Efficient algorithms for the identification of miRNA motifs in DNA sequences

Nuno D Mendes

UNIVERSIDADE TÉCNICA DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

UNIVERSITÉ CLAUDE BERNARD
LYON 1

Efficient algorithms for the identification of miRNA motifs in DNA sequences

Nuno Miguel Dias Mendes
(Mestre)

Dissertação para obtenção do Grau de Doutor em Engenharia Informática e de Computadores

Orientadores:	Doutora	Ana Teresa Correia de Freitas
	Doutora	Marie-France Sagot

Júri

Presidente:	Presidente do Conselho Científico do IST	
Vogais:	Doutor	Christian Gautier
	Doutor	Arlindo Manuel Limede Oliveira
	Doutora	Marie-France Sagot
	Doutora	Cecília Maria Pais de Faria Andrade Arraiano
	Doutora	Ana Maria Nobre Vilhena Nunes Pires de Melo Parente
	Doutora	Ana Teresa Correia de Freitas
	Doutora	Christine Gaspin

Junho de 2011

Acknowledgments

This thesis was the product of several years of work which, not unlike other academic endeavours, was punctuated by alternating periods of inspiration and frustration. It would not have been possible to complete this work without the support of several other people, at both the academic and personal levels.

This journey started at the Instituto Gulbenkian de Ciência, in the context of its pioneering PhD Program in Computational Biology. This program gave me the opportunity to contact many world-class experts in several topics of this field of research, paving the way to the choice of a project which would later result in this manuscript. In this regard, I have to thank the PhD Program and all the lecturers who took on to train a very heterogeneous group of young researchers which auspiciously included me. I must also thank the Fundação para a Ciência e a Tecnologia for funding my work by granting me a PhD scholarship.

I would also like to thank everyone at the BAOBAB Team of LBBE for kindly receiving me for the greater part of my doctoral work and the fellow researchers at KDBIO who accompanied me throughout these last few months and who created a truly welcoming environment.

I am in debt to several researchers who were kind enough to share their insights into my work over the years, namely Eric Westhof, Robert Giegerich, Eduardo Rocha, Rolf Backofen, just to name a few.

I am also deeply thankful to my family and friends for their relentless support, even in the face of my occasional negligence. And a special word to my mother, who always encouraged me and who has been nothing but thrilled that I chose the career she once dreamt of for herself.

I would also like to express my gratitude to the members of the jury who have been kind enough to accept being part of the thesis committee.

A final word of appreciation to my advisors, who always believed in my ability to complete this project in spite of my occasional hesitations. A special thanks to Ana Teresa Freitas, who has always nurtured my passion for research since the early days of my graduation thesis and who has always made me see the positive outlook of things. And to Marie-France Sagot, whose wisdom and uncompromising passion for science taught me, above all, that no matter how much a problem has been studied, there is always a fresh perspective and new insights to contribute with.

Abstract

Unravelling biological processes is dependent on the adequate modelling of regulatory mechanisms that determine the timing and spatial patterns of gene expression. In the last decade, a novel regulatory mechanism has been discovered and its biological importance has been increasingly recognised. This mechanism is mediated by RNA molecules named miRNAs that are the product of the maturation of non-coding gene transcripts and act post-transcriptionally usually to dampen or abolish the expression of protein-coding genes.

Despite having eluded detection for such a long time, it is now clear that the elucidation of the expression pattern of many genes cannot be achieved without incorporating the effects of miRNA-mediated regulation.

The technical difficulties that the experimental detection of these regulators entailed prompted the development of increasingly sophisticated computational approaches. Gene finding strategies originally developed for coding genes cannot be applied since these non-coding molecules are subject to very different sequence restraints and are too short to exhibit statistical properties that can be easily distinguished from the background. As a result, computational tools came to rely heavily on the identification of conserved sequences, distant homologs and machine learning techniques.

Recent developments in sequencing technology have overcome some of the limitations of earlier experimental approaches, but pose new computational challenges. At present, the identification of new miRNA genes is therefore the result of the use of several approaches, both computational and experimental.

In spite of the advancement that this research field has known in the last several years, we are still not able to formally and rigourously characterise miRNA genes in order to identify whichever sequence, structure or contextual requirements are needed to turn a DNA sequence into a functional miRNA.

Efforts using computational algorithms towards the enumeration of the full set of miRNAs of an organism have been limited by strong reliance on arguments of precursor conservation and feature similarity. However, miRNA precursors may arise anew or be lost across the evolutionary history of a species and a newly-sequenced genome may be evolutionarily too distant from other genomes for an adequate comparative analysis. In addition, the learning of intricate classification rules based purely on features shared by miRNA precursors that are currently known may reflect a perpetuating identification bias rather than a sound means to tell true miRNAs from other genomic stem-loops.

In this thesis, we present a strategy to sieve through the vast amount of stem-loops found in metazoan genomes in search of pre-miRNAs, significantly reducing the set of candidates while retaining most known miRNA precursors. Our approach relies on precursor properties derived from the current knowledge of miRNA biogenesis, analysis of the precursor structure and incorporation of information about the transcription potential of each candidate.

Our approach has been applied to the genomes of *Drosophila melanogaster* and *Anopheles gambiae*, which has allowed us to show that there is a strong bias amongst annotated pre-miRNAs towards robust stem-loops in these genomes and to propose a scoring scheme for precursor candidates which combines four robustness measures. Additionally, we have identified several known pre-miRNA homologs in the newly-sequenced *Anopheles darlingi* and shown that most are found amongst the top-scoring precursor candidates for that organism, with respect to the combined score. The structural analysis of our candidates and the identification of the region of the structural space where known precursors are usually found allowed us to eliminate several candidates, but also showed that there is a staggering number of genomic stem-loops which seem to fulfil the stability, robustness and structural requirements indicating that additional evidence is needed to identify functional precursors. To this effect, we have introduced different strategies to evaluate the transcription potential of the remaining candidates which vary according to the information which is available for the dataset under study.

Keywords: miRNA, gene finding, single-genome, robustness, stability, secondary structure

Resumo

A compreensão dos processos biológicos está dependente da modelação adequada dos mecanismos reguladores que determinam os padrões temporais e espaciais da expressão génica. Na última década, um novo mecanismo regulatório foi descoberto e a sua importância tem sido crescentemente reconhecida. Este mecanismo é mediado por moléculas de RNA designadas de miRNAs que são o produto da maturação de transcritos de genes não-codificantes e actuam de forma pós-transcricional de modo a, em geral, atenuar ou suprimir a expressão de genes codificantes.

Apesar da sua existência não ter sido detectada durante muito tempo, é hoje evidente que a descrição dos padrões de expressão de muitos genes não pode ser feita sem incorporar os efeitos da regulação mediada por miRNAs.

As dificuldades técnicas associadas à detecção experimental destes reguladores levaram ao desenvolvimento de abordagens computacionais cada vez mais sofisticadas. Estratégias de procura de genes originalmente desenvolvidas para genes codificantes não podem ser aplicadas porques estas moléculas não-codificantes estão sujeitas a constrangimentos ao nível da sequência de natureza muito diferente e são demasiado curtas para exibirem propriedades estatísticas que possam facilmente ser distinguidas das sequências circundantes. Daqui resulta que as ferramentas computationais tenham acabado por depender da identificação de sequências conservadas, de sequências homólogas distantes e do recurso a técnicas de aprendizagem.

Desenvolvimentos recentes no domínio da tecnologia de sequenciação superaram algumas das limitações das primeiras abordagens experimentais, mas introduzem novos desafios computacionais. Actualmente, a identificação de novos genes de miRNA é o resultado da aplicação de diversas abordagens tanto computacionais como experimentais.

Apesar dos avanços que este domínio de investigação conheceu nos últimos anos, ainda não é possível caracterizar genes de miRNA de um modo formal e rigoroso por forma a identificar os requisitos ao nível da sequência, da estrutura e do contexto genómico que são necessários para transformar uma sequência de DNA num miRNA funcional.

Os esforços computacionais desenvolvidos no sentido de enumerar o catálogo completo de miRNAs de um organismo têm sido limitados pela forte dependência que estes apresentam em relação a argumentos de conservação de precursores ou de semelhança de características. No entanto, os precursores de miRNA podem surgir ou desaparecer no decurso da história evolutiva de uma espécie e um genoma recém-sequenciado pode estar evolutivamente demasiado distante de outro genoma para se proceder a uma análise comparativa adequada. Por outro lado, a aprendizagem de regras de classificação complexas baseadas apenas em características partilhadas por precursores de miRNA que são actualmente conhecidos pode reflectir antes uma tendência para perpetuar um enviezamento quanto ao tipo de precursores identificados do que constituir efectivamente um forma de distinguir miRNAs de outras estruturas

semelhantes presentes no genoma.

Nesta tese é apresentada uma estratégia para filtrar o vasto número de potenciais candidatos identificáveis em genomas animais, reduzindo significativamente o seu número muito embora retendo a grande maioria dos precursores conhecidos. A abordagem usada apoia-se em propriedades exibidas pelos precursores e que foram deduzidas a partir do actual conhecimento do processo de biogénese dos miRNAs, na análise da estrutura secundária dos precursores e na incorporação de informação sobre o potencial de transcrição de cada candidato.

A abordagem foi aplicada aos genomas de *Drosophila melanogaster* e *Anopheles gambiae*, o que permitiu mostrar que há uma forte tendência da parte dos pre-miRNAs anotados no sentido de apresentarem estruturas secundárias robustas e propôr um esquema de classificação para candidatos que combina quatro medidas de robustez. Adicionalmente, foi possível identificar vários homólogos de pre-miRNAs conhecidos no genoma do recém-sequenciado *Anopheles darlingi* e mostrar que estes se encontram entre os candidatos melhor classificados naquele organismo em relação à combinação das medidas de robustez. A análise estrutural dos candidatos e a identificação da região do espaço das estruturas secundárias onde os precursores conhecidos se encontram permitiu a eliminação de muitos candidatos, mas permitiu também observar que um grande número de estruturas no genoma parece cumprir os requisitos de estabilidade, robustez e de estrutura secundária, ilustrando a necessidade de recorrer a meios adicionais para identificar precursores funcionais. Para este efeito, foram introduzidas diversas estratégias para avaliar o potencial de transcrição dos candidatos, que variam consoante a informação disponível para o genoma em estudo.

Palavras-chave: miRNA, procura de genes, genoma único, robustez, estabilidade, estrutura secundária

Résumé

La compréhension des processus biologiques est dépendente de la modélisation adéquate des mécanismes de régulation qui déterminent la répartition spatiale et temporelle de l'expression génique. Au début de la dernière décennie, un nouveau mécanisme de régulation a été découvert et son importance biologique a été de plus en plus reconnue. Ce mécanisme est médié par des molécules d'ARN appelés miRNAs qui sont le produit de la maturation de transcrits de gènes non-codants et qui agissent d'une façon post-transcriptionnelle afin d'atténuer ou supprimer l'expression de gènes codants.

Quoi qu'ils aient longtemps échappé à la détection il est maintenant évident que les variations de l'expression de nombreux gènes ne peuvent être comprises sans intégrer les effets de la régulation mediée par des miRNAs.

Les difficultés techniques de la détection expérimentale de ces régulateurs a suscité le développement d'approches computationnelles de plus en plus sophistiquées. Les stratégies initialement développées pour la recherche des gènes codants ne peuvent pas être appliquées car ces molécules non codantes sont soumises à très différentes restrictions au niveau de la séquence et sont trop courtes pour présenter des propriétés statistiques peuvant être facilement distinguées de celles des séquences parmis lesquelles elles se trouvent. Par conséquent, les outils de calcul comptent beaucoup sur l'identification des séquences conservées, sur les homologues lointains et les techniques d'apprentissage automatique.

Les développements récents dans la technologie de séquençage ont réussi à surmonter certaines des anciennes limites des approches expérimentales, mais posent de nouveaux défis informatiques. À l'heure actuelle, l'identification de nouveaux gènes de miRNA est donc le résultat de l'utilisation de plusieurs approches, à la fois computationelles et expérimentales.

Malgré l'avancement que ce domaine de recherche a connu dans les dernières années, nous ne sommes toujours pas capables de caractériser les gènes de miRNA de façon formelle et rigoureuse, de manière à identifier les contraintes au niveau de la séquence, de la structure ou du contexte génomique nécessaires à la formation de miRNAs fonctionnels.

L'utilisation d'algorithmes de calcul pour l'énumération de l'ensemble des miRNAs d'un organisme est limitée par une forte dépendance sur des arguments de conservation de précurseurs et de la similitude de caractéristiques. En effet, les précurseurs de miRNA peuvent apparaître *de novo* ou être perdus au long de l'histoire évolutive d'une espèce. De plus, un génome récemment séquencé peut être évolutivement trop éloigné d'autres génomes pour que l'on puisse faire une analyse comparative adéquate. Finalement, l'apprentissage des règles de classification complexes en se basant uniquement sur des caractéristiques communes aux plusieurs précurseurs de miRNA qui sont actuellement connus peut perpétuer un biais d'identification plutôt que de trouver vraiment un moyen de distinguer les miRNAs d'autres tiges-boucles génomiques.

Dans cette thèse, nous présentons une stratégie pour filtrer la grande quantité de tiges-

boucles se trouvant dans les génomes de métazoaires, réduisant de manière significative l'ensemble des candidats, tout en conservant la plupart des précurseurs de miRNA connus. Notre approche repose sur les propriétés des précurseurs provenant de l'état actuel des connaissances de la biogenèse des miRNAs, de l'analyse de la structure des précurseurs et de l'incorporation d'informations sur le potentiel de transcription de chaque candidat.

Notre approche a été appliquée aux génomes de *Drosophila melanogaster* et *Anopheles gambiae* ce qui nous a permis de montrer que dans ces génomes les pre-miRNAs annotés ont tendance à correspondre aux tiges-boucles robustes et aussi à proposer un systéme de notation des candidats qui combine quatre mesures de robustesse. Notamment, nous avons identifié plusieurs homologues de pre-miRNAs connus dans le génome récemment séquencé d'*Anopheles darlingi* et nous avons montré que la plupart se trouvent parmi les candidats à plus haute notation (selon notre combinaison de mesures) pour cet organisme. L'analyse structurelle de nos candidats et l'identification de la région de l'espace des structures où se trouvent généralement les précurseurs connus nous a permis d'éliminer plusieurs candidats, mais a également montré qu'il existe un nombre impressionnant de tiges-boucles génomiques qui semblent répondre aux exigences de stabilité, de robustesse et de structure. Ceci indique que des méthodes supplémentaires sont nécessaires pour identifier les précurseurs fonctionnels. Pour cela nous avons mis en place différentes stratégies pour évaluer le potentiel de transcription des candidats restants qui varient en fonction des informations qui sont disponibles pour les données analyseés.

Mots-clés : miRNA, recherche de gènes, génome isolé, robustesse, stabilité, structure secondaire

Contents

List of Figures

List of Tables

Glossary

DNA	-	Deoxyribonucleic acid molecule composed of two strands of nucleotides forming a double helix. It is the carrier of genetic information necessary to all cellular activities.
RNA	-	Ribonucleic acid molecule.
Prokaryotes	-	Organisms which lack a nuclear envelope.
Eukaryotes	-	Organisms in possession of a nuclear envelope.
Chromatin	-	Complex of DNA and proteins found in eukaryotic cells.
Nucleus	-	Organelle found in eukaryotic cells which contains most of their genetic material.
Cytosol	-	Consists of the internal fluid of the cell where most of its metabolism occurs.
Aminoacid	-	Molecule that contains an amino and a carboxylic acid functional group.
Protein	-	Organic compound consisting of aminoacids joined by peptide bonds. It is essential to the structure and function of all living cells.
Enzyme	-	Biopolymer that catalyzes chemical reactions. Most enzymes are proteins although some are made of RNA or DNA.
Primary structure	-	the sequence of nucleotides of a nucleic acid molecule.
Secondary structure	-	the two-dimensional organisation of a nucleic acid molecule (usually RNA) specifying which nucleotides are forming base-pairs with one another. A common representation of secondary structure is the dot-parenthesis notation. Each unpaired nucleotide is represented by a dot '.', whereas a left-hand side of a base-pair is represented by a left parenthesis '(', and the right-hand side of a base-pair is represented by a matching right parenthesis ')'.
Stem-loop	-	also called hairpin, is a single-strand RNA structure folded on itself giving rise to a stem (a stack of base-pairs with possibly some intervening unpaired nucleotides called bulges or inner loops) and a terminal loop (a string of unpaired nucleotides at the middle portion of the sequence, flanked by both stem arms).

Part I

Background

Chapter 1

Introduction

Despite a growing list of miRNAs, identified either by experimental assays or using current computational tools, the goal of enumerating the full catalogue of miRNAs of any single organism has proven to be difficult, requiring different approaches to identify a decreasing number of novel regulators. A recent thorough experimental study of mammalian miRNAs did find new regulators, but it also showed that several annotated sequences were likely not miRNAs [24]. The difficulties are two-fold. On the one hand, purely experimental detection is limited to miRNAs which are expressed at relatively high levels and in broad cellular types/conditions. Recent deep-sequencing techniques tackle these limitations but require extensive computational analyses [30]. On the other hand, computational miRNA gene finding tools are strongly dependent on conservation criteria and other sequence/structure similarities with previously identified miRNA precursors which limits their power to identify novel miRNAs, particularly those which are not conserved [83].

Single-genome approaches are increasingly necessary given the fact that a growing number of genome sequencing projects are under way for which no evolutionarily close genome is available and for which one cannot otherwise hope to thoroughly explore the miRNA landscape.

Considering that we have but rudimentary models of miRNA precursor evolution, which makes it hard to interpret the biological significance of conservation data, and that we lack a deep understanding of the structural requirements for efficient pre-miRNA processing, we believe that if we are to increase our knowledge of the miRNA repertoire of an organism, our efforts should privilege general properties that are known to characterise miRNA precursors. These properties should not necessarily emerge from rules learnt from the detailed analysis of previously known precursors, but should rather focus on features that, in principle, distinguish pre-miRNAs for other hairpins.

1.1 Problem

From amidst the several problems in the field of miRNA-mediated gene regulation, this thesis focus on the identification of miRNA precursors without recourse to conservation information by developing efficient and sensitive candidate extraction and evaluation methods.

From a computational biology perspective the pursuit of this objective entails the definition of a computational model to answer the question "What is a miRNA precursor?".

In order to fulfil this goal the following objectives are to be achieved:

1. Specification of a computational model to identify pre-miRNAs, integrating our current knowledge of miRNA biogenesis, and known characteristics of miRNA precursors.

2. Extraction of miRNA gene candidates using either an *ab initio* whole-genome approach or a set of candidate transcripts from RNA extraction assays, using a method that is both efficient and highly sensitive and which can serve as a starting point for several filtering and ranking procedures.

3. Development of measures and statistical methods to evaluate the quality of the extracted candidates and their putative biological function which are able to provide a series of evidence-based arguments for the likelihood that a given candidate is a functional pre-miRNA.

4. Integration of the developed algorithms in a support tool for research in functional genomics.

1.2 Approach

The approach proposed in this thesis is based on a three-pillar evaluation of candidates extracted from a single-genome, which is illustrated in Figure 1.1.

The first pillar refers to the combination of four measures of stability and robustness, collectively called *intrinsic measures*. These measures purport to identify candidates that have the features known to be a hallmark of miRNA precursors and refer to structural stability, but also mutational and contextual robustness. The second pillar consists in the attempt to characterise the region of the folding space that the cellular machinery involved in the miRNA maturation pathway is likely to recognise. Finally, the third pillar concerns the integration of transcription information to assess whether candidates are likely to be efficiently transcribed which is, of course, a necessary condition for the precursor to ever arise.

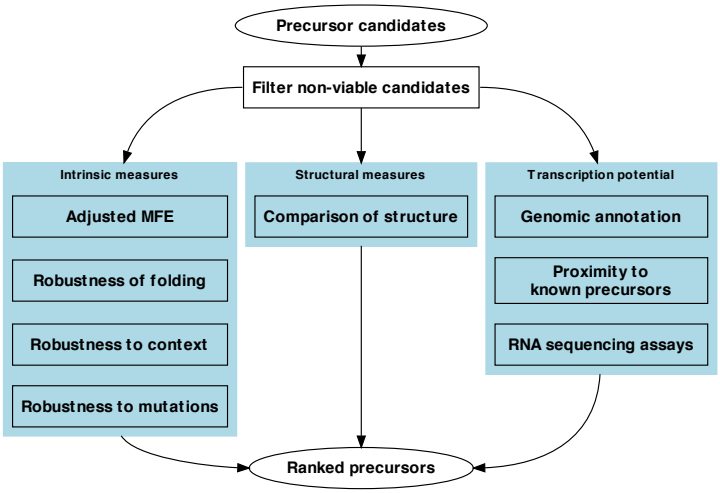

Figure 1.1: Diagram of the three-pillar candidate classification procedure

1.3 Contributions

The main contributions of this thesis refer to the development of an evidence-based computational model to the identification of miRNA precursors in animal genomes. To this effect we developed the following:

1. A genome-wide candidate extraction method, as well as a method to identify and statistically evaluate genome hits of transcripts originating from small RNA sequencing assays, which can be used to either infer precursor candidates from those hits or identify *ab initio* candidates with transcriptional evidence.

2. A method to combine stability and robustness measures in order to evaluate precursor candidates which was eventually published in *BMC Genomics* [84].

3. A method to represent RNA structures in a multidimensional space where relative distances reflect sequence/structural similarity as evaluated by existing structural clustering methods which allows the identification of a limited region of the multidimensional space where miRNA precursors tend to occur.

4. The integration of the developed tools in a publicly accessible framework, allowing a web-based exploration of the results

The developed methods have been applied to the genomes of *Drosophila melanogaster*, *Anopheles gambiae* and, partially, to the newly-sequenced *Anopheles darlingi*.

Additionally, we have contributed with the most recent survey on current tools for miRNA gene finding and miRNA target identification which was published in *Nucleid Acids Research* [83].

1.4 Thesis outline

This thesis is organized as follows:

Chapter 2 gives the necessary context to understand the biological questions addressed by this thesis. The elementary notions of miRNA biology are presented with a special focus on the aspects that are more relevant to the adopted approach.

Chapter 3 presents the current state of the art systematically describing and classifying in broad families the existing approaches to computational miRNA gene finding.

Chapter 4 presents our single-genome approach to distinguishing pre-miRNAs from other genomic hairpins including the whole-genome candidate extraction procedure, a set of stability and robustness measures used to evaluate miRNA precursor candidates as well as the procedure used to combine them into a single score, a vectorial representation of sequence/structure features of precursor candidates which is used to identify the candidates most likely to exhibit the structural requirements associated with pre-miRNAs, a few strategies to evaluate the transcription potential of a given precursor candidate and a procedure to map sequenced small RNAs back to the genome and, additionally, the CRAVELA framework is presented with a discussion of implementation details and a presentation of the web-based tool used to visualise the information produced by the procedures discussed in this chapter.

Chapter 5 describes and discusses the results obtained with our approach for the genomes of *Drosophila melanogaster*, *Anopheles gambiae*, and *Anopheles darlingi*.

Chapter 6 presents the final conclusions of this thesis, discusses future improvements and new developments to the CRAVELA framework and elaborates on the current perspectives for the research field on miRNA-mediated gene regulation in the context of computational biology.

Chapter 2

Biological background

Not so long ago, in the 19th century, Darwin gave us an account of how such complicated things like living beings can naturally occur, placing every extant organism in the same grand family tree and uniting all Biology under the same theoretical framework. His ideas have been enriched by many scientists since then, particularly with the rediscovery of Mendel's work and the characterization of DNA as a digital repository of information. This epistemological revolution elicited a novel approach to Biology which would henceforth seek mechanistic and, later, quantitative models rather than mere descriptions of the natural world.

Many researchers outside the field of Biology have been drawn to this area namely mathematicians, computer scientists, chemists, and physicists. This multidisciplinary trend is explained not only by the increasing complexity and depth of the investigations in the life sciences, but also by the fact that biological systems have proven to be a remarkably rich field for the application of theoretical methods developed in the context of other disciplines. On the other hand, Biology has also, on many occasions, served as an inspiration to novel approaches to problems arising from its tributary sciences.

Since this work is at the crossroads of Computer Science and Biology, always privileging the underlying biological questions, it is important to give enough detail about the biological problems that both motivate and inform our investigations.

In this chapter, we present the key biological concepts behind the questions we purport to address. First, we discuss basic notions of molecular biology and then we proceed to present the central aspects of what is currently known about microRNA biogenesis and function.

2.1 Fundamentals of molecular biology

2.1.1 Structure of nucleic acids

The field of molecular biology greatly benefited from the discovery of the three-dimensional structure of DNA by Watson and Crick in 1953 [124]. The DNA molecule, present in all living

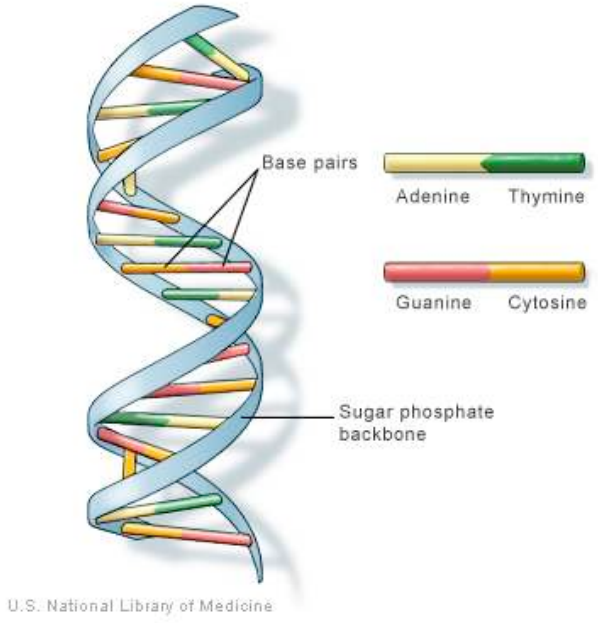

Figure 2.1: DNA double-helix structure. [90]

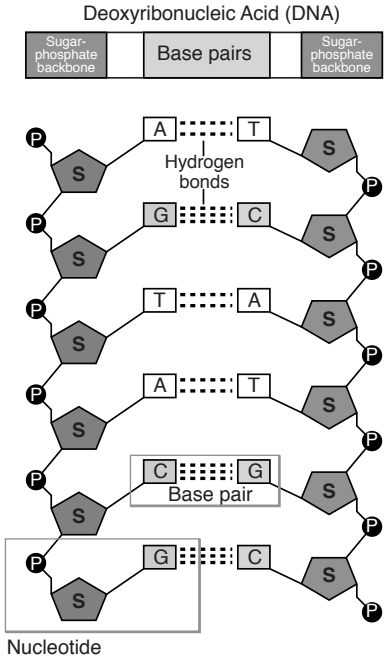

Figure 2.2: Nucleotides and the structure of DNA. [48]

cells, is the carrier of genetic information which is necessary to control cellular activities. This information is passed down to each new generation almost flawlessly. DNA is composed of two strands of nucleotides forming a double helix (Figure 2.1). A nucleotide is a molecule formed by a pentose (deoxyribose in DNA), a phosphate group and a nitrogenous base. There are four such nucleotides found in DNA, differing only on their nitrogenous base: Adenine (A), Guanine (G), Cytosine (C) and Thymine (T).

The pentose sugar-phosphate links form the backbone of the DNA molecule and are located in the exterior of the double helix. The two strands of DNA are kept together by hydrogen bonds linking each pair of bases. In a complete helix, Adenine always pairs with Thymine and Cytosine always pairs with Guanine. Because of this, the two strands are said to be complementary (Figure 2.2). Each strand of DNA (and RNA) has two ends designated 5'-end and 3'-end. The 5'-end refers to the end which has the fifth carbon of the pentose at its terminus. Similarly, the 3'-end (or tail end) terminates at the hydroxyl group of the third carbon of the pentose.

The information contained in DNA is represented by the specific sequence of nucleotides in either strand (the sequence of nucleotides in the complementary strand can be inferred considering the base-pairing scheme discussed earlier). It is thus a digital repository of information consisting of a text written with a four-letters alphabet.

Although DNA is structurally identical in all living cells, in prokaryotes consists of a single circular molecule whereas in eukaryotes is found associated with several proteins to form a complex named chromatin, which is located in the nucleus [26].

However, not all regions of DNA seem to carry information. Those regions which do carry information are named *genes* and if the information is used to produce proteins these are said to be coding regions. Genes are expressed as final products that generally consist of proteins which can serve different purposes: they can form part of the cell wall, act as catalytic components (enzymes) or influence the expression of genes and are, therefore, actors in virtually all cellular activities. In eukaryotic cells it is common to find genes which contain large amounts of noncoding regions. In these genes, coding regions named exons are separated by noncoding regions named introns. Additionally, there are two regions at the 5' and 3' ends of a transcribed gene that are also noncoding and are named 5' and 3' UTR (untranslated region), respectively.

RNA is another nucleic acid related to DNA. There are some important differences between these two molecules. Firstly, unlike DNA, RNA is a single stranded molecule. The pentose found in RNA is ribose and not deoxyribose (hence the name of the molecules) and the nucleotide Thymine is substituted by Uracil (U) (Figure 2.3). Despite being a single stranded molecule, RNA sometimes presents loops where homologous portions of the molecule self-hybridize. Neither the different sugar nor the base substitution alter the base-pairing scheme found in DNA, but additional non-canonical pairings are frequently observed, particularly

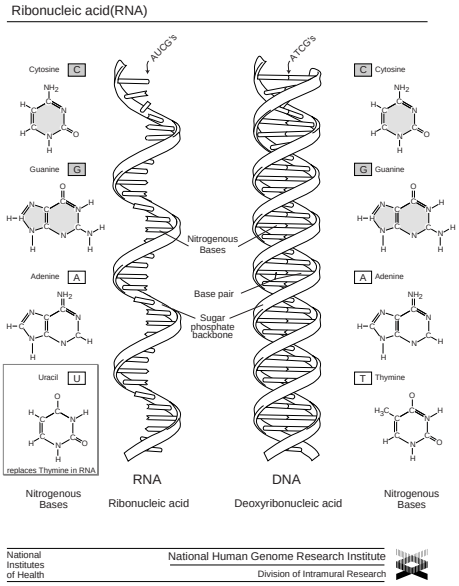

Figure 2.3: RNA versus DNA. [70]

G:U. In livings cells, one can always find larger quantities of RNA than of DNA and the amount of RNA varies with changing metabolic conditions whereas the amount of DNA is constant (in cells which are not in the process of cell division). This is consistent with the fact that RNA is a fundamental intermediary in the expression of genetic information.

2.1.2 Gene expression

The central dogma of molecular biology [26] establishes a pathway for the flow of genetic information: DNA → RNA → protein, i.e., from the DNA repository to the final products of gene expression. The first process in which RNA molecules are synthesized from a DNA template is called *transcription*. The RNA molecule thus obtained is called Messenger RNA (mRNA). The subsequent process in which mRNA is used as a template for protein synthesis is called *translation.*

In prokaryotes, transcription and translation occur almost simultaneously whereas in eukaryotes the two processes take place in different parts of the cell. In these organisms the transition from transcription to translation involves the migration of mRNA from the nucleus to the cytosol alongside with certain modifications to the mRNA molecule in a process called maturation. Figure 2.4 shows a schematic representation of the different processes involved in gene expression for both prokaryotes and eukaryotes.

