UNIVERSITÀ DEGLI STUDI DI MILANO
Facoltà di Scienze e Tecnologie
Dipartimento di Informatica

DEEPMIRNA: A NOVEL APPROACH TO MICRORNA TARGET PREDICTION USING DEEP LEARNING

Relatore: Prof. Sebastiano Vigna

Correlatore: Prof. Paolo Boldi

Tesi di:
Simone Sinigaglia
Matricola: 902960

Anno Accademico 2018/2019

I would like to thank . . .

Contents

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 [3]. 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.

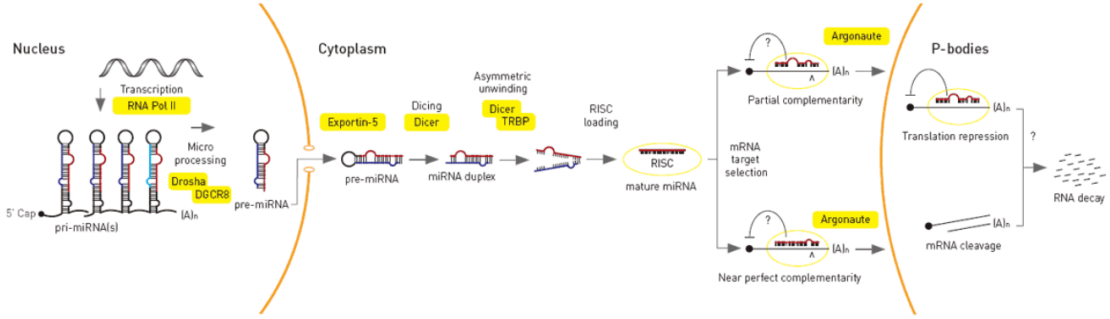

Figure 1: MiRNAs genesis and functionalities

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)

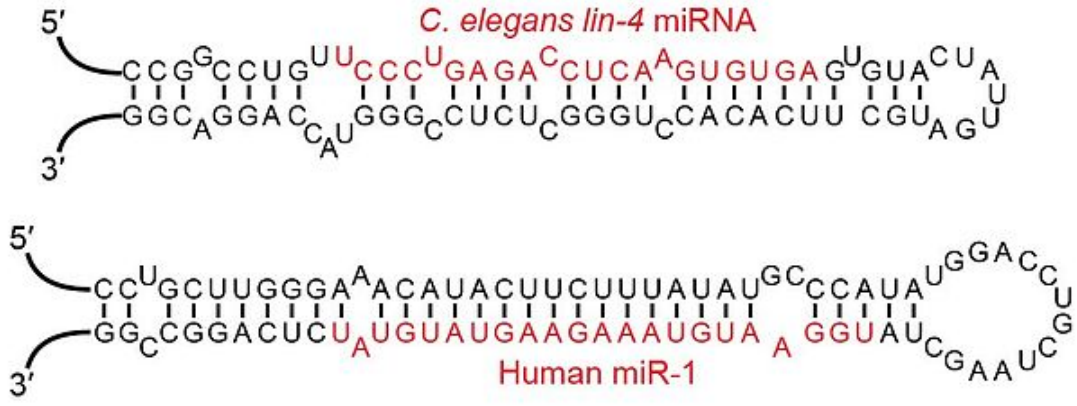

Figure 2: Examples of miRNA stem-loops. In red is shown the mature miRNA

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 [20].

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 [19].

Many miRNAs are evolutionarily conserved, which implies that they have important biological functions [19]. 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 two 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. [6]

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 [32]. Consequently, miRNAs are intensely studied as candidates for clinical diagnosis and predictors of drug response [21].

1.3 Why miRNAs target prediction is a difficult task?

1.4 acaso

1.4.1 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

1.4.2 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\\
```

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

