). However, studies such as Lytle *et al.* (2007); Moretti *et al.* (2010); Ørom *et al.* (2008); Qin *et al.* (2010); Fang et Rajewsky (2011) suggested that the association can be functional in the 5′ UTR also or even in the CDS.

It has been shown that the repression is even more effective when there are multiple miRNAs binding to the same mRNA target, suggesting that the regulation is controlled by multiple miRNAs (Fang et Rajewsky, 2011; Krek *et al.*, 2005; Grimson *et al.*, 2007). One more characteristic that was found to contribute to the regulatory effect is the AU content in the 3′ of the seed region (Jing *et al.*, 2005; Grimson *et al.*, 2007). Figure 1.4 shows a schema of the interaction between a miRNA and its target.

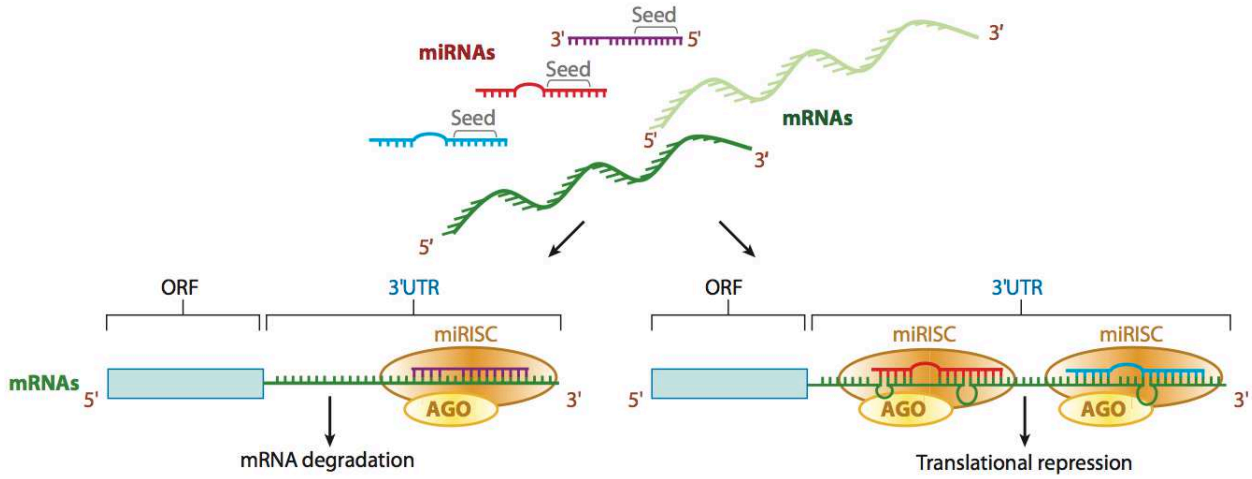


Figure 1.4: *Mechanisms of interaction between miRNA and target. The interaction of miRNA and target can cause either mRNA degradation, when the complementarity between miRNA and target is near perfect, or repression of translation, when there is partial sequence complementarity in relation to the whole miRNA sequence. A basic feature required for the interaction is seed, a region starting at the second nucleotide of the 5' end of the miRNA with perfect base pairing. In relation to the target, the interaction frequently occurs at the 3'UTR. On the right side of this figure, the miRNAs act in a synergistic way: multiple miRNAs (the ones in red and blue) bind to the same target to cooperatively regulate it. The figure was taken from Sun* et al. *(2010).*

Although regulation by miRNAs has been widely studied, a model describing in detail the mechanisms of the different modes of actions is still being debated (Lytle *et al.*, 2007; Moretti *et al.*, 2010; Ørom *et al.*, 2008; Qin *et al.*, 2010; Fang et Rajewsky, 2011).

1.1.4 Plant microRNAs

For the sake of concision, we will highlight just the differences between animal and plant miRNAs. Starting by the transcription, it seems that the great majority of plant miRNAs are produced from their own transcription units, while animal miRNAs can also be produced from introns of spliced genes (see Section 1.1.2). Just like in animals, plant miRNAs can also appear in clusters. However, this polycistronic organisation is much more frequent in animals than in plants. In plants, instead of requiring two different enzymes to process pri-miRNA and pre-miRNA (Drosha and Dicer respectively), it seems that Dicer-Like 1 (DCL1)

performs both roles in the nucleus, producing in the first step longer and more variable stem-loop pre-miRNAs. Once processed, the miRNA:miRNA* duplex is exported to the cytoplasm by the transporter HASTY. It will then be loaded into the RISC complex, and by near-perfect complementarity with its target, it will induce endonucleolytic cleavage of the mRNA. Translation repression, result of a weak base paring, is yet to be explored in plants. Target sites can be located either in coding exons or in 3' UTRs (Jones-Rhoades *et al.*, 2006; Axtell *et al.*, 2011; Rogers et Chen, 2013; Pasquinelli, 2012).

1.2 Methodological background

1.2.1 miRBase: a reference for microRNA studies

To begin this second section, we present an important miRNA resource that serves as a reference for miRNA research. Most of the miRNA studies use miRBase as a gold standard.

To validate the miRNA predictions, either an experimental method must be performed or a gold standard must be adopted. Given that the first option is much more expensive, the great majority of the authors use miRBase as a reference for the validation. miRBase is a simple and practical database for published miRNA sequence and annotation. It is currently the main source for miRNA annotation with frequent updates. As a consequence, the number of annotated miRNAs grows exponentially (see Figure 1.5), reaching in 2014 (release 21) 28,645 annotated miRNAs. miRBase is available at http://www.mirbase.org/.

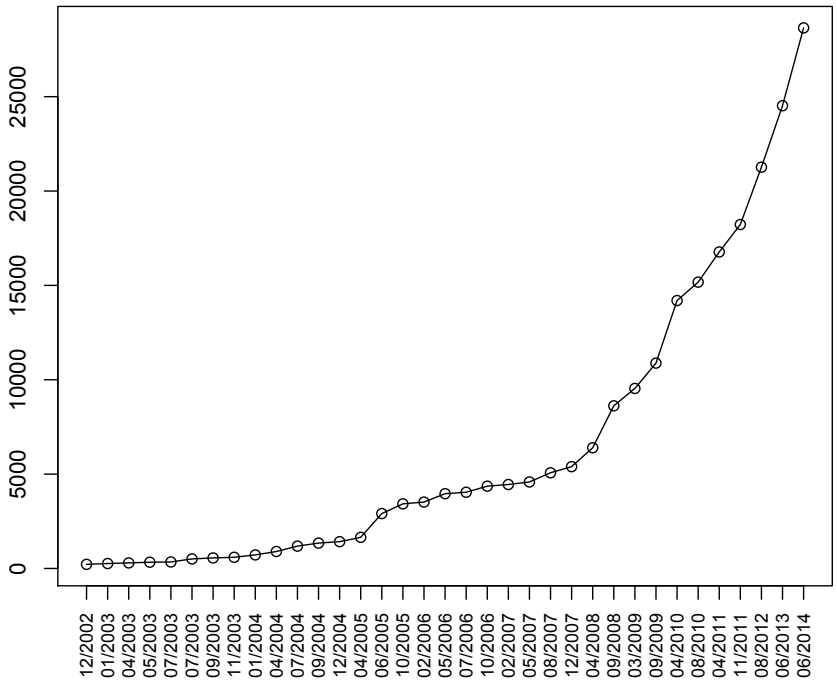


Figure 1.5: *Growth rate of annotated miRNAs in* miRBase *from 2002 (218 miRNAs) to 2014 (28,645 miRNAs).*

A miRNA gene is identified by a prefix (3 or 4 letters) corresponding to the species, followed by the sequence type (mature or precursor miRNA), with a sequential number at the end. For instance, the identifier "hsa-miR-101" corresponds to the human (hsa) mature miRNA sequence (miR) 101. While mature sequences are identified by "miR", precursor miRNA sequences are labelled "mir". Some more informations are aggregated to the name of the mature miRNA. For instance, on the same example as before, miRNAs "hsa-miR-101" and "mmu-miR-101" are similar genes appearing in different species, while "hsa-miR-101a" and "hsa-miR-101b" are similar genes (of the same species) differing at one or two bases. Furthermore, if the same miRNA sequence arises from different pre-miRNA loci, a numbered suffix is added to the end of the miRNA name. For instance, "dme-mir-281-1" and "dme-mir-281-2" are two identical miRNA sequences from *Drosophila melanogaster* derived from different positions of the same pre-miRNA. If two mature miRNAs are excised from both arms of a same pre-miRNA, for instance "hsa-miR-17", they are then called "hsa-miR-17-5p" (from the 5' arm) and "hsa-miR-17-3p" (from the 3' arm). The nomenclature for virus and plant miRNAs is slightly different: (i) for viruses, the genes are named according to the locus where the miRNA originates (for instance, "ebv-mir-BART1" is the miRNA from the virus Epstein Barr deriving from the BART locus); (ii) for plants, the names are in the form "ath-MIR166a", where "ath" is the plant species, "MIR166" is the name of the miRNA, and the suffix composed of one letter stands for the different loci that express the related mature miRNAs (Griffiths-Jones *et al.*, 2006, 2008).

For each entry, several features concerning the pre-miRNA(s) and the respective mature miRNA(s) are provided. We present in what follows a few features that are worth highlighting. When it is available, the alignment of deep sequencing reads is given, showing the regions in the stem-loop that were mostly expressed. Frequently these regions correspond to the miRNA(s) loci. This is important because it proves that the miRNA was indeed transcribed and reached its mature stage. The clustered miRNAs are another feature that shows miRNAs that are close in location to the current entry, that is, < 10kb away from the current miRNA. This is relevant because it allows the user to identify miRNA genes that can be related to each other, since they are very probably being co-expressed. Finally, one functional feature is the list of predicted and validated targets. This information is pertinent since the user can go further in the analysis by knowing which genes are potentially regulated by the given miRNA.

As concerns the organisation of the data in the miRBase ftp, the information provided is separated in the following files:

- miRNA.dat: all miRNA entries in EMBL format.
- hairpin.fa: predicted pre-miRNA sequences in fasta format.
- mature.fa: mature miRNA sequences in fasta format.
- miRNA.dead: removed entries from the database.
- miRNA.diff: differences between the current and the last release.
- miFam.dat: family classification of related hairpin sequences.

In a separated directory (genomes), the gff files with the genome coordinates of the miRNAs and pre-miRNAs are indicated.

1.2.2 Computational methods for microRNA identification

Mendes *et al.* (2009) classify the methods for miRNA prediction in five categories: (i) filter based approaches; (ii) machine learning approaches; (iii) target centred approaches; (iv) mixed approaches; and (v) homology search methods. In filter based approaches, the (pre-)miRNA features are verified in different filtering steps. In machine learning approaches, the methods are trained with the features of known (pre-)miRNAs to be later used for prediction. Target centred approaches are based on a more functional perspective, that is, target sequences are used to determine a potential miRNA as functional or not. Mixed approaches use high throughput experimental data and computational strategies for miRNA prediction. Finally, the homology based approach consists simply in searching for homologous miRNA sequences and/or structures, in the great majority of the cases using alignment methods. It is worth observing that many of the approaches actually use homology in one way or another (e.g. to verify sequence conservation).

In general, the methods implement characteristics originating from the biogenesis process of a miRNA to determine if a sequence is functional. The features are mainly related to the pre-miRNA hairpin, such as free energy, length of the stem-arms and terminal loop, percentage of paired nucleotides within the miRNA duplex; if small RNA sequencing data is used, the pattern of the read stacks is verified to be consistent or not with the one of an expressed miRNA. We thus provide an overview of the current methods for miRNA prediction by describing how the different methods perform this task. The described methods are summarised in Table 1.1 together with the categories to which they mainly belong.

MiRscan was one of the first methods developed for the prediction of miRNAs that is still available. It uses seven features to characterise a miRNA, such as the number of base pairs involving the miRNA candidate, conservation between related species, bulge symmetry between the two species, etc. For each of these features, the authors compute a log-odd score, and then sum them up to obtain an overall score that represents the miRNA. In that time, the authors detected 30 new genes in *Caenorhabditis elegans* (Lim *et al.*, 2003).

Triplet-SVM implements a support vector machine (SVM) classifier trained with the features extracted from every 3 adjacent nucleotides within the hairpin structure, the authors call it triplet element features. Instead of computing the hairpin structure during the execution of the method, it requires it a priori, as an input to the software. The authors trained and tested their method on a human dataset (Xue *et al.*, 2005).

ProMiR is a probabilistic co-learning method based on a paired hidden HMM implemented for the prediction of miRNA with either close or distant homologs. It incorporates both sequence and structural information in a probabilistic framework and also checks for the presence of signals, such as 3' overhang, left by Drosha (Nam *et al.*, 2006). As dataset, the authors used miRNAs from human chromosomes 16, 17, 18 and 19 (Nam *et al.*, 2005).

miRAlign identifies novel miRNAs based on sequence and structure alignment. It differentiates itself from other homology search methods because it is able to identify distant homologs, assuming little conservation of the mature miRNA. Moreover, it considers more properties of the miRNA structural conservation for the prediction of new candidates. The method was applied to *Anopheles gambiae* and 59 new miRNA genes were detected(Wang *et al.*, 2005).

RNAmicro is the implementation of a SVM classifier that evaluates the information of a multiple sequence alignment. To identify the miRNAs the authors use a sliding window approach to extract segments from the genome. Then for each segment, the consensus sequence and structure are computed, and an automaton is used to evaluate the consensus secondary

structure. The alignments that do not respect a few criteria are eliminated and the remaining ones are used to build the feature vector for the SVM classifier. The authors applied their method to the genomes of mammals, urochordates, and nematodes (Hertel et Stadler, 2006).

miRFinder scans whole-genomes for hairpin candidates and, by means of a SVM, evaluates the robustness of these candidates based on 18 parameters, including the minimum free energy (MFE), the frequency of the different kinds of motifs inside the hairpin structure, base pairing of the mature miRNA, etc. The search is performed in a pairwise manner, meaning that the user should provide a closely related genome in addition to the query genome. The authors applied their method to the genome pairs of chicken/human, and to *Drosophila melanogaster*/*Drosophila pseudoobscura* (Huang *et al.*, 2007).

MiPred is the implementation of a random forest method that uses a hybrid feature by incorporating the local contiguous structure-sequence composition, the MFE of the secondary structure, and the P-value of a randomization test. For training and testing their method, the authors used human pre-miRNA data; real pre-miRNAs were obtained from miRBase (at that time, called the miRNA Registry database), and the pseudo pre-miRNAs were the same as those used by the authors of Triplet-SVM (Jiang *et al.*, 2007).

miRank is based on a random walk ranking algorithm to characterise novel miRNAs from genomes with just a few annotated miRNAs. Differently from other machine learning approaches, this model can generalise with just a few samples. The method requires positive miRNA samples for the training step, but no negative samples are necessary. For training and validation, the authors used the genome of *Anopheles gambiae*, which is the vector of malaria (Xu *et al.*, 2008).

SSCprofiler is a probabilistic method based on profile hidden Markov models to predict novel miRNAs. The model is trained over a set of features arising from the sequence, structure, and conservation of known miRNAs. The authors trained the model with human pre-miRNAs and applied it to cancer-associated genomic regions in search of novel miRNAs (Oulas *et al.*, 2009).

HHMMIR is a *de novo* predictor based on a hierarchical hidden Markov model that does not require evolutionary conservation. To predict the miRNAs, the authors set a template for the structure of a typical pre-miRNA hairpin from publicly available data. They then build the HHMM model over this template that is comprised by the following regions: terminal loop, extension (area between the terminal loop and the miRNA duplex), the miRNA duplex itself, and the pri-miRNA extension. The model was trained over a human dataset and was tested on mammals, birds, fishes, worms, flies and plants (Kadri *et al.*, 2009).

MIReNA finds miRNAs, given a genome and a set of known miRNAs, using a filter-based approach with no learning at a genomic scale. It uses five (physical and combinatorial) conditions to define an acceptable pre-miRNA: the miRNA cannot fold itself into a hairpin structure, there is a strong pairing between miRNA and miRNA*, the percentage of unmatched nucleotides within the hairpin, and the MFE and MFE indices are below a certain threshold. Additionally, the authors use a RepeatMasker filter, an EST data filter, and another filter that eliminates other types of RNAs. The option of using deep sequencing data is also available. To compare and validate their method, the authors used, besides the human genome, six other eukaryotic species, including *Caenorhabditis elegans* and *Arabidopsis thaliana* (Mathelier et Carbone, 2010).

CSHMM uses a Context-Sensitive Hidden Markov Model to represent pre-miRNA structures with estimated transition probabilities. Initially, it uses a file with the secondary structure of human pre-miRNAs to set its parameters. It is then trained with the sequences of the same positive human pre-miRNAs and with a set of negative or pseudo pre-miRNAs. Once

the model is set, the authors compute the most likely sequence and its likelihood (Agarwal *et al.*, 2010).

miRD is a webserver which runs an implementation of two independent SVM models based on two different sets of features. To combine these two models, a boosting method was used. In practice, miRD has two applications: (i) to compute the probability of a candidate pre-microRNA to be a real one; and (ii) to extract the probable pre-microRNAs from deep sequencing data. The authors predicted 92 novel pre-miRNA candidates from a small RNA sequencing dataset of the human fetal ovary (Zhang *et al.*, 2011).

MiRPara is an SVM implementation trained with sequences from miRBase. It makes available a script to generate the model according to the miRBase release and to the desired organism(s). The authors used a set of 77 features as input to the SVM classifier; these features were based on characteristics of the miRNA, pre-miRNA and pri-miRNA which are important to the biogenesis of a miRNA (Wu *et al.*, 2011).

miR-BAG is a set of three complementarity approaches (naive Bayes, Best First Decision tree and SVM) which employs different miRNA features such as matrices with specific miRNA guided structural profile and structural triplet density variation profiles with respect to the position of the miRNA. The prediction can be performed at both genomic scale or by using deep sequencing data. The genomes of six species human, mouse, rat, dog, nematode, and fruit fly were used by the authors. (Jha *et al.*, 2012).

miRDeep is a package for the discovery of miRNAs from deep sequencing data. It first eliminates reads which map to many loci in the genome, and optionally it can remove reads mapping to rRNAs, tRNAs, etc. To obtain potential pre-miRNAs, the authors use the information of the mapped reads against the genome. Pre-miRNAs with an unlikely structure are discarded and the core algorithm computes a probabilistic score related to the structure and signature of the pre-miRNA candidate. To validate their method, miRNAs from *Caenorhabditis elegans* and *Homo sapiens* were used (Friedländer *et al.*, 2012).

miRNAFold is an *ab-initio* method that, given a sequence as input, directly searches for pre-miRNA hairpins. The main idea of the algorithm is to find a long stem, which is then taken as an anchor to predict the hairpin structure, attempting to improve the search time. To test their method, the authors used chromosome 19 of the human genome, chromosome 2 of mouse, chromosome 4 of the zebrafish, and chromosome 7q of the sea squirt (Tempel et Tahi, 2012).

RNA secondary structure prediction

Since the great majority of the methods for miRNA prediction use in one way or another the secondary structure of the hairpin pre-miRNA, we focus in this section on this issue.

RNA folding consists in intra-strand base-pairing to produce a secondary structure. As concerns RNA, guanine and cytosine (GC) pair by forming a triple hydrogen bond, adenine and uracil (AU) pair by a double hydrogen bond; additionally, guanine and uracil (GU) can pair by forming a single hydrogen bond. The stability of a given secondary structure depends on: (i) the number of GC versus AU and GU base pairs (the higher the energy bonds, the more stable the structures are, eg, GC is more stabilising than AU); (ii) the number of base pairs in a stem region (longer stems result in more bonds); (iii) the number of bases in a hairpin loop region (the formation of loops with more than 10 or less than 5 bases requires more energy); and (iv) the number of unpaired bases within the structure, either interior loops or bulges (unpaired bases decrease the stability of the structure) (Gesteland et Atkins, 1993; Mathews et Turner, 2006) .

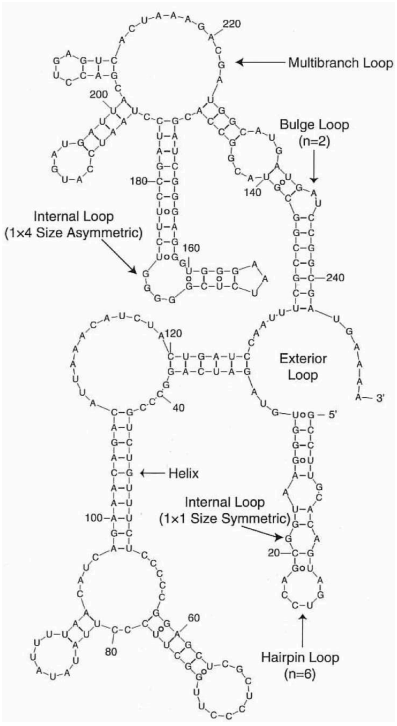
Method	Type	Category	Website*	Author/Paper
MiRscan	Webserver / Standalone on demand	i	MiRscan website	Lim *et al.* (2003)
miRAlign	Webserver	i, v	miRAlign website	Wang *et al.* (2005)
ProMiR	Standalone	ii	ProMiR website	Nam *et al.* (2005)
Triplet-SVM	Standalone	ii	Triplet-SVM website	Xue *et al.* (2005)
RNAmicro	Standalone for Linux	i, ii	RNAmicro website	Hertel et Stadler (2006)
miRFinder	Standalone for Windows	ii	miRFinder website	Huang *et al.* (2007)
MiPred	Webserver	ii	MiPred website	Jiang *et al.* (2007)
miRank	Standalone for Windows / Linux	ii	miRank website	Xu *et al.* (2008)
HHMMIR	Standalone for Linux	ii	HHMMIR website	Kadri *et al.* (2009)
SSCprofiler	Webserver	ii	SSCprofiler website	Oulas *et al.* (2009)
CSHMM	Webserver / Standalone on demand	ii	CSHMM website	Agarwal *et al.* (2010)
MIReNA	Standalone for Linux	i, iv, v	MIReNA website	Mathelier et Carbone (2010)
miRD	Webserver	ii, iv	miRD website	Zhang *et al.* (2011)
MiRPara	Standalone for Linux	ii	MiRPara website	Wu *et al.* (2011)
miR-BAG	Weserver / Standalone	ii	miR-BAG website	Jha *et al.* (2012)
miRDeep	Standalone for Linux	iv	miRDeep website	Friedländer *et al.* (2012)
miRNAFold	Webserver	i	miRNAFold website	Tempel et Tahi (2012)

*The full links to the corresponding websites are presented in the appendix of this thesis.

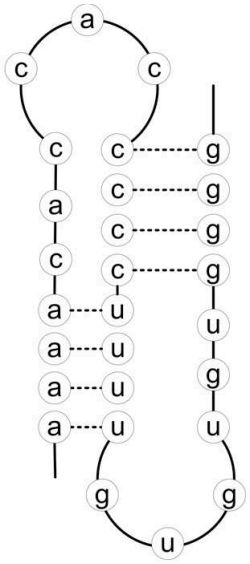
Table 1.1: *Main information on the current methods for the prediction of miRNAs. It includes the type of the prediction software (standalone or webserver), the respective website and reference, and the category in which the method was classified: (i) filter based approaches; (ii) machine learning approaches; (iii) target centred approaches; (iv) mixed approaches; and (v) homology search methods. Such classification scheme was defined by Mendes* et al. *(2009).*

The stability of a secondary structure is quantified as the amount of free energy released or used by forming base pairs. Positive free energy requires work to form a configuration; negative free energies release stored work. Therefore, the more negative the free energy of a structure, the more likely is the formation of that structure because more stored energy is released. Free energy changes of coupled reactions are additive, so one can determine the total free energy of a secondary structure by adding all the component free energies associated to each two consecutive base pairs (units are kilocalories per mole, kcal/mol). This is used to predict the secondary structure of a given sequence. Finding a base pair configuration with the minimum possible free energy is the aim of most secondary structure prediction algorithms (Nelson *et al.*, 1981).

To compute the minimum free energy of a sequence, empirical energy parameters are used. These parameters summarise the free energy change (positive or negative) associated to all possible pairing configurations (Turner et Mathews, 2010). The energy parameter depends on the place (motif) in the structure where the bases are located. An RNA secondary structure can have the following motifs: (i) a helix is the stacking of canonical base pairs (GC, AU and GU); (ii) a loop is a set of non-canonical pairs (that is, unpaired nucleotides): a terminal loop has one appended helix; an internal loop has two appended helices; a bulge loop is similar to an internal one, however, the non-canonical pairs appear just in one strand of the loop; a multibranch loop (junction) is a loop with at least three appended helices; and an exterior loop is a series of adjacent unpaired bases which are not accessible by any base pair; and (iii) a dangling end is the stacking of nucleotides at the end of helices (Turner et Mathews, 2010; Serra *et al.*, 1997). All the previously described motifs are presented in Figure 1.6a. A structure can also have a more complex motif called pseudoknot, which is formed by at least two hairpin structures, in which half of one of the stems is intercalated with the other hairpin, a hairpin being a structure formed by a stem and a terminal loop (see Figure 1.6b).



(a) *The different motifs a RNA secondary structure can have: helix, internal loop, bulge loop, hairpin loop, exterior loop, multi-branch loop. Figure taken from Gesteland et Atkins (1993).*



(b) *A pseudoknot is formed by at least two hairpin structures, in which half of one of the stems is intercalated with the other hairpin. Figure taken from Akutsu (2000).*

Figure 1.6: *Types of RNA secondary structures.*

Algorithms for RNA secondary structure prediction

The problem of predicting an RNA secondary structure is defined as follows. If $S = s_1 s_2 ... s_m$ is an RNA sequence of length m, then the secondary structure of S is defined as a set R of base pairs that satisfies the following criteria: (i) if s_i base pairs with s_j, then $i < j$; (ii) a base pair can only be established if the two bases are, at least, 3 nucleotides apart from one another; (iii) s_i can base pair with one, and only one other base s_j. Pseudoknots are usually not permitted because of the complexity that this leads to. The goal is thus to maximise the number of base pairs within R, or minimise the energy associated to the set of base pairs.

One of the first methods for RNA secondary structure prediction was described by Nussinov *et al.* (1978). The algorithm proposes the maximisation of the number of base pairs to find the best structure. For each position i in the sequence, one should verify all the possible cases: (a) i, j base pair; (b) i is unpaired; (c) j is unpaired; (d) i, j base pair with, respectively, k and $k+1$. The recurrence for this algorithm is presented in Equation 1.1 (Eddy, 2004):

$$E(i,j) = max \begin{cases} E(i+1, j-1) \text{ if i and j base pair} \\ E(i+1, j) \\ E(i, j-1) \\ max_{i<k<j}[E(i,k) + E(k+1,j)] \end{cases} \tag{1.1}$$

Clearly, filling each cell in the DP matrix takes $\mathcal{O}(n)$ time, and since there are $\mathcal{O}(n^2)$ cells, the complexity for the whole procedure is in $\mathcal{O}(n^3)$. However, maximising the number of base pairs is a naïve approach; a more realistic one minimises the free energy of the structure, as proposed for example in Mathews *et al.* (2004). The recurrence for the latter algorithm is presented in Equation 1.2:

$$E(i,j) = min \begin{cases} E(i+1, j) \\ E(i, j-1) \\ min_{i<k<j}[E(i,k) + E(k+1,j)] \\ P(i,j) \text{ if i and j base pair} \end{cases} \tag{1.2}$$

To minimise the free energy, one more table P is required to store the different types of motifs a structure can have, although the complexity in the worst case remains the same, namely in $\mathcal{O}(n^3)$ (Mathews *et al.*, 2004, 2006).

A minimum energy folding algorithm will return only one secondary structure, though there are many candidates for the natural structure. To address this problem, some algorithms (such as Zuker's MFOLD) are designed to provide a set of suboptimal solutions. Inferring what structure is truly representative of the natural structure requires additional information. Phylogenetic information is often used to constrain the search by identifying highly conserved motifs. Some programs allow the user to specify constraints on the secondary structure, by specifying paired, single-stranded, or non-pairable regions, or (Gesteland et Atkins, 1993).

Suboptimal folding One sequence can have several different secondary structures with very similar free energies, which can also be quite close to the minimum. Instead of providing a single optimum structure, a suboptimal approach provides all the partial structures which can be later refined to complete structures. This is done during the traceback step, in which suboptimal structures are chosen. Algorithms implementing this strategy were described in Williams et Tinoco (1986) and Wuchty *et al.* (1999).

Partition function McCaskill (1990) aggregated more quality and robustness for the folding by using a partition function in the prediction of RNA secondary structures. A partition function considers the statistical properties of a system, in this case a secondary structure, in relation to thermodynamics. The Boltzmann factor is defined by $e^{-\Delta G^\circ/RT}$, where ΔG° is the free energy of the structure, R is the constant of gas, and T is the temperature given in kelvin. Then, the probability of a given structure is defined by its Boltzmann factor divided by the partition function Z, which is defined by the sum of all the Boltzmann factors.

Obviously, there are a number of limiting assumptions to existing folding algorithms. These include the kinetics of folding during transcription, the difficulty in predicting pseudoknots, the role of chaperone proteins in folding, and the importance of modified bases (e.g. methylated bases). Some algorithms attempt to incorporate these considerations (e.g., Rivas et Eddy (1999) and Ruan *et al.* (2004) for pseudoknots).

Algorithm for global sequence alignment (Needleman-Wunsch)

We also present here the Needleman-Wunsch (Needleman et Wunsch, 1970) algorithm for global sequence alignment, since our method for the prediction of miRNAs makes use of it to approximate the free energy of a hairpin structure. Global alignments are applied to sequences of similar length, for which the algorithm will try to align every nucleotide in the sequences. The base case and recurrence for the algorithm are presented in Equations 1.3 and 1.4:

$$W(i,0) = W(0,j) = 0, i, j \in 0..n \tag{1.3}$$

$$W(i,j) = max \begin{cases} W(i-1,j-1) + f(s_i, s_j) \\ W(i,j-1) + \gamma \\ W(i-1,j) + \gamma \end{cases} \tag{1.4}$$

where n is the length of the aligned sequences, $f(s_i, s_j)$ is the function returning the score or penalty for, respectively, a match or a mismatch, and γ is the penalty for a gap. Using this recurrence one should take, in the worst case, $\mathcal{O}(n^2)$ time to align two sequences of length n (Needleman et Wunsch, 1970). Algorithm 1 contains this forward-filling step of the algorithm for two sequences A and B of length m and n, respectively.

Algorithme 1: *Forward step of Needleman-Wunsch's algorithm.*

```
Data : Two sequences A and B
Result : Dynamic matrix fulfilled
for i ← 0 to m do
    F(i,0) ← 0
for i ← 0 to n do
    F(0,j) ← 0
for i ← 1 to m do
    for j ← 1 to n do
        match ← F(i-1,j-1) + f(A_i, B_j)
        delete ← F(i-1, j) + γ
        insert ← F(i, j-1) + γ
        F(i,j) ← max(match, insert, delete)
```

Once the dynamic programming matrix F is filled up, the next task consists in recovering the alignment by backtracking along the matrix. The recovery is performed by means of a

recursion starting in cell $F(m, n)$ and ending when the left or the top part of the matrix is reached, as shown in Algorithm 2.

Algorithme 2: *Backtracking step of Needleman-Wunsch's algorithm.*

Data : Dynamic programming matrix F
Result : Alignment and its score

```
AlignmentA ← ""
AlignmentB ← ""
i ← m
j ← n
while i > 0 or j > 0 do
    if i > 0 and j > 0 and F(i, j) == F(i − 1, j − 1) + f(A_i, B_j) then
        AlignmentA ← A_i + AlignmentA
        AlignmentB ← B_j + AlignmentB
        i ← i - 1
        j ← j - 1
    else if i > 0 and F(i, j) == F(i − 1, j) + γ then
        AlignmentA ← A_i + AlignmentA
        AlignmentB ← "-" + AlignmentB
        i ← i - 1
    else if j > 0 and F(i, j) == F(i, j − 1) + γ then
        AlignmentA ← "-" + AlignmentA
        AlignmentB ← B_j + AlignmentB
        j ← j - 1
```

Nearest neighbour energy model

To model the free energy change for the folding of RNAs, one can use the thermodynamic Nearest-Neighbour (NN) energy associated to each type of motif in the structure. By summing up the energy increment of each motif, it is possible to obtain a reasonable approximation of the free energy change for folding an RNA or, in other words, to obtain a measure of the stability of an RNA molecule (Mathews *et al.*, 2004, 2006).

The motifs forming an RNA structure are determined by the base-pairs AU, GC and GU. The arrangement of these base pairs can shape into the different types of motifs, such as helices, bulge loops, and internal loops. The stabilising motifs are: the Watson-Crick helix represented by the stacking of at least two base pairs; and a dangling end which is a single base at the end of a helix. The destabilising motifs are of three types: the hairpin loop which is composed of non-canonical base pairs closed by one canonical base pair; the bulge loop which is an arrangement of unpaired nucleotides in one of the strands of a helix; and finally, the internal loop which includes unpaired nucleotides in both strands of a helix. There exist three more types of motifs which are the multi-branch loop, the exterior loop, and pseudoknots. However, they are not present in a pre-miRNA stem-loop structure, and will therefore not be explored in detail here (Mathews *et al.*, 2006; Turner et Mathews, 2010).

As mentioned before, to compute the free energy of an RNA structure, it is necessary to sum the increments according to the type of the motif. The equations presented hereafter describe how to compute the free energy associated to each kind of motif.

The energy of a dangling end depends only on the base-pair before the dangling nucleotide

and on this latter. For all the other types of motifs, the equations are given below. The energy of an internal loop is computed by means of Equation 1.5:

$$\Delta G_{Internal} = \Delta G_i(n) + (\Delta G_a * |n_1 - n_2|) + \Delta G_{m1} + \Delta G_{m2} + (\Delta G_{ru} * \lambda) \tag{1.5}$$

where $\Delta G_i(n)$ is the initiation energy to form an internal loop of $n \leq 30$ unpaired nucleotides; $\Delta G_a = 0.6$ is the asymmetry penalty multiplied by the absolute value of the difference between the number of unpaired nucleotides in each strand; ΔG_{m1} and ΔG_{m2} are the energy of the first and the last mismatches in the internal loop; and $\Delta G_{ru} = 0.7$ is the penalty for an RU closure, where $R = \{A, G\}$ and λ is the lambda function which returns 0 or 1, corresponding, respectively, to the presence or absence of, in this case, the RU closure.

