UNIVERSITÀ DEGLI STUDI DI MILANO
Facoltà di Scienze e Tecnologie
Computer Science

MICRORNA TARGET PREDICTION USING DEEP LEARNING

Relatore: Prof. Sebastiano Vigna

Correlatore: Prof. Paolo Boldi

Tesi di:
Simone Sinigaglia
Matricola: 902960

Anno Accademico 2019/2020

Vorrei ringraziare soprattutto ...

Contents

1 MicroRNAs and their importance in living beings **1**
1.1 What are microRNAs? 1
1.1.1 Transcription and processing of miRNAs 2
1.1.2 RNA-induced silencing complex (RISC) 2
1.2 Why are they important? 3
1.3 acaso 3
1.3.1 Folders 3
1.3.2 Tables 4
1.4 In Closing 5

2 MicroRNAs target prediction computational methods **7**
2.1 Introduction 7
2.2 MiRNA target prediction 8
2.2.1 Features and methodologies 8

A Frequently Asked Questions **12**
A.1 How do I change the colors of links? 12

Chapter 1

MicroRNAs and their importance in living beings

1.1 What are microRNAs?

MicroRNAs (abbreviated miRNAs) are a family of $\approx$ 22-nucleotide small non-coding RNAs that regulates gene expression at the post-transcriptional level [2]. This means that they act by binding to partially complementary sites on target genes, which had been previously transcribed from the DNA of the cell, to induce cleavage or repression of productive translation, preventing this way the target gene to be able to exit the cell and start the translational process that produces peptides and proteins.

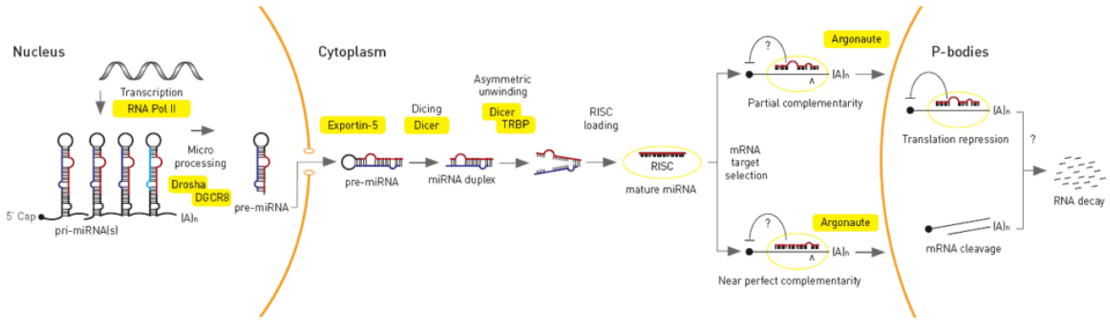

Figure 1: MiRNAs genesis and functionalities

1.1.1 Transcription and processing of miRNAs

As shown in figure 1 miRNAs genes are transcribed by the RNA polymerase II as as large primary transcripts (pri-miRNA) that are processed by a protein complex containing the enzyme Drosha, to form an approximately 70 nucleotide precursor miRNA (pre-miRNA). This precursor is subsequently transported to the cytoplasm where it is processed by a second enzyme, called DICER, to form a mature miRNA of approximately 22 nucleotides. The mature miRNA is then incorporated into a ribonuclear particle to form the RNA-induced silencing complex, RISC, which mediates gene silencing.

It's important to note that, generally, only one of the two strands of the stem loop is incorporated into the silencing process, and it's selected on the basis of its thermodynamic instability and weaker base-pairing on the 5′ end relative to the other strand. The latter, called the passenger strand due to its lower levels in the steady state, is usually denoted with an asterisk (*) and is normally degraded. However, in some cases, both strands of the duplex are viable and become functional miRNAs that target different mRNA populations. (see figure 2)