```
2 & 0.17 & 0.7\\
3 & 0.24 & 0.75\\
4 & 0.68 & 0.3\\
\bottomrule\\
\end{tabular}
\end{table}
```

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:

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

1.5 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 [32]. 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 [25]. 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 [2], 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.

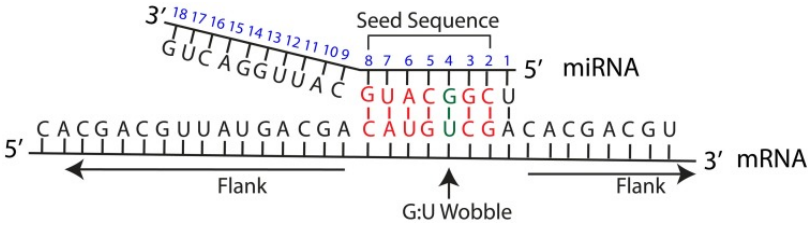

Figure 3: Example of miRNA targeting.

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 [23].

Originally gene targeting by miRNAs was believed to be the result of their binding to the 3'UTR of the target mRNA [2], however recent studies [8] 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 [28].

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 [22].

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 average performance of target prediction tools, which typically identify approximately 80% of known miRNA targets, indicates that the mechanisms associated with miRNA-regulated processes remain poorly understood. Thus, there is a room for novel approaches to improve the knowledge of the rules that govern their targeting process [14]

The various methodologies implemented use several different approaches and analyze a wide range of features for this task. Almost all target prediction methods are rule-based or adopt machine learning methodology with varying success. Rule-based systems incorporate various human-crafted descriptors to represent miRNA:gene target binding (e.g. type of pairs in the site, binding stability, or conservation of the target site among species). Machine learning techniques also use those descriptors, but as input features to machine learning models. The limitation

of both these approaches is indeed the process of feature selection and representation, which is constrained by the use of human selected descriptors to model a process that is not fully understood.

The most common characteristics used in miRNA targets identification are [26]:

- 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 [30].

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 [29] and a widely cited report [5] 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 [10].

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

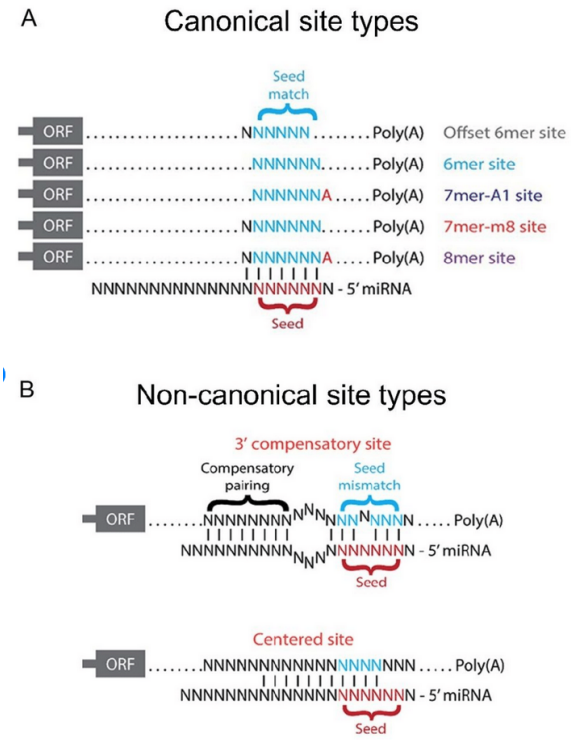

Figure 4: Example of canonical and non-canonical binding sites.

can be inferred (Figure 5) and the overall ΔG can be used as an indicator of how strongly bound they are [35].

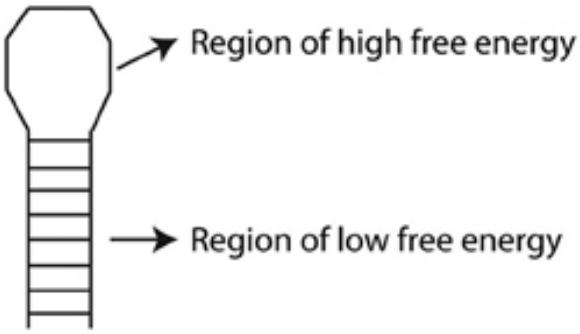

Figure 5: A hairpin loop is shown with the loop corresponding to a region of high free energy (a positive ΔG) and the stem corresponding to a region of low free energy (a negative ΔG)

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 [13].

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).

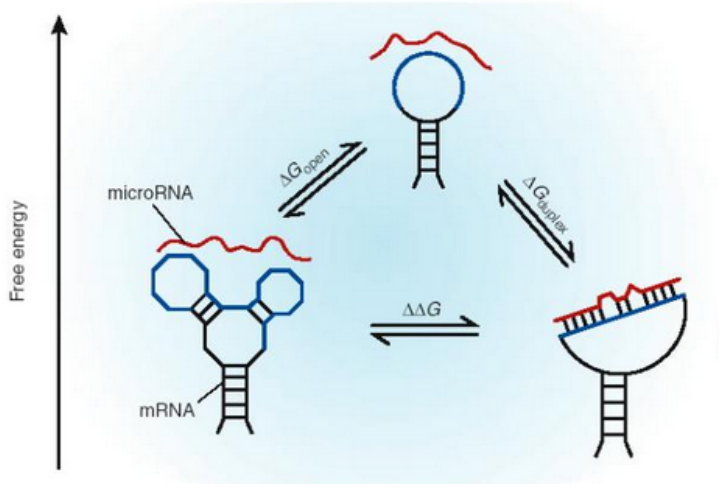

Figure 6: Binding of a miRNA to its target mRNA is depicted as a 2-step process. Portion of the mRNA structure must be open before miRNA:mRNA base pairing can be established.

Conservation

Conservation refers to the maintenance of a sequence across species. According to many reports [32] 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 [19].

Chapter 3

Experimental setup

3.1 Introduction

The majority of prediction tools are based on the assumption that it is the miRNA seed region that contains almost all the important interactions between a miRNA and its target and their focus are on these canonical sites.

In this thesis, we aim at using deep learning techniques to investigate the role of those non-canonical sites and pairing beyond the canonical seed region in miRNA targets. Recent increases in computational power have permitted the rise of methods that can dispense with human-crafted features, making it possible to deal directly with raw data and autonomously learn and identify patterns to appropriately represent it. In particular, deep learning has been shown to be an effective method for classification tasks in domains with complex feature representation [17].

Deep Learning (DL) has already been applied to the miRNA target prediction problem. Cheng et al. [4] used convolutional neural networks to analyze matrices of miRNA:site features, but the selected features were still human-crafted descriptors and thus the method faces similar problems as rule-based and ML approaches. A more recent work, DeepTarget [18], relied on recurrent neural networks to identify potential binding sites and assess their functionality. However this work is still oriented to the identification of canonical sites and relies on a limited small data set for the training phase. Another approach using DL is DeepMirTarSdA [33], that explore the use of stacked de-noising auto-encoders (SdA) to predict human miRNA-targets at the site level, but the network obtained is huge and the small availability of input data (about 8000 samples between positives and negatives) results in a model that performs well on the data being used but generalizes quite poorly.

