FACULTEIT BIO-INGENIEURSWETENSCHAPPEN

Academiejaar 2012–2013

THE PREDICTION OF INTERACTION BETWEEN mRNA AND miRNA USING MACHINE LEARNING TECHNIQUES

Ayla De Paepe

Promotoren: Prof. Dr. Bernard De Baets
Dr. Willem Waegeman
Tutor: Ir. Michiel Stock

Masterproef voorgedragen tot het behalen van de graad van
MASTER IN DE BIO-INGENIEURSWETENSCHAPPEN: CEL- EN GENBIOTECHNOLOGIE

Copyright

Acknowledgements

To me, this thesis represents the cherry on a huge pie of studying. I must admit, not all parts of this pie were as sweet as other, and sometimes I wished I chose a humble cookie instead. However, my love for science and the joy of understanding and questioning the things we encounter in our daily lives has only grown during these past five years.

The growing of this thesis cherry required more than just my own love and time, and I would like to take this opportunity to thank some people who helped me in the realisation of it. First of all, I would like to thank Michiel and Willem, who acted as the light feeding and guiding the growth of my very own research tree, which will hopefully bear many more juicy cherries. I would also like to thank Prof. De Baets for providing me with ideal growing conditions. Since growing is impossible without the necessary nutritions, I would like to thank Gert Van Peer and Prof. Vandesompele for making their data available to me.

On a more emotional basis, I would like to thank my parents for creating an ideal soil for my science tree to root in, and Laurens, who's love and attention supported me during the whole process.

For the realisation of this thesis, I would also like to thank my computer, who, in a way, has done most of the work.

Samenvatting

Dit werk beschrijft ons onderzoek omtrent het voorspellen van de interactie tussen messenger RNAs (mRNAs) en micro RNAs (miRNAs) met behulp van machine learning methoden. Deze biologische vraagstelling is van groot belang in biomedisch onderzoek in het algemeen en bij kankeronderzoek in het bijzonder (Farazi et al., 2011). De interactie tussen mRNAs en miRNAs is namelijk één van de wijzen waarop genexpressie gereguleerd wordt in een cel (Pillai et al., 2007; Costa, 2010), en disregulatie van normale cellulaire processen kan tot problemen als kanker leiden (Sato et al., 2011; Taft et al., 2010).

In de hoop nieuwe inzichten in het kanker proces te verwerven en bij uitbreiding nieuwe therapieën te ontwikkelen wordt veel onderzoek gedaan naar welke mRNAs en miRNAs met elkaar interageren (Gong et al., 2005; Iorio and Croce, 2012). Het testen van deze interacties in labo experimenten is echter duur en tijdrovend, waardoor het interessant is een slimme experimentele opzet te hanteren, en niet blindelings alle mogelijke combinaties uit te testen. Om een dergelijke opzet te bekomen kunnen voorspellingsalgoritmen gebruikt worden (Rajewsky, 2006), die onder andere op machine learning technieken gebaseerd kunnen zijn. Deze algoritmen kunnen dan *in silico* predicties maken van welke mRNA-miRNA combinaties een hoge waarschijnlijkheid hebben te interageren, zodat een veelbelovende selectie van combinaties in labo experimenten gevalideerd kan worden.

Tot dit doel zijn al meerdere algoritmen ontwikkeld (Zotos et al., 2012), maar er is duidelijk nog verbetering mogelijk, vooral op het vlak van het aantal mRNA-miRNA combinaties die ten onrechte als interagerend voorspeld worden. Er zijn echter een aantal praktische moeilijkheden verbonden aan de structuur van het probleem dat we wensen te onderzoeken. We willen namelijk een voorspelling maken omtrent een relatie tussen twee moleculen, hetgeen minder voor de hand liggend is dan een voorspelling betreffende één type van moleculen (Waegeman et al., 2012). In het geval van een relatie kunnen er zowel invloeden van de beide moleculen zelf van belang zijn, als invloeden eigen aan de interactie tussen beiden. Deze onderliggende structuur heeft ook grote invloed op hoe deze methoden geëvalueerd dienen te worden, wat ook in dit werk behandeld wordt.

Ondanks dat de informatie zowel van het niveau van de moleculen als van het niveau van specifieke interactiemogelijkheden afkomstig kan zijn, willen we voor elke mRNA-miRNA combinatie slechts één conclusie trekken: zal deze interactie vertonen of niet. De kunst zit er in een model te ontwikkelen dat correct omgaat met deze structuur, dit zowel op biologisch als machine learning vlak. In de literatuur zijn modellen te vinden waarbij op een bepaald deel van de beschikbare informatie gestoeld wordt (Wang and El Naqa, 2008), om op deze wijze de complexe structuur van het probleem te omzeilen. Het lijkt echter wenselijk een model te ontwikkelen dat meer, en liefst alle, data kan gebruiken om zijn voorspellingen op te baseren. Het maken van een dergelijk model was dan ook het doel van dit onderzoek. We evalueerden

verschillende mogelijke oplossingen onder de vorm van uiteenlopende model-structuren.

Welk van deze modellen de voorkeur draagt is afhankelijk van de exacte onderzoeksvraag. Over het algemeen kan echter geconcludeerd worden dat onze modellen goede resultaten bekomen bij het vergelijken met enkele veel gebruikte modellen beschreven in de literatuur.

Contents

Chapter 1

Introduction

The increasing availability and decreasing cost of molecular analyses results in an ever increasing amount of information (Mardis, 2011). Although brave researchers used to construct sequence alignments by hand, it is now common practice to use computer programs for these kinds of tasks (Nayeem et al., 2006; Sauder et al., 2000). The greater our insight in biological systems, the complexer the problems we want to analyse become. This gives rise to the need of computer algorithms that can cope with such problems, including the vast amount of complex data that comes with it. In some cases, this is no longer possible with algorithms explicitly programmed by humans, since this implies that one must perfectly understand the problem one wants to solve, including everything that might influence it. So, why not take the use of computers a step further, and make them able to learn so they might deduce information we ourselves are not able to? One must appreciate the simplicity of the term invented for this concept: machine learning. The idea of machines that are able to think and learn is not new, although public opinion generally does not consider this as a good thing. Sadly more movies are released handling the terrible nightmare of the technological singularity, when robots become self aware and smarter than humans, than about how great this technology can be. Luckily, but perhaps also sadly, most people do not know how far computer science has already evolved. Now, the good news: the past half century has brought us an enormous number of exiting machine learning applications that you encounter daily, maybe even without knowing it. Machine learning applications can be used to read text (Sebastiani, 2002), detect fraud (Fawcett, 1997), drive cars in real traffic situations (Desouza and a.C. Kak, 2002), detect spam email (Guzella and Caminhas, 2009), beat champion players in games and quizzes like *Jeopardy!* (Ferrucci et al., 2010) and so forth. Still, no need to worry, we are not reaching the technological singularity yet.

When we turn to applications in bio-informatics, machine learning can be a valuable tool in, for example, biomarker discovery (Calvo et al., 2010; Lyons-Weiler et al., 2003). Here, one could try to link the absence or presence of a molecule to an outcome of a disease or therapy. In this work, we have used machine learning techniques to predict the interaction between messenger RNAs (mRNAs) and micro RNAs (miRNAs). This situation is more complicated, since we are not trying to predict an outcome based on features, but a relation between two molecules. This relation can be influenced by characteristics of both the molecules themselves and the specific combination of these molecules. It is interesting to spend a moment thinking about the size of it all: mRNAs and miRNAs have sizes measurable in nanometres, which is 10^{-9} m or one

billionth of a meter. In order to translate this to a more daily context, think of it in this way: if one meter would be as big as the diameter of the earth then a nanometre would be about the size of a pebble. So, when thinking about the human body and mRNA or miRNA molecules, keep the earth and the pebble in mind.

The prediction of mRNA-miRNA interaction is of great interest to medical researchers, since it can regulate gene expression and thus influences many cellular processes (Pillai et al., 2007; Costa, 2010). Although cells are under constant regulation, sometimes things go wrong, which might cause diseases such as cancer (Sato et al., 2011; Taft et al., 2010). As a consequence, researchers wish to know which mRNAs are influenced by which miRNAs. This information might allow them to guide cell processes and hopefully prevent or cure illnesses Tong and Nemunaitis (2008). However, since wet-lab tests to validate this interaction can be very expensive, it is not desirable to test all possible combinations of mRNAs and miRNAs one might be interested in. Predictive models can be used to make *in silico* predictions of the interactions for a large number of mRNA-miRNA combinations, allowing researcher to test only the most promising combinations in wet-lab tests. In literature, multiple algorithms are described that can make such predictions (Zotos et al., 2012), some of which are also machine learning based models. However, the performance of these algorithms is often rather poor, especially when considering the high number of false positives. With the research described in this work, we try to improve this performance. Due to the structure of this unusual relation problem, some of the current methods choose to focus on a part of all available information, further explained in Chapter 4, in order to simplify the problem. In search of a better solution, we explore some approaches that can enable us to use more, if not all, of the available information.

The biological background on mRNA and miRNAs, including how they regulate the functions of a cell in the human body, is given in Chapter 2, while Chapter 3 provides a basic introduction to the machine learning side of this work. Both chapters are written as accessible as possible, allowing one to understand the basics of both topics, even if these are not one's speciality. In this light, Chapters 2 and 3 are more or less standalone and written independently form the rest of this work. However, readers are advised to go trough the chapters outside their field of study, in order to understand the "why" and 'how" of this work. Chapter 4 describes the dataset we have used for this work and Chapter 5 handles upon the models we have constructed. Chapter 6 summarises the results and discussion of this research. Some general conclusions and perspectives are included in Chapter 7. We hope all this can cast a light on the opportunities, and the pitfalls, of combining biological and technological research.

Chapter 2

Epigenetic regulation of gene expression mediated by miRNA

2.1 Introduction

The goal of this chapter is to give a general biological introduction to the topics concerning this master thesis, revolving around micro RNA (miRNA) and its influence on gene expression. After a brief introduction to the basic molecules and mechanisms of molecular biology, further information is given on the epigenetic regulation of gene expression and the role miRNAs play in this process. In a subsequent section, the measurement of miRNA and its effects will be discussed. The chapter will be concluded with some interesting research topics and applications involving miRNA.

2.2 Meet the molecules

In order to provide the necessary background to the biology behind this research, some of the basic molecules of life are mentioned. The two types of molecules that are most important for the remainder of this work are nucleic acids and proteins.

2.2.1 Nucleic acids

Nucleic acids are biopolymers that contain the genetic information of an organism. They consist of only a few basic nucleotides: adenine (A), tyrosine (T), guanine (G), cytosine (C) and uracil (U). Combining these nucleotides in long sequences makes it possible to store enormous amounts of complex information. Such a sequence of nucleotides is also called a strand, of which the start is called the 5' end and the end the 3' end. The two most important types of nucleic acids are DNA and RNA, each having their own structure and function.