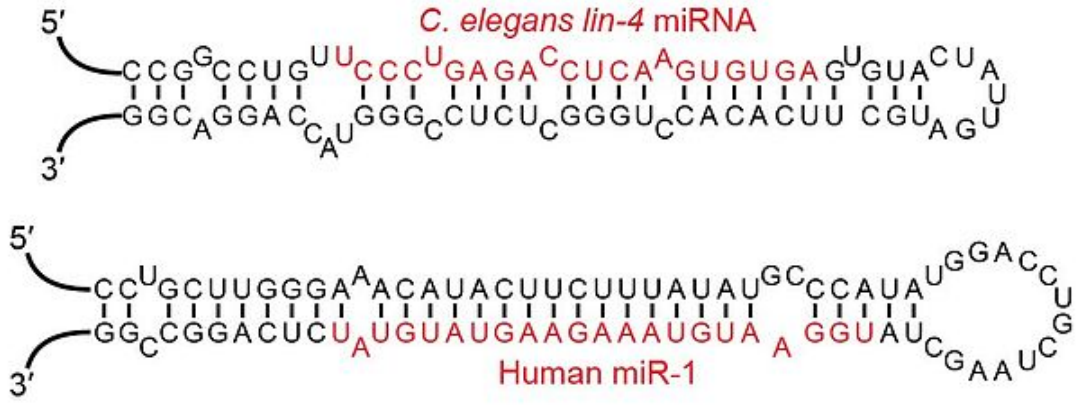

Figure 2: Examples of miRNA stem-loops. In red is shown the mature miRNA

1.1.2 RNA-induced silencing complex (RISC)

As mentioned before the mature miRNA is part of an active RNA-induced silencing complex (RISC). This process represents their main functionality in both animals and plants. The RISC it is a key process in gene silencing and can act in two

different ways as depicted in the right-hand side of picture 1: via mRNA degradation or by preventing mRNA translation. It has been demonstrated that given complete complementarity between the miRNA and target mRNA sequence, Ago2 can cleave the mRNA and lead to direct mRNA degradation. In the presence of only partial complementarity instead, silencing is achieved by preventing translation [9].

1.2 Why are they important?

MiRNAs are particularly abundant in many mammalian cell types and appear to target about 60% of the genes of humans and other mammals [8].

Many miRNAs are evolutionarily conserved, which implies that they have important biological functions [8]. For example, 90 families of miRNAs have been conserved since at least the common ancestor of mammals and fish, and most of these conserved miRNAs have important functions.

The discovery of the first miRNA over 20 years ago has ushered in a new era in molecular biology. There are now over 2000 miRNAs that have been discovered in humans and it is believed that they collectively regulate one third of the genes in the genome.

The repressive action of miRNAs has a huge impact on many biological processes such as cell cycle control and several developmental and physiological processes including stem cell differentiation, cardiac and skeletal muscle development, neurogenesis, insulin secretion, cholesterol metabolism, aging, immune responses and viral replication. [4]

In addition to their important roles in healthy individuals, microRNAs have also been implicated in a number of diseases including a broad range of cancers, heart and neurological diseases. In fact it has been discovered that their expression patterns are highly specific in respect to external stimuli, developmental stage or tissue and this can be used to diagnose diseases in which the expression levels of miRNAs are known to change considerably [18]. Consequently, miRNAs are intensely studied as candidates for clinical diagnosis and predictors of drug response [10].

1.3 acaso

1.3.1 Folders

This template comes as a single zip file that expands out to several files and folders. The folder names are mostly self-explanatory:

Appendices – this is the folder where you put the appendices. Each appendix should go into its own separate `.tex` file. An example and template are included in the directory.