In this thesis we present DeepMiRNA, a miRNA target prediction tool that attempts to take advantage of the learning capacity of a neural network to extract abstract patterns from raw input data. Unfortunately, to the best of our knowledge, there is no suitable raw-data representational method for miRNA-target prediction. This is very likely the consequence of the very small quantity of validated data available for this task. For this reason, rather than making assumption about suitable descriptor, we created a set of rules to help finding the best candidate binding sites leaving the classification decision to the neural network (see CSSM in the next section).

More precisely, DeepMiRNA scans the 3'UTR of the gene identifying potential target sites according to the chosen rules. It then uses the previously trained network to identify the relevant patterns by directly examining the whole mature miRNA transcript, rather than focusing on the seed region and analyzing precomputed descriptors. In the following text we denote a miRNA binding site with MBS.

3.2 Preparing for training

Two of the fundamental properties in deep neural network theory states that:

1. with sufficient data samples and a correct network design a NN can approximate any mathematical function

2. a NN has the capacity to automatically learn the relevant features of complex data structures by means of its hidden layers [17]

For these reasons, in our approach we sought to minimize potential biases introduced by handcrafted features by working with the miRNA and the mRNA transcripts, feeding them directly to the neural network.

The DeepMiRNA working pipeline for the identification of functional targets for miRNAs can be summarized as follows (Figure 7: a 30-nucleotide sliding window with stride of 5 nucleotides is used to scan the 3'UTR of a given gene. Both values (size and stride) has been empirically computed during the training stage (see the results section for more informations). For each mRNA 30-nucleotide long subsequence, the VIENNA RNACofold package [11] is used to compute the stability of the binding between the miRNA transcript and the fragment. If the computed value is below a predetermined threshold, the primary structure of the mRNA and the miRNA are examined to see if the criteria defined by the candidate site selection rules (CSSR) are met. If so, the duplex is vectorized and fed into the network for classification. The prediction is then further refined using an a posteriori filter that computes the site accessibility of the mRNA region surrounding

the predicted binding site. This last step has revealed an important role in false positive reduction and it's used only for positive predictions.

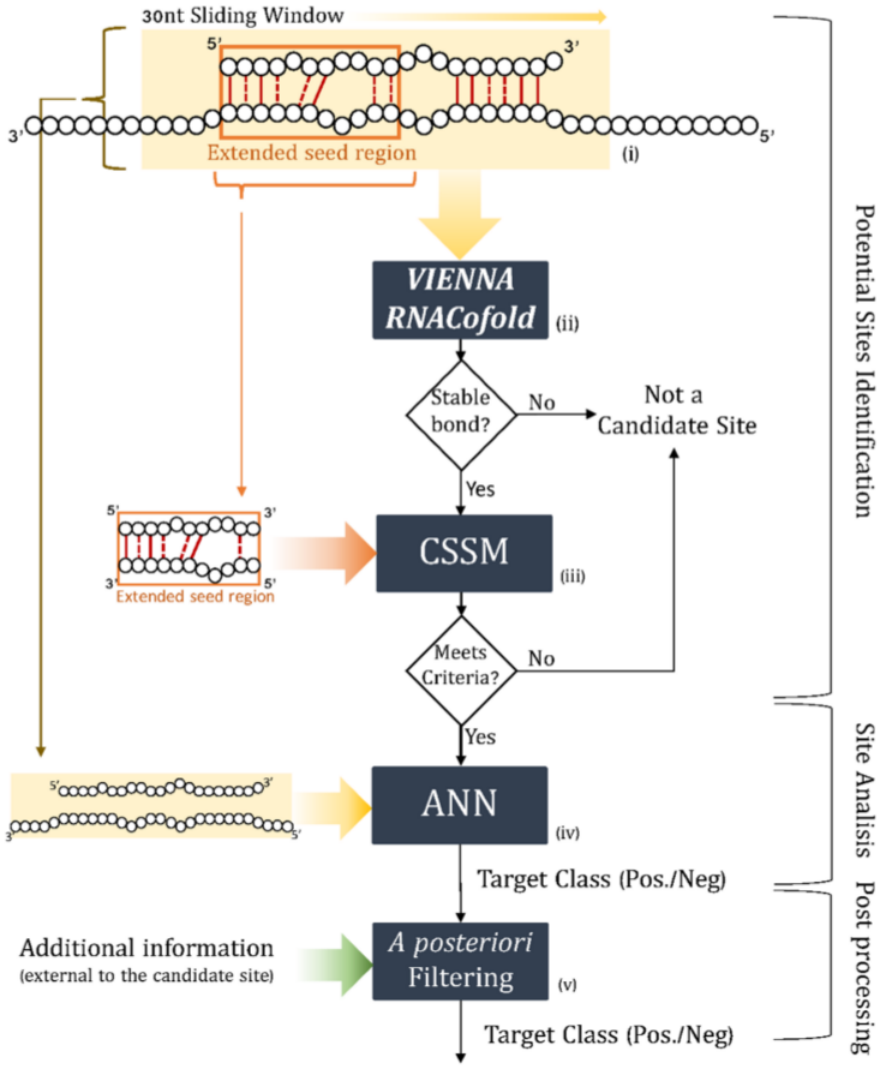

Figure 7: **DeepMiRNA pipeline**. (i) a 30nt sliding window with a 5nt step is used to scan the gene transcript; (ii) The Vienna Cofold software is used to compute the binding stability; (iii) if the bond is predicted to be stable a partial complementarity according to the rules defined in the CSSR is verified; (iv) if all previous checks passed the duplex is fed to the NN for prediction; (v) for positive predictions a filter is used to compute the site accessibility of the miRNA: if the energy needed to access the site is above a certain threshold the prediction is changed to negative.

3.2.1 Data preprocessing

A key factor for successful application of any Machine Learning technique and one of the most important aspect to consider before feeding the network with data is how we prepare the input data and the targets.

The main aspect to consider during this process is the use of suitable techniques to make the data more amenable for the NN: this includes vectorization, normalization and handling of missing values.

vectorization

In order to be accepted by the neural network all inputs must be tensors, that is they must be expressed by numerical arrays with suitable shape and dimension. The process to encode categorical data, in our case the transcript sequences, into tensor is called *vectorization*. In this thesis we actually employed two different techniques for this operation:

- one hot encoding of the sequences [1].
- sequence embedding using a Word2Vec approach [7].

One hot encoding is a process by which categorical data such as strings are converted into binary vectors. In our case, each nucleotide is translated to a binary vector of size 4, corresponding to the four possible nucleotide values as described in Table 2

The main problem using this method is that not all duplexes have the same size. This is in particular due to the different miRNA's transcript length (ranges from 18 to 30), The network, instead, requires that all inputs have the same shape. Hence, in order to meet this requirement, every miRNA sequence, when needed, has been padded with 'empty' letters to reach the maximum size length (in this case 30). Regarding the site transcript, each fragment has size 40: 30 corresponding to the

Table 2: One Hot encoding of a nucleotide.

Nucleotide	**Encoding**
A	[1, 0, 0, 0]
C	[0, 1, 0, 0]
G	[0, 0, 1, 0]
U or T	[0, 0, 0, 1]
Empty	[0, 0, 0, 0]

window size plus 5 additional nucleotides upstream and downstream. These additional nucleotides seek to capture any influence that the flanking sequence may exert on the target [19]. With these adjustments each duplex is represented by a binary vector of (fixed) size 280.