DNA The hard copy of the genetic information in a cell is stored as DNA inside the nucleus. The nucleotides present in DNA are A, T, G and C, which are organised in a double helix, as can be seen in Figure 2.1. This structure is stabilised due to the formation of base pairs between

the nucleotides of the different strands of the DNA helix: bonds can be formed between A and T and between G and C. This means that both strands of the helix hold the same information, but in complementary sequences.

RNA A less bulky and more versatile version of DNA is RNA. It generally consists of only a single strand of nucleotides, making it smaller and less stable than DNA and allowing it to bind to other molecules. In addition to single stranded RNA (ssRNA), double stranded RNA (dsRANA) also exists, and is a typical intermediate form during viral replication in a cell. Although largely similar, the nucleotides present in RNA are not all the same as those in DNA: T is replaced by U.

Many different types of RNA exist, all fulfilling different roles in the cell, but we will limit ourselves to those relevant to the processes regarding miRNA. This summation can be seen as a general overview of the RNAs related to this work, some functions will only be explained in Section 2.3.

- *Messenger RNA* (mRNA) is used to transport the information encoded in the DNA out of the nucleus. It is formed by transcription of the DNA, and functions as a template to form a protein during transcription. The information stored in RNA is encoded based on groups of three nucleotides, called a codon.
- *Transfer RNA* (tRNA) is used to link nucleotide sequences with their corresponding amino acids during translation of mRNA into proteins. The bottom of a tRNA molecule holds an anti-codon, a nucleotide sequence complementary to a specific codon. The correct amino acid can bind to the top of the tRNA, linking the correct codon to the correct amino acid.
- *Ribosomal RNA* (rRNA) is RNA that is present in ribosomes. These ribosomes are a type of cellular machinery, made up of both rRNA and proteins, which mediate translation of mRNA into proteins. They are formed by the binding of two parts, called the small ribosomal subunit and the big ribosomal subunit.
- *Micro RNA* (miRNA) is a small strand of non-coding RNA, meaning it does not form proteins. Most miRNAs are about 21 nucleotides long and play important roles in the regulation of gene expression. In plants, miRNAs usually form a perfect match with their target mRNAs, whereas in animals this is not the case. Here, the region at the 5' end of the miRNA is considered to be the most important region, holdings patches where miRNA and mRNA are complementary called “seeds".
- *Small interfering RNA* (siRNA) is a second type of small non-coding RNA. They generally originate form exogenous dsRNA, meaning they where not formed in the cell by transcription of DNA but, for example, introduced by viral infection. They are about the same size as miRNA, but form a perfect match with their target mRNA. In contrast to miRNAs, siRNAs target a very limited number of mRNAs, usually only one.

2.2.2 Proteins

Proteins are another type of biopolymers and are the workhorses of the cell. They are made up of amino acids, of which there are 21 common types. Proteins have structural or active functions, other than storing genetic information, such as catalysing reactions and transferring

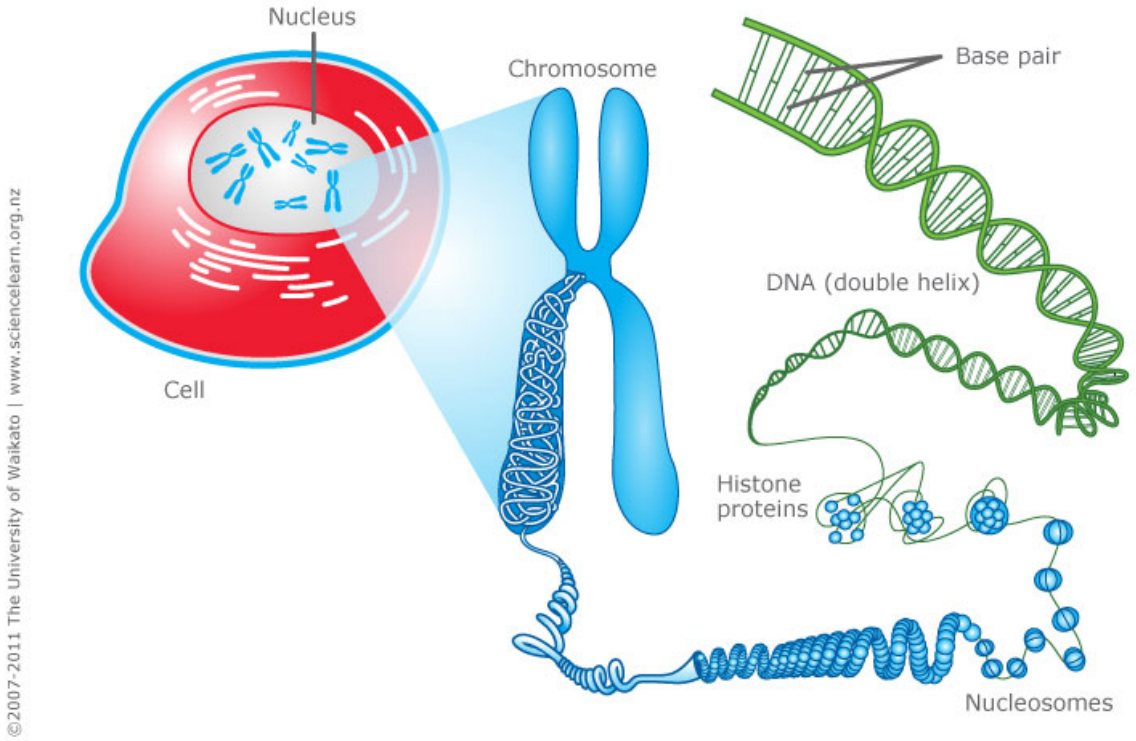


Figure 2.1: Overview of DNA in the cell. The DNA is present in the nucleus of the cell, wound around histon proteins. Together they are organised in nucleosomes and can be structured as chromosomes (sciencelearn.org).

other molecules. Their structure is defined at four levels. The primary structure is made up of the sequence of amino acids and will define the subsequent structures due to the availability of side chains of the amino acids. The secondary structure is defined as the way the strand of amino acids folds together in helices or sheets. The tertiary structure is defined by the final folding of one strand of amino acids, whereas the quaternary structure is the structure obtained if all the necessary subunits come together and form one big functional protein. Predicting this fold structure of protein sequences is one of the major challenges in structural bioinformatics.

Histones are of particular interest in this work, as they are involved in epigenetic regulation of gene expression. Histones are globular proteins that bind DNA and give structure to it, as can be seen in Figure 2.1. The complex of DNA and histone proteins is called chromatin (Cooper, 2000). Depending on how this binding occurs and how the histones are modified, DNA can either be packed very tightly or can be more accessible. This altering of DNA accessibility is a form of epigenetic regulation.

2.3 Central dogma of molecular biology

The central dogma of molecular biology describes how information is transferred from DNA through RNA to proteins, as illustrated in Figure 2.2, and is extremely relevant to biology, biotechnology and bioinformatics. As mentioned above, DNA holds the genetic information of an organism, but this information needs to be transcribed into mRNA in order to be translated

into proteins. It implies, however, that the study of DNA and its products can give insights in all molecular processes taking place inside a cell.

Replication Replication is the process in which DNA is copied and occurs whenever cells duplicate to make sure each of the daughter cells has the same amount of genetic information. The two strands of the original DNA are locally separated and are used as templates for the new DNA. New nucleotides are paired with the ones in the template strands by a protein called DNA polymerase. This results in two identical copies of the original DNA, each having one original strand and one newly generated one.

Transcription The process in which an RNA strand is formed using DNA as a template is called transcription. Like in replication, the two DNA strands separate locally by means of a protein complex. In this case a protein called RNA polymerase scans one of the DNA stands and generates a complementary mRNA strand.

Translation To create a protein, the mRNA needs to be translated. Three subsequent nucleotides of a mRNA molecule are called a codon. Each codon corresponds to one amino acid of the future protein. Since nucleotides of mRNA cannot bind to amino acids themselves, an intermediate molecule is needed. This molecule is tRNA, which has the nucleotides complementary to a specific codon on one side, and the corresponding amino acid on the other. When a mRNA molecule is to be translated, it needs to reach a ribosome. This ribosome will scan the mRNA strand, allowing one codon at a time to interact with its complementary tRNA. In this way all the necessary amino acids are provided and linked, one by one, in the correct order, resulting in a self-assembling protein.

2.4 Epigenetic regulation of gene expression

Epigenetics is a term for all processes that influence gene expression or cellular phenotype without changing something to the nucleotide sequence of the DNA or RNA involved (Goldberg et al., 2007). Although most of these processes were until recently considered mysteries, general principles are being discovered thanks to the growing research in this field.

DNA methylation and chromatin remodelling A fundamental way to influence the expression of a gene is by regulating transcription. To be transcribed, the DNA in the region of the gene of interest needs to be accessible for transcription factors (Cooper, 2000). This is determined by the DNA and chromatin structure in that region, which can be altered by DNA methylation and histone modification, respectively. In the case of DNA methylation, methyl groups are added to Cytosine residues in the DNA, preventing transcription. This methylation pattern is called an imprint and is stably passed on during cell division. In the case of histone modification, different outcomes are possible, depending on the modifications that are present. Acetylation of certain amino acids will relax the chromatin structure, making transcription possible, whereas other modifications can result in the silencing of the genes in that region.

