The Microsoft Research - University of Trento
Centre for Computational and Systems Biology

Technical Report CoSBi 24/2007

An integration of miRNA target predictions for the characterization of human miRNAs

Ilenia Fronza
Centre for Integrative Biology
University of Trento
`fronza@science.unitn.it`

Alessandro Quattrone
Centre for Integrative Biology
University of Trento
`quattrone@science.unitn.it`

Angela Re
Centre for Integrative Biology
University of Trento
`angelare@dit.unitn.it`

UNIVERSITÀ DEGLI STUDI DI TRENTO

The Microsoft Research - University of Trento
Centre for Computational and Systems Biology

Second Level International Master in Computational and Systems Biology

Master Thesis

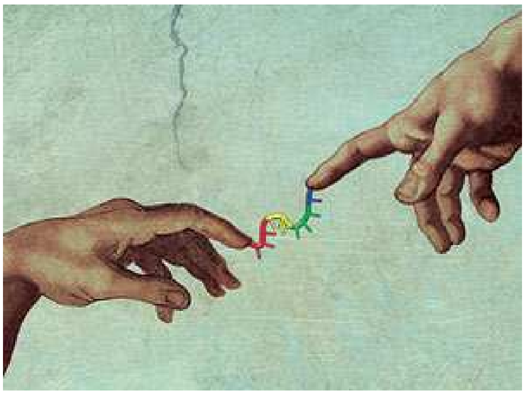
An integration of miRNA target predictions for the characterization of human miRNAs

Advisor
Prof. Alessandro Quattrone

Student
Ilenia Fronza

Co-advisor
Dott.ssa Angela Re

Academic year 2006/2007



Acknowledgments

Now that this experience is over, I wish to thank prof. Alessandro Quattrone for always believing in me and for constantly proving it with his support; another big thank you to Angela Re for her invaluable cooperation.

I could not neglect acknowledging my parents, always ready to share my choices along with all the happiness as well as the darker times and discomforts which they entail.

Thanks to Arianna, for the simple but great reason of being there. I often don't follow her advice, but she still finds herself helping me out in some way or another; and in this way she involved herself in this work, staying close to me and encouraging me, sharing all the hardships met in this period. So, Arianna, this is yours as well . . . like back then, and more than back then.

Ringraziamenti

Al termine di questa esperienza, desidero ringraziare di cuore il prof. Alessandro Quattrone per aver sempre creduto in me e avermelo sempre dimostrato con il suo appoggio; preziosissima la collaborazione con Angela Re alla quale va un altro enorme grazie.

Come non essere riconoscente ai miei genitori, sempre pronti a condividere le mie scelte con tutte le gioie, ma anche i periodi bui e gli sconforti, che comportano.

Grazie ad Arianna, per la semplice ma grande ragione di esserci. Spesso non seguo i suoi consigli, ma lei poi in un modo o nell'altro si ritrova ad aiutarmi; e così si è impegnata con tutta sè stessa in questo lavoro, standomi vicina e incoraggiandomi, facendo sue tutte le difficoltà trovate in questo periodo. E allora, Arianna, anche questa è tua . . . come allora, e più di allora.

“. . . One is like ten thousand for me,
if it is the best. . . ”
Eraclito

“. . . Uno è per me diecimila,
se è il migliore. . . ”
Eraclito

To Arianna

Contents

List of Figures

List of Tables

Introduction

In the last twenty years it has become clear that post-transcriptional regulation is an elaborate pathway of equal, if not greater, complexity to the one controlling mRNA transcription. Proteins that are involved in transcription, for example, represent as much as 10% of the coding sequence in the genome of metazoans. By contrast, specificity for RNA-mediated silencing-based regulation is conferred by small RNA guides (for example, micro RNAs (miRNAs), small interfering RNAs (siRNAs) and PIWI-interacting RNAs (piRNAs)) that are generated through distinct biogenesis pathways and function in collaboration with specialized effector proteins.

Here we shall consider the miRNA class of animal small silencing RNAs. MiRNAs are an ancient innovation among animals: for example, both of the first two miRNAs discovered, lin-4 and let-7, are conserved from nematodes to humans.

MiRNAs in general have captivated the scientific world, bringing new genetic tools to model organisms, new explanations for regulatory interactions, new methods to pharmaceutical discovery (the commercial potential of miRNA and related drugs is expected to exponentially increase within the next few years), and new life to the biotechnology industry. The Nobel Prize in Physiology or Medicine 2006 was awarded to Andrew Fire and Craig Mello for their discovery of RNA interference (RNAi)-gene silencing by double-stranded RNA. Even popular press highlighted this field of research; the cover of the international business magazine, the *Economist* [1], recently proclaimed small RNAs to be the “ Biology’s Big Bang”.

Although the current evidence is limited and hundreds of miRNAs of unknown function exist, it is known that miRNAs take on unique properties that allow them to regulate diverse biological processes such as cellular differentiation, proliferation, and apoptosis; because of their role in gene regulation, miRNAs are linked to several diseases due to defects in the regulation of mRNA translation which can result in an abnormal production of a protein.

Microarrays containing all known human miRNAs allow to find a downregulation of some miRNAs in various tumors, and also the lack of regulation of protein components of the miRNA biosynthetic pathway seems to be involved in cancer formation. Consistent with these studies, reports of altered miRNA expressions in human cancers are beginning to emerge in the literature, suggesting the role of miRNAs as a novel class of oncogenes or tumor suppressor genes.

Therefore, discovery of the role of miRNAs in various pathological processes has opened up possible applications in molecular diagnostics and prognostics, particularly for cancer; recent studies are trying to combine a high-throughput target analysis with genomics and proteomics in order to gain accurate predictions of target genes so as to aid cancer research.

The aim of the first part of this work (Chapter 1) is to give a biological introduction to the current information about miRNAs, their biogenesis and the pathways they are involved in; plus, functions of miRNAs and their involvement in diseases will be underlined. At the end of this part, it will be clear that the prediction of the miRNA target genes is a big challenge for bioinformatics and molecular biologists because of the imperfect base-pairing of the duplex miRNA:mRNA; the result is a blooming of target-prediction approaches and of several tools and resources that provide updated informations useful to study miRNAs, a selection of which is reported in Chapter 2 and in Chapter 3.

As we will see, there is not program for miRNA target prediction that can be considered consistently better than the others. For this reason in Chapter 4 we suggest not to arbitrarily prefer one algorithm to the others, but to consider the prediction quality of differently combined subsets of algorithms. In particular, in order to identify the best combination, we decide to evaluate pair-wise intersections of aforementioned programs and to describe the confidence of each combination by evaluating the statistical significance of the overlap in predictions.

Once identified the best combination of prediction softwares, we provide a characterization of miRNAs through the annotation of some significant targets testing the enrichment of any Gene Ontology term; links between our results in miRNA characterization and the available literature will be described in order to evaluate the consistency of the method we are proposing.

Chapter 1

Noncoding RNAs

The aim of this Chapter is to introduce the biological problem this work will talk about and also to provide necessary biological notions and notations; an overview on small RNAs will be presented, focusing on miRNAs and on the process they are involved in (post-transcriptional regulation).
Biological functions of miRNAs and their involvement in diseases will be also underlined, suggesting why it is so important to gain accurate predictions of their target genes.

1.1 The fate of eukaryotic mRNAs

Messenger RNAs (mRNAs) could be described as carriers of the genetic information from DNA to proteins; mRNAs synthesis takes place in the nucleus, where also the following processing events are located: $5'$-end capping, splicing, $3'$-end cleavage and polyadenylation. Once mature, mRNAs are exported to the cytoplasm, where they can reach ribosomes so as to be translated to produce proteins (Figure 1.1).
In the description of this process, the regulation of each step has to be taken into account; attention has traditionally been focused on trascriptional regulation, but the wide variability in the degree to which mRNA and protein abundances correlate in vivo [2] suggests to give more importance to post-transcriptional regulatory mechanisms in the control of eukaryotic gene expression.
It was initially believed that major post-transcriptional regulation involved only selected mRNA populations or was limited to highly specialized cell types (germ cells, neurons); nowadays, it is clear that the majority of mRNAs in multiple cell types is subject to regulatory activities affecting essentially every aspect of their lives [3].
Recent results show that an alternative fate for mRNAs exists: it can be either

translated or *silenced*, i.e. there is a post transcriptional regulation process capable of inhibiting translation. The *small RNAs* are crucial elements of this pathway; in addition to the associated factors (which always escort mRNAs), small RNAs can bind mRNAs together with specific proteins (called RBPs) to form the messenger ribonucleoprotein particle (mRNP). The majority of mRNA binding factors target particular structures or specific recognition sequences that commonly occur in the untranslated regions (UTRs) at the $5'$ and $3'$ ends of the mRNA.

Since containing binding sites for diverse mRNPs, mRNA can respond to many inputs (that is to say, activation/repression of processes); the result is an elaborate network of equal, if not grater, complexity to those controlling initial mRNAs synthesis [3] (Figure 1.2).

Just to give an idea, proteins that are involved in transcription represent as much as 10% of the coding sequence in the genomes of metazoans; by contrast, small RNAs are generated through distinct biogenesis pathways and function with specialized effector proteins [4].

From their discovery [1], many research works were dedicated to miRNAs and siRNAs (and small RNAs in general, Figure 1.3) because of their possible use in new genetic tools to model organisms, in new explanations for regulatory interactions, in new methods to pharmaceutical discovery, and in the biotechnology industry [7]; the Nobel Prize in Physiology or Medicine 2006 was awarded to Andrew Fire and Craig Mello [8] for their discovery of RNA interference (RNAi)-gene silencing by double-stranded RNA.

The existence of miRNAs may also help to explain the complexity of biplogical systems, which was something of a paradox until their discovery. Knowing that DNA stores data and that the complexities of development must require much information, biologists naturally expected that the more complex an organism is, the more genes it would have in its cells. It was surprising to find out that *C. elegans* has about 20000 genes and that this seems to be a widespread number for animals. But genes considered in this count are only the protein-coding ones: adding the genes whose RNA has other functions can help to explain complexity.

In Section 1.2 the major aspects concerning noncoding RNAs involved in mRNA silencing are presented, focusing on miRNAs (Section 1.3).

[1] MiRNAs were discovered by Victor Ambros in 1993 [5]; siRNAs were found just in 1999 by Hamilton and Baulcombe [6].

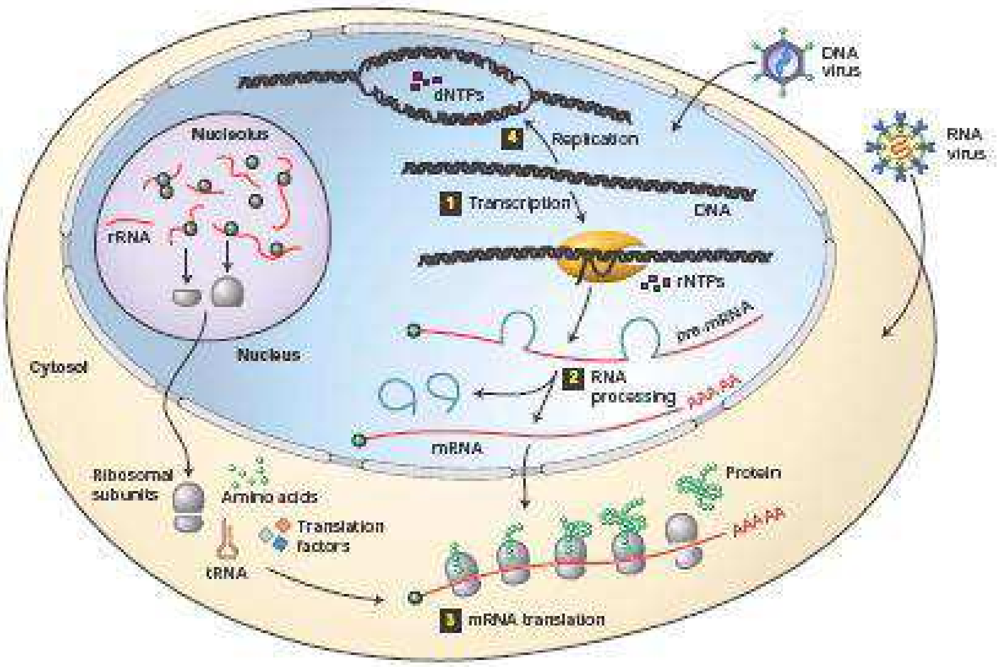


Figure 1.1: During transcription of a protein-coding gene by RNA polymerase (see number 1 in the Figure), the four-base DNA code specifying the amino acid sequence of a protein is copied into a precursor messenger RNA (pre-mRNA). A sequence of events (see number 2 in the Figure), known as RNA processing, produces a functional mRNA, which is transported to the cytoplasm. During translation (see number 3 in the Figure), the four-base code of the mRNA is decoded into the twenty-amino acid "language" of proteins. Ribosomes, which are composed of two subunits made of ribosomal RNAs (rRNAs) and multiple proteins, translate the mRNA code; ribosomal subunits associate with an mRNA and carry out protein synthesis with the help of transfer RNAs (tRNAs) and various translation factors. In cells which are preparing to divide, also DNA replication (see number 4 in the Figure) occurs: deoxyribonucleoside triphosphate monomers (dNTPs) are polymerized to produce two identical copies of each chromosomal DNA molecule. Each daughter cell receives one of the identical copies.(from [9]).