For the bulge loops, one should use Equation 1.6:

$$\Delta G_{Bulge(n=1)} = \Delta G_i(1) + \Delta G_C + \Delta G_s - \text{ RT ln(t)} + (\Delta G_{ru} * \lambda)$$
$$\Delta G_{Bulge(n>1)} = \Delta G_i(n) \tag{1.6}$$

where $\Delta G_i(n)$ is the energy required to form a bulge with $n \leq 30$ unpaired nucleotides; if the bulge is comprised of the nucleotide C only, and there is at least one more C not in the bulge (meaning, it is paired with a G), one should add the C bulge penalty $\Delta G_C = -0.9$ kcal/mol; ΔG_s is the base pair stacking around the bulge; t is the number of possible loop conformations with identical sequence; $R = 8.3144621$ J/mol K is the gas constant and $T = 310.15$ K is the temperature in kelvin. Notice that for bulges and helices, $\Delta G_{ru} = 0.45$ and is referred to as the penalty for a RU end (and not closure as for internal loops). For bulges and internal loops larger than 30 nucleotides ($n > 30$), Equation 1.7 should be applied instead:

$$\Delta G_{n>30} = \Delta G_i(30) + 1.75 \times RT \times ln(n/30) \tag{1.7}$$

Finally, for a helix, one should apply Equation 1.8:

$$\Delta G_{helix} = \sum \Delta G_{stck} + \Delta G_{sym} + (\Delta G_{ru} * \lambda) \tag{1.8}$$

where ΔG_{stck} is the stacking energy of each two consecutive base pairs; ΔG_{sym} is the symmetry correction for self complementarity duplexes; and $\Delta G_{ru} = 0.45$ is, as mentioned before, the RU end penalty.

All the thermodynamic NN energies used in this work, as well as the equations described above, were obtained in the Nearest Neighbor Database (NNDB) (Turner et Mathews, 2010; Zuker *et al.*, 1999).

1.2.3 Experimental methods for microRNA detection and quantification

As mentioned in Section 1.1.1, a sequence must fulfill three criteria to characterise a miRNA: (i) the mature miRNA must be expressed as a transcript of ~22nt (the expressed transcript is detected by means of experimental techniques such as northern blot, small RNA sequencing, etc.); (ii) the mature miRNA must derive from a precursor miRNA with a typical hairpin structure containing small bulges and internal loops; and (iii) the pre-miRNA should be processed by Dicer, as an increased accumulation of the precursor is noticed when Dicer is absent. The experimental methods for discovering miRNAs are based on these definitions

with variations among them (Berezikov *et al.*, 2006). A brief description of these methods, such as PCR based methods, microarray, northern blot, and RNA sequencing is given in what follows. It is important to observe that the first three methods—PCR, microarray, and northern blot—require the miRNA sequence, that is, these methods validate specific sequences known a priori while for RNA sequencing it is not necessary to know the miRNA sequence (Chaudhuri et Chatterjee, 2007).

PCR based methods

The reverse transcription polymerase chain reaction (RT-PCR) protocol is based on the reverse transcription of the small RNA to cDNA. The reverse transcription is generated either by the addition of a poly A queue or by means of a stem-loop primer. The reverse transcript cDNA is then submitted to a PCR for amplification and quantification. The accumulation of the reaction product can be monitored in real time (it is then called real time RT-PCR).

The method is widely used because of its ease of incorporation; consequently, it is a very well established method with a good sensitivity and specificity. The disadvantages are the medium-throughput concerning the number of samples processed per day, and the inability to detect novel miRNAs (Pritchard *et al.*, 2012; Aldridge et Hadfield, 2012). The available assay/platforms for RT-PCR are: TaqMan by ABI, miRCURY LNA qPCR by Exiqon, Biomark HD system by Fluidigm, SmartChip human microRNA by Wafergen, and miScript miRNA PCR array by SABiosciences/ Qiagen.

MicroRNA microarray

A microarray is a chip composed of several microscopic spots. Each spot is filled with DNA/RNA molecules for the measurement of their expression level. These molecules are specific oligonucleotide sequences (e.g., miRNAs, genes, etc.) known as probes. These probes will then be submitted to hybridisation with specific cDNA/cRNA targets under specific conditions. The occurrence of hybridisation between probe and target will be detected by a label that is linked to the target and will emit a fluorescent light. The different light spectra will quantify the level of expression (Yin *et al.*, 2008; Pritchard *et al.*, 2012).

The advantage of microarray experiments is that it is not an expensive method, while allowing for the parallel profiling of a large number of molecules. On the other hand, microarray technology has low specificity if the miRNAs are similar, and the absolute quantification of miRNA expression is not easily performed, while it is better suited for detecting the relative abundance of specific miRNAs in 2 different states (Pritchard *et al.*, 2012; Liu *et al.*, 2008).

The available platforms for microarrays are: Geniom Biochip miRNA by CBC, GeneChip miRNA array by Affymetrix, GenoExplorer by Genosensor, MicroRNA microarray by Agilent, miRCURY LNA microRNA array by Exiqon, NCode miRNA array by Invitrogen, nCounter (not a microarray but hybridization-based) by Nanostring, OneArray by Phalanx Biotech, and Sentrix array matrix and BeadChips by Illumina.

Northern blot

Northern blot is a technique to identify a specific RNA from a bunch of RNAs. The total RNA is denatured and after the addition of an agent, the RNA remains unfolded in its linear conformation. The collection of RNAs is then sorted by size by means of an electrophoresis gel and moved to a nitrocellulose filter in which the RNAs are attached. A labelled probe

is added to the filter, and it is finally submitted to autoradiography that will quantify the expression level of the given RNA (Lodish *et al.*, 2000).

The majority of the methods are specialised in the detection of smaller mature miRNAs. The advantage of northern blot is that it can detect a wide range of sizes from primary miRNAs to mature ones. However, the approach is low throughput and has low sensitivity, besides being time consuming and requiring a large amount of total RNA (Pritchard *et al.*, 2012; Liu *et al.*, 2008).

RNA sequencing

Since we produced and analysed small RNA sequencing data during this thesis, we give more details about this experimental technique, focusing on the Ilumina platform. Details about the sRNAseq data produced and analysed are provided in Chapter 3.

RNA sequencing makes use of NGS technology to verify the presence of and to quantify the RNAs from a genome in a specific condition. For miRNA sequencing, the input library is enriched for this kind of molecule. The advantages of using NGS for miRNA profiling are identification of known and novel miRNAs and precision in identifying very similar miRNAs, such as isomiRs of different length and miRNAs differing by a single nucleotide. On the other hand, small RNA sequencing can produce several putative small RNAs of novel sequence, and they are not necessarily bona fide miRNAs (Pritchard *et al.*, 2012).

The first step of the experiment consists in the isolation of the total RNA from a sample by means of a specific reagent, which depends on the kit used. The total RNA is then filtered by size using a polyacrylamide gel which is submitted to a process of electrophoresis. In the case of miRNAs, the selected size range is from 17nt to 25nt. Adapters are thus ligated on both 5' and 3' ends of the RNA sequence to act as the binding sites for the primers used during the next step (RT-PCR). Since sequencing technology is designed to sequence only DNA, it is necessary to convert the RNA into cDNA by means of a reverse transcriptase. The total amount of cDNA is amplified with a PCR, and it is finally submitted for sequencing. Figure 1.7 presents a summary of the above-mentioned process.

For sequencing, different kinds of platforms exist, we mention the most used ones: sequencing-by-synthesis on the Illumina[1] platform, pyrosequencing on the 454 Life Sciences[2] platform, and ABI Solid Sequencing[3] platform. Since for this thesis, the miRNAs of the *Acyrthosiphon pisum* (pea aphid) were sequenced on an Illumina platform, we provide more details about it.

The cDNA library is given as an input to the Illumina sequencer. The first step occurs in a device called Cluster Station over a 8-channel flow cell where amplification of the reads occurs. Oligos, that are complementary to the adaptors ligated to the cDNAs, are attached to this flow cell. The cDNA fragments will thus bind to these oligos and the DNA polymerase will produce approximately one million copies of the original fragments, which is a sufficient amount to generate the required signal intensity of the incorporated bases. After that, the single cDNA fragments are replaced by clusters of fragments. The four nucleotides, enriched with a unique fluorescent label, are then added to the channels of the flow cell, together with the DNA polymerase, to be incorporated into the clustered fragments. Each base incorporation is followed by an imaging step that scans the emitted light associated to each base. Each base incorporation corresponds to a cycle; consequently, the number of cycles is equivalent to the length of the fragments.

[1] http://www.illumina.com/
[2] http://www.454.com/
[3] http://solid.appliedbiosystems.com/

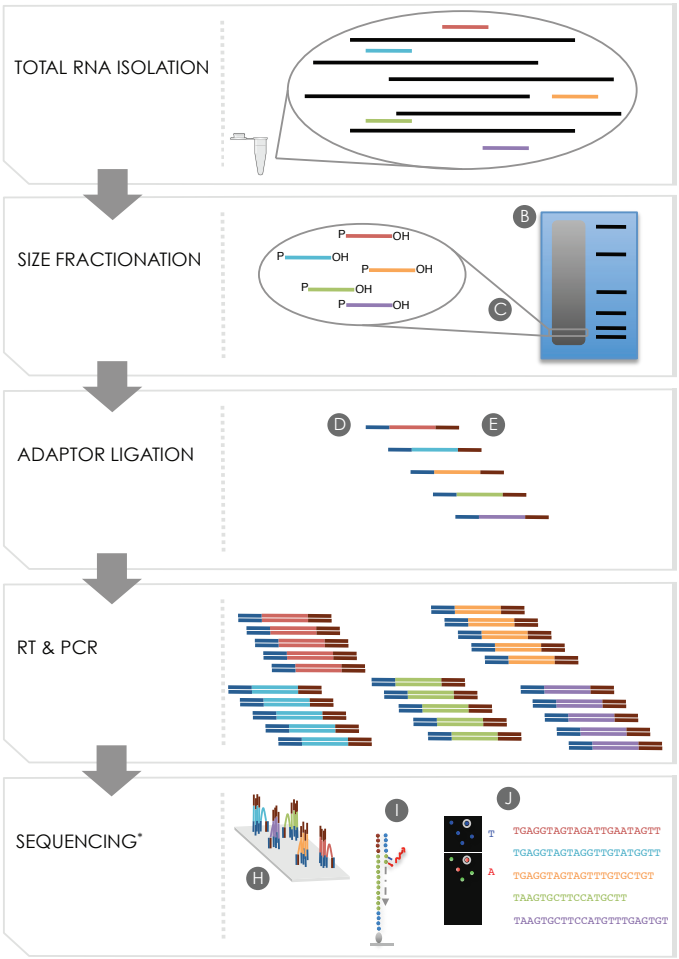


Figure 1.7: *Preparation of a miRNA Illumina sequencing library. The steps are described in order as follows: (i) isolation of total RNA from the sample; (ii) size fractionating of total RNA using denaturing PAGE, and selection of small RNA by size (17-25 nt); (iii) 3' and 5' adapter ligation; (iv) reverse transcription of RNA sequences, and PCR amplification; (v) flow cell attachment, bridge amplification, annealing of sequencing primers and base extension, base calling till the number of cycles is finished. Figure taken from Wikipedia.*

As mentioned before, a plethora of putative novel miRNAs is produced, and they are not necessarily bona fide miRNAs. In order to retain the real miRNAs, further criteria must be applied for the annotation of small RNA sequences as a miRNA: length of ∼22nt, hairpin structure of the corresponding precursor miRNA, sequenced reads aligning to both arms, -3p and -5p of the precursor, and, when a close species is available, conservation across species. All the details concerning the sRNAseq data analysis of the pea aphid is provided in Chapter 3.

1.2.4 Computational methods for target prediction

As briefly mentioned before in Section 1.1.3, the exact mechanism used by miRNAs to regulate target gene expression is still uncertain. Cases have been reported of target mRNA cleavage, translation repression, and also of activation of gene expression. There are even evidences, in both plants and animals, that miRNAs can reduce protein (and not mRNA) levels. More specific mechanisms are not clear, the decrease of gene expression can be associated to the prevention of translation initiation or elongation, and also to the proteolysis of peptides (Liu *et al.*, 2014; Pasquinelli, 2012).

Although the mechanism of miRNA target regulation is not clear, the problem is modelled focusing either on characteristics that are specific of some cases (such as cleavage of mRNA), or on features that are in principle common to all the cases.

Most of the computational methods incorporate features related to the base pairing interaction between miRNA and target. These include the presence of perfect complementarity of the *seed* region located at the 5' end of the miRNA (nucleotides 2-7), and 3' UTR for the target mRNA. Accessibility of the 3' UTR in its secondary structure is also verified and it is associated to AU content in the flanking regions. Target conservation is used to eliminate false positives. Even if these general rules have been successful in many predictions, a substantial part of the methods diverge in their results, with levels of false predictions that are not easy to evaluate. One of the causes of the previous mentioned problem is the lack of experimentally validated miRNA-mRNA interactions (Witkos *et al.*, 2011; Pasquinelli, 2012).

As mentioned before, methods for target prediction basically employ two features: (i) base pairing between miRNA and target, specially considering the seed region; and (ii) target conservation across related species. The currently available methods are TargetScanS, miRanda, DIANA-microT, PicTar, and RNAHybrid.

TargetScanS requires a seed region of length 6nt from positions 2 to 7 in relation to the miRNA. It also demanded target site conservation across all the five genomes the authors studied: human, mouse, rat, dog, and chicken. The authors observed the presence of conserved adenosines flanking the seed region in the target mRNA, suggesting that these nucleotides can play a decisive role in the recognition of miRNA targets (Lewis *et al.*, 2005).

miRanda looks for sequence match between miRNA and target, permitting GU wobble pairs and moderate insertions and deletions, while giving a stronger weight to complementarity at the 5' end of the miRNA. The free energies of the duplexes are then computed with the Vienna package (Lorenz *et al.*, 2011). Conservation is thus verified according to three factors: (i) a miRNA should match orthologous UTRs in the three species the authors studied (i.e., *Drosophila melanogaster*, *Drosophila pseudoobscura* and *Anopheles gambiae*); (ii) the target sequences in all three species should respect a threshold identity with each other; and (iii) the positions of both target sites are equivalent according to a cross-species UTR alignment (Enright *et al.*, 2004).

DIANA-microT requires the interactions to meet two criteria. The first is high-affinity

measured on the basis of free energy. The second criterion considers the proteins associated to the interaction between miRNA and target; the authors verify it by analysing the position and sizes of the loops (bulges and internal loops) within the miRNA:target duplex that are imposed by the associated proteins (Kiriakidou *et al.*, 2004).

PicTar defines a seed as a perfect base pair of length 7nt starting at position 1 or 2 at the miRNA 5' end. Insertions or mutations are allowed only if the free energy of the duplex does not increase and does not form GU wobbles. A combined score is computed for the mRNA target; it is comprised of the maximum likelihood of the given target to be regulated by a set of miRNAs, plus a few other features observed in the results of experimental interactions. If a miRNA seed aligns to overlapping positions of the UTR sequences across the different species, conservation is considered to be verified (Krek *et al.*, 2005).

RNAHybrid is an adaptation of the classical RNA secondary structure prediction described by Zuker et Stiegler (1981). Instead of folding a single sequence in the energetically most favourable conformation, it determines the most favourable hybridisation site between two sequences. The presence of a seed is also verified by this method, however the properties (such as length, position) are set by the user (Krüger et Rehmsmeier, 2006).

PITA predicts target sites by verifying the presence of seed regions (allowing for GU wobbles and mismatches). It next uses target accessibility, the core of their algorithm, a concept that is strictly related to the secondary structure of the target transcript. The hypothesis is based on the fact that the mRNA structure is an important factor in the recognition of the target, by thermodynamically favouring or not the interaction. The free energy gained from the formation of the miRNA-target duplex, and the energetic cost of unfolding the target to make it accessible to the miRNA are computed, and the tradeoff between these two measures is assessed to classify an interaction as functional or not (Kertesz *et al.*, 2007).

1.2.5 Experimental methods for microRNA target identification

This section is included for the sake of completeness, since we did not performed any wet experiments involving target identification. It is then a concise section providing only an overview of the methods available.

Interactions between miRNA and target are often validated by fusing the target site to a reporter gene and verifying, in the presence or absence of the miRNA, if regulation occurs. In this case, the original cellular context is lost. Nevertheless, a recent technology called cross-linking immunoprecipitation sequencing (CLIP-seq) allows the identification of endogenous target sites by means of the sequencing of those targets that co-immunoprecipitate with RISC components (Pasquinelli, 2012).

The use of microarray experiments, after miRNA overexpression or knockdown, consists in another method for identifying genes regulated by miRNAs. Since miRNAs reduce the levels of gene transcripts, measuring the expression of a given mRNA after an abnormal miRNA expression provides an effective manner to verify functional interactions (Thomas *et al.*, 2010).

Using a stable isotope labeling with amino acids (SILAC) in cell culture followed by mass spectrometry based proteomics, one can evaluate the effect of down or overexpression of miRNAs on global protein expression (Thomas *et al.*, 2010).

There are mainly two types of regulation performed by miRNAs: translation repression and mRNA cleavage. Parallel analysis of RNA ends (PARE), also known as degradome sequencing, detects interactions originating from the second case. It detects the products of mRNA cleavage. An RNA adaptor is ligated on the mRNA 3' fragments resulted from the Argonaute-mediated cleavage. These fragments are submitted to RT-PCR for enrichment,

followed by deep sequencing (German *et al.*, 2009).

Other methods exist, which are mainly improvements of the mentioned ones; we do not detail them here as this would be beyond the scope of this thesis. For a complete survey of the current methods, see (Thomson *et al.*, 2011).

Chapter 2

MIRINHO: Efficient precursor miRNA predictor

Contents

2.1 Introduction 25
2.2 Material and methods 26
2.2.1 Algorithm 26
2.2.2 Dataset 27
2.2.3 Compared methods 29
2.2.4 Measuring sensitivity and precision 30
2.3 Results and discussion 30
2.3.1 Regression analysis of the free energies 30
2.3.2 Time efficiency 30
2.3.3 miRNA hairpin structure prediction in sRNA-seq of plant 33
2.3.4 Sensitivity and precision 34
2.4 Conclusion 40

This chapter is strongly based on the paper Higashi *et al.* (ress). Here, we address the problem of prediction of miRNAs, focusing on three main issues: (i) efficiency of the algorithm for free energy due to the use of a quadratic algorithm (instead of a cubic one as used so far by other methods); (ii) high quality hairpin secondary structure when sRNA-seq data is available; and (iii) dependence on as little information as it is possible to compute the free energies. These items were defined in details and implemented in a software called MIRINHO, which is available at `http://mirinho.gforge.inria.fr`. Besides the better time complexity of the algorithm itself, a speed-up was implemented during the Master internship of a bioinformatics student, Cyril Fournier, whom I co-advised together with Marie-France Sagot.

2.1 Introduction

Given the ubiquity of miRNAs and their functional importance, it became crucial to develop methods for the prediction and analysis of miRNAs. As a consequence a plethora of such methods have been developed (as shown in Chapter 1). Despite all the effort put in developing them, there remain a number of issues that need to be addressed: (i) to predict the characteristic hairpin structure of a pre-miRNA, the vast majority of the existing softwares

rely on a folding algorithm of cubic time complexity which is suitable when the input is small enough, but it can become impracticable when the size of the input increases; (ii) for longer pre-miRNAs (such as in plant), such folding methods moreover can produce hairpin structures different from the ones provided in MIRBASE (Kozomara et Griffiths-Jones, 2011), which uses sRNA-seq data to do so; (iii) together with folding, most methods then rely on further information that must be learned from previously validated miRNAs of closely related genomes (at a minimum within the same clade, plant or animal) for the final prediction of new miRNAs in order either to set the parameters of the model or to restrict the search to a limited space.

We therefore developed MIRINHO to address all three issues. The search for pre-miRNAs is concentrated on regions with the same length as the two stem-arms separated by the length of the terminal loop. The direct application to *sRNA-seq* data guarantees a better quality in the prediction of the pre-miRNA structures. A *quadratic* time complexity algorithm improves the practical efficiency of the free energy computation. As neither of the two attributes used (length of stem-arm and terminal loop) are species-specific within the animal or the plant kingdom (they differ only between these two kingdoms), the method can easily be applied for predicting pre-miRNAs in either clade.

Importantly, while the method we provide is thus much simpler, faster, and general to use, we also show for tested examples that it has a sensitivity and precision as good as other methods, in some cases even better. Moreover, we show that the secondary structures predicted by MIRINHO are much closer to the ones available in MIRBASE than for the other compared methods.

2.2 Material and methods

2.2.1 Algorithm

Screening the genome to identify potential pre-miRNAs A pre-treatment of the data is performed in order to identify all the inputs for our algorithm. This was done without loss of information at this step, meaning that all sequences that are potential candidates for being a pre-miRNA were selected.

We set a sliding window w of length $\ell = 25$, that is the mean length of a miRNA sequence. For each stem-arm represented by $st_1 = [w_i, ..., w_{i+\ell-1}]$, we looked for its putative stem-arm pair $st_2 = [w_{i+\ell+n-1}, ..., w_{i+2*\ell+n-1}]$, n nucleotides away from the first one, with n between 5 and 20. We thus have that w represents the potential stem-arms and n the length of the terminal loop, as shown in Figure 2.1. Each such pair (w, n) will be an input for the alignment algorithm described next.

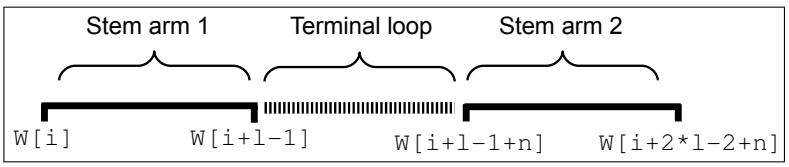


Figure 2.1: *Stem-loop coordinates and representation. The black lines represent the stem-arms, and the stripped line represents the terminal loop.*

Assessing the potential pre-miRNAs Each pair of putative stem-arms screened in the previous step was given to an alignment algorithm in order to evaluate whether it is a stable stem-loop structure. For that, we implemented the Needleman & Wunsch global alignment algorithm (Needleman et Wunsch, 1970) (Section 1.2.2) with a scoring strategy based on the Nearest Neighbour energy model (Section 19). Instead of using the sum of the integer penalties for gaps, matches and mismatches, the alignment is assessed according to the free energy related to each two consecutive nucleotides in the alignment.

We define the alphabet $\Sigma = \{M_{xy}, S_{xy}, I_{xy}, D_{xy}\}$, where the symbols correspond, respectively, to *Match*, *Mismatch*, *Insertion* and *Deletion*, and $x, y \in \{A, U, C, G, -\}$. The definition of an alignment of two putative stem-arms, st_1 and st_2, is a vector comprised by the symbols in Σ, such that $align(st_1, st_2) = v$ and $v = [v_i, v_{i+1}, ..., v_n]$, where $v_i \in \Sigma$.

To determine the stability of a pre-miRNA stem-loop, we go through vector v and sum up the free energy of each pair (v_i, v_{i+1}) according to the type t of the motif it is inserted in. For that, we use Equation 2.1 below to compute the energy of each motif in the structure:

$$\epsilon(t) = k(t, m) + \sum_{i}^{m-1} e(v_i, v_{i+1}) \tag{2.1}$$

where t is the motif type that can be an internal loop, a bulge loop or a helix. The value $k(t, m)$ accounts for the penalties associated to the motif t, which appears m times in the structure. For example, for a motif of type $t = helix$, one should consider the symmetry correction ΔG_{sym} for self-complementary duplexes (see Equation 1.8). Finally, the function e returns the energy associated to the pair (v_i, v_{i+1}).

We then sum all the energies related to the different types of motifs to obtain the final free energy E of the structure using Equation 2.2:

$$E = \sum \epsilon(t) \tag{2.2}$$

where t is again the different types of motifs a given structure can have.

Alignment speed-up Considering that a stable hairpin structure should not contain very large bulges neither internal loops, an ideal alignment should be concentrated around the main diagonal of the dynamic programming (DP) matrix. Instead of using the whole matrix, the user can therefore constrain the alignment to this diagonal and prune parts of the bottom-left and top-right corners of the matrix, thus saving time in the computation of the free energies with a small loss, as shown in Figure 2.2.

A parameter *dw* (diagonal width) is established that depends on the length of the aligned sequences and on a compromise between sensitivity and precision in relation to the version that uses the full matrix (see the Section 2.3.2 to determine how to set an appropriate value for this parameter).

2.2.2 Dataset

To set an appropriate energy threshold for MIRINHO, we chose chromosomes from six different metazoan genomes with the respective MIRBASE miRNA annotations.

- Chromosome 25 from *Bos taurus* (27 miRNAs)

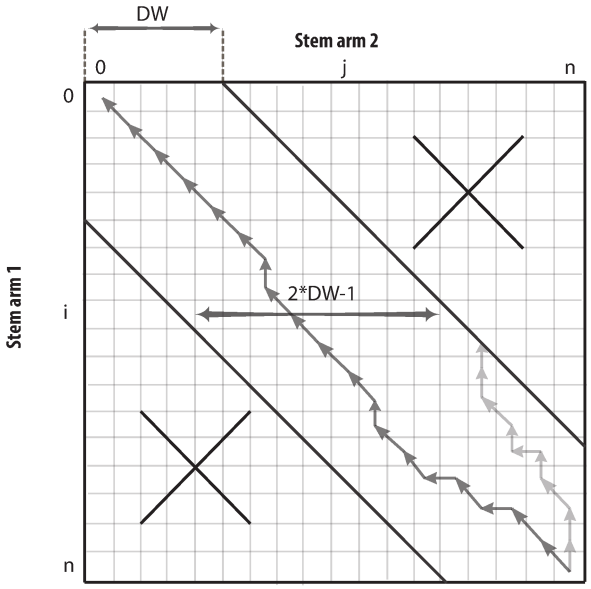


Figure 2.2: *Pruned dynamic programming matrix according to the parameter dw (diagonal width). The alignment is concentrated only in the diagonal portion of the matrix. Alignments touching the border of the diagonal portion are disregarded.*

- Chromosome I from *Caenorhabditis briggsae* (14 miRNAs)
- Chromosome 2R from *Drosophila simulans* (36 miRNAs)
- Chromosome 25 from *Gallus gallus* (6 miRNAs)
- Chromosome 22 from *Gorilla gorilla* (8 miRNAs)
- Chromosome 19 from *Mus musculus* (60 miRNAs)

To compare our method to other pre-miRNA predictors, we applied it to:

- the prediction of plant pre-miRNAs: we used the sequence of chromosome 4 of *Arabidopsis thaliana* (version 2.0) as well as 340,114 reads of high-throughput small RNA sequencing data from *Arabidopsis thaliana* (GEO accession number GPL3968).
- three animal chromosomes for which the miRNAs are well characterised:
  - Chromosome III of *Caenorhabditis elegans* (44 miRNAs)
  - Chromosome 2R of *Drosophila melanogaster* (92 miRNAs)
  - Chromosome 19 of *Homo sapiens* (234 miRNAs)

In the latter case, as two of the softwares to which MIRINHO was compared are too slow, the predictions were performed on smaller sets of sequences obtained in the following way: for each of the three chromosomes (III in *Caenorhabditis elegans*, 2R in *Drosophila melanogaster*, and 19 in *Homo sapiens*), 10 miRNAs were randomly chosen together with 100nt both up and downstream. Each fragment (miRNA+extension) of length n was flanked by sequences of the same length, which were generated based on the nucleotide distribution of the given chromosome. In the end, we obtained three different sequences of ∼4265nt that were given as input to CSHMM, MIReNA, MIRINHO, and miRPara (mentioned in the Section 2.2.3).

For computing sensitivity and precision in a genomic scale, we used the genomes of three insects that are of special interest for our group:

- *Acyrthosiphon pisum* genome assembly version 2 (123 miRNAs)
- *Culex quinquefasciatus* genome assembly version 1 (120 miRNAs)
- *Heliconius melpomene* genome assembly version 1.1 (101 miRNAs)

All the chromosomes, genomes, and sRNA-seq data were obtained from the NCBI. The annotations concerning the known (pre-)miRNAs were obtained from miRBase (release 20) (Kozomara et Griffiths-Jones, 2011).

2.2.3 Compared methods

To compare the accuracy of our method with other predictors, we first made an extensive search of the available ones (see Table 1.1 in Section 1.2.2). We put aside the predictors that required other kinds of input files than just the fasta sequence and/or sRNA-seq data, as well as those incompatible with the Unix system. Web-servers were also disregarded because there always is a restriction to the length of the sequence that may be input. The methods that remained were CSHMM, MIReNA, and miRPara. Notice that as one of our main contributions is the efficiency in the prediction of pre-miRNAs in relation to other methods that use cubic complexity algorithms, it was natural to compare Mirinho to methods that adopt this kind of algorithm. However, we also included in the comparison a method such as CSHMM which does not use the same cubic algorithm for the prediction of miRNAs.

Since the set of input parameters differs for each method, it is not a trivial task to set them accordingly to the data, and at the same time be fair in the comparison. We then applied the methods with default parameters. However, we adapted one aspect that was common to all the methods: the set of known (pre-)miRNAs. All the methods were trained, when required, with animal (pre-)miRNA sequences. The description of each method, and how they were trained and performed is given below.

To set the initial parameters for CSHMM, we used the secondary structures of the kingdom metazoan that are available in miRBase release 20. To generate the likelihood score, the same metazoan hairpin sequences were given as the positive training set, and as the negative instances the sequences used by the authors were employed.

MIReNA provides different starting points for the prediction based on the different kinds of input files. We then chose the one that allows genomic inputs (-M option), and the set of known mature miRNAs required was from the same metazoan kingdom, taken from miRBase release 20.

miRPara makes available a script to generate the model according to the miRBase release and to the desired organism(s)/clade. In our case, we chose the model trained with metazoan pre-miRNAs of the latest release 20.

To analyse the quality of the predicted structures, we used RNAfold (Hofacker *et al.*, 1994) and miRNAFold (Tempel et Tahi, 2012). The first is a classical method for predicting an RNA secondary structure through energy minimisation. If one has access to GPU facilities, we may cite two papers that implemented algorithms for such a kind of technology: Rizk et Lavenier (2009) and Steffen *et al.* (2010). The second is a method for predicting a hairpin structure that takes into account specific criteria (such as length of the stem, percentage of nucleotides, size of terminal loops) related to known hairpins from miRBase, and verifies if these are present in the query structure. miRNAFold is moreover, as far as we know, the

only other method that has quadratic complexity for predicting a miRNA structure. For more details on how each of the methods were used, see Section 2.3.3.

2.2.4 Measuring sensitivity and precision

To evaluate the performance of each method in the prediction of pre-miRNAs, we used as measures sensitivity and precision (besides the stem-loop structures available in MIRBASE). The first is the proportion of true pre-miRNAs that are correctly predicted while the second is the fraction of predicted pre-miRNA candidates that are real pre-miRNAs:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.3}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.4}$$

where TP stands for True Positive, FP for False positive, and FN for False Negative.

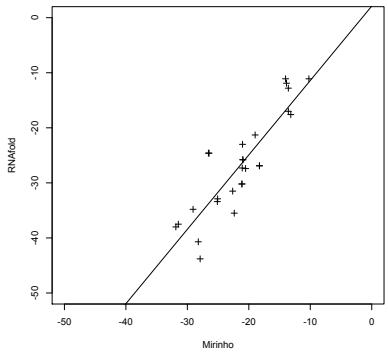
To compute the number of true pre-miRNAs predicted by each method, we do the following. For each of the six species, there is a control set $C = \{c_j, c_{j+1}, ...\}$ of the miRNAs that are considered to be true miRNAs following according to MIRBASE, where $j \in 1..n$ and n is the number of true miRNAs for a given species. Ideally, to compute the number of TPs, one should compare a predicted miRNA pm with the control miRNAs cm_j that has at least one position in common with it. However, not all the softwares provide the exact coordinate of the predicted miRNA. Instead, all of them give the coordinates of the respective predicted pre-miRNA ppm. In order to compute the number of true miRNAs for a given species, we therefore verified, for each ppm, whether it fully covered a control cm_j. If that was the case, we accounted for one TP. If the same ppm covered more than one control miRNA, we considered just the one with the best prediction score according to each method.

2.3 Results and discussion

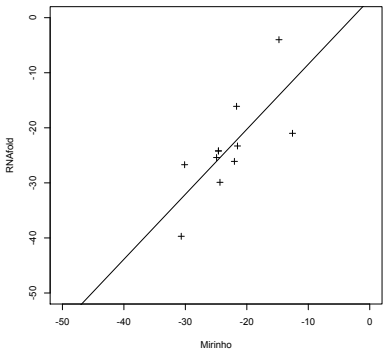
2.3.1 Regression analysis of the free energies

To verify how close we get to the algorithms based on a secondary structure prediction, we present a regression analysis between the energies of the pre-miRNAs corresponding to the true positive pre-miRNAs predicted by MIRINHO and their energies when predicted by RNAFOLD (Hofacker *et al.*, 1994).