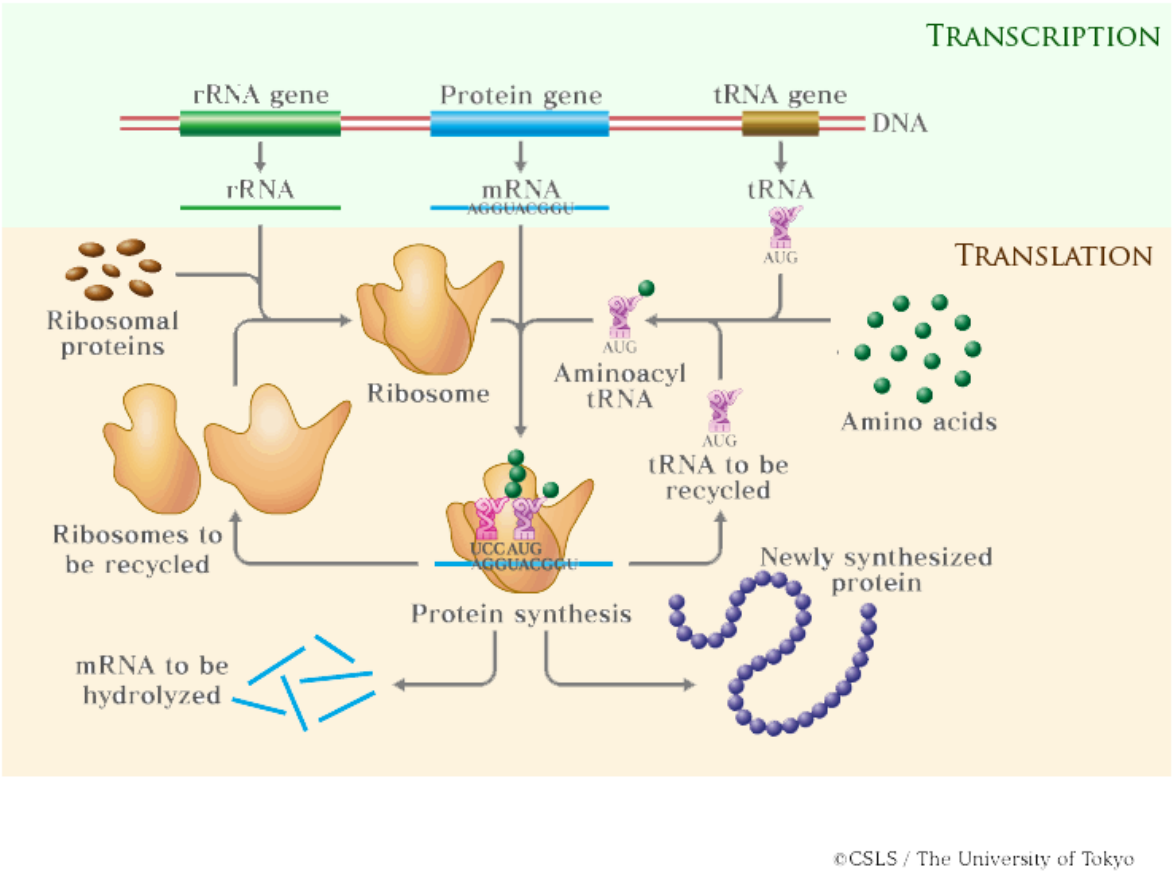


Figure 2.2: Overview of translation and transcription of the central dogma of molecular biology (based on *http://csls-text.c.u-tokyo.ac.jp/active/03_02.html*).

RNA interference Whenever cells detect dsRNA they will try to break it down due to the resemblance to viral dsRNA. The cell also uses this mechanism to regulate the expression of its own genes by producing small strands of RNA, including miRNA, which can bind to mRNA present in the cell, forming dsRNA (Bartel et al., 2004). This process is called RNA interference (RNAi). The dsRNA will be recognised by the cell as alien and will be degenerated in the same way as if it were viral dsRNA. The dsRNA, whether originating from viruses or from endogenous miRNA, will be cut into smaller pieces. Each strand can subsequently bind to multiple proteins and form an RNA-induced silencing complex (RISC). This complex can bind to target mRNAs that are complementary to the miRNA strand incorporated, allowing the proteins of the complex to cleave these mRNAs.

Binding of miRNA to mRNA can also lead to inhibition of translation. In this case the mRNA-miRNA duplex is not degraded, but normal translation in the ribosomes cannot occur due to the hindrance of protein binding to the mRNA. This mRNA will thus not be translated and will be degenerated.

Since some miRNA targets are genes needed for chromatin structure remodelling like DNA methylation and histon modification, miRNAs can indirectly regulate these processes (Sato et al., 2011). Regulation also works the other way around: the regions of the DNA that encode miRNAs can be made accessible or not, due to the structure of the chromatin in that region, and

thus DNA accessibility regulates miRNA formation. An overview of the epigenetic regulation by miRNAs and their own epigenetic regulation can be seen in Figure 2.3.

2.5 Measuring miRNA expression

To analyse miRNAs one needs to be able to measure their levels in a cell. However, this is not enough to grasp their effects. To actually detect the regulation due to miRNAs one needs to measure the difference in concentration of the mRNAs or proteins it regulates. An analysis generally goes as follows: one must first measure the level of mRNA or protein in the relative absence of the miRNA. Secondly, the miRNA must be overexpressed in the cell, after which the measurement of the mRNA or protein levels need to be repeated, allowing some time for the regulation to take place. If only the level of miRNA is changed in this second setting, the observed difference in mRNA or protein level is due to the regulation by miRNA. Note that the measurement of mRNA will only show the influence due to mRNA degradation, whereas protein measurement will also show the influence due to blocking of transcription.

2.5.1 RNA-based techniques

The techniques used to detect miRNAs and mRNA are generally the same, although the detection of miRNA is harder due to its size and lower stability (Cissell and Deo, 2009).

Northern blot A sample containing RNA is loaded on a gel and separated by electrophoresis. After separation, the RNA is transferred to a blotting membrane. Bands of RNA can be visualised by hybridisation with a marked RNA strand, complementary to the RNA that has to be detected. Notwithstanding a simple technique, northern blotting is labour intensive and not suitable for high throughput analysis.

qRT-PCR Polymerase chain reaction (PCR) is generally used to amplify DNA. This method consists of multiple cycles in which a selected part of the DNA present in the original sample is exponentially amplified. To deal with RNA, however, reverse transcriptase PCR (RT-PCR) is needed. Since only DNA is stable enough to be used in PCR, the RNA needs to be transcribed to complementary DNA (cDNA) by a reverse transcriptase enzyme. This cDNA can then be amplified as normal DNA would be during PCR. If quantification is wanted however, a normal RT-PCR is not enough, since it will only show the presence of an RNA molecule. Quantitative reverse transcriptase PCR (qRT-PCR) however makes this possible. Levels of cDNA are measured in real time during amplification, using fluorescent probes that are only detectable if incorporated in a synthesised DNA strand. Based on this information, the original level of RNA can be determined through comparison to the result of known levels of a DNA standard.

Microarray In order to measure many RNA molecules at the same time, microarrays are a good option. For each RNA molecule that has to be testes, small complementary DNA (cDNA) strands are generated. All these cDNA types are fixed on a substrate while making sure that the exact spot of every type of cDNA is known. A sample containing labelled RNAs is brought

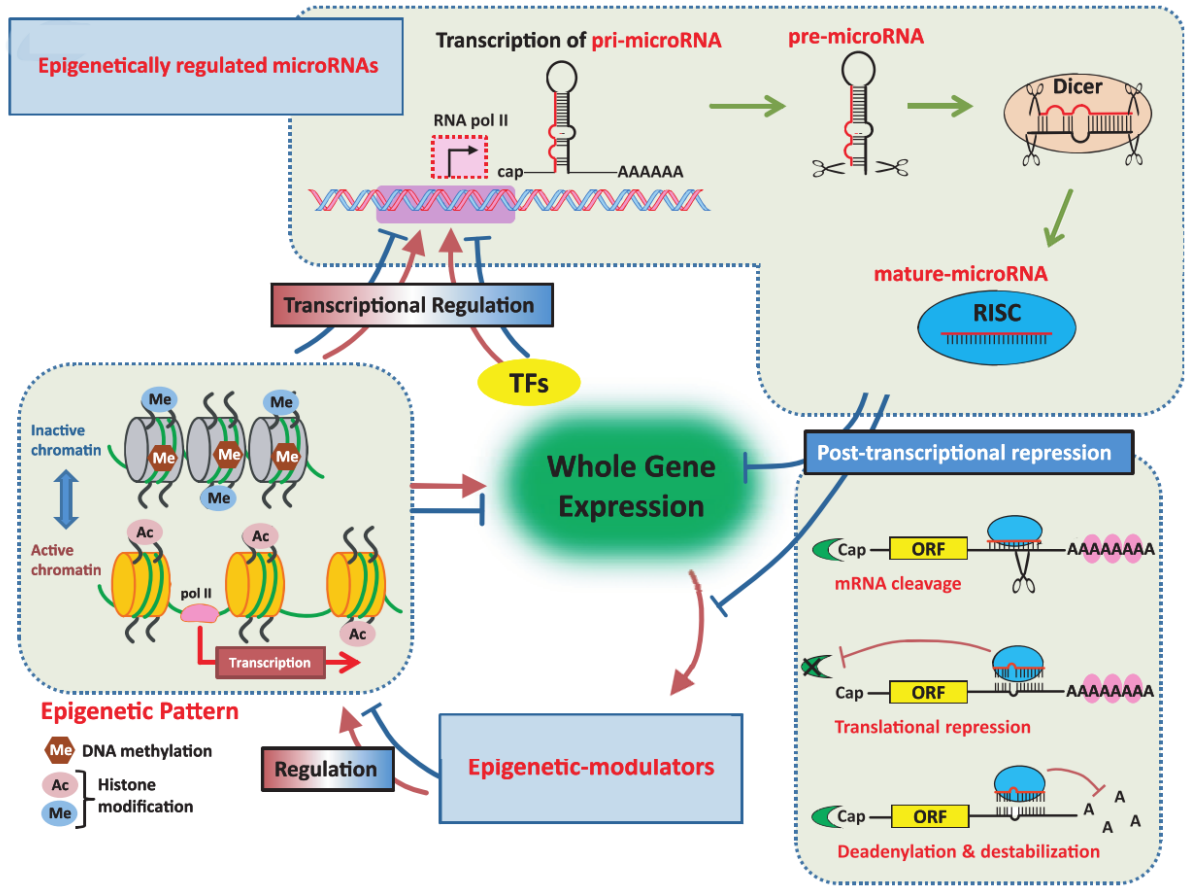


Figure 2.3: Overview of the role of miRNAs in epigenetic regulation. Starting at the top and evolving clockwise: miRNAs are transcribed from DNA by RNA polymerase II, resulting in a pre-miRNA stemloop. This loop is processed and cut by Dicer, after which it is included in an RNA-induced silencing complex (RISC). This complex can now target mRNAs complementary to the embedded miRNA, resulting in post-transcriptional regulation by mRNA cleavage, transcriptional repression or destabilisation. If the targeted mRNAs are coding for epigenetic regulators, such as proteins that perform DNA methylation, miRNA can indirectly regulate DNA availability. To complete the circle, the DNA regions coding for miRNAs have to be available and the right transcription factors (TF) have to be present in order for the miRNA to be transcribed (based on Sato et al. (2011)).

onto this microarray. After some time in which hybridisation can take place, the microarray is washed and only the molecules with strong interactions, indicating full complementarity, will be retained. By analysing the strength of the signal on each spot of the microarray, one can determine the level of each RNA present in the sample for which there was a cDNA on the microarray.

2.5.2 Protein-based techniques

To grasp the full influence of miRNA-mediated regulation of gene expression, the proteins encoded by the gene of interest need to be quantified. However, one has to keep in mind that there is a time gap between miRNA expression and the visibility of its influence. Multiple detections in time might be needed.

General techniques Many techniques can detect total protein level of a sample. There is for example the Bradford assay, which uses a dye called Coomassi Blue that changes colour when bound to proteins. By spectroscopically analysing the colour of the sample, one can determine the total level of proteins. If one wants to measure the level of a specific protein however, purification is needed before these methods can be uses. This can be very time consuming and labour intensive and good results cannot always be obtained.

Using antibodies Antibodies are proteins made by the cell to specifically bind antigenes, which are recognizable parts of other proteins. The cell can use these antibodies to, for example, recognize proteins of pathogens. Antibodies can also be used to detect proteins of interest in research. Methods using this principle are for example sandwich-ELISA and western blot.