The typical products of gene expression – proteins – consist of sequences of aminoacids.

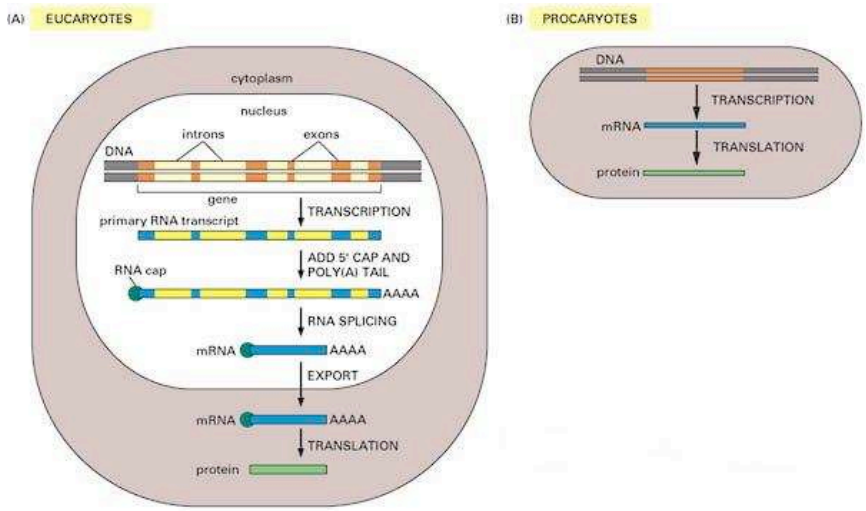

Figure 2.4: Schematic representation of the processes involved in gene expression in prokaryotes and eukaryotes [2]

Proteins, as we mentioned earlier, have a central role in all cellular activities and their function depends on their three-dimensional structure which, in turn, is derived from the specific linear ordering of their constituent aminoacids.

The genes of an organism are not all simultaneously expressed. Their activation depends on the current needs of the cell and is subjected to various regulatory mechanisms. One of the most important mechanisms is the transcriptional regulation. Some of the noncoding regions of DNA play a fundamental role in the regulation of transcription. These regions (regulatory regions) contain small sequences of nucleotides, which are recognized by proteins associated with the transcription machinery. The most common regulatory regions are located upstream of the start of transcription and are called promoter regions or, in a broader sense, cis-regulatory regions. The presence of these sequence motifs is essential for the efficient binding of the cellular transcription machinery. Different motifs can play different roles in gene expression. While some are critical for eliciting the start of transcription others recruit proteins which act as activators or repressors.

RNA polymerase is responsible for the transcription process. This enzyme, when examined *in vitro*, transcribes DNA into RNA but initiates at nonspecific sites on the DNA.

Transcriptional regulation in eukaryotes [26] is considerably complex. In fact, eukaryotes use different types of RNA polymerase for different purposes. The most studied type which is also responsible for transcribing most genes is RNA polymerase II (RNA pol II). This type of RNA polymerase requires several general transcription factors to form a functional transcription initiation complex.

The regulation of transcription in eukaryotes is primarily made at the level of initiation

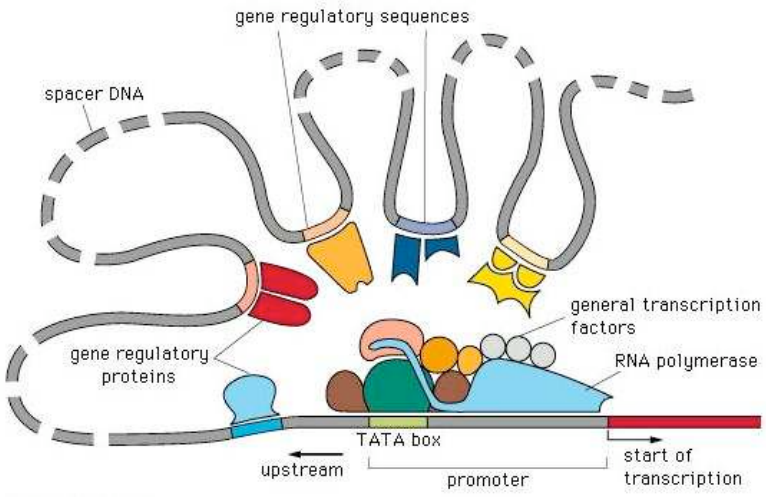

Figure 2.5: Schematic representation of regulation in eukaryotes [26]

of transcription although it may be attenuated or stimulated at subsequent steps. Many genes in eukaryotic cells are controlled by regulatory sequences located far upstream from the transcription start site (sometimes over 10 000 nucleotides). These sequences, called enhancers, were found to stimulate transcription and are binding sites for transcription factors which are allowed to interact with the transcription machinery because the intervening DNA can form loops (Figure 2.5). Interestingly, enhancers are active regardless of orientation with respect to the direction of transcription and can be located either upstream or downstream of the transcription start site. In addition to these regulatory mechanisms, eukaryotic cells can also regulate transcription by modifying the state of condensation of chromatin.

However, during the last quarter of the 20th century, evidence was accumulating that post-transcriptional regulation might be extremely important in eukaryotes.

2.2 MicroRNA biology

2.2.1 Discovery

Conventionally, post-transcriptional gene regulation had been considered to act upon the stability of mRNA transcripts, modulating their half-life, by way of protein-RNA interactions. Possible regulatory targets in eukaryotes were the 5' cap or the length of the polyadenylated tail, as these are common features of mRNA transcripts which delay their degradation in the cytoplasm by exonucleolytic digestion.

New possibilities arose with the discovery of naturally occurring antisense RNA control, also known as natural antisense transcription (NAT), which could act upon mRNA processing,

transport, translation and degradation[1].

Antisense RNA control is not new. It has been known in prokaryotes since the 1980's, having been postulated back in the 1970's [120]. Researchers also started to wonder whether eukaryotes could have widespread antisense RNA control mechanisms, and indeed, in the 1990's, some suggested this was the case [56, 117, 58], this being now an established fact [133]. These suggestions were based, among other things, on the fact that there were many enzymes that seemed to detect the presence of double-stranded RNA (dsRNA) which is a hallmark of antisense transcription.

In the final years of the last century, the phenomenon of RNA interference (RNAi) was described as a mechanism that could provide anti-viral defense, modulation of transposition, or regulation of gene expression [37, 10]. By this mechanism, the introduction of dsRNA in a diverse group of organisms including the nematode, the fruit fly, fungi, and plants, originated the inhibition of expression of cognate genes.

The discovery that small antisense RNAs of about 22 nucleotides in length, called small interfering RNAs (siRNAs), are central to RNA interference suggested that tiny RNAs have major regulatory roles in eukaryotes and that they may all share parts of their effector mechanism. RNAi is an evolutionarily conserved genetic surveillance mechanism that triggers the specific degradation of mRNAs in response to the presence of dsRNA corresponding to the target mRNA [4]. It is now understood that RNAi in animals, PTGS (Posttranscriptional Gene Silencing) in plants and quelling in fungi are just different names for similar post-transcriptional RNA silencing phenomena common to virtually all eukaryotes [137]. Archea and bacteria lack the proteins known to be required for the RNA silencing pathways, so RNA silencing is probably an eukaryotic innovation.

A novel post-transcriptional silencing process was discovered at the turn of the century. It is elicited by tiny endogenous RNAs called microRNAs (miRNAs). MicroRNAs are a large class of small non-coding RNA molecules that have early on been recognised to be numerous and phylogenetically extensive [65, 63]. Many of these molecules originate from non-coding genes which produce mature transcripts of $\sim$22 nucleotides in length and are thought to function primarily as antisense regulators of other RNAs [4]. A detailed history of the discovery of these regulatory molecules is available in [60].

The initial members of the miRNA class of non-coding RNAs were *lin-4* and *let-7* of *Caenorhabditis elegans*. They were termed heterochronic or small temporal RNAs [65, 63] because all known instances seemed to be involved in controlling the timing of larval development. Most known miRNAs are very well conserved in close species and some can be found across very large taxonomic groups, notably *let-7* of *C. elegans* [94].

[1]NAT is a phenomenon observed in both prokaryotes and eukaryotes where a single gene locus is transcribed in both directions yielding both a sense and antisense RNA species, the latter being usually untranslated. Some of these antisense transcripts have been implicated in transcriptional and posttranscriptional regulation of gene expression by way of mechanisms that have not been yet fully characterised.

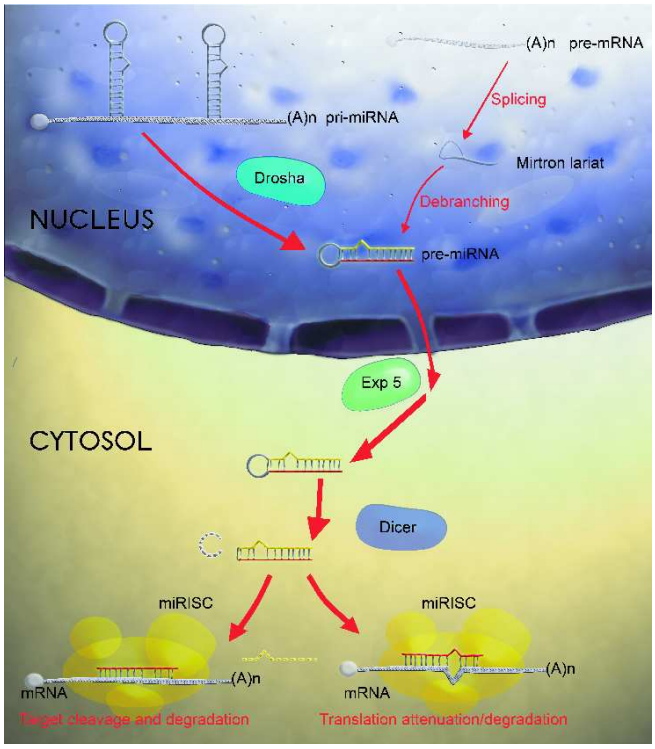

Figure 2.6: MiRNA biogenesis in metazoans. [83]

2.2.2 Biogenesis and function

MiRNA genes are frequently expressed individually, but many exist in clusters of 2–7 genes with small intervening sequences. Experimental results suggest that they are expressed co-transcriptionally, which indicates that they are under control of common regulatory sequences [63, 67, 9].

Other miRNA genes are excised from the introns of protein-coding genes [75, 61], introns and exons of non-coding genes [102], or even from the 3' UTR of protein-coding genes [22]. In mammalian genomes, it is also possible to find miRNAs in repetitive regions, and some studies suggest that transposable elements may be involved in the creation of new miRNAs [112].

MicroRNA biogenesis in animals is a two-step process [67], as shown in Fig. 2.6. The nascent transcript, which is several hundred nucleotides long, is called primary microRNA (pri-miRNA). Although some miRNAs are transcribed by RNA pol III [17], most rely on RNA pol II [68, 22], therefore pri-miRNAs can be subject to elaborate transcriptional control.

In a first step, the primary transcript is processed in the nucleus by a multiprotein complex (Microprocessor) containing an enzyme called Drosha [66] to give rise to the ∼70-nucleotides long miRNA stem-loop precursor (pre-miRNA) which is then exported to the cytoplasm. Secondary structure, rather than primary sequence, seems to be a critical feature for Drosha substrate recognition [27], however it is not known how this enzyme discriminates pre-miRNAs

from the great variety of cellular RNA stem-loops. It is known, however, that efficient processing by Drosha is dependent upon the presence of unstructured regions flanking the stem-loop [38]. The nuclear export is elicited by a complex of Exportin 5 (Exp5) and Ran-GTP which selectively binds pre-miRNAs while also protecting them from exonucleolytic digestion [138, 80].

In the cytoplasm, a second step takes place where the pre-miRNA matures into a ∼21-nucleotides long miRNA:miRNA* duplex, with each strand originating from opposite arms of the stem-loop. The cleavage is produced by the action of an enzyme called Dicer [47], which recognises the double-stranded stem [141].

In general, the miRNA strand is then integrated in a ribonucleoprotein complex known as the miRNA-induced silencing complex (miRISC) or miRNA-containing ribonucleoprotein particles (miRNPs) and the miRNA* is degraded [63]. Sometimes both strands can be detected [108], in which case the miRNA* designates the less predominant form of the mature miRNA.

Studies have shown that the intermediate miRNA duplexes exhibit a biased internal strand stability due not only to base-pair composition but also to structural features like mismatches or bulges [53]. These destabilising elements are thought to facilitate unwinding of the duplex and subsequent integration in the silencing complex. The strand that is less stable on its 5' end is preferably loaded onto miRISC [57]. When both ends exhibit similar stability, each strand is selected for integration with similar frequency [108]. Other studies, however, have suggested the existence of additional strand selection determinants [76, 32].

MicroRNA biogenesis in plants follows a similar process, but the miRNAs seem to be fully matured into a single stranded miRNA before being exported to the cytoplasm by an homolog of Exp5 termed HASTY (HST) and integrated onto the silencing complex, which partially explains why intermediate forms of plant miRNAs are only rarely detected [93, 8].

All maturation steps of plant miRNAs are processed by Dicer-like proteins. The predicted miRNA precursors in plants are much more variable in size than those of animals, ranging from around 60 to a few hundred nucleotides, whereas those in animals are typically ∼70-nucleotides long [7].

Given the stepwise process by which miRNAs are matured, and hence the diverse opportunities of regulation at each step, one can expect to find regulatory mechanisms at this level and, in fact, there is some evidence of post-transcriptional control of miRNA expression [27, 6, 79].

As mentioned, some miRNAs originate from the introns of other genes, usually being located in the same strand, which suggests that they are transcribed with the host genes and subsequently excised [75, 61, 102, 134]. Studies show that the expression of these miRNAs and their host genes is coupled, indicating a possible mechanism by which a protein and a miRNA are coordinately expressed [9, 8], presumably as part of a common biological process.

However, some intronic miRNAs occur in antisense orientation and may thus be transcribed under the influence of an independent promoter [27, 126].

An alternative pathway for intron-derived miRNAs has been recently identified in animals [104]. These introns, termed *mirtrons*, bypass Drosha processing and exhibit structural features similar to those of pre-miRNAs, thus entering the miRNA biogenesis pathway at the end of the first step. Unlike other intron-derived miRNAs which are excised from unspliced transcripts [54], *mirtrons* are dependent on the splicing machinery for maturation.

MicroRNAs in animals are thought to act primarily as translational repressors by pairing with specific partially-complementary 3' UTR regulatory elements on mRNAs [59], although target sites in the coding region and 5' UTR can also be functional [55, 81]. Another major miRNA silencing mechanism in animals leads to target mRNA destabilisation through a cleavage-independent process with a clear impact on transcript level [74, 96]. Some authors have suggested that miRNAs may have either a negative or positive regulatory effect [4]. In fact, recent evidence indicates that positive transcriptional regulation can be produced by miRNAs that target sites in promoter regions by an otherwise unknown mechanism [97]. Moreover, there are reports that in some circumstances and in certain cell-types, miRNAs can also enhance translation [118].

Plant miRNAs, on the other hand, frequently cleave and thus induce immediate degradation of the target mRNAs and are often almost perfectly complementary to sites in the coding region [98], as well as in the 3' UTR [99], and even in the 5' UTR [116]. However, some of these target sites may only be present after mRNA maturation since they span intron/exon boundaries [85]. It is also important to note that, *a priori*, nothing seems to prevent miRNAs from regulating RNAs other than mRNAs. They may also bind and regulate non-coding RNAs, perhaps even other miRNAs [4]. This possibility is illustrated by a study done with *Arabidopsis thaliana*, which suggests that miRNAs may bind fake targets in other non-coding RNAs thereby establishing a mechanism of negative regulation of miRNA activity [25].

Unsurprisingly, some large DNA viruses have evolved ways to explore the RNA silencing machinery of the host by coding for miRNAs [95]. These viral miRNAs can be expressed either individually or in clusters from pol II or pol III promoters. Interestingly, these miRNAs show no resemblance to other viral miRNAs, nor to the miRNAs of the host.

Chapter 3

MiRNA gene finding

Large-scale experimental approaches to miRNA gene finding met with some difficulties in the beginning, illustrated by the fact that these regulators escaped detection for so long. The short length of miRNAs and their ability to act redundantly, or to have only a subtle phenotypical impact imposes a limitation to the use of mutagenesis and other conventional genetics techniques [4]. Direct cloning, on the other hand, may not detect miRNAs that have very low expression levels or that are expressed only in specific conditions and cell-types. This is partially mitigated by the use of deep-sequencing techniques which nevertheless require extensive computational analyses to distinguish miRNAs from other non-coding RNAs of similar size [30]. It is clear, therefore, that computational approaches are essential for a more thorough catalogue of miRNA genes in sequenced genomes [75, 34].

Conventional *in silico* gene finding approaches are of limited use since miRNAs and non-coding genes in general do not exhibit the characteristic statistical properties of coding regions due to codon usage. The same can be said of homology-based searches in the absence of a clear evolutionary model for these genes. Obtaining such a model is particularly difficult due to the comparatively small size of the precursor and mature sequences.

The different characteristics of miRNAs in animals and plants have justified different approaches and, therefore, we discuss the methods developed for the two cases separately.

3.1 Computational approaches to miRNA gene finding in animals

Lee and Ambros [65] established the paradigm of what would become the typical strategy for miRNA gene search. According to these authors, future miRNA genes ought to share some features with *lin-4* and *let-7* of *C. elegans*, namely the expression of a mature RNA sequence of the appropriate length (∼22-nucleotides), which should have its origin in intergenic sequences and be processed from a stem-loop precursor transcript of around 65-nucleotides in length.

Furthermore, there should be extensive sequence similarity with orthologs in closely related species.

These observations prompted the adoption of several criteria for the annotation of novel miRNA genes [5]. First, expression criteria establish that new miRNA genes should be supported by experimental evidence that detects the ∼22-nucleotides RNA transcript, or that these small molecules should be found in cDNA libraries. Second, at least one of the following biogenesis criteria should be met: 1) the mature miRNA should be included in one arm of a predicted minimum free energy fold-back precursor structure with extensive base-pairing in the miRNA region which should not contain any large internal loops or bulges, especially asymmetric bulges; 2) the fold-back structure should be phylogenetically conserved; and 3) the precursor should be shown to accumulate in organisms with impaired Dicer function. Expression criteria need not be met in the case of obvious homologs.

It is clear that expression criteria alone are not sufficient for a confident annotation since they cannot distinguish miRNAs from other cellular RNAs with approximately the same size, or from spurious degradation products of other RNAs. On the other hand, the fact that expression evidence cannot be found does not necessarily exclude a candidate due to the limitations of experimental methods.

Known microRNA precursors have a typical stem-loop secondary structure which is essentially conserved amongst metazoa but heterogeneous within plants. Some of the first miRNA gene searches were carried out considering both this typical secondary structure and structure/sequence conservation between two closely related species (*C. elegans* and *C. briggsae*) [65]. However, it soon became clear that there were much more conserved stem-loops than miRNA genes, and additional criteria had to be put in place if we were to identify good candidates [60].

Moreover, although a significant fraction of known miRNAs seems to be very well conserved phylogenetically, this may reflect the bias of the search procedures used so far, which privilege phylogenetic conservation in order to validate miRNA candidates. It may also illustrate a general limitation of current computational approaches which can only predict candidates which resemble previously identified miRNAs [21]. Furthermore, strong conservation may be a sign of the existence of multiple conserved target sites which would constitute an overwhelming selective force against mutation. As more organisms are being analysed, more miRNA genes are identified and an increasing number is shown to be lineage or species-specific [11].

Despite the caveats, the annotation criteria have inspired most current computational methods for miRNA gene finding. Therefore, many tools share the same overall strategy but use different approaches to phylogenetic conservation, and different features to identify good stem-loop candidates. These methods can thus be distinguished roughly by the way they identify the initial candidate set, the structural criteria they use to further restrict

precursor candidates, the conservation criteria they adopt and any additional filters they may implement. We refer to these approaches as *filter-based* methods. Later approaches use conventional *machine learning* methods that try to generalise from a positive set of previously known miRNAs and a negative set of stem-loops presumed not to be miRNA precursors. *Target-centered* approaches use a putative set of miRNA targets derived from conservation analyses which are then used to seek new miRNAs. *Mixed* approaches use a combination of computational tools and high-throughput experimental procedures. Finally, *homology-based* searches try to identify stem-loops similar to previously identified pre-miRNAs that may have been missed by *ab initio* methods.

	Initial set	**Structural criteria**	**Conservation criteria**	**Additional filters**
Grad *et al.* [34]	Stem-loop structures in repeats-masked intergenic regions	MFE, GC content, matches, mismatches, gaps and occurrence of multi-loops	Homologous stem-loops transitively identified in two additional genomes	Hairpins containing short repeats or with low quality structure are eliminated
MirScan [75]	Folded structures identified sliding a 110-nt window along the genome	Number of bp, MFE, no overlap with repeats, no skewed base composition	Homologous stem-loops identified in an additional genome	Log-odds score for several features of the miRNA region of the stem-loop
Berezikov *et al.* [12]	Regions exhibiting a typical conservation pattern identified using phylogenetic shadowing	Only highly probable stable stem-loops are retained	Implicitly considered in the initial set	
MirSeeker [61]	Aligned non-coding non-annotated regions from two species	Metrics involving length of longest stem-arm, MFE, internal loops, asymmetric loops and bulges applied to predicted structures in aligned regions	Typical divergence pattern	

Table 3.1: Comparison of some filter-based approaches to miRNA gene finding in animals

3.1.1 Filter-based approaches

Early miRNA gene finding methods, summarised in Tab. 3.1, focused on the identification of small high-quality sets of conserved miRNA candidates which would have a better chance of being experimentally confirmed as true miRNAs. One of these methods, described in [34], identified several new miRNAs in *C. elegans*. An initial candidate set of imperfect stem-loops obtained from all repeats-masked intergenic regions of the genome of *C. elegans* was filtered according to criteria that accounted for matches, mismatches and gaps on the stem region, as well as GC content, MFE (minimum free energy) and the occurrence of multi-loops. The cut-offs for these parameters were chosen to reflect the characteristics of previously known miRNAs from the studied organism. Most of the filtering was achieved with the conservation criterion that required homologous stem-loops on two additional genomes. It is interesting to note that, from a universe of 61 known genes, only 29 out of 39 *C. elegans* miRNAs included

in the initial set passed the structural criteria and not more than 6 miRNAs were conserved in two additional genomes, illustrating the emphasis on specificity rather than sensitivity.

Another approach, also published in 2003, makes some improvements on sensitivity. This method called MiRscan [75], produced an initial set of candidates by scanning the genome of *C. elegans* with a sliding-window of 110 nucleotides. The regions were folded and filtered according to more permissive structural criteria. Potential homologs were sought in *C. briggsae* sequences and only conserved hairpins were retained, yielding a total of ∼36 000 candidates. With this procedure, 50 of the 53 miRNAs known at the time to be conserved in both species were recovered. Using these 50 miRNAs and the background set of over 36 000 hairpins, the authors developed a sophisticated log-odds scoring scheme that considered several features of the mature miRNA portion of the stem-loop. All candidate hairpins were scored and ranked according to this scheme. However, MiRscan was still not able to recover more than half of the previously known *C. elegans* miRNAs from the top scoring candidates.

The authors would later improve this method with MiRscanII [91] which, in addition to the features considered by MiRscan, took into account the presence of conserved motifs and blocks of sequence conservation up- and downstream of the predicted stem-loop precursors, presumably involved in transcriptional regulation. The authors observed that independently transcribed microRNA genes in *C. elegans* contained a well-conserved motif upstream of the stem-loop, with respect to homologous sequences in *C. briggsae*, and used it as an additional feature. Similar upstream motifs were found in *H. sapiens*, *M. musculus*, and *D. melanogaster*.

An approach described in [12] also considers conservation around the precursor region and was used in the search for mammalian miRNAs. In this study, the authors could not identify clearly conserved motifs in the flanking regions immediately adjacent to the pre-miRNA stem-loops, but they were able to observe a distinctive pattern of diminishing conservation that was used as a characteristic profile aiding the search for miRNA genes.

The method miRseeker [61] represents the first attempt to identify conserved stem-loops due to selection, and not as an artefact of considering genomes that are not sufficiently distant. The authors aligned the non-annotated intergenic and intronic sequences of the genomes of *D. melanogaster* and *D. pseudoobscura*. The conserved regions were then folded in order to identify and score potential stem-loop structures. The evaluation of the hairpins considered the length of the longest stem-arm and its MFE, as well as a set of metrics penalising internal loops, particularly asymmetric loops and bulges. By analysing a reference set of known miRNA genes of two drosophilid species, the authors derived a typical divergence pattern. In general, divergence was observed in the terminal loop, or in either one of the stem arms. A good miRNA candidate should exhibit a pattern such that divergence occurs in at most one stem arm, and the mutation rate at the stem arm should not exceed that seen in the terminal loop. This is justified by the fact that mutations in the terminal loop have a lesser impact on pre-miRNA structure and identity and, consequently, its processing efficiency and target

specificity, than mutations on the stem arm.

The methods described so far and variations thereof have been able to recover a substantial part of the known miRNAs and have been useful in identifying several new regulators [6, 73, 105]. Some of these methods have benefited from a growing number of sequenced species allowing more extensive and sophisticated studies of conservation patterns [106, 103]. However, they have failed to produce a set of rules capable of recovering all known miRNAs without leading to too many false positives. Additionally, they are critically dependent on conservation criteria to attain reasonable levels of specificity. This approach effectively prevents the identification of non-conserved candidates, and makes several assumptions in the absence of a clear evolutionary model for these structures.

3.1.2 Machine learning methods

One attempt to use a single-genome approach to miRNA gene finding was ProMIR [87]. The initial set of candidates are stem-loops that are present on human ESTs, therefore restricting the search to sequences with verified expression. Candidate stem-loops were filtered using very permissive structural criteria concerning stem length, loop size and MFE. This probabilistic method relies on an HMM (Hidden Markov Model) that models characteristics of the stem portion of the stem-loop viewed as a paired sequence. These characteristics concern the pattern of base-pairing and the location of the mature miRNA. The positive training set consisted in all known human pre-miRNAs and the negative set corresponded to 1 000 extended stem-loops randomly extracted from the human genome. A stem-loop is found to be a good pre-miRNA candidate if it contains a sequence with probability of being a mature miRNA above a certain threshold. However, the candidate set was still too large and additional filters had to be used, including the assessment of the statistical significance of the predicted secondary structure, and the verification of a decaying conservation pattern in the regions flanking the putative pre-miRNA by comparison to other vertebrate genomes, as done in [12].

The first successful single-genome approach came from a method developed to identify miRNAs in viral genomes [95]. Using conservation criteria in this case is not an option as most viral pre-miRNAs show no detectable conservation with respect to either other viral pre-miRNAs or to the precursors of the infected host. The method starts by identifying robust stem-loops, i.e., stem-loops which retain the typical folding structure regardless of the precise location of the start/end of the folded transcript. This is justified by the observation that a pre-miRNA should be robust with respect to the genomic context where it lies. These candidate stem-loops were then scored by an SVM (Support Vector Machine) classifier trained on a set of positive examples derived from known human miRNA precursors and a set of negative examples derived from mRNAs, tRNAs, rRNAs and random regions of the human and viral genomes. The features considered included folding free energy, nucleotide count in

the symmetrical stem, and number of A-U, G-C and G:U pairs in the predicted structure. The authors forced the misclassification of positives to be eight times more penalising than the misclassification of negatives, thus sacrificing sensitivity for higher specificity.

The same approach was then used to predict clustered pre-miRNAs in *H. sapiens*, *M. musculus*, and *Rattus norvegicus* [109] following the observation that many animal miRNAs indeed occur in clusters. The high false positive rate that the approach, in general, could entail is partially mitigated by the fact that only regions close to previously identified miRNAs are scanned, so it is reasonable to assume that these regions are indeed transcribed and can represent instances of clustered miRNAs.

Another single-genome approach is HHMMiR [52] which uses hierarchical hidden Markov models to identify the sequence/structure characteristics of different portions of the pre-miRNA hairpin (the terminal loop, the miRNA duplex, the remaining sequences of the miRNA precursor, additional sequences pertaining to the pri-miRNA which happen to be included in the extended stem-loop). The HHMMiR model parameters are estimated by training over the positive set of known human pre-miRNAs and the negative set of hairpins randomly extracted from the coding regions.

Several other machine learning methods have been proposed to tackle the problem of identifying good miRNA candidates. SVMs have been a popular framework used to learn the distinctive characteristics of miRNAs. Most approaches use sets of features concerning sequence composition [132, 41, 88], topological properties of the stem-loop [41, 88, 46], thermodynamic stability [41, 88, 46], and sometimes other properties including entropy measures [88].

A somewhat different approach called miRCoS-a [110] chains three different SVM classifiers, each focusing on different features of the candidate stem-loops obtained from conserved regions of vertebrate genomes. The aspects covered by each SVM concerned: 1) sequence conservation; 2) secondary structure conservation; and 3) location and structure of the mature miRNA in the hairpin structure. By using a chained-filter approach the authors were able to compute complex features for the SVMs downstream in the pipeline, which would have been prohibitively time-consuming if applied to all the initial candidates.

An SVM method specifically designed to predict Drosha processing sites is described in [40]. The classifier uses 11 features concerning sequence/structure properties in different regions of the stem-loop. This method not only can serve as a pre-processing tool of miRNA candidates as it can also generate additional features for precursor classifiers concerning metrics about the potential processing sites.

Other machine learning methods rely on Random Forests [49] (a method that uses a set of tree-based classifiers combining sampling of training data with random feature selection), a Naïve Bayes classifier [136], or genetic programming [20].

The methods described in this section are natural approaches to the miRNA gene finding problem. The latter is cast as a classification problem and powerful methods are used to

generalise from positive and negative examples, as is customary. In this case, however, there are a few questions raised by the positive and negative datasets adopted.

Negative datasets usually include randomly chosen stem-loops extracted from the genome, under the assumption that there is a very low-density of pre-miRNAs and therefore there is a small chance of a true miRNA precursor being recruited as a negative example. The number of miRNAs in any given genome is still an open problem and consequently we cannot confidently evaluate the impact of this assumption. Additionally, there may be many stem-loop structures in the genome that would be able to enter the miRNA processing pathway but are not efficiently transcribed, or are simply in the wrong genomic context. Since these machine learning approaches do not usually incorporate any information regarding transcription potential or genomic context, but rather concentrate on stem-loop features, they may be misclassifying an important portion of the search space, despite the fact that cross-validation procedures or validations with independent test sets have given very good measures of sensitivity and specificity.