Chapters – this is the folder where you put the thesis chapters. A thesis usually has about six chapters, though there is no hard rule on this. Each chapter should go in its own separate `.tex` file and they can be split as:

- Chapter 1: Introduction to the thesis topic
- Chapter 2: Background information and theory
- Chapter 3: (Laboratory) experimental setup
- Chapter 4: Details of experiment 1
- Chapter 5: Details of experiment 2
- Chapter 6: Discussion of the experimental results
- Chapter 7: Conclusion and future directions

1.3.2 Tables

Tables are an important way of displaying your results, below is an example table which was generated with this code:

```
\begin{table}
\caption{The effects of treatments X and Y on the four groups studied.}
\label{tab:treatments}
\centering
\begin{tabular}{l l l}
\toprule
\tabhead{Groups} & \tabhead{Treatment X} & \tabhead{Treatment Y} \\
\midrule
1 & 0.2 & 0.8\\
2 & 0.17 & 0.7\\
3 & 0.24 & 0.75\\
4 & 0.68 & 0.3\\
\bottomrule\\
\end{tabular}
\end{table}
```

Table 1: The effects of treatments X and Y on the four groups studied.

Groups	Treatment X	Treatment Y
1	0.2	0.8
2	0.17	0.7
3	0.24	0.75
4	0.68	0.3

You can reference tables with `\ref{<label>}` where the label is defined within the table environment. See `Chapter1.tex` for an example of the label and citation (e.g. Table 1).

There are many different LaTeX symbols to remember, luckily you can find the most common symbols in The Comprehensive LaTeX Symbol List.

You can write an equation, which is automatically given an equation number by LaTeX like this:

```
\begin{equation}
E = mc^{2}
\label{eqn:Einstein}
\end{equation}
```

This will produce Einstein's famous energy-matter equivalence equation:

$$E = mc^2 \tag{1}$$

All equations you write (which are not in the middle of paragraph text) are automatically given equation numbers by LaTeX. If you don't want a particular equation numbered, use the unnumbered form:

```
\[ a^{2}=4 \]
```

1.4 In Closing

You have reached the end of this mini-guide. You can now rename or overwrite this pdf file and begin writing your own `Chapter1.tex` and the rest of your thesis. The easy work of setting up the structure and framework has been taken care of for you. It's now your job to fill it out!

Good luck and have lots of fun!

Guide written by —
Sunil Patel: www.sunilpatel.co.uk
Vel: LaTeXTemplates.com

Chapter 2

MicroRNAs target prediction computational methods

2.1 Introduction

Earlier in Chapter 1 we described how miRNAs play a fundamental role in gene regulation. It is common belief that the final and probably most relevant step in their regulatory pathway is targeting [18]. Targeting is intended as the binding of the mature miRNA to the messenger RNA via the RNA Induced Silencing Complex (see figure 3). Hence, valid targets need to be identified for miRNAs in order to properly understand their role in cellular pathways.

However, many of the discovered miRNAs do not yet have identified targets. This is especially the case in animals where the miRNA does not bind to its target with a nearly perfect matching as it does in plants [13]. Experiments have proved that a single miRNA has the potential to regulate hundreds of target mRNAs and multiple miRNAs may compete for the regulation of the same mRNA [1], however target validation is difficult, expensive, and time consuming. Thus, having considered all these facts, it is of crucial importance to have accurate computational miRNA target predictions.

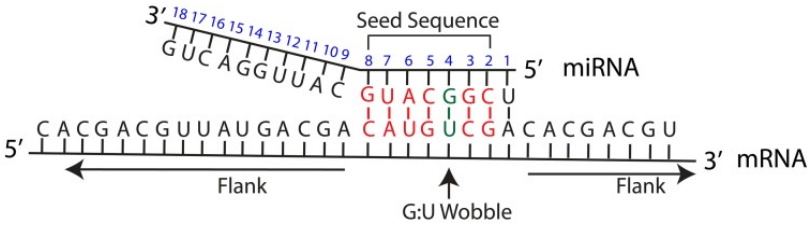

Figure 3: Example of miRNA targeting.

2.2 MiRNA target prediction

Before miRNA target prediction tools were available, possible miRNA binding sites were determined manually. These target sites were later confirmed by laborious and inefficient techniques such as site-directed mutagenesis and other experimental methods. The identification of the first targets for the let-7 and lin-4 miRNAs led to the idea that miRNAs have a pattern in targeting genes which could be used to develop target prediction algorithms [12].