In a sandwich-ELIAS essay, antibodies that bind the protein of interest are fixed to a substrate. The sample is added and if the target protein is present, it binds to the antibodies. This result is the capture of the protein of interest, but to detect it a second antibody that binds this protein needs to be added. This antibody is tagged with an enzyme, that can catalyse a reaction changing the colour of a substrate. After proper washing and addition of the substrate, the level of protein can be determined based on the colour of the sample.

The second method mentioned, western blotting, is the protein equivalent of northern blotting, described above. In this case however, the protein of interest is visualised by the interaction with a labelled antibody instead of a labelled nucleotide sequence. Although these methods might be able to selectively detect a protein of interest, design and production of the needed antibodies is very costly.

Using a reporter gene To prevent the use of costly consumables like antibodies, one can use a reporter gene, which has easily detectable gene products such as enzymes that can catalyse reactions like producing light signals. If a reporter gene, under influence of the same regulatory sequences as the gene of interest, is introduced in the cell, the reporter gene will have the same expression as the gene of interest. By measuring the signal generated by the proteins derived from a reporter gene, one can determine its expression level.

2.5.3 Bioinformatics-aided testing

Although all methods have their strong and weak points, good wet lab experiments are either labour intensive, expensive or time consuming, and often all of these. This limits the possibilities to detect new miRNA targets or perform other large-scale experiments. Thanks to bioinformatics, more sensible experiments can be designed, performing only the tests that actually have potential according to *in silico* predictions. However, the performance of these algorithms depends on the quality of the data used to build them, which originate from wet lab experiments themselves. This brings us to the challenge of collecting informative data and making good, reliable predicting algorithms, since the success of the resulting research will be determined by their performance.

2.6 Applications in miRNA research

Recent years have witnessed an explosion of miRNA research. In this section a few topics and applications will be highlighted to illustrate the diverse functions of miRNA.

Plants and viruses In plants, the miRNA pathway is part of a very efficient defence mechanism against pathogens. When a cell is infected with a virus and dsRNA is formed, the cell will recognise this viral dsRNA and degenerate it. In this way the virus can be stopped from reproducing before spreading. This process occurs in both animal and plant cells, but since the latter have plasmodesmata, small channels connecting neighbouring cells, its impact is larger in plants. Here, the produced miRNAs can diffuse through the plasmodesmata and spread throughout the whole plant, preventing other infections from the same virus anywhere in the plant (Voinnet, 2001). Obviously, evolution has resulted in viruses that can fool plants and are not targeted by certain processes, resulting in a never ending battle to be one step ahead of the other.

Metabolism miRNAs have been shown to have widespread regulatory influence on multiple aspects of the metabolism (Heneghan et al., 2010). The presence or absence of some miRNAs can lead to health problems and have been linked to diabetes, because of their role in insulin production and uptake (Poy et al., 2007). Other miRNAs can regulate the differentiation of fat cells, and thus influence fat storage and obesity (Esau et al., 2004).

Memory and behaviour The idea that gene regulation facilitated the creation of memories and influences our behaviour is not new (Blaze and Roth, 2013). However, the underlying mechanisms are still not very clear. Recent research has shown that experiences can influence epigenetic regulation by miRNAs and remodel DNA. For example, this adaptation can be observed in rodents after exposure to different behavioural tests, involving fear conditioning, novel object recognition and spatial memory tasks.

Cancer Because of the major interest of public health, most of the research concerning miRNA has some connection to cancer. As mentioned in the previous paragraphs, miRNA and other epigenetic mechanisms can influence the expression of many kinds of genes. Since cancer is a state in which the cell loses control of its division rate, it is easy to see that this can also be mediated by altering gene expression (Croce, 2008). Specific genes of interest are genes promoting growth and division on one the hand, and gene controlling the cell cycle on the other hand. The first are called oncogenes, although their expression is also needed in healthy cells, because of the tendency to induce cancer if overexpressed. The second type are tumor suppressor genes, since the presence of their gene products reduce the chance of cancer induction. Research has shown that multiple mRNAs regulate the activity of these genes, and thus can induce or prevent cancer (Iorio and Croce, 2012).

Medicine Since one miRNA regulates multiple mRNAs and even whole pathways, miRNAs find more general application in biotechnology than siRNAs, which only regulate one gene. This is both an advantage and a disadvantage compared to siRNA. An advantage, because

complex processes involving multiple genes and pathways can now be influenced, which is nearly impossible with siRNAs. But also a disadvantage, since great caution has to be taken only to influence the desired processes. Different approaches are possible, since miRNAs can be used as a biomarker, a target and a tool.

- *Biomarker* A molecule is called a biomarker if its presence or absence can be linked to a specific type, state or outcome of, for example, a disease. It is often very hard to determine the exact cancer type, the responsiveness of a patient to a specific type of therapy and the overall survival expectancy. Multiple miRNAs have been shown to be good biomarkers, indicating for example poor survival chances (Yanaihara et al., 2006) or relapse free recovery (Li et al., 2010) in cancer therapy if present.
- *Target of therapy* If the overproduction of mRNA causes cell disregulation resulting in an illness, degradation of this mRNA by the introduction of antisense RNA can restore normal functioning (Iorio and Croce, 2012). This application is based on the RNAi mechanism discussed earlier, in which dsRNA is degraded. Breaking down miR-21, for example, has been shown to reduce tumor development and spreading in breast cancer tissue (Si et al., 2007).
- *Therapeutic tool* If a miRNA is down regulated or just not present in a specific cell type, resulting in illnesses, it can be introduced as therapeutics. Reintroduction of miR-15a or miR-16-1, for example, results in the programmed cell death of leukaemic tumor cells in mice (Calin et al., 2008).

Chapter 3

Machine learning overview

In this chapter we will give an intuitive introduction to machine learning, using examples where possible. After the explanation of some basic machine learning concepts, we explain the algorithms used in this work. The most important part of this chapter is the introduction of some terminology, tools and tricks to estimate the performance of a model.

3.1 Introduction

The term “machine learning‘” has received many definitions, but one of the first was made by Arthur Samuel, being: “Field of study that gives computers the ability to learn without being explicitly programmed" (Samuel, 2000). Given the ambiguity of what the terms “thinking" and “machines" actually imply, Turing proposed not to ask ourselves “Can machines think?", but “Can machines do what we (as thinking entities) can do?". This detachment results in more formal definitions, like the one by Mitchell: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell, 1997).

3.2 Learning types

The type and availability of the data will greatly influence the type of learning and methods one will use, but before we come to this, we explain some machine learning terminology used in this work. When one has a dataset on which to use machine learning methods, every observation or record in this dataset is called an *instance*. The variable we want to predict is called the *label* or *output*, while all other variables are *features* or *inputs* (Hastie et al., 2009). These features hold the information that we want a model to learn in order to make predictions for the label. The first thing to ask oneself is whether the label that has to be predicted actually is available for all instances in the dataset. Depending on this availability, two main groups of machine learning algorithms can be defined: *supervised* and *unsupervised* learning algorithms (Larranaga, 2006).

3.2.1 Supervised machine learning

Supervised algorithms need the observed label of every instance in order to learn, since they will use this label to build a model. This makes supervised learning very similar to how we learn ourselves. To illustrate this, consider the following example: you want to predict whether it will rain tomorrow, based on the weather of today. To do this, you daily check if it rains and whether this can be linked to the weather of the day before, making every day an instance with as label: "rain" or "no rain". In other words, you use the label of a particular day to see if there is a link with the features regarding the weather on the previous day, making this supervised learning. From your own experience, you have made up some rules on which you base your predictions: there is a good chance that the weather of tomorrow will be the same as that of today. However you have noticed that there is a higher chance on rain the following day if the wind is coming from the west. And finally, you consider the season, somehow you feel like there is more rain in autumn and spring than in the rest of the year. This way, you have actually made your own model to predict whether it will rain tomorrow, which is the label, based on three features: whether it rains today, the direction of the wind and the current season. However, it is doubtful whether this particular invented model will result in very good predictions. The good thing about leaving this task to a computer is that you can gather huge amounts of data on all kinds of features which might be important, and let the algorithm decide how to use and combine them, probably resulting in better and more complex models than the one you would build yourself. Depending on the type of label, i.e. categorical or quantitative, different approaches are needed, which will be illustrated by the rain prediction example.

Classification If the label that one wants to predict is categorical, one will use a classification model. In this case, the label can take on two or more values, representing the class of an instance. The terms *binary classification* and *multi-class classification* are used when the label has respectively two or more than two classes. Consider again the rain prediction problem: we want to make predictions for a label with two classes, in this case: "rain" or "no rain", and thus we will use binary classification. However, if we would like to discriminate between rain, snow and hail, this example becomes a multi-class classification problem with four classes: "no precipitation", "rain", "snow" and "hail".

Ranking A ranking model will, as the name suggests, rank all instances for which it has to make predictions. Assume the rain prediction model is valid for different locations, for example: quite accurate predictions can be made regarding the rain in Paris, Brussels and Amsterdam. You want to go on a last-minute city trip tomorrow and the city you will visit will be the one with minimal rainfall. So you use a ranking model to rank the expected rainfall in each of the three cities and the one which ends up as the bottom ranked city will be your destination. Note that this model does not tell you whether it will actually rain in that city tomorrow and it might be sunny or rainy in all three of them.

Regression A regression model predicts a quantitative label, so actual continuous values can be predicted. With such a model, you could not only decide where to go on your city trip, but also see how much the expected rainfall will be. The output could be the expected rainfall in cm/m^2, showing you whether it will rain and how much. Although this is very attractive,

it is also the hardest model to train, and a lot of data and insight in weather forecasting will be needed. Weather is also inherently chaotic, making this a complicated problem for any prediction algorithm.

3.2.2 Unsupervised machine learning

For the rain prediction problem, daily observations were made to see whether it rained or not, making the label available. In other cases, the true labels might be unknown and these will thus have to be handled with *unsupervised* machine learning methods. This can be the case when one has data on the shopping behaviour of customers in a supermarket, and one wants to categorise all customers in customer types based on related shopping habits. In this case, every customer is an instance, the label is the type of customers this person belongs to and the features might be related to shopping frequency and purchases. No customer categories are defined upfront, which makes it impossible to assign a category to each person in the dataset, even if one would be willing to spend a lot of time to do so. This also makes it impossible for an algorithm to use the real label in order to build a model, an thus unsupervised machine learning methods, such as clustering, will have to be used.

3.3 Main methods used

For this work, we have considered multiple machine learning methods such as linear regression, logistic regression, decision trees, random forests, boosting and support vector machines. However, only logistic regression and random forests will be discussed here, since these are the ones used in the final models.