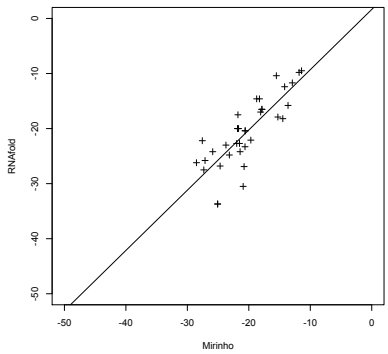
Figures 2.3a-2.3f shows the relationship of the energies for the true positive pre-miRNAs of chromosome 25 of *Bos taurus*, chromosome I of *Caenorhabditis briggsae*, chromosome 2R of *Drosophila simulans*, chromosome 25 of *Gallus gallus*, chromosome 22 of *Gorilla gorilla*, and chromosome 19 of *Mus musculus*. We consider as the dependent variable the energies of MIRINHO and as the independent variable the energies of RNAFOLD. As we can see, the energies are quite close to each other with, in general, bigger energies predicted by MIRINHO. It is expected that RNAFOLD produces energies that are more negative than MIRINHO since it minimises the free energy while the algorithm used by MIRINHO maximises the number of base pairs. This provides reasonable evidence that our method approximates well the free energy of hairpins.



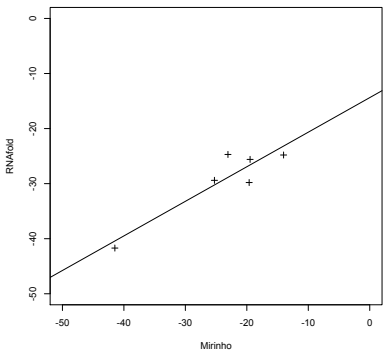
(a) *Chromosome 25 of Bos taurus* ($\rho = 0.8742886$).



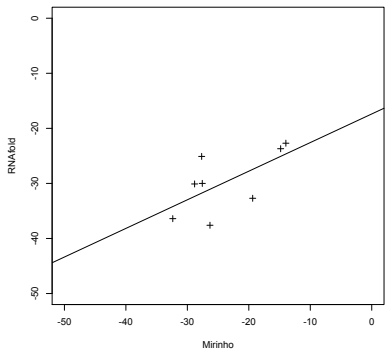
(b) *Chromosome I of Caenorhabditis briggsae* ($\rho = 0.7408282$).



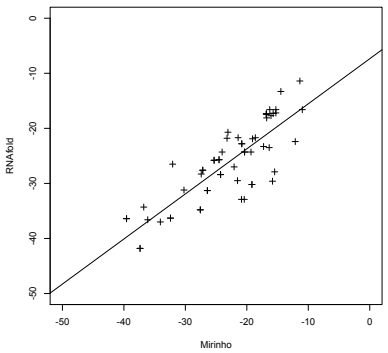
(c) *Chromosome 2R of Drosophila simulans* ($\rho = 0.8097171$).



(d) *Chromosome 25 of Gallus gallus* ($\rho = 0.9178415$).



(e) *Chromosome 22 of Gorilla gorilla* ($\rho = 0.6336104$).



(f) *Chromosome 19 of Mus musculus* ($\rho = 0.8320502$).

Figure 2.3: *Regression analysis of the energies predicted by* MIRINHO *and* RNAFOLD *for six different species.*

2.3.2 Time efficiency

As mentioned (in Section 2.2.1), we further improved the alignment algorithm by pruning the DP matrix and focusing on its diagonal only.

To establish the size of the diagonal portion of the DP matrix we should compute, we assessed different values for the parameter *dw* (diagonal width). The values for *dw* were evaluated empirically; they varied from 4 to 6 (see Table 2.1). A very small value for *dw* means to constrain the alignment to a very limited space around the diagonal part of the DP matrix, that is to permit a few or almost no bulges nor internal loops. This situation would not represent the real structure of a stem-loop and that is why we chose as minimum value $dw = 4$. On the other hand, a very large value for *dw* would not achieve the goal of the pruning strategy, that is time efficiency. In our experiments, the best results were obtained when using $dw = 5$ or $dw = 6$, which corresponds to the maximum number of unpaired nucleotides in the stem formed by both strands. The default value for the *dw* parameter was then set to 6.

	Sensitivity (%)	Precision (%)	*dw*
Aedes aegypti	38.85	0.02	4 (16%)
	38.85	0.02	5 (20%)
	39.57	0.02	**6** (24%)
Acyrthosiphon pisum	48.78	0.15	4 (16%)
	49.59	0.15	**5** (20%)
	49.59	0.14	6 (24%)
Culex quinquefasciatus	41.67	0.06	4 (16%)
	42.50	0.05	**5** (20%)
	42.50	0.05	**6** (24%)
Heliconius melpomene	63.37	0.28	4 (16%)
	63.37	0.27	5 (20%)
	64.36	0.27	**6** (24%)
Nasonia vitripennis	65.38	0.01	4 (16%)
	69.23	0.01	**5** (20%)
	69.23	0.01	**6** (24%)
Tribolium castaneum	43.45	0.25	4 (16%)
	43.69	0.24	5 (20%)
	44.42	0.23	**6** (24%)

Table 2.1: *Experiment to define the most appropriate value for the parameter dw (diagonal width, see Figure 2.2) of the dynamic programming (DP) matrix for sequence alignment, which is described in Section 1.2.2. The numbers in bold represent the values of dw with the best sensitivity and precision. The numbers in parenthesis, on the right side of the dw values, represent the percentage of the DP matrix that is used during the alignment of stem arms of 25nt. The energy threshold used is* $e = -20.6$.

The user of MIRINHO is given the freedom to compute the whole matrix instead of only its diagonal for a given value of *dw*. In this case, *dw* should be set equal to the length of the stem-arm (option -a).

Using this pruning strategy, the region exploited by the alignments is much smaller and the method performs, in general, 30% faster than the original version. Sensitivity and precision remain similar between the original and the optimised versions, in the great majority of the cases it remained the same.

Time efficiency is even more evident when comparing our method to other predictors, such as CSHMM, MIRENA, and MIRPARA. Table 2.2 presents the computation times for the prediction of putative pre-miRNAs in a sequence of length 4,951nt, running under a Mac OS X 10.6.8, 2.7 GHz Intel. As one can see, our method is indeed much faster than the others, making the prediction of pre-miRNAs much more feasible.

Method	**Time (in sec)**
MIRINHO	0.998
MIRPARA V6.2	68.008
MIRENA	989.958
CSHMM	1824.474

Table 2.2: *Running time comparison. Running time (in seconds) for the prediction of putative pre-miRNAs in a sequence of length 4,951nt, on a Mac OS X 10.6.8, 2.7 GHz Intel Core i7.*

To show that MIRINHO is much more applicable, we compared the time of prediction of MIRINHO, CSHMM, and MIRPARA. To facilitate the comparison of the predicted pre-miRNA candidates, we used the human chromosome 19, as the authors of CSHMM did. All three softwares were then submitted in a cluster queue of 29 hours (maximum job time without special bureaucratic request). MIRINHO finished its job after 5 hours, while the other two exceeded the 29 hours without finishing their prediction, with no reported result. Clearly, one can fragment the input in smaller pieces to finish the prediction with CSHMM and MIRENA, however the message here is to show that no fragmentation is required for long sequences since MIRINHO can finish its prediction in a smaller amount of time.

2.3.3 miRNA hairpin structure prediction in sRNA-seq of plant

To obtain a high quality structure, MIRINHO needs the information on the length of the stem-arms and terminal loop. It is clear that, when the search is made at a genomic scale, the precise information about length is unknown. However with sRNA-seq data, the length of the stem-arms and terminal loop may be naturally inferred from the alignment of the reads against the genome. This characteristic of our method thus allows its direct application to sRNA-seq data, enriching the prediction and quality of the hairpin structures.

To demonstrate this, we started by mapping the 340,114 reads of high-throughput small RNA sequencing data from *Arabidopsis thaliana* (GEO accession number GPL3968) to chromosome 4 of *Arabidopsis thaliana* using BOWTIE2 (Langmead *et al.*, 2009). We considered only the mapped regions that verified the expression profile of a pre-miRNA: high coverage on (at least one of) the stem arms and lower coverage in the terminal loop. It is easy to see that the length l of a stem-arm and the length t of the terminal loop can be naturally inferred from these alignments. We then gave l, t, and the respective pre-miRNA sequence as input to MIRINHO, RNAFOLD, and MIRNAFOLD.

For MIRINHO, we set the stem-arm length to l (option -a), and the minimum and maximum length of the terminal loop to t (options -n and -x respectively). Given that RNAFOLD is a method for predicting the secondary structure of an RNA in general, we used the option -C to force the structure to be a hairpin. We then required that the stem-arm regions, each of length l, were paired, and that the terminal loop region of length t was unpaired. For MIRNAFOLD, we gave as the sliding window parameter the length of the whole pre-miRNA, that is, l+t+l.

As MIRBASE is the basis for miRNA studies, we took its hairpin structures as a gold standard. In order to compare the structures predicted by the three methods, we then considered three criteria: (i) number of internal loops and bulges within the stem; (ii) length of the predicted stem-arm; (iii) length of the predicted terminal loop. For each predicted structure, we verified which method produced the best result. This corresponded to the predicted structure that produced values that are closest to those of the structure in MIRBASE. For example, if the MIRBASE structure s has 3 bulges, and RNAFOLD predicted a structure with 2 bulges while MIRINHO predicted one with 1 bulge, the first method would be considered the best one.

From the set of 50 pre-miRNAs of chromosome 4 of *Arabidopsis thaliana*, we randomly chose 10 structures from MIRBASE, for which such structures were predicted with the three methods from the sequences. In the end, MIRINHO obtained the closest structure in 80% of the cases, RNAFOLD was the second with 50%, and MIRNAFOLD the third with 40%.

Figures 2.4, 2.5, and 2.6 show, respectively, cases in which the closest structure was found by RNAFOLD, MIRINHO, and MIRNAFOLD. As we can see, even in the cases where MIRINHO was not the best, it was very close to the best.

2.3.4 Sensitivity and precision

To determine an appropriate energy threshold for the prediction of pre-miRNAs, we used randomly generated genomes. The reasoning behind this strategy is that, if the energy model is robust enough, there should exist a certain energy that is able to differentiate the stable hairpin structures from the randomly generated ones in which the base pairs would be established by chance.

To choose the different genomes for setting the threshold, we mainly considered the GC content as it plays an important role in determining a hairpin structure. We thus chose chromosomes with different percentages of GC varying from 37% to 54%, as shown in Table 2.3.

For each of the genomes, the nucleotide frequency distribution was used to generate the respective random genomes. After that, the prediction of the pre-miRNAs was performed in both versions (original and random) of each genome. We then chose as threshold the biggest energy for which the number of true miRNAs remains zero in the random genome, as can be seen in Figures 2.7a-2.7f. To define a true positive miRNA in the random genome, we simply verified if a given true miRNA in the original genome was present in the respective random region. Using this approach, the selected genomes had the thresholds presented in Table 2.3.

We provide to the user of MIRINHO an "automatic" way to set the threshold. In addition to the query genome G_q, the user should give as input a similar genome G_s, and an annotation file with the coordinates of the respective (true) miRNAs. MIRINHO will then generate a random genome G_r based on the nucleotide distribution of G_s, predict its pre-miRNAs, and compute the energy threshold. If the user chooses not to provide these additional files, the default energy is set to -20.6 kcal/mol, that is the mean of the previously mentioned energies

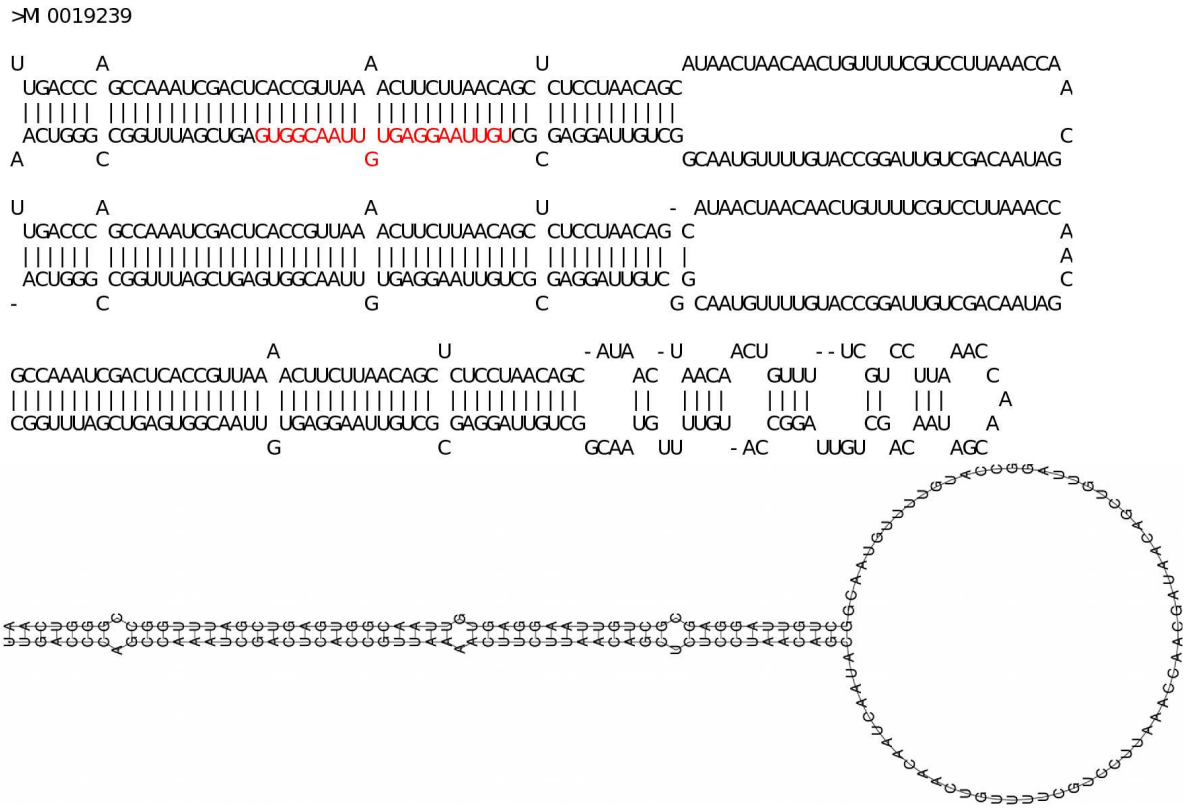


Figure 2.4: *From top to bottom: standard secondary structure of the pre-miRNA in* miRBase *(with miRNA coloured in red), and structures respectively predicted by* Mirinho, miRNAfold, *and* RNAfold. RNAfold *obtained the best prediction for the pre-miRNA MI0019239, with the closest values of stem length, terminal loop length, and number of bulges and internal loops as in* miRBase.

Species	Energy threshold	Chromosome	GC%
Caenorhabditis briggsae	−16	I	37,76
Mus musculus	−19	19	42,73
Gorilla gorilla	−19	22	47,74
Drosophila simulans	−21	2R	43,93
Gallus gallus	−24	25	54,96
Bos taurus	−25	25	46,96
Caenorhabditis elegans	–	III	35,75
Drosophila melanogaster	–	2R	41,84
Homo sapiens	–	19	50,06

Table 2.3: *Energy threshold obtained with the methodology mentioned in this section, and the GC% of the different chromosomes, including the ones for test (three last lines).*

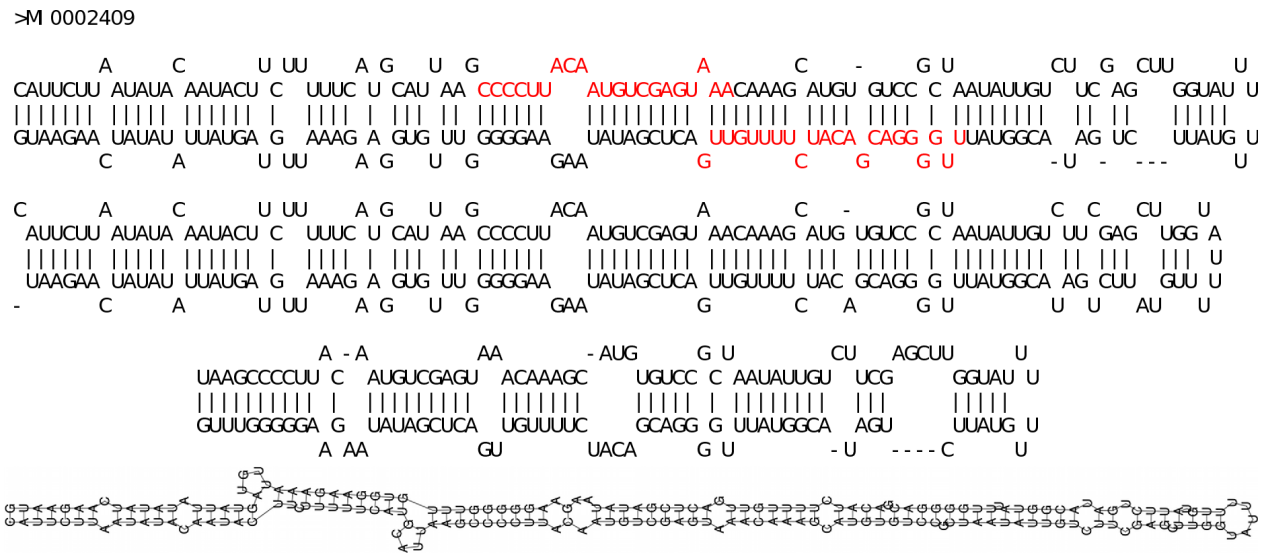


Figure 2.5: *From top to bottom: standard secondary structure of the pre-miRNA in* MIRBASE *(with miRNA coloured in red), and structures respectively predicted by* MIRINHO, MIRNAFOLD, *and* RNAFOLD. MIRINHO *obtained the best prediction for the pre-miRNA MI0002409, with the closest values of stem length, and number of bulges and internal loops as in* MIRBASE.

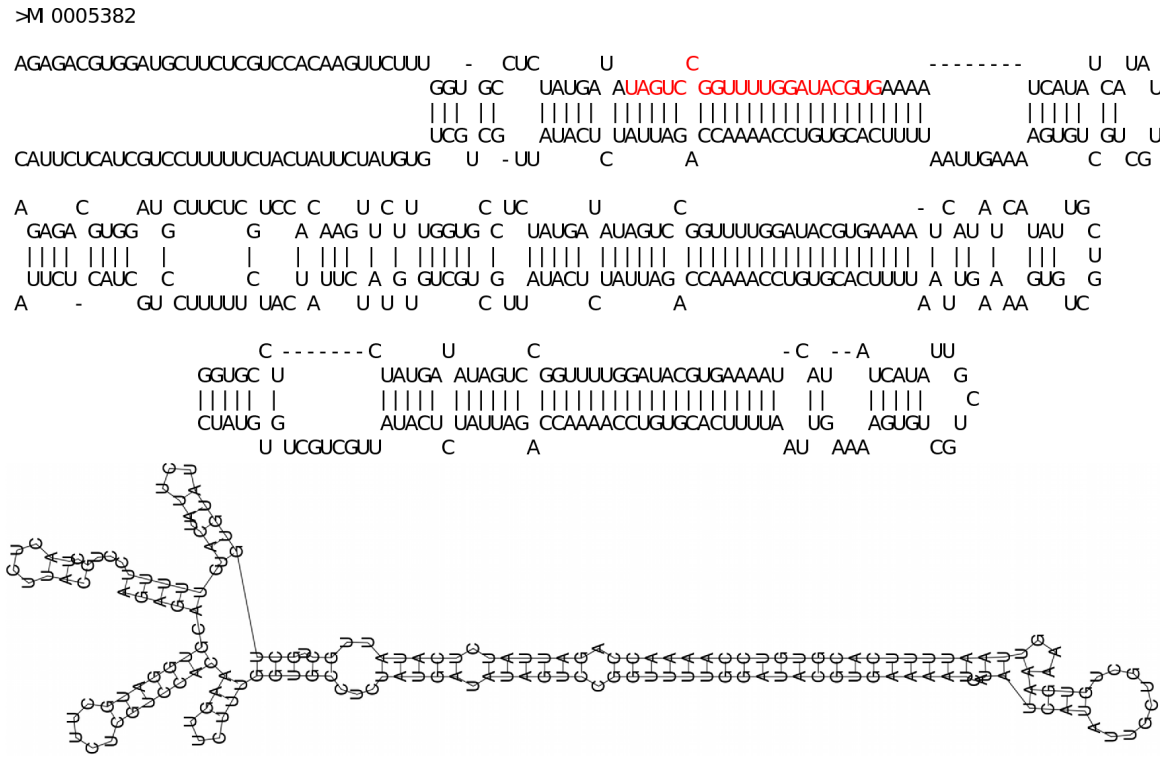
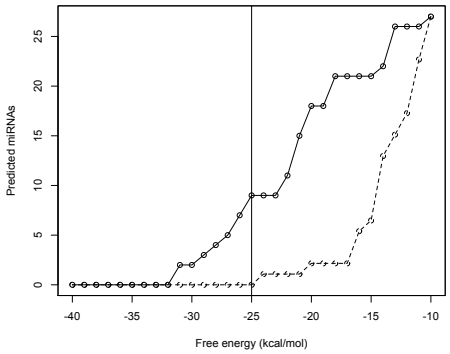
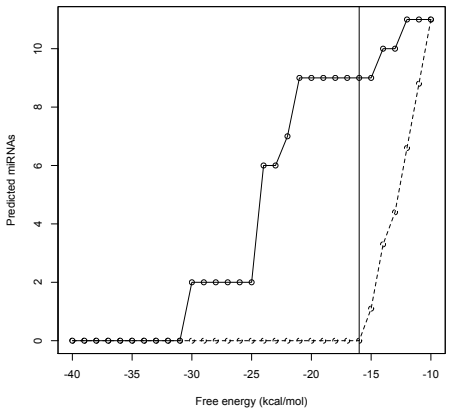


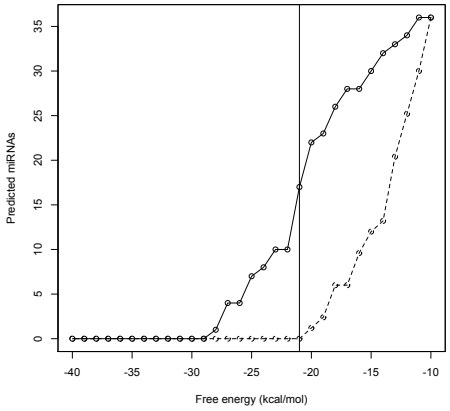
Figure 2.6: *From top to bottom: standard secondary structure of the pre-miRNA in* MIRBASE *(with miRNA coloured in red), and structures respectively predicted by* MIRINHO, MIRNAFOLD, *and* RNAFOLD. MIRNAFOLD *obtained the best prediction for the pre-miRNA MI0005382, with the closest values of terminal loop length, and number of bulges and internal loops as in* MIRBASE.



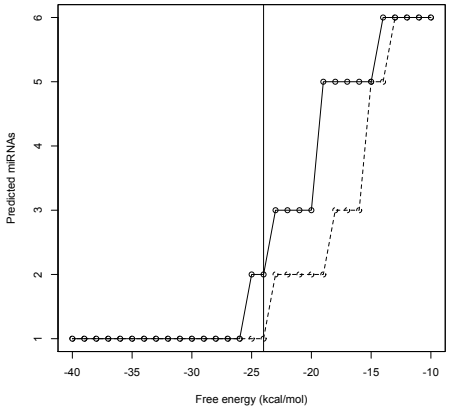
(a) *The energy threshold for Bos taurus is* -25 *kcal/mol.*



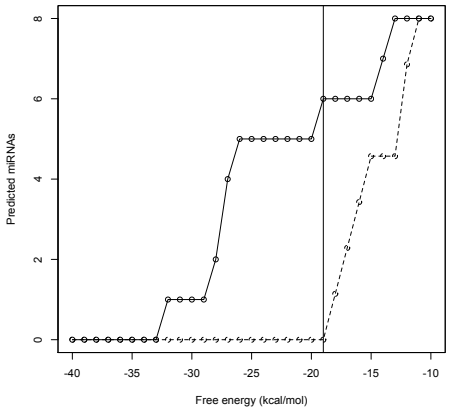
(b) *The energy threshold for Caenorhabditis briggsae is* -16 *kcal/mol.*



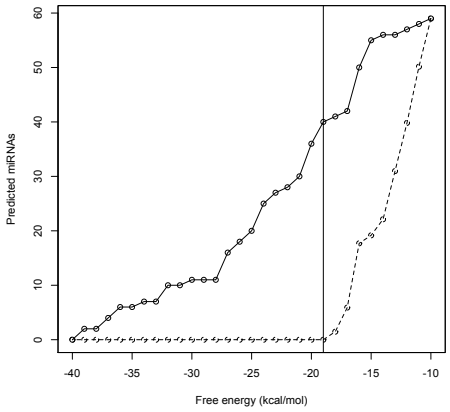
(c) *The energy threshold for Drosophila simulans is* -21 *kcal/mol.*



(d) *The energy threshold for Gallus gallus is* -24 *kcal/mol.*



(e) *The energy threshold for Gorilla gorilla is* -19 *kcal/mol.*



(f) *The energy threshold for Mus musculus is* -19 *kcal/mol.*

Figure 2.7: *Number of TP miRNAs predicted when using the original and the random genomes for the different species. The vertical line represent the energy threshold that better distinct true from false pre-miRNAs.*

generated with the same approach.

Table 2.4 presents a comparison between the different methods and MIRINHO with the mean energy threshold. As we can see, in humans MIRINHO has the best sensitivity (70%) and precision (50%) together with CSHMM. As concerns *Drosophila melanogaster*, MIRINHO also has the best sensitivity (80%), while MIRENA gets the best precision (75%). For *Caenorhabditis elegans*, CSHMM obtains the best sensitivity (70%), and MIRENA the best precision (44.44%).

	CSHMM 32202.854s		MIReNA 918.588s		miRPara 110.261s		Mirinho 1.667s	
	Precision	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**
Homo sapiens	23.08	60.00	*50.00*	10.00	13.00	60.00	*50.00*	*70.00*
Drosophila melanogaster	26.92	70.00	*75.00*	30.00	08.00	60.00	61.54	*80.00*
Caenorhabditis elegans	29.17	*70.00*	*44.44*	40.00	04.00	20.00	35.71	50.00

Table 2.4: *Comparison of the sensitivity, precision, and computing time of* CSHMM, MIReNA, Mirinho, *and* miRPara *using as input the dataset generated as described in Section 2.2.2. The energy threshold used in* Mirinho *was* $e = -20.6$. *The values for sensitivity and precision are given in percentage. Values in italic represent the best result for the given measure. The low precision for all the methods may be due to two reasons. One is that the model used for predicting (pre-)miRNAs needs refinement. The other is that the precise definition of a FP miRNA is completely dependent on the known miRNAs, which could represent just a small fraction of those that really exist.*

One should also remember that the only characteristics used by MIRINHO in the prediction of pre-miRNAs are the length of the terminal loop and stem-arms and the width of the diagonal. The other methods apply additional criteria that are based on other attributes, such as AU content, sequence homology, number of unpaired nucleotides, etc. Despite this, MIRINHO performs as well as the other compared methods and is at least 100 times faster than the quickest one (MIRPARA).

To analyse the sensitivity and precision at a genomic scale, we used the genomes of three insects, one of which, *Acyrthosiphon pisum*, is of particular interest to us. The results are shown in Table 2.4. Notice that the prediction is often far from being perfect for all methods; in particular, there is as usual a delicate choice to be made between sensitivity and precision, in as much as we are currently capable of accurately measuring the latter. The low precision for all the methods may be due to two reasons. One is that the model used for predicting (pre-)miRNAs needs refinement. The other is that the precise definition of a FP miRNA is completely dependent on the known miRNAs, which could represent just a small fraction of those that really exist.

Organism	Method	Sensitivity	Precision
Acyrthosiphon pisum	MIRINHO	*69.92*	0.52
	CSHMM	23.58	0.05
	MIRPARA	36.59	0.14
	MIRENA	24.39	*3.42*
Culex quinquefasciatus	MIRINHO	*69.17*	0.25
	CSHMM	48.51	0.10
	MIRPARA	28.33	0.07
	MIRENA	18.33	*2.00*
Heliconius melpomene	MIRINHO	*78.22*	0.94
	CSHMM	48.51	0.10
	MIRPARA	58.42	0.23
	MIRENA	31.68	*7.88*

Table 2.5: *Sensitivity and precision of three insect genomes. The energy threshold used in* MIRINHO *was* $e = -20.6$. *Values are given in percentage, and the ones in italic represent the best value for the given measure.*

2.4 Conclusion

With MIRINHO, we propose a faster and flexible method for the prediction of pre-miRNAs, using minimal information about known pre-miRNAs. Concerning the prediction results, we obtain very reasonable sensitivity and precision similar to the other tested methods, and in some cases even better. As concerns the quality of the predicted structures, the hairpins predicted by MIRINHO are much closer to the ones available in MIRBASE than the ones predicted by RNAFOLD and MIRNAFOLD.

Our method is faster because we employ a quadratic time complexity algorithm to predict the free energy of the hairpin, instead of the so used cubic algorithm for prediction of RNA secondary structure. We are flexible in two aspects. First, as concerns the input type we

accept both whole genome sequence and sRNA-seq data. Second, MIRINHO may be used for the prediction of either plant or animal pre-miRNAs, with a minimal adjustment (of the length of the stem-arm and terminal loop only). Finally, the only a priori knowledge we use is the length of the stem-arm, the length of the terminal loop, and the width of the diagonal.

Chapter 3

MicroRNA expression profile during embryonic development in *A. pisum*: combining deep sequencing data and MIRINHO to identify miRNAs

Contents

3.1 Introduction 44
3.2 Material and methods 44
3.2.1 Aphid rearing and embryo isolation 44
3.2.2 RNA extraction 46
3.2.3 Next-generation Illumina Sequencing 46
3.2.4 Treatment of the small RNA sequencing data 46
3.2.5 MicroRNA expression profile 49
3.2.6 Target prediction 49
3.3 Results and discussion 50
3.3.1 Statistical summary of the sequenced reads 50
3.3.2 MicroRNAs Expressed in *Acyrthosiphon pisum* 53
3.3.3 MicroRNA gene expression profile 63
3.3.4 MicroRNA target prediction 63
3.4 Conclusion 67

This chapter is strongly based on the paper Higashi *et al.* (tion) (in preparation). It presents an analysis of the small RNA sequencing (sRNA-seq) data of *Acyrthosiphon pisum*, the pea aphid, which were obtained during this thesis at ProfilExpert Genomic Platform (Université de Lyon, France) under the supervision of Dr Stefano Colella and with financial support of Dr Marie-France Sagot—European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. [247073]10. The persons involved in the wet experiments were Gabrielle Duport, Karen Gaget, Federica Calevro, and Hubert Charles from the SymTrophique team at BF2I (Biologie Fonctionnelle, Insectes et Interactions, UMR0203). To treat the data in order to guarantee a more accurate set of reads, as well as to detect the expressed miRNAs, three approaches were used: (i)

MIRINHOPIPE, specially developed for this analysis; (ii) sRNA-PLAN, a pipeline designed for the annotation of small RNAs; and (iii) miRDEEP, a classical method for the discovery of miRNAs from deep sequencing data (Friedländer *et al.*, 2008). The detected miRNAs were submitted to the prediction of mRNA targets. Together with such predictions, the gene expression profile of *Acyrthosiphon pisum* was analysed and compared to the miRNA expression profile, leading to very interesting results.

3.1 Introduction

The unique feeding habit of aphids combined with their ability to rapidly reproduce makes of them one of the most damaging pests of crops with economical importance worldwide. Considering their impact on agriculture and the role miRNAs play in gene regulation, it is imperative to better characterise and understand the function of these miRNAs.

One first effort has already been made by Legeai *et al.* (2010a) in *Acyrthosiphon pisum* (the pea aphid), a laboratory model for the study of these pests whose genome was sequenced. The authors combined small RNA sequencing data from parthenogenetic females and bioinformatics approaches to identify 103 *Acyrthosiphon pisum* miRNAs. It is worth noting that in Legeai's work, the miRNAs of parthenogenic females were sequenced and analysed, while we focus on the miRNAs expressed in three embryonic developmental and one larval stages; all the details concerning their methodology is presented in Section 3.3.2. Furthermore, the potential mRNA targets of the detected miRNAs were identified by the overlapped predictions of two methods (PITA and miRANDA).

Another effort published by Hansen et Degnan (2014), that is not directly related to small RNAs in *Acyrthosiphon pisum* but instead to small RNAs in its symbiont *Buchnera aphidicola*, provides evidence of protein regulation and of a reasonable number of conserved small RNA and UTR sequences among different *Buchnera* strains. The authors predicted small RNAs involved in the post-transcriptional mechanisms of the bacterium, as well as other types of mechanisms, for instance involving proteases at the same post-transcriptional level.

3.2 Material and methods

3.2.1 Aphid rearing and embryo isolation

A long-established parthenogenetic clone (LL01) of *Acyrthosiphon pisum* was maintained at 21°C, with a 16 hour photoperiod, on *Vicia faba* (L. cv. Aquadulce). In order to have a supply of synchronised aphids and embryos, around one hundred mass-reared winged adults were maintained on young plants and removed after 24 h. The resulting apterous insects were maintained on *Vicia faba* plants for a nine-day period, until they reached the adult stage. Embryos were dissected from synchronised parthenogenetic viviparous adult aphids, removing the ovariole sheath in a buffer kept on ice. We used an RNase-free buffer composed of 35 mM Tris-HCl (pH 7.5), 25 mM KCl, 10 mM $MgCl_2$, 250 mM sucrose, in 0.1% diethyl pyrocarbonate water. Following a stereoscopical analysis (Olympus IX-81, Olympus, France), embryos were classified according to their length and morphological characteristics into 3 groups (see Table 3.1 and Figure 3.1): early embryos (EE) ($\leq$ 0.4 mm), intermediate embryos (IE) (0.4 to 0.8 mm), and late embryos (LE) ($>$ 0.8 mm) corresponding, respectively, to the developmental stages $\leq$ 15, 16-18 and 19-20 as described by Miura *et al.* (2003). For L1 aged from 0 to 24h, viviparous adults were maintained on young plants for 24 hours and the resulting L1s were collected.