Originally gene targeting by miRNAs was believed to be the result of their binding to the 3'UTR of the target mRNA [1], however recent studies [5] have confirmed gene regulation as a result of the binding of the miRNA to the coding region as well as to the 5'UTR. Furthermore, computational evidence suggests that regulation via the binding of the miRNA to the coding region differs in comparison to the binding pattern seen at the 3'UTR. In particular, it's suggested that miRNAs target the coding regions of mRNAs with short 3'UTRs [15].

Another key factor in target prediction is that 3'UTRs are prone to change under different conditions which might result in the elimination of the target site. Binding in the coding region on the other hand may instead present an evolutionary advantage for the cell as it could help in the preservation of the miRNA binding site [11].

2.2.1 Features and methodologies

While many miRNA targets have been computationally predicted only a limited number have been experimentally validated. Moreover, although a variety of miRNA target prediction algorithms are implemented, results amongst them are generally inconsistent and correctly identifying functional miRNA targets remains a challenging task.

The various methodologies implemented use several different approaches and analyze a wide range of features for this task, however, the most common characteristics are [14]:

- seed region complementarity
- free energy
- site accessibility
- conservation

The seed region

Targeting patterns are very different between plants and animals. Plants, in fact, show a near perfect complement between their miRNA and the respective target mRNA. On the other hand animal miRNAs bind their targets with only partial complementarity. In particular, a region of about 6 to 8 nucleotides in length at the beginning of the miRNA is of crucial importance in the targeting. This short subsequence is called *seed region* and it comprises the nucleotides between the second at the eighth (the seed sequence in Figure 3) starting from the 5′ end. The seed region is very important because it binds to the target mRNA leading to the regulation of the gene in question [17].

Undoubtedly the seed region is one of the most commonly used miRNA traits for target prediction. This seed-centric view, in fact, has been supported by structural studies [16] and a widely cited report [3] that investigated the importance of other (non-canonical) regions within a miRNA concluding that their contributions had relatively low relevance compared to the (canonical) seed region. More recent experiments, however, have highlighted a role for the entire miRNA, suggesting that a more flexible methodology is needed [6].

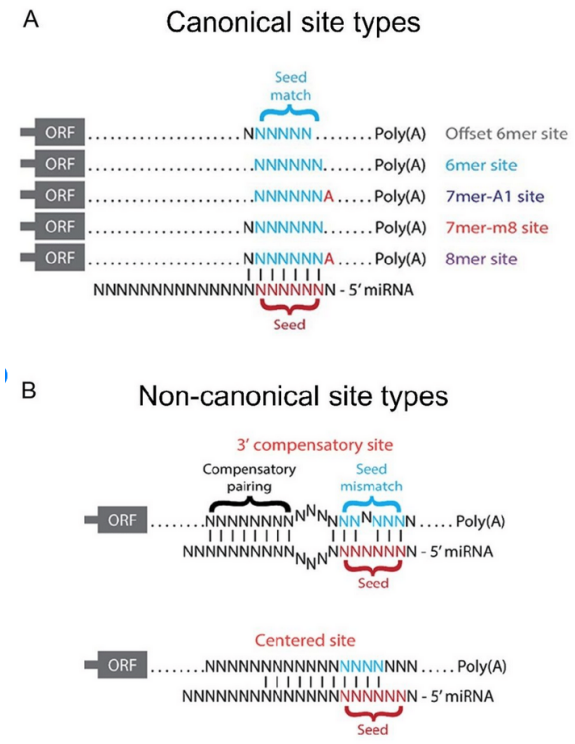

Figure 4: Example of canonical and non-canonical binding sites.

Free energy

The free energy, also called hybridization energy, is defined as the energy released by the pairing between the miRNA and mRNA and it can be used as the measure of the stability of the bond. In fact a stable bond is considered more likely to be a functional target of the miRNA. However, since measuring this quantity directly is difficult, usually the change of free energy during a reaction is considered (ΔG). Reactions with a negative ΔG have less energy available to react in the future, hence they result in systems with an increased stability. By predicting how the miRNA and its candidate target hybridize, regions of high and low free energy can be inferred (Figure 5) and the overall ΔG can be used as an indicator of how strongly bound they are [19].