3.3.1 Logistic regression

Logistic regression will fit a linear model by assigning weights to every feature, in such a way that the made predictions reflect a probability for the outcome (Hastie et al., 2009). We will explain some of the mathematics for a binary classification problem with one feature, x, where the label of an instance can either be 0 or 1. The feature x receives a weight: β_1, while the intercept, the value if $x = 0$, receives a value β_0. The linear combination thus becomes $\beta_0 + \beta_1 x$. Now, the logistic function must be applied, which looks as follows:

$$\sigma(t) = \frac{1}{1+e^{-t}} \tag{3.1}$$

which is visualised in Figure 3.1 for $\beta_0 = 0$ and $\beta_1 = 0.5$. If we replace t by our linear combination of x, this becomes:

$$y = \sigma(\beta_0 + \beta_1 x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \tag{3.2}$$

The outcome, y, is now the probability that the class for a given instance is 1.

In case of the rain prediction example, the model holds three features, i.e. the rain, direction of the wind and season today, which makes the model slightly more complex. However, since

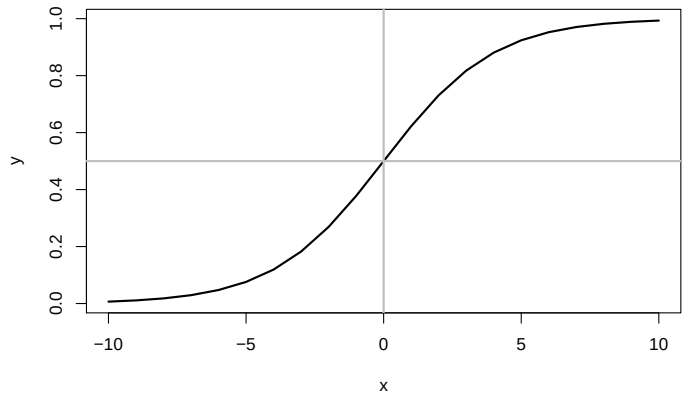


Figure 3.1: Logistic function for $\beta_0 = 0$ and $\beta_1 = 0.5$, which results in $y = \sigma(\frac{x}{2})$.

a real-life problem is easier to imagine, we will illustrate the same reasoning for this example, where 'rain" is class 1 and "no rain" class 0. Each of the three features will receive coefficients, respectively β_1, β_2 and β_3, which represent the influence of a feature on the chance of rain the following day. The linear combination of these features thus results in $\beta_0 + \beta_1 rain + \beta_2 wind + \beta_3 season$. Applying the logistic function gives the final model:

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 rain + \beta_2 wind + \beta_3 season)}} \tag{3.3}$$

The predictions made by the model indicate the predicted probability that it will rain the following day.

3.3.2 Random forests

In order to explain random forests, we will first introduce the classification tree, which is a machine learning algorithm on itself and can be combined to form random forests.

A classification tree makes subsequent binary splits of the dataset, based on the available features, in order to find rules that result in a good classification (Mitchell, 1997). The groups resulting from a split are called leaves, the full grown model a tree, due to the resemblance in structure. We will illustrate this with the rain prediction example. The data of all days is gathered in one big group, representing the stem of the tree. Now the algorithm will try to split this group according to the features. Splitting by whether or not it rained during the previous day, for example, results in one group where all instances have "rain" for this feature, and a second group where this feature is "no rain". When the algorithm has tried to split the data for all three features, it will actually use the feature that gives the "best" split. There are multiple measures that can be used in order to decide which split is best, but generally the "pureness" of the leaves, how well the two classes of the label are separated, is important. This principle will be repeated, and now the season might be most informative for a group. The algorithm will stop when no further splitting is possible or when every training instance is classified correctly.

In a tree, every decision depends on the previous splits, making this algorithm very susceptible

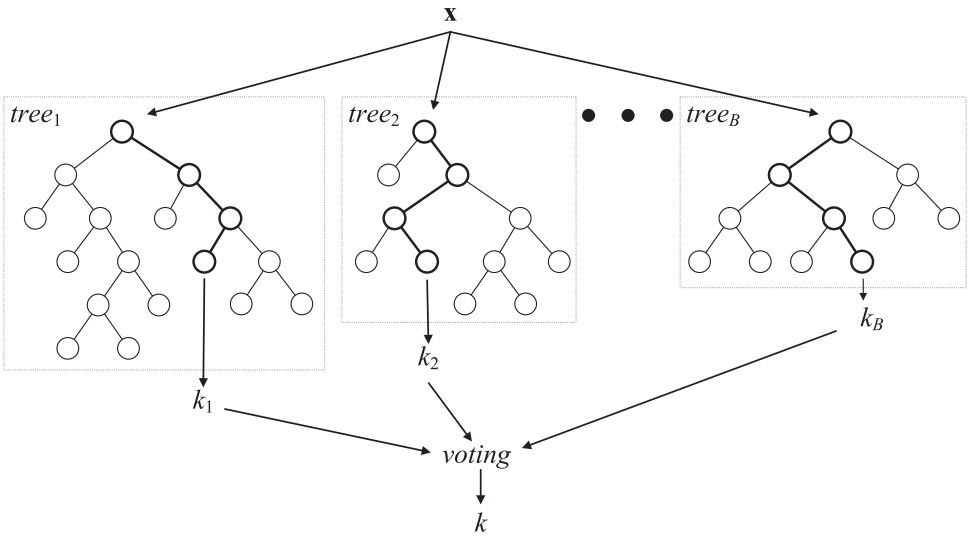


Figure 3.2: General architecture of a random forests model (Verikas et al., 2011). All models, B in total for this example, make predictions for an instance, denoted by k_b for the bth model. A majority vote is made on all these predictions, which results in the final class prediction denoted by k. For example: if three models are considered for a binary classification problem, the predictions made by these trees for a given instance might be $k_1 = 0$, $k_2 = 1$ and $k_3 = 1$. The random forests model will thus predict the outcome k for this instance to be 1.

to errors in the data: one instance misclassified in the dataset can result in a totally different tree. To solve this, the idea of democracy is used: build multiple models and let them vote in order to decide on the outcome, as illustrated in Figure 3.2. Of course, something has to be done to prevent all models from being exactly the same. Instead of allowing the model to select any of the given features to make a split, only a randomly selected subsection of the features will be considered in each split (Hastie et al., 2009). Since splits are forced to be different, the resulting trees will also differ. All these trees together, linked by a final vote to decide the overall outcome for every instance, is a random forests model. Instead of making a final majority vote, one could also consider the fraction of trees that classify an instance as "1", which gives some kind of scaled probability of this outcome.

3.4 Performance estimation

This section will handle questions like how the data should be used in order to build a model and how the performance of the resulting model can be estimated.

3.4.1 Performance measures for binary classification

Since the problem considered in this work is approached as a binary classification problem with ranking, we will discuss two performance measures that can be used to estimate the performance of this type of model.

Accuracy The accuracy represents the percentage of correctly classified instances. To see whether an accuracy can be considered as good or not, one can compare it to the accuracy of

a random classifier, which will randomly assign “rain” or “no rain” to every instance. This is just random guessing, giving the classier a 50% chance to be correct if it rains about half of the time, resulting in both an accuracy and AUC of 0.5. The more the accuracy of a model differs from the one by a random model, the more it has learned from the data. As an example, we will calculate the accuracy for a fictive outcome of the rain prediction problem, given in Table 3.1. The first row represents the true label for the five days considered in this example, while the second row shows the prediction scores resulting from a model. The third and fourth row show whether the model classified the instance correctly if a threshold of respectively 0.5 and 0.4 is applied to the predictions score. A correct classification is represented by “V” and an incorrect one by “X”. When the threshold is set to 0.5, the algorithm is correct in 3 out of 5 days, resulting in an accuracy of 0.6. If a cutoff of 0.4 is considered, the accuracy rises to 0.8. This shows the first problem related to using accuracy: it greatly depends on the threshold set, and since random forests give a scaled probability that the label will be “rain”, it is very plausible that the ideal threshold is not 0.5. The second problem with using accuracy as a performance measure is that it can give an incorrect view if the dataset is very unbalanced. If 90% of the days in our dataset happened to be rainy days, a model that always predicts rain will have an accuracy of 0.9. A model with an accuracy like this would be considered to perform well, even though this is clearly not the case since it only has one rule: “always predict rain”. Notwithstanding these problems, the accuracy of a model might be exactly what one wants to know.

ROC curves and AUC In order to avoid the need of setting a threshold, one can use a receiver operating characteristic (ROC) curve to analyse the performance of a model, where the information for all possible thresholds is represents (Fawcett, 2004). A ROC curve is the result of plotting the true positive rate (TPR) as a function of the false positive rate (FPR). The true positive rate is the percentage of positives for which the model made a correct prediction:

$$\begin{aligned} True\ positive\ rate &= \frac{true\ positives}{positives} && (3.4)\\ &= \frac{true\ positives}{true\ positives\ +\ false\ negatives} && (3.5) \end{aligned}$$

The false positive rate is the percentage of negatives for which the model made an incorrect prediction:

$$\begin{aligned} False\ positive\ rate &= \frac{false\ positives}{negatives} && (3.6)\\ &= \frac{false\ positives}{false\ positives\ +\ true\ negatives} && (3.7) \end{aligned}$$

A typical ROC curve is visualised in Figure 3.3. The more the ROC curve tends to the top left, the better the performance of the model, since the true positive rate will increase more rapidly than the false positive rate. A random classifier shows a ROC curve coinciding with the first bisector. The area under the curve (AUC) is, as the name suggests, the area under this curve, which is greater with better performances and is 0.5 for a random classifier.

To illustrate this with a real example, Figure 3.4 represents the ROC curve for the rain prediction problem. Since this example only contains five instances, this curve is not as smooth as the one

Table 3.1: Toy data used for accuracy calculation on the rain prediction problem. This example considers five days, or instances, for which the prediction scores, the model outputs, are compared to the true label. This coparison is done for a cutoff of 0.5 and 0.4 in row three and four respectively, where correct predictions are represented by "V" and incorrect prediction by "X". It illustrates that the accuracy depends on the threshold set on the prediction scores.

	Day 1	Day 2	Day 3	Day 4	Day 5
True label	rain	rain	no rain	rain	no rain
Model output	0.9	0.6	0.5	0.4	0.1
cutoff 0.5	V	V	X	X	V
cutoff 0.4	V	V	X	V	V

in the previous plot. As mentioned before, the ROC curve represents the performance for all thresholds, starting with a threshold equal to the maximum prediction score and in turn considering each following score as the next threshold. For a threshold of 0.9, the first day is classified as positive, giving a rise in the TPR. The same happens for the next threshold, being 0.6, but when the threshold is set to 0.5, the associated instance is a negative instance. If one considers all instances with prediction scores equal or above this threshold as interacting, this is a false positive, resulting in an increase of the FPR. To illustrate this, we calculate the TPR and FPR for this threshold, which can also be read from the ROC curve:

$$TPR = \frac{3}{3+1} = 0.75 \tag{3.8}$$

$$FPR = \frac{1}{1+1} = 0.5 \tag{3.9}$$

At a threshold of 0.4, we again encounter a positive instance, causing a rise in the TPR, followed by a rise in the FPR for a threshold of 0.1. The AUC for this example is 0.833, giving a measure for the overall performance of the model instead of just a reflection for a given threshold, as is the case with the accuracy.