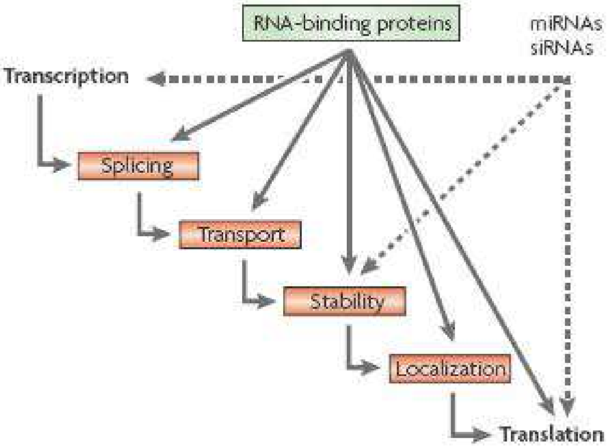


Figure 1.2: In eukaryotic cells, mRNAs undergo several steps of regulation from transcription to translation. RNA-binding proteins and small non-coding RNAs (miRNAs and siRNAs) regulate at different levels the coordination of multiple mRNAs (from [10]).

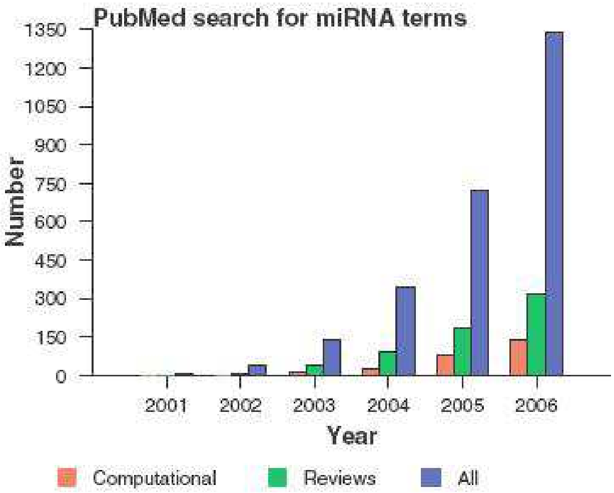


Figure 1.3: Throughout the short time span in which miRNAs have been known, a rapidly increasing amount of information about miRNAs has been accumulated. In this plot, the number of published papers (all, reviews, computational) containing the PubMed keywords *micro RNA*, *microRNA* or *miRNA* as a function of time (adapted from [11]).

1.2 Small silencing RNAs

The classical classification of the Ribonucleic Acid (RNA) in ribosomal (rRNA), messenger (mRNA) and transfer (tRNA) has become not sufficient because of the results of recent studies (as introduced in Section 1.1); new types of RNA molecules has been in fact characterized, which are involved in regulation processes of the cell, alone or as parts of ribonucleoproteic complexes.

The new family of *small silencing RNAs* has to be added to the three existing categories, in order to consider those regulatory molecules of non coding RNA, $19 - 28$ nucleotides long, which derive from a double stranded RNA and control the expression levels of their target genes.

Small RNA-mediated gene silencing has been observed in a number of eukaryotes for almost two decades but the mechanisms that yields to mRNA silencing is becoming clear only in these years. Since the observed phenomena were thought to be unrelated, several different terms refer to the same pathway, such as RNA interference (RNAi), co-suppression, quelling or gene silencing. Nowadays, it is extimated that the $40 - 50\%$ of the mammalian genome is regulated at the translational level by the action of small RNAs.

In mammalian brain these small silencing molecules of RNA are involved in neuronal differentiation as well as in learning and memory; they can be grouped in small nucleolar RNAs, small cytoplasmic RNAs, and miRNAs, which appear to be critical for dictating neuronal cell identity during development and to play an important role in neurite growth, synaptic development and neuronal plasticity [12].

In general, the three dominant classes of animal small silencing RNAs are *micro RNAs* (miRNAs), *small interfering RNAs* (siRNAs) and *PIWI- interacting RNAs* (piRNA; Table 1.1). Since the active forms are sometimes biochemically or functionally undistinguished, the classification is based on their origin [2] and on specific biogenesis factors [13] [4].

SiRNAs arise from some loci that contain inverted repeats; transcripts from these loci form foldback structures with near-perfect complementarity. Loci that generate bidirectional transcript pairs can produce siRNAs, such as natsiRNAs in *Arabidopsis thaliana*, and small RNAs that are associated with genome rearrangements in *Tetrahymena thermophila*. Specific mechanisms for small RNA biogenesis from these loci differ among lineages, but each is likely to involve processing of double stranded RNA into siRNA by members of the Dicer family. TasiRNAs are encoded at Trans-acting siRNAs (TAS) loci in plants and result from Dicer-like 4 (DCL4) processing of RNA-dependent RNA Polymerase 6 (RDR6)-dependent

[2] While miRNAs are generated from the dsRNA region of the hairpin-shaped precursors, siRNAs are derived from long dsRNAs.

dsRNA; dsRNA synthesis is triggered by miRNA-guided cleavage of the primary transcript, and is processed in a phased, 21-nt register from the miRNA-guided cleavage site.

The proposed piRNA-biogenesis model involves the initial targeting of transcripts from transposons and retroelements by a Piwi-like protein that is programmed with a small RNA; the cleavage of the transcript generates the $5'$ end of a new piRNA. Further $3'$-end processing might require a distinct Piwi-like protein, such as *Drosophila melanogaster* Argonaute 3 (Ago3), which generates a new piRNA with a $3'$-end that is offset by 10 nucleotides from the initial small RNA [4].

We will focus our attention on miRNAs (Section 1.3), describing in particular their biogenesis (Section 1.4) and action pathway (Section 1.5). We will also mention biological functions of miRNAs and their role in various pathological processes (Section 1.6).

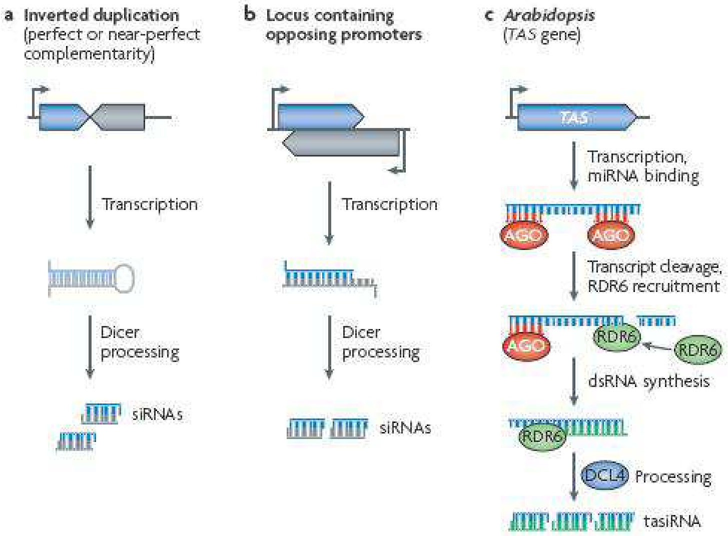


Figure 1.4: (a) siRNAs arise from some loci containing inverted repeats. (b) Loci that generate bidirectional transcript pairs can produce natsiRNAs in *Arabidopsis thaliana* and small RNAs that are associated with genome rearrangement in *Tetrahymena thermophila*. (c) tasiRNA are encoded at TAS loci in plants. In the figure: AGO, Argonaute protein (from [4]).

1.3 miRNAs

MiRNAs are small molecules of RNA involved in the regulation of gene expression, through a specific pairing with a target sequence in the $3'$ UTR of the mRNA

Table 1.1: Classes of small RNA identified in eukaryotes (from [4]).

Class	Description	Biogenesis and genomic origin	Function
miRNA	micro RNA	Processing of foldback miRNA gene transcripts by members of the Dicer and RNaseIII-like families	Post-transcriptional regulation of transcripts from a wide range of genes
Primary siRNA	Small interfering RNA	Processing of dsRNA or foldback RNA by members of the Dicer family	Binding to complementary target RNA; guide for initiation of RdRP-dependent secondary siRNA synthesis
Secondary siRNA	Small interfering RNA	RdRP activity at silenced loci (*Caenorhabditis elegans*); processing of RdRP-derived long dsRNA or long foldback RNA by members of the Dicer family (*Arabidopsis thaliana*)	Post-transcriptional regulation of transcripts; formation and maintenance of heterochromatin
tasiRNA	Trans-acting siRNA	miRNA-dependent cleavage and RdRP-dependent conversion of TAS gene transcripts to dsRNA, followed by Dicer processing	Post-transcriptional regulation of transcripts
natsiRNA	Natural antisense transcript-derived siRNA	Dicer processing of dsRNA arising from sense- and antisense-transcript pairs	Post-transcriptional regulation of genes involved in pathogen defense and stress responses in plants
piRNA	Piwi-interacting RNA	A biogenesis mechanism is emerging, which is Argonaute-dependent but Dicer-independent	Suppression of transposons and retroelements in the germ lines of flies and mammals

RdRP, RNA-dependent RNA polymerase.

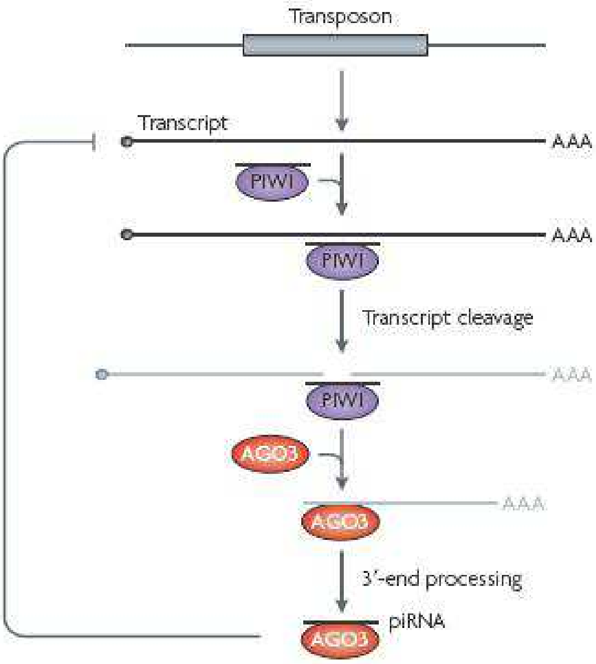


Figure 1.5: Model for Piwi-interacting RNA (piRNA) biogenesis (from [4]).

that causes an inhibition of the translation process or the degradation of the messenger. They are a family of 21-25 nucleotides (nt)-long RNAs expressed in a wide variety of organisms ranging from plants to worms and humans [3]. Many miRNAs are highly conserved across species, and components of the miRNA machinery have been found even in archaea and eubacteria, revealing their old ancestry [14]. The first known miRNA, the lin-4, was discovered in 1993 by Victor Ambros [5] and his colleagues through the study of heterochronic gene lin-14 in worms. Acting as a developmental repressor of the accumulation of lin-14 protein, lin-4 RNA controls the timing of *Caenorhabditis elegans* larval development [15].
Two major strategies led to the identification of the verified mammalian miRNAs: sequencing libraries of cloned small RNAs from various cell types and bioinformatic analysis of genomic sequences. Recent studies have suggested that as many as several hundreds of additional miRNAs are likely to exist. Even if the targets and the roles of miRNAs in mammalians are not clear yet, it is supposed that they are involved in deciding the fate of progenitor cells as in other orghanisms (*C. elegans*, *Drosophila*); this is a crucial role especially in the nervous system, where many varieties of cells can derive from a single type of cell.
Because of their role in gene regulation, miRNAs are linked to several diseases

[3] For example, both of the first two miRNAs discovered, lin-4 and let-7, are conserved from nematodes to humans.

due to defects in the regulation of mRNA translation which can result in an abnormal production of a protein. Downregulation of some miRNAs can be observed in various tumors, and also the lack of regulation of protein components of the miRNA biosynthetic pathway seems to be involved in cancer formation.

Recent studies on plants show that RNA silencing plays a role in the mechanism of defense against viruses, and also suggest a similar role in humans (Section 1.6.1) [16].

Many eukaryotic miRNAs have to be discovered and the prediction of their target genes (Chapter 3) is a big challenge for bioinformatics and molecular biologists because of their imperfect base-pairing with the target mRNAs (Section 1.5.1).

1.4 Biogenesis of miRNAs

The biogenesis of miRNAs is more deeply understood in animals. The miRGen database (Section 2.1) contains 471 human miRNAs. MiRNAs are embedded in either independent noncoding RNAs or introns of protein-coding genes; additionally, to allow coordinated expression, some miRNAs are clustered in polycistronic transcripts [14] (Figure 1.6). Information about the location of this collection of miRNAs with respect to UCSC genome can be obtained from the *genomics* interface of miRGen (Tables 1.2 and 1.3).

MiRNAs are transcribed (Figure 1.8) from different genomic locations as long primary transcripts (pri-miRNAs) by RNA polymerase II; pri-miRNAs contain a stem of 33 nt, a terminal loop and flanking single-stranded (ss)RNA sequences. The RNA pol II, which guides the transcription of the coding RNAs (mRNAs), is also the major transcriptional unit for the noncoding RNAs; the primary transcripts contain a $5'$-end cap structure and a poliA-tail sequence, which are proper to the RNA pol II transcripts [17].