Group	Group abbreviation	Developmental stages	Size (length or weight)	External morphological features
Early embryos	EE	0-15	$\leq 400\mu$m	No visible eyes, very slight body pigmentation
Intermediate embryos	IE	16-18	$400 - 800\mu$m	Developing eye spots in many individuals, pigmented bodies
Late embryos	LE	19-20	$> 800\mu$m	Developed eye spots in all individuals, highly pigmented bodies
First instar larvae	L1	1st larval	≤ 0.2 mg	0-24 hours old

Table 3.1: *Description of embryonic and larval stages used for the extraction of the total RNA (subsequently submitted to Ilumina sequencing). The first column presents the four developmental groups of Acyrthosiphon pisum, followed by their abbreviation, and the developmental stage itself as described by Miura* et al. *(2003). The size of the organism is subsequently presented together with the morphological features associated to the developmental stage; the external morphological features can be observed in Figure 3.1.*

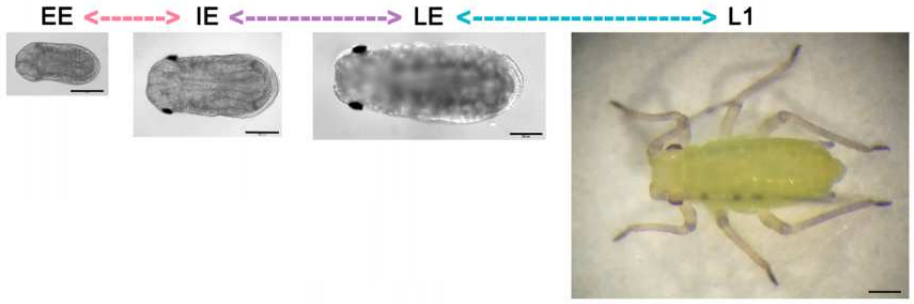


Figure 3.1: *Micro-photographs of the four stages, the scale bar represents 200 µm in all photographs to allow for size comparison. The microphotographs show just one embryo stage among those belonging to the corresponding groups (see Table 3.1 for details). Figure taken from Rabatel* et al. *(2013).*

3.2.2 RNA extraction

Total RNA was prepared using the mirVana™ miRNA Isolation Kit (Ambion, Austin, TX, USA). Three independent extractions were prepared for each group starting with 60 embryos for the EE group, 30 embryos for both the IE and LE groups, and 30 larvae for the L1 group (0-24h). The extraction was followed by a step of DNase treatment using DNA-*free*™ DNase Treatment and Removal Reagents (Ambion, Austin, TX, USA). Total RNA concentration and quality were initially checked using the NanoDrop® ND-1000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and samples had to meet the following quality parameters: A260/A280 $\geq$ 1.8 and A260/A230 $\geq$ 1.8, in order to be used in the subsequent analysis. The RNA samples were then run using the Agilent RNA 6000 Nano Kit on the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA) to check their integrity. Degraded samples appeared as significantly lower intensity traces, with the main peak area shifted to the lower molecular weights, and they typically exhibited much more noise on the trace. Only good quality samples were sent for sequencing.

3.2.3 Next-generation Illumina Sequencing

Total RNA was shipped to ProfilExpert Genomic Platform (Université de Lyon, France). RNA concentration was verified using the RiboGreen® Assay Thermo Scientific, Wilmington, DE, USA) for precise quantification before sequencing. Barcoded small RNA libraries were created from 1 ug of total RNA for each sample according to Illumina TruSeq small RNA Sample Preparation Guide (Illumina, San Diego, CA, USA): adaptor ligation was followed by RT-PCR amplification. The small RNA libraries were gel purified and they were validated on the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA) to check for quality and size. The 12 samples were sequenced with a Single Read 50 cycles run on one lane of the flow cell v3 (150 million raw reads) of the Illumina HiSeq-2500 (Illumina, San Diego, CA, USA).

3.2.4 Treatment of the small RNA sequencing data

To identify the expressed miRNAs in the four developmental stages of the pea aphid, three methods were used: MIRINHOPIPE, SRNA-PLAN, and MIRDEEP. The first was developed specially for the purpose of analysing *Acyrthosiphon pisum* sRNAseq data by the author of this thesis at the BAMBOO-BAOBAB team (head Dr Marie-France Sagot) of the LBBE-UMR5558. SRNA-PLAN was developed as a pipeline for small RNA annotation by one of our collaborators, Oliver Rue, at UBIA & PF GenoToul Bioinfo (head Dr Christine Gaspin). MIRDEEP2 is a classical method for the discovery of miRNAs from deep sequencing data; it uses the read stacks, that are consistent to the ones of an expressed miRNA, to select the best candidates to further verify other characteristics, such as the free energy (Friedländer *et al.*, 2012). To document and describe the technical details of the pipelines, a wiki was created and is available at `http://mirinho.gforge.inria.fr/mirinhopipe.html`.

To present the details on how the methods were developed and/or performed, we use Figure 3.2 as a reference guide. As shown in the figure, the first three steps were common to MIRINHOPIPE, SRNA-PLAN and MIRDEEP2. From raw data, we used CUTADAPT (version 1.4.1) to trim the adapters from the 3' end (option -a) of the reads, and to filter out reads with less than 16nt (option -m). The redundancy was removed by collapsing redundant reads within each of the four samples, followed by the copy number computation of the unique reads. Only reads with a copy number greater than ten (10X) were kept. As mentioned before, for

each sample three biological replicates were made. If a miRNA transcript is indeed expressed, it should thus appear in all the three replicates. Based on that, only the reads that appeared in all the three replicates were considered for subsequent analysis.

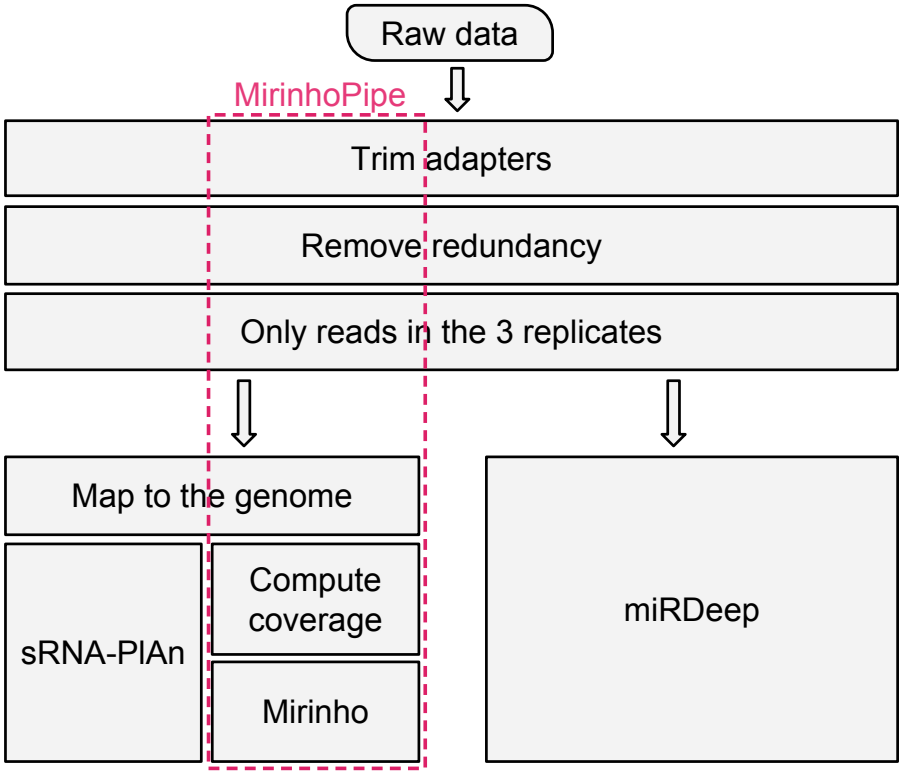


Figure 3.2: *Flowchart describing the steps of the methods for the treatment of small RNA sequencing (sRNA-seq) data. The information flows from top to bottom: (i) trimming the adapters from the 3' end and filtering out reads smaller than 16nt using* Cutadapt *(version 1.4.1); (ii) collapsing redundant reads and computing their copy number; (iii) only reads appearing in the three replicates remained for subsequent analysis; (iv) on the left: mapping the reads to the genome with* Bowtie2*; (v) computing coverage of mapped regions with* genomeCov *and excising potential pre-miRNAs sequences; (vi) computing the free energies of pre-miRNA hairpins with* MirinhoPipe*. The set of unique reads appearing in the three replicates were also given as input to* miRDeep *(on the right). The details concerning this last method are provided in the text.*

After the cleaning process previously mentioned, the more accurate set of reads were mapped to the genome of *Acyrthosiphon pisum* (assembly version 2) using Bowtie2 (version 2.1.0). It was required that the reported reads mapped to at most 5 different loci in the genome (option -k), and only alignments with at most 1 mismatch were permitted (field "XM:i:<N>" from the Bowtie2 output represents the number N of mismatches). The very same set of accurate reads was also given as input to miRDeep2. The details concerning the subsequent steps of each method are provided in what follows.

MirinhoPipe

It is worth noting that all the preceding steps are included in MirinhoPipe; however, to maintain a structured presentation we split the description of the steps. From the filtered

Bowtie mappings, we first computed the coverage of each position in the genome using a tool called GENOMECOV from the toolset Bedtools (version 2.17). If a region has a coverage of a minimum height of one and a minimum length of 20nt (length of a miRNA), it is considered as a region with potential to harbour a (pre-)miRNA. To guarantee that the whole pre-miRNA is identified, flanking portions are also taken into account: if the region is smaller than 70nt, a flanking portion of 60nt down and upstream is considered and the final pre-miRNA locus is extracted.

These potential pre-miRNA loci are then given as input to MIRINHO for the computation of their secondary structures and free energies. The energy threshold used is -20.6 kcal/mol, which is set as described in Section 2.3.4.

sRNA-PLAN

As for MIRINHOPIPE, the reads mapping to a same locus in the reference genome are assembled into a longer region resulting in a potential miRNA locus. Each locus is submitted to the annotation process and prediction of miRNA(s). To annotate a locus, non-coding RNA (ncRNA) databases are used to assign one or more putative function(s). All the loci are then submitted to a prediction step that will determine if it is a potential miRNA or not. For each potential miRNA, the up and downstream flanking portions are accounted to extract the pre-miRNA sequence. A glocal alignment is then computed between the most represented read (the putative miRNA) and its 5' or 3' neighbour region, in order to mimic the hybridisation between both strands of the potential hairpin. Each alignment is scored according to a few criteria related to the expression profile. The top ranked miRNAs are thus classified in three classes: (i) miRNA-annotated/predicted; (ii) other-function-annotated/predicted; (iii) annotation-orphan/predicted. These putative candidates can be sorted according to their score to be further submitted to experimental validation. The details about this pipeline were omitted in this thesis manuscript because it is not yet publish.

miRDeep2

The miRDeep package is composed of two modules, MAPPER and miRDeep2. The first module maps the reads to the genome with BOWTIE (version 1), keeping only the alignments with 0 mismatches (option -n) in the seed region. The seed region, set to 18nt (option -l), is defined as the n first nucleotides of a read. A maximum of 2 mismatches (option -e 80) occurring after the seed region were allowed (option -n). Only reads that do not map to more than five different loci in the genome were kept (option -m). Option -best-strata was used to order the mappings (from best to worse) according to the strata definition of BOWTIE (Langmead *et al.*, 2009; Friedländer *et al.*, 2012).

In the second module, potential miRNA precursors are excised from the genome using the read mappings as guidelines. Then the two genome strands of each genome sequence are scanned separately, from 5' to 3' . Excision is initiated when a stack of reads (height one or more) is encountered. If there is a higher read stack within 70nt downstream of the current read stack, then this is chosen instead. In this way, the highest local read stack is identified. Then the sequence covered by the highest local read stack is excised twice, once including a 70nt upstream and a 20nt downstream flanking sequence, and once including a 20nt upstream and a 70nt downstream flanking sequence. The second step of the module is to prepare the signature file. The BOWTIE-BUILD tool is used with default options to build a Burrows-Wheeler transform index of the excised potential precursors. Then the set of sequencing reads is mapped to the index, using BOWTIE (version 0.12.7). The set of known

mature miRNAs for the reference species is also mapped to the index. The RNA secondary structures of the potential precursors are then predicted with RNAFOLD with default options. Finally, the potential precursors are individually scored or discarded by the core algorithm of MIRDEEP2.

3.2.5 MicroRNA expression profile

Before verifying the expression profile of the pre-miRNAs, a normalisation of the read counts is necessary. To that purpose, the RPM (reads per million) number was computed according to Equation 3.1 (modified from RPKM—reads per kilobase per million—equation described in (Ammar *et al.*, 2012)):

$$RPM = r/(R_s * 10^{-6}) \tag{3.1}$$

where r is the number of reads mapped to the given transcript (in this case the pre-miRNA), and R_s is the total number of reads from sample s that mapped to any locus of the genome. The normalisation originally includes the length of the transcript; however, we did not consider it here because miRNAs are roughly of the same length (i.e., ∼22nt).

To visualise the pre-miRNA gene expression in each sample, we used MEV (MultiExperiment Viewer), a software that was originally developed for microarray data analysis. MEV incorporates algorithms for clustering, visualisation, classification, statistical analysis, and biological theme discovery from single or multiple experiments (Howe *et al.*, 2010). The HCL option in MEV allows for the visualisation of the datasets by means of a heatmap, a graphical representation of the expression data in which the values in the matrix are represented by colors, in an organised manner, via a dendrogram, to look for emergent trends.

To build the dendogram, a hierarchical clustering is implemented in MEV. It is based on the average-linkage method developed by Sokal et Michener (1958) for clustering correlation matrices. The algorithm assembles all elements, in this case the pre-miRNAs, into a single tree. For a set of n pre-miRNAs, an upper-diagonal similarity matrix is computed by using the Pearson correlation as a metric to score all pairs of pre-miRNAs. The pair of pre-miRNAs with the most similar expression profile is determined by finding the largest value in the matrix (i.e., the best correlation). A new cell is created by joining the two pre-miRNAs, and a new expression profile (normalised read count) is computed for the cell by averaging the expression of the joined elements. The similarity matrix is updated with this new cell replacing the two joined elements, and the process is repeated $n - 1$ times until only a single element remains. We used a similar strategy to cluster motifs as shown in Section 5.2.2.

3.2.6 Target prediction

The 3'UTR target sequences were obtained at APHIDBASE (Legeai *et al.*, 2010b), requiring a minimum length of 50nt for a target sequence. As a consequence, from the 40,336 sequences, 32,127 remained. Together with the targets, the set of detected miRNAs were given as input to two algorithms for the prediction of targets, PITA and MIRANDA. We also used RNAHYBRID, however, as it produced too many interactions and none of them were common to the ones of the two other, we decided to put is aside. The main difference between MIRANDA and PITA, is that the latter uses accessibility to predict functional interactions; more details about each method are provide below.

MIRANDA uses classical features such as sequence matching, highly scoring matches at the miRNA 5' end (seed location), free energy, and target site conservation in three insect

species *Drosophila melanogaster*, *Drosophila obscura*, and *Anopheles gambiae*; for validation the authors used targets from *Caenorhabditis elegans* and *Caenorhabditis briggsae* (Enright *et al.*, 2004).

PITA models the target site accessibility by defining a score, $\Delta\Delta G$, which is computed as the difference between the energy gained to form the duplex ΔG_{duplex} and the energetic cost to unpair the target secondary structure ΔG_{open} (Kertesz *et al.*, 2007). To validate their method, the authors used a quantitative luciferase assay in *Drosophila melanogaster* tissue. More details about the methods are presented in Section 1.2.4.

3.3 Results and discussion

3.3.1 Statistical summary of the sequenced reads

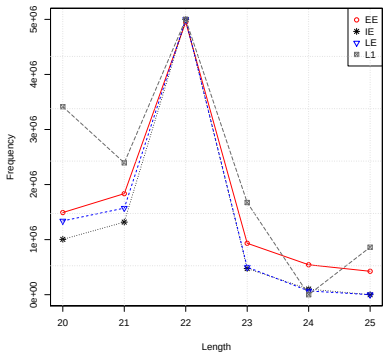
As mentioned in the previous sections, the sRNA-seq data were submitted to a series of processing steps before the miRNA detection itself. Table 3.2 shows the evolution of the read counts after each processing step. Filtering reads out by length (column "CUTADAPT") eliminated an average of ∼36.8% of the total reads—31.7% for EE, 30.3% for IE, 35.5% for LE, and 49.7% for L1. Collapsing the redundant reads (column "Unique reads") resulted in an average removal of ∼97.2% of the reads (in relation to the previous step). Disregarding reads with a copy number smaller than 10X eliminated ∼82.2% of the reads. Finally, considering only the reads appearing in all the three replicates and only the reads that mapped to the genome, removed respectively ∼30% and ∼40,5% of the total reads. Initially the dataset was comprised of 187,357,260 reads; after all the processing and pre-treatment of this dataset, a more accurate set of 352,061 reads remained.

Figures 3.3a-3.3c present the read length distribution across the four samples EE, IE, LE, and L1. The length distribution is analysed from four different perspectives: (i) considering *all* the reads (Figure 3.3a), i.e., the reads obtained after step "Trim adapters" from Figure 3.2; (ii) considering *collapsed unique* reads (Figure 3.3b), i.e., the reads acquired after the step "Remove redundancy" from the same figure; (iii) considering *all* the reads that appear in all the *three replicates* (Figure 3.3c) – from the reads used in step i, only the ones appearing in all the three replicates remained; and (iv) considering *collapsed unique* reads that appear in all the *three replicates*. These four perspectives were chosen to first verify how the "noise" of redundant reads could bias the distribution, and how a more accurate set of reads (i.e. the ones appearing in all three replicates) could affect the same distribution. In the distributions with no redundancy, it is easier to see the points (which correspond to the different miRNA lengths) where the reads are concentrated and to see differences between each sample. The same occurs when we consider only the reads in the three replicates, mainly for the case "all" reads (left side of Figure 3.3). We can notice that the majority of the reads are of length 22nt, which is indeed the mean length of a miRNA. This peak at 22nt occurs for samples EE, IE and LE. For sample L1, differently from what is observed for the other samples, the peak is at 20nt, with a number of reads similar to the ones at 21nt and 22nt. Since the larval stage is developmentally farther from the three embryo stages, it is natural to expect a different behaviour for L1 in relation to the three others, and a similar behaviour within the three embryo stages. This heterogeneity of the miRNAs in L1 may be explained by the heterogeneity of the organisms from which the RNA was extract. The organisms in L1 live in a different environment than the embryos: L1 larvae are exposed to an *in vivo* medium, feeding on the plants, while embryos are inside their progenitor in a more stable environment. This difference in the environment results in a variation in the organism, and as a consequence

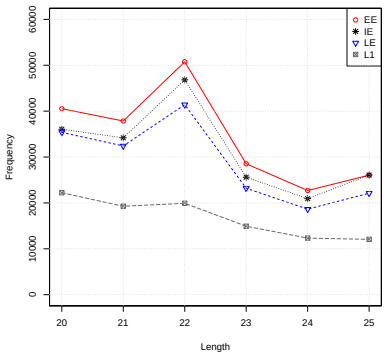
	Raw	Cutadapt	Unique reads	10X	3 replicates	Mapped reads
EE-1	17,009,637	12,510,677				
EE-2	15,740,057	10,338,438				
EE-3	17,214,015	11,321,579				
EE	49,963,709	34,170,694	1,065,896	192,172	186,354	105,271 (0.031%)
IE-1	14,857,272	10,743,632				
IE-2	14,639,000	9,920,251				
IE-3	18,103,230	12,531,817				
IE	47,599,502	33,195,700	1,045,952	182,050	176,256	105,582 (0.032%)
LE-1	17,260,430	10,881,355				
LE-2	16,128,653	10,579,881				
LE-3	13,227,401	8,623,332				
LE	46,616,484	30,084,568	903,452	161,709	157,979	92,622 (0.031%)
L1-1	15,509,610	8,235,329				
L1-2	13,611,983	7,130,465				
L1-3	14,055,972	6,366,278				
L1	43,177,565	21,732,072	448,818	80,125	77,318	48,586 (0.011%)

Table 3.2: *Read counts at the different steps of the treatment workflow for the four samples EE, IE, LE, and L1—the number after the dash represents the replicate. From raw reads, the first step is to trim the adapters from the 3' end and to remove the reads smaller than 16nt with* Cutadapt. *The number of unique reads and the respective copy number is computed. Reads with copy number smaller than 10X are discarded. Only the reads appearing in all the three replicates remain. Finally, the number of mapped reads, with at most 1 mismatch, is presented. The number in parenthesis is the percentage of reads that remained in relation to the column "Cutadapt".*

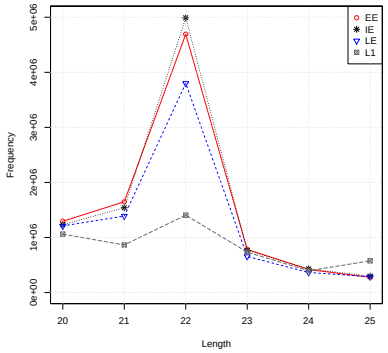
more variation in the miRNA transcripts.



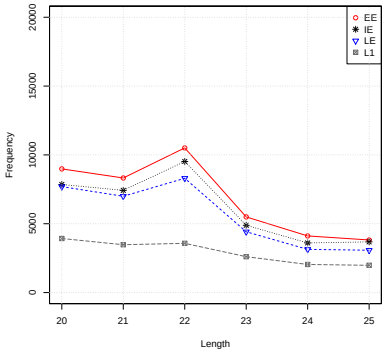
(a) Considering *all* the reads.



(b) Considering *collapsed unique* reads.



(c) Considering *all* the reads that appear in all the *3 replicates*.



(d) Considering *collapsed unique* reads that appear in all the *3 replicates*.

Figure 3.3: *Read length distribution across the four samples EE, IE, LE, and L1 from four different perspectives.*

3.3.2 MicroRNAs Expressed in *Acyrthosiphon pisum*

In 2010, Legeai and co-workers identified 103 mature miRNAs in *Acyrthosiphon pisum* using three different approaches. In the first approach, the authors blasted insect miRNAs from MIRBASE (release 14) against the genome of *Acyrthosiphon pisum* (assembly version 1.0). The second approach consisted in sequencing small RNAs from a mixed generation sample of *Acyrthosiphon pisum* parthenogenetic females, and mapping the 850,000 unique reads against the same genome used in the first approach. The mappings were thus given as input to MIRDEEP. In the third approach, the authors implemented a machine learning classifier trained with 30 pea aphid miRNAs. The ensemble of these three methods initially produced 149 mature miRNAs that were deposited in MIRBASE. From MIRBASE release 14 to the current release 21, 46 pea aphid miRNAs were removed from the database, thus remaining 103 miRNAs of *Acyrthosiphon pisum*.

While Legeai *et al.* (2010a) sequenced parthenogenic females, we sequenced the small RNAs extracted at four different developmental stages of *Acyrthosiphon pisum*: early embryo (EE), intermediate embryo (IE), late embryo (LE), and larvae (L1) stage. Moreover, to

guarantee quality and consistency of the sRNA-seq data, for each sample three biological replicates were made (for the details about the experimental procedure, see Section 3.2). After treating the sequenced reads, the expressed miRNAs discovered with our methodology were classified in three categories: (i) miRNAs known in *Acyrthosiphon pisum*; (ii) miRNAs known in other species but not present in *Acyrthosiphon pisum*; and (iii) potential novel miRNAs. Table 3.3 summarises the number of miRNAs in each of these categories identified using the three analysis methods.

	Known in *Acyrthosiphon pisum*	Known in other species	*Potential Novel
MirinhoPipe	70	26	4908
miRDeep	65	21	454
sRNA-PlAn	56	21	826
Predicted by all the 3	40	16	23

Table 3.3: *Summary of the miRNAs predicted by* MirinhoPipe, miRDeep, *and* sRNA-PlAn *organised in three categories: (i) miRNAs known in Acyrthosiphon pisum; (ii) miRNAs known in other species but not present in Acyrthosiphon pisum; and (iii) potentially novel miRNAs. In this table, we refer to all the discovered miRNAs disregarding the sample(s) from which they originated. The last line contains the number of strict consensus miRNAs (i.e. predicted by all the three methods). *These are the predicted miRNAs that did not fit into any of the two categories i and ii, for the final list of potential novel miRNAs more criteria were verified (see Section 3.3.2).*

MicroRNAs known in *Acyrthosiphon pisum*

In miRBase (release 21), there are currently 103 pea aphid mature miRNAs deriving from 123 precursors. To identify the known miRNAs in our data, we used Blastn (version 2.2.28+) to align the 103 mature miRNA sequences against the pre-miRNAs identified by the three methods. To define a miRNA as a known miRNA in the pea aphid, we used the following criteria: (i) glocal (global+local) alignment required, local for the pre-miRNA and global for the miRNA, that is, the miRNA must be fully covered by the pre-miRNA; and (ii) maximum of one mismatch in the alignment.

From the 103 known miRNAs, MirinhoPipe retrieved 70, miRDeep 65, and sRNA-PlAn 56 miRNAs. As we can see in Figure 3.4, the number of strict consensus miRNAs between each two methods is close to the number of miRNAs predicted when considering each of the two methods separately; for example miRDeep and sRNA-PlAn find 51 strict consensus miRNAs, while miRDeep and sRNA-PlAn alone predicts, respectively, 65 and 56 miRNAs. It means that the methods are converging in their predictions and are consistent in their results. The combination MirinhoPipe+miRDeep detected 53 and MirinhoPipe+sRNA-PlAn 43 strict consensus miRNAs. The combination of all the three methods resulted in a high confidence set of 40 miRNAs known in *Acyrthosiphon pisum*.

From the 103 *Acyrthosiphon pisum* miRNAs in miRBase, 100 miRNAs have an annotation with an experimental evidence ("Evidence: experimental; Illumina"), while three miRNAs (api-miR-1923, api-miR-281, api-miR-iab-4) are annotated with an evidence obtained by similarity ("Evidence: by similarity"). It is worth noting that from these three miRNAs, miRNA api-miR-281 appears in our high confidence list (i.e., identified by the three methods), while

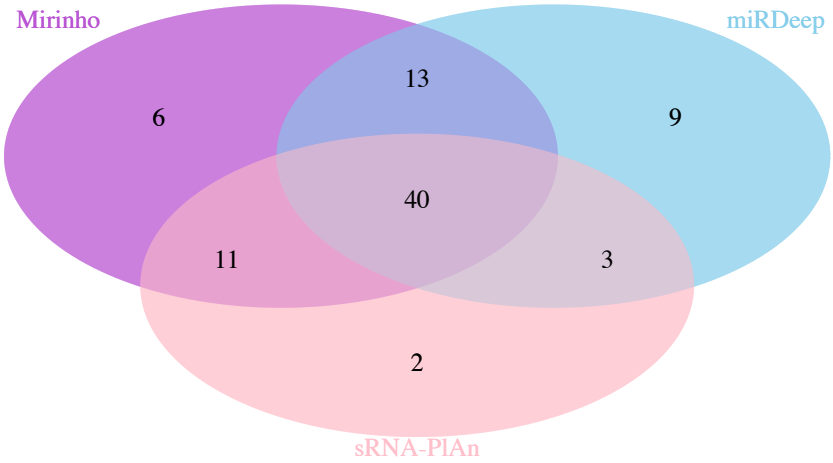


Figure 3.4: *Venn diagram of the miRNAs (known in Acyrthosiphon pisum) recovered by* MIRINHOPIPE *(70),* MIRDEEP *(65), and* SRNA-PLAN *(56). The preceding numbers in parenthesis represent the total number of miRNAs in each set.*

miRNA api-miR-iab-4 was retrieved by both MIRINHOPIPE and MIRDEEP; miRNA api-miR-1923 was not recovered by any method. One difference between these three miRNAs that is important to highlight is that the high confidence miRNA api-miR-281 is expressed in 41 different species and miRNA api-miR-iab-4 in 24, while the miRNA api-miR-1923 that was not found in our predictions appears in only one species *Bombyx mori*. Based on these results and considering that conservation is a strong argument, the annotation field "Evidence" in MIRBASE, for the miRNAs api-miR-iab-4 and api-miR-281 should be updated to experimental, as we have obtained it from sRNA-seq data.

MicroRNAs known in other species and not (yet) identified in *Acyrthosiphon pisum*

There are currently 32,488 miRNAs in MIRBASE (release 21) that are known in other species and were not (yet) identified in *Acyrthosiphon pisum*. From these miRNAs, 26 were identified in our data by MIRINHOPIPE, 21 by MIRDEEP, and 21 miRNAs by SRNA-PLAN, as shown in Figure 3.5. It is worth noting that to compute the preceding numbers, the unique mature miRNA sequence was considered instead of miRNA families. Although miRNAs miR-2**a** and miR-2**b** are very similar (differing in one or two bases), they are two unique miRNA sequences.

The miRNAs in this category are of special interest, since they were not known to be expressed in *Acyrthosiphon pisum* before. To proceed with further analyses, we focused on the miRNAs identified by all the three methods (see Table 3.5), and with no family member in the pea aphid. A "family member" is a miRNA very similar in sequence; for example, miRNA dme-miR-184 was identified as "known in other species", however, the pea aphid expresses miRNA miR-184b that differs in two bases in relation to dme-miR-184. Using this definition, the only miRNA with no family member in the pea aphid is miR-79 (see Table 3.5).

We now focus on the expression profile of miRNA miR-79: first the 17 precursors giving rise to this miRNA were recovered and then their read coverage was computed. A expression profile is consistent with Dicer and Drosha processing if a few criteria are met. According to MIRBASE, a sequence must meet the criteria below to be annotated with high confidence:

1. At least 10 reads must map with no mismatches to each of the two possible mature

MirinhoPipe miRDeep (53)		MirinhoPipe sRNA-PlAn (43)	miRDeep sRNA-PlAn (51)		In all the three methods (40)
let-7	miR-92a	miR-13a	bantam	miR-92b	miR-13a
miR-10	miR-971	miR-184a	miR-1	miR-971	miR-184a
miR-124	miR-981	miR-184b	miR-137	miR-981	miR-184b
miR-13a	miR-993	miR-190	miR-13a	miR-993	miR-190
miR-14	miR-996	miR-210	miR-184a	miR-9a	miR-263a
miR-184a	miR-998	miR-263a	miR-184b	miR-iab-4	miR-276
miR-184b	miR-9a	miR-276	miR-190		miR-278
miR-190	miR-iab-4	miR-278	miR-263a		miR-279b
miR-263a		miR-279b	miR-275		miR-2a
miR-263b		miR-2a	miR-276		miR-2b
miR-276		miR-2b	miR-277		miR-3015c
miR-278		miR-2c	miR-278		miR-3016
miR-2796		miR-3015c	miR-279b		miR-3017a
miR-279a		miR-3016	miR-281		miR-3018
miR-279b		miR-3017a	miR-29		miR-3019
miR-2a		miR-3018	miR-2a		miR-3024
miR-2b		miR-3019	miR-2b		miR-3026
miR-3015a		miR-3024	miR-3015c		miR-3031
miR-3015c		miR-3026	miR-3016		miR-3032
miR-3016		miR-3031	miR-3017a		miR-3033
miR-3017a		miR-3032	miR-3018		miR-3036
miR-3018		miR-3033	miR-3019		miR-3037
miR-3019		miR-3036	miR-3020		miR-3040
miR-3024		miR-3037	miR-3024		miR-3041
miR-3026		miR-3040	miR-3026		miR-3042
miR-3031		miR-3041	miR-3027		miR-3043
miR-3032		miR-3042	miR-3031		miR-3047
miR-3033		miR-3043	miR-3032		miR-307
miR-3035		miR-3047	miR-3033		miR-315
miR-3036		miR-3050	miR-3036		miR-317
miR-3037		miR-307	miR-3037		miR-87a
miR-3040		miR-315	miR-3040		miR-87b
miR-3041		miR-317	miR-3041		miR-927
miR-3042		miR-87a	miR-3042		miR-929
miR-3043		miR-87b	miR-3043		miR-92a
miR-3047		miR-927	miR-3047		miR-971
miR-3053		miR-929	miR-3051		miR-981
miR-3055		miR-92a	miR-307		miR-993
miR-307		miR-971	miR-315		miR-9a
miR-315		miR-981	miR-317		miR-iab-4
miR-317		miR-993	miR-87a		
miR-87a		miR-9a	miR-87b		
miR-87b		miR-iab-4	miR-927		
miR-927			miR-929		
miR-929			miR-92a		

Table 3.4: *List of known miRNAs in Acyrthosiphon pisum predicted by the combination of each two methods, and by all the three methods. The number in parenthesis represents the miRNAs in that set. The official names of these miRNAs in* miRBase *are all preceded by the prefix "api-".*

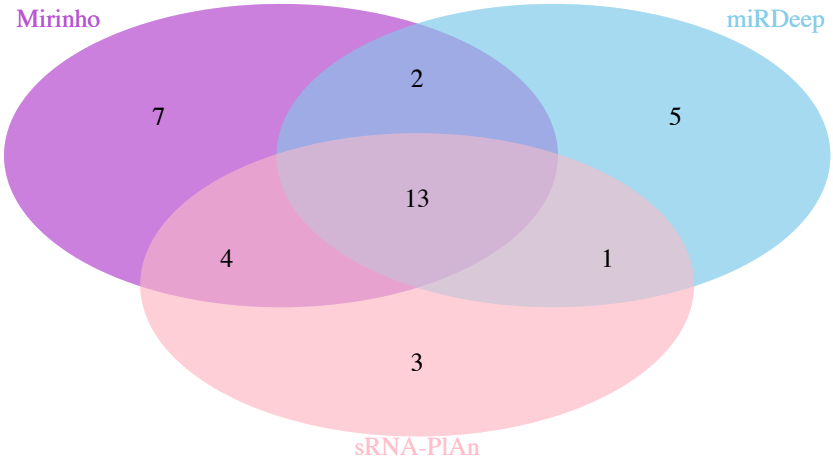


Figure 3.5: *Venn diagram of the miRNAs (known in other species but not (yet) identified in Acyrthosiphon pisum) recovered by* MIRINHOPIPE *(26),* MIRDEEP *(21), and* SRNA-PLAN *(21). The numbers in parenthesis represent the total number of miRNAs in each set.*

miRNA name	No. of species	miRNA family	miRNA family members
miR-263	1	miR-263	api-miR-263a, api-miR-263b
miR-79-5p	**1**	**miR-79**	-
miR-184-3p	20	miR-184	api-miR-184a, api-miR-184b
miR-184	19		
miR-2-3p	2	miR-2	api-miR-2a, api-miR-2b, api-miR-2c
miR-2	8		
miR-87-3p	5	miR-87	api-miR-87a, api-miR-87b
miR-87	5		
miR-9-1	1	miR-9	api-miR-9a, api-miR-9b
miR-9-2	1		
miR-9-3p	2		
miR-9-5p	24		
miR-9	17		

Table 3.5: *List of the miRNA genes known in other species and found in Acyrthosiphon pisum by all the three methods. The number of species expressing the miRNA is given (No. of species), followed by the corresponding miRNA families and their members present in the pea aphid. The line in bold represents the only miRNA, miR-79-5p, with no family member in Acyrthosiphon pisum.*

microRNAs derived from the hairpin precursor.