3.4.2 Data usage

Since a model is fitted to make a good prediction for an instance on which it is trained, it is easier to make predictions for such an instance than it is for a previously unseen instance following the same distribution. As a result, one cannot make a good assessment of the performance of a model if all available data is used to train this model: no data is left to make a realistic assessment of its performance. To avoid this, the dataset is split in a training and a test set during model-building, and the model is trained only on the training set. Since the instances of the test set are previously unseen by the model, the performance on these instances gives a good idea of the actual performance of the model. For the same reason, a third set is needed if hyper-parameters of a model have to be tuned: the validation set. In this case, the model is trained on the training set and the performance for different hyper-parameter settings is calculated on the validation set. The parameters that result in the best performance on this set will be used in the final model, for which the performance is estimated on the test set. However, in order to do

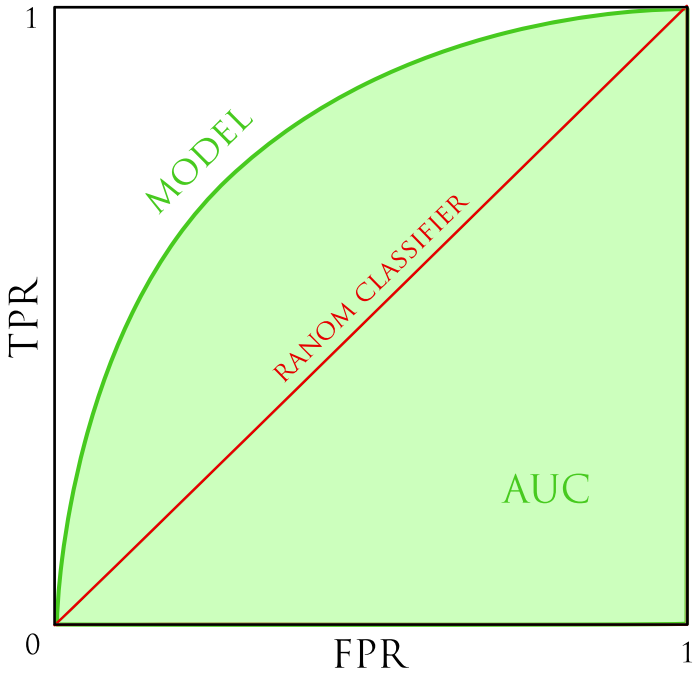


Figure 3.3: Representation of a typical ROC curve. The green curve shows a ROC curve of a realistic model, tending to the top left of the graph, and the light green area under the ROC curve represents the AUC. The red curve represents the ROC curve of a random classifier, which coincides with the first bisector since TPR and FPR increase proportionally.

this, the dataset has to be big enough, since the number of instances that can be used to train the model becomes smaller.

By splitting the dataset, one only makes predictions for a small part of the data. In order to make predictions for all instances in the data, which also gives the possibility to calculate a performance on the complete dataset, one can use cross-validation. Here, the data is split into n folds and each fold in turn is used as a test set, while the other $n-1$ folds form the training set. After n cross-validation runs, each fold has been used as training set $n-1$ times and once as test set, which results in predictions for all instances.

3.4.3 Under- and overfitting

The complexity of a model will influence its performance, where a complexer model will generally show a better fit to the training data. However, since real data is not perfect and noise can be present, a perfect fit to the training data usually does not result in the best model. This concept is illustrated in Figure 3.5, showing a classification problem and the decision boundaries of three models with different complexity. All instances above such a boundary are in this case predicted to be negative, represented by red circles, and the others to be positive, represented by green discs. When looking at the data plotted with respect to two parameters, represented in Figure 3.5, one might conclude that the best fit, following the general trend of the data, is probably the one in Figure 3.5d. Figure 3.5c shows the decision boundary of a model that is too simple: if all instances above the decision boundary are predicted to be negative and all other positive, multiple instances are misclassified and the general trend of the data is not present

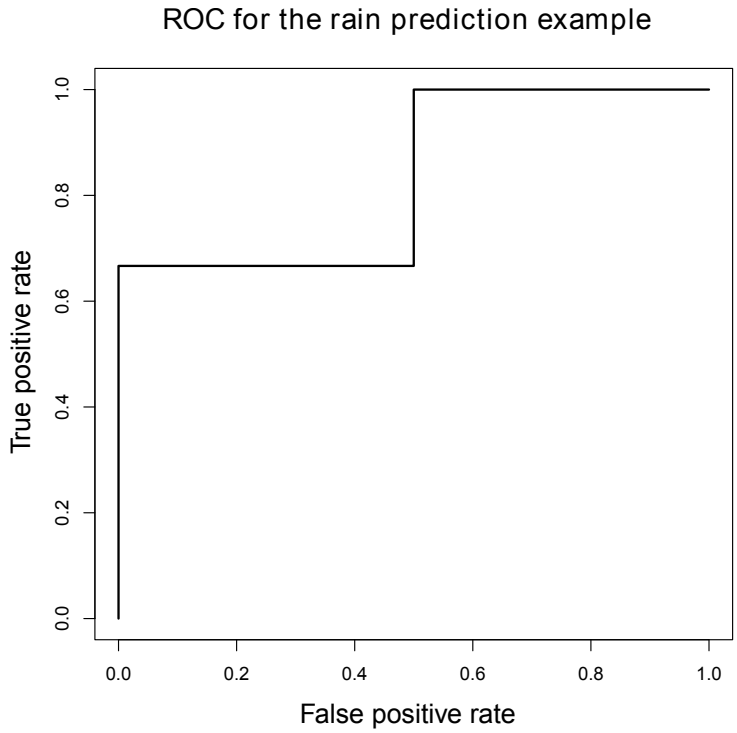


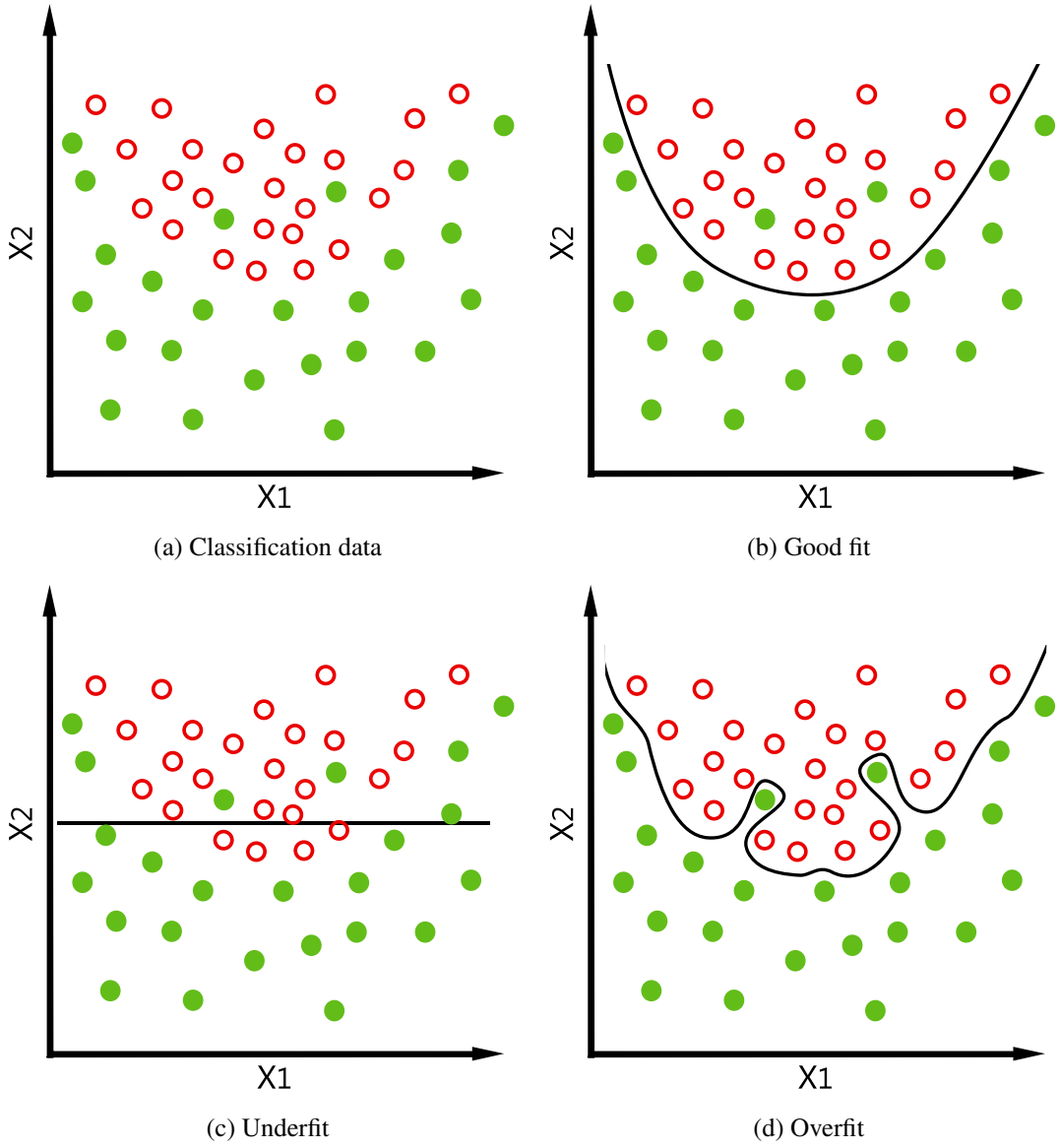
Figure 3.4: ROC curve for the rain prediction example, based on the data in Table 3.1. The AUC for this example is 0.833.

in the model. This model is underfit: a more complex model might make a better separation. Figure 3.5c shows a model that classifies every instance of the training data correctly. However it seems likely that the two positive instances that cause the decision boundary to make extreme turns are actually just noisy data. Since this model focusses too much on the training data it is unable to make a good generalisation: the model is overfit.

Underfitting can be amended by using a more complex method, like using random forests instead of logistic regression, or adding features to the model. Overfitting is avoided by simplification of a model, which can be done in different ways, depending on the method used to build the model. In case of logistic regression, one can use regularisation. One kind of regularisation is the Lasso or L1 regularisation, which will make a selection of features to use in the final model and drop the others, thus simplifying the model. For tree-based methods, like random-forests, the trees can be "pruned" by removing some of the terminal splits, which also simplifies the model. The concept of random forests itself also reduces the chance of overfitting, since it uses a vote of multiple trees, resulting in a more reliable prediction.