The pri-miRNA is then processed by the microprocessor (a complex composed of the nuclear RNase III Drosha and the DGCR8 protein) into a $60 - 70$ nt stem loop miRNA precursor (pre-miRNA). Patients with genetic disorders regarding the information for the DGCR8 (Table 1.4) protein show congenital heart defects, characteristic facial appearance, immunodeficiency, and behavioural problems [16].

The pre-miRNA is then bound by the Exportin-5 and exported in the cytoplasm (a process that needs RanGTP) where it is cleaved at the base of the loop by a second RNase III enzyme located in the cytoplasm, Dicer [4], to generate a duplex of $21 - 24$ nt. After a not completely understood process of separation and selection, one of the two strands is cleaved while the other one, the mature miRNA,

[4] Localized mainly in the cytoplasm or in the endoplasmic reticulum, human Dicer is a large protein composed of several domains essential for mammalian development, as Dicer-deficient mice die at the embryonic stage [16].

binds to the RISC ribonucleoproteic complex [5] and guides this miRNP complex to find the target mRNA; the fate of the target mRNA is either cleavage or silencing depending on the perfect/imperfect Watson-Crick complementarity of the duplex miRNA:mRNA (Figures 1.8 and 1.9).

A single miRNA species can bind to many different mRNA targets and, conversely, several different miRNAs can cooperatively control a single mRNA target.

We have to mention that, a work published in July 2007 [18] suggests an alternative pathway for miRNA biogenesis, in which certain debranched introns mimic the structural features of pre-miRNAs to enter the miRNA-processing pathway without Drosha-mediated cleavage. These pre-miRNAs/introns are called *mirtrons*; 14 mirtrons have been identified in *Drosophila melanogaster* and four in *Caenorhabditis elegans*. In [18] it is also proposed that the mirtron pathway (Figure 1.7) could have provided an early avenue for the emergence of miRNAs before the advent of Drosha.

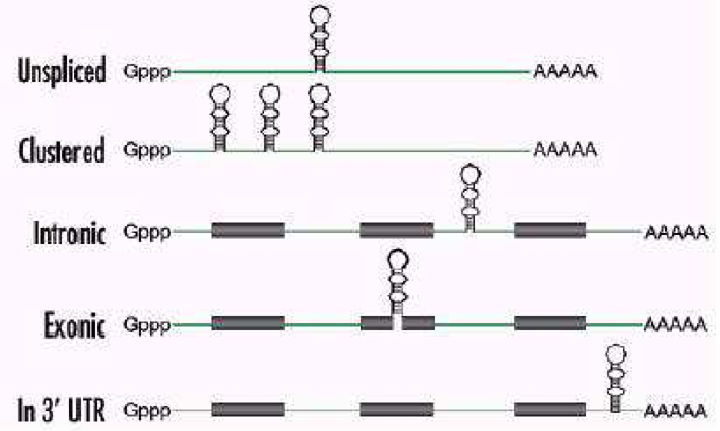


Figure 1.6: miRNA host transcripts may be unspliced, or miRNAs may be located within introns, exons, or untranslated regions (UTRs) of spliced transcripts (from [19]).

[5] The human RISC is made at least of the three proteins Dicer, TRBP and Argonaute 2 (Ago2).

Table 1.2: Location of miRNA precursors in human genome (from miRGen database, last update: January 1, 2007).

Location	Number of miRNAs
In genes	213 [a]
Out of genes	58 [a] within 5000 nt 204 [a] beyond 5000 nt

[a] *Out of* 475.

Table 1.3: Location of the miRNA precursors in the genes (from miRGen database, last update: January 1, 2007).

Location of the precursor	Number of genes
On same strand overlapping start of exon	2
On same strand inside exon	2
On opposite strand inside exon	5
On same strand overlapping end of exon	6
On same strand inside intron	325
On opposite strand inside intron	48
On same strand in $5'$ UTR	58
On opposite strand in $5'$ UTR	4
On same strand in $3'$ UTR	20
On opposite strand in $3'$ UTR	11

Table 1.4: Characteristics of the major protein components of the miRNA-guided RNA silencing pathway.

Protein	Role	Localization
Drosha	from pri-miRNA to pre-miRNA	nucleus
DGCR8	microprocessor with Drosha	nucleus
Exportin-5	pre-miRNA exporting from nucleus	nuclear membrane
Dicer	from pre-miRNA to miRNA	cytoplasm - ER
TRBP	processing complex with Dicer	cytoplasm
Ago2	processing complex with Dicer	cytoplasm

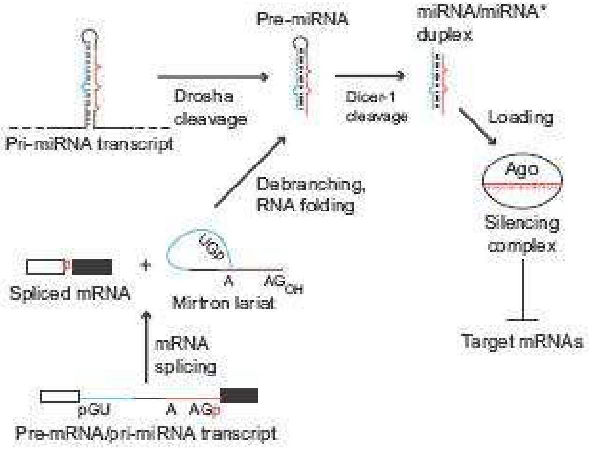


Figure 1.7: Model for the convergence of the canonical and mirtronic miRNA biogenesis pathways (from [18]).

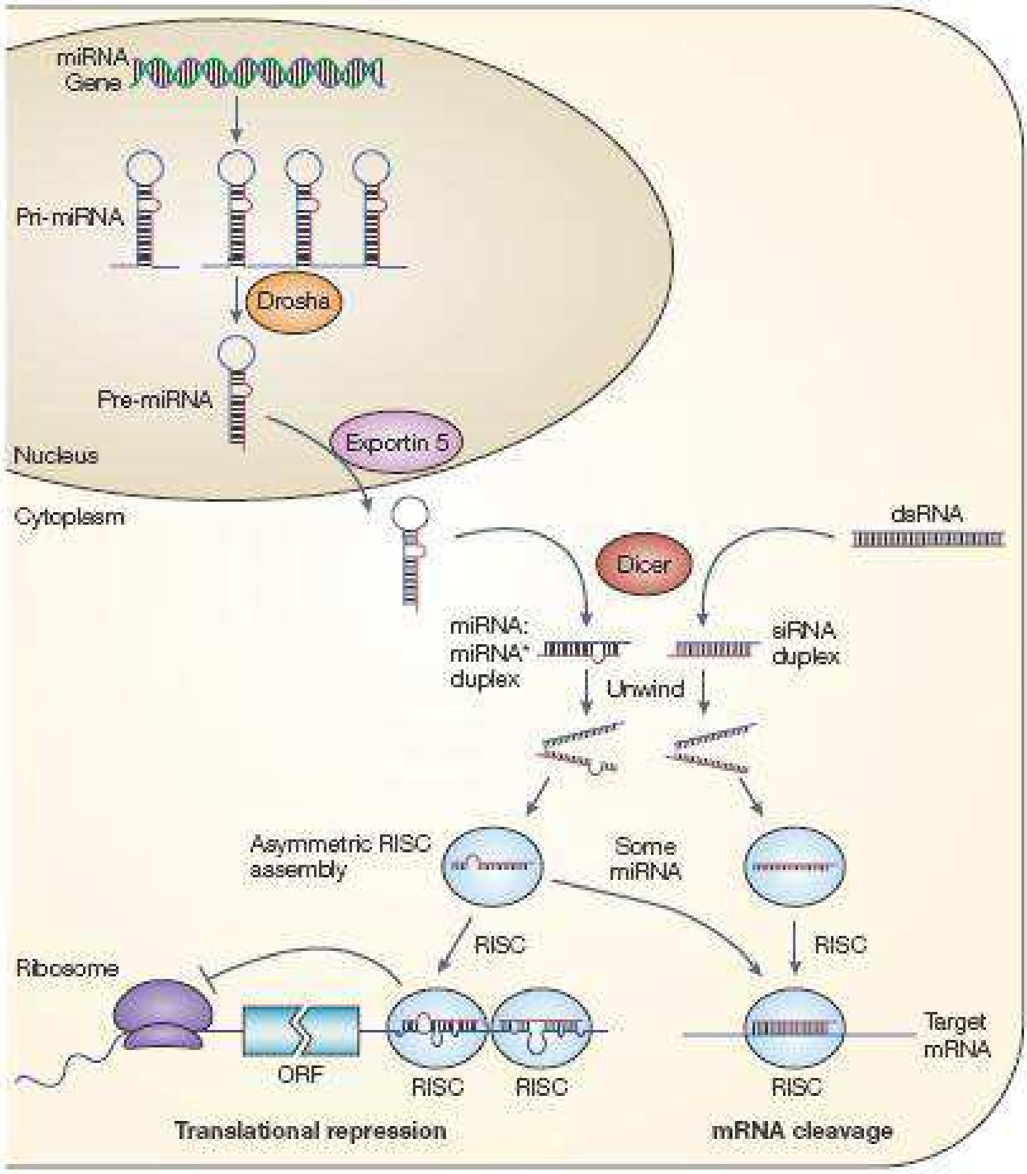


Figure 1.8: The current model for the biogenesis and post-transcriptional suppression of miRNAs and small interfering RNAs. The pri-miRNA are initially cleaved by Drosha and DGCR8 to produce the pre-miRNA which is transported to the cytoplasm by Exportin-5. A second cleavage event, performed by Dicer, generates an $18 - 24$ nt RNA duplex. One strand of this duplex is incorporated into RISC and guides the silencing of the target mRNA [19] (from [20]).

1.5 miRNAs in action

When perfectly base-paired to their target, miRNAs direct cleavage of a single phosphodiester bond in the mRNA causing its degradation; the human Ago2 protein of the RISC complex supports this event. An incomplete complementarity with the target causes the inhibition of the translation of the messenger, avoiding the accumulation of the protein for which it codes for.

MiRNAs act in particular as a guide for the proteins of the miRNP complex, so as to stop the translation of the bound mRNA. It was proposed that translation may be blocked after the initiation phase, for example by inducing proteolysis of the polypeptides as soon as they exited the ribosome [14].

If the RNA cleavage mechanism by Ago2 is a well understood event, the translational repression mechanism by miRNAs is less clear. Recent studies show that the target mRNAs binding to RISC through partial base pairing are accumulated in cytoplasmic foci, named *Processing bodies* (P-bodies) [17].

The available evidence indicates that miRNAs can repress translation at both initiation and post-initiation levels (Figure 1.10). The two-step mechanism starts with a block in translation of the m^{7}G-capped mRNA at the initiation step; run-off of ribosomes results in the second step, that is to say the aggregation of the repressed ribosome-free mRNA into P-bodies for either storage or degradation. The post-initiation mechanism is the repression of translation at the elongation or termination step, or the promotion of ribosome drop-off; in this model, repressed mRNAs remain largely associated with ribosomes and thus might not relocate to P-bodies. As yet, there is no experimental support for a post-initiation mechanism involving rapid proteolysis of nascent polypeptide chains. The mRNA decay machinery is enriched in P-bodies but it is not clear whether the degradation occurs outside or inside them [14] [21].

1.5.1 Target recognition rules

While in plants miRNAs and target mRNAs often show perfect complementarity, in animals many structures of functional duplexes are possible (short matching sequence, with gaps and mismatches).

The *seed region* is defined in miRNA target sites as the consecutive stretch of 7 nucleotides starting from either the first or the second nucleotide at the $5'$-end. Experiments led to classify miRNA target sites into three categories, depending on the type of base pairing with the seed region (Figure 1.11):

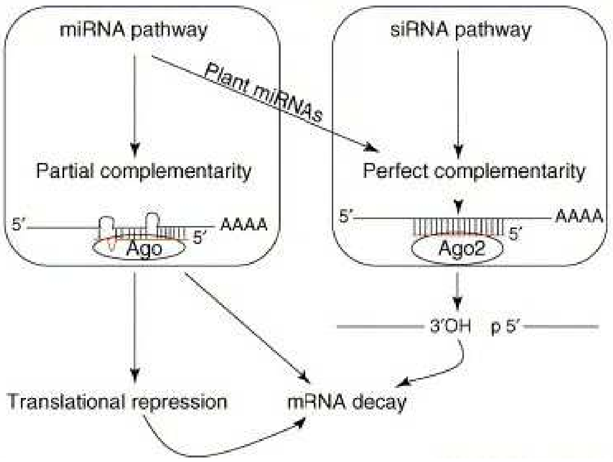


Figure 1.9: Perfect complementarity between siRNAs (and many plant miRNAs) and their target leads to an endonucleolytic cleavage, catalyzed by the human Ago2 in the RISC complex. Generally, animal miRNAs show only partial complementarity to their target mRNAs, which precludes endonucleolytic cleavage but promotes translational repression and exonucleolytic degradation of target mRNAs (from [21]).