On the other hand, positive examples are recruited from miRNAs previously identified by experimental procedures or other computational methods and these datasets are, therefore, strongly biased towards highly-expressed and extensively conserved miRNAs. This questions the critical assumption that the positive set is truly representative, as low-expression non-conserved miRNAs may have features that are substantially different. Despite this, with a growing number of miRNAs being identified, one can expect an increasingly better performance from these methods.

3.1.3 Target-centered approaches

An innovative strategy to predict miRNA genes is described in [131]. The authors aligned the 3' UTRs of several mammalian genomes and identified highly conserved short motifs showing properties reminiscent of miRNA target seeds. Subsequently, the authors identified hundreds of conserved and stable stem-loops containing conserved sequences complementary to the short motifs previously identified, including several known miRNAs.

Target-centered approaches have the benefit of making few assumptions about the structure of miRNA precursors, but are dependent on the identification of highly-conserved motifs in 3' UTRs which do not represent all the universe of possible targets.

3.1.4 Mixed approaches

Some approaches have combined high-throughput experimental methods with computational procedures in order to identify a wider range of miRNAs. These approaches can use two different strategies: 1) identification of a great number of low-confidence precursor candidates subsequently subject to high-scale experimental verification; 2) extensive cloning of

small RNAs that are then analysed with respect to their localisation in the genome and their ability to form stem-loops in the genomic context of the identified locations.

A method called PalGrade [11] followed the former strategy to identify several new conserved and non-conserved miRNAs in *H. sapiens*. Thousands of candidate stem-loops were selected based on a scoring scheme that considers thermodynamic stability and structural features. The potential expression of this set of candidates was then tested in several tissues with microRNA microarrays, and candidates with strong hybridisation signals were further subjected to directed cloning and sequencing. This approach has substantially expanded the catalogue of human miRNAs.

Methods following the second strategy usually consider the bulk of sequenced RNAs, determine their genomic location and apply filters similar to those used by *ab initio* methods [86, 105, 62].

As noted before, these approaches cannot, however, detect low-expression or tissue-specific miRNAs. Deep sequencing techniques have formidably expanded our ability to detect low-abundance transcripts but have also presented new challenges. While raising the ability to sequence rare miRNAs, other small transcripts are also amplified and more sophisticated approaches are required to sieve out miRNA transcripts. A method called miRDeep [30] uses a probabilistic model to assess the compatibility of the pattern of sequenced RNA transcripts with properties of miRNA biogenesis. According to this model, a true miRNA precursor should have a characteristic signature, with frequent sequence reads corresponding to the mature region of the stem-loop, and less frequent reads corresponding to other parts of the hairpin structure.

3.1.5 Homology-based searches

Homology-based approaches are a common way of detecting miRNAs that may have been missed by *ab initio* predictors, and in fact many miRNA gene prediction approaches incorporate an homology-based search as part of their protocol, in addition to the usual search for orthologs which is an integral part of the conservation requirements.

Many homology searches are alignment-based methods and can be applied to the members of the original candidate set that failed to pass some of the filters [34], or specifically directed to regions surrounding known miRNAs in the hope of finding new members of a gene cluster [3]. Alternatively, these methods can be used to scan newly-sequenced genomes for homologs of known miRNAs [23, 125, 89], or to further saturate miRNA gene predictions in previously studied genomes [126].

However, alignment-based methods rely exclusively on sequence conservation. More sensitive methods can be developed by considering structure conservation. An example is the approach described in [69] which proposes a profile-based method using an RNA comparison tool named ERPIN [31] to account for sequence/structure conservation and was able to

predict hundreds of new candidates from several different families of animal miRNAs. An alternative example is miRAlign [122].

Another powerful strategy is the use of structure-based clustering. In this approach, a set of candidate structures are clustered using a metric based on sequence/structure alignments. Potential homologs are found in clusters with known miRNAs [128, 100].

3.2 Computational approaches to miRNA gene finding in plants

Strategies similar to those used in animals have been applied to the prediction of plant miRNA genes. In this case, the problem is considerably more difficult due to the heterogeneous nature of plant pre-miRNA stem-loops which vary greatly in size and structure. Consequently, these methods rely more on the properties of the miRNA:miRNA* duplex within the variable precursor, and it is also not surprising that much fewer approaches have been proposed for plants than for animals.

3.2.1 Filter-based approaches

One of the first methods for identifying miRNAs in plants is described in [121]. The authors proposed a workflow that began by identifying all potential hairpins in the intergenic regions of *A. thaliana*. The hairpins were found by looking for imperfect inverted repeats of 21-nucleotides, representing the putative mature miRNA and corresponding star sequence, that were separated by a distance within a given window. The candidate hairpins were then filtered according to criteria concerning GC content and loop length. The putative miRNA sequences were checked against the rice genome and only those showing high conservation were retained. Finally, the remaining precursor candidates and their orthologs were folded to validate the characteristic stem-loop secondary structure. This procedure suggested 83 new and identified 12 previously known miRNAs. Amongst the miRNA candidates, 19 had their expression experimentally verified, or were found in public databases of small RNAs.

A similar approach is described in [50]. The candidate sequences are folded using a secondary structure prediction algorithm and given to a program called MIRcheck. This program receives a sequence/structure specification and the co-ordinates of a 20-mer within the hairpin and uses a series of metrics concerning the number of unpaired nucleotides and bulges in the miRNA mature regions and the length of the hairpin. Sequences overlapping repetitive elements are eliminated, and a strong conservation criterion is applied retaining only stem-loops where the mature miRNA appears in both genomes and exhibit high conservation in both the miRNA and miRNA* sequence. Additionally, stem-loops are tested for robust folding, indicating that their secondary structures do not change substantially in the presence of flanking sequences. An additional filter consisted in searching for conserved near-perfect complementary matches in the mRNAs of both genomes, presumably target sites for these

miRNA candidates. With this method, the authors were able to identify 379 good miRNA candidates in 228 unique loci, of which 23 had their expression experimentally verified.

A computational pipeline called MIRFINDER [15] identifies conserved hairpin structures in the genomes of *A. thaliana* and *O. sativa* and subsequently applies several filters, based on core features derived from known miRNAs. The features seen in the miRNA reference set suggested that the mature miRNA should be part of a stable continuous helix with no more than a few unpaired or G:U pairs in the miRNA region. The conservation requirements included extensive conservation of the mature miRNA sequence and location in the same stem arm. The authors observed that a large amount of sequences in both genomes could fold into hairpin structures, so a randomisation test was setup to assess the statistical significance of the predicted secondary structures. After applying filters for GC content and low complexity sequences, a total of 91 potential miRNA genes were identified, of which 58 had at least one nearly perfect target match.

The methods described so far make extensive use of conservation criteria and are therefore unable to identify miRNAs with less obvious patterns of evolutionary conservation. Other methods have taken advantage of the near-perfect complementarity observed between the miRNA and corresponding target sites in plant mRNAs and were able to identify several novel non-conserved plant miRNAs.

3.2.2 Target-centered approaches

A single-genome approach called FINDMIRNA [1] replaced the sieve of cross-species conservation of candidate stem-loops with the detection of potential targets within transcripts of the same species. The algorithm starts by indexing all the 7-mers of the intergenic regions, excluding repeats and low GC-content sequences. For each transcript, its overlapping 7-mers are tentatively matched against the index previously computed. For each match, an ungapped alignment of the surrounding areas is produced. The best length-normalised alignment score of size 18 to 25 is marked as a potential miRNA. If the score is above a given threshold, a dynamic programming algorithm is used to search for a complementary sequence in the vicinity. A secondary structure prediction algorithm is used to verify the presence of a stem-loop structure, and whether the length-normalised MFE is below a given threshold. An additional filter is then used for higher specificity, which exploits the expected typical divergence pattern of miRNA precursors of the same family, whose members have presumably arisen by duplication events. Precursor candidates are put in the same family cluster if they target the same transcript region. Clusters are then scored according to the degree of conservation of the miRNA, miRNA*, and intervening sequence, using a scoring function that privileges conservation of the miRNA sequence and penalises conservation in the intervening region.

A similar approach described in [78], unlike the previous method, does not require that miRNAs be clustered into families. This method takes each mRNA and a genome-wide

search is performed in order to identify regions of 20-27 nucleotides that match a portion of the mRNA with at most 2 mismatches. These matches, termed *micromatches*, are then used to identify miRNA candidates. The candidates are passed by six filters: 1) high sequence complexity; 2) no overlap with annotated exons; 3) no overlap with repeat sequences; 4) stable miRNA:mRNA duplex; 5) no more than 10 identical copies in the genome; 6) the putative miRNA is contained in a stable precursor stem-loop structure exhibiting some typical features. An additional sieve is then added that includes only miRNA candidates with more than one target, which is thought to be typical of most plant miRNAs.

3.2.3 Homology-based searches

Upon the identification of an ever increasing number of plant miRNA genes in several species, homology-based search methods begun to be developed seeking the complete enumeration of miRNAs in model organisms [72, 28]. In general terms, these methods first identify genome hits matching known miRNA mature sequences and then extract the genomic context of such hits and align the candidates with their putative miRNA families followed by the application of some criteria to determine a final list of candidate homologs. More recently, these protocols have been adapted to search for new miRNAs by analysing EST (expressed sequence tag) data [139].

3.2.4 Other approaches

Other methods for plant miRNA gene identification have been developed using a combination of high-throughput sequencing, filtering and machine learning approaches in similar ways to those discussed for animal miRNA prediction [115].

Part II

Methods

Chapter 4

Single-genome approach to miRNA gene finding

In this chapter, we present a three-pillar approach to the identification of miRNA precursors from a single genome which is illustrated by the schematisation in Figure 1.1.

We begin by presenting our candidate identification procedure which purports to identify viable genomic hairpins which will be used as our initial candidate set. The most significant contribution of this procedure is that it provides a means to avoid folding all the possible genomic windows under analysis by only considering those with locally maximal self-hybridising potential. The gains in efficiency come from the fact that calculating this potential is less computationally expensive than performing structural folding.

In the following section, we present a procedure to evaluate our set of candidates using a number of robustness and stability measures collectively called *intrinsic measures*. These measures have been shown to distinguish pre-miRNAs from other stem-loops but were never before used in the context of *ab initio* miRNA gene finding. Each individual measure is justified by a sound biological argument but any one of those measures in isolation is not sufficiently segregative. We present a procedure to combine those measures into a single combined score (*cscore*) which we later show to greatly improve over their individual performances.

In section 4.3 we present our strategy to perform a structural analysis of the candidates in an attempt to identify those which resemble known precursors. First, we discuss the use of a vectorial representation of the primary/secondary structure of the precursor candidates which is able to capture information about key aspects of the hairpin, particularly the number and scale of bulges, internal loops and other features which are known to compromise the recognition of the stem-loop by the appropriate cellular mechanisms. The empirical support for the choice of this particular vectorial representation and the demonstration that it resembles the results obtained with candidate samples using conventional structural clustering is presented later, in chapter 5. The vectorial representations of the candidates are then submitted to a principal components analysis procedure and mapped to the multidimensional principal

components space. We proceed to detail how, on this transformed space, one can identify the region most likely to contain candidates with the appropriate structural features and select the candidates contained therein.

In section 4.4 we discuss strategies to address the need to assess the transcriptional potential of a candidate precursor sequence. These strategies are dependent on the availability of additional data, namely annotation or experimental data and consist of filtering out candidates with annotation information inconsistent with being a pre-miRNA, the identification of candidates with genomic locations close to previously identified pre-miRNAs and the mapping of small sequences of RNA transcripts back to the genome.

Finally, in section 4.5 we introduce the framework that implements the methods described in the previous sections. We provide details of the implementation, an overview of the processing pipeline and a description of the web-based interface.

4.1 Efficient identification of candidate hairpins

The problem of identifying candidate stem-loop structures in a genome can be cast as the problem of finding an imperfect palindrome with a central intervening sequence. Scanning the entire genome of interest in an attempt to seek imperfect palindromes directly proved to be computationally unfeasible, especially considering the irregular nature of these stem-loop structures with potentially numerous and large bulges as well as non-canonical base pairings. Existing methods to enumerate imperfect palindromes are inefficient [36] for our purposes and usually require the specification of the number of gaps and their length, which cannot be effectively estimated beforehand. Instead, we adopted a filtering approach using a heuristic procedure which is based on the observation that a segment of a genome which, upon transcription, may adopt a stem-loop conformation should exhibit a higher degree of potential pairing between its two halves (if the midpoint of the segment falls within the region corresponding to the terminal loop) than a segment that either does not contain a stem-loop or only partially contains such a structure.

Let S be a string over an alphabet Σ, $|S|$ denotes the length of the string, and S_i denotes the character on the i-th position of the string, with $0 < i \leq |S|$, $\overleftarrow{S}$ denotes the reversed string such that $\overleftarrow{S} = S_{|S|} \dots S_1$. We denote by $L^S = S_1 \dots S_{\lfloor |S|/2 \rfloor}$ and $R^S = S_{\lceil |S|/2 \rceil} \dots S_{|S|}$, respectively, the left and right halves of string S so that $S = L^S R^S$ as illustrated in Figure 4.1.

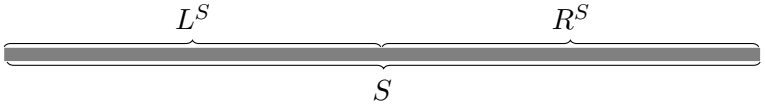

Figure 4.1: Two halves of a split string

Let $\Phi \subset \Sigma^2$ be the set of accepted pairings of characters in Σ. The set Φ induces the

predicate $F_\Phi : \Sigma^* \times \Sigma^* \mapsto \{0,1\}$ defined as follows: $F_\Phi(a\alpha, b\beta) = 1$ iff $(a,b) \in \Phi \wedge (F_\Phi(\alpha,\beta) = 1 \quad \vee \quad \alpha = \beta = \varepsilon)$, with $a, b \in \Sigma$, $\alpha, \beta \in \Sigma^*$, and ε being the empty string.

In order to consider acceptable RNA pairings, including all canonical pairs along with G:U base-pairs (which is the most common wobble base-pair) we have

$$\Phi = \{(\mathtt{A},\mathtt{U}), (\mathtt{U},\mathtt{A}), (\mathtt{C},\mathtt{G}), (\mathtt{G},\mathtt{C}), (\mathtt{G},\mathtt{U}), (\mathtt{U},\mathtt{G})\}$$

.

The best local alignment over a split string S is calculated using a dynamic programming matrix where each cell, $H(i,j)$, is computed using the following recurrence:

$$\max \left\{ \begin{array}{ll} 0 & \\ H(i-1,j-1) + \xi_0 & \text{if } F_\Phi({L^S}_i, \overleftarrow{{R^S}_j}) = 1 \\ H(i-1,j-1) - \xi_1 & \text{if } F_\Phi({L^S}_i, \overleftarrow{{R^S}_j}) = 0 \\ H(i-1,j) - \xi_2 & \\ H(i,j-1) - \xi_2 & \end{array} \right\} \tag{4.1}$$

where ξ_0, ξ_1, and ξ_2 represent the contribution to the alignment score of matches, mismatches, and gaps, respectively.

Using the Smith-Waterman algorithm [113] on a split string, S, one can determine the best alignment in $O(|S|^2)$.

Consider a genome with k chromosomes, seen as a collection of sequences $\mathcal{S} = \{C_1, C_2, \ldots, C_k\}$. The algorithm will slide a window of length w along each chromosome of the genome determining, for each position, the best local alignment under a model $M = (\xi_0, \xi_1, \xi_2)$.

Using the described sliding-window procedure, it is clear that the best alignments for all windows in the genome can be computed in $O(w^2 \sum_i (|C_i| - w + 1))$.

An alignment Λ of two sequences S_1, S_2 is a tuple (e_1, e_2, σ) where $0 \le e_1 \le |S_1|$, $0 \le e_2 \le |S_2|$, and $\sigma \in \{\uparrow, \leftarrow, \nwarrow\}^*$.

If $\Lambda = (e_1, e_2, \sigma)$ is a best local alignment of a split string, then:

- $\forall i,j \quad H(i,j) \le H(e_1,e_2)$, where $H(i,j)$ is the value of the ith row, jth column of the dynamic programming matrix of the Smith-Waterman algorithm.
- σ represents a path from (e_1, e_2) to a cell in the dynamic programming matrix containing the value 0 such that if the k-th cell in the path is (i_k, j_k) and $H(i_k, j_k) \neq 0$ then the $(k+1)$-th cell in the path is:
  - $(i_k - 1, j_k - 1)$ if $H(i_k, j_k) = H(i_k - 1, j_k - 1) + \xi_0$ and $\sigma_k = \nwarrow$
  - $(i_k - 1, j_k - 1)$ if $H(i_k, j_k) = H(i_k - 1, j_k - 1) - \xi_1$ and $\sigma_k = \nwarrow$
  - $(i_k - 1, j_k)$ if $H(i_k, j_k) = H(i_k - 1, j_k) - \xi_2$ and $\sigma_k = \leftarrow$

– $(i_k, j_k - 1)$ if $H(i_k, j_k) = H(i_k, j_k - 1) - \xi_2$ and $\sigma_k = \uparrow$

The rationale of the procedure is to take the best alignment of the two halves of each genome window considered and identify the windows where the pairing potential is locally maximal with respect to a normalised score.

The normalised score for a best local alignment $\Lambda = (e_1, e_2, \sigma)$ of a split string is defined as:

$$s(\Lambda) = \begin{cases} \frac{2H(e_1,e_2)}{e_1+e_2} & \text{if } e_1 + e_2 > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

The adopted score not only normalises the score of the best alignment with respect to the alignment length, but it also privileges a base pairing closer to the midpoint of the genome window under consideration. We can now define what is a candidate position in the genome.

Consider a chromosome, C, of a given genome. Let $S = C_p \dots C_{p+w-1}$ be the sequence of length w starting at position p of the said chromosome, and let Λ_p be the best local alignment in S. We have that S is a candidate sequence iff the normalised score is locally maximal at S, i.e.,

1. $\exists \hat{p} : \forall p' : \hat{p} < p' \leq p \implies s(\Lambda_{\hat{p}}) < s(\Lambda_p) = s(\Lambda_{p'})$
2. $\exists \hat{p} : \forall p' : \hat{p} > p' \geq p \implies s(\Lambda_{\hat{p}}) < s(\Lambda_p) = s(\Lambda_{p'})$

having $s(\Lambda_{\hat{p}}) = 0$ for every $\hat{p} < 0$ or $\hat{p} > |C| - w + 1$.

As several candidate positions may be identified in contiguous co-ordinates in the genome, presumably for each window whose midpoint falls within the terminal loop portion of the stem-loop, we aggregate them together in candidate regions as they will refer to the same stem-loop structure.

Let R^l_p be a region of length $l \geq w$ starting at position p of a chromosome C of a given genome. R^l_p is a candidate region iff $C_p, \dots, C_{p+l-w}$ are candidate positions and $C_{p-1}, C_{p+l-w+1}$ are not.

We have chosen a window length of 200, which approximately corresponds to the length of the largest annotated metazoan precursor sequence and is wide enough to accommodate the vast majority of known animal pre-miRNAs, and we have adopted a scoring model such that $\xi_0 = \xi_1 = \xi_2 = 1$. The choice of parameters for the model is important since it may affect the identity and number of candidate regions identified. Our scoring model was based on three observations. First, most DNA alignment methods prefer a linear model for gaps and an equal penalty for gaps and mismatches [114]. Second, miRNA precursors necessarily exhibit gaps and mismatches when aligning their stem portion due to the ubiquitous yet small bulges and inner loops which justifies that the penalty for a gap/mismatch be the same as the contribution given by matches. Finally, small variations in the scoring model did not produce

significantly different results, whereas more radical departures from the adopted model, such as having mismatches or gaps negatively contributing to the alignment score more than twice the contribution of a match, did have an impact on the sensitivity (data not shown).

Having identified the candidate regions, these are folded using RNAfold [45] with standard parameters and the largest stem-loop structure contained therein is extracted and re-folded. The final set of precursor candidates is made up of these refolded stem-loops after the elimination of non-viable and redundant candidates. Candidates are deemed non-viable if they exhibit a minimum free energy higher than -20 Kcal/mol or if they have one or both stem arms shorter than 16 nucleotides because otherwise the structure would be too unstable or not be long enough to accommodate a mature miRNA sequence. These filtering parameters capture the vast majority of known pre-miRNAs while significantly reducing the number of candidate stem-loops. The final list of candidates is subjected to an additional filtering step in order to identify redundant candidate subsets. A subset of candidates is said to be redundant if each member has identical terminal loop start/stop genomic positions although the length of the stem arms may vary. In this case, the candidate (or a random element of the subset of candidates) with the largest total sum of stem arm length is retained and the remaining members of the subset are discarded. This procedure creates a small bias towards larger stem-loop structures, but it avoids considering largely redundant sequences or sequences missing key portions of the stem or even the putative mature sequence of a possible precursor. This type of candidates, which are likely larger than the actual precursor (if one happens to be contained therein), are usually referred to as *extended stem-loops*, and correspond to the largest hairpin that can be folded in its genomic context.

4.2 Robustness and Stability Measures

The evaluation measures described in this section, which we collectively call *intrinsic measures*, purport to assess whether the candidate precursors possess certain features that have been shown to distinguish pre-miRNAs from other stem-loops and are related to the stability and robustness of their secondary structure.

It was shown that miRNAs have an *adjusted minimum free energy* (AMFE) that is lower than that of other stem-loop structures [140], i.e., when normalised for length, other genomic stem-loops tend to be less stable than miRNA precursors. Similarly, it was established that miRNA precursors tend to preserve roughly the same secondary structure in the face of variations in their genomic context [64], presumably as an evolved robustness to mutations in their flanking sequences (*Robustness to context*). Likewise, it was shown that miRNA precursor structures are usually also robust with respect to mutations (*Robustness to mutations*) [18], possibly as a result of second-order evolutionary processes. To these three measures, we add the requirement for *Robustness of folding* observing that a true miRNA precursor should fold

into a stable stem-loop structure for the most part of the structures in the thermodynamic ensemble where the molecule is found in physiological conditions.

From the combination of these four measures, it is possible to derive a single combined score (*cscore*) for each precursor candidate and rank the stem-loops extracted from a given genome. Our single score not only combines the information provided by each measure, but it does so against a background of hairpins extracted from a random sequence with the same dinucleotide distribution of the original genome. This procedure compensates for hairpin robustness provided by genome composition alone.

We describe the details of the computation of the intrinsic measures for each candidate, as well as the determination of the *cscore*, in the remainder of this section.

4.2.1 Adjusted MFE

The Adjusted MFE (AMFE) for a precursor p of length $|p|$ is defined as:

$$s_1(p) = \text{AMFE}(p) = 100\frac{\text{MFE}(p)}{|p|} \tag{4.3}$$

This measure consists in normalising the minimum free energy of the structure under study with respect to its length. The normalisation procedure is justified by the observation that larger structures can have lower free energies due simply to the fact that they have more opportunities to form base-pairs. GC content also plays a role in defining the lower limit of free energy a structure can exhibit, but additional normalisation by GC content as done in [140] using the MFEI (Minimum Free Energy Index) yields an ill-defined measure for candidates from AT-rich regions which will sometimes have a GC content of zero. The procedure used to calculate the *cscore*, described below, partially compensates for the absence of this normalisation step.

4.2.2 Robustness of folding

This measure refers to the fraction of base-pairs that are preserved across a set of sub-optimal structures and is implemented by RNAfold [45] as a measure of ensemble diversity [130].

The value of this measure is the average base-pair distance (d_{bp}) between the optimal (p_0) and each of the sub-optimal structures (p_i) in the thermodynamic ensemble, weighted by their probability. The base-pair distance is defined as the number of base-pairs present in only one of the two structures being compared, i.e, the number of base-pairs that have to be opened or closed to transform one structure into the other. The value of the robustness of folding measure is given by the following formula:

$$s_2(p) = \frac{100}{|p|} \langle d_{\text{bp}} \rangle = \frac{100}{|p|} \sum_i d_{\text{bp}}(p_0, p_i) \frac{e^{-\Delta G_i / kT}}{Z} \tag{4.4}$$

where ΔG_i is the i-th energy level, k is the Boltzmann constant, T is the temperature, and $Z = \sum_i e^{-\Delta G_i / kT}$ is a normalising constant.

As shown in the formula, the computed value is normalised for a sequence with a length of 100 nucleotides.

4.2.3 Robustness with respect to context

This measure is a modification of the Self-containment index [64] (SC index) and it evaluates the impact of genomic context variations in the secondary structure of the precursor candidate, and can be summarised by this formula:

$$s_3(p) = \underset{i=1,\ldots,100}{\text{median}} \left\{ 1 - \frac{d_{\text{H}}(p'_{c_0}, p'_{c_i})}{|p'_{c_0}|} \right\} \tag{4.5}$$

where d_{H} is the Hamming distance, p'_{c_0} is the dot-parenthesis representation of the secondary structure of the precursor in its original genomic context, p'_{c_i} is the dot-parenthesis representation of the secondary structure of the precursor folded in the i-th randomised genomic context with any mismatched parenthesis replaced by dots.

Given the large number of candidates and the need for computational efficiency, instead of strictly preserving dinucleotide frequencies, as proposed in [64], we trained two single-state first-order Markov chains with the up- and downstream genomic contexts of each candidate, covering a length identical to the size of the candidate. The candidates are then re-folded in 100 random contexts generated according to the Markov models previously obtained. The value of this measure is the median proportion of the original structure that is preserved in each refolded candidate. The measure takes values between 0 (none of the original structure is preserved in more than half of the randomised contexts) and 1 (the entire structure remains intact in more than half of the randomised contexts).

4.2.4 Robustness with respect to mutations

The original formulation of a measure of mutational robustness was proposed in [18], and it was shown to be a notable property of miRNA precursors. In order to compute the value of this measure, for each candidate precursor, p, the entire 1-mutation neighbourhood, $\left\{\mu_i^1(p)\right\}_{i=1,\ldots,3|p|}$, is generated (i.e. the set of all sequences obtained from p by performing all possible point mutations at each position so that each mutant differs from p by only one nucleotide) and each mutant is folded. The value of this measure is the median base-pair distance between the original and each mutant structure, normalised to a sequence with a length of 100 nucleotides, summarised in the following formula:

$$s_4(p) = \frac{100}{|p|} \underset{i=1,\ldots,3|p|}{\text{median}} \left\{ d_{bp}(p, \mu_i^1(p)) \right\} \tag{4.6}$$

where $\mu_i^1(p)$ is the i-th 1-mutation of p (i.e. the i-th member of its 1-mutation neighbourhood), and d_{bp} is the base-pair distance.

The authors who proposed the measures of robustness with respect to context and robustness with respect to mutations, discussed above, calculated their scores by averaging the proportion of preserved structure and base-pair distance to each mutant, respectively. We, instead, preferred the median of these values since it is a more robust centrality measure when dealing with values originating from possibly non-Gaussian distributions.

4.2.5 Combining measures

In order to combine the measures described above into a single score, we determine the significance of the value of the measures for each candidate against its empirical distribution in a random genome with similar dinucleotide frequencies. To that effect, a single-state first-order Markov chain is trained with the sequences of the genome whose candidates are to be evaluated, which is then used to generate a single 5 Mb sequence. Precursor candidates are extracted from these artificial genomes using the same procedures as the ones previously described for the original genomes and each of these artificial candidates is evaluated using the four robustness/stability measures. The use of empirical distributions for our measures is justified by the fact that the underlying probability distribution is unknown.

The combined score (*cscore*) of a precursor candidate is given by the product of the significance of each measure value against the corresponding empirical distribution in the artificial genome, i.e., the proportion of candidates in the artificial genome which have worse scores than the candidate under consideration. This notion of worse can either mean lower or higher values, depending on the biological interpretation of the measure. For instance, a worse value for the AMFE measure is, in fact, a higher value since it refers to less stable structures. On the other hand, for the *Robustness to context* measure, worse values are lower, since they refer to structures which are more poorly preserved in varying genomic contexts.

Let F_i be the empirical cumulative probability function for the distribution of the i-th evaluation measure on a randomised genome with the same statistical properties of the genome of interest, then the *cscore*, $s(p)$, for the candidate precursor p is

$$s(p) = \prod_{i=1}^{4} F_i(s_i(p)) \tag{4.7}$$

The *cscore* thus varies between 0 (the candidate scores worse than all candidates in the artificial genome for at least one measure) and 1 (the candidate scores better than all candidates in the artificial genome for all measures).

4.3 Structural analysis

An important feature of pre-miRNAs that elicits their recognition by the miRNA-processing machinery is their secondary structure. As we stated before, pre-miRNAs typically exhibit a stem-loop structure with few internal loops or asymmetric bulges but the variety of structures that are efficiently recognised has escaped any strict characterisation [77].

The intrinsic measures presented in section 4.2 can be used to identify and select robust and stable stem-loop structures. However, the candidates thus identified are still impractically numerous to be subjected to experimental confirmation (see Chapter 5), and there is no guarantee that a robust and stable stem-loop meets the structural requirements of the enzymes involved in miRNA maturation.

The most immediate approach to analysing the variety of pre-miRNAs in our candidate set is to seek the identification of families amongst the precursor candidates. Albeit miRNAs have been grouped into families according to their sequence similarity in the miRBase database [35], this approach does not give enough insight as to the structural features that are important for the recognition by the miRNA-processing machinery. Hence, the grouping has to be performed according to sequence *and* structure. Various algorithmic approaches have been introduced to determine structural similarities and to derive consensus structure patterns for structural RNAs with low sequence identity [111, 42, 107, 33, 39, 82, 19, 43, 127, 14]. However, all these approaches suffer from a high computational complexity, with a time requirement typically between $O(n^4)$ and $O(n^6)$.

A first approach towards clustering of microRNAs has been achieved in [51], where the sequence-structure alignment tool `Foldalign` [33, 39] was used to cluster 220 microRNAs into structural classes. However, it is computationally unfeasible to cluster hundreds of thousands of candidates using this approach.

In this section, we present a vectorial representation which purports to summarise the structural and sequence features of each candidate, a method used to map the candidates onto a multidimensional feature space, and we describe the procedure used to identify the region of the feature space most likely to contain candidates with the appropriate structural requirements. Additionally, we describe the procedures used in Chapter 5 to validate our choice of vectorial representation and to demonstrate that it approximates the results of conventional structural clustering.