3.4.4 Settings for learning relations

Special considerations have to be made when one wants to predict relations between two objects. This is the case for our research question, since we want to predict the interaction between a mRNA and a miRNA. We consider every mRNA-miRNA combination to be an instance for which we want to predict whether it will interact or not, resulting in a binary classification problem. When a model has to make predictions to be used in real research, we would like to have an idea of the performance. As discussed previously, this can be assessed during model-



(a) Classification data

(b) Good fit

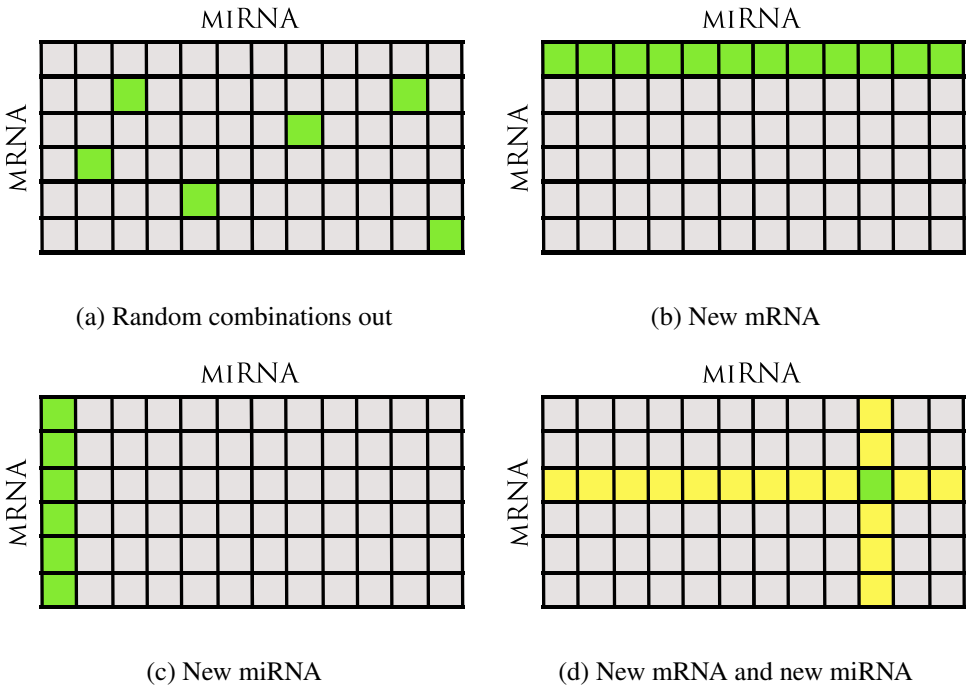
(c) Underfit

(d) Overfit

Figure 3.5: Toy data to illustrate under- and overfitting for a classification problem.

building by checking the performance on a test set. However, this performance will depend on whether the mRNA and miRNA are new to the model or whether some information on these molecules was already present in the training set. In this light, we can consider four different cases, which are discussed in the following paragraphs. These cases will form now on be referred to as "the settings". An overview of the data usage for each setting can be seen in Figure 3.6, where an example for 6 mRNAs and 12 miRNAs is given.

Random combinations out In this case, predictions are made for mRNA-miRNA combinations that where not present in the training set. However, information on how both molecules interact with other mRNAs and miRNAs is provided. This means that the exact combination which has to be predicted is new to the model, but the molecules themselves are not. In order to assess the performance for this setting using cross-validation, every fold will contain a ran-



(a) Random combinations out

(b) New mRNA

(c) New miRNA

(d) New mRNA and new miRNA

Figure 3.6: Overview of the data usage in the different settings, represented for one run in the cross-validation. All training data for this run is represented in grey, wile the test data is represented in green. The yellow mRNA-miRNA combinations in the setting where both mRNA and miRNA are new cannot be considers in order to train or test the model that will predict the interaction between the mRNA and miRNA in question.

dom selection of mRNA-miRNA combinations, as illustrated in Figure 3.6a. Since no specific structure is present in the selection, the performance can be estimated on the pooled predictions. This means we will no longer pay attention to the mRNA and miRNA a combinations belongs to and just compare the true label and the prediction score for each instance.

New mRNA When predictions are made for a mRNA-miRNA combination of which the mRNA is previously unseen by the model, this is considered to be the "New mRNA" setting. To test this on our data, every fold of the cross-validation will contain the combinations involving one mRNA, visualised in Figure 3.6b. As a result, every run in the cross-validation will build a model on the data of all but one mRNAs, and make predictions for the instances involving this unseen mRNA. In this setting, structure is present in the instances assigned to different folds, complicating the performance estimation. Since every model is built per mRNA, the resulting models are also optimised to make good predictions per mRNA, and the performance measured per mRNA will hopefully be quite satisfying. However, this does not guarantee that these models will not discriminate a certain miRNA, which can only be seen if the performance is measured per miRNA. This is illustrated by the data in Table 3.2. The AUC of the pooled data is 0.667 and the AUCs per mRNA are all 0.75. However, the AUCs per miRNA differ greatly form one another, reaching form 1 all the way to 0. An AUC of 0 means that all instances are systematically misclassified, where a model will predict interaction for all non interaction combinations and vice versa. This illustrates that it is important to check both the macro AUC, which is the AUC calculated on the pooled data, and the micro AUCs, which are calculated per mRNA or miRNA. In the case of the "New mRNA" setting it is especially advised to check

the AUC over the miRNAs. For good models, given enough data is available, micro and macro AUCs will hardly differ from one another.

Table 3.2: Toy data used to illustrate the need of computiong both macro and micro AUCs. The number proceding the slash symbol represents the true label, while the numer following the slash symbol represents the prediction score made by a fictive model.

	miRNA 1	miRNA 2	miRNA 3	miRNA 4	miRNA 5	AUC per mRNA
mRNA 1	1 / 1	1 / 0.5	1 / 0.75	1 / 0	0 / 0.25	0.75
mRNA 2	1 / 0.5	0 / 0.25	1 / 1	1 / 0	1 / 0.75	0.75
mRNA 3	0 / 0.5	1 / 0.75	0 / 1	0 / 0.25	0 / 0	0.75
AUC per miRNA	1	1	0.25	0	1	pooled AUC: 0.667

New miRNA The "New miRNA" setting is equivalent as the previous setting, but here a new miRNA is considered instead of a new mRNA. The folds of a cross-validation must now be based on the miRNAs, as can be seen in Figure 3.6c, and the performance should be estimated per mRNA.

New mRNA and new miRNA In this setting, both the mRNA and miRNA are previously unseen by the model. Since absolutely no information on the molecules of interest is present, this is the hardest of the four settings. To make this happen in a cross-validation, all combinations involving the mRNA and miRNA in question have to be removed form the training set when predicting this mRNA-miRNA combination. This setting is visualised in Figure 3.6d, where the mRNA-miRNA combinations that have to be removed in order to predict the green instance are indicated in yellow. In this case, it is advised to check both the performance per mRNA and per miRNA.

3.4.5 Testing with random data

A simple test to check whether a good performance is due to actual learning or a possible mistake, is to replace the data by random information. One could for example replace a binary label by a random sampling between 0 and 1. In this case there is no link between the features and the random label, so a binary classifier should display an accuracy and AUC of 0.5 on a test set. The same should be true when using randomly sampled features or when both labels and features are random. If this is not the case, something is wrong. There might be a "leak" between training and test data, making it possible for a totally overfitted model to predict previously seen instances and thus do better than random.

Chapter 4

The mRNA-miRNA dataset

In this chapter we will explore the characteristics of the dataset used in this work. Section 4.1 explains how the research questing can be translated into a machine learning questing, explaining the features which will be used. Section 4.2 discusses how the dataset was generated, involving both wet-lab experiments an bioinformatics analyses. In the main section of this chapter, Section 4.3, multiple statistical exploration techniques are applied to the data in order to form a general idea of its characteristics.

4.1 A multi-level problem

The goal of this research is to predict interaction between mRNAs and miRNAs in humans. However, to realise such predictions, one has to obtain a notion of what might cause this interaction. As mentioned in Chapter 2, the 3' region of mRNAs may contain seed matches or sites for a miRNA of interest, meaning there is a small patch of sequence complementarity. Different types of sites are described, but the ones most frequently used are the 6mer, 7mer-A1, 7mer-m8 and 8mer sites, all positioned at the 5' end (the beginning) of the miRNA. Figure 4.1 shows the difference between these sites. The first number in the site name represents the number of nucleotides in this site, for example: a 6mer consists of six subsequent nucleotide matches between mRNA and miRNA. In the case of multiple sites of the same length, as with the 7mer sites, the last part of the name gives additional information about the site. A 7mer-A1 site has an adenine (A) nucleotide that does not have to be a match at position one of the site, whereas a 7mer-m8 site has a matching nucleotide at position 8 of the site.

Note that the position of a site is defined for the miRNA, but the matching part on the mRNA can occur anywhere in its 3' regulatory region. Because of this, multiple site matches between one mRNA and one miRNA can occur and can belong to different site types. To show the complexity of this situation, a small example is given, which is also visualised in Figure 4.2 for clarity. A given mRNA may have two 6mer sites and one 8mer site for a specific miRNA-1, three 6mer sites for miRNA-2, no sites for miRNA-3 to 7, one 7mer-A1 site and one 7mer-m8 site for miRNA-8 and so on. To see how potent these sites are to induce interaction, information on each site, like the conservation of the nucleotides over different animal species, can be used. This results in information on two levels: the level of a mRNA-miRNA combination and the site level. The first holds information on the interaction for a mRNA-miRNA combination,

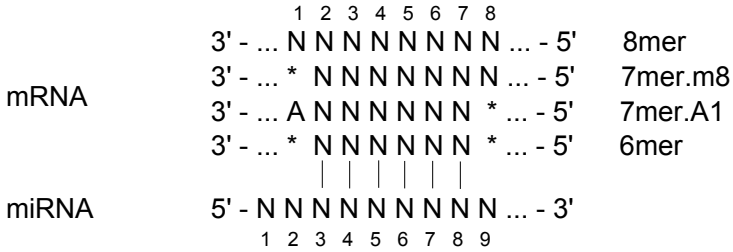


Figure 4.1: Overview of seed site types used in this work. All sites start at the second nucleotide of the miRNA, being position one of the seed site. A matching nucleotide is represented by the letter N, an adenine nucleotide by the letter A and any non-matching nucleotide by a star. Three dots indicate that the RNA strand continues in that direction. A 6mer site has six subsequently matching nucleotides, a 7mer-A1 site has the same matching nucleotides and additionally holds an adenine nucleotide at position 1 of the site. A 7mer-m8 site has seven matching nucleotides and a 8mer site has eight.

mRNA	miRNA	6mer	7mer-A1	7mer-m8	8mer
1	1	2	0	0	1
1	2	3	0	0	0
1	3	0	0	0	0
1	4	0	0	0	0
1	5	0	0	0	0
1	6	0	0	0	0
1	7	0	0	0	0
1	8	0	1	1	0

mRNA	miRNA	site type	...	features	...
1	1	6mer	...	...	...
1	1	6mer	...	...	...
1	1	8mer	...	...	...
1	2	6mer	...	...	...
1	2	6mer	...	...	...
1	2	6mer	...	...	...
1	8	7mer-A1	...	...	...
1	8	7mer-m8	...	...	...