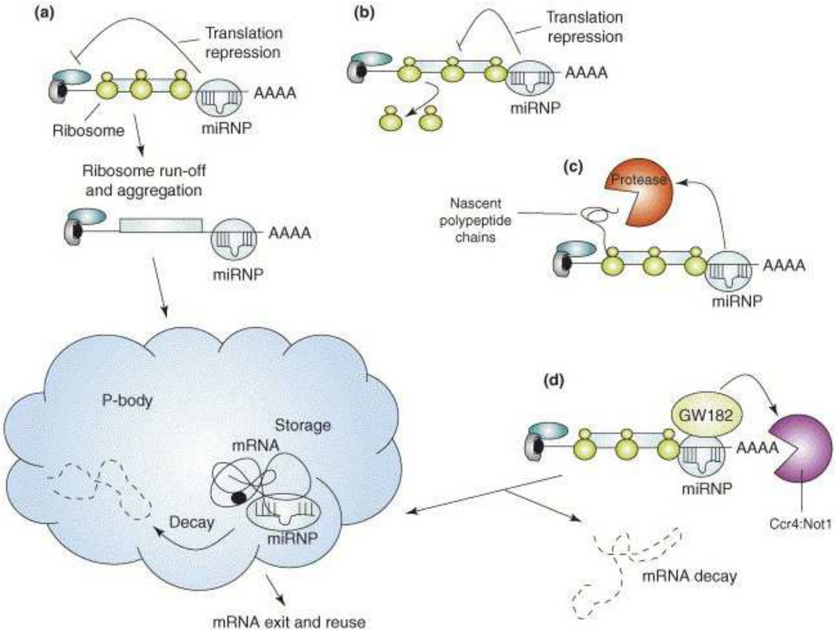


Figure 1.10: (a) The two-step mechanism; the turquoise oval represents the translation initiation complex protein eIF4G. (b) Post-initiation mechanism. (c) Post-initiation mechanism involving rapid proteolysis of nascent polypeptide chains. As yet, there is no experimental support for this model. (d) miRNA-mediated deadenylation and decay of target mRNAs (from [21]).

1. $5'$-dominant canonical: perfect base pairing to at least the seed portion of the $5'$-end of the miRNA and extensive base pairing to the $3'$-end of the miRNA;

2. $5'$-dominant seed only: base pairing is perfect at least to the seed portion of the $5'$-end of the miRNA while it is limited to the $3'$-end of the miRNA;

3. $3'$-compensatory: extensive base pairing to the $3'$-end of the miRNA to compensate for an imperfect or a shorter stretch of base pairing to the seed portion of the miRNA [22].

Experiments show that genome-wide $5'$-dominant sites occur 2- to 3-fold more often in conserved $3'$ UTR sequences than it would be expected at random. Moreover, because of the increased stability of the duplex they form, canonical sites seem to be more effective than other types, and may function in one copy. Conversely, due to their lower pairing energies, seed sites are expected to be more effective when present in more than one copy [22].

It has also be shown that even if seed match can confer regulation by itself, additional $3'$ pairing increases miRNAs functionality; however, a high degree of matching in $3'$-end can not be functional without a minimal element of $5'$ complementarity.

Moreover, matches between miRNAs and target mRNAs seem to follow some experimentally deduced rules:

1. complementarity between nt 2 and 8 of miRNAs and their targets is critical for target recognition by miRNAs [6] [23];

2. confirmed miRNA:mRNA duplexes can have mismatches in the $5'$-end of miRNA region [24];

3. G:U wobble base pairs are less common in the $5'$-end of a miRNA:mRNA duplex. Experiments show that a single G:U base-pair considerably reduces the effectiveness of a seed site, while the presence of more than one G:U base-pair compromises the activity of all the sites [22];

4. many miRNAs can bind to one gene (cooperativity of binding) [25] and target sites may overlap to some degree [26];

5. bases surrounding the seed sequence are important for target recognition [27];

6. A:T terminal match is often observed in miRNA target sites [27].

[6] This observation led to the definition of the seed sequence.

These are the main rules which are taken into account in the development of target prediction algorithms (Chapter 3).

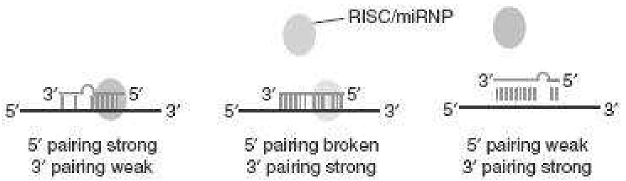


Figure 1.11: Depending on the type of base pairing with the seed region, miRNA target sites are classified in: (*left*) $5'$-dominant canonical, if there is perfect base pairing to at least the seed portion of the $5'$-end of the miRNA and extensive base pairing to the $3'$-end of the miRNA; (*centre*) $5'$-dominant seed only, if base pairing is perfect at least to the seed portion of the $5'$-end of the miRNA while it is limited to the $3'$-end of the miRNA and (*right*) $3'$-compensatory, if there is extensive base pairing to the $3'$-end of the miRNA to compensate for an imperfect or a shorter stretch of base pairing to the seed portion of the miRNA (adapted from [28]).

1.6 Biological functions of miRNAs

Although the functions of very few miRNAs have been worked out in detail, their highly conservation and the developmental-stage-specific and/or tissue-specific expression patterns suggest that miRNAs are critical regulators of physiologic processes. The few characterized miRNAs allow the attribution to these molecules of a significant role in regulating development and basic cellular pathways that control differentiation, proliferation and death [19].

Since the silencing pathway plays a crucial role, it is believed that miRNAs (and in general all the small RNAs) are important for the preservation and the management of the genome; fore example some miRNAs seem to interact directly with the DNA blocking the transcription of some genes.

A famous example of growth regulation by miRNAs is the *bantam* gene discovered in *Drosophila* because of its effect on tissue growth: tissues are larger when bantam is overexpressed and smaller, even without loss of proportions, when bantam expression is suppressed. Now, it is known that the bantam gene does not encode for a protein but for a miRNA that control the proapoptotic gene *hid*; thus, bantam promotes proliferation while inhibiting apoptosis [16].

Recent works discovered many viral-encoded miRNAs, whose function is still not documented; however, it is emerging that viral evolution has taken advantage of the miRNA pathway to generate effectors that enhance the probability of successful infection [15].

1.6.1 miRNAs and diseases

Although the current evidence is limited and hundreds of miRNAs of unknown functions exist, it is becoming clear that miRNAs regulate pathways such as cellular differentiation, proliferation, and apoptosis that are known to be disregulated in human malignancies. Consistent with these studies, reports of altered miRNA expression in human cancers (brest and lung cancer, to give some examples) are beginning to emerge in literature [19]. Recently, it has become possible to analyze the entire *miRNome* by microarrays containing all known human miRNAs; it has been shown that miRNAs are over-expressed in cancer, suggesting their role as a novel class of oncogenes or tumor suppressor genes [15].

Important information about the role of a miRNA in a disease can be deduced by mimicking or inhibiting its activity and examining its impact on the phenotype/behaviour of the cell or organism. If the modulation of the activity of a miRNA leads to improvement in disease symptoms, this implies that the target miRNA plays an important role in the disease.

Thanks to the discovery of the role of miRNAs in various pathological processes it is possible to develop miRNA-based therapeutic products that can either increase or decrease the levels of proteins in pathophysiological conditions such as cancer, cardiovascular diseases, viral diseases, metabolic disorders and programmed cell death [29]. To give an example, target oncogenes can be regulated by chromatin modifying drugs, and this may lead to novel cancer therapies in the future; moreover, miRNAs can complement other genomic and proteomic biomarkers of cancer diagnosis and prognosis [30]. Herein lies the therapeutic potential of miRNAs, as it may now be possible to induce or inhibit RNA interfering in a given diseased cell population by controlling the miRNA expression profile of the cells [31].

Since the benefit of understanding the role of miRNAs in cancer and other diseases is potentially enormous, especially if it helps in providing new therapies for patients, a combination of high-throughput target analysis with genomics and proteomics in order to gain an accurate prediction of target genes could lead to applications in clinical research.

Chapter 2

Computational tools and resources

The discovery of novel miRNAs, the characterization of their biogenesis and the identification of their functions are subjects of many research projects and there are several existing tools and resources that provide updated information regarding each of these areas of research.
This Chapter contains a description of the main computational tools and resources useful for the purpose of studying miRNAs (Table 2.1).

2.1 miRGen

The aim of the miRGen database is to integrate data about:

- positional relationships between animal miRNAs and genomic annotation sets;
- animal miRNA targets according to combinations of widely used target prediction programs.

The database consists of three integrated interfaces representing 11 genomes [1]. The *genomics* interface allows the user to explore where whole-genome collections of miRNAs are located with respect to UCSC genome browser (Section 2.5) annotation sets. The *targets* interface allows to connect these miRNAs to their experimentally supported target genes from TarBase (Section 2.2), as well as to computationally predicted target genes from optimized intersections and unions of several widely used mammalian target prediction programs (Chapter 3). Finally, the *clusters* interface provides predicted miRNA clusters at any given inter-miRNA distance.

[1] Last genome data update : January 1, 2007.

2.2 TarBase

Tarbase collects a comprehensive set of experimentally supported miRNA targets in at least 8 organisms. For every target site that has experimental support, TarBase describes the miRNA which binds to it, the binding (alignment) picture, the kind of inhibition that it induces, its single site sufficiency status, its genomic location and the types of experiments that were conducted to support it; moreover, the papers from which all of these data were extracted can be found.
It is also possible to submit new targets in order to keep the resource updated; nowadays, the database contains 128 miRNAs, 570 target genes and 763 target sites.

2.3 MirBase

The miRBase Sequence Database is a searchable database of published miRNA sequences and annotations, whose data were previously provided by the miRNA Registry.
This resource contains three main sections:

- Sequences: a searchable database of published miRNA sequences and annotations. It contains all published miRNA sequences, genomic locations and associated annotations;
- Targets: a database of predicted miRNA target genes;
- Registry: provides a confidential service assigning official names to novel miRNA genes prior to the publication of their discovery.

2.4 Gene Ontology

Since the knowledge of gene and protein roles in cells is accumulating and changing, the goal of the Gene Ontology (GO) Consortium is to produce a dynamic, controlled vocabulary for cell biology [32]. For this purpose, the GO project has developed three ontologies that describe gene products in terms of their associated biological processes [2], cellular components and molecular functions [3] in a species-independent manner. Queries to the controlled vocabularies can be done

[2] A biological process is defined as a series of events accomplished by one or more ordered assemblies of molecular functions; for example, the pyrimidine metabolism.

[3] Molecular functions are activities, such as catalytic or binding activities, that occur at the molecular level.

at different levels: it is possible, for example, to find all the gene products in the mouse genome that are involved in signal transduction.
The terms in an ontology are linked by two types of relationships:

- *is_a*: it is a class-subclass relationship, where A is_a B means that A is a subclass of B; for example, nuclear chromosome is_a chromosome;
- *part_of*: C part_of D means that whenever C is present, it is always a part of D, but C does not always have to be present. An example could be nucleus part_of cell; nuclei are always part of a cell, but not all cells have nuclei.

The ontologies are structured as Directed Acyclic Graphs (DAGs), which are similar to hierarchies but differ in that a child term can have many parent terms. For example, the biological process term hexose biosynthesis has two parents: hexose metabolism and monosaccharide biosynthesis.

2.5 UCSC Genome Bioinformatics

The UCSC Genome Bioinformatics tool shows any requested portion of the human genome at any scale and displays for example assembly contigs and gaps, mRNA and expressed sequence tag alignments, multiple gene predictions, cross-species homologies, single nucleotide polymorphisms and so on. Links to sequence details and supplementary off-site databases are also provided [33].
In particular, the following utilities are present:

- *The Genome Browser*: zooms and scrolls over chromosomes;
- *The Gene Sorter*: shows expression, homology and other information about on groups of genes;
- *Blat*: quickly maps a sequence to the genome;
- *The Table Browser*: provides convenient access to the underlying database;
- *VisiGene*: lets the user browse through a large collection of in situ mouse and frog images to examine expression patterns;
- *Genome Graphs*: allows to upload and display genome-wide data sets.

Table 2.1: Computational tools and resources.

Resource	Website	References
miRGen	http://www.diana.pcbi.upenn.edu/miRGen	[36]
Tarbase	http://www.diana.pcbi.upenn.edu/tarbase.html	[37]
miRBase	http://microrna.sanger.ac.uk	[38][39]
GO the Gene Ontology	http://www.geneontology.org	[32]
UCSC	http://http://genome.ucsc.edu	[33]
e! Ensembl	http://www.ensembl.org	[34]
UTResource	http://bighost.area.ba.cnr.it/BIG/UTRHome	[35]

2.6 Ensembl

Ensembl is a project that wants to provide an accurate, automatic analysis of genome data, whose annotation is currently maintained, concentrating on vertebrate genomes; the result is a comprehensive and integrated source of annotation of chordate genome sequences (33, nowadays). Some of the major features recently added include the first complete gene sets for genomes with low-sequence coverage, the introduction of new strain variation data and also of new orthologous/paralogous annotations based on gene trees [34].