The second vectorization method uses a Word2Vec approach. Word2Vec [7] is a Natural Language Processing (NLP) methodology to map words into numeric vectors based on their context. Being the context defined as the words surrounding the word to encode. For DNA sequences, however, there is no clear definition for words, so usually a k-mer (that is a set of k continuous nucleotides)is used to define a word (see Figure 8). Therefore, in case of biological sequences the context is defined as the set of n adjacent k-mers (being n a parameter to validate). For this thesis use the software available at `https://github.com/pnpnpn/dna2vec` to train the model used to encode the k-mers. This encoding has two important advantages compared to one hot encoding:

1. each k-mer of length comprised between 3 and 8 is mapped to an equal size vector of size 100.

2. similar k-mers are mapped to close points in the features space according to a specific distance metric (usually Euclidean distance).

In our case each variable lenght miRNA sequence has been split into 4 different size k-mers each mapped into a 100-dimension vector, while each fixed size site transcript (plus the flanking nucleotides) has been split into 5 8-mers. This way each duplex is mapped into a $9x100$ matrix obtained concatenating the resulting 9 vectors. It's important to note that this vectorization requires a different design and implementation of the neural network to use as we will describe in the next section.

normalization and missing values

Another crucial part of data preprocessing concerns their normalization. In general, it's not safe to feed the network with data that takes relatively large values or that are heterogeneous (i.e have very different values ranges). In our case, however, the vectorization process guarantees that the numerical values resulting from the encoding are both small and homogeneous. Regarding missing or incomplete data, we found a very small quantity of them in the datasets retrieved, hence we simply decided to discard them.

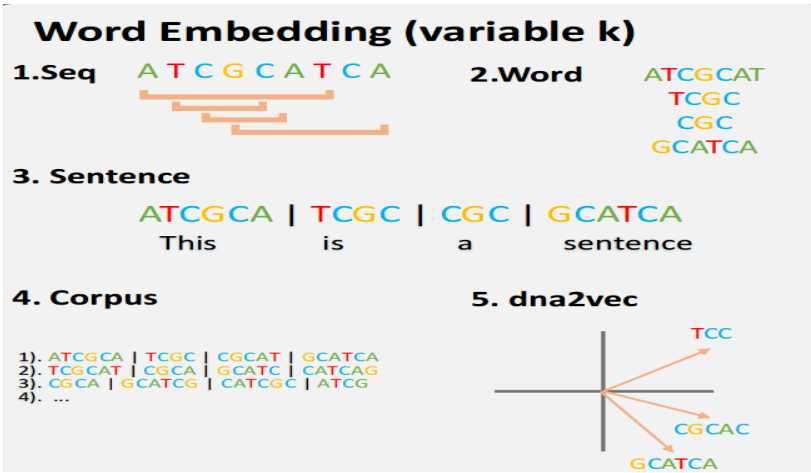

Figure 8: Dna2Vec mapping example using different length k-mers

3.2.2 Dataset preparation

One of the biggest challenge to face in any Machine Learning application is to get a good set of significant data. The main purpose is to have a sufficiently variable and representative dataset that allows the model to generalize well on new and unseen examples. This phase is probably the most important together with the network design as it's crucial for achieving good performances. It is of fundamental importance that the data is representative for the problem we want to solve. It does not matter how big is the quantity of data we obtain if that is not aligned with the goal we want to pursue. One should focus on finding data with features that matter to what we're trying classify or predict and discard unrelated features. Hence, the first step of a classification process should be proper data collection, and until we achieve this, we will find ourselves constantly coming back to this step [15].

For the miRNA target prediction task, this requires a comprehensive dataset of verified positive and negative targets that encompass both canonical and non-canonical examples. While there are multiple data repositories providing information regarding experimentally validated positive miRNA targets[31] [12] there are significantly fewer experimentally verified negative targets. This represents a major concern for ML-based approaches that require similar numbers of labeled examples for both classes.

In this thesis we only considered human data and we used the version 8 of the Diana TarBase[1] and the release 7.0 of miR TarBase[2] datasets. Moreover, in order to annotate the datasets with lacking informations such as molecules transcripts,

[1] http://diana.imis.athena-innovation.gr/DianaTools/index.php
[2] http://mirtarbase.mbc.nctu.edu.tw/php/index.php

Table 3: Example of Diana TarBase entries.

gene name	miRNA name	functionality
AB002442	hsa-miR-192-5p	POSITIVE
DAD3R	hsa-miR-34a-5p	NEGATIVE
E2IG4	hsa-miR-200c-3p	POSITIVE

we downloaded gene sequences from Ensembl[3] and miRNA transcripts from miRBase[4]

Diana TarBase

The Diana TarBase is the most widely used database for validated miRNA interactions. The latest version at time of writing is *v*8 and asking for a bulk download from their website is possible to obtain a dataset of 526658 positive and 63559 negative validated targets. This dataset provides informations about miRNA:mRNA interactions in the form of miRNA name, gene name and functionality as it's illustrated in Table 3. Unfortunately there is no information about transcript sequences of the molecules or binding sites location.

Mir TarBase

This database contains interactions of about 2000 miRNA and around 20000 genes. In total the downloaded dataset comprises 380091 positive and 382 negative duplexes available from the Download section of the website. It's possible to find it under the name *hsa_MIT.xlsx*. It's important to note that every single miRNA can interact with hundreds of different genes. Again, exactly as with the Diana TarBase dataset, no sequence or binding sites transcripts are provided.

Ensembl