2. The most abundant reads from each arm of the precursor must pair in the mature microRNA duplex with 0-4nt overhang at their 3' ends.

3. At least 50% of the reads mapping to each arm of the hairpin precursor must have the same 5' end.

4. The predicted hairpin structure must have a folding free energy of <-0.2 kcal/mol/nt.

5. At least 60% of the bases in the mature sequences must be paired in the predicted hairpin structure.

To verify if the criteria applied to the 17 pre-miRNAs, their indexes were first built with Bowtie (2.2.0), and the reads from the four samples EE, IE, LE, and L1 were mapped to the precursors. It is worth noting that the packages of reads used were the ones obtained after trimming out adapters and filtering out reads $<$ 16nt; unique reads were not used because the real expression profile would be "hidden" by the removal of the copies. Our set was then comprised of 34,170,694 reads for EE, 33,195,700 reads for IE, 30,084,568 reads for LE, and 21,732,072 reads for L1.

The EE reads mapped to 13 precursors, the IE reads to 13 too, the LE reads to 11, and the L1 reads to 9 precursors. Filtering out these precursors according to the criteria mentioned above, only two remained: one belonging to contig GL350203 (471,709..471,793) and the other to contig GL349650 (1,158,472..1,158,559). To make it simple, we call mir-79-GL350203 the first precursor that appeared in all the samples, and mir-79-GL349650 the second one that appeared only in sample LE.

The three first criteria are related to the pattern of the mapped reads, while the two last are related to the precursor sequence. Table 3.6 presents the different values for these criteria considering the four samples. The three first criteria apply to sample EE only, while criterion 3 did not apply to the IE, LE, and L1 samples. Furthermore, the EE sample has the largest number of reads aligning to the precursor mir-79-GL350203, as shown in Figure 3.6. To check criteria 4 and 5, which refer to characteristics of the secondary structure, we use Figure 3.7) as a guide. The first precursor mir-79-GL350203 has a secondary structure with a free energy of -27.6 kcal/mol, thus $-27.6/84$nt $= -0.328$ kcal/mol/nt (criterion 4), and 68% of the nucleotides of the miRNA duplex were paired (criterion 5). The second precursor mir-79-GL349650 has a secondary structure with a free energy of -24.52 kcal/mol, so $-24.52/84$nt $= -0.291$ kcal/mol/nt, and 86% of paired nucleotides in the duplex. As one can notice, only precursor mir-79-GL350203 together with the reads of sample EE fulfilled all the criteria.

These facts provide a strong evidence that the miRNA api-miR-79 derives from the precursor mir-79-GL350203 since it verifies all the biological criteria. Moreover, when considering the number of reads mapping to this precursor, the highest stacks are obtained with the reads from sample EE. This means that the expression profile of the miRNA api-miR-79 is more prominent during the early embryo stage of *Acyrthosiphon pisum*. This make us believe that api-miR-79 has an important function in the developmental process of early embryos. Although only the precursor mir-79-GL350203 (more strongly expressed in the EE stage) fulfilled all the criteria, we do not discard the hypothesis that the same precursor is being expressed in other stages, since the great majority of the criteria also applied to the IE, LE and L1 stages. Based on that, we present in Figures 5.4-5.6 (in the Appendix 5.4), the expression profile of precursor mir-79-GL350203, and in Figures 5.7-5.9 expression profile of precursor mir-79-GL349650 during the three remaining stages.

Figure 3.6: *EE reads mapped against the precursor mir-79-GL350203 (miRNA miR-79). For the sake of the presentation, the reads used in this figure were the unique reads of the EE sample because otherwise it would not fit in the page.*

```
>GL350203:471709:471793:premirna_12522:85:+:-27.62:forward
CAATGTTGATCTCTTTGGTACTTTAGCTGTAGGTATATTTTAAAGAGACGCCCTAAAGCTTCTGTACCAATGTTATTGGCAATT
C A       CT T       TTT    G    TATATTTT
 A TGTTGAT  C TTGGTAC   AGCT TAGG        A
 | |||||||  | |||||||   |||| ||||
 T ACGGTTA  G AACCATG   TCGA ATCC        A
T A       TT T       TCT    A    CGCAGAGA
```

```
>GL349650:1158472:1158559:premirna_3079:88:+:-24.52:forward
GCGTTGTGTTATCTGGCTGTTGACTTTTTCCGAAACATTCAGCCTGGTTTTTCGGAAAATCAACGGGCTCGGTGCTGTGAAAAA
-GCG     T    T  -       CT          CATTCA
    TTGTG TATC GG CTGTTGA  TTTTCCGAAA      G
    ||||| |||| || |||||||  ||||||||||
    AGTGT GTGG TC GGCAACT  AAAAGGCTTT      C
AAAA     C    C  G       --          TTGGTC
```

Figure 3.7: *In the upper box is shown the secondary structure of the first precursor mir-79-GL350203 with a free energy of* -27.6 *kcal/mol, thus* $-27.6/84nt = -0.328$ *kcal/mol/nt, and 68% of paired nucleotides in the miRNA duplex; this precursor appeared in all the samples. In the bottom part is shown the secondary structure of the other precursor mir-79-GL349650 with a free energy of* -24.52 *kcal/mol, so* $-24.52/84nt = -0.291$ *kcal/mol/nt, and 86% of paired nucleotides in the duplex; this precursor appeared only in the LE sample. Secondary structures and respective free energies were computed with* MIRINHO.

	EE	IE	LE		L1
	GL350203	GL350203	GL350203	GL349650	GL350203
Criterion 1	4,844 / 263	4,527 / 233	3,626 / 204	3,626 / 204	803 / 65
Criterion 2	2nt / 2nt	2nt / 1nt	2nt / 1nt	2nt / 1nt	3nt / 3nt
Criterion 3	66% / 84%	47% / 86%	46% / 85%	46% / 84%	47% / 82%
Criterion 4	−0.328	−0.328	−0.328	−0.291	−0.328
Criterion 5	68%	68%	68%	86%	68%

Table 3.6: *Summary of the criteria of a high confident precursor, for precursor mir-79, together with the reads from the four samples. The second column (from left to right), for example, represents the EE reads mapped to the precursor mir-79 originated from the contig GL350203 of Acyrthosiphon pisum. The Criterion 1 stands for the number of reads aligning to, respectively, the 5p-arm and the 3p-arm. Criterion 2 refers to the number of overhanging nucleotides in the 3' ends. Criterion 3 is related to the number of reads in each of the arms that have the same end. On the left side of the slash are the values relative to the 5p-arm and on the right side the values relative to the 3p-arm of the precursor. Criterion 4 stands for free energy associated to each nucleotide. Criterion 5 refers to the percentage of mature miRNA, within the hairpin stem, that is paired. The criteria are precisely described in the beginning of this section. Values in gray are the ones that did not reach the minimum threshold for the given criterion.*

Novel precursor microRNAs in *Acyrthosiphon pisum*

Only the potential novel pre-miRNAs retrieved by all the three methods were considered for downstream analysis. To verify if a given pre-miRNA was a consensus among two or three methods, we applied a global sequence alignment (instead of a local as used for known miRNAs), since we are comparing two sequences with similar lengths (pre-miRNA sequences). For that, we used the tool GGSEARCH36 implemented in the FASTA package. A pre-miRNA sequence was considered as a consensus between n methods if the alignment between the sequences outputted by the n different methods had no mismatches. Note that the sequences are not necessarily identical since there can exist gaps. Using this strategy, the number of consensus pre-miRNAs was computed for each two methods and for all the three, as shown in Figure 3.8. As we can see, MirinhoPipe obtains a larger number of potential novel pre-miRNAs because we consider all the mapped regions for prediction, while miRDeep and sRNA-PlAn eliminate unlikely regions before the prediction. This means that MirinhoPipe will also consider low expressed miRNAs while the two other methods will preferentially detect the highly expressed ones. miRDeep uses the pattern of the read stack to constrain the prediction to a smaller region while sRNA-PlAn uses a database of other kinds of RNAs (e.g. ribosomal RNAs) to eliminate "non-miRNAs".

Considering that there is no a priori knowledge about novel pre-miRNAs, an additional criterion was used: only pre-miRNAs holding a pattern of read coverage consistent with the Dicer and Drosha processing were deemed as strong candidates. As pointed out in Kozomara et Griffiths-Jones (2011), a typical pre-miRNA would meet the five criteria mentioned in the previous section: at least 10 mapped reads, 3' end overhangs, 50% of the reads with a same 5' end, a minimum free energy, and 60% of paired bases. We thus selected only the pre-miRNAs respecting these criteria. For that, we used the same approach as for the miRNA miR-79. First, the indexes of the pre-miRNAs were built with Bowtie2 (version 2.2.0) and the reads

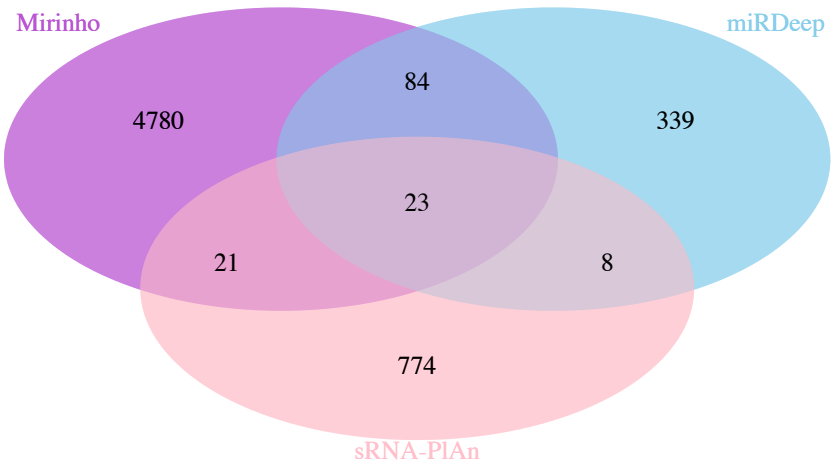


Figure 3.8: *Venn diagram of the novel pre-miRNAs identified by* MIRINHOPIPE *(4908),* MIRDEEP *(454), and* SRNA-PLAN *(826). The numbers in parenthesis represent the total number of pre-miRNAs identified by each method.*

from samples EE, IE, LE, and L1 were mapped to the pre-miRNAs. From the 23 potential novel pre-miRNAs, 14 had EE read mappings, 16 had IE read mappings, 13 had LE read mappings, and 6 had L1 read mappings, all of them with more than 10 reads aligning to each mature miRNA (criterion 1). After applying the other criteria (2-5), 10 precursors remained for EE, 15 for IE, 12 for LE, and 5 for L1, giving a total of 14 unique precursor sequences, as shown in Tables 3.7 and 3.8. The first table presents the developmental stages during which the pre-miRNA is being expressed, while the second table presents the sequence of the precursor miRNA.

Precursor	EE	IE	LE	L1
1		X		
2	X	X	X	
3	X	X		
4	X			
5	X		X	
6			X	
7	X		X	
8	X	X	X	X
9		X		
10	X	X	X	
11				X
12	X	X		
13	X			
14				X

Table 3.7: *The 14 novel precursor-miRNA sequences organised by the developmental stages in which they are expressed. To recover the precursor sequence, see Table 3.8.*

1	ACACGCACGCACGAACACGATCCGTTCGCAGTCGTATTATTCGAGTACGCGAGTGTGAAGCGATCGTGTGCGTGCGCGC
2	AGACTGATAGCAGCGACTGTTACGAGGCCCTGTTTCCTTTGTGCTATTTAGTATACTTATAAGAAACGGGGCCTAGTAACAGTCGCTGCCGTCAGT
3	ATACCAGAATCGAAGTTCTGTGGTAGTGGGCCACTCGAATACAAACAGTGGCTCACAAACACATCACATCATTAATGTATTT
4	CAATGTTGATCTCTTTGGTACTTTAGCTGTAGGTATATTTTAAAGAGACGCCCTAAAGCTTCTGTACCAATGTTATTGGCAATT
5	CGTCAACGTAAACTCGCTTTAAATCCATCTTGAATATAATATTTGAAATTCCAGATAGTTATAAAGCGAGTCTAAGTTGACGAT
6	CTTTTATTTTTGGGTGTTTTTCATCAGGTTAGTAGTGATTATATACATACTACTTGATGAAAAATATCCTAAAATGGAAG
7	CTTTTATTTTTGGGTGCTTTTCATCAGGTTAGTAGTGATTAAATACATACTACTTGATGAAAAATATCCTAAAATGGAAG
8	CTTTTATTTTTGGGTGCTTTTCATCAGGTTAGTAGTGATTTAATACATACTACTTGATGAAAAATATCCTAAAATGGAAG
9	CTTTTATTTTTGGGTGCTTTTCATCAGGTTAGTAGTGATTAAATACAATATACTACTTGATGAAAAATATCCTAAAATGGAAG
10	CTTTTATTTTTGGGTGCTTTTTATCAGGTTAGTAGTGATTAAATGCATACTACTTGATGAAAAATATCCTAAAATGGAAG
11	GATTCAAGCTGTGGTAACTCCAAACCATTGCCGGCGTTTTATTTTGTATCCCGCAATGGTTGGAAGTTCCTCACTTTGGTCACGCAA
12	GTAACTGAGGACATCATTACCTGACAGTATTAGACATATCAATTGTCACTCTAATCCTGCCCAAGTAAGACGTTAACAGTT
13	TCAGGTCGTTACTCCAATATGCCTCCTTCAATGTGTTTTGATAATGTAGGACAGCACATTCAAGGACACATACTGAAGAAAAAAC
14	TTCTCAGGCTGTGATTGTCCAAACGCAATTCTTGTTAAACGTATATATGCAATCAAGGATTGAGTAGGGACGTCAACGCTTGAGACG

Table 3.8: *The 14 collapsed sequences corresponding to the new precursor-miRNAs identified by all the three methods and applying to all the five criteria for a high confidence annotation. The sequences are in a 5' to 3' orientation, and the mature miRNAs are highlighted in red. Although sequences 6-10 appear to be the same, they differ from each other by one or two nucleotides. That is for instance the case of sequences 7 and 8 for which in position 41 (underlined) there is a an "A" for 7 and a "T" for the other sequence.*

3.3.3 MicroRNA gene expression profile

All the previous analyses considered the ensemble of discovered miRNAs disregarding the samples from which they originated. In this section, we focus on the differential miRNA expression in each sample (EE, IE, LE, and L1). It is important to identify the miRNA genes specific to a certain stage to understand which miRNAs are biological determinants in the development of the pea aphid. To identify these genes, the expression profile of the miRNAs found with our methodology was computed with MEV (MultiExperiment Viewer), which uses an unsupervised hierarchical clustering—generated by an average linkage method with euclidean distance and no leaf order optimisation (Howe *et al.*, 2010).

To measure the expression, we first normalised the read counts using Equation 3.1, which is simply the ratio between the number of reads, specific to one sample, that aligned to the given pre-miRNA and the total number of reads that aligned to the genome. These normalised counts are then submitted to MEV, and using the option "HCL", the result is a heatmap of the expression of the miRNAs; the expression profiles are arranged in clusters within a dendogram, as shown in Figure 3.9.

The expression profiles are organised in the heatmap by miRNAs vs. samples: the lines are the different miRNAs found with our methodology, and the columns are the samples (including the replicates) in which these miRNAs were expressed. In Figure 3.9, the clustering shows that our samples can be classified based on the miRNA expression levels: there are four clusters of expression profiles that precisely agree with the four different samples (see the top of the figure). We can also observe that the profiles of the stages IE and LE are clustered together, meaning that they are biologically closer to each other. Moreover, the height of the branches represents the differences between the clusters, i.e., the more distant are the samples the longer are the branches. As one can notice in the dendogram, the branches of cluster L1 are longer, and this reflects the biological conditions in which the samples were obtained. As mentioned before, embryos live inside their progenitor in a stable environment, while L1 larvae are exposed to an *in vivo* environment feeding on plants. The samples from the L1 stage are thus more heterogeneous (longer branches) than the ones extracted from embryos, due to this environmental conditions and/or to their age (varying from 0 to 24 hours after their birth).

When we consider the clusters from the perspective of the miRNAs, using a distance cutoff of 0.396 results in eight clusters, from which four have one single miRNA gene, while the others have, respectively, 3, 10, 14 and 39 miRNAs, as shown in the same Figure 3.9 (left part of the dendogram in blue). The miRNAs within each of the different clusters may be regulating a specific gene (or a specific set of genes), since synergism is known to play an effective role in the regulation (Xu *et al.*, 2011; Lutter *et al.*, 2010). We will confirm this with the expression and functional analysis of the predicted targets.

3.3.4 MicroRNA target prediction

The three sets of miRNAs, comprised of 40 known miRNAs, 1 known miRNA in other species but not present in the pea aphid, and 14 novel miRNAs, were submitted to the prediction of targets by two methods, PITA and MIRANDA (see Section 1.2.4 for an introduction of these softwares). We set an energy threshold of -10 kcal/mol for MIRANDA, while for PITA, a negative $\Delta\Delta G$ was required. As mentioned in Section 3.2.6, the accessibility is measured by the difference between the free energy gained from the formation of the microRNA-target duplex and the energetic cost of unpairing the folded target to make it accessible to the microRNA ($\Delta\Delta G = \Delta G_{duplex} - \Delta G_{open}$). The ideal situation would be a small cost to open

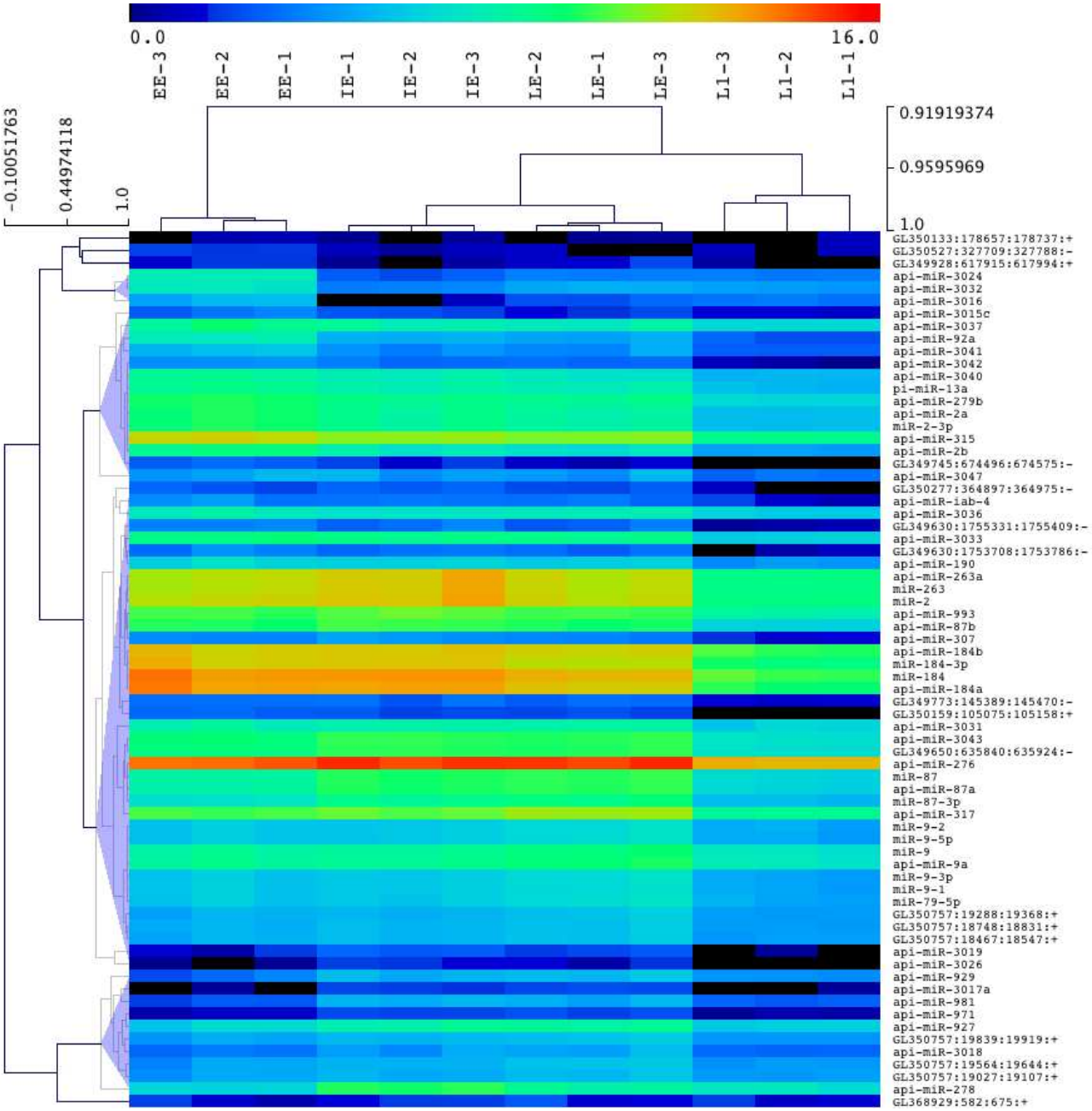


Figure 3.9: *An unsupervised hierarchical clustering, generated by an average linkage method with euclidean distance and no leaf order optimisation, of the number of reads mapping to each one of the 81 identified miRNAs. The colour chart indicates expression intensities using a base 2 logarithmic scale: blue and red represent, respectively, lower (2.0) and upper (16) expression intensities. This expression profile was computed with* MeV *(MultiExperiment Viewer) (Howe* et al., *2010).*

the target structure, combined with a strong base pairing between the duplex (more negative energy). Based on that, we set the threshold to $\Delta\Delta G$ to be at least negative. Table 3.9 presents the total number of interactions found for each *Acyrthosiphon pisum* miRNA by the two methods, and the number of common predictions between both.

To guarantee a more reliable set of interactions, we used a similar strategy as the one used for finding miRNAs: only those interactions predicted by the two methods were kept. While MIRANDA and PITA separately predicted respectively 80,191 and 1,408,997; 68,787 interactions were found by both methods for the 40 pea aphid miRNAs. Considering the only miRNA known in other species, api-miR-79, MIRANDA predicted 1,336 and PITA 8,967, while the overlap consisted in 980 interactions for this miRNA. For the 23 potential novel miRNAs, miRanda found 358,360 interactions and PITA 2,217,052, with 204,163 in common.

Known miRNAs in *Acyrthosiphon pisum*	miRanda	PITA	Overlap
api-miR-13a_MIMAT0014713_Acyrthosiphon_pisum_miR-13a	1665	10751	1474
api-miR-184a_MIMAT0014132_Acyrthosiphon_pisum_miR-184a	742	6186	637
api-miR-184b_MIMAT0014715_Acyrthosiphon_pisum_miR-184b	789	5932	656
api-miR-190_MIMAT0014127_Acyrthosiphon_pisum_miR-190	7855	58837	6011
api-miR-263a_MIMAT0014718_Acyrthosiphon_pisum_miR-263a	1376	12086	1211
api-miR-276_MIMAT0014133_Acyrthosiphon_pisum_miR-276	1027	14450	961
api-miR-278_MIMAT0014723_Acyrthosiphon_pisum_miR-278	507	7376	443
api-miR-279b_MIMAT0014725_Acyrthosiphon_pisum_miR-279b	963	12465	894
api-miR-2a_MIMAT0014727_Acyrthosiphon_pisum_miR-2a	1936	12685	1693
api-miR-2b_MIMAT0014821_Acyrthosiphon_pisum_miR-2b	2225	13122	1915
api-miR-3015c_MIMAT0014794_Acyrthosiphon_pisum_miR-3015c	1471	9401	1308
api-miR-3016_MIMAT0014751_Acyrthosiphon_pisum_miR-3016	2495	24546	2201
api-miR-3017a_MIMAT0014752_Acyrthosiphon_pisum_miR-3017a	547	8431	507
api-miR-3018_MIMAT0014754_Acyrthosiphon_pisum_miR-3018	5259	38274	4293
api-miR-3019_MIMAT0014755_Acyrthosiphon_pisum_miR-3019	1487	10046	1373
api-miR-3024_MIMAT0014760_Acyrthosiphon_pisum_miR-3024	1638	14786	1424
api-miR-3026_MIMAT0014762_Acyrthosiphon_pisum_miR-3026	691	8022	615
api-miR-3031_MIMAT0014767_Acyrthosiphon_pisum_miR-3031	2206	26479	1992
api-miR-3032_MIMAT0014768_Acyrthosiphon_pisum_miR-3032	3833	39125	3267
api-miR-3033_MIMAT0014769_Acyrthosiphon_pisum_miR-3033	543	3982	452
api-miR-3036_MIMAT0014773_Acyrthosiphon_pisum_miR-3036	2166	28761	1917
api-miR-3037_MIMAT0014774_Acyrthosiphon_pisum_miR-3037	3067	16039	2665
api-miR-3040_MIMAT0014778_Acyrthosiphon_pisum_miR-3040	700	5874	585
api-miR-3041_MIMAT0014779_Acyrthosiphon_pisum_miR-3041	2007	26757	1832
api-miR-3042_MIMAT0014780_Acyrthosiphon_pisum_miR-3042	612	8371	537
api-miR-3043_MIMAT0014781_Acyrthosiphon_pisum_miR-3043	5151	22135	4135
api-miR-3047_MIMAT0014786_Acyrthosiphon_pisum_miR-3047	1250	13014	1177
api-miR-307_MIMAT0014729_Acyrthosiphon_pisum_miR-307	645	4519	574
api-miR-315_MIMAT0014730_Acyrthosiphon_pisum_miR-315	5663	49567	4595
api-miR-317_MIMAT0014732_Acyrthosiphon_pisum_miR-317	2007	12888	1740
api-miR-87a_MIMAT0014739_Acyrthosiphon_pisum_miR-87a	2051	29023	1829
api-miR-87b_MIMAT0014738_Acyrthosiphon_pisum_miR-87b	1631	27591	1471
api-miR-927_MIMAT0014740_Acyrthosiphon_pisum_miR-927	1525	28680	1430
api-miR-929_MIMAT0014741_Acyrthosiphon_pisum_miR-929	1812	27343	1674
api-miR-92a_MIMAT0014742_Acyrthosiphon_pisum_miR-92a	1342	21341	1243
api-miR-971_MIMAT0014745_Acyrthosiphon_pisum_miR-971	1924	20095	1695
api-miR-981_MIMAT0014736_Acyrthosiphon_pisum_miR-981	1435	14583	1280
api-miR-993_MIMAT0014135_Acyrthosiphon_pisum_miR-993	496	8033	468
api-miR-9a_MIMAT0014748_Acyrthosiphon_pisum_miR-9a	2868	25384	2435
api-miR-iab-4_MIMAT0014129_Acyrthosiphon_pisum_miR-iab-4	2584	15768	2178
Total	80191	742748	68787

Table 3.9: *Number of target interactions found for the 40 miRNAs known in Acyrthosiphon pisum, using the methods for target prediction* Pita *and* miRanda.

3.4 Conclusion

The first result of this work is a pipeline for the analysis of small RNA sequencing data. We preferred to develop a more flexible method, in which all the steps may be adjusted according to the data, rather than completely automating the process, and leave it like a "black box". The documentation of MIRINHOPIPE is available at http://mirinho.gforge.inria.fr/mirinhopipe.html.

The combination of MIRINHOPIPE, together with sRNA-PLAN and miRDEEP, allowed us to analyse, for the first time, the miRNAs on the pea aphid parthenogenesis, that revealed several novel miRNAs with potential to play key roles in the transcription regulation during the development of this insect.

From the miRNAs discovered in this work, forty were known in *Acyrthosiphon pisum*. Two among these, api-miR-iab-4 and api-miR-281, were annotated with an evidence of "by similarity" in miRBase. We thus suggest that their annotation should be changed to "experimental". We found a miRNA, api-miR-79, that was not known to be expressed in the pea aphid, mainly during the early embryo developmental stage, suggesting that it may play an important role at such stage. We do not discard the possibility that api-miR-79 is also expressed during the other stages, IE, LE, and L1, since it was detected during these stages; however, the expression pattern did not fulfill all the five criteria for a (pre-)miRNA high confidence annotation (Kozomara et Griffiths-Jones, 2011). Twenty-three further potentially novel (pre-)miRNAs were found in our data, out of which 14 were annotated with high confidence (based on the criteria mentioned above). A few were specific to certain stage(s), while others were common to the four stages.

A clustering of the normalised expression profiles of the detected miRNAs allowed to verify the quality of the samples. Those belonging to a same stage remained clustered in sub-groups, while the samples obtained during closer stages (i.e., IE and LE) were also found within a same sub-group. The sample clusters are a reflection of development and this result indicates indirectly an important role of miRNAs in the pea aphid development.

Target prediction using two methods (miRanda and Pita), resulted in 68,787 interactions between the 40 pea aphid miRNAs and the 3'UTR sequences; 980 interactions between the miRNA api-miR-79 and its putative targets; and finally 204,163 interactions for the 23 potentially novel miRNAs.

The study of miRNAs showing differential expression in different stages, and a more detailed analysis of the predicted targets (including a comparison with the mRNA microarray based profiles (Rabatel *et al.*, 2013), will allow to characterise the underlying regulatory network.

Chapter 4

Prediction of non-coding RNAs and targets in *Mycoplasma hyopneumoniae*

Contents

4.1 Introduction 70
4.2 Material and methods 70
4.2.1 Prediction of non-coding RNAs 70
4.2.2 Alvinho: An algorithm for the prediction of non-coding RNA targets 74
4.2.3 Conservation analysis 75
4.3 Results and discussion 77
4.3.1 Identified ncRNA candidates 77
4.3.2 Predicted non-coding RNA targets 77
4.3.3 Conserved ncRNAs 83
4.4 Conclusion 83

This chapter is strongly based on the paper Godinho *et al.* (tion). In the context of a collaboration, made possible through LIRIO, an International Associated Laboratory (LIA – Laboratoire International Associé) between the LBBE-UMR5558, and notably the BAMBOO-BAOBAB team (head Dr Marie-France Sagot), and the Laboratório de Bioinformática of the LNCC/MCT, Brazil (head Dr Ana Tereza Vasconcelos), the Master student Caio Padoan de Sá Godinho came to Lyon to work on problems related to the prediction of non-coding RNA (ncRNA) in *Mycoplasma hyopneumoniae* (that was the main topic of his project). The problems included the segmentation of numerical sequences that represented a predicted ncRNA, the prediction of ncRNA targets, and the analysis of conservation between intergenic regions in Mycoplasmas. Although the problems were not directly related to the regulation in eukaryotes, as miRNAs are, it is important to understand how the regulation in bacteria works since one of the perspectives for future works is to understand the regulatory interactions between the partners in a symbiotic relationship, for example the interaction model between the bacterium *Buchnera aphidicola* and its eukaryotic host *Acyrthosiphon pisum* (the pea aphid), and between the bacterium *Mycoplasma hyopneumoniae* and its host *Suc scrofa* (the swine). It is this latter case that interested Caio in his Master, and that will therefore concern us also in this chapter.

4.1 Introduction

The bacterium *Mycoplasma hyopneumoniae* strain 7448 is a pathogenic and obligate parasite of porcine respiratory systems. It lives adhered to the epithelium of its host respiratory tract, and together with other bacteria and viruses, it is considered one of the ethiologic agents of swine enzootic pneumonia. The disease can cause a decrease in the productivity of these animals, sometimes resulting in their death (BYRT *et al.*, 1985; DeBey et Ross, 1994; Brockmeier *et al.*, 2002).

Although some effort has already being put on understanding the infection process, the specific mechanisms relating the bacterium and the disease remain unknown. Between the different sequenced strains of the same species, only *Mycoplasma hyopneumoniae* J (ATCC 25934) was deemed non-pathogenic (Gardner et Minion, 2010; Hsu et Minion, 1998; Nicolás *et al.*, 2007; Siqueira *et al.*, 2013).

Mycoplasma hyopneumoniae 7448 has only one known transcription factor (TF) and a complex gene expression pattern. The incomparability between the number of regulatory elements and the complexity of the gene expression of the bacterium, together with increasing evidences that ncRNAs are involved in this phenomenon, strongly encourage the search for ncRNAs in the genome of *Mycoplasma hyopneumoniae* 7448.

After predicting the regions with a potential to harbour ncRNA genes, additional analyses were performed in an attempt to provide more evidences to carry on with experimental validation of the ncRNAs. The first problem was related to the output of the pipeline for the prediction of ncRNAs: the pipeline was generating one single assembled ncRNA sequence where two or more different ncRNA candidates were in fact present. We solved this by applying a segmentation algorithm on these outputs. To then provide stronger evidence that the candidates were indeed functional, we performed the prediction of the ncRNA targets with a method, called Alvinho, that was specially developed for this purpose. Finally, to verify if conservation could play any role in the functionality of ncRNAs, the identity of intergenic regions was assessed between closely-related *Mycoplasma* species by means of a k-partite graph. Genomic motifs surrounding the ncRNA, such as promoters and terminators, were also verified to reinforce the functional evidence of the ncRNA candidates. All the three steps of the pipeline are available in the form of a script or a C++ implementation.