2.7 UTRdb

UTRdb is a specialized database of $5'$- and $3'$-untranslated sequences of eukaryotic mRNAs cleaned from redundancy. UTRdb entries are enriched with specialized information not present in the primary databases, including the presence of functional patterns already demonstrated by experimental analysis to have some functional role. A collection of such patterns is being collected in UTRsite database which can also be used with appropriate computational tools to detect known functional patterns contained in mRNA untranslated regions [35].

Chapter 3

Mammalian target prediction algorithms

As reported in Chapter 1, miRNAs can mediate post-transcriptional regulation by the RNA-interference pathway, which decides for translational inhibition or for the cleavage of the target transcript. Hundreds of miRNAs are known in complex eukaryotes and genome-wide statistical analysis showed that usually miRNAs have hundreds of evolutionarly conserved target sites, indicating that miRNAs regulate a large fraction of protein-coding genes [22].
Recent studies tend to a systematic bioinformatic identification of animal miRNA targets; here we focus on mammalian target prediction.

3.1 Shared features of target prediction algorithms

To date, most studies on miRNA target interactions have focused primarily on the characteristics of sequence complementarity between the miRNA and the putative target sites in the mRNA. Although perfect Watson-Crick base pairing of seven or eight consecutive bases in the $5'$-ends of miRNAs is important for target regulation, animal miRNAs are typically only partially complementary to their targets: bulges and loops are in fact frequent. Therefore, due to the relatively short length of miRNAs and the tolerance for mismatches, nearly all the predictive algorithms (Table 3.1) are based on phylogenetic conservation of the predicted duplexes.
More recently, new studies incorporate site accessibility in the identification of miRNA targets (Section 3.2.10). This approach might encompass more authentic targets and identify species-specific targets.

Table 3.1: The considered mammalian target prediction algorithms and their availability.

Program	Website	References
DIANAmicroT	http://www.diana.pcbi.upenn.edu	[40]
miRanda [a]	http://www.microrna.org	[25] [41]
PicTar [a]	http://pictar.bio.nyu.edu	[42]
PITA [a]	http://genie.weizmann.ac.il/pubs/mir07	[43]
TargetScan [bc]	http://www.targetscan.org	[27] [23]

[a] *Precompiled lists of predictions are also provided.*

[b] *All the versions are available.*

[c] *Precompiled lists of predictions are also provided for all the versions.*

3.2 Predicting miRNA targets

3.2.1 miRanda

The software miRanda uses a modified dynamic programming approach that recognizes the importance of seed binding (but does not require perfect seed complementarity), searching for maximal local complementarity corresponding to a double-stranded antiparallel duplex.

The algorithm allows for G:U basepairs along with certain mismatches and/or bulges in $5'$ region in the context of dinamically weighted pairing [28]. A score of $+5$ is assigned for G:C and A:T pairs, $+2$ for G:U wobble pairs, -3 for mismatches pairs, -8 and -2 for the gap-open and gap-elongation, respectively. Complementarity scores at the first 11 positions, counting from the miRNA $5'$-end, are multiplied by a scaling factor of 2.0 trying to reflect the experimentally observed $5' - 3'$ asymmetry. The total score S is defined as the sum of single-residue-pair match scores over the alignment trace.

The RNA folding program described in [44] [45] is used to consider thermodynamic properties calculating ΔG, the free energy of duplex formation from a completely disassociated state. The thresholds for candidate target sites are $S > 90$ and $\Delta G < -17$ kcal/mol [41] [25].

Conservation is defined as an entire target site occurring at least 90% identity at exactly corresponding positions in a cross species UTR alignment. The algorithm looks for conservation of target site position and sequence in the $3'$ UTRs of orthologous genes; conservation is in particular required only between human and rodent [41] [25], providing the additional option of using extensive conservation (more than two species).

3.2.2 DIANAmicroT

As in miRanda, in DIANAmicroT a modified dynamic programming algorithm is implemented to calculate free energies of both canonical (Watson-Crick) and G:U wobble dinucleotide base pairs. An empirically determined set of binding rules is required allowing mismatches, that is to say that strong base pairing to a miRNA seed region is required, but perfect seed complementarity is not necessary.

The miRNA:mRNA duplex can show a single miRNA-recognition element or miRNA central bulge or a loop; moreover the $5'$ first nucleotides (and the last nucleotides toward the $3'$-end) of the miRNA may or may not base pair with their targets. Binding energies are evaluated for every three consecutive nucleotide pairs [40].

The definition of conservation is the same used in miRanda algorithm, and it is required only between human and rodent.

3.2.3 TargetScan 1.0

Given a miRNA that is conserved in multiple organisms and a set of orthologous $3'$ UTR sequences from these organisms, the algorithm:

1. searches for perfect seed matches to the UTRs in the first organism;

2. extends each seed match with additional base pairs to the miRNA as far as possible in each direction, allowing G:U pairs, but stopping at mismatches;

3. optimizes base pairing of the remaining $3'$ portion of the miRNA to the 35 bases of the UTR immediately $5'$ of each seed match using the RNAfold program [1]. The seed match is therefore extended to a longer target site, allowing the presence of bulges (Figure3.1);

4. assigns a folding free energy G (kcal/mol) to each miRNA:mRNA duplex using the RNAeval function [47];

5. assigns a Z score to each UTR defined as $Z = \sum_{k=1}^{n} e^{\frac{-G_k}{T}}$ where n is the number of seed matches in the UTR and G_k is the free energy of the duplex for the k^{th} target site. The value of T parameter influences the relative weighting of UTRs with fewer high-affinity target sites to those with larger numbers of low-affinity sites;

[1] RNAfold program can be found in the Vienna RNA secondary structure server [44], which provides a web interface to the most frequently used functions of the Vienna RNA software package for the analysis of RNA secondary structures. Requiring as input a single sequence, RNAfold predicts the minimum free energy structure of a single sequence using a classic algorithm [46].

6. sorts the UTRs in this organism by Z score and assigns a rank R_i to each;
7. repeats this process for the set of UTRs from each organism;
8. predicts as targets those genes for which both $Z_i \geqslant Z_c$ and $R_i \leqslant R_c$ for an orthologous UTR sequence in each organism, where Z_c and R_c are presetted Z score and rank cutoffs.

The only free parameters in this algorithms are T, Z_c and R_c [23]. Conservation is required among human, chimp, rodent and dog.

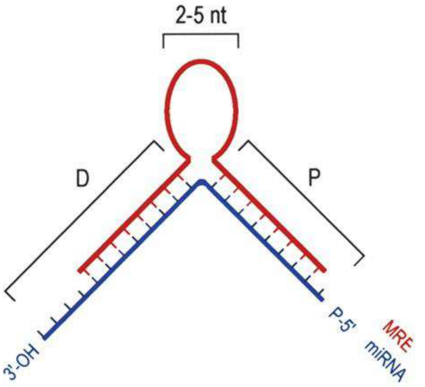


Figure 3.1: Schematic representation of a *bulge*, one of the possible binding types between the miRNA (blue) and its target (red) [40]. Because of the possible presence of this type of conformation, the target site can be longer than the miRNA; for this reason, some algorithms look for base pairing of additional flanking nucleotides (adapted from [40]).

3.2.4 TargetScan 2.0

The algorithm requires target site conservation among human, mouse, rat, dog and chicken; from the previous versions, the noise (estimated number of false-positive predictions) is reduced, such that the TargetScan score and rank cutoffs can be relaxed, or even eliminated. Moreover, the requirement of a 7-nt match to the seed region of the miRNA (nucleotides 2-8) is relaxed to require a 6-nt match to a reduced seed comprising nucleotides 2-7 of the miRNA while still retaining modest specificity. As a result, a target is predicted simply by virtue of the presence of at least one 6-nt seed match to the miRNA in orthologous UTRs of each of the five mentioned genomes [27].

3.2.5 TargetScan 3.0

This version of TargetScan displays predicted regulatory targets of mammalian miRNAs, which are mostly the same as those predicted by the previous ones. Changes include a user-friendly interface and the possibility to predict additional miRNA families.

Nonconserved sites, which can often mediate repression when the miRNA and the mRNA are both present in the same cell, are annotated.

Moreover, the program provides the depiction of the UTRs (showing the position of each site), of the local alignment surrounding each site and of the predicted pairing between each miRNA and each site.

3.2.6 TargetScan 3.1

In this release the following changes are introduced since version 3.0:

- target conservation is based on overlapping targets in the UTRs of four species;
- the start and the end of each species no longer need to exactly match, removing genome alignment artifacts.

3.2.7 TargetScanS

TargetScanS requires the conserved seed matches to be at conserved positions within the UTRs, focusing only on an 8-nt segment of the UTR centered on the seed match, without consideration of other criteria, such as predicted thermodynamic stability of pairing, pairing outside the immediate vicinity of the seed, or presence of multiple complementary sites per UTR.

The algorithm defines a seed as positions 2-7 of a mature miRNA; a miRNA family is comprised of miRNAs with the same seed region (positions 2-8 of the mature miRNA, also called seed+m8). The algorithm predicts biological targets of miRNAs by searching for the presence of conserved 8mer [2] and 7mer sites [3] that match the seed region of each miRNA.

[2] Exact match to positions 2-8 of the mature miRNA (the seed + position 8) with a downstream 'A' across from position 1 of the miRNA.

[3] 7mer sites can be divided in two classes. *7mer-m8*: showing an exact match to positions 2-8 of the mature miRNA (the seed + position 8); *7mer-1A*: showing an exact match to positions 2-7 of the mature miRNA (the seed) with a downstream 'A' across from position 1 of the miRNA.

3.2.8 TargetScan 4.0

Targets are predicted using the TargetScanS algorithm; a context score is also evaluated [48] in order to predict site performance taking into account the context in which the target is inserted on the mRNA, which seems to influence significantly the effectiveness of the site.
Sites within 15 nt of a stop codon are typically less effective than other sites in the $3'$ UTR; the algorithm flags this type of sites, analyzes the context around and gives a score, considering the following four features [48]:

- site-type contribution: reflects the type of seed match (8mer, 7mer-m8, and 7mer-1A);

- $3'$ pairing contribution: reflects consequential miRNA-target complementarity outside the seed region. One of the important context determinants that influences efficacy of sites is in fact their proximity to sites for coexpressed miRNAs;

- local AU contribution: reflects transcript AU content 30 nt upstream and downstream of predicted site, because effective sites seem to reside within a locally AU-rich context;

- position contribution: reflects distance to nearest end of annotated UTR of target. Experiments show that, to be more effective, sites in the $3'$ UTR are not too close to the stop codon. Moreover, sites near the two ends of long UTRs are generally more effective than those near the centre and more sites are selectively maintained near the ends than in the central region.

More negative is the score, more favorable is the associated site; the sum of the scores gives the context score, and the context score percentile is the percentile rank of each site compared to all sites for this miRNA family. An high value of this percentile (between 50 and 100) means that the specific site is more favorable than most other sites of this miRNA.
In a gene with multiple sites for one miRNA family, a total context score is evaluated as the sum of context scores for the most favorable (most negative) miRNA in this family. The representative miRNA is the one in this set with the most favorable total context score.
MiRNA families have been updated to include recent additions to miRBase; also the $3'$ UTR dataset has been updated to include chicken and to be based on current RefSeq annotation and genome coordinates.
Conservation is imposed as in [27] and includes three levels:

- highly conserved = conserved across human, mouse, rat, dog, and chicken;

- conserved = conserved across human, mouse, rat, and dog;
- poorly conserved = not conserved across all four mammals.

3.2.9 PicTar

PicTar defines the *perfect seed* as a perfectly Watson-Crick-base-paired stretch of 7 nt starting at either the first or the second base of the miRNA (counted from the $5'$-end). The sequence of a perfect seed can vary, but its free energy of binding (determined by standard RNA secondary structure prediction software) can not increase and it can not contain G:U basepairings; these mutated seeds are called *imperfect nuclei*.

The algorithm also requires the free energy of the miRNA:mRNA duplex to lie below a cutoff value. A probability p to be a binding site for the miRNA is evaluated for the perfect seeds which pass the filtering steps; the probability for imperfect nuclei is $1 - p$ divided by the total number of imperfect nuclei [42].

Target genes containing several predicted binding sites are considered to be more probably downregulated by miRNAs.

PicTar also requires target conservation across several species: human, chimpanzee, mouse, rat, dog, chicken, and pufferfish.

3.2.10 PITA

This thermodynamic model scores miRNA:mRNA interactions by the difference between the energy of the bond (ΔG_{duplex}) [4] and the energy required to make the target region accessible for miRNA binding (ΔG_{open}); this difference represents an energy score and it is defined as $\Delta\Delta G$.

The ΔG_{duplex} value is computed selecting the minimum free energy structure; ΔG_{open} is defined as the difference between the free energy (computed using RNAfold) of the ensemble of all secondary structures of the target region and the free energy of all target-region structures in which the target nucleotides (and 70 additional nucleotides [5] upstream and downstream in the case of flanking) are required to be unpaired (Figure 3.2). The total interaction score $\Delta\Delta G$ is equal to the difference between ΔG_{duplex} and ΔG_{open}.

[4] MiRNA and target are paired according to pairing constraints imposed by the seed, which is a sequence of $6 - 8$ bases, beginning at position 2 of the miRNA. No mismatches or loops are allowed, but a single G:U wobble is allowed in 7- or 8-mers.