From Ensembl[34] is possible to obtain the transcripts of all human genes. In this thesis we focused on the 3'UTR and for the download we used the BioMart tool available from the main menu. More specifically, from the BioMart webpage we followed these steps:

- from the drop down menu choose database: *Ensembl Genes 96*;

[3] `https://ensembl.org/index.html`
[4] `http://mirbase.org`

- then choose dataset: *human genes GRCh38.p12*;
- from the left-hand side of the page click attributes and select *sequences*;
- from this submenu select *3'UTR* and then click on *Header Information*;
- now from the header submenu select the following attributes: *Gene stable ID*, *Gene name*, *Transcript stable ID*, *3'UTR start* and *3'UTR end*;
- lastly click on *Results* from the top left and export the generated fasta file selecting the box *Unique results only* and pressing *Go*.

Note that a single gene identified with a unique ID may produce multiple RNA's transcripts each denoted with its own transcript ID. The actual transcript observed will depend on on the tissue, developmental time point, and environmental or hormonal factor. Typically there's a single major transcript expressed in a given cell at a given time, but not always. Unfortunately in both the Diana TarBase and miR TarBase the transcript ID of the miRNA:mRNA pair is not present, hence, according to what is used in [33], we kept the transcript with the longest sequence.

mirBase

The mirBase database allowed us to retrieve all known miRNAs transcripts. For the homo sapiens species there currently around 2000 miRNAs sequences and in order to collect them we followed these steps:

- from the homepage click *Browse* from the top menu;
- click on *human* and a preview of the data will appear on screen as depicted in figure 9;
- on the bottom of the page select sequence type *Mature sequence* and output format *Unaligned fasta format*;
- press *Select all* and then *Fetch sequences*;
- the query result will be printed on the next page where it can be copied and pasted to a regular text file and saved as a fasta file.

As a preliminary step, the Diana and Mir TarBase data were parsed to remove inconsistent entries that were marked both as positive and negative targets due to contradictory results in different experimental validations and combine entries that were validated more than once by different verification methods. This produced a final dataset of 646206 positive (+) and 38464 negative (-) miRNA:mRNA

Homo sapiens miRNAs

ID	Accession	RPM	Chromosome	Start	End	Strand	Confidence
hsa-let-7a-1	MI0000060	145261	chr9	94175957	94176036	+	✓
hsa-let-7a-2	MI0000061	142652	chr11	122146522	122146593	-	✓
hsa-let-7a-3	MI0000062	142757	chr22	46112749	46112822	+	✓
hsa-let-7b	MI0000063	83403	chr22	46113686	46113768	+	✓
hsa-let-7c	MI0000064	135622	chr21	16539828	16539911	+	✓
hsa-let-7d	MI0000065	4649	chr9	94178834	94178920	+	✓

Figure 9: A sample of the mirBase database.

interactions containing bindings for about 16000 different genes and around 2000 miRNAs. This data was then split into two equal parts for the training and testing phases.

3.2.3 Training dataset

Structuring a proper training dataset is an essential aspect of effective deep learning models but one that is particularly hard to solve. Part of the challenge comes from the intrinsic relationship between a model and the corresponding training dataset. If the performance of a model is below expectations, it is often hard to determine whether the causes are related to the model itself or to the composition of the training dataset.

The purpose of this stage is to train and validate the neural network that has the responsibility for distinguishing between functional (positive) and non-functional (negative) target sites. A positive experimentally validated gene can exhibit tens of potential positive binding sites but not necessarily all them are actually functional. Hence, the training set must be composed of miRNA:MBS pairs rather than miRNA:mRNA duplexes. However, the retrieved data, while consistent and sufficiently various, do not contain such informations. Searching the Internet we were able to find only 2 publicly available validated datasets providing information regarding experimentally identified miRNA binding site locations: the Helwak [10] and the Grosswendt [8] datasets.

Positive binding sites

The two above datasets contain miRNA:MBS locations obtained through PAR-Clip [9] and CLASH [16] experiments, however the binding site identified might not be functional. Hence, in order to consider a site as positive we cross-reference them with the Diana TarBase and miR TarBase. In particular we considered a given duplex miRNA:MBS as functional if:

- form a stable bond, that is it has a free energy below a predetermined threshold accoriding to Vienna Cofold. The threshold used was $-10\ kcal/mol$.+
- correspond to a miRNA:gene pair marked as functional in either the Diana or the miR dataset.

This process produced a total of 33142 positive miRNA:MBS duplexes for the training stage.

Negative binding sites