4.3.1 Vectorial representation of primary/secondary structure

In this thesis, we use a vectorial representation for candidate precursors which summarises key features of the primary/secondary structure of a given stem-loop. The representation we chose, after considering several options and selecting the one that best matched the results of conventional clustering (see Chapter 5), consists of a vector of normalised counts. In order

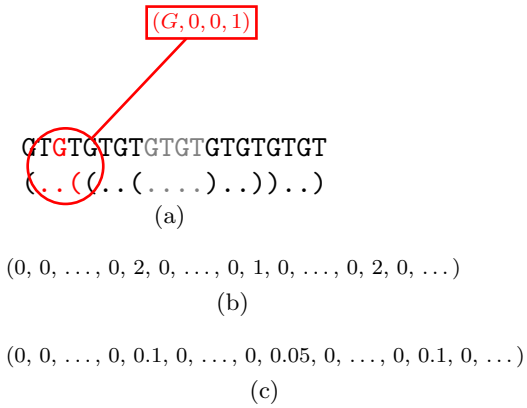

Figure 4.2: Example of a vectorial representation.
(a) The characteristics of a single position are determined, which include the nucleotide and whether the previous, current and following positions in the secondary structure are left/right paired, unpaired or located in the terminal loop. **(b)** Portions of the final vector illustrating the counts. Each vector position refers to a particular nucleotide type and the neighbouring pairing status, from $(\mathtt{A}, \mathtt{0}, \mathtt{0}, \mathtt{0})$ to $(\mathtt{G}, \mathtt{3}, \mathtt{3}, \mathtt{3})$. **(c)** Portions of the normalized vector obtained from (b), each position is divided by a constant such that the sum of all components is 1.

to build this vector, we use a sliding window of length 3 (a triplet) that scans the precursor candidate. At each step, a position in the vector is incremented. The appropriate vector position is mapped considering whether each nucleotide within the window and with respect to the MFE structure is the left/right-hand side of a base-pair, an unpaired nucleotide on the stem, or part of the terminal loop, and, additionally, which base is present at the midpoint of the window. We have, thus, a vector with 256 positions. After scanning the entire precursor, each position in the vector is normalised by dividing its counts by the length of the sequence.

Formally, the features used in our vectorial representation can be represented as a 4-tuple $(\alpha, \phi_{-1}, \phi_0, \phi_{+1})$ where $\alpha \in \{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}$ refers to a nucleotide in the sequence, each $\phi_i \in \{0, 1, 2, 3\}$ is a function indicating whether the structural position at coordinates $i \in \{-1, 0, 1\}$ (with respect to the nucleotide α) is unpaired, left-paired, right-paired, or located at the terminal loop. Fig. 4.2 gives an example of the vectorial representation for a particular stem-loop.

A similar representation has already been used to represent feature vectors of RNA stem-loops in the context of training a support-vector machine [132] and is amongst the representations we have evaluated. The representation we use here is richer than the one proposed by the authors in the sense that it distinguishes the situation where a given position is the left or right-hand side of a base-pair instead of simply being a paired position and it also represents unpaired nucleotides in the stem region or the terminal loop differently. This way, information about asymmetrical loops and bulges in the stem is captured by the vector counts, and the number of nucleotides involved in the terminal loop is also accounted for.

4.3.2 Feature space

The vectorial representation of hairpins used in the structural analysis of our precursor candidates captures information about sequence/structure features but, in general, the dimensions of these feature vectors are not independent making it difficult to draw conclusions from the analysis of the spatial distribution of the candidates in the vectorial space. Furthermore, all vectors will always have zero values in some dimensions as some combinations of left/right-hand paired and unpaired nucleotides are not possible in actual RNA structures.

In order to reduce the number of dimensions of the feature space and to ensure that the structures are represented in a space with linearly independent dimensions we perform some modifications to the space defined by the vectorial representations. First, we readily eliminate dimensions with zero variance. In practice, for a sufficiently large dataset, this will only eliminate dimensions for which all vectors have value zero. On the remaining dimensions we apply a scaling procedure to make sure each dimension exhibits unit variance and then we perform a principal components analysis (PCA). The principal components thus obtained become the dimensions of the feature space.

The representation of each candidate in this feature space, by itself, does not elicit the identification of a region of interest corresponding to the portion of the multidimensional space satisfying the necessary structural requirements. The identification of that region, which we will call *acceptance region*, has to be seeded by some point or points referring to structures which are known to pertain to the set of accepted hairpins.

If we assume that there is a single connected acceptance region, we can calculate the centroid of the known precursor structures for a given dataset and use it as the seed to select all the candidates inside a sphere centred at the calculated centroid and use the radius of the sphere to define how conservative or how inclusive the selection should be. If, on the other hand, we allow for an acceptance region with disconnected components, we can take the set of points corresponding to the known precursor structures, $\mathcal{P}$, and select every candidate, c, such that

$$\min_{p \in \mathcal{P}} d(c, p) < r, \qquad r > 0 \tag{4.8}$$

where $d(c, p)$ is the Euclidian distance between the representation of $c, p \in \mathcal{P}$ in the feature space. Any of these two approaches has the advantage of not requiring the specification of negative examples which, lacking experimental confirmation, would arguably be arbitrary and, in any event, would also likely be unrepresentative. On the other hand, by relying on the prior identification of known precursors, our approaches are predicated on the assumption that the acceptance region is not fragmented across the feature space but rather concentrated in at most a small number of disconnected components and that the known precursors are representative of the acceptance region. This last concern can be mitigated by the inclusion of

precursors from other species in the seed set, under the additional assumption that structural requirements for precursor hairpins are sufficiently similar across species.

4.3.3 Validation of the vectorial representations

Our approach to the structural analysis of candidate hairpins is based on the observation that performing conventional structure-based clustering is unfeasible for a large set of hairpins as the one we are faced with when analysing the stem-loops extracted from a genome-wide scan. The solution we present in the previous sections assumes that the adopted vectorial representation and the space transform operated by the PCA places the candidates distributed in such a way that their relative distances reflect their structural similarities. In order to test this assumption it is necessary to compare our approach to a conventional structural clustering procedure, which is only possible if small samples of our candidate set are used. In Chapter 5 we present the results of the validation of our approach using the procedures discussed below.

4.3.3.1 Randomisation procedure

A randomisation procedure is used during the analysis of how well the distances in the feature space match the results obtained for several samples of the candidate sets using a conventional structural clustering method, in terms of the proportion of correctly assigned cluster members. A member of a structural cluster is said to be correctly assigned if, upon calculating the centroid of all the cluster members in the feature space, the centroid closer to the structure under consideration is the one pertaining to its structural cluster rather than the centroid of some other cluster. This procedure allows us to estimate the likelihood that our values were obtained by chance or as a result of the particular way our candidates are spatially distributed in the feature space.

To obtain the proportion of correctly assigned cluster members in the randomised version of the samples, we keep each candidate in the same position of the feature space but we shuffle their identities. So, in other words, we randomly select two candidates and we swap their co-ordinates, repeating the process until all candidates have had their co-ordinates swapped. After having performed the random swapping of candidates we re-calculate the centroids of each cluster, but now in the shuffled space, and the resulting proportion of correctly assigned cluster members.

The significance of the proportion of correctly assigned cluster members obtained in the original feature space is determined by performing a two-sample Welch t-test using the original values against the values in the shuffled space over several samples of the candidate set.

4.3.3.2 Conventional structural clustering

In order to generate reliable partitions of a dataset, we apply a clustering procedure based on RNA sequence-structure alignment. For this purpose we use `LocARNA`, which is one of the fastest and most accurate tools for multiple RNA sequence alignment [127]. `LocARNA` performs Sankoff-style simultaneous alignment and folding [107]. This approach generates high quality alignments that take structural similarity into account. Notably, the structural information is not required *a priori* but can be inferred, in parallel to the alignment process, based on an RNA free energy model. `LocARNA` achieves its short run-times for pairwise alignment because it needs to consider only significant base-pairs.

A hierarchical cluster tree is generated by applying an average-linkage clustering (UPGMA) to the matrix of pairwise `LocARNA` distances. This procedure in combination with `LocARNA` was validated by a re-clustering of Rfam [127]. At high average recall, the Rfam families were reproduced with good precision.

In the case of clustering miRNA candidates, we do not have any prior knowledge of clusters. Therefore, we need to define a reasonable partitioning of the cluster tree into an optimal number of clusters. For this purpose, we apply a variant of the Duda rule [29]. Under this rule, a subtree is reported as an optimal cluster if the sum-of-squared error for two clusters is not significantly smaller than what would be expected by chance [51]. The significance level can be controlled by k. The larger is the value of k, the larger the difference of squared error allowed before a subtree is split into two clusters. In our case, the error of a cluster is determined via the free energy of its consensus structure and the minimum free energies of its individual sequences. The minimum free energy of single sequences is calculated by RNAfold [45]. The consensus structure and energy is calculated by RNAalifold [44] based on a multiple `LocARNA` alignment of the subtree.

4.4 Transcription potential

An essential condition for a given stem-loop to be a functional miRNA is that it is on an adequate genomic context so that it is efficiently transcribed when needed. An hairpin may otherwise fulfil all the requirements for effective recognition by the cellular miRNA-processing machinery but if it is not appropriately transcribed it will never direct the silencing of a target gene in a physiologically relevant manner.

From a computational biology perspective, the approaches to the determination of the transcription potential of a precursor candidate are dependent on the available data, such as genome annotation or experimental data from transcriptomics assays.

In this section, we discuss three approaches to evaluating transcription potential: annotation data, genomic location and genome mapping of sequenced small RNAs.

4.4.1 Annotation data

Although not directly implicated in the assessment of the transcription potential, the availability of annotation information can help us exclude candidates which may actually be transcribed but whose corresponding transcripts are associated with other functional roles. This includes candidates which overlap same-strand regions annotated as transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), transposable elements and other repeat sequences or exons of protein coding genes, since these regions are unlikely to contain functional miRNA precursors despite the fact that they may otherwise mimic features usually associated with pre-miRNAs.

On the other hand, annotation information can also help us identify intronic miRNAs as well as *mirtron* candidates, both originating from introns in a splicing-independent and -dependent manner, respectively. Candidates contained in introns are certainly transcribed and by this criterion alone should be paid special attention.

4.4.2 Genomic clusters

When some pre-miRNAs are already known for a given organism it is possible to identify a group of candidates which are more likely to be transcribed [109]. These refer to candidates whose genome location falls near previously annotated miRNA precursors. This is based on the observation that pre-miRNAs in animals tend to occur in tandem and be, occasionally, simultaneously transcribed in a unique pri-miRNA containing several precursors. The most obvious shortcoming of this approach is that it is limited to the identification of new members of previously described miRNA genomic clusters.

4.4.3 Mapping sequenced small RNAs

Despite the limitations of the use of experimental data to systematically identify novel miRNAs, due to the expression profiles of some of these regulators, which may be transcribed only in particular tissues, physiological conditions or at very low-rates, but also as a result of the persistence of biases in RNA sequencing techniques, they have been successfully used to identify a great number of miRNAs.

In the following paragraphs, we describe our approach to mapping small RNA transcripts to the genome. In this thesis, we use this procedure in order to validate the transcriptional potential of pre-miRNA candidates, but it could easily be used to identify an initial set of candidates by seeking to find an appropriate stem-loop surrounding the genome hit.

4.4.3.1 Genome localization

The objective of our method is to determine the most likely genome location of the sequenced RNA transcripts. To this effect, we use an approximate matching procedure reporting

all hits with up to 2 errors. The choice of a tolerance of 2-nucleotide mismatches in order to identify a hit is made to account for the possibility of sequencing errors or sequence edit events. Only transcripts with a length between 16 to 27 nucleotides are considered, as these capture all the known variability in size for mature miRNAs.

A transcript may have no genomic hits, in which case it may either be foreign RNA, the result of a contamination, the sequencing process might have suffered more errors than our tolerance or the transcript may have been extensively edited. In general, the most likely genomic origin for a transcript corresponds to the best match. Considering, additionally, that mature miRNAs are much more commonly sequenced in RNA sequencing assays than any other portion of the intermediaries (notably, the pre-miRNA), the fact that a match falls within the stem portion of a candidate stem-loop is an important indication as to whether our transcript could be a *bona fide* miRNA.

For transcripts with genomic hits it is important to determine whether the number of hits is statistically significant, given the expected the number of occurrences for the corresponding sequence characteristics of both the transcript and the genomic context of the hits.

4.4.3.2 Statistical model for genome hits

The probability of occurrence of each word w, $\mu(w)$, is determined in their genomic context using a single-state first-order Markov chain, $\mathcal{M}$. The Markov chain captures the background distribution of nucleotides and dinucleotides despite the fact that we do not distinguish between annotated and non-annotated regions. This is mitigated by the choice of a context of appropriate length.

The probability of occurrence of a word w, $\mu(w)$ has to account for the fact that the genome occurrences can have up to 2 mismatches and can occur in the reverse strand. Let $p^{\mathcal{M}}(w)$ be the probability of w according to the Markov chain $\mathcal{M}$, let $\nu_e(w)$ be the e-mismatch neighborhood of w, i.e., the set of all words with up to e mismatches of w and let $\overline{w}$ be the reversed-complement of w, then:

$$\mu(w) = \sum_{v \in \nu_e(w)} \left(p^{\mathcal{M}}(v) + p^{\mathcal{M}}(\overline{v}) \right) \tag{4.9}$$

The expected number of occurrences of a word w in a given portion of the genome is bounded by $l_c\mu(w)$ where l_c is the size of the genomic context. The number of occurrences can be modeled by a Poisson distribution $X \frown Poisson(\lambda)$ with parameter $\lambda = l_c\mu(w)$. The p-values determined for each sequence are obtained by calculating the probability of having a higher number of occurrences given our distribution, $Pr\{X > N(w)\}$, where $N(w)$ is the actual number of occurrences.

This distribution is not exact, since the words we seek can only rarely overlap and therefore each occurrence is not strictly independent of the others, as is assumed by our computation

of the expected number of occurrences. As a result, the p-values of the exact distribution are smaller than the ones we determine.

We recall that the null hypothesis in this case is that the hits occur by chance. By selecting a significance level of 5% we should be able to sieve out most spurious hits. However, we have to adjust our significance level due to the fact that we perform extensive multiple testing. To that effect, we apply the Bonferroni correction whereby the adjusted significance level is $\frac{\alpha}{N}$ where α is the baseline level (in our case, 5%) and N is the number of tests, which in our case refers to the number of genome-wide hits of each sequence with up to e errors.

4.5 CRAVELA framework

The CRAVELA framework consists of the public tool for the analysis and presentation of data regarding the identification and evaluation of miRNA precursor candidates in metazoan genomes. This tool integrates most of the contributions that have been described in the previous chapters.

This framework comprises three distinct aspects: 1) the database model where information is represented and kept, 2) the processing pipeline, 3) the web-based presentation of the data.

4.5.1 Database model

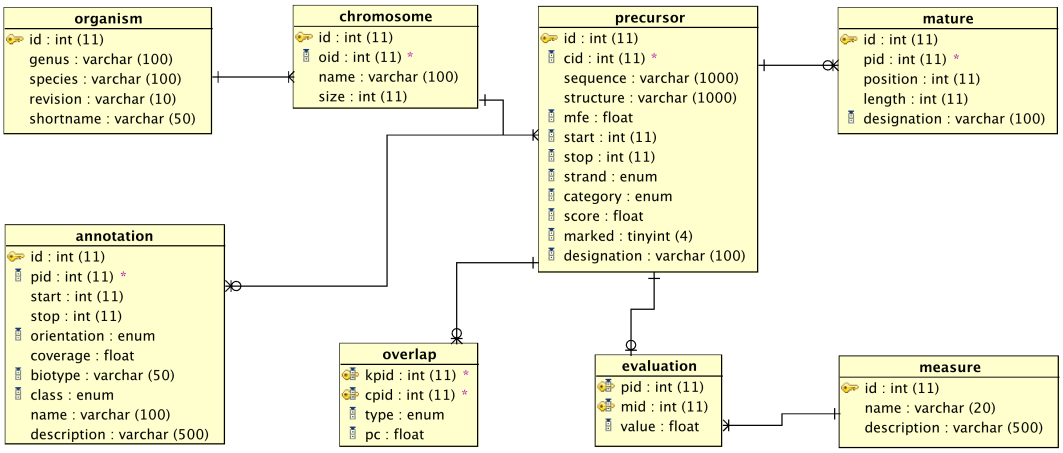

Figure 4.3: Database ER model

Figure 4.3 shows the Entity-Relationship model of the database used in the CRAVELA framework. The most important table is `Precursor` which represents both precursor candidates and annotated miRNAs. The genome location of each precursor is anchored on a chromosome or a contig read represented by an entry on the `Chromosome` table. Each chromosome is, in turn, associated with an entry on the `Organism` table. Each entry on this table represents a single dataset. Associated with the main `Precursor` table we have the `Mature`

table which represents the location of the mature transcript (when known) in the annotated precursor sequence, the `Overlap` table which represents the superposition of candidates and known precursors and the `Evaluation` table, which associates a value to each precursor according to a measure described in the `Measure` table. Appropriate indices were created for the most common update and selection operations, in order to speed up such tasks.

Additionally, the `Precursor` table allows for a master filter using the `marked` field so that any group of precursor candidates may be discarded for the purposes of any portion of our analysis based, for instance, on additional external information which may authoritatively indicate that some candidates cannot possibly be pre-miRNAs.

4.5.2 Processing pipeline

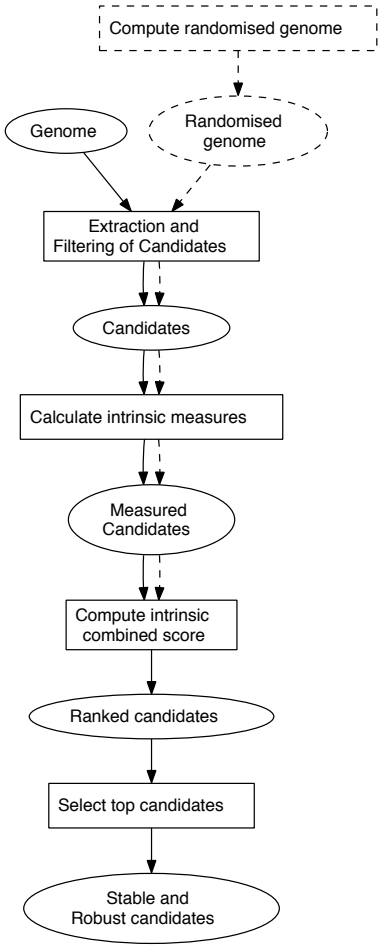

Figure 4.4: CRAVELA processing pipeline: extraction, evaluation and selection of stable and robust candidates using a combination of measures

Figure 4.4 shows the first portion of the CRAVELA processing pipeline that was described in section 4.2. All the steps represented in the figure are implemented using PERL, except for the procedure to combine all different measures into the *cscore*, which is implemented using R scripts. The step which calculates intrinsic measures requires the use a computer cluster with several cores due to the intensive computational needs which include folding several hundred different versions of each of the few millions of candidates which are to be analysed.

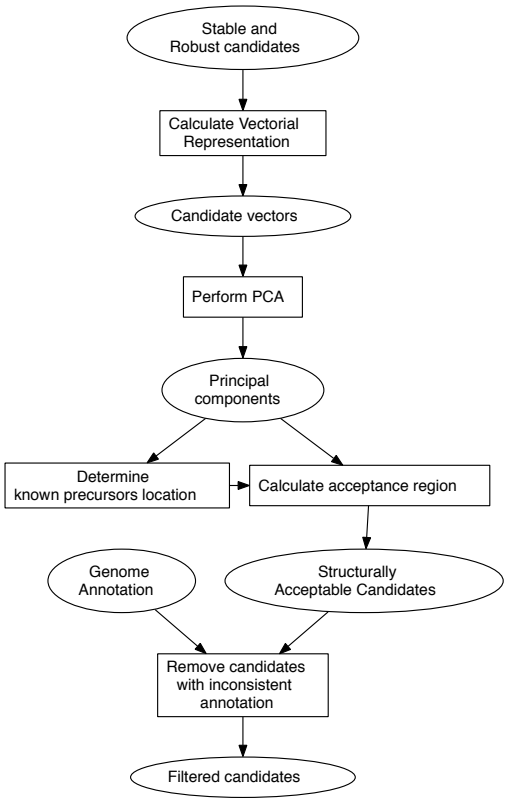

Figure 4.5: CRAVELA processing pipeline: structural analysis and annotation filtering

Figure 4.5 illustrates the second portion of the pipeline showing all the steps involved in evaluating both the structural requirements of the candidates and the consistency of the annotation data with the possibility of a candidate being a miRNA precursor. All steps are implemented using PERL programs except for the PCA calculations and the statistical validation procedures which are done using R scripts.

The third portion of the pipeline consists in the assessment of the transcriptional potential of each candidate, which ultimately depends on the additional information which might be available and is, therefore, tailored to each dataset. Examples of types of data which can be used for this purpose are deep sequencing data, EST data or small RNA libraries, each requiring a specific type of specialised treatment.

4.5.3 Web interface

The web interface, which is under active development, is a means to navigate the wealth of data produced by the CRAVELA pipeline. At present, besides a presentation of the *intrinsic measures* discussed in section 4.2, it is possible to browse through the known precursors of the listed organisms and compare them to the overlapping candidates, in terms of structure, score, and genome location, as is illustrated in Figure 4.6.

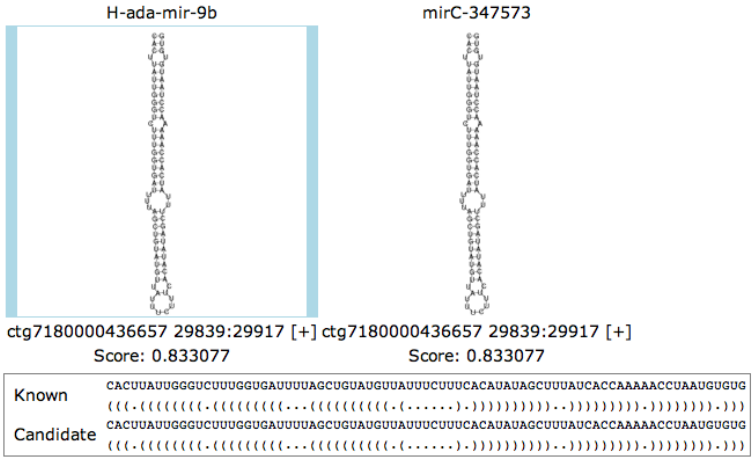

Figure 4.6: CRAVELA web interface: Example of a known precursor and best matching candidate

Part III

Results and Discussion

Chapter 5

Results and Discussion

In this chapter, we present and discuss the results obtained by applying our pipeline to metazoan genomes, in particular, to the genomes of *Drosophila melanogaster*, a very well-studied organism, for which many researchers believe most miRNAs have been identified thus providing one of the most complete miRNA catalogues for a single organism, and *Anopheles gambiae*, which is the main vector of malaria and where miRNAs are thought to play a role in parasite resistance [129]. Additionally, some exploratory results are shown for the recently sequenced genome of *Anopheles darlingi*, the principal vector of dengue and which is a species close to *A. gambiae*, including the detection of clear pre-miRNA homologs.

We present the results obtained at each step of the processing pipeline, including the enumeration of candidates, the evaluation using a combination of intrinsic measures, the structural analysis, and the assessment of the transcriptional potential of our candidate set for each organism, in particular, the use of a small RNA library from *A. gambiae* to identify the most promising candidates. In addition, we present some results that justify certain decisions that were made, particularly the parameters of our models and the way the secondary structure of our precursor candidates is represented.

5.1 Data preparation

The genome sequences for the chromosomal arms X, 2R, 2L, 3R, 3L and UNKN of *A. gambiae* (assembly AgamP3) were obtained from the Ensembl ftp site (`ftp://ftp.ensemblgenomes.org/pub/metazoa`). Both the euchromatic and heterochromatic genomic sequences of *Drosophila melanogaster* (release 5) were obtained from the BDGP project website (`http://www.fruitfly.org/sequence/release5genomic.shtml`). All the sequences concerning known pre-miRNAs and their respective mature sequences were recovered from the miRBase webserver (`http://www.mirbase.org`, release 13).

The Whole Genome Shotgun project of the newly-sequenced *A. darlingi* has been deposited at DDBJ/EMBL/GenBank under the accession ADMH00000000. The version de-

scribed in this thesis is the first version, ADMH01000000. It consists of 18 629 contigs with a total size of 173 473 443 nucleotides.

The small RNA library for *A. gambiae* was kindly made available by C. Brunel and collaborators and it was produced in the context of a work published in [129].

The annotation data for both *A. gambiae* and *D. melanogaster* were obtained from EnsEMBL, using their PERL APIs.

5.2 Enumeration of candidates

In order to evaluate the sensitivity of this procedure, it is necessary to assess whether known precursors are found amongst the extracted candidates. The number of known precursors and precursor candidates extracted from the datasets using the enumeration procedure described in section 4.1 are shown in Table 5.1.

Dataset	**Positive Set**	**Negative Set**	**Overlapping**	**Extracted candidates**
A. gambiae	67	2 244 922	92	2 245 014
D. melanogaster	157	1 316 105	200	1 316 305
A. darlingi	44	1 748 087	66	1 748 153

Table 5.1: The number of elements in the positive and negative sets, the number of candidates overlapping elements of the positive set and the total number of extracted candidates for each dataset.
For *A. darlingi*, the positive set is the set of clear homologs to *A. gambiae* pre-miRNAs whereas in all other datasets the positive set corresponds to the known precursors. In all cases, the negative set consists in the presumptive non-precursor candidates which do not overlap precursors in the positive set.

Annotated sequences may or may not include sequences flanking the pre-miRNAs because the precise co-ordinates of the precursor hairpins are not always experimentally determined. The stem-loops identified by our enumeration strategy are extended stem-loops in the sense that they are the largest stem-loops contained in their local genomic contexts, and will therefore tend to be larger than precursor hairpins. In both cases, if the flanking sequences are short with respect to the actual precursor, the impact on the candidate evaluation procedure is likely to be modest.

Only very few known precursors are not matched by any candidate (1 out of 67 in *A. gambiae*, and 6 out of 157 in *D. melanogaster*). Since the enumeration procedure can only identify canonical stem-loops, some of these cases refer to multi-loop structures that share a common stem but with relatively small secondary stem-loops which fail to pass the minimum length criterion. Others are short structures which are filtered by the -20 kcal/mol stability criterion (see description of the enumeration procedure in section 4.1).

Additionally, the vast majority of annotated precursors with known mature sequences has

a best match which includes the mature sequence in one of its stem arms (59 out of 67 in *A. gambiae*, and 134 out of 152 in *D. melanogaster*). It is worth to point out that in most cases a best match noted as not including the mature sequence in the stem arm in fact only misses the start or end of the mature sequence by a few nucleotides (beyond a 2-nucleotides tolerance) because of a missed dangling end.

5.3 Robustness and stability measures

In order to separate the problem of correctly identifying candidate precursors with that of assessing the performance of our evaluation measures, we take the known pre-miRNAs in each dataset as the positive set, and the negative set is made of all the candidates which do not overlap the co-ordinates of known precursors. The number of elements in the positive/negative sets for each dataset is shown in Table 5.1.

The negative sets may include several yet unidentified precursors whose identification would have an impact on our performance assessment. To mitigate this problem and to assess the stability of the cut-off value for each measure as well as the combined score, and, more importantly, to deal with the greatly uneven positive/negative set sizes, we have adopted an undersampling procedure for the negative sets. In this procedure, we randomly extract 1 000 samples from the negative set each having the size of the positive set.

5.3.1 Performance of the combined score vs. individual measures

Figures 5.1 and 5.2 show the performance of the evaluation measures for the *A. gambiae* and *D. melanogaster* datasets, respectively. Table 5.2 summarises the sensitivity and specificity values obtained for the *cscore*.

Dataset	Avg Optimal Cut-off	Sensitivity	Specificity
A. gambiae	0.41	0.90	0.88
D. melanogaster	0.30	0.83	0.80
A. darlingi	0.32	0.89	0.84

Table 5.2: The average optimal *cscore* cut-off value for each dataset over the 1 000 samples, alongside the average sensitivity and specificity values at each sample's optimal cut-off.

The AMFE [140] measure performs best in the *A. gambiae* dataset. The fact that this measure does not compensate for GC content, which has a significant impact on folding free energy values, may explain the disparity of the results. The *D. melanogaster* dataset includes both euchromatic and heterochromatic sequences with different GC content, the latter having considerably lower values. The procedure used to combine the evaluation measures partially compensates for the lack of GC content normalisation because the randomised dataset is

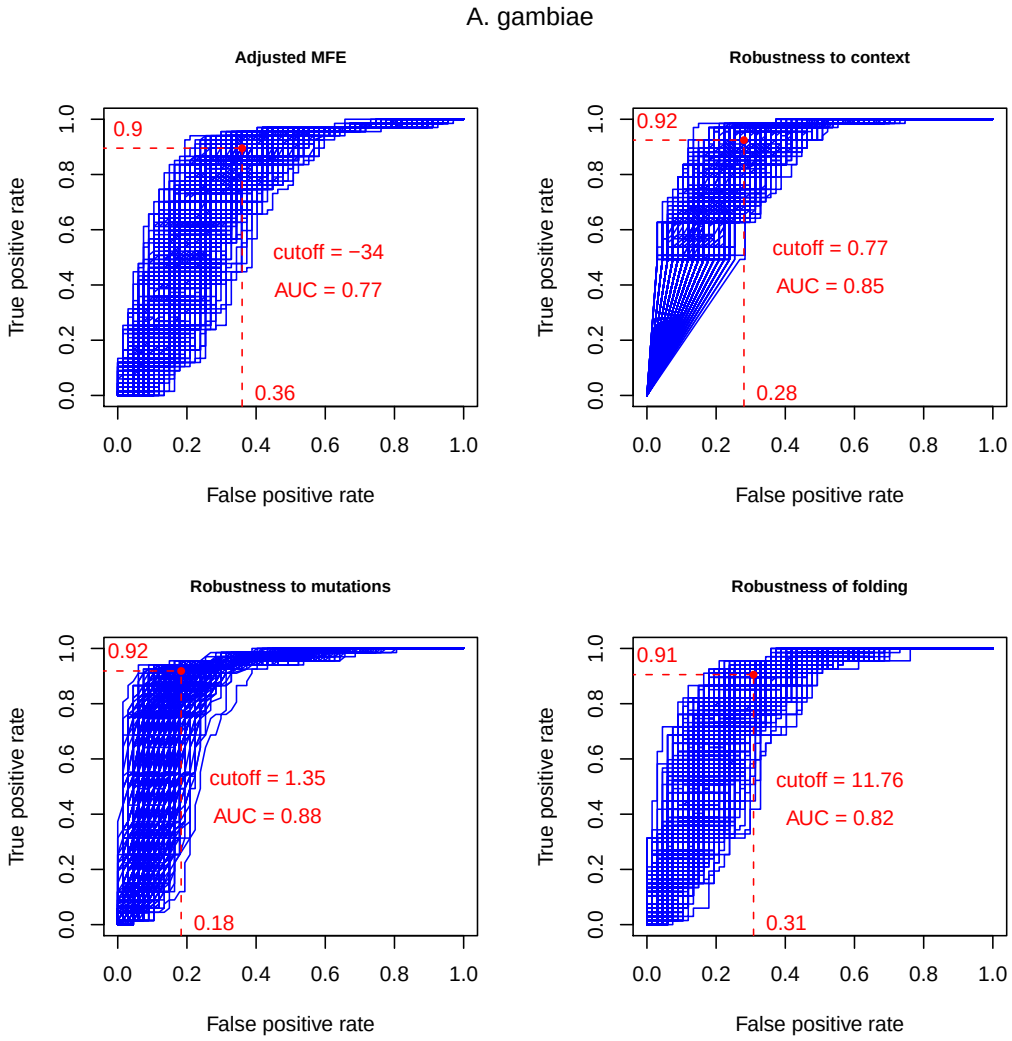

Figure 5.1: ROC curves for the evaluation measures in the *A. gambiae* dataset. The dashed lines indicate the true/false positive rates for the average optimal cut-off, i.e., the average cut-off value that maximises the difference between true and false positive rates ($\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}), \mathrm{FPR} = \mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$). The negative sets consist of 1000 samples, each of the size of the positive set, drawn from the non-overlapping candidates. The average optimal cut-off value and the average AUC (area under the curve) value are also shown.

generated maintaining the same dinucleotide distribution of each of the original sequences. Replacing the AMFE with a modified version of the MFEI [140] measure, which does compensate for GC content (discussed in section 4.2), had no discernible impact on the combined score (data not shown).