[5] The value of 70 nucleotides is based on the fact that there is a low probability of secondary structure base-pairing interactions between nucleotides that are separated by more than 70 nucleotides.

An overall miRNA:UTR interaction score (T) is also evaluated to consider multiple sites with $\Delta\Delta G$ scores $s_1, \ldots, s_n$ for one miRNA on the same UTR; T is defined as the statistical weight of all configurations in which exactly one of the sites is bound by the miRNA according to $T = log \sum_{i=1}^{n} e^{s_i}$.
This simple method of integrating multiple sites over computing the actual probability of miRNA binding, and over computations that include configurations in which two sites can be bound simultaneously, is chosen because such computations would require knowledge of an additional (unknown) miRNA free concentration parameter [43].

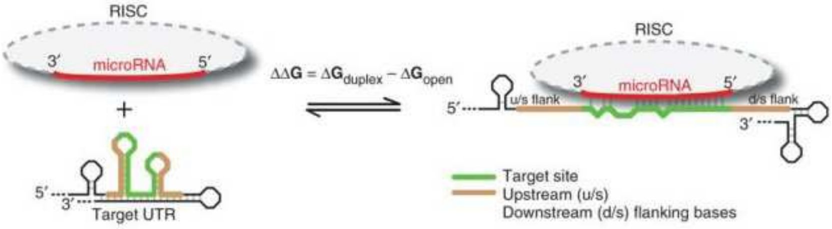


Figure 3.2: PITA algorithm explains variability in target strength due to differences in accessibility. $\Delta\Delta G$ is computed as the free energy gained by transitioning from the state in which miRNA and target are unbound (*left*) to the state in which the miRNA binds its target (*right*). The region of the target site that needs to be unpaired to let a miRNA-target interaction occur includes the miRNA bound region (green) and likely additional flanking nucleotides (brown; adapted from [43]).

Chapter 4

Integration of miRNA target predictions

4.1 Combining target sets

Despite the blooming of target-prediction approaches (a selection of which is described in Chapter 3), it is currently not possible to accurately assess the relative rates of false negative predictions by each method compared with the rest. Evaluating sensitivity would in fact require systematic testing of targets not predicted by each method, and such data are not yet available.

In [49] the authors evaluate specificity by comparing the output of several programs on experimentally supported miRNA target interactions. No program has turned out to be consistently superior to the others.

Therefore it might be reasonable not to arbitrarily prefer one algorithm to the others. Rather, it may be helpful to consider the prediction quality of differently combined subsets of algorithms. In order to identify the optimized combination, we decide to evaluate pair-wise combinations of aforementioned programs.

Predictions can be combined by:

- *union*, searching all targets predicted by any of the listed prediction programs (it is the most sensitive search type);
- *intersection*, searching targets which are predicted by all of the prediction programs listed in the set (it can provide more specific searches).

Since we reason that unifying predictions would lead to a lower signal-to-noise ratio and thus a greater chance of false positive predictions, we prefer to limit to pair-wise intersections.

In [49] the authors adopt a similar approach. They focus on TargetScan, DIANA-microT, miRanda, TargetScanS and PicTar methods, test their possible pair-wise

combinations and evaluate the improvement achieved in each case by counting the percentage of shared targets that are experimentally supported (Figure 4.1).
Putative miRNA target interactions are far from being extensively tested in vivo. Therefore, we foresee in the current small rate of experimental validations a severe limit of the approach in [49] to assess increased specificity of a combination compared to the initial methods.
Here we describe the confidence of each pair-wise combination by evaluating the statistical significance of overlap in predictions. Turned around, we select the pair of methods that is characterized by the highest number of single miRNAs for which it is unlikely to get a number of shared predictions equal or greater to the observed one by pure chance effect.

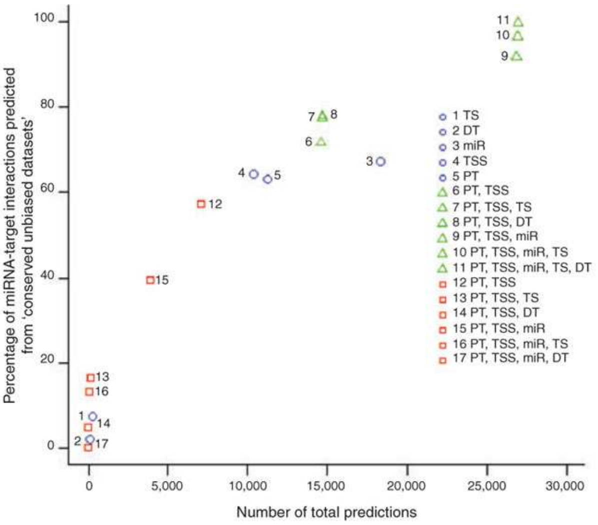


Figure 4.1: Performance spectrum of target prediction programs. The sensitivity is plotted versus the number of genes predicted as target. The blue symbols represent the individual programs: TargetScan (TS), DIANA-microT (DT), miRanda (miR), TargetScanS (TSS) and PicTar (PT). The red symbols represent the intersections, while the green ones the unions. The intersection between PicTar and TargetScanS seems to provide a good trade-off between sensitivity and specificity for human miRNA gene prediction programs (from [49]).

4.2 Selecting the best combination of algorithms

In order to identify which of the pair-wise intersections obtains the highest confidence, the following steps are followed:

Table 4.1: Human miRNAs and precursors in miRBase 9.2, which is the miRBase release we refer to (Section 2.3).

Human miRNAs and precursors	number
mature miRNA ids	471
miRNA precursors	475

1. for each miRNA (Table 4.1) the following combinations between target prediction sets are evaluated, using miRGen resource (Section 2.1):
    - miRBase [1] $\cap$ PicTar;
    - miRBase $\cap$ PITA;
    - miRBase $\cap$ TargetScan 3.1;
    - PicTar $\cap$ PITA;
    - PicTar $\cap$ TargetScan 3.1;
    - TargetScan 3.1 $\cap$ PITA.

2. For each pair-wise combination and each miRNA, the Fisher Test evaluates the statistical significance of the number of shared targets (Tables 4.2 and 4.3); this statistical test gives the probability P of obtaining an equal or greater number of genes predicted as target in a set made of the same number of genes but selected at random from the full list of predicted targets. Let J and j be the overall and miRNA-specific numbers of target prediction by one method, respectively. If $K(m)$ is the number of target of m predicted by the other method taken into account and $k(m)$ the number of targets shared by both the methods, then the probability P is given by the right tail of the hypergeometric distribution:

$$P(J, K(m), j, k(m)) = \sum_{h=k(m)}^{min(j,K(m))} F(J, K(m), j, h), \tag{4.1}$$

where

$$F(J, K, j, k) = \frac{\binom{K}{h}\binom{J-K}{j-h}}{\binom{J}{j}}. \tag{4.2}$$

[1] Predictions obtained running miRanda algorithm.

Table 4.2: Number of microRNAs which the softwares (Chapter 3) predict targets of.

Software	number of miRNAs
miRanda (miRBase)	470
PicTar	178
TargetScan 3.1	238
PITA	470

Table 4.3: Results of the pair-wise intersections of the prediction algorithms.

Intersection	miRNAs [a]	skew [b]	mean [b]	sd [b c]	median[b]
miRBase ∩ PicTar	178	1.07	23.67	16.14	20.38
miRBase ∩ PITA	470	0.86	98.63	26.78	96.07
miRBase ∩ TargetScan 3.1	235	1.29	19.60	14.47	14.94
PicTar ∩ PITA	178	0.44	84.56	54.54	77.00
PicTar ∩ TargetScan 3.1	172	−0.61	211.38	107.08	249.75
TargetScan 3.1 ∩ PITA	238	0.55	77.41	50.36	68.74

[a] *Number of miRNAs shared by the two algorithms.*

[b] *Of the distribution of the* $p - values$.

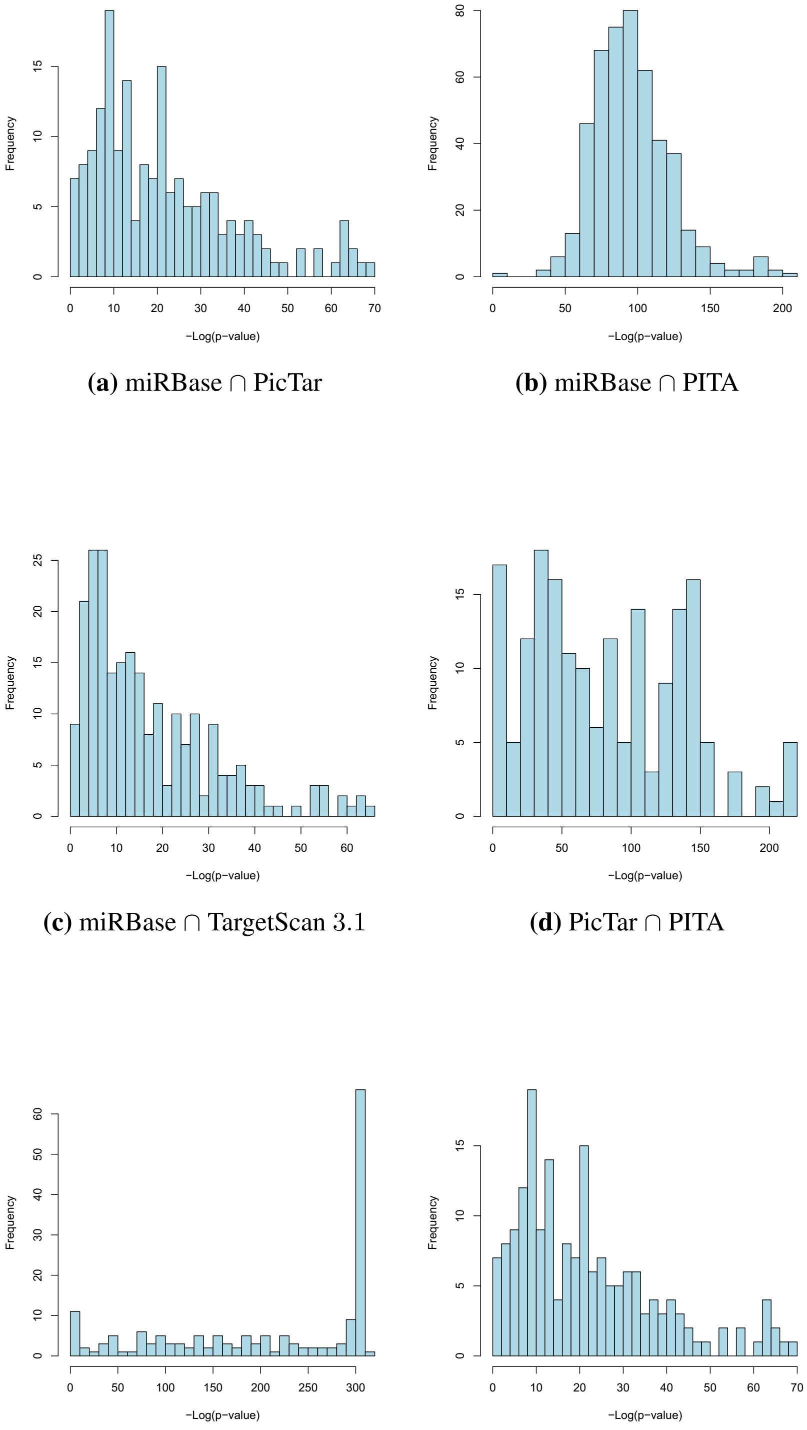
[c] *Standard Deviation.*

For each combination of algorithms, P is evaluated for the target intersection sets of all the miRNAs.

3. The combination showing the highest right skewness [2] of the distribution of $-Log(p - value)$ is selected as the best one (Figure 4.2).

Following this criterion, we find the combination of PicTar and TargetScan 3.1 (Figure 4.2e) let us be more confident than otherwise.

[2] Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable.



(a) miRBase ∩ PicTar

(b) miRBase ∩ PITA

(c) miRBase ∩ TargetScan 3.1

(d) PicTar ∩ PITA

(e) PicTar ∩ TargetScan 3.1

(f) TargetScan 3.1 ∩ PITA

Figure 4.2: Distribution of $-Log(p - value)$ of the target intersection sets.

Table 4.4: PicTar $\cap$ TargetScan 3.1: number of predicted miRNA targets.

	miRNAs [a]	min [b]	median [b]	mean [b]	max [d]	sd [b c]
PicTar $\cap$ TargetScan 3.1	172	3.0	181.5	222.4	595.0	162.7

[a] *Number of miRNAs shared by the two algorithms.*

[b] *Of the distribution of the number of predicted targets.*

[c] *Standard Deviation.*

[d] *Value corresponding to: miR-30a-5p, miR-30b, hsa-miR-30c, hsa-miR-30d, hsa-miR-30e-5p.*

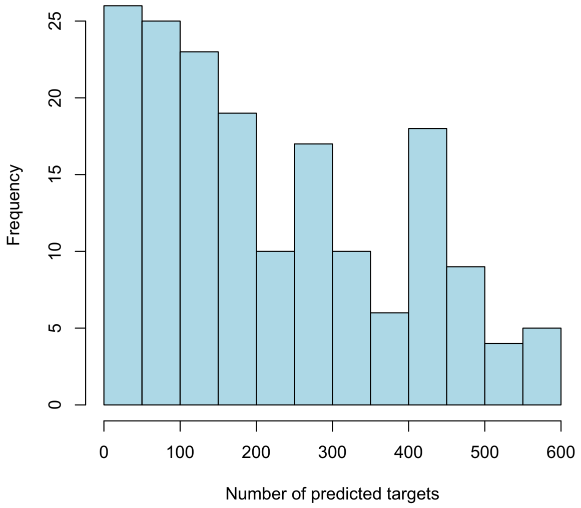


Figure 4.3: Number of targets in the intersection set PicTar $\cap$ TargetScan 3.1; according to current literature, a single miRNA is predicted to have up to hundreds of targets.