The small number of non-functional experimentally validated binding sites makes the construction of a representative negative dataset a very difficult task. Some of the examined existing tools overcome this issue by creating 'mock' targets [24] continuously shuffling the sequence of a real functional binding site until the resulting transcript does not appear in any positive miRNA target repository. This method, however, may lead the neural network to learn the function used to create the negative examples resulting in a model trained to discriminate between artificial and real data rather than discerning functional targets from non-functional. Besides, there is no guarantee that the generated sequence is indeed a true negative because it has not been experimentally validated.

The solution we implemented opted for using the validated miRNA:mRNA pairs to extract the negative binding sites and add them the ones provided by the CLASH and CLIP experiments. The idea is that any sequence of approximately 30 nucleotides in the mRNA of a negatively validated miRNA:mRNA pair may represent a potential negative target site. However, in order to avoid the introduction of noise including such sequences, we decided to keep only subsequences of mRNA where the associated miRNA has the potential to form a stable bond. In order to check this requirement we used a sliding window of 30 nt along the entire 3'UTR region. For each negative miRNA:mRNA pair we kept only MBS with the a binding energy below a certain threshold, chosen according to RNACoFold tool from the ViennaRNA package [11]. This process resulted in 32284 negative target sites.

Training stage

Once the positive and the negative datasets have been created, we shuffled and merged them to create a unique dataset. The whole process of creation of the training set is illustrated in figure 10. In order to train and validate the network we used an 8 fold cross-validation approach keeping 66% of the data for training and the rest for validation and test set. During this split we made sure to exclude miRNA:MBS pairs that shared miRNA and MBS with data instances in the training dataset.

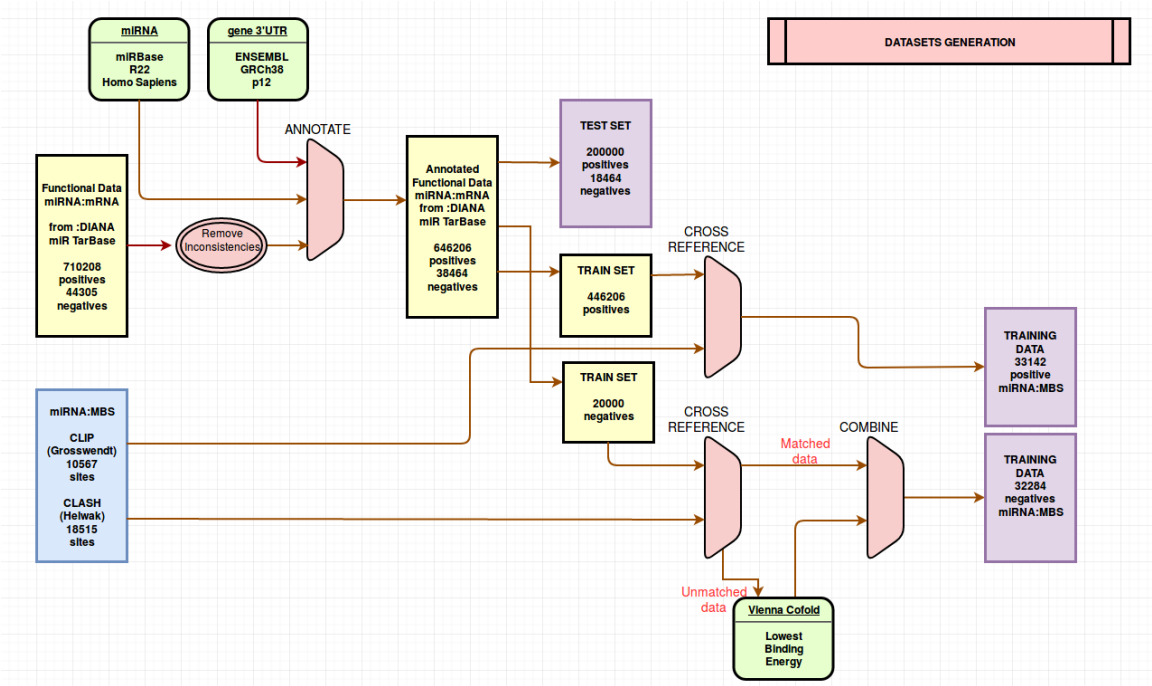

Figure 10: The datasets generation process.

3.3 The testing stage

There is a big difference regarding the procedure adopted for the testing stage compared to the training phase. Here the purpose is to predict if a given miRNA targets a certain gene, hence the testing data consists of pairs containing the miRNA and the whole 3'UTR transcript, rather than a specific MBS. This step is essential to be able to evaluate the whole pipeline process of DeepMiRNA. To correctly evaluate the network performance we used the experimentally verified miRNA:mRNA

pairs excluded from the training set. We now describe all the steps involved in the test stage.

3.3.1 Candidate site selection methods: CSSM

Often, in the academic world, people tend to affirm that, in order to properly train a neural network, one needs to have a huge amount of data available. While we partially disagree with this common belief, convinced that quality and variety prevail over quantity, we still agree that in the case of miRNA targets prediction the available experimentally validated data is still not sufficiently representative of the task and the candidate site selection step effectively narrows the search space to simplify the NN classification task.

For this reason, the selection of candidate sites in a mRNA becomes a key step for a miRNA target sites predictor because it helps identifying which regions within the mRNA have the potential to accommodate a binding site. We found out that most of the publicly available algorithms follow a similar approach: they scan the gene's 3'UTR looking for sites that are partially complementary to the miRNA transcript; if a site meets certain criteria, it is considered to be a candidate site and is subjected to further analysis.

In the light of new recent results, stating that functional miRNA targets can arise either by a single strong binding site (i.e. 6 consecutive complementary nucleotides) or by multiple weak binding sites [10], we believe it's essential to consider the whole 3'UTR sequence using CSSM willing to accept both canonical and non-canonical sites. These methods are less conservative and allow accepting bulges, mismatches or wobble pairs in the seed region (see figure 11).

The CSSM adopted in DeepMiRNA use a very similar approach to miRAW [27]: we consider a 30-nucleotide window to scan the 3'UTR and we extend the typical 7mer seed region (figure 11a) considering the first 10 nucleotides of the miRNA transcript (figure 11d) to look for partial complementarity.

In particular, we consider a site to be a potential candidate site if there is a minimum number of base pairs, both Watson-Crick and Wobble, within the extended seed region. In this thesis we investigated two different configurations:

1. CSS-6.10: a candidate site must contain at least 6 base pairs between the extended seed region;

2. CSS-7.10: a candidate site must contain at least 7 base pairs between the extended seed region.

In each case, base pairs do not need to be consecutive in order to accommodate

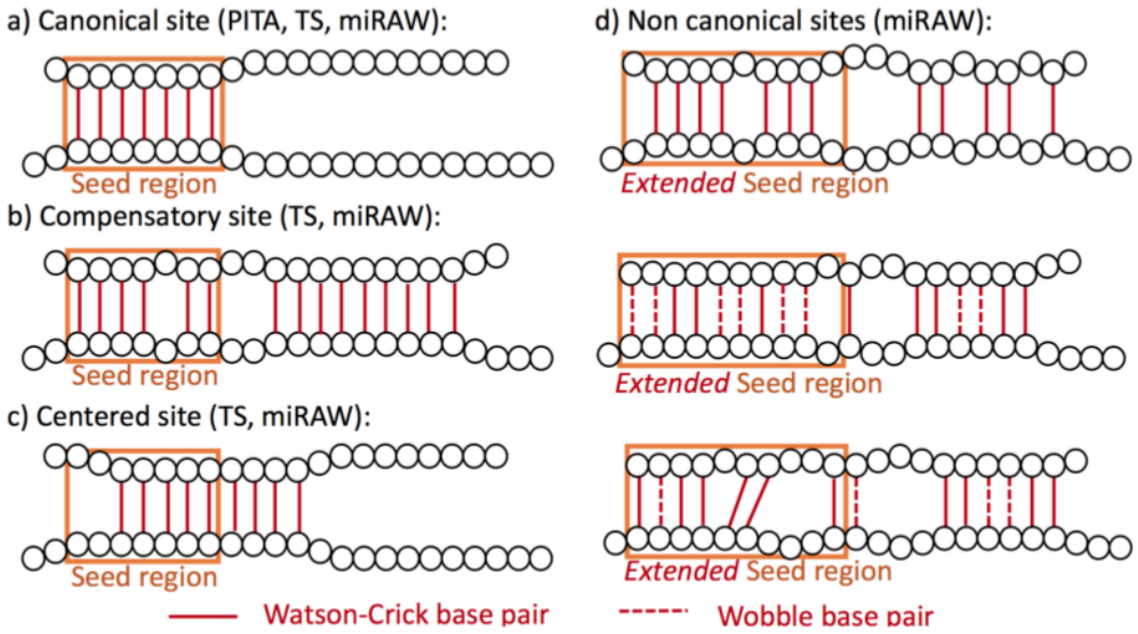

Figure 11: Example of canonical and non-canonical sites used in three available tools: Pita, TargetScan and miRAW.

the presence of gaps and bulges. The task of finding non-consecutive complementary nucleotides can be mapped to the problem of finding the longest common subsequence (LCS) between two strings as illustrated in figure 12.