In both the *A. gambiae* and *D. melanogaster* datasets, the *Robustness of folding* and the *Robustness to context* measures have comparable performances in terms of average AUC, which summarises the relation between the true/false positive rates across all possible cut-off

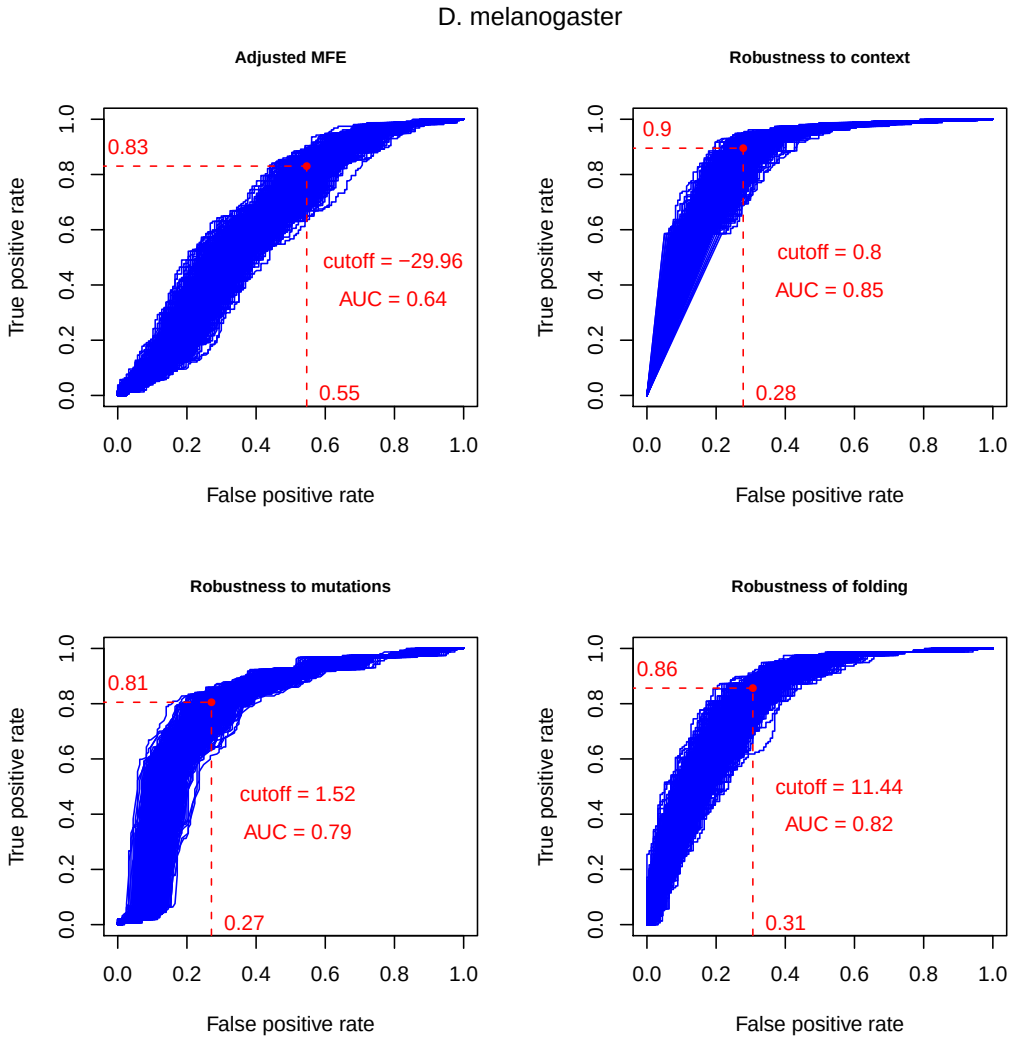

Figure 5.2: ROC curves for the evaluation measures in the *D. melanogaster* dataset. The dashed lines indicate the true/false positive rates of the average optimal cut-off, i.e., the average cut-off value that maximises the difference between true and false positive rates ($\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}), \mathrm{FPR} = \mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$). The negative sets consist of 1000 samples, each of the size of the positive set, drawn from the non-overlapping candidates. The average optimal cut-off value and the average AUC (area under the curve) value are also shown.

values for each of the samples of the negative set.

The *Robustness to mutations* measure performs well with the *A. gambiae* dataset but the performance in the *D. melanogaster* dataset is negatively influenced by the presence of several long inverted repeats (mainly due to the inclusion of heterochromatic sequences) that are resilient to point mutations and thus attain a high score for this measure. These sequences should not be summarily excluded as they can include true miRNA precursors.

The results for the combined score (*cscore*) for both datasets are shown in Figures 5.3, and

5.4. In both cases, the *cscore* performs better than any individual measure in terms of average AUC, which means that, in general, for the same false positive rate (FPR = FP/(FP + TN)) one can attain higher sensitivity (Sensitivity = TPR) or, conversely, for the same true positive rate (TPR = TP/(TP+FN)) one can expect better specificity (Specificity = TN/(TN+FP) = 1 − FPR). The optimal cut-off in each sample is calculated with respect to the Youden index [135], J, defined as $\max_c \{\mathrm{TPR}(c) - \mathrm{FPR}(c)\}$, i.e., the maximum difference between the true positive rate (TPR) and the false positive rate (FPR) over all cut-off values, c, which is a standard method to select the best sensitivity/specificity compromise in ROC curves. The optimal cut-off value, c^*, is therefore the cut-off for which $J = \mathrm{TPR}(c^*) - \mathrm{FPR}(c^*)$.

If we take the average optimal cut-off value for the *cscore* on each dataset and discard all candidates with a score below that cut-off, we obtain a reduced set of 328 829 candidates for *A. gambiae* and 287 469 for *D. melanogaster*. If we further observe that many of these candidates overlap and that it is very unlikely that two overlapping candidates are both true miRNA precursors, we can eliminate all candidates whose genomic locations are overlapped by other candidates with higher *cscore*, thus reducing the total number of candidates to 290 133 for *A. gambiae* and 240 751 for *D. melanogaster*. However, it is not clear that we can safely discard these candidates because the actual boundaries of the transcription unit harbouring the precursor rather than the *cscore* are the ultimate determinants of which overlapping stem-loop will be available for processing. In fact, we have identified a few cases of candidates overlapping known pre-miRNAs which have higher *cscore* than the precursors themselves (data not shown).

5.3.2 Performance comparison to other methods

Only a few classification methods can be readily compared to our combination of measures due to both the lack of use of conservation information and the need to evaluate a large number of candidates. Most methods, as we have shown in Chapter 3, either rely on conservation data or simply take too much time to determine whether a given candidate is likely to be a miRNA, making them unsuitable to classify millions of precursor candidates. TripletSVM [132] is a fast and well-known binary classification method that uses a support vector machine (SVM) to learn sequence/structure features of pre-miRNAs in order to distinguish them from other genomic stem-loops. The feature vector used to train the SVM considers the pairing states of every three nucleotides (triplet) plus the identity of the nucleotide at the middle. The results presented here were obtained using the method with default parameters and the SVM model provided by the authors. Being a binary classification method, TripletSVM cannot be used to generate ROC curves for a direct comparison with our method. HHMMiR [52] is a sophisticated method based on hierarchical hidden-Markov models. This method tries to learn the distinctive sequence/structure characteristics of different regions of the miRNA precursor. HHMMiR scores each candidate by calculating the ratio of the log-likelihoods

generated by the positive and negative models (learnt from known pre-miRNAs and random hairpins, respectively). Unlike with TripletSVM, the fact that HHMMiR can associate a score with each candidate elicits a direct comparison with our approach using ROC curves. Like before, the results presented for HHMMiR were obtained using default parameters and the maximum likelihood models provided by the authors.

The results presented in Figures 5.3 and 5.4 show the comparative performance of the *cscore*, TripletSVM and HHMMiR. The graphs illustrate the fact that TripletSVM tends to sacrifice sensitivity in order to obtain better specificity. In all datasets, the average performance of *cscore* always outperforms the average performance attained by TripletSVM.

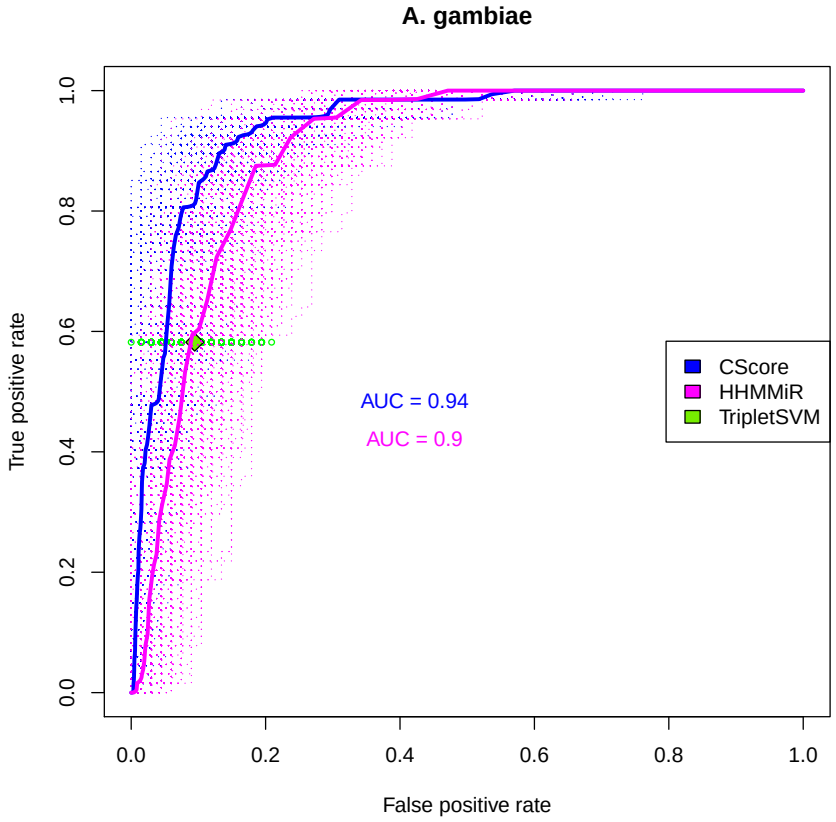

Figure 5.3: ROC curves for the *cscore* and HHMMiR in the *A. gambiae* dataset. The negative sets consist of 1000 samples, each of the size of the positive set, drawn from the non-overlapping candidates. The dashed lines are the individual ROC curves for each sample. The solid lines are the average ROC curves. The average AUC (area under the curve) values are also shown. The green diamond represents the average performance of the TripletSVM pre-miRNA classifier and the smaller green circles represent its performance on each sample.

The performances of the *cscore* and HHMMiR are quite similar in terms of average AUC. The *cscore* slightly outperforms HHMMiR for the *A. gambiae* dataset, whereas the reverse

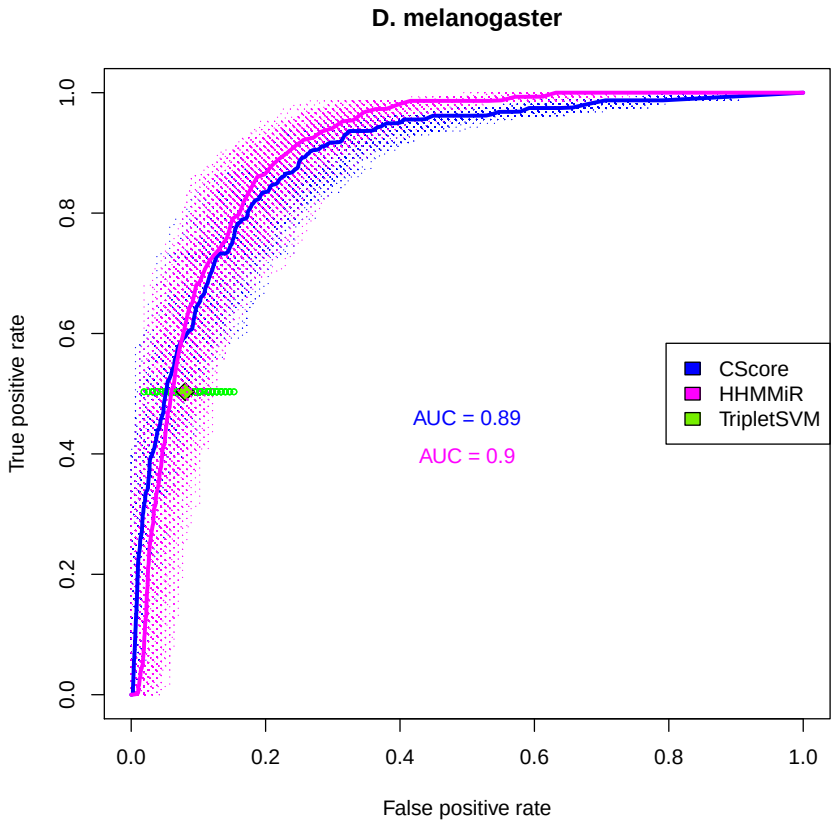

Figure 5.4: ROC curves for the *cscore* and HHMMiR in the *D. melanogaster* dataset. The negative sets consist of 1000 samples, each of the size of the positive set, drawn from the non-overlapping candidates. The dashed lines are the individual ROC curves for each sample. The solid lines are the average ROC curves. The average AUC (area under the curve) values are also shown. The green diamond represents the average performance of the TripletSVM pre-miRNA classifier and the smaller green circles represent its performance on each sample.

is seen in *D. melanogaster*. It is nonetheless surprising that a scoring scheme such as *cscore*, which makes no prior assumptions about precursor stem-loops except that they ought to be stable and robust, exhibits a performance comparable to a classifier that has been trained on known pre-miRNAs and is capable of sophisticated modelling of precursor sequences.

Both TripletSVM and HHMMiR are supervised learning methods which rely on training sets to produce a decision rule. In both cases, their ability to find novel miRNAs is dependent on how representative positive and negative examples turn out to be. The results presented here show that an approach that requires no prior training performs as well as the best of the two methods.

5.3.3 Exploration of precursor candidates in *A. darlingi*

Our candidate enumeration and evaluation procedures were applied to the newly-sequenced *A. darlingi*. A total of 1 748 153 precursor candidates were identified as shown in Table 5.1. To test our approach on a non-annotated genome, we analyse the performance of our *cscore* on three groups of candidates: those that are identified as homologs of known precursors from *A. gambiae*, those that are conserved in both genomes (excluding the homologs), and those which show low or no conservation (see section 5.3.4).

We found clear homologs of 44 *A. gambiae* pre-miRNAs supported by both high-quality mutually best alignments and the observation that in all cases the mature sequence is perfectly conserved. The list of homologs and the alignment of the mature sequence with the homologous precursors is shown in Appendix A. The number of homologs identified corresponds to 67% of the pre-miRNAs known in *A. gambiae*, which is the closest sequenced genome to that of *A. darlingi*. All remaining known precursors in *A. gambiae* except one, despite not having clear homologous precursor sequences, do have identical mature sequences within the stem-arm of a precursor candidate in *A. darlingi*, which could indicate homology through conservation at a structural level. Additionally, we identified 7 855 precursor candidates conserved in both genomes (see section 5.3.4).

Fig. 5.5 shows the analysis of the distribution of the *cscore* for the three groups of candidates. The median scores are 0.613, 0.034, and 0.033 for the homologs, conserved and non-conserved candidates, respectively. The conserved and non-conserved stem-loops have similar *cscore* distributions, but the scores for the set of homologs, however, are distinctively higher. The fact that the score distribution for conserved and non-conserved candidates is very similar reinforces the idea that conservation criteria alone are not sufficient to identify good precursor candidates.

Figure 5.6 shows the ROC curve for the performance of the *cscore* in the *A. darlingi* dataset using the pre-miRNA homologs as the positive set. The performances of TripletSVM and of HHMMiR are also shown. The results replicate what was observed in the other datasets. The *cscore* again outperforms TripletSVM and performs only slightly worse than HHMMiR.

There are 305 681 candidates above the average optimal cut-off for the *A. darlingi* dataset, which can be reduced to 248 970 by eliminating candidates overlapped by candidates with higher scores. Of these, 422 are found amongst the list of candidates conserved with respect to *A. gambiae*.

5.3.4 Identification of homologs and conserved stem-loops

Pre-miRNA homologs were found by performing a `Blastn` search and identifying a two-way best alignment with respect to the set of known pre-miRNAs of *A. gambiae* and the set of candidates from *A. darlingi*. Only homologous sequences that folded into stem-loops with MFE < -20 Kcal/mol and with both stem arms longer than 16 nucleotides were considered

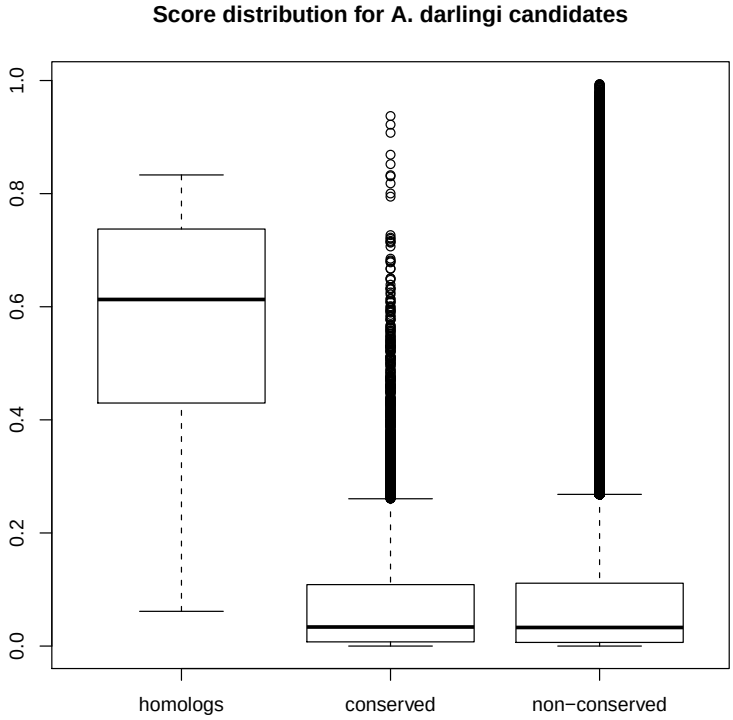

Figure 5.5: *Cscore* distribution in *A. darlingi* candidates broken down in homologs, conserved and non-conserved.

homologous precursors. Conserved stem-loops were also determined by a two-way best alignment between the candidates of both genomes, restricted to alignments with E-value below $1e-20$ and excluding all *A. gambiae* candidates that overlap known precursors.

5.4 Structural analysis

The computational search for novel miRNA precursors often involves some sort of structural analysis with the aim of identifying which type of structures are prone to being recognised and processed by the cellular miRNA-maturation machinery. A natural way to tackle this problem is to perform structural clustering over the candidate structures along with structures known to be recognised as pre-miRNAs and to try to identify which clusters contain known precursors and pay closer attention to candidates found therein. Given the large number of candidate pre-miRNAs that were identified by our approach, even after applying several stability and robustness filters, a conventional structural clustering approach is unfeasible.

In this section, we present the results of applying our method which represents candidate structures in a feature space summarising key sequence/structures characteristics of each

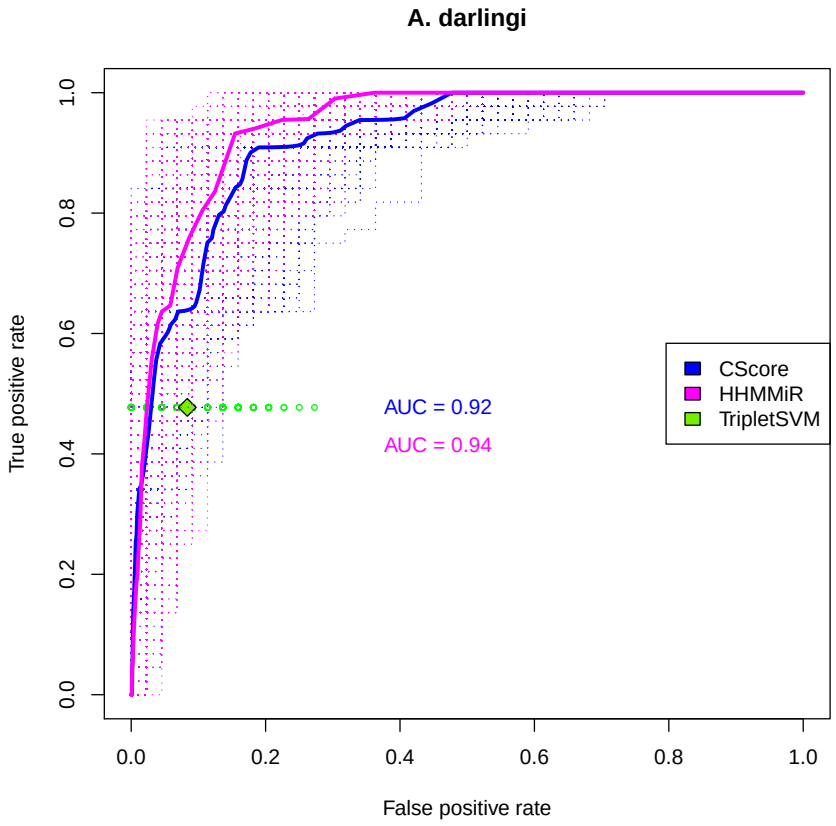

Figure 5.6: ROC curves for the *cscore* and HHMMiR in the *A. darlingi* dataset. The positive set consists of 44 clear homologs to *A. gambiae* pre-miRNAs. The negative sets consist of 1000 samples, each of the size of the positive set, drawn from the non-overlapping candidates. The dashed lines are the individual ROC curves for each sample. The solid lines are the average ROC curves. The average AUC (area under the curve) values are also shown. The green diamond represents the average performance of the TripletSVM pre-miRNA classifier and the smaller green circles represent its performance on each sample.

candidate. The results are shown for the two different approaches presented in section 4.3, the first assuming a single connected *acceptance region* and the second allowing for an *acceptance region* made up by smaller disconnected components. Additionally, we demonstrate that proximity in this feature space is related to sequence/structure similarity, which justifies the selection of candidates from the region populated by known pre-miRNAs as the best candidates, and we compare several different representations of our candidates in order to justify the adoption of the representation used throughout this section.

5.4.1 Structural similarity vs. proximity in the feature space

In order to assess the adequacy of our approach with respect to its ability to identify regions of structural similarity in a way that resembles conventional sequence/structure clustering we adopt the following procedure. We use `LocARNA` [128] to perform hierarchical structural clustering over 100 samples of 1000 randomly chosen stem-loops drawn from the *D. melanogaster* and *A. gambiae* datasets, always including the entire set of known miRNAs for each organism, and we determine the optimal partition into clusters applying a tree node evaluation rule for various significance levels called k-levels (the details of this procedure are described in section 4.3). For low values of k, the procedure produces clusters with highly similar structures. An increasing value of k allows for more dissimilar structures to be included in the same cluster, therefore producing a lower number of clusters with an increasing number of structures.

We then represent each structure from the samples using a vectorial representation summarising sequence/structural features, in a effort to capture the key elements distinguishing the various hairpins. These feature vectors contain, however, both interdependent dimensions and dimensions with different variance. In order to obtain a linearly independent set of dimensions, we perform a principal components analysis (PCA) over the vectorial representations mapping them to their principal components representation which we call the *feature space*.

To determine whether our representation of the candidates in the feature space reflects the structural clusters found by `LocARNA` for the different k-levels, we calculate the proportion of correct assignments, which, we recall, measures the ratio of cluster members that are closer to the centroid of their assigned cluster as opposed to a centroid of another cluster. The cluster centroid is calculated by determining the average position of the cluster members each dimension at a time. The distribution of this measure in our 100 samples is then compared to its distribution in a randomised version of our spacial representation of the candidates, where candidate positions are kept but candidate identities are shuffled.

The comparison with the randomised version of the spatial distribution of candidates allows us to address the fact that some clusters have only one member and will therefore always coincide with their cluster centroid and that variations in distance of a candidate to its assigned cluster centroid for different k-levels, or even different vectorial representations, may reflect, in part, the overall density of the candidates rather than a better evaluation of structural similarity.

Table 5.3 shows that, for both datasets, a large proportion of cluster members are found closer to their cluster centroid than to the centroid of any other cluster. For the most heterogeneous clusters which are obtained at k-level 0.90 the proportion of correctly assigned cluster members is about two thirds, and it rises above 80% for the structurally more homogeneous clusters obtained at k-level 0.00. The comparison with the randomised datasets shows that

the results are statistically significant, i.e., these results are well above what one would hope to obtain by chance or simply due to the way candidates are spatially distributed.

k-level	*A. gambiae*			*D. melanogaster*		
	Corr. assign.	*p*-value	Avg. cl. size	Corr. assign.	*p*-value	Avg. cl. size
0.00	83.60%	8.58e-87	3.05	82.10%	3.45e-125	3.29
0.10	82.50%	1.69e-87	3.33	81.24%	1.92e-120	3.54
0.20	81.21%	8.68e-84	3.70	80.01%	4.31e-113	3.89
0.30	79.30%	2.37e-76	4.27	78.24%	3.31e-108	4.44
0.40	77.09%	9.22e-65	5.31	76.08%	1.64e-96	5.44
0.50	74.12%	2.24e-55	7.61	72.80%	1.43e-84	7.54
0.60	71.23%	2.76e-42	12.09	69.45%	1.01e-61	11.44
0.70	68.70%	1.05e-31	17.24	67.41%	1.74e-54	15.32
0.80	68.14%	6.01e-26	19.52	66.07%	9.62e-44	17.77
0.90	67.37%	2.57e-22	21.08	65.72%	4.01e-37	20.09

Table 5.3: Evaluation of vectorial representations. For each k-level, the table shows the percentage of correct assignments in the datasets of *A. gambiae* and *D. melanogaster*, the p-value of Welch's two-sample t-test comparing the observed correct assignments with a randomised version of each dataset shuffling the correspondence between candidates and their vectorial representation, and the average number of cluster members.

5.4.2 Selecting the most adequate vectorial representation

The vectorial representation used above and in the remainder of this section was chosen from a set of several different vectorial representations of the primary/secondary structure of a given hairpin. The eight representations we considered differ on the amount of information they represent and thus on their ability to distinguish the structural characteristics of different stem-loops.

The first representation is called TripletS. This representation consists of a vector of normalised counts. To build this representation, a sliding window of length 3 is passed through the structure. At each step, a count position in the vector is incremented. The appropriate position in the vector is mapped considering whether each nucleotide in the window is paired or unpaired in the MFE structure and which base is present at the midpoint of the window. In the end, the counts on each position of the vector are divided by the length of the structure. The vector has thus 32 positions.

The second representation is called TripletB and is built in a way similar to that of TripletS, except that it distinguishes whether the paired nucleotide is in the $5'$ or $3'$ stem arm, i.e., whether it is the left/right-hand side of a base-pair. In this case, the vector has 108 positions.

The third representation is called TripletL and it extends TripletB by distinguishing nucleotides at the terminal loop from other unpaired nucleotides. This mapping yields a vector with 256 positions.

Three additional representations called QuintupletS, QuintupletB, and QuintupletL are calculated in a way similar to those previously discussed except that they scan the

structural information of five consecutive positions yielding vectors with 128, 972, and 4096 positions respectively.

Finally, three representations termed StructureS, StructureB, and StructureL are also similar to the first three representations but the identity of the nucleotide at the midpoint is not considered. These representations, therefore, only include structural information and give rise to vectors with 8, 27, and 64 positions, respectively.

These vectorial representations allow us to capture different types of information about the sequence/structure of our candidates and position them across a hyperplane on a multidimensional space. In order to use the Euclidian distance consistently as a measure of similarity between these vectorial representations it is preferable to represent our candidates using a set of independent and scaled dimensions. A straightforward way to guarantee these conditions is to perform a Principal Components Analysis (PCA) as described in section 4.3.

To determine which of these representations better reflects the results of conventional structural clustering we take the structural clusters obtained using `LocARNA` for 100 samples of 1000 randomly chosen stem-loops from each of the datasets (*D. melanogaster* and *A. gambiae*). As described before, the optimal partition into clusters is done by performing a node evaluation rule for various significance levels (k-levels) where low values for k produce clusters of highly similar structures and increasing values of k allow for increasingly heterogeneous clusters.

After calculating the centroid of each `LocARNA` cluster on the principal components space we can then calculate the rate of correct assignments. We repeat the procedure on a randomised version of our samples in order to assess the statistical significance of our results against a random background where the identity of each precursor is shuffled, thereby randomising the position of a precursor on the principal components space, but preserving the `LocARNA` cluster it belongs to. The statistical significance of the results is determined by comparing the results obtained for the regular and randomised samples using Welch's two-sample t-test. In each case, the normality of both sets of results (regular and randomised) is checked using the Kolmogorov-Smirnov test for a Gaussian distribution with the same mean and variance of each sample.

In order to compare the results for all the considered vectorial representations across the k-levels ranging from 0.0 to 0.9 we take the symmetric of the logarithm of the p-values of our statistical test. The larger this value the more significant are the results. Tables 5.4 and 5.5 show these values for the *A. gambiae* and *D. melanogaster* datasets, respectively.