4.3 Functional characterization of miRNAs

Once identified the best combination of prediction softwares, we aim to functionally characterize miRNAs by the annotation of their targets; for this purpose, only the miRNAs whose $p - value$ is lower than 10^{-3} are selected. Bonferroni correction [3] is then applied as follows:

$$\text{cut-off on } p - value = \frac{10^{-3}}{\text{number of miRNAs}} = \frac{10^{-3}}{172} = 5.81 \times 10^{-6} \quad (4.3)$$

and

$$\text{cut-off on } - Log(p - value) = 5.24 \quad (4.4)$$

and a set of 166 miRNAs fulfills the cut-off.

Since Gene Ontology (Section 2.4) is one of the most important standardized biological vocabulary in a sharable and computationally amenable form, it serves our annotation task perfectly.

Given a set, the standard Fisher's Test evaluates the existence of any functional enrichment [4] of the set by separately testing it on all of the annotations that are held within the GO ontological structure. A set is reported as GO annotated if its corresponding $p - value$ is lower than 10^{-4} (Bonferroni-corrected). A set of 144 miRNAs shows to have shared targets between PicTar and TargetScan that are enriched in at least one GO term.

One of the main difficulties with this approach, although strictly rigorous under a statistical point of view, is that it assumes that all of GO terms have the same semantic value, which appears to be a not fully exact assumption.

Generally, we expect that the greater the distance from the root of the GO graph, the more semantically meaningful and biologically specific the terms are. However, GO varies widely in the distance of nodes from the root; GO terms more semantically precise can occur to lay less deep than less informative GO terms. It would appear that the depth of GO reflects mostly the vagaries of biological knowledge, rather than anything intrinsic about the terms.

Nonetheless, any critical interpretation of our annotation outcomes can not help to identify the annotations that are the richest in information content.

For this purpose, attempting to define among all of the GO terms resulted over-represented in a given set exclusively the deepest GO terms, does not seem to us

[3] The Bonferroni correction is a multiple-comparison correction used when several dependent or independent statistical tests are being performed simultaneously (since while a given cut-off may be appropriate for each individual comparison, it is not for the set of all comparisons). In order to avoid a lot of spurious positives, the cut-off needs to be lowered to account for the number of comparisons being performed; the simplest and most conservative approach is the Bonferroni correction, which sets the cut-off for the entire set of comparisons equal to by taking the cut-off for each comparison equal to cut-off/number of tests.

[4] Fisher's Test (Section 4.2) is applied in order to evaluate functional enrichment. In this case, J and j are the number of genes in the genome and the number of genes in the genome annotated in the considered GO function, respectively; $K(m)$ is the number of targets belonging to one of the intersection sets, while $k(m)$ represents the number of these targets annotated in the considered GO function.

an effective approach. Instead, we prefer to inspect all the *is_a* links for each GO term and keep the GO terms for which more than half of the parent GO terms are over-represented as well. This approach should manage to identify the most robust and biologically consistent annotation outcomes.
Let us stress that this choice does not aim at providing GO terms with additional statistical value; rather, it let us assess the consistency of our outcomes within the GO annotation structure. A set of 122 miRNAs fulfilled this last requirement.

4.4 Significant results

We can say that our functional characterization of miRNA targets mostly agrees with the current available literature. In our analysis miRNAs are in fact shown to control fundamental cellular processes, such as differentiation of cells and timing of development of the organisms, being GO functions concerning *binding*, *regulation* and *development* frequently targeted (Table 4.5); we have to precise that in order to have a comparable number of genic associations, GO functions we refer to are at quite at the same level of the ontological tree (Figure 4.5).
Since their discovery, the emerging characterization of miRNAs suggested that their aberrations could be involved in various human diseases, including cancer; afterward, more than half of the known miRNAs have been reported to be located in cancer-associated genomic regions and to show copy number alterations in cancer [50] [51], and deregulation of several miRNAs has been detected in various cancers [52] [53]. Although targets of most miRNAs have not been identified, functional studies of individual miRNAs have shown that they can negatively regulate well-known oncogenes [30]; according to [54] [55] [56] which describe miR-15a and miR-16 as repressors of the antiapoptotic factor BCL2 involved in the development of the ovarian cancer and implicated in chronic lymphocytic leukemia, BCL2 gene is predicted to be target of these two miRNAs in our analysis, too. In [50] miR-15 and miR-16 are shown to be located at chromosome the site 13q14, a region deleted in more than half of B-cell chronic lymphocytic leukemia (CLL); detailed deletion and expression analysis showed that these miRNAs are located within a 30 kb region of loss in CLL, and that both genes were deleted or downregulated in approximately 68% of CLL cases [57].
It is also interesting to point out that 24 miRNAs (out of 122) are predicted to target genes annotated in GO *binding* related functions. One of the considered function is *sequence-specific DNA binding*, i.e. miRNAs are predicted to target genes which code for DNA binding proteins. In regard of this, a recent work [58] has investigated the role of miRNAs in mammalian mid brain dopaminergic neurons (DNs); miR-133b is shown to be specifically expressed in mid brain DNs and deficient in mid brain tissues from patients with Parkinson's disease. In particu-

lar [58] suggests that miR-133b can regulate the maturation and function of mid brain DNs within a *negative feedback circuit* that includes the transcription factor Pitx3; a model is proposed, with miR-133b functioning within a negative feedback circuit that normally suppresses Pitx3 expression post-transcriptionally, and, in turn, Pitx3 activates mid brain DN gene expression and induces transcription of miR-133b (Figure 4.4).

The available studies on miR-17-5p and miR-20a are also confirmed in our study, since they are predicted to target the E2F1 gene. In [59] the transcription factor E2F1 is shown to be a target of c-Myc [5], a protein that promotes cell cycle progression. As reported in [59] expression of E2F1 is negatively regulated by miR-17-5p and miR-20a. These findings reveal a mechanism through which c-Myc simultaneously activates E2F1 transcription and limits its translation, allowing a tightly controlled proliferative signal.

Moreover, the transcription factor Hand2, that promotes ventricular cardiomyocyte expansion, is predicted to be a target of miR-1 [60]; an excess of miR-1 in the developing heart is shown to lead to a decreased pool of proliferating ventricular cardiomyocytes. As suggested in this work, miR-1 genes titrate the effects of critical cardiac regulatory proteins to control the balance between differentiation and proliferation during cardiogenesis. According to [60], our analysis also leads to predict Hand2 gene as a target of miR-1.

Simply analyzing some predicted targets (by consulting UCSC Genome Bioinformatics (Section 2.5)) and looking for literature about the current research on them, we can say that our combination of predictions agrees with the reported miRNA/target involvement in processes of growth, development, learning, memory establishment and so on. An imperfect working of miRNAs machinery in regulating the expression of these genes could reasonably be the cause of diseases.

To give an example, hsa-miR-206 is predicted to target the mRNA of SNAP-25, which is a presynaptic plasma membrane protein involved in the regulation of neurotransmitter release and playing an integral role in synaptic transmission. Recent studies have suggested a possible involvement of SNAP-25 in learning and memory [61], both of which are key components of human intelligence.

The LHX8 and KLF11 genes are predicted to be targets of hsa-miR-30b .The first of them codes for a transcriptional regulator that is preferentially expressed in germ cells, it is critical for mammalian oogenesis and it is connected with premature ovarian failure [62]; the other gene (KLF11) is referred to be a tumor suppressor inactivated in pancreatic cancer [63].

[5] The proto-oncogene c-Myc encodes a transcription factor that regulates cell proliferation, growth and apoptosis; dysregulated expression or function of c-Myc is one of the most common abnormalities in human malignancy [59].

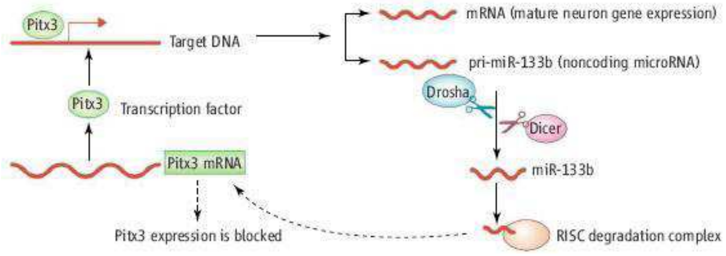
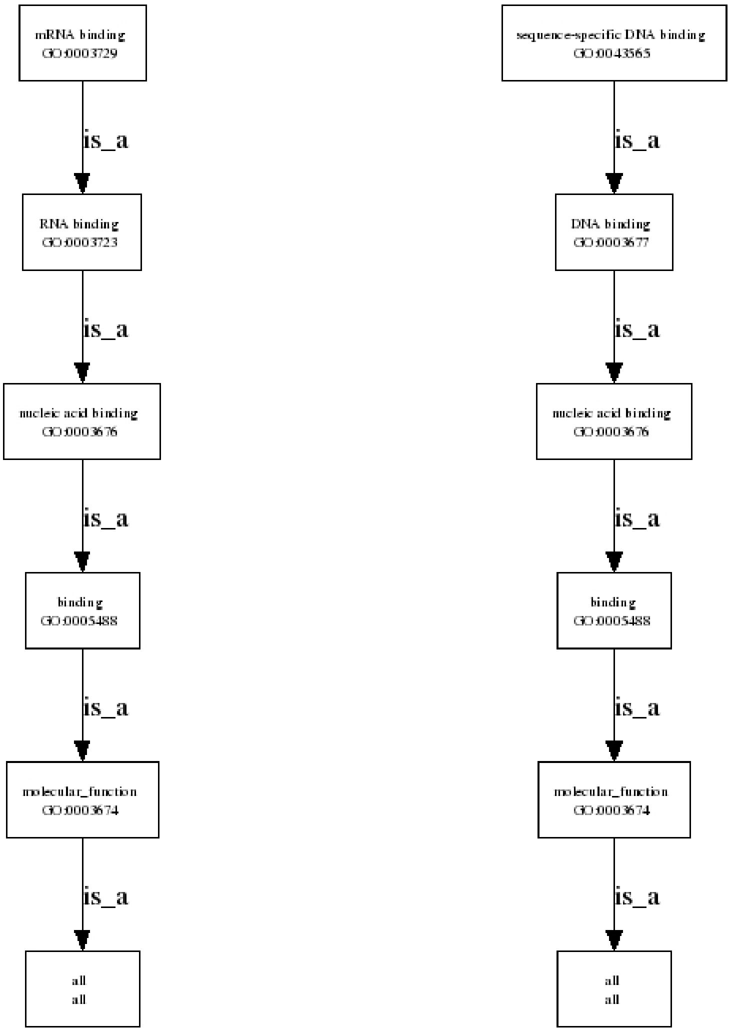

Pitx3
Target DNA
mRNA (mature neuron gene expression)
pri-miR-133b (noncoding microRNA)
Pitx3
Transcription factor
Drosha
Dicer
Pitx3 mRNA
miR-133b
Pitx3 expression is blocked
RISC degradation complex

Figure 4.4: An autoregulatory feedback loop composed of the transcription factor Pitx3 and miR-133b is implicated in dopaminergic neuron maturation and survival in the brain. miR-133b is deficient in the mid brain of Parkinsons disease patients and in mouse models of dopamine neuron deficiency (from [64]).

Table 4.5: GO functions of predicted miRNA target genes.

GO function	number of miRNAs [a]
Binding	24
Regulation	10
Development	5

[a] *Number of miRNAs (out of* 122*) with predicted target genes annotated in the considered GO functions.*



(a) GO:0003729

(b) GO:0043565

Figure 4.5: GO functions at the same level of the ontologic tree show a comparable number of genic associations. To give an example, the GO functions (4.5a) *mRNA binding* and (4.5b) *sequence-specific DNA binding* have 431 and 486 annotated genes, respectively.

Discussion

As we have seen in this work, obtaining an accurate prediction of miRNA target genes can answer to more than one challenge; besides of representing a challenge for bioinformatics because of the lack of certain rules which the algorithms can be based on, a good approximation of target sets can lead to the functional characterization of miRNAs. Because and above all of that, this type of prediction can lead to applications in clinical research.
We have also shown how the importance of the goal and the use of experimentally deduced rules is resulting in a blooming of softwares; since any of them turns out to be consistently superior to the others, we suggested to consider the predictions of differently combined (by intersection) subsets of programs describing the confidence of each pair-wise combination by evaluating the statistical significance of overlap in the predictions.
Once identified the best combination of prediction softwares we suggested a functional characterization of miRNAs by the annotation of their targets. In order to realize the most solid and biologically consistent annotation we preferred to select among all of the GO terms resulted over-represented in a given set only the terms for which more than half of the GO terms that subsume it are over-represented as well.
Our functional characterization of miRNA targets mostly agrees with the current available literature; also in our analysis miRNAs are indeed shown to control fundamental cellular processes, such as differentiation of cells and timing of the organisms development, being GO functions concerning *binding*, *regulation* and *development* frequently over-represented. Moreover, the selected intersection of predictions keeps in the target set some significant genes which are known to be involved in the progression of some diseases.