Those configurations accept both standard canonical MBSs as well as a broader range of non-canonical target site structures including the vast majority of experimentally validated sites from Diana TarBase and CLIP/CLASH binding site datasets. Moreover, while these relaxed conditions for the seed region generate a much larger number of candidate sites and potentially an increased quantity of false positives, the decision of whether a site represents a functional target is delegated to the neural network. This way, we ensure that minimal assumptions, and hence bias, are incorporated into the analysis.

3.3.2 Filtering predictions

In chapter 2 we highlighted the importance of site accessibility for miRNA targets prediction. In fact, many studies [10] [26] have proved genes accommodate site accessibility by preferentially positioning targets in highly accessible regions [13] thus demonstrating that target accessibility is a critical factor in miRNA function.

Moreover, the use of a greedier CSSM approach gives rise to an increase on the

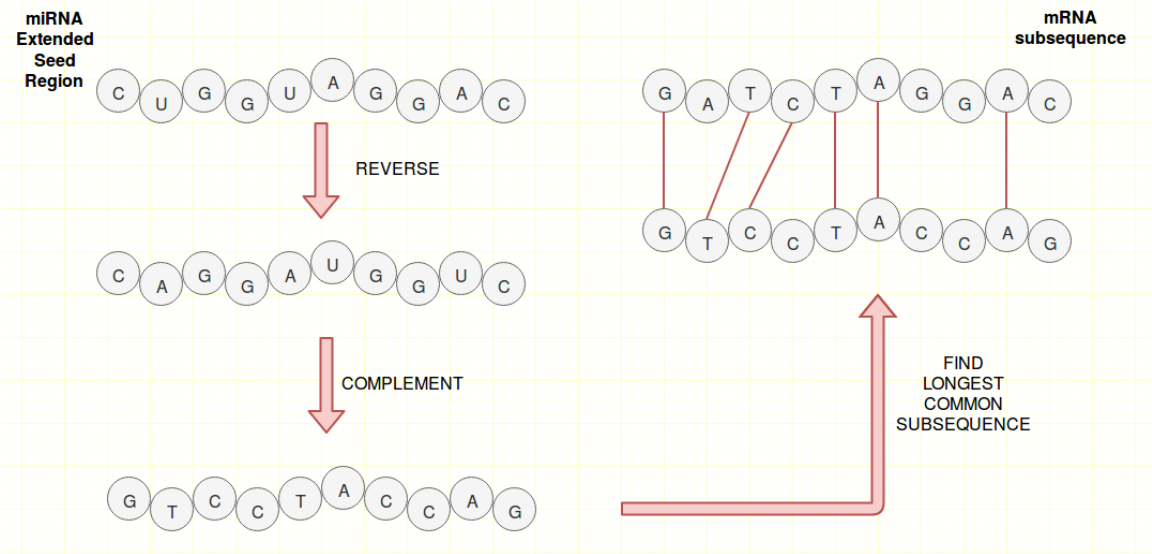

Figure 12: Mapping partial non-consecutive complementarity to longest common subsequence search.

number of potential candidate sites for each miRNA:mRNA duplex. For example, on average, where a strict canonical approach identifies $\approx$ 3-4 sites per duplex, the CSS-6:10 identifies approximately 32 MBSs while the CSS-7:10 about 21. As a consequence the chance of obtaining more false positives is very likely to increase.

Furthermore, the neural network itself may not be able to completely discern false from true positive duplexes. This is due, in particular, to two main factors: first of all the number of validated negative MBSs is extremely low; in fact, despite having a good number of negatively validates miRNA:mRNA pairs, we only have a small quantity of negatively validated binding sites. Thus, we had to artificially select the negative sites for the training stage as explained in the previous section. Second, functional and non-functional sites are very similar in terms of complementarity with the pairing miRNA and, hence, we believe it's important to also consider the characteristics of the secondary structure of the duplex to improve the accuracy of the prediction and reduce the number of false positives.

To this purpose, we investigated the possibility of using an a-posteriori filter that considers the secondary structure and compute the site accessibility of the target site. The site accessibility has been computed as follows:

- for each potential site identified using the CSSM, we considered the region of 200 nucleotides surrounding the MBS and we call it *folding chunk*. For example, if the MBS has length 30 and comprises the region between nucleotide

100 and 129, we consider for the computation of the site accessibility the region between nucleotide 15 and 214. If either on the left or on the right-hand side of the MBS there were not enough nucleotides we considered a shorter fold;

- the free energy ΔG_{free} of the folding region is computed, using Vienna Cofold, to check the amount of energy released during the reaction;
- next we computed the opening energy ΔG_{open}, that is the energy needed to unfold the mRNA and allow the miRNA binding;
- the site accessibility energy $\Delta\Delta G$ is given by the difference between the free energy released with the binding and the opening energy (see figure 13). The bigger the value the more accessible the site:

$$\Delta\Delta G = \Delta G_{free} - \Delta G_{open}$$

In order to use this feature as a filter we set a site accessibility threshold meaning that sites with a low value are discarded while only MBS with a site accessibility greater than the threshold are considered as functional.

One may wonder why we used this feature as an a-posteriori filter rather than using it as another candidate selection rule. While the final result is the same, computing the secondary structure of a folding chunk, and hence its accessibility, is a pretty difficult task involving a lot more computation than asking the network for a prediction. Thus, we decided to use this feature as the last pipeline step.

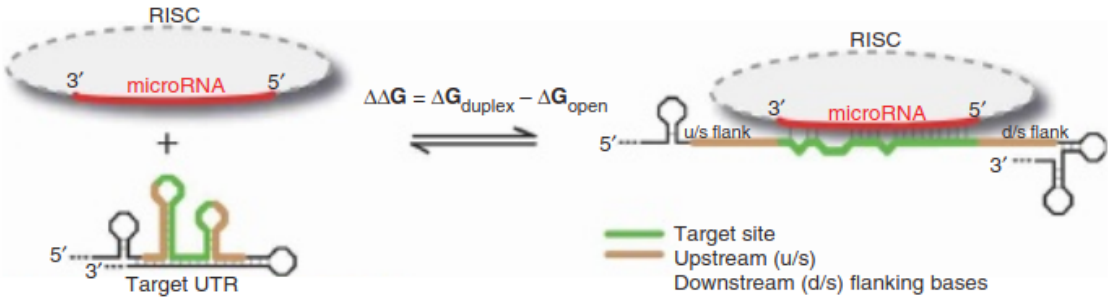

Figure 13: Illustration of site accessibility energy for miRNA:MBS interactions.

Chapter 4

Neural network design and implementation

4.1 External libraries

DeepMiRNA has been developed using *Python3*. In particular we used Pandas and

4.2 Choosing the right model

Building deep learning applications in the real world is a never-ending process of selecting and refining the right elements of a specific solution. Among those elements, the selection of the correct model and the right structure of the training dataset are, arguably, the two most important decisions that any data scientist needs to make when designing deep learning solutions. How to decide what deep learning model to use for a specific problem? How do we know whether we are using the correct training dataset or we should gather more data? Those questions are the common denominator across all stages of the life cycle of a deep learning application. Even though there is no magic answer to those questions, there are several ideas that could guide the decision process.