For low values of k (up to 0.4), in both datasets, TripletL obtains the best results, which means that, for mostly homogeneous clusters, this vectorial representation outperforms all others. If we allow for more heterogeneous clusters (larger values of k), other vectorial representations take the lead but in an inconsistent way, since we obtain different results on both datasets or for different values of k.

	k									
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TripletS	207.9	204.9	191.4	165.4	147.3	108.1	83.9	63.4	56.2	52.5
TripletB	198.2	199.8	191.3	174.1	147.4	125.8	95.7	71.3	58.1	49.7
TripletL	251.9	242.3	224.0	197.0	180.0	157.2	117.8	83.5	65.9	60.7
QuintupletS	140.0	149.2	146.4	134.6	116.5	89.4	55.8	43.0	37.7	32.5
QuintupletB	139.2	140.2	129.9	120.9	113.4	85.9	54.0	31.4	26.8	22.3
QuintupletL	105.3	105.7	97.8	94.0	91.4	80.0	58.8	38.7	34.0	30.3
StructureS	14.3	12.8	12.8	12.4	12.7	10.0	8.4	7.9	7.8	8.3
StructureB	79.5	74.8	72.5	65.1	58.7	44.7	30.0	27.9	27.7	27.7
StructureL	213.4	212.5	204.3	184.7	150.2	121.3	81.3	74.6	73.9	73.3

Table 5.4: Table showing the -log(p-values) for the statistical significance of the correct assignment rate across all considered vectorial representations and k-levels for the *A. gambiae* dataset

	k									
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TripletS	266.1	256.7	244.4	223.1	201.3	182.7	148.2	131.1	114.2	96.3
TripletB	286.6	275.7	258.7	247.5	220.6	193.1	140.4	123.8	99.1	83.8
TripletL	322.1	302.3	279.5	253.9	224.6	184.9	145.7	125.9	101.0	89.2
QuintupletS	229.7	222.2	217.3	204.4	174.5	139.1	113.2	103.2	90.3	78.8
QuintupletB	212.4	204.0	195.4	179.0	162.3	129.6	95.8	72.2	56.1	50.1
QuintupletL	149.8	154.6	148.4	142.3	122.8	101.3	81.5	60.3	41.5	36.7
StructureS	33.1	28.2	26.0	22.8	24.1	22.6	17.6	17.0	15.7	14.0
StructureB	113.3	107.7	102.4	97.8	93.6	79.8	53.6	43.1	38.4	32.1
StructureL	233.9	220.4	211.9	206.4	187.2	162.0	114.0	99.5	95.8	85.4

Table 5.5: Table showing the -log(p-values) for the statistical significance of the correct assignment rate across all considered vectorial representations and k-levels for the *D. melanogaster* dataset

For the *D. melanogaster* dataset, the best vectorial representation for $k = 0.5$ becomes TripletB and then TripletS for $k > 0.5$, whereas for the *A. gambiae* dataset, the best vectorial representation changes for $k > 0.7$ to StructureL. In both cases, the transition is to a vectorial representation encoding less information about the hairpins (either structural information in the case of *D. melanogaster* or sequence information for *A. gambiae*), which is consistent with clusters grouping increasingly heterogeneous hairpins.

The TripletL representation emerges as the best choice, since it exhibits the best results for the greater range of k levels and, even though it is outperformed by other representations for the larger values of k, it maintains a very good relative performance.

It is interesting to note that all representations including quintuplets, although encoding more structural information, fail to yield top performances. This might be explained, in part, by the very large number of dimensions and also by the sparsity of the information across the vectors (where most positions will have zeroes) and the implications it has on the principal components analysis. On the representations that exclude sequence information, all except the one distinguishing left/right-hand pairings and stem arm/terminal loop unpaired positions have relatively poor performances, which underlines the importance of including sequence information.

5.4.3 Distribution of known pre-miRNAs in the feature space

Using the same samples from the datasets presented above, we can observe that despite the fact that not all known precursors are grouped together in the same cluster by `LocARNA` at any k-level (data not shown), they are however significantly close and restricted to a limited portion of the feature space. In fact, if we take the centroid of the known precursors and calculate the median distance of each known precursor to the centroid we obtain a value which is much smaller than what would be expected by chance (p-value $= 8.00e-60$ for *A. gambiae*, p-value $= 8.94e-71$ for *D. melanogaster*). Figures 5.7 and 5.8 illustrate this observation in the the full datasets by showing that, in the three-dimensional space defined by the first three principal components, known precursors populate the same limited region.

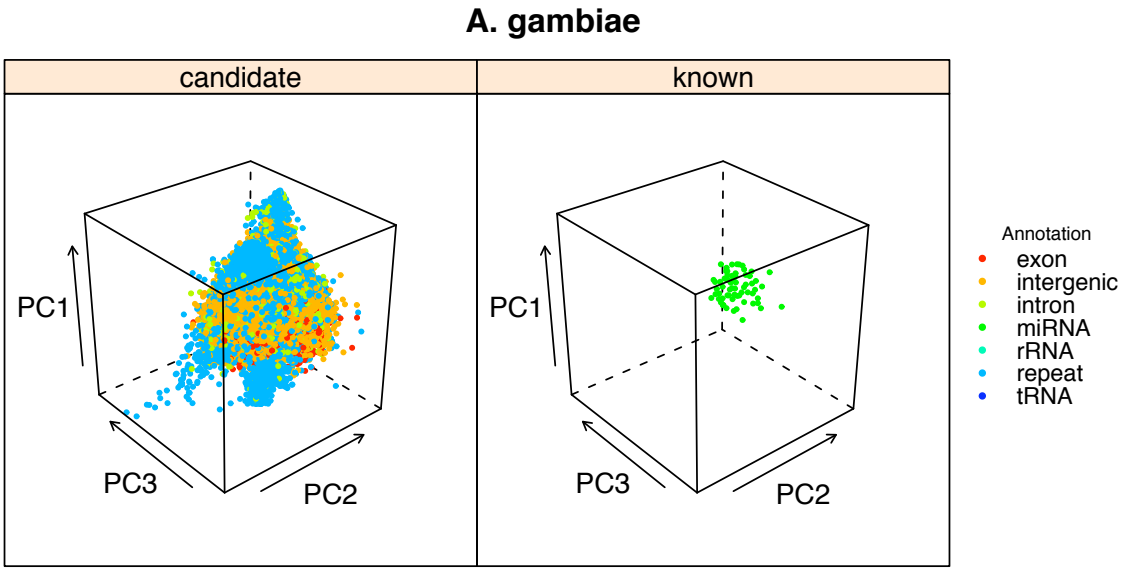

Figure 5.7: The spatial distribution of candidates and known precursors across the three-dimensional space defined by the first three principal components of the vectorial representation of the hairpins of *A. gambiae*

5.4.4 Identification of candidates structurally similar to known precursors

The results shown above suggest that known precursors tend to concentrate on a particular region of the feature space. That region, however, is also densely populated by other precursor candidates. Since the region where known precursors occur is inserted in an area of great density it cannot be identified by a purely unsupervised approach. As we stated before, this region can either be thought of as a single connected component or several disconnected components. If we take the first possibility we can take the centroid of all known precursor

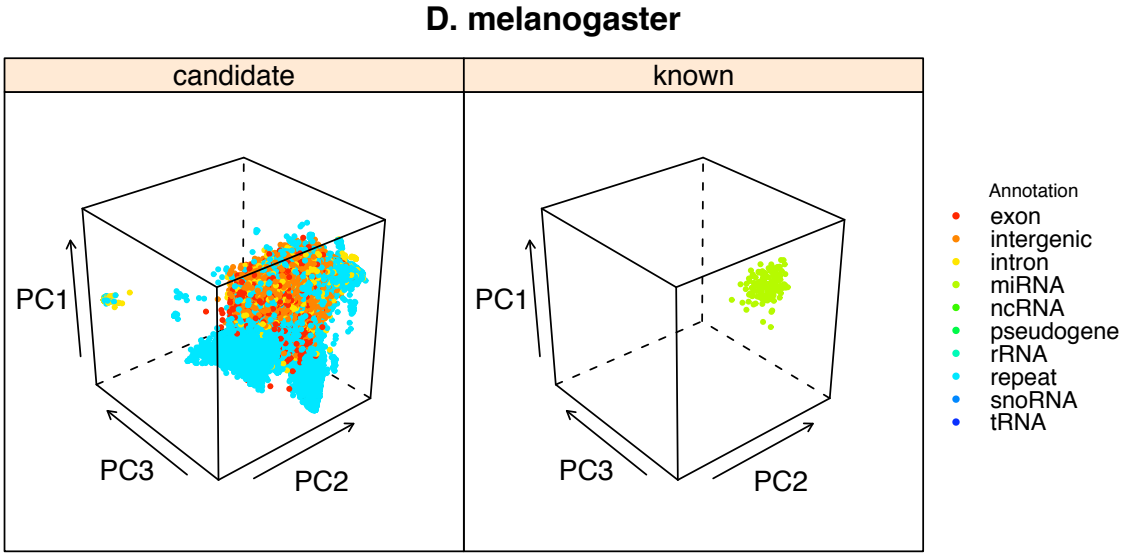

Figure 5.8: The spatial distribution of candidates and known precursors across the three-dimensional space defined by the first three principal components of the vectorial representation of the hairpins of *D. melanogaster*

co-ordinates and define a sphere around it with a varying radius. Then, all the structures inside the sphere are deemed part of the acceptance region.

The trade-off between considering a large radius thus including the greatest possible number of known precursors and restricting the size of the sphere not to include too many candidates is better represented using a ROC curve. Figures 5.9 and 5.10 show the ROC curves for *A. gambiae* and *D. melanogaster*, respectively. The figures also show the true/false positive rates for the optimal radius (calculated using the Youden index in order to identify the optimal cut-off).

Since these results depend on the proper identification of the precursor-containing region and that a reduced number of known precursors might hinder the calculation of an adequate centroid, we studied the impact of using each known precursor as the centre of the sphere instead of the centroid. These results are also shown in Figures 5.9 and 5.10. Unsurprisingly, for those precursors which are farther from the centroid we obtain a poor performance. However, for most other precursors the results tend to approach those obtained using the centroid and, in fact, some precursors outperform the centroid as the centre of the candidate selecting sphere.

The results of this approach are relatively poor in light of the fact that we are evaluating not its predictive power, but merely its ability to recover the same precursors which were used to determine the centre of the sphere where interesting candidates presumably can be found.

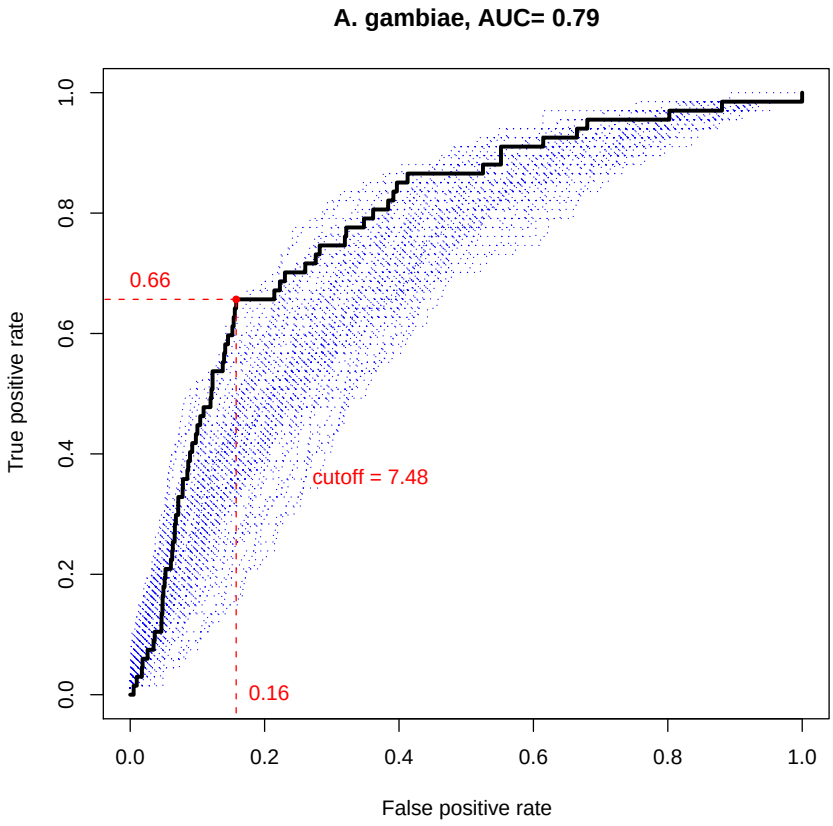

Figure 5.9: ROC curve for the distance to the centroid of known precursors for the *A. gambiae* dataset (solid line). ROC curves for the distance to each known precursor (dashed lines).

The performance of this approach, which presupposes the existence of a single connected acceptance region, justifies considering the competing possibility that the acceptance region might, instead, be better represented by several disconnected components. In this new model for the acceptance region, we take the co-ordinates of each known precursor and use them to identify the closest candidates. This method, which we here refer to as MinDist, has the advantage of allowing for different pre-miRNA structural classes to emerge around subsets of known precursors. The number of candidates that are included in the acceptance region is controlled by the maximum distance allowed to the closest pre-miRNA.

The larger the permitted distance, the greater the chance of selecting a region that includes all interesting candidates, but at the expense of enlarging the number of false positives. Again, we use the Youden index to determine the best trade-off and identify the optimal cut-off for this distance. However, in this approach, in order to estimate the optimal cut-off value, we have to consider subsets of known precursors as reference and to calculate the true/false positive rate with respect to the remaining known precursors and other candidates. Figures 5.11

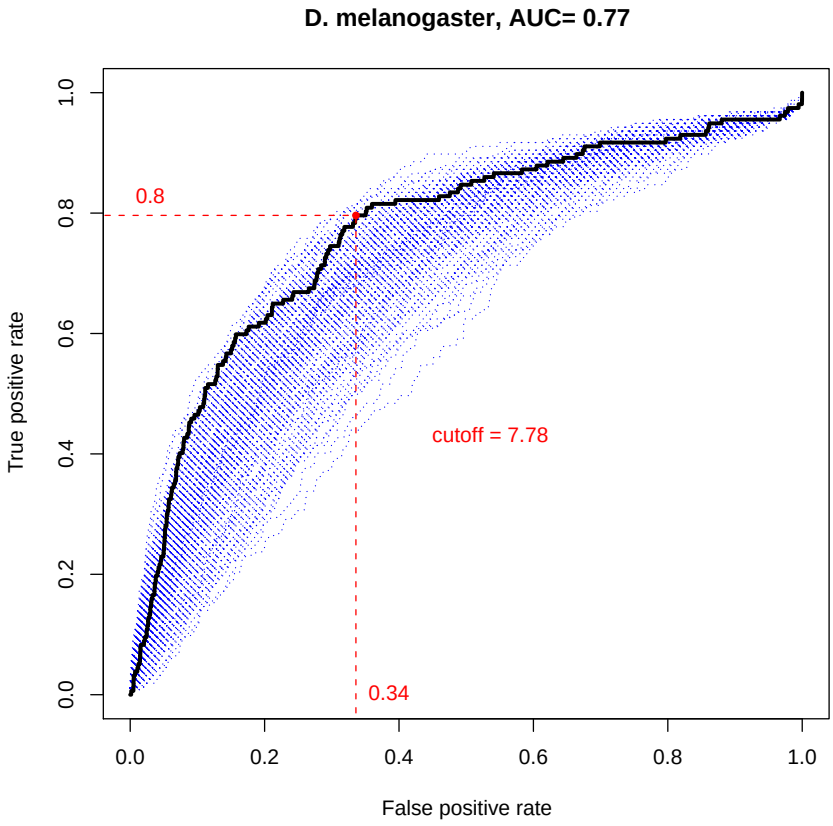

Figure 5.10: ROC curve for the distance to the centroid of known precursors for the *D. melanogaster* dataset (solid line). ROC curves for the distance to each known precursor (dashed lines)

and 5.12 show the ROC curves for *A. gambiae* and *D. melanogaster*, respectively, when using samples of 5%, 10%, 20%, and 50% of known precursors as reference and computing the trade-off between the true/false positive rates with respect to the remaining precursors and an equal number of sampled candidates. The figures show the ROC curves of 1000 such samples as well as the average curve, computed as the average performance over all samples across the full range of cut-off values. Additionally, the figures also show the average performance of our method, computed as the average TPR and FPR across all samples for the optimal cut-off on each sample (note that this may be significantly different from the optimal cut-off calculated on the average ROC curve).

The optimal cut-off in each of these ROC curves can be interpreted as the best choice of maximum distance allowed between a structure and the closest precursor so that the former may be included in the acceptance region. We have observed that there is a log-linear relation between the value of the average optimal cut-off and the proportion of known precursors that

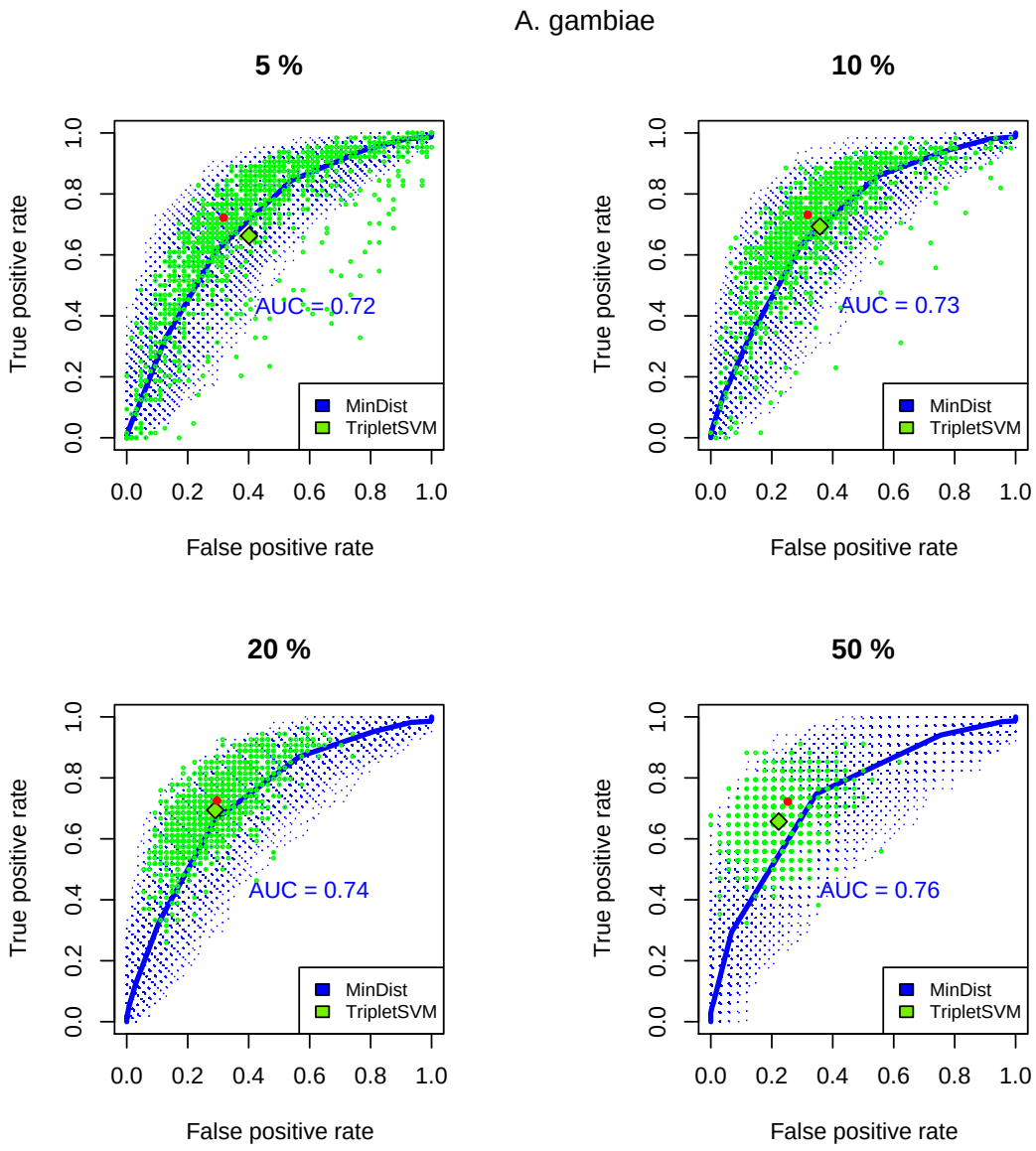

Figure 5.11: ROC curves for the minimum distance (MinDist) to pre-miRNAs method and the performance of TripletSVM over 4000 samples equally divided into 4 groups.
Each group uses 5%, 10%, 20% or 50% of the known precursors of *A. gambiae* to set up the positive examples of the training set. The positive examples of the testing set are made up by the remaining precursors. In both sets, the negative examples are samples of the set of candidates. ROC curves for each individual sample are shown in dashed lines and the average curve across the range of cutoff values is shown in a solid line. The red dot represents the average performance of the MinDist method over all samples considering the optimal cutoff for each sample. The green dots represent the performance of TripletSVM on each sample, whereas the green diamond refers to its average performance.

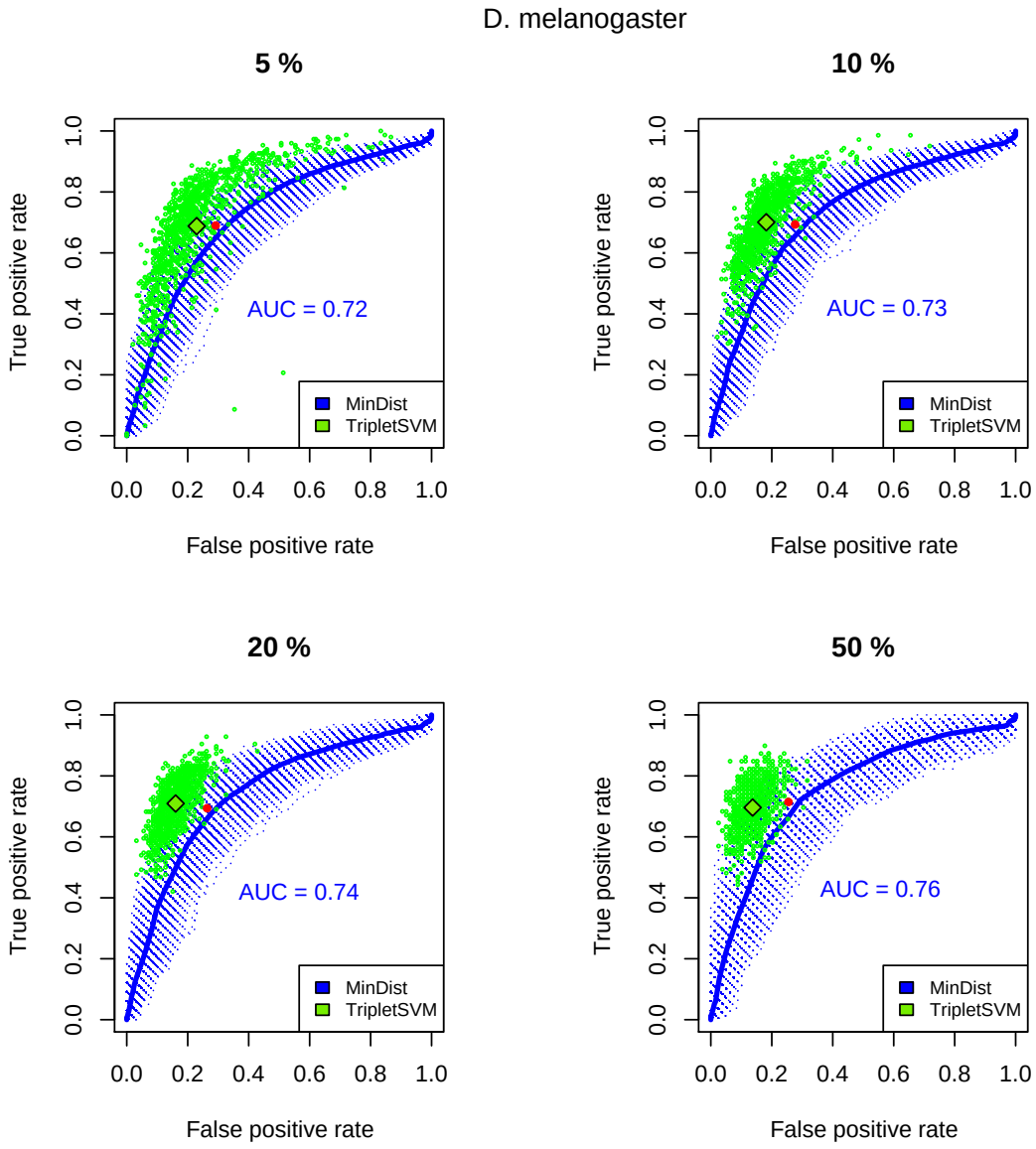

Figure 5.12: ROC curves for the minimum distance to pre-miRNAs method and the performance of TripletSVM over 4000 samples equally divided into 4 groups.
Each group uses 5%, 10%, 20% or 50% of the known precursors of *D. melanogaster* to set up the positive examples of the training set. The positive examples of the testing set are made up by the remaining precursors. In both sets, the negative examples are samples of the set of candidates. ROC curves for each individual sample are shown in dashed lines and the average curve across the range of cutoff values is shown in a solid line. The red dot represents the average performance of the MinDist method over all samples considering the optimal cutoff for each sample. The green dots represent the performance of TripletSVM on each sample, whereas the green diamond refers to its average performance.

is used as reference ($R^2 = 0.998$, for *A. gambiae*, and $R^2 = 0.989$, for *D. melanogaster*). Since the best choice of cut-off cannot be directly determined for the entire set of known precursors, we estimate it by extrapolating the log-linear model. The estimated optimal cut-off can be interpreted as the best choice of maximum distance to include additional (yet unknown) precursors with the least number of false positives.

Using the estimated optimal cut-offs, the acceptance regions include 23.5% (77 366) and 23.5% (67 619) of all candidates from the *A. gambiae* and *D. melanogaster* datasets, respectively.

5.4.5 Performance comparison to other methods

As we have seen before, TripletSVM [132] is a classifier based on a support vector machine that purports to determine whether a given stem-loop is a pre-miRNA. The features considered by this support vector machine are quite similar to those of the TripletS vectorial representation that is described above. It is also, to our knowledge, the only single-genome method whose source code is made available and which includes the necessary routines to re-train the model with new data. TripletSVM was trained using positive examples from samples of known precursors and negative examples from samples from the candidate set. Four groups of samples of different sizes were prepared for each dataset. Each sample group was divided in training sets and testing sets both with the same number of positive and negative examples. Each sample group is made of 1000 samples. The positive examples in the training set of each sample are a random subset of the known pre-miRNAs (either 5%, 10%, 20% or 50% of all known precursors) and the remaining pre-miRNAs make up the positive examples of the corresponding testing set. The negative examples in both the training and testing sets of each sample are random subsets of the candidates of the same size of the corresponding positive examples. Our method uses only the positive examples in the training set as a reference from which to compute the distance to the elements in the testing set, whereas TripletSVM, for each sample, is trained using both the positive and negative examples of the training set and is evaluated against the testing set. A graphical representation of the performance of TripletSVM in each of the 4000 samples (evenly distributed between training sets using 5%, 10%, 20%, and 50% of the annotated pre-miRNAs), as well as the average performance in each group of samples, is shown in Figures 5.9 and 5.10. Table 5.6 shows the sensitivity and specificity of TripletSVM and MinDist across training sets including a greater range of varying proportions (from 5 to 95%) of known precursors.

In general, the average performance of TripletSVM is comparable to the average performance of our method despite a tendency for having comparatively lower sensitivity and higher specificity. However, two clear shortcomings are apparent. TripletSVM is markedly less robust for small training sets than our method, which is manifest from the comparatively poor performance on the 5% sample group in the *A. gambiae* dataset, and by the fact that the

% known	*A. gambiae*				*D. melanogaster*			
	MinDist		TripletSVM		MinDist		TripletSVM	
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
5 %	0.72	0.68	0.66	0.60	0.69	0.71	0.69	0.77
10 %	0.73	0.68	0.69	0.64	0.69	0.72	0.70	0.82
20 %	0.73	0.68	0.69	0.71	0.69	0.74	0.71	0.84
50 %	0.72	0.75	0.66	0.78	0.71	0.75	0.70	0.86
80 %	0.74	0.80	0.65	0.80	0.72	0.77	0.69	0.88
90 %	0.78	0.83	0.65	0.88	0.73	0.81	0.68	0.88
95 %	0.75	0.89	0.66	0.81	0.75	0.83	0.68	0.88

Table 5.6: Sensitivity (TPR) and Specificity (1 - FPR) of TripletSVM and MinDist computed as the average performance across all samples for training sets whose positive examples consist of a fraction of known pre-miRNAs in *A. gambiae* and *D. melanogaster*

smaller the training sets (in all datasets) the more variable the performance in each individual sample. This variability is particularly visible in both 5% sample groups. TripletSVM also bears the inconvenience of requiring negative examples which are inevitably chosen under the assumption – however plausible and defensible – that miRNA precursors are rare with respect to the overall number of candidates, but one cannot generally guarantee that hairpins which would normally be processed by the miRNA-maturation pathway are not being included in the negative training set. Our approach, despite assuming all candidates to be false positives for the purposes of the performance evaluation, does not use this information to shape the acceptance region.

The average performance of our method in the *D. melanogaster* dataset is slightly worse than TripletSVM, possibly due to the inclusion of heterochromatic sequences in this dataset, which introduce greater variability in terms of sequence/structure, and as a consequence, the region where pre-miRNAs are found is more densely populated with candidates exhibiting more regular features.

Unlike TripletSVM, our approach allows us to control the number of candidates we wish to select by adjusting the cut-off level and therefore either privileging sensitivity or specificity. Additionally, since it reflects sequence/structure similarity in a way comparable to conventional structural clustering, our method offers a better interpretation of the decision rule that is made when selecting candidates.

5.5 Transcription potential

The number of candidates obtained after our structural analysis (77 366 for *A. gambiae*, and 67 619 for *D. melanogaster*), albeit considerably lower than the original candidate set (328 829 for *A. gambiae*, and 287 469 for *D. melanogaster*), is still quite numerous. A plausible interpretation of these results is that, despite their structural similarity to known precursors, the majority of these candidates are not pre-miRNAs due to other factors. Chiefly among these is the fact that most remaining candidates are probably not efficiently transcribed or

are playing different biological roles. This illustrates the often ignored distinction between having an adequate secondary structure and actually being transcribed and processed.

In this section, we present the results concerning the assessment of the transcriptional potential of the remaining candidates considering annotation data, potential genomic clusters and experimental data.

5.5.1 Identification of candidates with viable annotation and forming potential genomic clusters

A straightforward way to address the need to assess the transcriptional potential of the candidates is the observation that many fall within regions that have been annotated. Genomic locations with no annotation or which have been annotated as introns may contain miRNA precursors, but candidates which overlap regions annotated as exons, repeats or other non-coding RNAs are very unlikely to contain pre-miRNAs. The variety of annotated candidates (using the annotation data provided by EnsEMBL) present in our candidate sets can be seen in Figures 5.7 and 5.8. If we filter out non-viable candidates by this criterion, our candidate set is reduced to 44 210 for *A. gambiae* and 40 582 for *D. melanogaster*.

If we additionally restrict our search to putative miRNA cluster members, we can lower the number of candidates by considering only those which are found in the vicinity of known precursors. By selecting candidates with viable annotation and at a genomic distance not greater than 50 kb [9] from known precursors, we reduce our candidate set to 439 for *A. gambiae* and 1604 for *D. melanogaster*.

Using a single-genome approach it is difficult to further assess the transcriptional potential of a precursor candidate in the absence of experimental data.

5.5.2 Using experimental data as evidence of transcription

In this thesis, we use the small RNA library done for *A. gambiae* by Brunel and collaborators [129] against our reduced set of candidates (stable, robust, structurally close to known pre-miRNAs and with viable annotation).

All the sequenced small RNAs (length between 16 and 27 nucleotides) were mapped to the genome using an approximate matching procedure allowing up to 2 errors. The choice of a tolerance of 2 mismatches accounts for the possibility of both sequencing errors and RNA editing events.