Even if we have focused on post-transcriptional regulation of mRNA by the action of miRNAs, we have to mention that regulation by proteins is the classical mean for this type of activity, while repression by miRNAs has been the topic of more recent studies. It is supposed that miRNAs and regulatory proteins can interact to regulate certain mRNAs. We are convinced that integration can be the key to gain prediction results as close as possible to the real biological scenario; in this work we suggested to integrate prediction softwares, but we think that informations about binding proteins, sequence and so on have to be added.

We have also to mention that alternative prediction methods are candidates in literature; to give an example, machine learning algorithms have been proposed in order to create classifiers that capture the characteristics of verified examples to determine whether genomic hairpins are similar to verified miRNA genes or if message $3'$ UTRs possess known target characteristics [65]. A support vector machine (SVM) classifier (one of the most popular machine learning algorithms which has good performance in classification problems) for miRNA target gene prediction is proposed in [66]; the idea is to gather all of the known miRNA-target associations, to determine a certain number of features describing these associations and to build a statistical model that would fit these features. The SVM presented in [66] considers three types of features: structural, thermodynamic and position-based. Although the first two categories correspond to properties already described, the position-based features attempt to describe more accurately the mechanism of the seed pairing; the introduction of these position-based features could potentially increase the specificity SVM predictions. Since requiring a large negative training set, which is not currently available for miRNA targets, SVMs are promising solutions which will increase their reliability and efficiency with the availability of additional data.

Bibliography

[1] The RNA revolution. Biology's Big Bang. *The Economist*, July 2007. Figure at page *i* from http://www.economist.com/opinion/displaystory.cfm?story_id=9339752.

[2] Greenbaum D., Colangelo C., Williams K., and Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.*, 4(9):117, 2003.

[3] Angela Re. *Two applications of complex systems methods to cell biology*. PhD thesis, University of Torino, 2007.

[4] Elisabeth J. Chapman and James C. Carrington. Specialization and evolution of endogenous small RNA pathways. *Nature*, 8:884–896, 2007.

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

[6] Hamilton A., Voinnet O., Chappell L., and Baulcombe D. Two classes of short interfering RNA in RNA silencing. *EMBO J.*, 21(17):4671–4679, 2002.

[7] Christian Matranga and Phillip D. Zamore. Small silencing RNAs. *Current Biology*, 17(18).

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

[9] Arnold Berk, Chris A Kaiser, Harvey Lodish, James Darnell, Lawrence Zipursky, Matthew P Scott, Monty Krieger, and Paul Matsudaira. *Molecular Cell Biology*. W H Freeman, fifth edition, 2003.

[10] Jack D. Keene. RNA regulons: coordination of post-transcriptional events. *Nature Reviews Genetics*, 8:533–543, 2007.

[11] Morten Lindow and Jan Gorodkin. Principles and limitations of computational microRNA gene and target finding. *DNA and Cell Biology*, 26(5):339–351, 2007.

[12] C. Presutti, J. Rosati, S. Vincenti, and S. Nasi. Non coding RNA and brain. *BMC Neuroscience*, 7(suppl. 1).

[13] V. Narry Kim. Small RNAs: Classification, biogenesis and function. *Molecules and Cells*, 19(1).

[14] Ramesh S. Pillai. MicroRNA function: Multiple mechanisms for a tiny RNA? *RNA*, 11:1753–1761, 2005.

[15] Branislav Kusenda, Marek Mraz, Jiri Mayer, and Sarka Pospisilova. MicroRNA biogenesis, functionality and cancer relevance. *Biomedical papers of the Medical Faculty of the University Palack, Olomouc, Czechoslovakia*, 150(2):205–215, 2006.

[16] D. L. Ouellet, M. P. Perron, L. A. Gobeil, P. Plante, and P. Povrost. MicroRNAs in gene regulation: when the smallest governs it all. *Journal of Biomedicine and Biotechnology*, 2006:1–20, 2006.

[17] S. Tsuchiya, Y. Okuno, and G. Tsujimoto. MicroRNA: biogenetic and functional mechanisms and involvements in cell differentiation and cancer. *Journal of Pharmacological Sciences*, 101:267–270, 2006.

[18] J. Graham Ruby, Calvin H. Jan, and David P. Bartel. Intronic miRNA precursors that bypass Drosha processing. *Nature*, 448:83–87, 2007.

[19] Joshua T. Mendell. MicroRNAs: Critical regulators of development, cellular physiology and malignancy. *Cell Cycle*, 4(9):1179–1184, 2005.

[20] Lin He and Gregory J.Hannon. microRNAs: small RNAs with a big role in gene regulation. *Nature*, 5:522–532, 2004.

[21] Ramesh S. Pillai, Suvendra N. Bhattacharyya, and Witold Filipowicz. Repression of protein synthesis by miRNAs : how many mechanisms? *TRENDS in Cell Biology*, 17(3):118–126, 2007.

[22] Brennecke J., Stark A., Russell R. B., and Cohen S. M. Principles of microRNA target recognition. *PLoS Biol.*, 3:e85, 2005.

[23] Benjamin P. Lewis, I-Hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. Prediction of mammalian microRNA targets. *Cell*, 115:787–798, 2003.

[24] Johnston R. J. and Hobert O. A microRNA controlling left/right neuronal asimmetry in *Caenorhabditis elegans*. *Nature*, 426:845–849, 2003.

[25] Bino John, Anton J Enright, Alexei Aravin, Thomas Tuschl, Chris Sander, and Debora S Marks. Human microRNA targets. *PLoS Biology*, 2(11):e363, 2004.

[26] Doench J.G. and Sharp P.A. Specificity of microRNA target selection in translational repression. *Genes Dev*, 18:504–511, 2004.

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

[28] Eric C. Lai. Predicting and validating microRNA targets. *Genome Biology*, 5(9):A 115, 2004.

[29] Jeyaseelan K., Herath W.B., and Armugam A. MicroRNAs as therapeutic targets in human diseases. *Expert Opin. Ther. Targets*, 11(8):1119–1129, 2007.

[30] Cherie Blenkiron and Eric A. Miska. miRNAs in cancer: approaches, aetiology, diagnostics and therapyregulons. *Human Molecular Genetics*, 16:R106–R113, 2007.

[31] Mathupala S.P., Mittal S., Guthikonda M., and Sloan A.E. MicroRNA and brain tumors: a cause and a cure? *DNA Cell Biol.*, 26(5):301–310, 2007.

[32] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[33] Kent W.J., Sugnet C. W., Furey T. S., Roskin K.M., Pringle T. H., Zahler A. M., and Haussler D. The Human Genome Browser at UCSC. *Genome Res.*, 12(6):996–1006, 2002.

[34] T.J.P. Hubbard el al. Ensembl 2007. *NAR*, 00:D1–D8, 2007.

[35] G. Pesole, S. Liuni, G. Grillo, and C. Saccone. UTRdb: a specialized database of $5'$- and $3'$-untranslated regions of eukaryotic mRNAs. *NAR*, 26(1):192–195, 1998.

[36] Molly Megraw, Praveen Sethupathy, Benoit Corda, and Artemis G. Hatzigeorgiou. miRGen: a database for the study of animal microRNA genomic organization and function. *Nucleic Acids Research*, 35:D149–D155, 2007.

[37] Praveen Sethupathy, Benoit Corda, and Artemis G. Hatzigeorgiou. TarBase: a comprehensive database of experimentally supported animal microRNA targets. *RNA*, 12:192–197, 2006.

[38] Griffiths-Jones S. The microRNA registry. *NAR*, 32:D109–D111, 2004.

[39] Griffiths-Jones S., Grocock RJ, van Dongen S., Bateman A, and Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. *NAR*, 34:D140–D144, 2006.

[40] Marianthi Kiriakidou, Peter T. Nelson, Andrei Kouranov, Petko Fitziev, Costas Bouyioukos, Zissimos Mourelatos, and Artemis Hatzigeorgiou. A combined computational-experimental approach predicts human microRNA targets. *Genes & Dev*, 18:1165–1178, 2004.

[41] Anton J Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S Marks. microRNA targets in *Drosophila*. *PLoS Biology*, 2(11):e363, 2004.

[42] Azra Krek, Dominic Grün, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J. Epstein, Philip MacMenamin, Isabelle da Piedade, Kristin C. Gunsalus, Markus Stoffel, and Nikolaus Rajewsky. Combinatorial microRNA target predictions. *Nature Genetics*, 37(5):495–500, 2005.

[43] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microRNA target recognition. *Nature Genetics*, 39:1278–1284, 2007.

[44] Ivo L. Hofacker. The Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431, 2003.

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

[46] M. Zuker and P. Stiegler. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids research*, 9:133–148, 1981.

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

[48] Andrew Grimson, Kyle Kai-How Farh, Wendy K. Johnston, Philip Garrett-Engele, Lee P. Lim, and David P. Bartel. microRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell*, 27(1):91–105, 2007.

[49] Praveen Sethupathy, Molly Megraw, and Artemis G. Hatzigeorgiou. A guide through present computational approaches for the identification of mammalian microRNA targets. *Nature Methods*, 3(11):881–886, 2006.

[50] Calin G.A. and Croce C.M. MicroRNA signatures in human cancers. *Nat. Rev. Cancer*, 6:857–866, 2006.

[51] Zhang L., Huang J., and Yang N. et al. microRNAs exhibit high frequency genomic alterations in human cancer. *Proc Natl Acad Sci USA*, 103:9136–9141, 2006.

[52] Michael M.Z., O' Connor S.M., van Holst Pellekaan N.G., Young G.P., and James R.J. Reduced accumulation of specific microRNAs in colorectal neoplasia. *Mol Cancer Res*, 1:882–891, 2003.

[53] einhart B.J., Slack F.J., and Basson M. et al. The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, 403:901–906, 2000.

[54] Cimmino A., Calin G.A., and Fabbri M. et al. miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc Natl Acad Sci U S A*, 102:13944–13949, 2005.

[55] Marilena V. Iorio, Rosa Visone, Gianpiero Di Leva, Valentina Donati, Fabio Petrocca, Patrizia Casalini, Cristian Taccioli, Stefano Volinia, Chang-Gong Liu, Hansjuerg Alder, George A. Calin, Sylvie Menard, and Carlo M. Croce. MicroRNA signatures in human ovarian cancer. *Cancer Res*, 67(18):8699–8707, 2007.

[56] Kati P. Porkka, Minja J. Pfeiffer, Kati K. Waltering, Robert L. Vessella, Teuvo L.J. Tammela, and Tapio Visakorpi1. microRNA expression profiling in prostate cancer. *Cancer Res*, 67(13):6130–6135, 2007.

[57] Cho W.C. OncomiRs: the discovery and progress of microRNAs in cancers. *Mol. Cancer*, 6(1):60, 2007.

[58] Jongpil Kim, Keiichi Inoue, Jennifer Ishii, William B. Vanti, Sergey V. Voronov, Elizabeth Murchison, Gregory Hannon, and Asa Abeliovich. A microRNA feedback circuit in midbrain dopamine neurons. *Science*, 317:1220–1224, 2007.

[59] Kathryn A. O'Donnell, Erik A. Wentzel, Karen I. Zeller, Chi V. Dang, and Joshua T. Mendell. c-Myc-regulated microRNA modulate E2F1. *Nature*, 435:839–843, 2005.

[60] Yong Zhao, Eva Samal, and Deepak Srivastava. Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature*, 436:214–220, 2005.

[61] M.F. Gosso, E.J.C. de Geus, M.J. van Belzen, T.J.C. Polderman, P. Heutink, D.I. Boomsma, and D. Posthuma. SNAP-25 gene is associated with cognitive ability: evidence from a family-based study in two independent dutch cohorts. *Molecular Psychiatry*, 11:878–886, 2006.

[62] Qin Y., Zhao H., Kovanci E., Simpson J.L., Chen Z.J., and Rajkovic A. Analysis of LHX8 mutation in premature ovarian failure. *Fertil Steril*, 2007.

[63] Buck A., Buchholz M., Wagner M., Adler G., Gress T., and Ellenrieder V. The tumor suppressor KLF11 mediates a novel mechanism in transforming growth factor beta-induced growth inhibition that is inactivated in pancreatic cancer. *Mol Cancer Res*, 4(11).

[64] Sebastien S. Hebert and Bart De Strooper. miRNAs in neurodegeneration. *Science*, 317(5842):1179–1180, 2007.

[65] Saetrom P. and Snove O. Jr. Robust machine learning algorithms predict microRNA genes and targets. *Methods Enzymol.*, 427C:25–29, 2007.

[66] Sung-Kyu Kim, Jin-Wu Nam, Je-Keun Rhee, Wha-Jin Lee, and Byoung-Tak Zhang. miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics*, 7:411–423, 2006.