4.2 Material and methods

4.2.1 Prediction of non-coding RNAs

The main component of the pipeline for the prediction of ncRNAs is a method called Single Genome ncRNA Search (SIGRS), developed by Larsson *et al.* (2008). The method uses a set of known ncRNAs, provided by the user, to guide the search for new ncRNAs with a similar nucleotide composition profile. If an annotation file is also provided, the coding regions of the genome are masked according to the annotation of known genes, and the search is concentrated in smaller regions with lesser noise.

In this work, we used a set of 816 ncRNAs from species of the class Mollicutes (the class to which the bacterium *Mycoplasma hyopneumoniae* 7448 belongs to) and the gene annotation of the organism. To focus the search in a more specific space, all the regions containing an annotated gene were masked, that is, the nucleotides in these positions were replaced by X's. The ncRNA sequences were obtained at the Bacterial Small RNA Database (BSRD) (Li *et al.*, 2013) and the annotation file at the NCBI. The set of known ncRNAs and the masked genome

are provided to SIGRS, which creates a scoring system, based on the nucleotide composition of the known ncRNAs, to transform the genome in a numerical sequence. The segments with a high cumulative sum are thus considered as an ncRNA candidate.

Scoring system

In (SIGRS) (Larsson *et al.*, 2008), a scoring system is first built and then evaluated. It considers the dependency between two consecutive nucleotides. The nucleotides of a given sequence are said to be independent if the corresponding dinucleotide frequency does not differ significantly from the one generated by chance; the test used to compute it is called G-test (Zar *et al.*, 1999). The scoring system is computed according to the result of the G-test: either the nucleotides are independent, and in this case a model $\mathcal{M}0$ is used to compute the scores, or the nucleotides are dependent on the preceding adjacent neighbour, and as a consequence a model $\mathcal{M}1$ is used.

To build the scoring system, the frequencies of the (di)nucleotides of both the ncRNA sequences and the masked genome sequences must be computed. We denote by f_α the frequency of a 1-mer word α, with $\alpha \in \mathcal{N} = \{A, T, C, G\}$. The same notation stands for a 2-mer word $\alpha\beta \in \mathcal{N} \times \mathcal{N} = \{AA, AC, AG, \ldots, TT\}$, for which the frequency is represented by $f_{\alpha\beta}$.

These frequencies are then used in SIGRS for the construction of a stochastic model that enables to compute the score. The random variable X_t represents an element of $\mathcal{N}$ at time t, with probability $P(X_t = \alpha)$, with $\alpha \in \mathcal{N}$. At every instant t, the element X_t is concatenated with its preceding elements $X_0X_1X_2 \ldots X_{t-1}$, thus forming a chain of nucleotides. To compute the probability $P(X_t = \alpha)$, two stochastic models may be used, $\mathcal{M}0$ and $\mathcal{M}1$. Model $\mathcal{M}0$ assumes that the probabilities $P(X_t = \alpha)$ are constant and do not depend on earlier events. The vector $\mathbf{p}(\alpha) = [p_A, p_C, p_T, p_G]$ is then defined and the chain of nucleotides can be built in an iterative manner. For the model $\mathcal{M}0$, it is clear that the transition probability p associated with state α is simply equal to f_α. The scores s are thus assigned to each state α as shown in Equation 4.1:

$$s_\alpha = 10\, log_2 \left(\frac{p_\alpha^{nc}}{p_\alpha^{gf}} \right) = 10\, log_2 \left(\frac{f_\alpha^{nc}}{f_\alpha^{gf}} \right). \tag{4.1}$$

where f_α^{nc} is the frequency of the nucleotide α in the ncRNA sequence and f_α^{gf} is the frequency of the same nucleotide in the genome sequence. Figure 4.1 shows an example for model $\mathcal{M}0$.

Time	Sequence	States α	p(α)
0	*T*	A	0.2
1	*TA*	T	0.4
2	*TAC*	C	0.3
3	*TACT*	G	0.1
4	*TACTG*		

Figure 4.1: *An example of the stochastic model $\mathcal{M}0$ for the construction of a nucleotide sequence.*

As for model $\mathcal{M}1$, the probabilities $P(X_t = \alpha)$ are conditioned to the previous event X_{t-1}, i.e., $P(X_t = \alpha | X_{t-1} = \beta)$. This is exactly the Markov property, which is described in Equation 4.2:

$$P(X_t = \alpha | X_{t-1} = \beta, X_{t-2} = \gamma, \ldots, X_0 = \omega) = P(X_t = \alpha | X_{t-1} = \beta). \tag{4.2}$$

These additional criteria also apply for the definition of a Markov chain:

(*i*) The initial probabilities of the states $P(X_0 = \alpha) = p_\alpha$ are given by the vector $\mathbf{p}(\alpha)$, to all $\alpha \in \mathcal{N}$;

(*ii*) The conditional probabilities of all other states $P(X_t = \alpha | X_{t-1} = \beta) = T_{\beta\alpha}$, $t > 0$ are determined by the transition matrix $\mathbf{T}(\beta, \alpha)$, for all $(\beta, \alpha) \in \mathcal{N} \times \mathcal{N}$;

(*iii*) The sum of all probabilities of the same conditional status should result in $\sum_{\alpha \in \mathcal{N}} T_{\beta\alpha} = 1$.

For the model $\mathcal{M}1$, the transition probabilities $T_{\alpha\beta}$ are also related to $f_{\alpha\beta}$, however, they are normalised with respect to f_α. The scores $s_{\alpha\beta}$ are therefore computed using Equation 4.3.

$$s_{\alpha\beta} = 10\, log_2 \left(\frac{T^{nc}_{\alpha\beta}}{T^{gf}_{\alpha\beta}} \right) = 10\, log_2 \left(\frac{f^{nc}_{\alpha\beta} / f^{nc}_{\alpha}}{f^{gf}_{\alpha\beta} / f^{gf}_{\alpha}} \right). \tag{4.3}$$

It is worth noting that the scoring system is built in such a way that: (i) at least one value of s is positive and the transition probability p is not null; (ii) positive scores are assigned to profiles similar to known ncRNAs, and negative scores are assigned otherwise; and (iii) the average of the scores is negative. A formalism of the previous can be found in Karlin et Altschul (1990); Karlin et Dembo (1992). An example of scoring S^0 and S^1, generated respectively by the models $\mathcal{M}0$ and $\mathcal{M}1$, is presented in Figure 4.2.

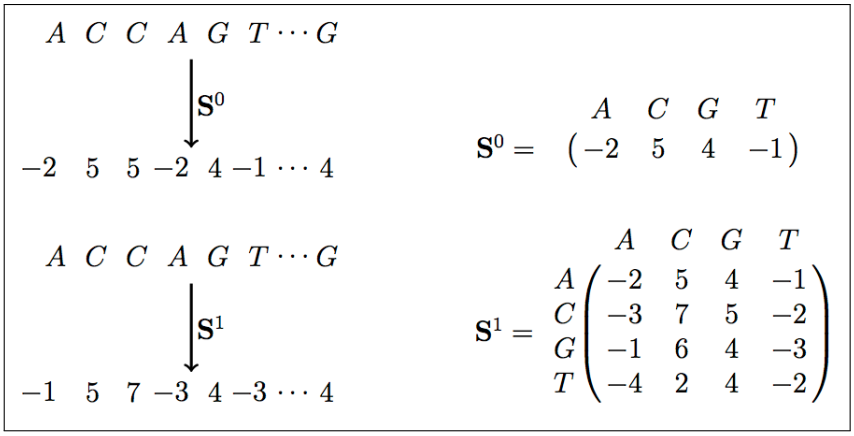


Figure 4.2: *An example of the transformation of a nucleotide sequence into a numerical sequence. The system at the top of the figure assumes that the nucleotides are independent (model $\mathcal{M}0$) from adjacent neighbours. Each nucleotide is replaced by a score according to the scoring system S^0. The system at the bottom of the figure assumes that the nucleotides are dependent (model $\mathcal{M}1$) and the scoring scheme S^1 is used instead.*

Computing the scores of ncRNA candidates

Once the query genome is converted to a numerical sequence Λ of length n, SIGRS can identify the subsequences with high cumulative score, which is obtained by the partial sum H_i^j:

$$H_i^j = \sum_{x=i}^{j} \Lambda_x, \qquad 0 \leq i \leq j \leq n, \tag{4.4}$$

where Λ_x is the score in position x of the sequence; variable H_i^j then provides the cumulative score of the region from positions i to j.

To determine a high representative value for a partial sum, a probability density function (PDF) is used. A PD function gives the probability to find a sequence with a larger or equal score to the one generated randomly, given the score S and the associated transition probabilities p. This probability is represented by a classical measure called e-value.

Segmentation of the outputs of SIGRS One problem with this scoring system is that, instead of outputting two regions (for instance 0-2800 and 3400-5000 in Figure 4.3) each associated with a distinct ncRNA candidate, it considered the whole region (in this case 0-5000) as a unique ncRNA candidate, which would provide the wrong answer. One solution to this problem requires the segmentation of the numerical sequence that represents the ncRNA candidate in order to identify the largest local slopes in a given sequence. In order to do this, the algorithm by Kadane (Bentley, 1984) designed to identify the largest cumulative sums, was adapted to find these slopes, thereby allowing for the fragmentation of the output into the correct number of candidates. The adapted algorithm is defined in what follows.

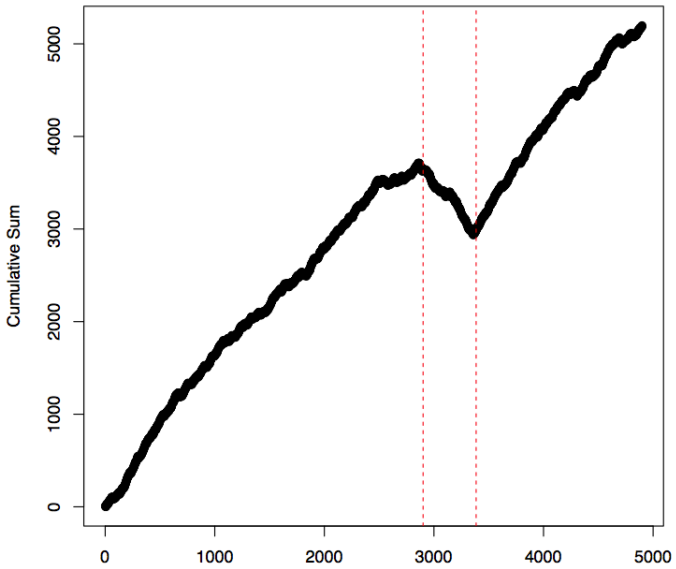


Figure 4.3: *Scoring system of* SIGRS*: the genome coordinates are presented in the x-axis, and the cumulative sum of the scoring is shown in the y-axis. The striped red lines represent the largest local slope.*

Let Λ be a numerical sequence, with $\Lambda_i \in \mathbb{R}$, which represents the scoring system shown above. It is easy to see that the product $k \times \sum_{i=0}^{n} \Lambda_i$ may represent the most negative score a sequence can reach, where n is the length of the numerical sequence and k a parameter optimised as follows. To measure how the different values of k modified the segmentation, we used two other parameters related to the proportion of nucleotides in accordance with known ncRNAs. The first parameter α is associated to the incorrect rejection of a true nucleotide (i.e., a nucleotide that should be in the ncRNA sequence and was discarded); in statistical

hypothesis testing it is called "type I error". The second parameter β is related to the failure to reject a false nucleotide (i.e., a nucleotide that should not be in the ncRNA sequence, but it was not discarded); this measure is called "type II error". It is clear that the ideal situation is to commit no error, i.e., $\alpha = \beta = 0$. To thus choose the best k, we varied its value from $(0, 1]$ with a spacer of 0.01, and at each interaction the euclidean distance to the point $(0, 0)$, that represents "no error", was computed.

4.2.2 ALVINHO: An algorithm for the prediction of non-coding RNA targets

ALVINHO was initially developed for the prediction of miRNA targets, which in turn was inspired in MIRINHO (see Section 2) (Higashi *et al.*, ress). Since base pair interaction is a common characteristic of the regulatory system of both eukaryotes and prokaryotes, the software was adapted to detect base pair interactions in bacterial systems. Although it may seem a much too simplified approach, this is the only characteristic that is precisely known and well defined. Moreover, in bacteria the regulation mediated by small RNAs is in general performed by different mechanisms involving proteins that are specific to certain types of bacteria, such as the RNA chaperone Hfq that is present only in gram-negative bacteria (which is not the case of *Mycoplasma hyopneumoniae* 7448) (Storz *et al.*, 2011). CRISPR (clustered regularly interspaced short palindromic repeats) is another mechanism of regulation in bacteria; however, this mechanism has not yet been described in *Mycoplasma hyopneumoniae* 7448 (Hale *et al.*, 2009). Based on these facts, we decided, at least in a first step, to use base pair interaction as the only feature.

As mentioned before, ALVINHO is based on MIRINHO, and therefore, for the sake of concision, we will focus only on the differences between the two methods. The main one is that MIRINHO is designed to compute the free energy of one single sequence that folds with itself (i.e., the free energy of hybridising the stem-arms of a hairpin) while in the case of target identification, we are dealing with the interaction between two different sequences (i.e. the ncRNA sequence, and the mRNA target sequence).

As a consequence, the alignment algorithm used in MIRINHO has to be modified leading to a new one in ALVINHO. In the case of MIRINHO, the aligned sequences (the two stem-arms) have the same length, and a global alignment must be used. When the aligned sequences are of different lengths, which is the case for targets, a local alignment is applied. Concerning the algorithmic aspect, there are two main differences between these two approaches: (i) the base conditions; and (ii) the starting and ending point of the backtracking step. The base condition of a local alignment is presented in Equations 4.5 and 4.6 below, and the base case for a global one is presented in Equation 1.3.

$$W(i, 0) = \sum_{k=0}^{i} \gamma \tag{4.5}$$

$$W(0, j) = \sum_{k=0}^{j} \gamma \tag{4.6}$$

where γ is the penalty for gaps. The recurrence of the local alignment is the same as for a global one (see Equation 1.4).

Given a DP matrix of size $m \times n$, to recover a local alignment, the starting point in the backtracking step is the largest value in the row $i = m - 1$ and the ending point is the first cell with a zero value $W(i, j) = 0$ or $W(1, 1)$. The starting point of a global alignment is the cell $W(m - 1, n - 1)$ and the ending point is necessarily the cell $W(1, 1)$.

Once the local alignment applied, the same nearest neighbour energy model can be used over each two consecutive base pairs to recover the corresponding free energy; all the details concerning the energy model is presented in Section 1.2.2. As concerns the energy model, there is one small difference between the one used here and the one used for a global alignment. Since the hybridisation is given between two different sequences, here the energy model does not account for the symmetry correction for self complementary duplexes ΔG_{sym} in Equation 1.8. The implementation of ALVINHO produces an output like the one shown in Figure 4.4. It is available at https://sourceforge.net/p/alvinho/code/ci/master/tree/.

```
5'- ncRNA -3':  SIGRS_34 (1,10)
3'- Target -5':  AAZ53378.2|dnaA|chromosomal (188,197)
Energy:  -9.8 kcal/mol
GAGGAAAGT
|||||||||
TTTTTTTCA
```

Figure 4.4: *An example of an output produced by* ALVINHO. *The base pair interaction between the ncRNA SIGRS_34 and the mRNA target sequence AAZ53378.2 has an hybridisation energy of -9.3 kcal/mol.*

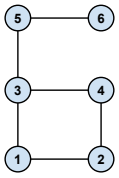
4.2.3 Conservation analysis

To verify if conserved ncRNAs were more susceptible to be functional in *Mycoplasma hyopneumoniae* 7448, a conservation analysis was performed.

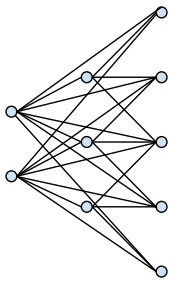
Before describing how this was done, a few definitions are necessary. A graph $G = (V, E)$ composed of a set V of vertices and a set E of edges is said to be *undirected* if the edges have no direction, that is, the relations between pairs of adjacent vertices are symmetric. Two vertices are said to be *adjacent* if there is an edge connecting them. A *k-partite* graph $G = (V, E)$ is a graph whose vertices can be decomposed into k disjoint sets so that a pair of vertices is adjacent if and only if the two vertices belong to two different sets. A *clique* in an undirected graph $G = (V, E)$ is a subset C of V such that the subgraph G' of G induced by C is complete, that is, for every two vertices in C, there exists an edge connecting them in G' (and thus in G). A k-partite graph may have cliques of size at most k (i.e., having k vertices). Figures 4.5a, 4.5b, and 4.5c show examples of respectively a graph, a k-partite graph, and a clique in a graph.

To identify the conserved ncRNAs, four species were considered: *Mycoplasma hyopneumoniae* 7448, *Mycoplasma hyorhinis* HUB1, *Mycoplasma synoviae* 53, and*Mycoplasma agalactiae* PG2. The set of intergenic regions (IGRs) where the ncRNAs may be found in each species was composed of, respectively, 567, 518, 511, and 630 IGRs. Each of the four IGR sets represents one subset (one partition) of the set of vertices of the k-partite graph; in this case $k = 4$. Two vertices u and v belonging to two different subsets (partitions) are adjacent, if and only if an identity of $I(u, v) > 70\%$ between the two IGR sequences labelling the vertices was verified. To compute the identity between the sequences, BLAST was used (Altschul *et al.*, 1990). A sequence was considered as conserved if and only if a clique of size four was associated to it. The algorithm of Bron et Kerbosch (1973) was used to list all the cliques of size $k = 4$.

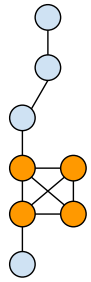
The authors first define three sets that are essential for the core algorithm: (i) the set



(a) *An example of an undirected graph.*



(b) *An example of a k-partite graph, in this example* $k = 3$.



(c) *Example of a clique of size* $n = 4$ *(in orange).*

Figure 4.5: *Illustrated graph concepts.*

compsub is the set to be extended by a new vertex or shrunk by one vertex on travelling along a path of the graph. The points that are eligible to extend *compsub*, i.e., that are connected to all points in *compsub*, are collected recursively in the remaining two sets; (ii) the set *candidates* contains all vertices that will in due time serve as an extension to the current configuration of *compsub*; and (iii) the set *not* is the set of all vertices that have at an earlier stage already served as an extension of the current configuration of *compsub* and are now explicitly excluded. The algorithm generates all the extensions of a given configuration of *compsub* using the elements in the set *candidates* that are not contained in the set *not*. The algorithm can be summarised in five steps:

1. Select a vertex candidate.
2. Add the candidate to *compsub*.
3. Create new sets *candidates* and *not* from the old sets by removing all the vertices not connected to the selected candidate, keeping the old sets intact.
4. Call the extension operator to extends the formed sets.
5. Upon returning, remove the selected candidate from *compsub* and add it to the old set *not*.

A necessary condition to have a clique is that the set *candidates* be empty, otherwise *compsub* could still be extended. This condition, however, is not sufficient, because if *not* is non-empty, the current configuration of *compsub* is contained in another and is therefore not maximal. *compsub* is thus a clique as soon as both *not* and *candidates* are empty.

4.3 Results and discussion

4.3.1 Identified ncRNA candidates

After applying SIGRS to the genome of *Mycoplasma hyopneumoniae* 7448, and segmenting the output when it was necessary, 48 regions susceptible of harbouring ncRNA genes were identified. From these 48 regions, 36 resulted from the segmentation process of 25 potential new ncRNAs; the remaining 12 regions were known ncRNAs. Table 4.1 presents the characteristics of the 48 ncRNA candidates, including the ncRNA identifier, start, end, and length of the ncRNA sequence, the GC content, the free energy of the secondary structure of the ncRNA, and the strand from which it originated. The free energy of a folded sequence is the sum of the energies associated to each base base; frequently, the methods implement an approach to minimise this energy, since the most negative this energy, the more stable a molecule is. The free energy of the sequences were computed with RNAFOLD and were normalised by the length of each sequence.

4.3.2 Predicted non-coding RNA targets

The interaction between a ncRNA and its target may occur mainly in the UTR region; however, a fewer cases have also been observed in the coding regions. Based on this, the whole CDS was considered for target prediction. For each annotated gene, a flanking portion of 150nt downstream of the start codon and 50nt upstream of the stop codon were taken into account. The 48 putative ncRNAs together with the 698 annotated genes were then given as input to ALVINHO. From the outputted interactions, only the best ones were considered, that is, for each ncRNA the interaction with the most negative free energy was taken.

From these interactions, 41,7% are associated to proteins annotated as hypothetical; Tables 4.3-4.5 present the characteristics of the identified interactions. This large percentage of hypothetical proteins agrees with the number of annotated genes (294) with the same classification, that is 42,12% (294/698) of the annotated genes. The proteins are related to the following biological functions: hybridisation, RNA translation, ABC transporters, carbohydrate metabolism, adhesins, and lipoproteins. All of these biological functions are of extreme importance to the survival of *Mycoplasma hyopneumoniae*, and some of them can be directly related to its pathogenicity, such as adhesins. Adhesins are cell-surface components of a bacterium that facilitate adhesion to other cells. The regulation of adhesins is thus very susceptible to be related to the process of infection of the bacterium in the swine (Madsen *et al.*, 2008). Lipoproteins are known to be related to the immune evasion system in the swine (Kelesidis, 2014). The regulation of lipoproteins may then be relevant to pathogenicity (Razin, 2006). These results sustain the hypothesis of the existence of ncRNAs as regulatory elements in the studied bacterium with fundamental roles in its survival and pathogenesis.

ID	Start	End	%GC	Length (nt)	ΔG (kcal/mol)	Strand
SIGRS_34	138331	138427	47,42	97	-0,229	−
SIGRS_82	139192	139291	33	100	-0,191	−
SIGRS_80	139364	139448	38,82	85	-0,136	−
SIGRS_36	139841	140042	41,09	202	-0,222	+
SIGRS_35	220605	220672	42,65	68	-0,224	−
SIGRS_15	303056	303334	39,43	279	-0,248	−
SIGRS_26	353281	353424	45,83	144	-0,162	−
SIGRS_3	353927	354072	44,52	146	-0,114	−
SIGRS_59	354179	354261	38,55	83	-0,222	−
SIGRS_30	371715	371810	39,58	96	-0,182	−
SIGRS_20	388732	388832	43,56	101	-0,144	+
SIGRS_69	389273	389361	38,2	89	-0,175	−
SIGRS_66	389384	389513	33,85	130	-0,139	+
SIGRS_6	389727	389789	55,55	63	-0,287	+
SIGRS_12	407854	407952	38,38	99	-0,192	+
SIGRS_38	407972	408050	39,24	79	-0,125	+
SIGRS_72	421281	421362	41,46	82	-0,233	+
SIGRS_16	427966	428039	39,19	74	-0,143	−
SIGRS_8	428042	428151	42,73	110	-0,141	+
SIGRS_18	434330	434389	53,33	60	-0,29	−
SIGRS_23	488344	488437	43,62	94	-0,185	+
SIGRS_40	488512	488582	36,62	71	-0,075	+
SIGRS_17	489408	489488	45,68	81	-0,202	+
SIGRS_31	515596	515739	38,89	144	-0,177	+
SIGRS_14	516015	516096	53,66	82	-0,257	−
SIGRS_75	516107	516212	34,91	106	-0,177	−
SIGRS_33	516286	516705	0,35	420	-0,183	−
SIGRS_11	517984	518076	45,61	93	-0,266	+
SIGRS_100	523316	523410	31,58	95	-0,205	+
SIGRS_118	523517	523617	32,67	101	-0,126	−
SIGRS_27	569512	569661	38	150	-0,173	−
SIGRS_7	570424	570484	49,18	61	-0,251	+
SIGRS_5	571166	571236	46,48	71	-0,168	−
SIGRS_43	574204	574274	39,44	71	-0,155	+
SIGRS_64	574341	574513	32,37	173	-0,148	+
SIGRS_9	583898	584023	38,89	126	-0,275	+
SIGRS_1	585178	585304	38,58	127	-0,184	+
SIGRS_19	585272	585351	40	80	-0,239	+
SIGRS_22	585384	585474	40,66	91	-0,218	−
SIGRS_29	585779	586036	38,76	258	-0,227	+
SIGRS_13	585954	586072	42,86	119	-0,252	+
SIGRS_52	586154	586255	32,35	102	-0,172	−
SIGRS_25	605409	605471	52,38	63	-0,162	+
SIGRS_4	624366	624583	42,66	218	-0,234	+
SIGRS_24	624684	624829	41,78	146	-0,142	+
SIGRS_32	624876	625140	39,62	265	-0,296	−
SIGRS_10	625172	625595	35,85	424	-0,223	−
SIGRS_2	625621	625731	40,54	111	-0,167	+

Table 4.1: *Characteristics of the 48 ncRNA candidates including the normalised free energy.*

ncRNA ID	ncRNA start	ncRNA end	Target gene	Target start	Target end	ΔG (kcal/mol)
SIGRS_36	1	202	AAZ53573.1|protein P102-copy 1	225238(982)	228023(1183)	-388,3
SIGRS_15	1	142	AAZ53622.1|hypothetical	300768c(2289)	303197c(2430)	-266,9
SIGRS_29	4	29	AAZ53412.1|gap|Glyceraldehyde 3-phosphate dehydrogenase	45040(1015)	46250(1039)	-259,54
SIGRS_40	1	18	AAZ53430.1|atpD|ATP synthase subunit beta	63960(267)	65443(283)	-219,54
SIGRS_27	1	116	ABP01119.1|hypothetical	569546(413)	570073(528)	-215,1
SIGRS_30	1	96	AAZ53679.1|ABC transporter ATP-binding protein	369350c(2366)	371830c(2461)	-183,3
SIGRS_25	4	63	AAZ53673.2|hypothetical	361983(1722)	363913(1783)	-116,6
SIGRS_72	1	82	AAZ53712.1|ABC transporter ATP-binding protein	421614c(1971)	424038c(2053)	-115,5
SIGRS_52	1	56	AAZ53815.1|hypothetical	586200(409)	586663(464)	-95,6
SIGRS_3	4	27	AAZ53673.2|hypothetical	361983(1748)	363913(1771)	-49,5
SIGRS_26	32	51	AAZ53673.2|hypothetical	361983(1754)	363913(1773)	-40,6
SIGRS_23	20	53	AAZ53394.1|hypothetical	17210(740)	18213(773)	-39,9
SIGRS_7	1	57	AAZ53587.2|oppC-1|Oligopeptide transport system permease protein	244370(519)	245337(575)	-38,45
SIGRS_16	4	36	AAZ53997.1|gcp|tRNA N6-adenosine threonylcarbamoyltransferase	853368c(1105)	854536c(1137)	-35,5
SIGRS_18	1	19	ABP01100.1|hypothetical	110613(329)	111085(348)	-35,2
SIGRS_17	1	34	AAZ53749.1|lipoprotein	477696c(1341)	479710c(1375)	-30,35
SIGRS_100	1	17	AAZ53782.2|putative ICEF-II	519944c(3373)	523332c(3389)	-30
SIGRS_69	1	23	AAZ53898.2|polC|DNA polymerase III polC-type	709474c(219)	713967c(241)	-30
SIGRS_11	4	32	AAZ53927.1|hypothetical	748087c(224)	748563c(252)	-28,85
SIGRS_32	4	30	AAZ53488.1|Amino acid permease	155942(1034)	157629(1060)	-28,65
SIGRS_5	1	30	AAZ53522.1|conserved hypothetical protein	187957(350)	189258(377)	-28,5
SIGRS_38	1	19	AAZ53659.1|trmD|tRNA (guanine-N(1)-)-methyltransferase	344932c(505)	345687c(523)	-28,4
SIGRS_20	1	32	AAZ53467.2|conserved hypothetical protein	115482(89)	116398(120)	-28,15
SIGRS_8	1	15	AAZ54025.1|adhesin like-protein P146	890363c(446)	894522c(460)	-27,8
SIGRS_75	1	36	AAZ53596.1|nrdI| ribonucleoprotein	262905c(405)	263409c(440)	-27,3
SIGRS_1	8	25	AAZ53873.1|pdhD| Dihydrolipoamide dehydrogenase	672369(1148)	674299(1165)	-26,7
SIGRS_118	1	21	AAZ53817.2|conserved hypothetical protein	587575c(930)	588572c(950)	-26,5
SIGRS_14	1	24	AAZ53749.1|lipoprotein	477696c(1797)	479710c(1820)	-26,2

Table 4.2: *Identified interactions between the predicted ncRNAs and target genes in the forward strand.*

ncRNA ID	ncRNA start	ncRNA end	Target gene	Target start	Target end	ΔG (kcal/mol)
SIGRS_19	1	14	AAZ54025.1|Adhesin like-protein	890363c(3516)	894522c(3529)	-26,2
SIGRS_4	4	32	AAZ53482.1|Protein P102-copy 2	142634c(960)	145548c(988)	-25,9
SIGRS_33	20	48	AAZ54028.2|hypothetical	903202(162)	903662(190)	-25,8
SIGRS_6	12	34	AAZ53979.1|rpoC|DNA-directed RNA polymerase subunit beta	817365c(2350)	821667c(2373)	-25,7
SIGRS_12	21	49	AAZ53855.1|hypothetical	637216c(84)	640064c(112)	-24,95
SIGRS_22	1	17	AAZ53581.2|lysS|Lysine-tRNA ligase	236186(436)	237777(452)	-24,85
SIGRS_64	1	30	AAZ53477.2|hypothetical	131792(755)	134781(785)	-24,8
SIGRS_35	1	22	AAZ53796.1|tkt|Transketolase	541342(574)	543278(595)	-24,55
SIGRS_34	1	15	AAZ53468.2|uvrA|Excinuclease ABC subunit A	116682c(642)	119568c(656)	-24,15
SIGRS_31	1	16	AAZ53589.2|oppF-1|Oligopeptide ABC transporter ATP-binding protein	246749(262)	248057(278)	-24,1
SIGRS_82	16	37	AAZ53737.2|lipoprotein	457488(1630)	459872(1652)	-23,8
SIGRS_2	1	20	AAZ53812.2|hypothetical	574659c(460)	581494c(479)	-23,6
SIGRS_10	259	293	AAZ53952.1|thrS|threonyl-tRNA ligase	781617c(1)	783353c(37)	-23,2
SIGRS_13	1	15	ABP01100.1|hypothetical	110613(303)	111085(317)	-22,4
SIGRS_9	4	22	AAZ53589.2|oppF-1|Oligopeptide ABC transporter ATP-binding protein	246749(1035)	248057(1053)	-22,15
SIGRS_80	1	14	AAZ53827.1|rplK|50S ribosomal protein L11	602635c(424)	603236c(438)	-21,6
SIGRS_24	1	20	AAZ53424.2|atpB|ATP synthase subunit a	59315(432)	60071(451)	-21,5
SIGRS_59	1	20	AAZ53514.2|hemK| Protoporphirogen oxidase	181851(143)	182604(162)	-21,1
SIGRS_66	4	36	AAZ53633.2|tmk|Thymidylate kinase	313085(249)	313704(279)	-20,25
SIGRS_43	14	29	AAZ53378.2|dnaA|Chromosomal replication initiator protein dnaA	57(483)	1648(498)	-19,95

Table 4.3: *Identified interactions between the predicted ncRNAs and target genes in the forward strand (continuation).*

ncRNA ID	ncRNA start	ncRNA end	Target gene	Target start	Target end	ΔG (kcal/mol)
SIGRS_35	68	1	AAZ53569.1|rpsJ|30S ribosomal protein S10	220227c(379)	220750c(446)	-130
SIGRS_22	67	1	AAZ53814.2|hypothetical	585408(350)	585823(416)	-120,8
SIGRS_2	111	66	AAZ53845.1|atpA|ATP synthase alpha chain	625686c(1)	627274c(46)	-88,1
SIGRS_29	258	214	AAZ53814.2|hypothetical	585408(1)	585823(45)	-83,6
SIGRS_15	279	250	AAZ53623.1|smf|DNA processing protein SMF	303305c(1)	304284c(30)	-64
SIGRS_30	96	66	ABP01106.1|hypothetical	371780c(1)	372231c(31)	-59,3
SIGRS_12	92	48	AAZ53894.2|gyrA|DNA gyrase subunit A	704311(182)	707072(226)	-46,85
SIGRS_32	40	5	AAZ53675.1|permease	364482(87)	366043(121)	-37,7037
SIGRS_5	25	1	AAZ53473.2|tpx|thiol	125114(123)	125805(147)	-32,15
SIGRS_4	18	1	AAZ53939.1|rpsD|30S ribosomal protein S4	765105(329)	765922(346)	-32,1
SIGRS_72	31	8	AAZ53687.1|ABC transporter ATP-binding protein	383903(1906)	386274(1929)	-31,25
SIGRS_14	44	10	AAZ53620.2|conserved hypothetical protein	294444c(92)	298126c(124)	-31,0037
SIGRS_16	21	1	AAZ53826.1|rplA|50S ribosomal protein L1	601890c(521)	602637c(541)	-30,9
SIGRS_17	28	5	AAZ53937.2|PTS system, N-acetylglucosamine-specific II ABC component	762317c(1229)	764123c(1252)	-29,9
SIGRS_38	45	20	AAZ53894.2|gyrA|DNA gyrase subunit A	704311(272)	707072(297)	-29,85
SIGRS_1	32	14	AAZ53684.2|conserved hypothetical protein	381480c(789)	382327c(807)	-29,35
SIGRS_23	24	4	AAZ53941.1|fpg| Foramidopyrimidine DNA gycosylase	766420(824)	767297(843)	-29,35
SIGRS_11	32	4	AAZ53929.1|hypothetical	751968c(224)	752539c(252)	-28,85
SIGRS_25	45	19	AAZ53669.1|rpsF|30S ribosomal protein S6	356707c(863)	357600c(889)	-28,8
SIGRS_31	33	8	AAZ53979.1|rpoC|DNA-directed RNA polymerase subunit beta	817365c(2347)	821667c(2373)	-28,8
SIGRS_33	32	1	AAZ53467.2|conserved hypothetical protein	115482(89)	116398(120)	-28,15
SIGRS_18	18	1	AAZ53595.2|nrdE| Ribonucleoside-diphosphate reductase	260688c(916)	262955c(933)	-27,4
SIGRS_9	19	1	AAZ53433.1|tsf|Elongation factor Ts	67405(194)	68391(213)	-27
SIGRS_24	40	1	ABP01126.1|hypothetical	582095(142)	582533(184)	-26,8
SIGRS_19	22	1	AAZ53873.1|pdhD|dihydrolipoamide dehydrogenase	672369(1002)	674299(1023)	-26,6
SIGRS_59	25	4	AAZ53469.2|conserved hypothetical protein	119575c(1531)	122183c(1551)	-26,5
SIGRS_100	29	1	AAZ53714.1|hypothetical	425927c(192)	426748c(221)	-26,05

Table 4.4: *Identified interactions between the predicted ncRNAs and target genes in the reverse strand.*

ncRNA ID	ncRNA start	ncRNA end	Target gene	Target start	Target end	ΔG (kcal/mol)
SIGRS_20	25	1	AAZ53764.2|DNA methylase	496624(154)	497377(179)	-26,05
SIGRS_6	23	5	AAZ53628.2|lplA|Lipoate-protein ligase A	308509c(781)	309525c(798)	-25,15
SIGRS_13	19	1	AAZ53997.1|gcp|tRNA N6-adenosine threonylcarbamoyltransferase	853368c(193)	854536c(211)	-25,1
SIGRS_69	24	1	AAZ53860.2|mannose-6-phosphate	644284(168)	645269(191)	-24,8
SIGRS_43	27	1	AAZ53385.1|ftsY|Cell recognition particle receptor FtsY	7934(523)	9075(548)	-24,7
SIGRS_75	22	1	AAZ53615.1|secD|Protein-export membrane protein	287568(2488)	290365(2509)	-24,65
SIGRS_27	36	24	AAZ53486.1|pfkA|6-phosphofructokinase	153256c(704)	154424c(716)	-24,1
SIGRS_8	16	1	AAZ53898.2|polC|DNA polymerase III polC-type	709474c(1763)	713967c(1777)	-24
SIGRS_34	20	1	AAZ53807.2|conserved hypothetical protein	554613c(1554)	556944c(1572)	-23,9
SIGRS_118	16	1	AAZ53714.1|hypothetical	425927c(578)	426748c(593)	-23,2
SIGRS_64	22	1	AAZ53468.2|uvrA|Excinuclease ABC subunit A	116682c(422)	119568c(443)	-23,1
SIGRS_26	18	1	AAZ53985.1|dam|DNA adenine methylase	831808c(1579)	833657c(1595)	-22,6
SIGRS_52	25	1	AAZ53953.2|trpS|Tryptophanyl-tRNA ligase	783349c(136)	784486c(158)	-22,55
SIGRS_40	19	1	AAZ53705.1|lipoprotein	412029c(121)	412861c(140)	-22,3
SIGRS_7	15	1	AAZ53482.1|Protein P102-copy 2	142634c(1384)	145548c(1397)	-22
SIGRS_36	168	147	AAZ53919.1|PTS system galactitol-specific enzyme IIB component	739396c(1)	739677c(22)	-21,35
SIGRS_3	22	9	AAZ53407.2|hypothetical	36517(472)	37220(485)	-21,15
SIGRS_10	17	4	AAZ53581.2|lysS|Lysine-tRNA ligase	236186(756)	237777(769)	-21,1
SIGRS_80	69	48	AAZ53507.1|rpmA|50S ribosomal protein L27	172033(1)	172346(22)	-19,95
SIGRS_82	15	4	AAZ53483.1|Protein P97-copy 2	145527c(231)	148856c(242)	-19
SIGRS_66	94	82	AAZ53467.2|conserved hypothetical protein	115482(4)	116398(16)	-16,5

Table 4.5: *Identified interactions between the predicted ncRNAs and target genes in the reverse strand (continuation).*

4.3.3 Conserved ncRNAs

Using the approach described in Section4.2.3, only four IGRs were observed to be conserved, representing 3% of the intergenic content of *Mycoplasma hyopneumoniae* 7448. From the four conserved IGRs, only one was found in the predicted ncRNAs. This result is coherent with other studies showing that for other gram-positive bacteria, the level of conservation is close to null, even for closely related species (Acebo *et al.*, 2012; Richter et Backofen, 2012).