Out of a total of 268 unique sequences, 253 had approximate matches in the genome of *A. gambiae* totalling 29 470 hits. One sequence was responsible for 20 731 hits occurring in 973 variants (with up to two errors) but only 35 exact matches. Many of these matches were to regions annotated as repeats rendering the sequence unlikely to originate from a miRNA and justifying its exclusion from further consideration.

The great number of hits and the 2-error tolerance justifies the need of a statistical validation procedure in order to distinguish spurious hits from statistically significant matches. The statistical significance of the match is computed given the genomic context, the number of errors of the match and the number of hits of a given sequence in each genomic context against the corresponding expected number of observations (see section 4.4 for details).

The choice of the size of the genomic context to adopt is a non-trivial matter. We considered windows of sizes 50k, 200k, 500k, 1M, and 2M bases. Fig. 5.13 shows the pairwise comparison of the p-values obtained for the different window sizes.

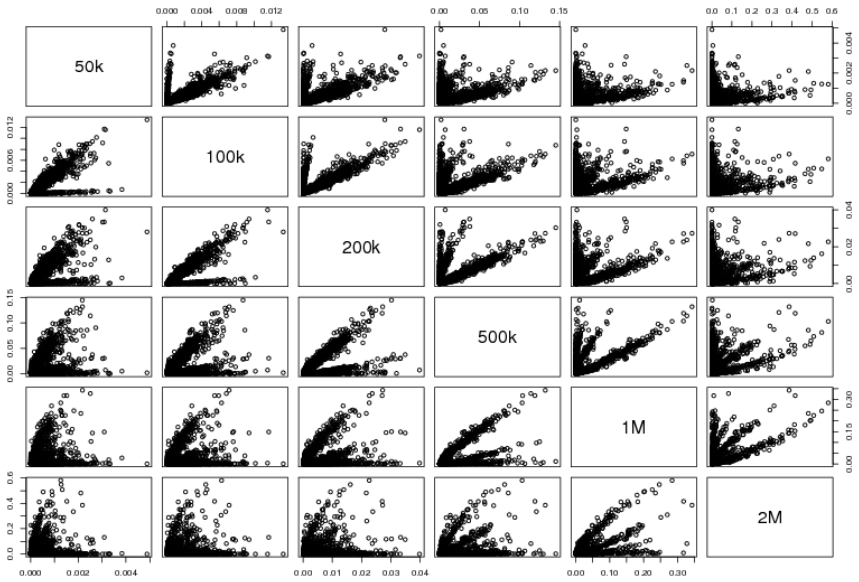

Figure 5.13: Pairwise comparison of the p-values obtained for genome occurrences considering genomic contexts of 50k, 100k, 200k, 1M, and 2M bases.

For a constant number of observed occurrences, the p-value of a hit should be proportional to the size of the genomic context, provided that the statistical characteristics of the context are stable. We do observe this general trend in Fig. 5.13, but we can also see that many hits deviate from this theoretical behavior by keeping or even lowering the corresponding p-values for increasing genomic context sizes. This can be due to either a non-constant number of observed occurrences or to sudden changes in the statistical characteristics of the genomic contexts under consideration.

Figs. 5.14 and 5.15 show the pairwise comparison of expected and actual number of occurrences, respectively, for several genomic context sizes. We can see that the number of observed occurrences varies greatly with different context sizes, but the expected number of occurrences varies almost linearly with context size amongst the larger contexts. We can therefore con-

clude that the behavior seen on Fig. 5.13, which departs from what is theoretically predicted, is largely due to non-homogeneous variation in the number of actual occurrences observed in the genomic context and not to variation in the local statistical characteristics, at least for larger contexts. In contexts of size above 500k base-pairs the subsequent evolution of the expected number of occurrences is almost perfectly linear which means that the local statistical characteristics of the genome do not vary much and therefore any context above this length could be used. We have decided to use the p-values of the 2M base-pairs contexts as the reference statistical assessment score because they exhibit a better resolution. To that effect it suffices to notice that the p-values for the 500k and 1M base-pairs contexts vary between 0-0.15 and 0-0.3, respectively, whereas the p-values for the 2M base-pairs contexts vary between 0 and 0.6.

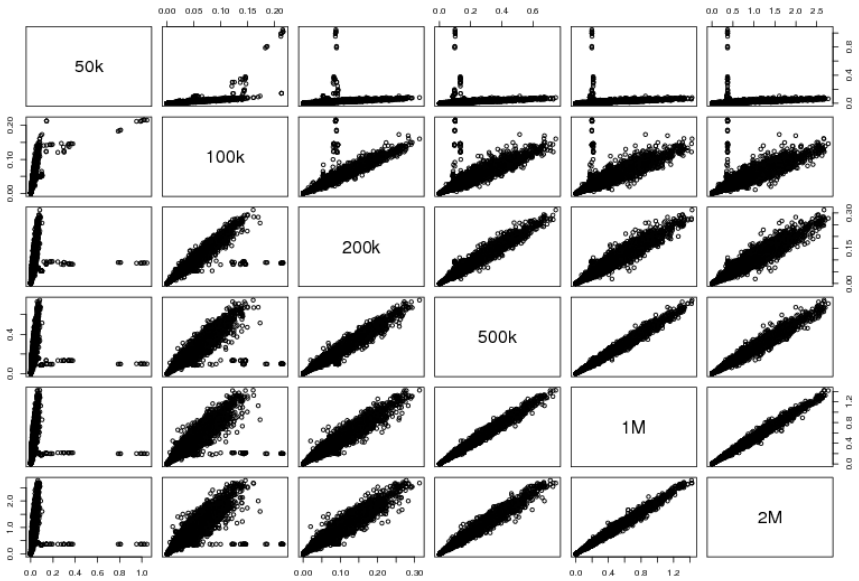

Figure 5.14: Pairwise comparison of the expected number of occurrences considering genomic contexts of 50k, 100k, 200k, 1M, and 2M bases.

A total of 75 candidates from the acceptance region in the feature space and with viable annotation contain hits within their start/stop co-ordinates, but only 54 have statistically significant hits and, amongst these, only 46 have hits on their stem arms, rather than on the terminal loop. All these hits are 2-error matches against sequenced small RNAs except for one which is a single-error match. In only 2 cases is the origin of the sequenced small RNA not better explained by hits elsewhere in the genome, which are either exact matches or have fewer mismatches. Table 5.7 identifies the *A. gambiae* candidates with most plausible direct transcriptional evidence.

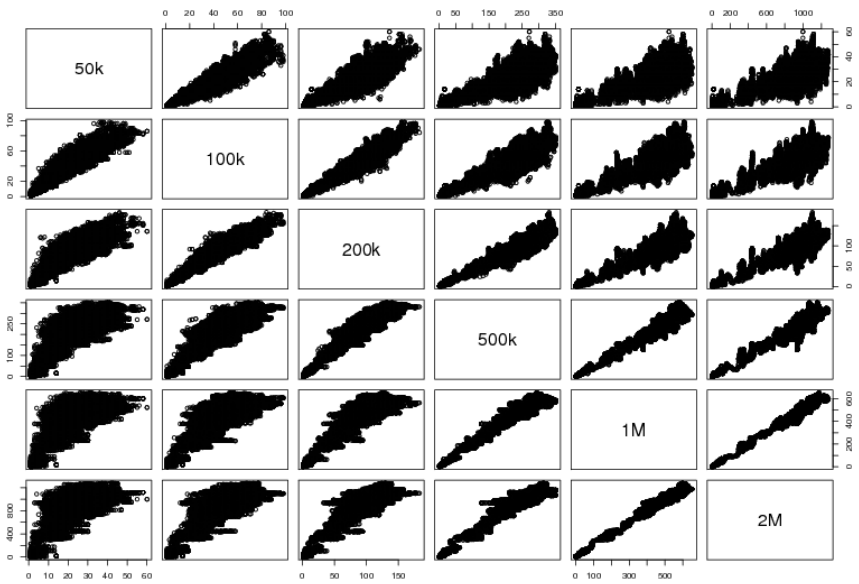

Figure 5.15: Pairwise comparison of the observed number of occurrences considering genomic contexts of 50k, 100k, 200k, 1M, and 2M bases.

Transcript sequence	Hit sequence	Errors	Chr	Strand	Start:Stop
GTGGTGCTCCCTCGACC	CTGCTGCTCCCTCGACC	2	2L	+	10481374 : 10481468
CCAGTCGGTAGCGCTTA	CCAGTCGTTAGCGCTTA	1	2L	+	36469979 : 36470097

Table 5.7: *A. gambiae* precursor candidates with better transcriptional evidence. The transcript sequence refers to the sequence present in the small RNA library and the hit sequence is the actual sequence in the genome.

Part IV

Conclusions and Perspectives

Chapter 6

Conclusions and Perspectives

6.1 Conclusions

Common computational strategies to address the problem of miRNA gene discovery usually involve the identification of a set of candidates that are subsequently filtered by their similarity to previously known pre-miRNAs and their degree of conservation in evolutionarily close species. Although these approaches have vastly expanded the list of known miRNAs, relying merely on our general knowledge of the miRNA silencing pathway and arguments of similarity and conservation, they generally fail to integrate and use a growing amount of knowledge about these regulators. More significantly, these methods have introduced a strong bias favouring pre-miRNAs exhibiting extensive conservation and sharing a number of features with previously identified precursors. To varying degrees, their ability to discover non-conserved or non-canonical pre-miRNAs is therefore greatly reduced if not completely suppressed. This is not to say that one expects pre-miRNA features to vary greatly, but rather that the bias may not reveal the adequate learning set. The fact that most computational approaches to miRNA gene finding make extensive use of evolutionary conservation illustrates our collective ignorance about the subtle rules presiding miRNA biogenesis. Since the cell cannot use the filter of evolutionary conservation [8] to choose amongst all potential stem-loops, we seem to be missing a significant part of the whole story. Additionally, if we consider that many authors argue that the identification of well-conserved and phylogenetically extensive miRNAs is reaching its saturation point, it is important to assess whether non-conserved, presumably more exotic, miRNA precursors would be processed as such in different organisms that may have small yet important differences in their miRNA-processing pathways. The elucidation of this question is crucial to methods which try to generalise from pre-miRNAs taken from several different species. Our efforts should therefore focus on the need to establish more accurate models of candidate selection.

The approach developed in the context of this thesis purports to be able to draw from heterogeneous sources of data and to offer a single-genome evidence-based tool to estimate

the likelihood that a given set of precursor candidates are indeed pre-miRNAs.

The search for the distinguishing characteristics of animal miRNA precursors is the central problem tackled by this thesis. We began by observing that it had been already established that pre-miRNAs have particularly low-energy structures compared to other RNAs, even when corrected for size and GC content [16, 140]. Additionally, these stem-loops seemed to be robust with respect to their genomic contexts, as should be expected for efficient Drosha recognition. They were also stable in the sense that a similar base-pairing pattern persisted for a set of sub-optimal structures making up the thermodynamic ensemble where the stem-loop should be found most of the time. Mutational robustness was another property that had been suggested for miRNA precursors [18], but it was most likely observed in ancient well-conserved pre-miRNAs rather than more recent non-conserved precursor hairpins

We show that, by combining variations of these different robustness and stability measures, it is possible to obtain an unsupervised scoring scheme that outperforms any individual measure, and performs at least as well as the best of two supervised approaches. The results for our combination of measures show that there is a strong bias towards robust miRNA precursors and that this information can be used to reduce the vast number of stem-loops that are found in metazoan genomes. Nonetheless, given the number of precursor candidates that remain after sieving through those which were identified in the three datasets we have presented, it is not yet possible to claim that we are in the presence of a miRNA gene finding method. We have, however, reduced the number of candidates by one order of magnitude, without resorting to classic approaches and retaining most known pre-miRNAs in each dataset. The main advantage of the adopted score is that it relies on intelligible criteria based on arguments of biological plausibility. Over this reduced set of candidates, it was then possible to perform a structural analysis of the set of candidate hairpins.

In the context of the structural analysis of our candidate set, we have presented an approach to evaluating the sequence/structure similarity of a very large number of structures with an application to the identification of a reduced set of pre-miRNA candidates. In a first step, we have demonstrated that our vectorial representation of RNA structures and the Euclidian distance in the feature space consequently defined was comparable to the sequence/structure similarities identified by `LocARNA`– a conventional and efficient structural clustering method. In a second step, we have observed that known pre-miRNAs tend to populate a specific region of the feature space defined by our vectorial representation and we used these known precursors to identify that region and to select candidates populating it. We have shown that as little as 5% of known precursors could be used to identify such a region and recover a substantial proportion of the remaining pre-miRNAs. The fact that this region is very dense in terms of the number of precursor candidates it contains tells us that a large number of genome loci have the potential to generate stable and robust structures which present sequence/structure similarities to known pre-miRNAs. The use of annotation

information helps to reduce the number of selected candidates but after this filtering step they remain in the tens of thousands. Therefore, there is either an exceedingly large number of pre-miRNAs in these datasets or, more plausibly, most of these candidates are not efficiently transcribed but could otherwise be recognised as miRNA precursors.

The portion of the feature space selected by our structural analysis approach purports to identify the region that includes all the structures which have the potential of being efficiently recognised by the cellular miRNA-processing machinery. As we said, this does not mean that all these structures are pre-miRNAs, but rather that they present a strong sequence/structure similarity to known precursors. Unlike many machine learning approaches to the identification of miRNA precursors that use features of the sequence and secondary structure to provide a classifier, our approach does not need to postulate a set of negative examples. In fact, we contend that if the purpose is to characterise the structures which have the potential of being recognised by the enzymes involved in miRNA maturation one needs to reduce one's dependence on the positive set as well, since it will most likely not be representative. To this effect, it suffices to observe that the set of recognisable structures is surely larger than the set of all the pre-miRNAs contained in a genome and that these two sets are subject to different evolutionary constraints. In our work, information about known precursors is used merely to pinpoint a region of interest in our multidimensional representation of sequence/structure features, rather than to learn the characteristics that distinguish pre-miRNAs from other stem-loops and, in this sense, it is not a machine learning method.

It is therefore clear that there is a distinction between having a secondary structure amenable to recognition and actually being a pre-miRNA. In practical terms, as we have seen, the distinction is most probably due to whether a given structure is transcribed. To this effect, it is important to have transcriptional data with which to determine whether a candidate could actually be a miRNA precursor. Thus, our third and final contribution is a set of strategies to incorporate an estimation of the transcription potential of each candidate since our results warrant the conclusion that animal genomes contain a large amount of potential hairpins that might enter the miRNA maturation pathway if only they were efficiently transcribed. Our results here are, however, limited by the availability and quality of the data that serve as the basis to assess the transcriptional potential. We have, nevertheless, provided a method to use a library of small RNAs and we have indicated two novel miRNA predictions for *A. gambiae* based on that information. We have also shown that, in the absence of transcriptional data, one can limit the candidates to those occurring in the genomic vicinity of known precursors and which could, therefore, be part of a miRNA cluster.

6.2 Future improvements

As with most research projects, we can identify several instances in which further improvements can be introduced. In particular, the procedure used to combine the different robustness and stability measures for each dataset relies on the measure distributions calculated for stem-loops extracted from artificial sequences generated with the same dinucleotide distributions observed in the original genomes. As suggested by the relatively poorer results obtained for the *D. melanogaster* dataset, which includes both euchromatic and heterochromatic sequences, genome heterogeneity may warrant the generation of multiple artificial sequences with their respective measure distributions in order to obtain a scoring scheme able to cope with variations in sequence features for different regions of the genome. This presupposes the identification of isochores, i.e., broad regions of the genome with highly uniform GC content. These measures, but also the structural analysis results, could also benefit from breaking down the analysis of the candidate set into candidates located in intergenic regions, introns, repetitive regions or regions containing protein-coding genes in the opposite strand, since candidate precursors arising from these diverse locations will likely be subject to different evolutionary constraints and may therefore exhibit distinct stability, robustness and structural characteristics.

In addition, the CRAVELA framework, in particular its web interface, can be expanded to allow for a more complete exploration of the data, plugging in the results of the structural analysis and permitting the direct visualisation of potential miRNA genomic clusters identified using our pipeline, as well as the information provided by the assessment of the transcriptional potential of the candidate precursors.

6.3 Perspectives

Understanding the complexity of gene regulation, we now know, cannot be achieved without integrating regulation at the post-transcriptional level, particularly miRNA-mediated regulation. Identifying miRNA genes and their targets is a crucial task in determining the biological role of these regulators and effectively modelling the expression of their target genes.

Future developments in the field of miRNA gene finding can probably benefit from the insight provided by considering what is not a miRNA. Thus, a given locus in the genome does not contain a viable miRNA if 1) it is not efficiently transcribed; 2) it does not contain a stem-loop structure amenable to efficient processing by all participants in the miRNA biogenesis pathway; 3) it cannot regulate a target gene in a physiologically relevant manner. Most available methods for miRNA gene finding have focused on the second aspect and little attention has been paid to the other two.

In the context of determining whether a locus is transcribed in an efficient manner, transcription data are an invaluable resource, although the treatment of this information, as we

have discussed earlier, poses its own challenges. A complementary approach to this problem is to seek the identification of the primary miRNA (pri-miRNA) and its promoter region (for miRNAs arising from their own transcription units), which could also contribute to understanding the transcriptional control behind the expression regulation of miRNA genes themselves.

The discovery of whether a potential miRNA has target genes is dependent on the breakthroughs achieved in this area. The area of target prediction has received a new impetus with the recent proposal of a thermodynamic model incorporating target accessibility. However, seed matches constitute an important sieve to control false positives. The seed hypothesis, adopted almost unanimously by current target prediction methods, was recently reinforced with a study that obtained the structure of an important component of the silencing complex bound to a DNA guide-strand, and which lays down the biochemical basis for the role of seed sites [123]. However, at least some experimentally confirmed targets seem to violate the seed rule by including mismatches or G:U pairs [71, 119]. The present scarcity of confidently validated miRNA targets, establishing not only miRNA-target associations, but specifically pinpointing the hybridisation sites, is the greatest obstacle not only to the development of better prediction methods but also to the systematic assessment of the performance of current tools. The increasing availability of degradome data could prove an essential contribution to this effort, although it is generally limited to the identification of cleavage-sites in potential targets, being unable to detect instances of cleavage-independent miRNA-mediated regulation.

Target prediction becomes even more challenging with the discovery that RNA editing is common in miRNAs [79, 13, 92]. This could substantially change the mature sequence and, consequently, the specificity of its targets. Moreover, a study conducted on human miRNA targets [101] shows that miRNAs tend to target genes with distinctively AT-rich 3' UTR regions, even when these genes are located in GC-rich isochores, suggesting an unknown function for this compositional bias. The authors argue that better knowledge of the background distribution of nucleotides in 3' UTR regions may lead to improvements in miRNA target predictions. Additionally, the study of preferred versus avoided motifs in accessible regions with potential target sites could help elucidate the various evolutionary forces behind miRNA target evolution.

The identification of miRNA-mediated regulatory models, constituted by a set of miRNAs jointly regulating a set of target genes, and its integration in gene regulatory networks in the ultimate goal of the research efforts developed in this area. The discovery and study of these regulatory modules could benefit from a better understanding of the evolutionary forces driving miRNA evolution, particularly the restrictions at the sequence, structure and genomic context levels, but also the influence of the evolution of the regulatory network itself.

Part V

Appendices

Appendix A

Pre-miRNA homologs from *A. gambiae* in *A. darlingi*

Table A.1: Homologs to pre-miRNAs of *A. gambiae* identified amongst the precursor candidates of *A. darlingi*

miRNA	Contig	Strand	Start	Stop	Length	E-value	Identity
aga-mir-281	ctg7180000455710	F	12304	12396	93	6e-44	98.92
aga-mir-137	ctg7180000423045	F	24973	25062	90	3e-42	98.89
aga-mir-125	ctg7180000436522	R	30069	30161	93	2e-41	97.85
aga-mir-9c	ctg7180000436657	F	27474	27563	90	8e-40	97.78
aga-mir-iab-4	ctg7180000409079	R	9295	9378	84	1e-38	98.81
aga-mir-278	ctg7180000393996	R	2139	2222	84	3e-36	98.81
aga-mir-8	ctg7180000296739	R	14269	14350	82	5e-35	97.56
aga-mir-957	ctg7180000502071	F	10129	10209	81	2e-34	97.53
aga-mir-1175	ctg7180000395096	R	20239	20316	78	1e-32	97.44
aga-mir-305	ctg7180000409513	R	69	156	88	5e-32	95.51
aga-mir-9a	ctg7180000394369	R	31986	32065	80	2e-31	97.50
aga-mir-79	ctg7180000436657	F	29361	29431	71	8e-31	98.59
aga-mir-263b	ctg7180000380439	F	18574	18666	93	8e-31	94.74
aga-mir-927	ctg7180000364658	R	2214	2301	88	1e-29	94.32
aga-mir-1891	ctg7180000436657	F	69525	69616	92	5e-29	92.39
aga-mir-1000	ctg7180000409240	F	26262	26333	72	5e-29	97.22
aga-mir-929	ctg7180000436895	R	35399	35472	74	7e-28	95.95
aga-mir-993	ctg7180000380779	R	2989	3093	105	1e-27	90.57

Continued on Next Page...

miRNA	Contig	Strand	Start	Stop	Length	E-value	Identity
aga-mir-307	ctg7180000501812	F	90845	90913	69	3e-27	97.10
aga-mir-7	ctg7180000456148	R	10858	10934	77	3e-27	94.81
aga-mir-283	ctg7180000394200	R	11363	11450	88	1e-26	92.05
aga-mir-14	ctg7180000394624	F	29822	29905	84	1e-26	94.05
aga-mir-210	ctg7180000325517	R	378	447	70	1e-25	95.71
aga-mir-92b	ctg7180000502202	R	52014	52091	78	2e-25	93.59
aga-mir-190	ctg7180000409440	F	26813	26893	81	8e-25	93.83
aga-mir-184	ctg7180000395192	F	12643	12726	84	2e-24	94.05
aga-bantam	ctg7180000380411	F	3920	4019	100	3e-24	93.07
aga-mir-263	ctg7180000422962	R	6187	6272	86	2e-22	95.35
aga-mir-277	ctg7180000436725	F	2857	2947	91	3e-21	90.11
aga-mir-124	ctg7180000394913	F	3659	3737	79	9e-21	90.36
aga-mir-10	ctg7180000299625	F	88710	88792	83	7e-19	89.16
aga-mir-13b	ctg7180000296969	F	18851	18926	76	3e-18	92.21
aga-mir-988	ctg7180000423020	R	24356	24424	69	3e-18	92.86
aga-mir-276	ctg7180000394910	R	4307	4390	84	1e-17	89.41
aga-mir-219	ctg7180000456051	R	64129	64209	81	4e-17	89.41
aga-mir-282	ctg7180000297228	F	31847	31926	80	6e-16	90.12
aga-mir-9b	ctg7180000436657	F	29840	29918	79	4e-14	87.21
aga-mir-1890	ctg7180000297175	R	19896	19966	71	4e-14	90.14
aga-mir-317	ctg7180000381136	F	4237	4319	83	2e-12	86.90
aga-mir-275	ctg7180000409513	R	6089	6155	67	2e-12	89.71
aga-mir-87	ctg7180000358126	R	296	383	88	1e-11	85.56
aga-mir-308	ctg7180000296848	R	3648	3718	71	1e-11	88.73
aga-mir-279	ctg7180000436869	R	24027	24088	62	2e-09	88.89
aga-mir-92a	ctg7180000502202	R	73930	73990	61	9e-09	88.52

Table A.2: Alignment of mature miRNAs from *A. gambiae* against precursor homologues identified amongst pre-miRNA candidates from *A. darlingi*

miRNA	Alignment	Identity
aga-mir-281	5′ AUCGAAUGUGAAAAUAAAGAGAGCUAUCCGUCGACAGUAGGGAUAUAAUUCACUGUCAUGGAAUUGCUCUCUUUAUGUACAAUUCGAUAUUCA 3′ 5′ UGUCAUGGAAUUGCUCUCUUUAU 3′	100.00
aga-mir-137	5′ AAAACUUGGUUGGCCACGCGUAUUCUUGGGUUACUAACACACUAUUUAUGUUGUUAUUGCUUGAGAAUACACGUAGUUGACUAGUGUUGU 3′ 5′ UAUUGCUUGAGAAUACACGUAG 3′	100.00
aga-mir-125	5′ GUAUCUGCUGAUUCCCUGAGACCCUAACUUGUGACUAUCGUUACAAAGUUUCACAAGUUUUGAUCUCCGGUAUUAGCGGUUGAGAUGCGACGG 3′ 5′ UCCCUGAGACCCUAACUUGUGA 3′	100.00
aga-mir-9c	5′ UUUCCGGCUGUGUCUUUGGUAUUCUAGCUGUAGAAUGUUGUUUUGAUUGUAAUAUCUCUAAAGCUUUAGUACCAGAGGUCCAACUGGGAA 3′ 5′ UCUUUGGUAUUCUAGCUGUAGA 3′	100.00
aga-mir-iab-4	5′ GUGCCGCUUCAUGAACGUAUACUGAAUGUAUCCUGAGUGCUACUUAUCCGGUAUACCUUCAGUAUACGUAACAGGAGGCGACAC 3′ 5′ ACGUAUACUGAAUGUAUCCUGA 3′	100.00
aga-mir-278	5′ GGUACGGUACGGACGGACGAUAGUCUUCAACGACCGUUCACGUUUGACACGAGGUCGGUGGGACUUUCGUCCGUUUGUAAGGCC 3′ 5′ UCGGUGGGACUUUCGUCCGUUU 3′	100.00
aga-mir-8	5′ GUCUGUUCACAUCUUACCGGGCAGCAUUAGAUAUUUUAUCGGAUACUUCUAAUACUGUCAGGUAAAGAUGUCGUCCGAGCCC 3′ 5′ UAAUACUGUCAGGUAAAGAUGUC 3′	100.00
aga-mir-957	5′ ACUGCGGGCGUUAGUUUUGGGCGGGUUUUAGUGUAUUUCGAUGAGAAUUCUAUUGAAACCGUCCAAAACUGAGGCCGGCAG 3′ 5′ UGAAACCGUCCAAAACUGAGGC 3′	100.00
aga-mir-1175	5′ GAUAUGGAAUAAGUGGAGUAGUGGUCUCAUCGCUUAGUUUUAGAAAAGUGAGAUUCUACUUCUCCGACUUAAUUCAUA 3′ 5′ UGAGAUUCUACUUCUCCGACUUAA 3′	100.00

Continued on Next Page...

miRNA	Alignment	Identity
aga-mir-305	$^{5'}$UUUGUCACAUGUCUAUUGUACUUCAUCAGGUGCUCUGGUGGAUUUCAGAAACCCGGCACAUGUUGGAGUACACUUAAUGUGCUGACAA$^{3'}$ $_{5'}$AUUGUACUUCAUCAGGUGCUCUG$_{3'}$	100.00
aga-mir-9a	$^{5'}$GUCAAUGUUCUCUUUGGUUAUCUAGCUGUAUGAGUGUAUUUAAAAACGUCAUAAAGCUAGCAUACCGAAGUUAAUAAUUG$^{3'}$ $_{5'}$UCUUUGGUUAUCUAGCUGUAUGA$_{3'}$	100.00
aga-mir-79	$^{5'}$GCUUUGGCGCUUUAGCUGUAUGAUAGAAUUUGAAGUAUUUCAUAAAGCUAGAUUACCAAAGCAUAGACGAA$^{3'}$ $_{5'}$UAAAGCUAGAUUACCAAAGCAU$_{3'}$	100.00
aga-mir-263b	$^{5'}$UGACCAAUUAUGGACCUUGGCACUGGGAGAAUUCACAGUGAUCGUACAUUAUCGUUCUGUGGAUCUUUUCGUGCCAUCGUUCAAAUUUGGUGC$^{3'}$ $_{5'}$CUUGGCACUGGGAGAAUUCAC$_{3'}$	100.00
aga-mir-927	$^{5'}$GUUAAUGGUUCGUUUUAGAAUUCCUACGCUUUACCCGUUAAAUUAAAGUAGUGCGGCAAAGCGUUUGGAUUCUGAAACGAAACAUUAA$^{3'}$ $_{5'}$UUUAGAAUUCCUACGCUUUACC$_{3'}$	100.00
aga-mir-1891	$^{5'}$UCUUUUUCUGUCAUGUUGAGGAGUUAAUUUGCGUGUUUUUGCAUACGAUUAAACACGUCCAUUAACUCUGGUACAUGAUGGAAAAACCGAGC$^{3'}$ $_{5'}$UGAGGAGUUAAUUUGCGUGUUUU$_{3'}$	100.00
aga-mir-1000	$^{5'}$GUCGAUGAUAUUGUCCUGUCACAGCAGUACUAUUUGCCUAGCUUACUGUUGUUUCGGGACAUUUCCAUCGAC$^{3'}$ $_{5'}$AUAUUGUCCUGUCACAGCAGU$_{3'}$	100.00
aga-mir-929	$^{5'}$UGGGAUUAAAUUGACUCUAGUAGGGAGUCCUUCCAUACGAGAGACUCCCUAACGGAGUCAGAUUGAUUCCGGUA$^{3'}$ $_{5'}$CUCCCUAACGGAGUCAGAUUG$_{3'}$	100.00
aga-mir-993	$^{5'}$...CGUGACCUACCCUGUAGUUCCGGGCUUUUGUGGGUUGAAAUACAAAAACAUGUAAAUUCAUAUUCUCUUAUCAGAAGCUCGUUUCUAUAGAGGUAUCUCA$^{3'}$ $_{5'}$GAAGCUCGUUUCUAUAGAGGUAUCU$_{3'}$	100.00
aga-mir-307	$^{5'}$UCUCUCGAUUACUCACUCAACCUGGGUGUGAUGCUUAUUUGAAUCAUCACAACCUCCUUGAGUGAGCGA$^{3'}$ $_{5'}$UCACAACCUCCUUGAGUGAG$_{3'}$	100.00
aga-mir-7	$^{5'}$UUGUAUGGAAGACUAGUGAUUUUGUUGUUUGGCUUAAGAUACUAACAAUAAAUCCCUUGUCUUUCUACAAAGAUUGC$^{3'}$ $_{5'}$UGGAAGACUAGUGAUUUUGUUGU$_{3'}$	100.00

Continued on Next Page...