First of all we need to start identifying the correct baseline model, in particular we should select what type of networks suits more the input dataset. In the case of miRNA targets predictions the topological structure of the available data are strictly correlated to the vectorization method selected. If we opt for the one-hot encoding that maps duplexes into fixed-size vectors, we should be thinking of using a feed-forward network with inter layer connectivity. While, if we select the Dna2Vec approach that transforms each duplex into a matrix, then the problem

could be tackled using convolutional neural networks (CNN)[17].

The second part concerns the selection of the optimization algorithm to use. The most popular are, arguably, SGD (Stochastic Gradient Descent) and its variation using momentum or learning decay, and Adam. The latter, in particular, is very often used combined with CNNs.

As mentioned in chapter 3 DeepMiRNA uses 2 different neural network for the training stage: a regular feed-forward network for the one-hot encoded sequences and a CNN for the Dna2Vec encoded duplexes. For both networks we used the Adam optimizer with the following parameters:

- learning rate = 0.002;
- beta_1 = 0.9;
- beta_2 = 0.999;
- epsilon = $1e-8$;
- decay = 0

4.2.1

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] Onehot encoding in python. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html. Accessed: 2019-06-15.

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

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

[4] CHENG, S., GUO, M., WANG, C., LIU, X., LIU, Y., AND WU, X. Mirtdl: a deep learning approach for mirna target prediction. *IEEE/ACM transactions on computational biology and bioinformatics 13*, 6 (2015), 1161–1169.

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

[6] 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.

[7] GOLDBERG, Y., AND LEVY, O. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722* (2014).

[8] 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.

[9] HAFNER, M., LANDTHALER, M., BURGER, L., KHORSHID, M., HAUSSER, J., BERNINGER, P., ROTHBALLER, A., ASCANO JR, M., JUNGKAMP, A.-C.,

MUNSCHAUER, M., ET AL. Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. *Cell 141*, 1 (2010), 129–141.

[10] 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.

[11] HOFACKER, I. L. Vienna rna secondary structure server. *Nucleic acids research 31*, 13 (2003), 3429–3431.

[12] HSU, S.-D., LIN, F.-M., WU, W.-Y., LIANG, C., HUANG, W.-C., CHAN, W.-L., TSAI, W.-T., CHEN, G.-Z., LEE, C.-J., CHIU, C.-M., ET AL. mirtarbase: a database curates experimentally validated microrna–target interactions. *Nucleic acids research 39*, suppl_1 (2010), D163–D169.

[13] 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.

[14] KIM, D., SUNG, Y. M., PARK, J., KIM, S., KIM, J., PARK, J., HA, H., BAE, J. Y., KIM, S., AND BAEK, D. General rules for functional microrna targeting. *Nature genetics 48*, 12 (2016), 1517.

[15] KOTSIANTIS, S., KANELLOPOULOS, D., PINTELAS, P., ET AL. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering 30*, 1 (2006), 25–36.

[16] KUDLA, G., GRANNEMAN, S., HAHN, D., BEGGS, J. D., AND TOLLERVEY, D. Cross-linking, ligation, and sequencing of hybrids reveals rna–rna interactions in yeast. *Proceedings of the National Academy of Sciences 108*, 24 (2011), 10010–10015.

[17] LECUN, Y., BENGIO, Y., AND HINTON, G. Deep learning. *nature 521*, 7553 (2015), 436.

[18] LEE, B., BAEK, J., PARK, S., AND YOON, S. deeptarget: end-to-end learning framework for microrna target prediction using deep recurrent neural networks. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (2016), ACM, pp. 434–442.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] MENOR, M., CHING, T., ZHU, X., GARMIRE, D., AND GARMIRE, L. X. mirmark: a site-level and utr-level classifier for mirna target prediction. *Genome biology 15*, 10 (2014), 500.

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

[26] 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.

[27] PLA, A., ZHONG, X., AND RAYNER, S. miraw: A deep learning-based approach to predict microrna targets by analyzing whole microrna transcripts. *PLoS computational biology 14*, 7 (2018), e1006185.

[28] 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.

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

[30] 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.

[31] VLACHOS, I. S., PARASKEVOPOULOU, M. D., KARAGKOUNI, D., GEORGAKILAS, G., VERGOULIS, T., KANELLOS, I., ANASTASOPOULOS, I.-L., MANIOU, S., KARATHANOU, K., KALFAKAKOU, D., ET AL. Diana-tarbase v7. 0: indexing more than half a million experimentally supported mirna: mrna interactions. *Nucleic acids research 43*, D1 (2014), D153–D159.

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

[33] WEN, M., CONG, P., ZHANG, Z., LU, H., AND LI, T. Deepmirtar: a deep-learning approach for predicting human mirna targets. *Bioinformatics 34*, 22 (2018), 3781–3787.

[34] YATES, A., AKANNI, W., AMODE, M. R., BARRELL, ET AL. Ensembl 2016. *Nucleic acids research 44*, D1 (2015), D710–D716.

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