4.4 Conclusion

The work described in this chapter is still ongoing, the results obtained *in silico* having now to be validated experimentally, however, from the results obtained so far, we may already conclude a few points. Using the approach implemented in SIGRS, that considers the nucleotide composition to detect potential ncRNA genes, 48 putative ncRNA were discovered in *Mycoplasma hyopneumoniae* 7448. The segmentation approach allowed the detection of fragmented ncRNAs that were "hidden" within longer sequences. Genes related to the life and pathogenicity of the bacterium were found to be interacting with the putative ncRNAs, an important additional evidence that reinforces the idea that the ncRNAs are indeed playing a regulation role in the bacterium. Finally, very few strong conservation was found between the IGRs of closely-related *Mycoplasma* species, something that was is in agreement with previous studies.

Chapter 5

Cluster analysis of structured motifs

Contents

5.1 Introduction 85
5.2 Materials and methods 86
5.2.1 A brief reminder on SMILE 86
5.2.2 Unweighted Pair Group Method with Arithmetic Mean 89
5.3 Initial results and discussion 90
5.4 Conclusion 90

In this chapter, we present a problem related to structured motifs, which is basically a pattern, that may be composed of one or more parts separated by a certain distance, that one may look for in a sequence or a set of sequences (a more formal definition is provided in the next sections). This is an issue that may seem unrelated to the study of miRNAs but the two may however appear combined in some studies. For instance, during the internship of an undergraduate student in the team, Evgueni Jacob, the motifs associated to the miRNAs that were exported from a human tissue were analysed and classified according to their statistical significance.

The problem of finding structured motifs was first addressed by Marsan et Sagot (2000) and implemented as a software called SMILE (Structured Motifs Inference and Evaluation). Depending on the parameters given to SMILE, the algorithm can generate a large output that may contain redundant information. For instance, if the characteristics of the motifs are not precisely known, one should choose more permissive parameters in an attempt to recover such motifs. Here we present some clustering solutions to group motifs that may correspond to the same biological "object", and to better identify the noise that may be present in such large outputs.

5.1 Introduction

Efficiently identifying biological sites or features in a set of sequences is an essential approach to identify functional elements in a genome. Example of such elements are DNA binding sites and miRNA families (i.e., all the isoforms of a same miRNA). There are two main problems related to this identification. One is the inference of a consensus sequence for such elements, the other is the prediction of the location of the sites or features that represent true positive representatives of the corresponding elements in the set of sequences. The algorithms

for the prediction of location often use the results produced by the consensus extraction methods to establish all true positive positions along a genome, although the two can also be extracted together. Indeed, this is the case of the algorithm SMILE (Marsan et Sagot, 2000): it simultaneously infers consensus motifs and and locates the corresponding elements in a set of sequences. The software is available at https://team.inria.fr/bamboo/en/softwares/smile/.

SMILE implements an exact algorithm for finding motifs in a set of sequences. A suffix tree is used to represent the input sequences, which together with the strategies implemented in the algorithm, result in an efficient method for the extraction of motifs. SMILE requires a number of parameters, such as the number p of parts, called *boxes*, that a (structured) motif may have, the minimum number of substitutions e (one per box) between the motif and its occurrence, and the minimum number of times q (which stands for *quorum*) that the motif has to appear among the sequences. Depending on the values of these parameters, the size of the output generated by SMILE may be very large, and may contain redundant motifs, or motifs that overlap and may be considered as one single functional element. For example, the larger is the value of e, or inversely the smaller the value of q, the larger will be the output. In an attempt to organise such output by eliminating the redundancy or by grouping together motifs that correspond to a same functional element, we implemented an UPGMA (Unweighted Pair Group Method with Arithmetic Mean) algorithm to cluster the motifs that are similar according to the positions in the sequences where these motifs appear. The implementation of this algorithm was performed during the internship of Thomas Balezeau, an undergraduate student in computer science, whom I co-advised together with Marie-France Sagot. Another approach that has been explored, but not yet implemented, is the use of hashing for list intersection as a quicker estimator to find redundant motifs, or motifs that represent a single biological entity.

5.2 Materials and methods

5.2.1 A brief reminder on SMILE

Basic definitions

A *motif* is a pattern that "appears" in a set of sequences. Each such "appearance" is called an *occurrence*. An occurrence is thus a word in a sequence, while a motif may be seen as a "representation" of a set of occurrences. Motifs thus serve to both locate and to describe certain words, their occurrences, in a set of sequences.

More formally, a *structured motif* (or simply *motif*) is defined as an ordered set of $p \geq 1$ "box(es)", with p maximum error rates (one for each box), and $p-1$ intervals of distance (one for each pair of successive boxes). Let Σ be the alphabet of nucleotides A, C, G, T. An element $m \in \Sigma+$ is said to be a *motif*, if there is at least one occurrence u in s such that: (i) $s = xuy$ for $x, y \in \Sigma*$, and (ii) the Hamming distance (i.e., minimum number of substitutions) between u and m is no more than e, a non-negative integer. Given N sequences $s_1, ..., s_N \in \Sigma*$ and an integer $1 \leq q \leq N$, an element $m \in \Sigma+$ is said to be a *valid motif* if it has at least one occurrence in a quorum q of distinct sequences. From now on, we will call simply *motif* any valid one, given a quorum q. Notice that if e is strictly greater than 0, a motif may never appear exactly in any of the sequences of the set.

Algorithm

The algorithm implemented in SMILE solves the problem of identifying motifs and is described as follows. Given a set of N sequences $s_1, ..., s_N$, a non-negative integer e, and a positive integer $q \leq N$, the goal is to find all the motifs $((m_1, ..., m_p), ((d_{min_1}, d_{max_1}, \delta_1), ..., (d_{min_{p-1}}, d_{max_{p-1}}, \delta_{p-1})))$ that are valid, where $p \geq 1$. The distances play a role only if $p > 1$: d_{min} and d_{max} stand respectively for the minimum and maximum distances between two successive boxes. In a same way, δ plays a role only if it is strictly greater than one: in this case indeed, it is not only one interval ($[d_{min}, d_{max}]$) that will be considered, but all intervals $[d_{min} - 1, d_{min} + 1]$ until $[d_{max} - 1, d_{max} + 1]$. Notice that we have a single motif when $p = 1$ and $d_{min} = d_{max} = 0$ (in this case, by default, $\delta_1 = 0$), otherwise we have a structured motif composed of p boxes.

As mentioned, a suffix tree $\mathcal{T}$ is used to represent the set of sequences $s_1, ..., s_N$. Suffix trees were introduced by McCreight (1976), and modified by Gusfield (1997) and Bieganski *et al.* (1994) to consider $N \geq 1$ sequences. To extract all the valid single motifs $m \in \Sigma^{k \geq 1}$ with a number e of substitutions allowed and appearing in at least q (quorum) sequences, Marsan et Sagot (2000) implemented an algorithm that traverses simultaneously and recursively the lexicographic trie $\mathcal{M}$ of all possible motifs of length k and the suffix tree $\mathcal{T}$ of the sequences. The algorithm is based on a recurrence that is stated by the following lemma:

Lemma 1 (Sagot, 1998) *A pair (v, e_v) is a node-occurrence of $m' = m\alpha$ with $m \in \Sigma^l$ for $1 \leq l < k$ and $\alpha \in \Sigma$ if, and only if, one of the following two conditions is verified:*

(match) *A pair $(parent(v), e_v)$ is a node-occurrence of m and the label of the arc from $parent(v)$ to v is α;*

(subst.) *A pair $(parent(v), e_v - 1)$ is a node-occurrence of m and the label of the arc from $parent(v)$ to v is $\beta \neq \alpha$.*

As for structured motifs, the lemma above together with extensions described in Marsan et Sagot (2000) are used.

Algorithm 1

Here we describe the procedure used in Marsan et Sagot (2000) to find structured motifs of the type $((m_1, m_2), (d_{min}, d_{max}))$, i.e. with $p = 2$ and $\delta_1 = 0$. In other words, we want to find a structured motif with two boxes separated by a fixed interval (that can be a fixed length if $d_{min} = d_{max}$).

Using the suffix tree $\mathcal{T}$, the first motif of length k can be found together with its set V_1 of $\mathcal{T}$-node-occurrences (which are nodes located at level k in $\mathcal{T}$). Once an occurrence of motif m_1 is found to finish at node v of the tree $\mathcal{T}$, a "jump" from $level(v)$ to $level(w)$, with $d_{min} \leq level(w) - level(v) \leq d_{max}$, is performed. The node w corresponds to the potential starts of node-occurrences of w of motif m_2, with $w \in V_2$, such that:

$$V_2 = \{(w, e_v = e_w) \mid \exists v \in V_1 \text{ with } d_{min} \leq level(w) - level(v) \leq d_{max}\} \tag{5.1}$$

From a node-occurrence v in V_1 of motif m_1, a jump is thus made in $\mathcal{T}$ to all potential start node-occurrences w of m_2. If the nodes v in V_1 and the nodes w in V_2 satisfy the recurrence formula given in lemma 1, the structured motif $((m_1, m_2), (d_{min}, d_{max}))$ is verified.

To find structured motifs with $p > 2$ and $\delta_1 > 0$, the authors extended the algorithm accordingly. For a detailed description, see Marsan et Sagot (2000).

Parameters and output

The user of SMILE has the option to generate a generic parameter file by providing the number of required boxes, or to manually specify the parameters in the command line. These parameters include: the name of the input and output files, the alphabet for the motifs that may be the same as for the sequences from which they are inferred or an extended IUPAC one, the quorum q (minimum number of sequences where the motifs have to appear), minimum and maximum length of the motif, number of substitutions e, and number of boxes. If motifs with more than one box are sought, additional informations must be provided, such as the minimum and maximum length of the spacer between the boxes (that is the minimum and maximum size of the interval separating the two boxes), together with the value of *delta* if this is strictly greater than zero (otherwise, *delta* does not need to be specified). Once the motifs are found and their occurrences extracted, a statistical measure is used to check whether the motifs may be considered potentially significant or not. Notice that statistical significance does not necessarily imply biological significance, but may be seen as a first filter for the latter. To that purpose, the authors compute a $\mathcal{X}^2$ test (with one degree of freedom) on two contingency tables, one corresponding to what is observed, the other to what was expected under the null hypothesis. To determine what would be expected under the null hypothesis, the idea is to shuffle the original sequence(s) from which the motifs were extracted, and to count how many times the motifs found in the original dataset are present, considering a Hamming distance with the same value of e, in the shuffled dataset. The user of SMILE is required to provide the number of shufflings to be performed and the size of the k-mer to be conserved when shuffling the sequences.

The output of SMILE is composed of the parameters summarised in the header, the sequence of the motif, followed by a numerical encoding of the motif sequence and the number of sequences in which it appeared. The source sequence and the positions of the occurrences are listed below the motif, and finally the total number of occurrences is presented. One example of output is presented in Figure 5.1.

As mentioned before, depending on how permissive are the input parameters, this output can be very large, possibly producing motifs that are redundant in the sense that many correspond to a same functional element. This procedure may also allow to reveal motifs that are clearly noise. We made a first attempt to address this problem by implementing an UPGMA algorithm, described as follows, to organise in clusters the motifs that are very similar.

5.2.2 Unweighted Pair Group Method with Arithmetic Mean

UPGMA is an agglomerative hierarchical clustering method that was initially proposed by Sneath *et al.* (1973) and improved by Murtagh (1984, 1983) into an ($\mathcal{O}(n^2)$ time and $\mathcal{O}(n^2)$ space) algorithm. As the name indicates, it is a method that is unweighted (all pairwise distances contribute equally), pair group (groups are combined in pairs, dichotomies only), and arithmetic mean (the pairwise distance to a group is the mean of all the distances to each member of that group).

From a distance matrix that provides the distances (e.g., euclidean) between the pairwise points, the algorithm first finds the smallest distance. The two corresponding points are inserted as leaves in a rooted tree (that will represent the structure of the pairwise matrix). The same two points are agglomerated in a single cell (a cluster) in the matrix and the distance of this new cell to all the other points is computed as the mean of the distances to all the members of the given cluster. For example, if the distance between the points a and b is the

```
%%% 1 7/43 46919 14 14 1 alphabet ACGT$


================================================================================
TCCAGCCTGG 3110211322 7
Seq    23   Pos   276
Seq    15   Pos   635
Seq    14   Pos   986
Seq    41   Pos    94
Seq     8   Pos   288
Seq     7   Pos   216
Seq    27   Pos   183
7
AAAAAAAAAT 0000000003 6
Seq    29   Pos    30
Seq    21   Pos   247
Seq    15   Pos   672
Seq     2   Pos   622
Seq    16   Pos   615
Seq     8   Pos   804
Seq     8   Pos   805
Seq     8   Pos   806
8
GGGGCTGGGG 2222132222 3
Seq    40   Pos   198
Seq    38   Pos  1016
Seq     0   Pos   493
3
User time : 0.52 sec.
```

Figure 5.1: *An example of output of* Smile. *The header is simply a summary of the input parameters provided. For each motif, its sequence and numerical encoding are presented, together with the number of sequences where the motif had occurrences. The source sequence and start positions of the occurrences are presented together, followed by the total number of occurrences.*

smallest in the matrix, these two points will form a single clustered cell ab. If the distance from a to c is 0.1 and the distance from b to c is 0.2, the distance between the new cluster ab and c will be 0.15 (i.e., (0.1+0.2)/2). The algorithm can be simply summarised in the following steps:

1. Determine all interpoints dissimilarities.
2. Form a cluster from the two closest points or clusters.
3. Redefine dissimilarities between the new cluster and the other points or clusters (all the other interpoint dissimilarities remaining unchanged).
4. Return to Step 2 until all points are in one single cluster.

Metric definition and matrix construction

To compute the dissimilarities between the points, in this case the motifs, we used a metric that we called "motif co-occurrence metric" and that takes into account the overlapping positions in the original sequence of two motifs. The idea is that similar motifs would more probably co-occur in the same positions. The clustering method would thus enable to group occurrences of a given motif. We used Equation 5.2 to compute the metric:

$$d = 1 - \frac{\cap(m_1, m_2)}{\cup(m_1, m_2)} \tag{5.2}$$

which is simply the ratio between the number of overlapping positions between motifs m_1 and m_2 and the sum of the lengths of both motifs. Figure 5.2 shows an example of the motif co-occurrence metric.

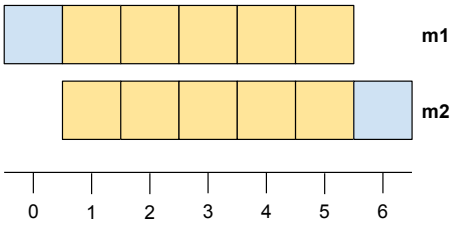


Figure 5.2: *An example of the co-occurrence metric for motifs m_1 and m_2. The distance between the two motifs is equal to $d = 1 - \frac{\cap(m_1,m_2)}{\cup(m_1,m_2)} = \frac{10}{12} = 0.83$, that is, motifs m_1 and m_2 are 83% similar to each other.*

To compute the dissimilarity matrix between the motifs of an output of Smile, one needs only to parse the output file recovering all the motifs per sequence, and for each pair of motifs, to compute the dissimilarity measure shown in Equation 5.2. Once the matrix is built, it can be given as input to the UPGMA algorithm.

5.3 Initial results and discussion

As this issue was addressed at the end of this thesis, we present here only initial results. First, Smile was applied over a set of miRNA sequences derived from a human tissue during the internship mentioned at the beginning of this chapter. The mature miRNA sequences were extended 500nt up and downstream, and were given as input to Smile. Four different

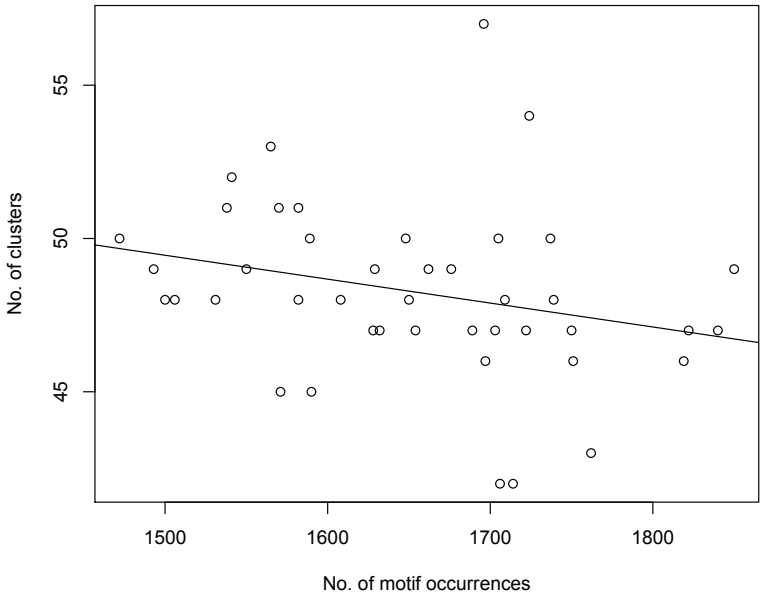


Figure 5.3: *Regression curve between the number of motif occurrences (per sequence) and the respective number of clusters grouping these occurrences. The regression was performed on the clustering of the* Smile*'s output obtained with the following configuration:* $p = 1$ *box, at most* $e = 1$ *substitution, and minimum and maximum motif lengths of 6nt and 11nt.*

configurations of the algorithm were used, all of them requiring motifs with one box ($p = 1$), a maximum of one substitution ($e = 1$), over the alphabet "ATCG". The minimum and maximum lengths were respectively: 6nt and 11nt, 10nt and 11nt, 15nt (for both minimum and maximum), 14nt (for both minimum and maximum), one for each of the four configurations.

The UPGMA algorithm was then fed with the four outputs of Smile, generated as described above. The number of motif occurrences per sequence and the respective number of clusters are presented in Figure 5.3 for the first configuration, that we call "configuration 611". As one may notice, when the number of occurrences increases, the number of clusters decreases, meaning that the variability is more apparent when the number of occurrences is smaller. This may be expected since the distance between the clusters is computed by the mean of the distances of their components.

5.4 Conclusion

Although this is the beginning of a study on clustering motifs, we may present a few conclusions. The number of clusters grouping the different occurrences seems to be coherent to what was expected. To verify the consistency of these clusterings, and more importantly, to determine if the grouped occurrences are biologically functionally related, we will explore different datasets for which the biological motifs are precisely described. As concerns the method and its performance, a substantial improvement, either in the implementation or in the method itself, must still be performed since it currently is time consuming. For instance, for an input of 1.6Mb, it took ∼16 minutes to compute the clusters, running under a Mac OS

X 10.6.8, 2.7 GHz Intel.

Conclusion and perspectives

The most important contribution of this thesis was the development of a reliable, flexible, and much faster method for the prediction of pre-miRNAs. MIRINHO predicts pre-miRNAs as well as the other tested methods, however it is orders of magnitude faster. Our method was used as the basis for other issues addressed during this thesis. It is at the heart of the pipeline MIRINHOPIPE for the treatment of sRNAseq data and was adapted inside the method ALVINHO for the prediction of ncRNA targets. Moreover, MIRINHO is currently been used in other projects of the team, for example that involve the prediction of pre-miRNAs in swines.

The efficiency and reliability of our method creates new perspectives related to the "miRNA world". The incorporation of a larger number of features for the detection of miRNAs is now possible due to the speed of our method. Such features have already been defined in Kozomara et Griffiths-Jones (2013) and appear to be very precise in determining a positive (pre-)miRNA. To our knowledge, they have not yet been incorporated in any software. Another characteristic we have been explored but need to develop further is the use of targets to eliminate false pre-miRNAs. Besides possibly providing a more accurate set of pre-miRNAs, this approach would enrich the results by providing a functional overview of such molecules.

The direct application of MIRINHO to sRNAseq data allows the processing of millions of reads in a more feasible time. As NGS is a constantly evolving technology, the quantity of such type of data can only increase. The efficient extraction of knowledge from such data is an essential task to provide a richer comprehension of how regulation is influencing species evolution. One point that deserves special attention as concerns sRNAseq is the large number of identified pre-miRNAs when low expression must be considered. One possible solution to this problem would be the incorporation of features in the efficient prediction of pre-miRNAs, such as the ones associated to the structure of the hairpin (e.g., minimum free energy per base) and to the location of the miRNA within the structure (e.g., in the stem with an overhang of ∼2nt at each 3' end of the miRNA duplex). As noticed during this thesis (Chapter 3), such features can be powerful in discriminating true from false pre-miRNAs.

Moving now to the sRNAseq data that we analysed: the miRNAs identified in the pea aphid, together with their putative targets, open perspectives that need to be addressed. One crucial task is the identification of the miRNAs that are being differentially expressed between the different developmental stages. If we are able to address this problem, we will be able to precisely determine which miRNA is playing a key role in each stage. Another question that needs to be treated is the huge number of interactions that were found. To this purpose, a functional analysis of the targets will be performed together with an analysis of the correlation between the expression of the miRNAs and the respective predicted targets. This will provide more accurate evidences for the potential functional interactions, and shed some light on the consequences on the development of the pea aphid.

The last issue addressed in this thesis was related to the clustering of motifs. To verify the consistency of the clusterings, and more importantly, to determine if the grouped occurrences

are biologically functionally related, different datasets for which the biological motifs are better described need to be explored. A starting point would be the datasets described in Vanet *et al.* (1999) and Vanet *et al.* (2000). Once the consistency and relevance of the clusters are verified, an interesting application of the clustering approach would be in providing additional evidence for the predicted targets. In this case, the motif is the predicted miRNA, for which we know a few characteristics such as its length and the number e of substitutions between its isoforms. The sequences where the motif would be searched are mRNAs, the occurrences being potential targets for the miRNAs. This could lead to an extra verification of the interaction between miRNA and target reinforcing the target prediction results. As concerns the method and its performance, a substantial improvement, either in the implementation or in the method itself, needs to be performed since it is currently time consuming. Alternatively, other approaches, using for instance a hashing intersection list, will be investigated in future and implemented to verify which approach gives the best performance.

Bibliography

Acebo, P., Martin-Galiano, A. J., Navarro, S., Zaballos, Á. et Amblar, M. (2012). Identification of 88 regulatory small RNAs in the tigr4 strain of the human pathogen streptococcus pneumoniae. *RNA*, 18(3):530–546.

Agarwal, S., Vaz, C., Bhattacharya, A. et Srinivasan, A. (2010). Prediction of novel precursor miRNAs using a context-sensitive hidden markov model (cshmm). *BMC bioinformatics*, 11(Suppl 1):S29.

Akutsu, T. (2000). Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 104(1):45–62.

Aldridge, S. et Hadfield, J. (2012). Introduction to miRNA profiling technologies and cross-platform comparison. *In Next-Generation MicroRNA Expression Profiling Technology*, pages 19–31. Springer.

Almeida, M. I., Reis, R. M. et Calin, G. A. (2011). MicroRNA history: discovery, recent applications, and next frontiers. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 717(1):1–8.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. et Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Ammar, A., Elouedi, Z. et Lingras, P. (2012). Rpkm: The rough possibilistic k-modes. *In Foundations of Intelligent Systems*, pages 81–86. Springer.

Axtell, M. J., Westholm, J. O., Lai, E. C. *et al.* (2011). Vive la difference: biogenesis and evolution of microRNAs in plants and animals. *Genome Biol*, 12(4):221.

Batista, P. J. et Chang, H. Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. *Cell*, 152(6):1298–1307.

Bentley, J. (1984). Programming pearls: algorithm design techniques. *Communications of the ACM*, 27(9):865–873.

Berezikov, E., Cuppen, E. et Plasterk, R. H. (2006). Approaches to microRNA discovery. *Nature genetics*, 38:S2–S7.

Bieganski, P., Riedl, J., Cartis, J. et Retzel, E. F. (1994). Generalized suffix trees for biological sequence data: Applications and implementation. *In System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on*, volume 5, pages 35–44. IEEE.

Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B. et Cohen, S. M. (2003). *bantam* encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene *hid* in *Drosophila*. *Cell*, 113(1):25–36.

Brennecke, J., Stark, A., Russell, R. B. et Cohen, S. M. (2005). Principles of microRNA–target recognition. *PLoS biology*, 3(3):e85.

Brockmeier, S. L., Halbur, P. G. et Thacker, E. L. (2002). *Polymicrobial Diseases*, chapitre 13, Porcine respiratory disease complex, pages 231–258. Washington (DC), ASM Press, USA, ISBN-10.

Bron, C. et Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577.

BYRT, D., Heap, P. et POINTON, A. (1985). Effect of enzootic pneumonia of pigs on growth performance. *Australian veterinary journal*, 62(1):13–18.

Castel, S. E. et Martienssen, R. A. (2013). RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nature Reviews Genetics*, 14(2):100–112.

Chaudhuri, K. et Chatterjee, R. (2007). MicroRNA detection and target prediction: integration of computational and experimental approaches. *DNA and cell biology*, 26(5): 321–337.

Chen, K. et Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics*, 8(2):93–103.

Darzacq, X., Jády, B. E., Verheggen, C., Kiss, A. M., Bertrand, E. et Kiss, T. (2002). Cajal body-specific small nuclear RNAs: a novel class of 2′-o-methylation and pseudouridylation guide RNAs. *The EMBO journal*, 21(11):2746–2756.

Davidson, B. L. et McCray, P. B. (2011). Current prospects for RNA interference-based therapies. *Nature Reviews Genetics*, 12(5):329–340.

de Planell-Saguer, M. et Rodicio, M. C. (2011). Analytical aspects of microRNA in diagnostics: a review. *Analytica chimica acta*, 699(2):134–152.

DeBey, M. C. et Ross, R. F. (1994). Ciliostasis and loss of cilia induced by *Mycoplasma hyopneumoniae* in porcine tracheal organ cultures. *Infection and immunity*, 62(12):5312–5318.

Dieci, G., Preti, M. et Montanini, B. (2009). Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics*, 94(2):83–88.

Eddy, S. R. (2004). How do RNA folding algorithms work? *Nature Biotechnology*, 22(11): 1457–1458.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., Marks, D. S. *et al.* (2004). MicroRNA targets in drosophila. *Genome biology*, 5(1):R1–R1.

Fang, Z. et Rajewsky, N. (2011). The impact of miRNA target sites in coding sequences and in 3′ utrs. *PloS one*, 6(3):e18067.

Filipowicz, W. et Pogačić, V. (2002). Biogenesis of small nucleolar ribonucleoproteins. *Current opinion in cell biology*, 14(3):319–327.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. et Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans. *nature*, 391(6669):806–811.

Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S. et Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using mirdeep. *Nature biotechnology*, 26(4):407–415.

Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W. et Rajewsky, N. (2012). mirdeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids research*, 40(1):37–52.

Gardner, S. W. et Minion, F. C. (2010). Detection and quantification of intergenic transcription in *Mycoplasma hyopneumoniae*. *Microbiology*, 156(8):2305–2315.

Geisler, S. et Coller, J. (2013). RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nature Reviews Molecular Cell Biology*, 14(11):699–712.

German, M. A., Luo, S., Schroth, G., Meyers, B. C. et Green, P. J. (2009). Construction of parallel analysis of RNA ends (pare) libraries for the study of cleaved miRNA targets and the RNA degradome. *Nature protocols*, 4(3):356–362.

Gesteland, R. et Atkins, J. (1993). The RNA world.

Ghosh, Z. et Mallick, B. (2012). Renaissance of the regulatory RNAs. *In Regulatory RNAs*, pages 3–22. Springer.

Godinho, C. P. S., Higashi, S., Sagot, M., Zaha, A. et Vasconcelos, A. T. R. (in preparation). Prediction and experimental validation of non-coding RNAs in the bacteria *Mycoplasma hyopneumoniae*.

Gommans, W. M. et Berezikov, E. (2012). Controlling miRNA regulation in disease. *In Next-Generation MicroRNA Expression Profiling Technology*, pages 1–18. Springer.

Gottesman, S. et Storz, G. (2011). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harbor perspectives in biology*, 3(12):a003798.

Griffiths-Jones, S., Grocock, R. J., Van Dongen, S., Bateman, A. et Enright, A. J. (2006). mirbase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research*, 34(suppl 1):D140–D144.

Griffiths-Jones, S., Saini, H. K., van Dongen, S. et Enright, A. J. (2008). mirbase: tools for microRNA genomics. *Nucleic acids research*, 36(suppl 1):D154–D158.

Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P. et Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell*, 27(1):91–105.

Guo, S. et Kemphues, K. J. (1995). *par-1*, a gene required for establishing polarity in c. elegans embryos, encodes a putative ser/thr kinase that is asymmetrically distributed. *Cell*, 81(4):611–620.

Gusfield, D. (1997). *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge University Press.

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., Terns, R. M. et Terns, M. P. (2009). RNA-guided RNA cleavage by a crispr RNA-cas protein complex. *Cell*, 139(5):945–956.

Hansen, A. K. et Degnan, P. H. (2014). Widespread expression of conserved small RNAs in small symbiont genomes. *The ISME journal*.

He, L. et Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, 5(7):522–531.

Hertel, J. et Stadler, P. F. (2006). Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, 22(14):e197–e202.

Higashi, S., Fournier, C., Gautier, C., Gaspin, C. et Sagot, M.-F. (in press). Mirinho: An efficient and general plant and animal miRNA predictor for genomic and deep sequencing data. *BMC Bioinformatics*.