miRNA	Alignment	Identity
aga-mir-283	5′UUCGACUGAAAGGUAAAUAUCAGCUGGUAAUUCUAGGCUAUCUAAACUUCGUGCACCCCGGAAUUUCAGCUGAUAUCCACUUUUCCGU3′ 5′UAAAUAUCAGCUGGUAAUUCU3′	100.00
aga-mir-14	5′GCCCGAUAAGCCUGUGGGAGCGAGAUUAAGGCUUGCUGGUUAUCAAUUUGAACUUUAGUCAGUCUUUUUCUCUCUCCUAUCGGU3′ 5′UCAGUCUUUUUCUCUCUCCUA3′	100.00
aga-mir-210	5′CAUUGCAGCUGCUGACCACUGCACAAGAUUAGAAUAUGACUCUUGUGCGUGUGACAACGGCUAUUGUGGG3′ 5′UUGUGCGUGUGACAACGGCUA3′	100.00
aga-mir-92b	5′GGGCUCCGGAUGUAGGGCGUGACUUGUGCAUAAUUUGCUGAUUUCCAAUGUCAAAUUGCACUUGUCCCGGCCUGCAGC3′ 5′AAUUGCACUUGUCCCGGCCUGC3′	100.00
aga-mir-190	5′UUUCGGUAAGAUAUGUUUGAUAUUCUUGGUUGUUAAAUUGUUCAAUUAUCACCCAGGAAUCAAACAUAUUAUUACUGUGAC3′ 5′AGAUAUGUUUGAUAUUCUUGGUUG3′	100.00
aga-mir-184	5′GGUGCACUCGAACCCUUAUCAUUCUUUCGCCCCGUGUGCAUUGCGAACCGACUGGACGGAGAACUGAUAAGGGCCCGGGUCACC3′ 5′UGGACGGAGAACUGAUAAGGG3′	100.00
aga-bantam	5′AAAUGUAAUCACAGAACCGGUUUUCAUUUUCGAUCUGACUUAUUCAUUUUACAACGAGUGAGAUCACUUUGAAAGCUGAUUUUGUACAGUUAACUCAACG3′ 5′UGAGAUCACUUUGAAAGCUGAUU3′	100.00
aga-mir-263	5′CCCUGGUACAUGUAAUGGCACUGGAAGAAUUCACGGGAUUUGUUUCAAUACUCCCGUGUUCUCUUAGUGGCAUACCCAGUACAGGG3′ 5′UGUAAUGGCACUGGAAGAAUUCAC3′	100.00
aga-mir-277	5′GUUUUGGGGUACGUGUCAGAAGUGCAUUUACAUCGGCAUUCCGCAGUUUGAGGUAUUUGUAAAUGCACUAUCUGGUACGACAUUCCAGAAU3′ 5′UAAAUGCACUAUCUGGUACGACA3′	100.00
aga-mir-124	5′CGUUUUUCUCCUGGUGUUCACUGUAGGCCUGUAUGUUACCUGAUUCCAUAAGGCACGCGGUGAAUGCCAAGAGCGAACG3′ 5′UAAGGCACGCGGUGAAUGCCAAG3′	100.00
aga-mir-10	5′UUAUGUUCUACAUCUACCCUGUAGAUCCGAAUUUGUUUGAAAUUUAACAAGCGACAAAUUCGGUUCUAGAGAGGUUUGUGUGG3′ 5′ACCCUGUAGAUCCGAAUUUGU3′	100.00

Continued on Next Page...

miRNA	Alignment	Identity
aga-mir-13b	5′ UCGUGGUCGGGUCGUAAAAAUGGUUGUGCUGUGUCGAUACUUACGAAAGUUCAUAUCACAGCCAUUUUGACGAGUU 3′ 5′ UAUCACAGCCAUUUUGACGAGU 3′	100.00
aga-mir-988	5′ CCGGUGUGUGCUUUGUGACAAUGAGAUUUUCAGUUGAAGUUCAUCCCCUUGUUGCAAACCUCACGCUGG 3′ 5′ CCCCUUGUUGCAAACCUCACGC 3′	100.00
aga-mir-276	5′ GGUGAUUGCCAUCAGCGAGGUAUAGAGUUCCUACGUUGUUCAUAUGAAAUCUGUAGGAACUUCAUACCGUGCUCUUGGAUAGCC 3′ 5′ UAGGAACUUCAUACCGUGCUCU 3′	100.00
aga-mir-219	5′ UUUCUAGCUCUGAUUGUCCAAACGCAAUUCUUGUUGAUACCAUUAGCUACUCAAGAGUUGUGACUGGACAUCCGUGGCUCG 3′ 5′ UGAUUGUCCAAACGCAAUUCUUG 3′	100.00
aga-mir-282	5′ CUAAUCUAGCCUCUCCUAGGCUUUGUCUGUAAAUGGUUUCACAAUCCAGACAUAGCCUGACAGAGGUUAGGUGAAAUCUG 3′ 5′ AAUCUAGCCUCUUCUAGGCUUUGUCUGU 3′	96.43
aga-mir-9b	5′ CACUUAUUGGGUCUUUGGUGAUUUUAGCUGUAUGUUAUUUCUUUCACAUAUAGCUUUAUCACCAAAAACCUAAUGUGUG 3′ 5′ ACUUUGGUGAUUUUAGCUGUAUG 3′	95.65
aga-mir-1890	5′ CAGAGCUAAUUGGAGCAUUUCUUGAAGAUAUAUUUUCUGCAAACUCAUGAAAUCUUUGAUUAGGUCUGGUU 3′ 5′ UGAAAUCUUUGAUUAGGUCU 3′	100.00
aga-mir-317	5′ CUCUGCCGCUGGGAUACACCUUGUGCUCGCUUUGCAAUUGAAAUACUCAAGUGAACACAUCUGGUGGUAUCUCAGUGGCCGGG 3′ 5′ UGAACACAUCUGGUGGUAUCUCAGU 3′	100.00
aga-mir-275	5′ CGCGCUAAGCAGGAACCGGGACUUGAUCCAUUUUGCAAACAGUCAGGUACCUGAAGUAGCGCGCGUU 3′ 5′ UCAGGUACCUGAAGUAGCGCGCG 3′	100.00
aga-mir-87	5′ GAUUGCUCCCGGCCCAGCCUGAAAUUUGCUAAACCUGCUGCAUAUGAGGAAAAGGUGAGCAAAUAUUCAGGUGUGUCGAAGAGUGGUC 3′ 5′ GGUGAGCAAAUAUUCAGGUGU 3′	100.00
aga-mir-308	5′ UGUUUCGCAGUAUAUUCUUGUGAGUUUGCUUCCUUUUUAUGGGCCAAAUCACAGGAGUAUACUGUGAGAUG 3′ 5′ AAUCACAGGAGUAUACUGUGAG 3′	100.00

Continued on Next Page...

miRNA	Alignment	Identity
aga-mir-279	$^{5'}$AAUGGGUGUGAAUCUAGUGCUUCACAUGUUUUUGCUACUGUGACUAGAUCCACACUCAUUAA$^{3'}$ $_{5'}$UGACUAGAUCCACACUCAUUAA$_{3'}$	100.00
aga-mir-92a	$^{5'}$UCGGCUGGAUCAAGGGCAAAAUUGUGUUUUUGAUACCAAUAUUGCACUUGUCCCGGCCUAU$^{3'}$ $_{5'}$UAUUGCACUUGUCCCGGCCUAU$_{3'}$	100.00

Bibliography

[1] A Adai, C Johnson, S Mlotshwa, S Archer-Evans, V Manocha, V Vance, and V Sundaresan. Computational prediction of miRNAs in Arabidopsis thaliana. *Genome Res*, 15(1):78–91, 2005. 26

[2] B Alberts, A Johnson, J Lewis, M Raff, K Roberts, and P Walte. *Molecular Biology of the Cell, Fourth Edition*. Garland Science, 2002. xi, 11

[3] Y Altuvia, P Landgraf, G Lithwick, N Elefant, S Pfeffer, A Aravin, M J Brownstein, T Tuschl, and H Margalit. Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res*, 33(8):2697–706, 2005. 24

[4] V Ambros. MicroRNAs: tiny regulators with great potential. *Cell*, 107(7):823–6, 2001. 13, 16, 17

[5] V Ambros, B Bartel, D P Bartel, C B Burge, J C Carrington, X Chen, G Dreyfuss, S R Eddy, S Griffiths-Jones, M Marshall, M Matzke, G Ruvkun, and T Tuschl. A uniform system for microRNA annotation. *RNA*, 9(3):277–9, 2003. 18

[6] V Ambros, R C Lee, A Lavanway, P T Williams, and D Jewell. MicroRNAs and other tiny endogenous RNAs in C. elegans. *Curr Biol*, 13(10):807–18, 2003. 15, 21

[7] B Bartel and D P Bartel. MicroRNAs: at the root of plant development? *Plant Physiol*, 132(2):709–17, 2003. 15

[8] D P Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–97, 2004. 15, 83

[9] S Baskerville and D P Bartel. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, 11(3):241–7, 2005. 14, 15, 76

[10] B L Bass. Double-stranded RNA as a template for gene silencing. *Cell*, 101(3):235–8, Apr 2000. 13

[11] I Bentwich, A Avniel, Y Karov, R Aharonov, S Gilad, O Barad, A Barzilai, P Einat, U Einav, E Meiri, E Sharon, Y Spector, and Z Bentwich. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet*, 37(7):766–70, 2005. 18, 24

[12] E Berezikov, V Guryev, J van de Belt, E Wienholds, R H A Plasterk, and E Cuppen. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, 120(1):21–4, 2005. 19, 20, 21

[13] M J Blow, R J Grocock, S van Dongen, A J Enright, E Dicks, P A Futreal, R Wooster, and M R Stratton. RNA editing of human microRNAs. *Genome Biol*, 7(4):R27, 2006. 87

[14] A F Bompfunewerer, R Backofen, S H Bernhart, J Hertel, I L Hofacker, P F Stadler, and S Will. Variations on RNA folding and alignment: lessons from Benasque. *J Math Biol*, 56(1-2):129–144, 2008. 39

[15] E Bonnet, J Wuyts, P Rouzé, and Y Van de Peer. Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. *Proc Natl Acad Sci USA*, 101(31):11511–6, 2004. 26

[16] E Bonnet, J Wuyts, P Rouzé, and Y Van de Peer. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20(17):2911–7, 2004. 84

[17] G M Borchert, W Lanier, and B L Davidson. RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol*, 13(12):1097–101, 2006. 14

[18] E Borenstein and E Ruppin. Direct evolution of genetic robustness in microRNA. *Proc Natl Acad Sci USA*, 103(17):6593–8, 2006. 35, 37, 84

[19] R K Bradley, L Pachter, and I Holmes. Specific alignment of structured RNA: stochastic grammars and sequence annealing. *Bioinformatics*, 24(23):2677–83, 2008. 39

[20] M Brameier and C Wiuf. Ab initio identification of human microRNAs based on structure motifs. *BMC Bioinformatics*, 8(1):478, 2007. 22

[21] J Brennecke and S M Cohen. Towards a complete description of the microRNA complement of animal genomes. *Genome Biol*, 4(9):228, 2003. 18

[22] X Cai, C H Hagedorn, and B R Cullen. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, 10(12):1957–66, 2004. 14

[23] R Chatterjee and K Chaudhuri. An approach for the identification of microRNA with an application to Anopheles gambiae. *Acta Biochim Pol*, 53(2):303–9, 2006. 24

[24] H R Chiang, L W Schoenfeld, J G Ruby, V C Auyeung, N Spies, D Baek, W K Johnston, C Russ, S Luo, J E Babiarz, R Blelloch, G P Schroth, C Nusbaum, and D P Bartel. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev*, 24(10):992–1009, May 2010. 3

[25] D H Chitwood and M C P Timmermans. Target mimics modulate miRNAs. *Nat Genet*, 39(8):935–6, 2007. 16

[26] G M Cooper and R E Hausman. *The Cell: A Molecular Approach, Third Edition*. Sinauer Associates, Inc., 2003. xi, 9, 10, 11, 12

[27] B R Cullen. Transcription and processing of human microRNA precursors. *Mol Cell*, 16(6):861–5, 2004. 14, 15, 16

[28] T Dezulian, M Remmert, J F Palatnik, D Weigel, and D H Huson. Identification of plant microRNA homologs. *Bioinformatics*, 22(3):359–60, 2006. 27

[29] R O Duda, P E Hart, and D G Stork. *Pattern Classification*. John Wiley & Sons, INC., 2001. 43

[30] M R Friedländer, W Chen, C Adamidi, J Maaskola, R Einspanier, S Knespel, and N Rajewsky. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol*, 26(4):407–15, 2008. 3, 17, 24

[31] D Gautheret and A Lambert. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J Mol Biol*, 313(5):1003–11, 2001. 24

[32] J Gorodkin, J H Havgaard, M Ensterö, M Sawera, P Jensen, M Ohman, and M Fredholm. MicroRNA sequence motifs reveal asymmetry between the stem arms. *Comput Biol Chem*, 30(4):249–54, 2006. 15

[33] J Gorodkin, L J Heyer, and G D Stormo. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res*, 25(18):3724–32, 1997. 39

[34] Y Grad, J Aach, G D Hayes, B J Reinhart, G M Church, G Ruvkun, and J Kim. Computational and experimental identification of C. elegans microRNAs. *Mol Cell*, 11(5):1253–63, 2003. 17, 19, 24

[35] S Griffiths-Jones, A Bateman, M Marshall, A Khanna, and S R Eddy. Rfam: an RNA family database. *Nucleic Acids Res*, 31(1):439–41, 2003. 39

[36] D Gusfield. *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, 1999. 32

[37] S M Hammond, E Bernstein, D Beach, and G J Hannon. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature*, 404(6775):293–6, Mar 2000. 10.1038/35005107. 13

[38] J Han, Y Lee, K-H Yeom, J-W Nam, I Heo, J-K Rhee, S Y Sohn, Y Cho, B-T Zhang, and V N Kim. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell*, 125(5):887–901, 2006. 15

[39] J H Havgaard, R B Lyngso, G D Stormo, and J Gorodkin. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, 21(9):1815–24, 2005. 39

[40] S A Helvik, O Snøve, and P Saetrom. Reliable prediction of Drosha processing sites improves microRNA gene prediction. *Bioinformatics*, 23(2):142–9, 2007. 22

[41] J Hertel and P F Stadler. Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, 22(14):e197–202, 2006. 22

[42] Matthias Höchsmann, Thomas Töller, Robert Giegerich, and Stefan Kurtz. Local similarity in RNA secondary structures. In *Proceedings of Computational Systems Bioinformatics (CSB 2003)*, volume 2, pages 159–168. IEEE Computer Society, 2003. 39

[43] I L Hofacker, S H Bernhart, and P F Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20(14):2222–7, 2004. 39

[44] I L Hofacker, M Fekete, and P F Stadler. Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, 319(5):1059–66, 2002. 43

[45] I L Hofacker, W Fontana, P F Stadler, S Bonhoeffer, M Tacker, and P Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte Chemie*, 125:167–188, 1994. 35, 36, 43

[46] T Huang, B Fan, M Rothschild, Z Hu, K Li, and S Zhao. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC Bioinformatics*, 8(1):341, 2007. 22

[47] G Hutvágner and P D Zamore. A microRNA in a multiple-turnover RNAi enzyme complex. *Science*, 297(5589):2056–60, 2002. 15

[48] National Human Genome Research Institute. Nucleotide. Website, 2006. `http://www.genome.gov/Pages/Hyperion/DIR/VIP/Glossary/Illustration/nucleotide.cfm?key=nucleotide`. xi, 8

[49] P Jiang, H Wu, W Wang, W Ma, X Sun, and Z Lu. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res*, 35(Web Server issue):W339–44, 2007. 22

[50] M W Jones-Rhoades and D P Bartel. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell*, 14(6):787–99, 2004. 25

[51] B Kaczkowski, E Torarinsson, K Reiche, J H Havgaard, P F Stadler, and J Gorodkin. Structural profiles of human miRNA families from pairwise clustering. *Bioinformatics*, 25(3):291–4, 2009. 39, 43

[52] S Kadri, V Hinman, and P V Benos. HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC Bioinformatics*, 10 Suppl 1, 2009. 22, 58

[53] A Khvorova, A Reynolds, and S D Jayasena. Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115(2):209–16, 2003. 15

[54] Y-K Kim and V N Kim. Processing of intronic microRNAs. *EMBO J*, 26(3):775–83, 2007. 16

[55] W P Kloosterman, E Wienholds, R F Ketting, and R H A Plasterk. Substrate requirements for let-7 function in the developing zebrafish embryo. *Nucleic Acids Res*, 32(21):6284–91, 2004. 16

[56] R Knee and P R Murphy. Regulation of gene expression by natural antisense RNA transcripts. *Neurochem Int*, 31(3):379–92, Sep 1997. 13

[57] J Krol, K Sobczak, U Wilczynska, M Drath, A Jasinska, D Kaczynska, and W J Krzyzosiak. Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. *J Biol Chem*, 279(40):42230–9, 2004. 15

[58] M Kumar and G G Carmichael. Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. *Microbiol Mol Biol Rev*, 62(4):1415–34, Dec 1998. 13

[59] E C Lai. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet*, 30(4):363–4, 2002. 16

[60] E C Lai. microRNAs: runts of the genome assert themselves. *Curr Biol*, 13(23):R925–36, 2003. 13, 18

[61] E C Lai, P Tomancak, R W Williams, and G M Rubin. Computational identification of Drosophila microRNA genes. *Genome Biol*, 4(7):R42, 2003. 14, 15, 19, 20

[62] P Landgraf, M Rusu, R Sheridan, A Sewer, N Iovino, A Aravin, S Pfeffer, A Rice, A O Kamphorst, M Landthaler, C Lin, N D Socci, L Hermida, V Fulci, S Chiaretti, R Foà, J Schliwka, U Fuchs, A Novosel, R-U Müller, B Schermer, U Bissels, J Inman, Q Phan, M Chien, D B Weir, R Choksi, G De Vita, D Frezzetti, H-I Trompeter, V Hornung, G Teng, G Hartmann, M Palkovits, R Di Lauro, P Wernet, G Macino, C E Rogler, J W Nagle, J Ju, F N Papavasiliou, T Benzing, P Lichter, W Tam, M J Brownstein, A Bosio, A Borkhardt, J J Russo, C Sander, M Zavolan, and T Tuschl. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129(7):1401–14, 2007. 24

[63] N C Lau, L P Lim, E G Weinstein, and D P Bartel. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294(5543):858–62, 2001. 13, 14, 15

[64] M T Lee, J Kim, and L A Meyers. Self containment, a property of modular RNA structures, distinguishes microRNAs. *PLoS Comput Biol*, 4(8):e1000150, Aug 2008. 35, 37

[65] R C Lee and V Ambros. An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294(5543):862–4, 2001. 13, 17, 18

[66] Y Lee, C Ahn, J Han, H Choi, J Kim, J Yim, J Lee, P Provost, O Rådmark, S Kim, and V N Kim. The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425(6956):415–9, 2003. 14

[67] Y Lee, K Jeon, J-T Lee, S Kim, and V N Kim. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J*, 21(17):4663–70, 2002. 14

[68] Y Lee, M Kim, J Han, K-H Yeom, S Lee, S H Baek, and V N Kim. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J*, 23(20):4051–60, 2004. 14

[69] M Legendre, A Lambert, and D Gautheret. Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, 21(7):841–5, 2005. 24

[70] D Leja and National Human Genome Research Institute. RNA ribonucleic acid - a more detailed description. Website. `http://www.accessexcellence.org/RC/VL/GG/rna2.php`. xi, 10

[71] B P Lewis, C B Burge, and D P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, 2005. 87

[72] Y Li, W Li, and Y-X Jin. Computational identification of novel family members of microRNA genes in Arabidopsis thaliana and Oryza sativa. *Acta Biochim Biophys Sin (Shanghai)*, 37(2):75–87, 2005. 27

[73] L P Lim, M E Glasner, S Yekta, C B Burge, and D P Bartel. Vertebrate microRNA genes. *Science*, 299(5612):1540, 2003. 21

[74] L P Lim, N C Lau, P Garrett-Engele, A Grimson, J M Schelter, J Castle, D P Bartel, P S Linsley, and J M Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–773, 2005. 16

[75] L P Lim, N C Lau, E G Weinstein, A Abdelhakim, S Yekta, M W Rhoades, C B Burge, and D P Bartel. The microRNAs of Caenorhabditis elegans. *Genes Dev*, 17(8):991–1008, 2003. 14, 15, 17, 19, 20

[76] S-L Lin, D Chang, and S-Y Ying. Asymmetry of intronic pre-miRNA structures in functional RISC assembly. *Gene*, 356:32–8, 2005. 15

[77] M Lindow and J Gorodkin. Principles and limitations of computational microRNA gene and target finding. *DNA Cell Biol*, 26(5):339–51, 2007. 39

[78] M Lindow and A Krogh. Computational evidence for hundreds of non-conserved plant microRNAs. *BMC Genomics*, 6:119, 2005. 26

[79] D J Luciano, H Mirsky, N J Vendetti, and S Maas. RNA editing of a miRNA precursor. *RNA*, 10(8):1174–7, 2004. 15, 87

[80] E Lund, S Güttinger, A Calado, J E Dahlberg, and U Kutay. Nuclear export of microRNA precursors. *Science*, 303(5654):95–8, 2004. 15

[81] J R Lytle, T A Yario, and J A Steitz. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. *Proc Natl Acad Sci USA*, 104(23):9667–72, 2007. 16

[82] D H Mathews and D H Turner. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol*, 317(2):191–203, 2002. 39

[83] N D Mendes, A T Freitas, and M-F Sagot. Current tools for the identification of miRNA genes and their targets. *Nucleic Acids Res*, 37(8):2419–33, May 2009. xi, 3, 6, 14

[84] N D Mendes, A T Freitas, A T Vasconcelos, and M-F Sagot. Combination of measures distinguishes pre-miRNAs from other stem-loops in the genome of the newly sequenced Anopheles darlingi. *BMC Genomics*, 11:529, Jan 2010. 5

[85] A A Millar and P M Waterhouse. Plant and animal microRNAs: similarities and differences. *Funct Integr Genomics*, 5(3):129–35, 2005. 16

[86] A Molnár, F Schwach, D J Studholme, E C Thuenemann, and D C Baulcombe. miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature*, 447(7148):1126–9, 2007. 24

[87] J-W Nam, K-R Shin, J Han, Y Lee, V N Kim, and B-T Zhang. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res*, 33(11):3570–81, 2005. 21

[88] K L S Ng and S K Mishra. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, 23(11):1321–30, 2007. 22

[89] T Norden-Krichmar, J Holtz, A Pasquinelli, and T Gaasterland. Computational prediction and experimental validation of Ciona intestinalis microRNA genes. *BMC Genomics*, 8(1):445, 2007. 24

[90] US National Library of Medicine. What is DNA? Website, 2008. `http://ghr.nlm.nih.gov/handbook/basics/dna`. xi, 8

[91] U Ohler, S Yekta, L P Lim, D P Bartel, and C B Burge. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA*, 10(9):1309–22, 2004. 20

[92] M Ohman. A-to-I editing challenger or ally to the microRNA process. *Biochimie*, 89(10):1171–6, 2007. 87

[93] M Y Park, G Wu, A Gonzalez-Sulser, H Vaucheret, and R S Poethig. Nuclear processing and export of microRNAs in Arabidopsis. *Proc Natl Acad Sci USA*, 102(10):3691–6, 2005. 15

[94] A E Pasquinelli, B J Reinhart, F Slack, M Q Martindale, M I Kuroda, B Maller, D C Hayward, E E Ball, B Degnan, P Müller, J Spring, A Srinivasan, M Fishman, J Finnerty, J Corbo, M Levine, P Leahy, E Davidson, and G Ruvkun. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–9, 2000. 13

[95] S Pfeffer, A Sewer, M Lagos-Quintana, R Sheridan, C Sander, F A Grässer, L F van Dyk, C K Ho, S Shuman, M Chien, J J Russo, J Ju, G Randall, B D Lindenbach, C M Rice, V Simon, D D Ho, M Zavolan, and T Tuschl. Identification of microRNAs of the herpesvirus family. *Nat Methods*, 2(4):269–76, 2005. 16, 21

[96] R S Pillai. MicroRNA function: multiple mechanisms for a tiny RNA? *RNA*, 11(12):1753–61, 2005. 16

[97] R F Place, L-C Li, D Pookot, E J Noonan, and R Dahiya. MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proc Natl Acad Sci USA*, 105(5):1608–13, 2008. 16

[98] B J Reinhart, E G Weinstein, M W Rhoades, B Bartel, and D P Bartel. MicroRNAs in plants. *Genes Dev*, 16(13):1616–26, 2002. 16

[99] M W Rhoades, B J Reinhart, L P Lim, C B Burge, B Bartel, and D P Bartel. Prediction of plant microRNA targets. *Cell*, 110(4):513–20, 2002. 16

[100] W Ritchie, M Legendre, and D Gautheret. RNA stem-loops: to be or not to be cleaved by RNAse III. *RNA*, 13(4):457–62, 2007. 25

[101] Harlan Robins and William H Press. Human microRNAs target a functionally distinct population of genes with AT-rich 3' UTRs. *Proc Natl Acad Sci USA*, 102(43):15557–62, 2005. 87

[102] A Rodriguez, S Griffiths-Jones, J L Ashurst, and A Bradley. Identification of mammalian microRNA host genes and transcription units. *Genome Res*, 14(10A):1902–10, 2004. 14, 15

[103] D Rose, J Hackermueller, S Washietl, K Reiche, J Hertel, S Findeiss, P Stadler, and S Prohaska. Computational RNomics of drosophilids. *BMC Genomics*, 8(1):406, 2007. 21

[104] J G Ruby, C H Jan, and D P Bartel. Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448(7149):83–6, 2007. 16

[105] J G Ruby, A Stark, W K Johnston, M Kellis, D P Bartel, and E C Lai. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res*, 17(12):1850–64, 2007. 21, 24

[106] T Sandmann and S M Cohen. Identification of novel Drosophila melanogaster MicroRNAs. *PLoS ONE*, 2(11):e1265, 2007. 21

[107] D Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J Appl Math*, 45(5):810–825, 1985. 39, 43

[108] D S Schwarz, G Hutvágner, T Du, Z Xu, N Aronin, and P D Zamore. Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, 115(2):199–208, 2003. 15

[109] A Sewer, N Paul, P Landgraf, A Aravin, S Pfeffer, M J Brownstein, T Tuschl, E van Nimwegen, and M Zavolan. Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, 6:267, 2005. 22, 44

[110] Y Sheng, P G Engström, and B Lenhard. Mammalian MicroRNA prediction through a support vector machine model of sequence and structure. *PLoS ONE*, 2(9):e946, 2007. 22

[111] S Siebert and R Backofen. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, 21(16):3352–9, 2005. 39

[112] N R Smalheiser and V I Torvik. Mammalian microRNAs derived from genomic repeats. *Trends Genet*, 21(6):322–6, 2005. 14

[113] T F Smith and M S Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147:195–197, 1981. 33

[114] D J States, W Gish, and S F Altschul. Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods*, 3(1):66–70, 1991. 34

[115] R Sunkar, X Zhou, Y Zheng, W Zhang, and J Zhu. Identification of novel and candidate miRNAs in rice by high throughput sequencing. *BMC Plant Biol*, 8(1):25, 2008. 27

[116] R Sunkar and J-K Zhu. Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. *Plant Cell*, 16(8):2001–19, 2004. 16

[117] C Vanhée-Brossollet and C Vaquero. Do natural antisense transcripts make sense in eukaryotes? *Gene*, 211(1):1–9, Apr 1998. 13

[118] S Vasudevan, Y Tong, and J A Steitz. Switching from repression to activation: microRNAs can up-regulate translation. *Science*, 318(5858):1931–4, 2007. 16

[119] M C Vella, K Reinert, and F J Slack. Architecture of a validated microRNA::target interaction. *Chem Biol*, 11(12):1619–23, 2004. 87

[120] E G Wagner and R W Simons. Antisense RNA control in bacteria, phages, and plasmids. *Annu Rev Microbiol*, 48:713–42, Jan 1994. 13

[121] X-J Wang, J L Reyes, N-H Chua, and T Gaasterland. Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. *Genome Biol*, 5(9):R65, 2004. 25

[122] Xiaowo Wang, Jing Zhang, Fei Li, Jin Gu, Tao He, Xuegong Zhang, and Yanda Li. MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, 21(18):3610–4, 2005. 25

[123] Y Wang, G Sheng, S Juranek, T Tuschl, and D J Patel. Structure of the guide-strand-containing argonaute silencing complex. *Nature*, 456(7219):209–13, 2008. 87

[124] J Watson and F Crick. A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953. 7

[125] D Weaver, J Anzola, J Evans, J Reid, J Reese, K Childs, E Zdobnov, M Samanta, J Miller, and C Elsik. Computational and transcriptional evidence for microRNAs in the honey bee genome. *Genome Biol*, 8(6):R97, 2007. 24

[126] Michel J Weber. New human and mouse microRNA genes found by homology search. *FEBS J*, 272(1):59–73, 2005. 16, 24

[127] S Will, K Reiche, I L Hofacker, P F Stadler, and R Backofen. Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLOS Computational Biology*, 3(4):e65, 2007. 39, 43

[128] S Will, K Reiche, I L Hofacker, P F Stadler, and R Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol*, 3(4):e65, 2007. 25, 64

[129] F Winter, S Edaye, A Hüttenhofer, and C Brunel. Anopheles gambiae miRNAs as actors of defence reaction against plasmodium invasion. *Nucleic Acids Res*, 35(20):6953–62, Jan 2007. 53, 54, 76

[130] S Wuchty, W Fontana, I L Hofacker, and P Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–65, Feb 1999. 36

[131] X Xie, J Lu, E J Kulbokas, T R Golub, V Mootha, K Lindblad-Toh, E S Lander, and M Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–45, 2005. 23

[132] C Xue, F Li, T He, G-P Liu, Y Li, and X Zhang. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6:310, 2005. 22, 40, 58, 74

[133] R Yelin, D Dahary, R Sorek, E Y Levanon, O Goldstein, A Shoshan, A Diber, S Biton, Y Tamir, R Khosravi, S Nemzer, E Pinner, S Walach, J Bernstein, K Savitsky, and G Rotman. Widespread occurrence of antisense transcription in the human genome. *Nat Biotechnol*, 21(4):379–86, Apr 2003. 13

[134] S-Y Ying and S-L Lin. Intronic microRNAs. *Biochem Biophys Res Commun*, 326(3):515–20, 2005. 15

[135] W J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, Jan 1950. 58

[136] M Yousef, M Nebozhyn, H Shatkay, S Kanterakis, L C Showe, and M K Showe. Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics*, 22(11):1325–34, 2006. 22

[137] P D Zamore. Ancient pathways programmed by small RNAs. *Science*, 296(5571):1265–9, May 2002. 13

[138] Y Zeng and B R Cullen. Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. *Nucleic Acids Res*, 32(16):4776–85, 2004. 15

[139] B Zhang, X Pan, C H Cannon, G P Cobb, and T A Anderson. Conservation and divergence of plant microRNA genes. *Plant J*, 46(2):243–59, 2006. 27

[140] B H Zhang, X P Pan, S B Cox, G P Cobb, and T A Anderson. Evidence that miRNAs are different from other RNAs. *Cell Mol Life Sci*, 63(2):246–54, 2006. 35, 36, 55, 56, 84

[141] H Zhang, F A Kolb, L Jaskiewicz, E Westhof, and W Filipowicz. Single processing center models for human Dicer and bacterial RNase III. *Cell*, 118(1):57–68, 2004. 15