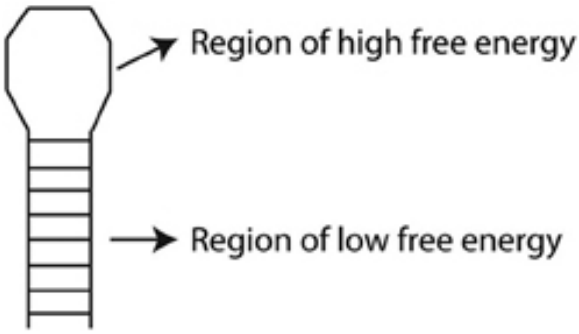

Figure 5: A hairpin loop is shown with the loop corresponding to a region of high free energy (a positive ΔG) and the stem corresponding to a region of low free energy (a negative ΔG)

Site accessibility

Site accessibility is the measure of the ease with which a miRNA can locate and hybridize with its target. After transcription, in fact, a mRNA assumes a certain secondary structure which can interfere with the miRNA ability to bind to its target site. To understand why this is important, we need to consider that the miRNA:mRNA hybridization involves a two-step process in which a miRNA firstly binds to a short accessible region of the mRNA and only after, while the secondary structure of the mRNA unfolds, completes the binding. It is likely that secondary structures contribute to target recognition, because there is an energetic cost to freeing base-pairing interactions within mRNA in order to make the target accessible for miRNA binding. Hence, to assess the likelihood that a mRNA is a target of a given miRNA, the predicted amount of energy required to make

the site accessible (the so called site accessibility energy SAE) should be taken into consideration [7].

The SAE can be computed as the difference between the free energy cost of opening the mRNA and free energy gained from the intermolecular interaction (Figure 6).

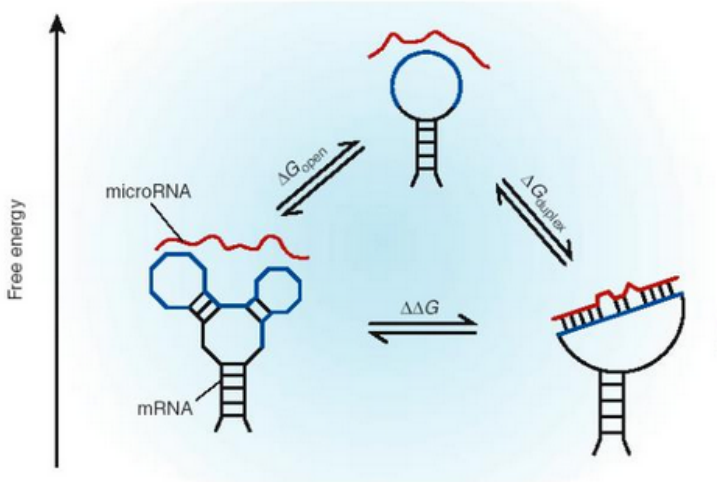

Figure 6: Binding of a miRNA to its target mRNA is depicted as a 2-step process. Portion of the mRNA structure must be open before miRNA:mRNA base pairing can be established.

Conservation

Conservation refers to the maintenance of a sequence across species. According to many reports [18] looking at conserved targets between different species helps reducing the number of false positive results. However, other more recent studies highlighted the fact that this may also increase the number of false negatively identified targets [8].

Appendix A

Frequently Asked Questions

A.1 How do I change the colors of links?

The color of links can be changed to your liking using:

`\hypersetup{urlcolor=red}`, or

`\hypersetup{citecolor=green}`, or

`\hypersetup{allcolor=blue}`.

If you want to completely hide the links, you can use:

`\hypersetup{allcolors=.}`, or even better:

`\hypersetup{hidelinks}`.

If you want to have obvious links in the PDF but not the printed text, use:

`\hypersetup{colorlinks=false}`.