Higashi, S., Rue, O., Gaget, K., Duport, G., Charles, H., Calevro, F., Colella, S., Gautier, C., Gaspin, C. et Sagot, M.-F. (in preparation). MicroRNA identification from small RNA sequencing data in four developmental stages of *A. pisum*.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. et Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188.

Howe, E., Holton, K., Nair, S., Schlauch, D., Sinha, R. et Quackenbush, J. (2010). Mev: multiexperiment viewer. *In Biomedical informatics for cancer research*, pages 267–277. Springer.

Hsu, T. et Minion, F. C. (1998). Molecular analysis of the p97 cilium adhesin operon of *Mycoplasma hyopneumoniae*. *Gene*, 214(1):13–23.

Huang, T.-H., Fan, B., Rothschild, M. F., Hu, Z.-L., Li, K. et Zhao, S.-H. (2007). Mirfinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC bioinformatics*, 8(1):341.

Jha, A., Chauhan, R., Mehra, M., Singh, H. R. et Shankar, R. (2012). mir-bag: bagging based identification of microRNA precursors. *PloS one*, 7(9):e45782.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X. et Lu, Z. (2007). Mipred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic acids research*, 35(suppl 2):W339–W344.

Jing, Q., Huang, S., Guth, S., Zarubin, T., Motoyama, A., Chen, J., Di Padova, F., Lin, S.-C., Gram, H. et Han, J. (2005). Involvement of microRNA in au-rich element-mediated mRNA instability. *Cell*, 120(5):623–634.

Jones-Rhoades, M. W., Bartel, D. P. et Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. *Annu. Rev. Plant Biol.*, 57:19–53.

Kadri, S., Hinman, V. et Benos, P. (2009). Hhmmir: efficient de novo prediction of microRNAs using hierarchical hidden markov models. *BMC bioinformatics*, 10(Suppl 1):S35.

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L. *et al.* (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316(5830):1484–1488.

Karlin, S. et Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences*, 87(6):2264–2268.

Karlin, S. et Dembo, A. (1992). Limit distributions of maximal segmental score among markov-dependent partial sums. *Advances in Applied Probability*, 24:113–140.

Kawamata, T. et Tomari, Y. (2010). Making risc. *Trends in biochemical sciences*, 35(7):368–376.

Kelesidis, T. (2014). The cross-talk between spirochetal lipoproteins and immunity. *Frontiers in immunology*, 5.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. et Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nature genetics*, 39(10):1278–1284.

Kim, V. N. (2005). Small RNAs: classification, biogenesis, and function. *Mol cells*, 19(1):1–15.

Kiriakidou, M., Nelson, P. T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z. et Hatzigeorgiou, A. (2004). A combined computational-experimental approach predicts human microRNA targets. *Genes & development*, 18(10):1165–1178.

Kozomara, A. et Griffiths-Jones, S. (2011). mirbase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research*, 39(suppl 1):D152–D157.

Kozomara, A. et Griffiths-Jones, S. (2013). mirbase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research*, page gkt1181.

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M. *et al.* (2005). Combinatorial microRNA target predictions. *Nature genetics*, 37(5):495–500.

Krüger, J. et Rehmsmeier, M. (2006). RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic acids research*, 34(suppl 2):W451–W454.

Kunej, T., Godnic, I., Horvat, S., Zorc, M. et Calin, G. A. (2012). Cross talk between microRNA and coding cancer genes. *Cancer journal (Sudbury, Mass.)*, 18(3):223.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S. L. *et al.* (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25.

Larsson, P., Hinas, A., Ardell, D. H., Kirsebom, L. A., Virtanen, A. et Söderbom, F. (2008). De novo search for non-coding RNA genes in the at-rich genome of dictyostelium discoideum: performance of markov-dependent genome feature scoring. *Genome research*, 18(6):888–899.

Lee, R. C., Feinbaum, R. L. et Ambros, V. (1993). The c. elegans heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75(5):843–854.

Legeai, F., Rizk, G., Walsh, T., Edwards, O., Gordon, K., Lavenier, D., Leterme, N., Méreau, A., Nicolas, J., Tagu, D. *et al.* (2010a). Bioinformatic prediction, deep sequencing of microRNAs and expression analysis during phenotypic plasticity in the pea aphid, acyrthosiphon pisum. *BMC genomics*, 11(1):281.

Legeai, F., Shigenobu, S., Gauthier, J.-P., Colbourne, J., Rispe, C., Collin, O., Richards, S., Wilson, A. C., Murphy, T. et Tagu, D. (2010b). Aphidbase: a centralized bioinformatic resource for annotation of the pea aphid genome. *Insect molecular biology*, 19(s2):5–12.

Lewis, B. P., Burge, C. B. et Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *cell*, 120(1):15–20.

Li, L., Huang, D., Cheung, M. K., Nong, W., Huang, Q. et Kwan, H. S. (2013). BSRD: a repository for bacterial small regulatory RNA. *Nucleic acids research*, 41(D1):D233–D238.

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., Burge, C. B. et Bartel, D. P. (2003). The microRNAs of caenorhabditis elegans. *Genes & development*, 17(8):991–1008.

Liu, B., Li, J. et Cairns, M. J. (2014). Identifying miRNAs, targets and functions. *Briefings in bioinformatics*, 15(1):1–19.

Liu, C.-G., Calin, G. A., Volinia, S. et Croce, C. M. (2008). MicroRNA expression profiling using microarrays. *Nature Protocols*, 3(4):563–578.

Lodish, H. F., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., Darnell, J. *et al.* (2000). *Molecular cell biology*. WH Freeman New York.

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., Hofacker, I. L. *et al.* (2011). ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1):26.

Lutter, D., Marr, C., Krumsiek, J., Lang, E. W. et Theis, F. J. (2010). Intronic microRNAs support their host genes by mediating synergistic and antagonistic regulatory effects. *BMC genomics*, 11(1):224.

Lytle, J. R., Yario, T. A. et Steitz, J. A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ utr as in the 3′ utr. *Proceedings of the National Academy of Sciences*, 104(23):9667–9672.

Madsen, M. L., Puttamreddy, S., Thacker, E. L., Carruthers, M. D. et Minion, F. C. (2008). Transcriptome changes in mycoplasma hyopneumoniae during infection. *Infection and immunity*, 76(2):658–663.

Marsan, L. et Sagot, M.-F. (2000). Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *Journal of Computational Biology*, 7(3-4):345–362.

Mathelier, A. et Carbone, A. (2010). Mirena: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics*, 26(18):2226–2234.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M. et Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7287–7292.

Mathews, D. H., Schroeder, S. J., Turner, D. H. et Zuker, M. (2006). Predicting RNA secondary structure. *Cold Spring Harbor Monograph Archive*, 43:631–657.

Mathews, D. H. et Turner, D. H. (2006). Prediction of RNA secondary structure by free energy minimization. *Current opinion in structural biology*, 16(3):270–278.

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

McCreight, E. M. (1976). A space-economical suffix tree construction algorithm. *Journal of the ACM (JACM)*, 23(2):262–272.

Meister, G. (2013). Argonaute proteins: functional insights and emerging roles. *Nature Reviews Genetics*, 14(7):447–459.

Mendes, N., Freitas, A. T. et Sagot, M.-F. (2009). Current tools for the identification of miRNA genes and their targets. *Nucleic acids research*, 37(8):2419–2433.

Miura, T., Braendle, C., Shingleton, A., Sisk, G., Kambhampati, S. et Stern, D. L. (2003). A comparison of parthenogenetic and sexual embryogenesis of the pea aphid acyrthosiphon pisum (hemiptera: Aphidoidea). *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 295(1):59–81.

Moretti, F., Thermann, R. et Hentze, M. W. (2010). Mechanism of translational regulation by mir-2 from sites in the $5'$ untranslated region or the open reading frame. *Rna*, 16(12):2493–2502.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359.

Murtagh, F. (1984). Complexities of hierarchic clustering algorithms: State of the art. *Computational Statistics Quarterly*, 1(2):101–113.

Nam, J.-W., Kim, J., Kim, S.-K. et Zhang, B.-T. (2006). Promir ii: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucleic acids research*, 34(suppl 2):W455–W458.

Nam, J.-W., Shin, K.-R., Han, J., Lee, Y., Kim, V. N. et Zhang, B.-T. (2005). Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic acids research*, 33(11):3570–3581.

Napoli, C., Lemieux, C. et Jorgensen, R. (1990). Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. *The Plant Cell Online*, 2(4):279–289.

Needleman, S. B. et Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Nelson, J. W., Martin, F. H. et Tinoco, I. (1981). Dna and RNA oligomer thermodynamics: The effect of mismatched bases on double-helix stability. *Biopolymers*, 20(12):2509–2531.

Nicolás, M. F., Barcellos, F. G., Nehab Hess, P. et Hungria, M. (2007). Abc transporters in *Mycoplasma hyopneumoniae* and *Mycoplasma synoviae*: insights into evolution and pathogenicity. *Genetics and Molecular Biology*, 30(1):202–211.

Nussinov, R., Pieczenik, G., Griggs, J. R. et Kleitman, D. J. (1978). Algorithms for loop matchings. *SIAM Journal on Applied mathematics*, 35(1):68–82.

Ogawa, Y., Sun, B. K. et Lee, J. T. (2008). Intersection of the RNA interference and x-inactivation pathways. *Science*, 320(5881):1336–1341.

Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M. et Lai, E. C. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. *Cell*, 130(1):89–100.

Okamura, K., Phillips, M. D., Tyler, D. M., Duan, H., Chou, Y.-t. et Lai, E. C. (2008). The regulatory activity of microRNA species has substantial influence on microRNA and 3' utr evolution. *Nature structural & molecular biology*, 15(4):354–363.

Ørom, U. A., Nielsen, F. C. et Lund, A. H. (2008). MicroRNA-10a binds the 5′ utr of ribosomal protein mRNAs and enhances their translation. *Molecular cell*, 30(4):460–471.

Oulas, A., Boutla, A., Gkirtzou, K., Reczko, M., Kalantidis, K. et Poirazi, P. (2009). Prediction of novel microRNA genes in cancer-associated genomic regions,Äîa combined computational and experimental approach. *Nucleic acids research*, 37(10):3276–3287.

Pasquinelli, A. E. (2012). MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Reviews Genetics*, 13(4):271–282.

Petersen, C. P., Doench, J. G., Grishok, A. et Sharp, P. A. (2006). 19 the biology of short RNAs. *Cold Spring Harbor Monograph Archive*, 43:535–565.

Pratt, A. J. et MacRae, I. J. (2009). The RNA-induced silencing complex: a versatile gene-silencing machine. *Journal of Biological Chemistry*, 284(27):17897–17901.

Pritchard, C. C., Cheng, H. H. et Tewari, M. (2012). MicroRNA profiling: approaches and considerations. *Nature Reviews Genetics*, 13(5):358–369.

Qin, W., Shi, Y., Zhao, B., Yao, C., Jin, L., Ma, J. et Jin, Y. (2010). mir-24 regulates apoptosis by targeting the open reading frame (orf) region of faf1 in cancer cells. *PLoS One*, 5(2):e9429.

Rabatel, A., Febvay, G., Gaget, K., Duport, G., Baa-Puyoulet, P., Sapountzis, P., Bendridi, N., Rey, M., Rahbé, Y., Charles, H. *et al.* (2013). Tyrosine pathway regulation is host-mediated in the pea aphid symbiosis during late embryonic and early larval development. *BMC genomics*, 14(1):235.

Razin, S. (2006). The genus mycoplasma and related genera (class mollicutes). *In The Prokaryotes*, pages 836–904. Springer.

Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R. et Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in caenorhabditis elegans. *nature*, 403(6772):901–906.

Richter, A. S. et Backofen, R. (2012). Accessibility and conservation: General features of bacterial small RNA-mRNA interactions. *RNA Biol*, 9(7):954–965.

Rivas, E. et Eddy, S. R. (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of molecular biology*, 285(5):2053–2068.

Rizk, G. et Lavenier, D. (2009). Gpu accelerated rna folding algorithm. *In Computational Science–ICCS 2009*, pages 1004–1013. Springer.

Rogers, K. et Chen, X. (2013). Biogenesis, turnover, and mode of action of plant microRNAs. *The Plant Cell Online*, 25(7):2383–2399.

Ruan, J., Stormo, G. D. et Zhang, W. (2004). An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 20(1):58–66.

Sagot, M.-F. (1998). Spelling approximate repeated or common motifs using a suffix tree. *In LATIN'98: Theoretical Informatics*, pages 374–390. Springer.

Scott, M. S. (2012). Diversity, overlap, and relationships in the small RNA landscape. *In Regulatory RNAs*, pages 23–48. Springer.

Sen, G. L. et Blau, H. M. (2006). A brief history of RNAi: the silence of the genes. *The FASEB journal*, 20(9):1293–1299.

Serra, M. J., Barnes, T. W., Betschart, K., Gutierrez, M. J., Sprouse, K. J., Riley, C. K., Stewart, L. et Temel, R. E. (1997). Improved parameters for the prediction of RNA hairpin stability. *Biochemistry*, 36(16):4844–4851.

Siomi, H. et Siomi, M. C. (2010). Posttranscriptional regulation of microRNA biogenesis in animals. *Molecular cell*, 38(3):323–332.

Siomi, M. C., Sato, K., Pezic, D. et Aravin, A. A. (2011). Piwi-interacting small RNAs: the vanguard of genome defence. *Nature reviews Molecular cell biology*, 12(4):246–258.

Siqueira, F. M., Thompson, C. E., Virginio, V. G., Gonchoroski, T., Reolon, L., Almeida, L. G., da Fonsêca, M. M., de Souza, R., Prosdocimi, F., Schrank, I. S. *et al.* (2013). New insights on the biology of swine respiratory tract mycoplasmas from a comparative genome analysis. *BMC genomics*, 14(1):175.

Sneath, P. H., Sokal, R. R. *et al.* (1973). *Numerical taxonomy. The principles and practice of numerical classification.*

Sokal, R. et Michener, C. (1958). A statistical method for evaluating systematic relationships. i'niv. kansas sci. bull., 38: 1409-1438. *Primary productivity and ecological factors in Lake Maggiore*, 127.

Steffen, P., Giegerich, R. et Giraud, M. (2010). Gpu parallelization of algebraic dynamic programming. *In Parallel Processing and Applied Mathematics*, pages 290–299. Springer.

Storz, G., Vogel, J. et Wassarman, K. M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. *Molecular cell*, 43(6):880–891.

Sun, W., Julie Li, Y.-S., Huang, H.-D., Shyy, J. Y. et Chien, S. (2010). microRNA: a master regulator of cellular processes for bioengineering systems. *Annual review of biomedical engineering*, 12:1–27.

Taft, R. J., Glazov, E. A., Lassmann, T., Hayashizaki, Y., Carninci, P. et Mattick, J. S. (2009). Small RNAs derived from snoRNAs. *Rna*, 15(7):1233–1240.

Tempel, S. et Tahi, F. (2012). A fast ab-initio method for predicting miRNA precursors in genomes. *Nucleic acids research*, 40(11):e80–e80.

Thomas, M., Lieberman, J. et Lal, A. (2010). Desperately seeking microRNA targets. *Nature structural & molecular biology*, 17(10):1169–1174.

Thomson, D. W., Bracken, C. P. et Goodall, G. J. (2011). Experimental strategies for microRNA target identification. *Nucleic acids research*, 39(16):6845–6853.

Turner, D. H. et Mathews, D. H. (2010). Nndb: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic acids research*, 38(suppl 1):D280–D282.

Van der Krol, A. R., Mur, L. A., Beld, M., Mol, J. et Stuitje, A. R. (1990). Flavonoid genes in petunia: addition of a limited number of gene copies may lead to a suppression of gene expression. *The Plant Cell Online*, 2(4):291–299.

Vanet, A., Marsan, L., Labigne, A. et Sagot, M.-F. (2000). Inferring regulatory elements from a whole genome. an analysis of *Helicobacter pylori* σ^{80} family of promoter signals. *Journal of molecular biology*, 297(2):335–353.

Vanet, A., Marsan, L. et Sagot, M.-F. (1999). Promoter sequences and algorithmical methods for identifying them. *Research in Microbiology*, 150(9):779–799.

Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X. et Li, Y. (2005). MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, 21(18):3610–3614.

Wightman, B., Ha, I. et Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in c. elegans. *Cell*, 75(5):855–862.

Williams, A. L. et Tinoco, I. (1986). A dynamic programming algorithm for finding alternative RNA secondary structure. *Nucleic acids research*, 14(1):299–315.

Winter, J. et Diederichs, S. (2011). MicroRNA biogenesis and cancer. *In MicroRNA and Cancer*, pages 3–22. Springer.

Winter, J., Link, S., Witzigmann, D., Hildenbrand, C., Previti, C. et Diederichs, S. (2013). Loop-mirs: active microRNAs generated from single-stranded loop regions. *Nucleic acids research*, 41(10):5503–5512.

Witkos, T., Koscianska, E. et Krzyzosiak, W. (2011). Practical aspects of microRNA target prediction. *Current molecular medicine*, 11(2):93.

Wu, Y., Wei, B., Liu, H., Li, T. et Rayner, S. (2011). Mirpara: a svm-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC bioinformatics*, 12(1):107.

Wuchty, S., Fontana, W., Hofacker, I. L., Schuster, P. *et al.* (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–165.

Xu, J., Li, C.-X., Li, Y.-S., Lv, J.-Y., Ma, Y., Shao, T.-T., Xu, L.-D., Wang, Y.-Y., Du, L., Zhang, Y.-P. *et al.* (2011). MiRNA–miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. *Nucleic acids research*, 39(3):825–836.

Xu, P., Vernooy, S. Y., Guo, M. et Hay, B. A. (2003). The *Drosophila* microRNA mir-14 suppresses cell death and is required for normal fat metabolism. *Current Biology*, 13(9):790–795.

Xu, Y., Zhou, X. et Zhang, W. (2008). MicroRNA prediction with a novel ranking algorithm based on random walks. *Bioinformatics*, 24(13):i50–i58.

Xue, C., Li, F., He, T., Liu, G.-P., Li, Y. et Zhang, X. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC bioinformatics*, 6(1):310.

Yang, J.-S., Phillips, M. D., Betel, D., Mu, P., Ventura, A., Siepel, A. C., Chen, K. C. et Lai, E. C. (2011). Widespread regulatory activity of vertebrate microRNA species. *Rna*, 17(2):312–326.

Yin, J. Q., Zhao, R. C. et Morris, K. V. (2008). Profiling microRNA expression with microarrays. *Trends in biotechnology*, 26(2):70–76.

Zamore, P. D. et Haley, B. (2005). Ribo-gnome: the big world of small RNAs. *Science*, 309(5740):1519–1524.

Zar, J. H. *et al.* (1999). *Biostatistical analysis*. Pearson Education India.

Zhang, Y., Yang, Y., Zhang, H., Jiang, X., Xu, B., Xue, Y., Cao, Y., Zhai, Q., Zhai, Y., Xu, M. *et al.* (2011). Prediction of novel pre-microRNAs with high accuracy through boosting and svm. *Bioinformatics*, 27(10):1436–1437.

Zheng, Y. et Zhang, W. (2010). Animal microRNA target prediction using diverse sequence-specific determinants. *Journal of Bioinformatics and Computational Biology*, 8(04):763–788.

Zuker, M., Mathews, D. H. et Turner, D. H. (1999). Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. *In RNA biochemistry and biotechnology*, pages 11–43. Springer.

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

Appendix A

List of the websites of the corresponding methods for miRNA prediction.

- miRFinder
  http://www.bioinformatics.org/mirfinder/
- MIReNA
  http://www.lgm.upmc.fr/mirena/index.htm
- miRD
  http://mcg.ustc.edu.cn/rpg/mird/mird.php
- RNAmicro
  http://www.bioinf.uni-leipzig.de/~jana/index.php/jana-hertel-software/65-jana-hertel
- SSCprofiler
  http://mirna.imbb.forth.gr/SSCprofiler.html
- MiRscan
  http://genes.mit.edu/mirscan/
- HHMMIR
  http://www.benoslab.pitt.edu/kadriAPBC2009.html
- MiRPara
  https://code.google.com/p/mirpara/wiki/miRPara
- CSHMM
  http://web.iitd.ac.in/~sumeet/mirna/
- miRank
  http://reccr.chem.rpi.edu/MIRank/
- miR-BAG
  http://scbb.ihbt.res.in/presents/mirbag/
- miRDeep
  https://www.mdc-berlin.de/8551903/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep
- ProMiR
  http://bi.snu.ac.kr/ProMiR/

- miRAlign
  http://bioinfo.au.tsinghua.edu.cn/miralign/
- miRNAFold
  http://evryrna.ibisc.univ-evry.fr/miRNAFold/

As mentioned in Section 3.3.2, we present the expression profile of the two precursors that give rise to miRNA api-miR-79. Figures 5.4-5.6 present the expression profiles of the reads mapping to the precursor mir-79-GL350203, during the three remaining developmental stages IE, LE, and L1. Figures 5.7-5.9 show the expression profiles for the precursor mir-79-GL349650 during the same three stages. For the latter precursor, the read counts in the figure are smaller than 10 (criterion 1), this is because these are the mappings of unique collapsed reads, when the reads are expanded the counts are larger than 10, hence fulfilling criterion 1.

Figure 5.4: *IE reads mapping to the precursor of miRNA mir-79-GL350203. For the sake of the presentation, the reads used in this figure were the unique reads of the IE sample because otherwise it would no fit in the page.*

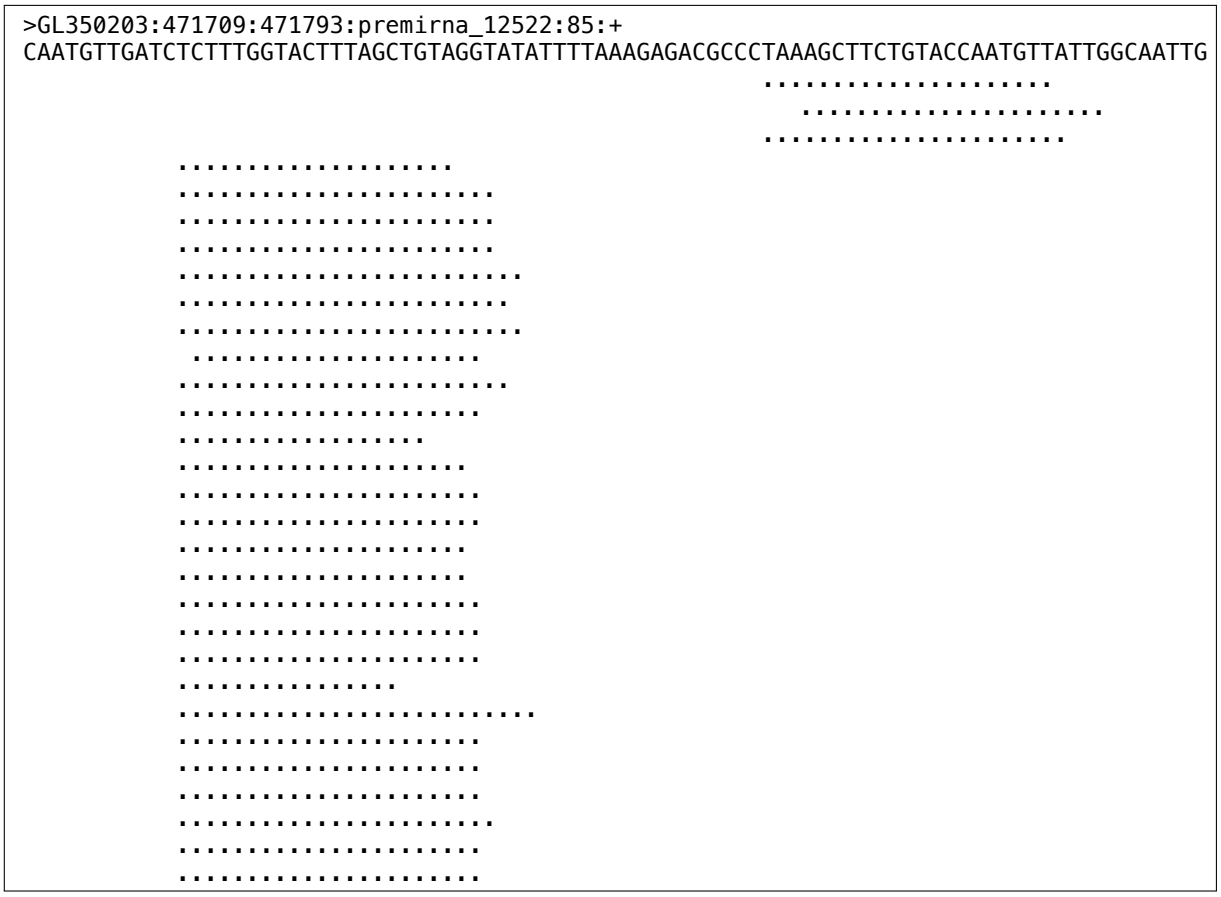


Figure 5.5: *LE reads mapping to the precursor of miRNA mir-79-GL350203. For the sake of the presentation, the reads used in this figure were the unique reads of the LE sample because otherwise it would no fit in the page.*

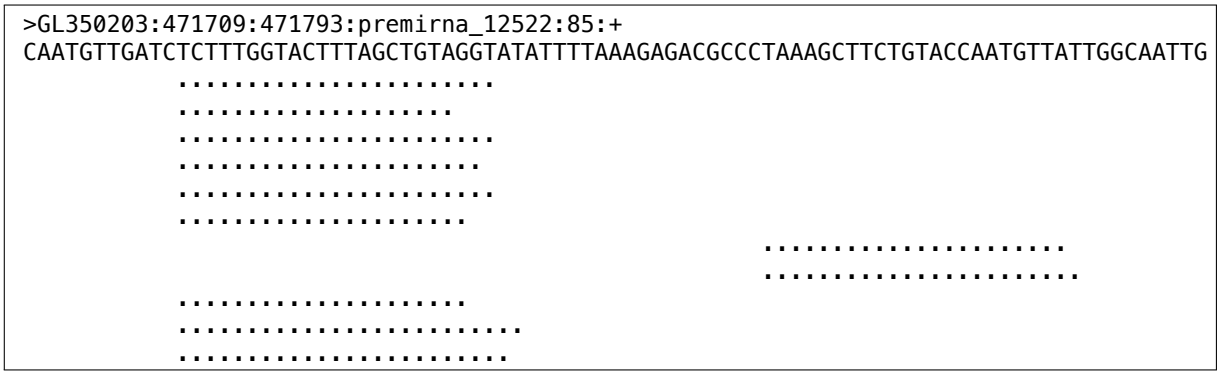


Figure 5.6: *L1 reads mapping to the precursor of miRNA mir-79-GL350203. For the sake of the presentation, the reads used in this figure were the unique reads of the L1 sample because otherwise it would no fit in the page.*

```
>GL349650:1158472:1158559:premirna_3079:88:+
TATGCGTTGTGTTATCTGGCTGTTGACTTTTTCCGAAACATTCAGCCTGGTTTTTCGGAAAATCAACGGGCTCGGTGCTGTGAAAAAA
                                                 ..........................
           ..........................
           ..........................
                                               ..........................
           ..........................
           ..........................
           ......................
                                               ..........................
                                                 ..........................
           ......................
           ......................
                                                 ..........................
           ......................
                                                  ..........................
```

Figure 5.7: *IE reads mapping to the precursor of miRNA mir-79-GL349650. For the sake of the presentation, the reads used in this figure were the unique reads of the IE sample because otherwise it would no fit in the page.*

```
>GL349650:1158472:1158559:premirna_3079:88:+
TATGCGTTGTGTTATCTGGCTGTTGACTTTTTCCGAAACATTCAGCCTGGTTTTTCGGAAAATCAACGGGCTCGGTGCTGTGAAAAAA
                                 ..........................
                                                 ..........................
           ..........................
                                               ..........................
           ......................
                                               ..........................
           ......................
           ..........................
                                               ......................
                                                 ..........................
           ......................
                                                  ..........................
```

Figure 5.8: *LE reads mapping to the precursor of miRNA mir-79-GL349650. For the sake of the presentation, the reads used in this figure were the unique reads of the LE sample because otherwise it would no fit in the page.*

```
>GL349650:1158472:1158559:premirna_3079:88:+
TATGCGTTGTGTTATCTGGCTGTTGACTTTTTCCGAAACATTCAGCCTGGTTTTTCGGAAAATCAACGGGCTCGGTGCTGTGAAAAAA
           ......................
           ......................
```

Figure 5.9: *L1 reads mapping to the precursor of miRNA mir-79-GL349650. For the sake of the presentation, the reads used in this figure were the unique reads of the L1 sample because otherwise it would no fit in the page.*

Titre: MiARN et compagnie: une exploration méthodologique du monde des petits ARNs

Résumé: La principale contribution de cette thèse est le développement d'une méthode fiable, robuste, et rapide pour la prédiction des pré-miARNs. Deux objectifs avaient été assignés : efficacité et flexibilité. L'efficacité a été rendue possible au moyen d'un algorithme quadratique. La majorité des prédicteurs publiés utilisaient un algorithme de complexité polynomiale de degré 3 pour évaluer la structure en épingle à tige-boucle des pré-miARNs, conduisant à des temps de calculs excessifs pour des données volumineuses. La flexibilité repose sur deux aspects, la nature des données expérimentales et la position taxonomique de l'organisme (en particulier plantes ou animaux). Mirinho accepte en entrée des séquences de génomes complets mais aussi les très nombreuses séquences résultant d'un séquençage massif de type NGS de "RNAseq". "L'universalité" taxonomique est obtenu par la possibilité de modifier les contraintes sur les tailles de la tige (double hélice) et de la boule terminale. Dans le cas de la prédiction des miARN de plantes la plus grande longueur de leur pré-miARN conduit à des méthodes d'extraction de la structure secondaire en tige-boule moins précises. Mirinho prend en compte ce problème lui permettant de fournir des structures secondaires de pré-miARN plus semblables à celles de miRBase que les autres méthodes disponibles. Mirinho a été utilisé dans le cadre de deux questions biologiques précises l'une concernant des RNAseq l'autre de l'ADN génomique. La première question a conduit à le traitement et l'analyse des données RNAseq de *Acyrthosiphon pisum*, le puceron du pois. L'objectif était d'identifier les miARN qui sont différentiellement exprimés au cours des quatre stades de développement de cette espèce et sont donc des candidats à la régulation des gènes au cours du développement. Pour cette analyse, nous avons développé un pipeline, appelé MirinhoPipe. La deuxième question a ermis d'aborder les problèmes liées à la prévision et l'analyse des ARN non-codants (ARNnc) dans la bactérie Mycoplasma hyopneumoniae. Alvinho a été développé pour la prédiction de cibles des miRNA autour d'une segmentation d'une séquence numérique et de la détection de la conservation des séquences entre ncRNA utilisant un graphe k-partite. Nous avons finalement abordé un problème lié à la recherche de motifs conservés dans un ensemble de séquences et pouvant ainsi correspondre à des éléments fonctionnels. L'originalité de la méthode réside dans la complexité des motifs recherchés qui peuvent être constitué de sous motifs séparés. Cela avait déjà été abordée dans une méthode robuste appelé Smile mais conduisant à des sorties très volumineuses et difficilement interprétables. Nous avons développé des solutions utilisant des méthodes de classification pour regrouper les motifs pouvant correspondre à un même élément biologique. Cette approche permet de mieux distinguer les motifs biologiquement pertinents de séquences apparaissant de manière aléatoire.

Mots-Clefs : pre-microARN; programmation dynamique; modèle de énergie du plus proche voisin; prédiction; sequecançage des petit ARNs; puceron du pois; cibles de ARN non-codants; motifs

Title: MiRNA and co: Methodologically exploring the world of small RNAs

Abstract: The main contribution of this thesis is the development of a reliable, robust, and much faster method for the prediction of pre-miRNAs. With this method, we aimed mainly at two goals: efficiency and flexibility. Efficiency was made possible by means of a quadratic algorithm. Since the majority of the predictors use a cubic algorithm to verify the pre-miRNA hairpin structure, they may take too long when the input is large. Flexibility relies on two aspects, the input type and the organism clade. Mirinho can receive as input both a genome

sequence and small RNA sequencing (sRNA-seq) data of both animal and plant species. To change from one clade to another, it suffices to change the lengths of the stem-arms and of the terminal loop. Concerning the prediction of plant miRNAs, because their pre-miRNAs are longer, the methods for extracting the hairpin secondary structure are not as accurate as for shorter sequences. With MIRINHO, we also addressed this problem, which enabled to provide pre-miRNA secondary structures more similar to the ones in miRBase than the other available methods. MIRINHO served as the basis to two other issues we addressed. The first issue led to the treatment and analysis of sRNA-seq data of *Acyrthosiphon pisum*, the pea aphid. The goal was to identify the miRNAs that are expressed during the four developmental stages of this species, allowing further biological conclusions concerning the regulatory system of such an organism. For this analysis, we developed a whole pipeline, called MIRINHOPIPE, at the end of which MIRINHO was aggregated. We then moved on to the second issue, that involved problems related to the prediction and analysis of non-coding RNAs (ncRNAs) in the bacterium *Mycoplasma hyopneumoniae*. A method, called ALVINHO, was thus developed for the prediction of targets in this bacterium, together with a pipeline for the segmentation of a numerical sequence and detection of conservation among ncRNA sequences using a k-partite graph. We finally addressed a problem related to motifs, that is to patterns, that may be composed of one or more parts, that appear conserved in a set of sequences and may correspond to functional elements. This had already been addressed in a robust method called Smile. However, depending on the input parameters, the output may be too large to be tractable, as was realized in other works of the team. We then presented some clustering solutions to group the motifs that may correspond to a same biological element, and thus to better distinguish the biologically significant ones from noise that may be present in what often are large outputs from many motif extraction algorithms.

Keywords: pre-microRNA; dynamic programming; nearest neighbor energy model; prediction; small RNA sequencing; pea aphid; non-coding RNA target; motifs