Bibliography

[1] AGARWAL, V., BELL, G. W., NAM, J.-W., AND BARTEL, D. P. Predicting effective microrna target sites in mammalian mrnas. *elife 4* (2015), e05005.

[2] BARTHEL, D. ibiology: Introduction to mirnas. `https://www.ibiology.org/genetics-and-gene-regulation/introduction-to-micrornas/`. Accessed: 2019-06-11.

[3] DU, T., AND ZAMORE, P. D. microprimer: the biogenesis and function of microrna. *Development 132*, 21 (2005), 4645–4652.

[4] FROMM, B., BILLIPP, T., PECK, L. E., JOHANSEN, M., TARVER, J. E., KING, B. L., NEWCOMB, J. M., SEMPERE, L. F., FLATMARK, K., HOVIG, E., ET AL. A uniform system for the annotation of vertebrate microrna genes and the evolution of the human micrornaome. *Annual review of genetics 49* (2015), 213–242.

[5] GROSSWENDT, S., FILIPCHYK, A., MANZANO, M., AND KLIRONOMOS. Unambiguous identification of mirna: target site interactions by different types of ligation reactions. *Molecular cell 54*, 6 (2014), 1042–1054.

[6] HELWAK, A., KUDLA, G., DUDNAKOVA, T., AND TOLLERVEY, D. Mapping the human mirna interactome by clash reveals frequent noncanonical binding. *Cell 153*, 3 (2013), 654–665.

[7] KERTESZ, M., IOVINO, N., UNNERSTALL, U., GAUL, U., AND SEGAL, E. The role of site accessibility in microrna target recognition. *Nature genetics 39*, 10 (2007), 1278.

[8] 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*, 1 (2005), 15–20.

[9] LIM, L. P., LAU, N. C., GARRETT-ENGELE, P., GRIMSON, A., SCHELTER, J. M., CASTLE, J., BARTEL, D. P., LINSLEY, P. S., AND JOHNSON, J. M. Microarray analysis shows that some micrornas downregulate large numbers of target mrnas. *Nature 433*, 7027 (2005), 769.

[10] LU, M., ZHANG, Q., DENG, M., MIAO, J., GUO, Y., GAO, W., AND CUI, Q. An analysis of human microrna and disease associations. *PloS one 3*, 10 (2008), e3420.

[11] LYTLE, J. R., YARIO, T. A., AND STEITZ, J. A. Target mrnas are repressed as efficiently by microrna-binding sites in the 5'utr as in the 3'utr. *Proceedings of the National Academy of Sciences 104*, 23 (2007), 9667–9672.

[12] MAZIERE, P., AND ENRIGHT, A. J. Prediction of microrna targets. *Drug discovery today 12*, 11-12 (2007), 452–458.

[13] PASQUINELLI, A. E. Microrna and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Reviews Genetics 13*, 4 (2012), 271.

[14] PETERSON, S. M., THOMPSON, J. A., UFKIN, M. L., SATHYANARAYANA, P., LIAW, L., AND CONGDON, C. B. Common features of microrna target prediction tools. *Frontiers in genetics 5* (2014), 23.

[15] RECZKO, M., MARAGKAKIS, M., ALEXIOU, P., GROSSE, I., AND HATZIGEORGIOU, A. G. Functional microrna targets in protein coding sequences. *Bioinformatics 28*, 6 (2012), 771–776.

[16] SCHIRLE, N. T., SHEU-GRUTTADAURIA, J., AND MACRAE, I. J. Structural basis for microrna targeting. *Science 346*, 6209 (2014), 608–613.

[17] SMITH, S. M., AND MURRAY, D. W. An overview of microrna methods: expression profiling and target identification. In *Molecular Profiling*. Springer, 2012, pp. 119–138.

[18] WATANABE, Y., TOMITA, M., AND KANAI, A. Computational methods for microrna target prediction. *Methods in enzymology 427* (2007), 65–86.

[19] YUE, D., LIU, H., AND HUANG, Y. Survey of computational algorithms for microrna target prediction. *Current genomics 10*, 7 (2009), 478–492.