Figure 4.2: Example of the mRNA-miRNA data tables and their relation. The table on the left represents the information of mRNA-miRNA combination level, the table to the right the information on the site level.

which is what we want to predict, and the number of sites of each site type occurring for this combination. The second level has information on each occurring site, such as where on the mRNA a site occurs and how conserved it is. This info cannot simply be added or mapped to the mRNA-miRNA combination level, since multiple sites per combination may occur, and it is not clear which of these induce a potential interaction. We will examine some solutions to this problem in Chapter 5, in order to be able to use as much of the available information as possible.

4.2 Generating the mRNA-miRNA dataset

All data is generated by Gert Van Peer as a part of his research at UZ Gent. Both wet-lab tests and bioinformatics analyses were performed.

Wet-lab tests HEK 293T cells, which originate form transformed human embryonic kidney cell cultures, were seeded and grown in 96-well plates. After 24 hours, these cells were co-transfected with three nucleotide fragments. The first is a 3'UTR reporter construct, which is a combination of the 3'UTR of the gene for which interaction has to be tested (the mRNA), followed by the Firefly luciferase gene that functions as reporter gene. The second fragment is a control reporter construct, holding only a Renilla luciferase gene. This results in a Firefly reporter gene, which is regulated as the mRNA of interest would be, and an unregulated Renilla reporter gene. The third fragment introduced in the cells is the miRNA of interest, which is introduced as a miRNA mimic. A miRNA mimic is a dsRNA molecule with the mature

miRNA of interest as one of its strands, which can immediately be integrated in the RISC complex to induce silencing. This type of co-transfection was performed 7990 times, each in a different well and with a different combination of the 17 mRNA 3'UTR sequences and the 470 miRNA mimics of interest. After transfection, the cells were incubated for 48 hours, followed by the measurement of the reporter gene activities. This was done by adding the required substrates, i.e. luciferine and ATP, to the wells and measuring the produced light signals with a luminescence plate reader. The signal resulting from the Firefly luciferase is proportional to the level of the mRNA of interest, whereas the signal resulting from the Renilla luciferase represents the level of expression of this mRNA in case no regulation can occur. The difference between these two signals is a measure of the interaction between the mRNA and miRNA for that particular co-transfection. If interaction has occurred, the light signal resulting from the Firefly luciferase is weaker than the one resulting from the Renilla luciferase.

Bioinformatics analyses To generate the information on the site level, bioinformatics analyses were performed. The presence of miRNA seed site matches in the mRNA 3'UTRs was detected by alignments, resulting in information on the presence and position of the sites. For each site, the percentage of adenine and uracil nucleotides in the region flanking the site was computed, since this is believed to influence the probability of interaction. In addition, the level of conservation of each nucleotide in these 3'UTR site matches was determined by alignments of the corresponding 3'UTR regions of different animal species. This was done with two different methods, i.e. phastCons (Siepel et al., 2005) and phyloP (Pollard et al., 2010), and computed with respect to different animal groups. These conservations are believed to be a measure for the importance of a site.

A cutoff was defined in order to be able to interpret the interaction score in a binary way. This results in interacting and non-interacting mRNA-miRNA combinations instead of a measure for the probability of interaction. A set of validated interactions was extracted form literature by text mining. Since validated non-interactions are hard to find, these were generated by screening a library of reporter genes without 3'UTR regulatory sequences, which are expected to be unregulated by miRNAs. The optimal cutoff was found to be -1.77: with 95% specificity and 50% sensitivity, all mRNA-miRNA interactions scores equal or below this cutoff are interacting and all others non-interacting.

4.3 Exploring the mRNA-miRNA dataset

The mRNA-miRNA dataset used for this work consists of 17 mRNAs and 470 miRNAs, resulting in 7990 observed interactions. If the cutoff of -1.77, discussed in Section 4.2, is applied on the interaction score, 5.7% of the mRNA-miRNA combinations interact. This results in a very unbalanced dataset, where the majority of mRNA-miRNA combinations does not interact.

As can be seen in Figure 4.3, 30% of the mRNA-miRNA combinations have sites, of which 14% interact. In these combinations, a total of 3726 sites occur: 51% 6mer, 21% 7mer-A1, 20% 7mer-m8 and 8% 8mer sites. Of the mRNA-miRNA combinations that do not contain sites (70% of the total number of combinations), only 2% interact. All mRNAs have multiple interacting miRNAs, whereas 194 miRNAs (37%) have no interacting mRNAs. All mRNAs have multiple sites for the included miRNAs, but 14 miRNAs (3%) have no sites for the in-

cluded mRNAs.

Label distribution The lower the interaction score the higher the probability of interaction. Interpretation is easier if this score is converted to a binary label: interaction or non-interaction. This is possible by using the cutoff for interaction defined in Section 4.2: all mRNA-miRNA combinations with interaction scores equal to or below -1.77 are said to interact, all others not. The lowest interaction score in the dataset is -7.19, the highest 6.26, with an average around 0. The distribution resembles a normal distribution but with a slight negative skew and skinny tails, as can be seen in Figure 4.4.

Influence of sites The presence of sites influences the interaction score of a mRNA-miRNA combination, resulting in different distributions for the part of the data where a site type is present and where it is absent. The resulting shift in interaction score is illustrated in Figure 4.5, where the influence of different site types can be seen. For each site type the shift in distribution is significant, with 8mer sites having the biggest influence, followed by 7mer-m8, 7mer-A1 and 6mer sites. This justifies the use of site presence as a feature to predict mRNA-miRNA interactions.

PCA A Principal Component Analysis (PCA) is performed on the data to see if the data can be represented in fewer dimensions. The data can be represented as a 470 x 17 matrix of the interaction scores, with miRNA as rows and mRNA as columns. The goal is to reduce to 17 columns and still withhold most of the information. The output of the PCA can be seen in Table 4.1 and shows that we need 10 Principal Components (PC) to be able to explain 80 % of the variance (or information) present in the data, and that the first two PC only explain 26%. It can thus be concluded that PCA is not very effective in this case, since dimensions cannot be dramatically reduced, and mRNAs are probably not highly correlated. To check this, a heat map of the correlation between mRNAs based on their interaction score is shown in Figure 4.6. Yellow and red shades indicate negative correlations, while green shades represent positive correlations. The more intense the colour, the more correlated the mRNAs are. However, the figure shows that most mRNAs are only slightly correlated, explaining why PCA might have a hard time reducing the dimensions.

Table 4.1: Output of the Principal Component Analysis: importance of the 17 components. Notice that only 26% of the variance, and hence of the information present in the data, is captured in the first two Principal Components. The ten first Principal Components are needed to explain 80% of the variance.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	1.6216	1.4701	1.3572	1.28859	1.17598	1.10665	1.06765	1.01264	0.95572
Proportion of Variance	0.1433	0.1178	0.1004	0.09048	0.07535	0.06673	0.06211	0.05587	0.04977
Cumulative Proportion	0.1433	0.2610	0.3614	0.45188	0.52723	0.59396	0.65607	0.71195	0.76172

	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Standard deviation	0.92545	0.8395	0.78564	0.74068	0.71972	0.66336	0.60187	0.57078
Proportion of Variance	0.04667	0.0384	0.03363	0.02989	0.02822	0.02398	0.01974	0.01775
Cumulative Proportion	0.80839	0.8468	0.88042	0.91031	0.93853	0.96251	0.98225	1.00000

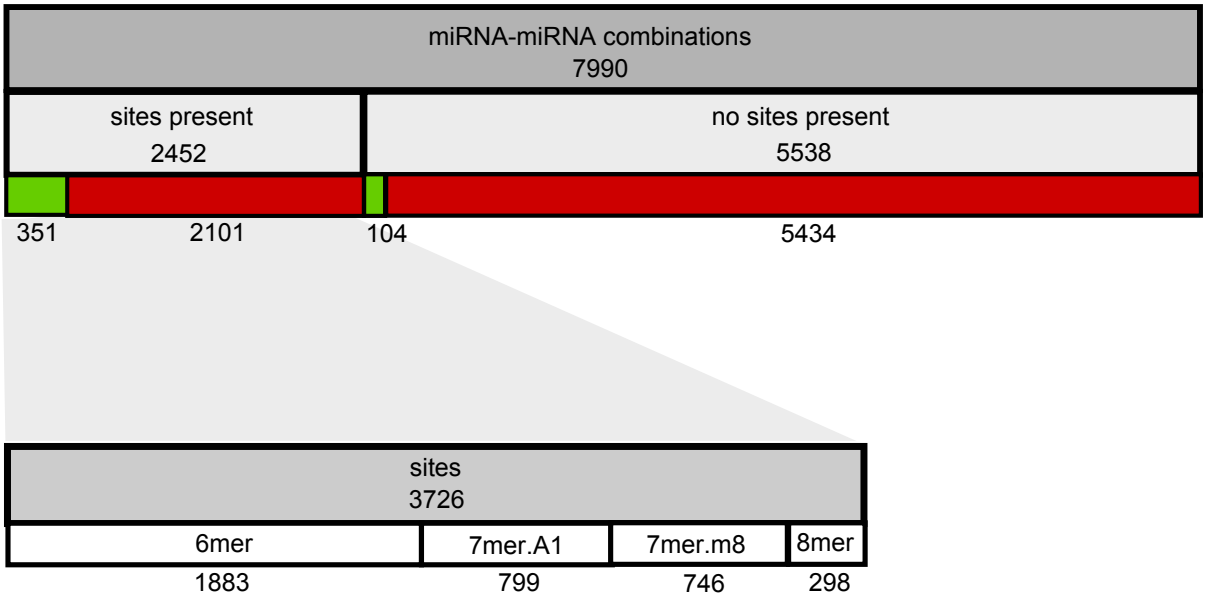


Figure 4.3: Overview of some aspects of the mRNA-miRNA dataset, including interaction and presence of sites. In total, 7990 mRNA-miRNA combinations are present in the dataset, 2452 of which (30%) have sites and 5538 (70%) do not. Among the combinations that have sites, 351 (14%) interact and a total of 3726 sites occur: 1883 (51%) 6mer, 799 (21%) 7mer-A1, 746 (20%) 7mer-m8 and 298 (8%) 8mer sites. Of the mRNA-miRNA combinations that do not contain sites, only 104 (2%) interact.

Cluster analysis To see which mRNAs are related with respect to their interaction score for miRNAs, the 17 mRNAs were clustered using hierarchical clustering based on the Ward criterion. Figure 4.7a illustrates this clustering. In order not to overcomplicate the interpretation, we will only look at the initial split in two clusters. The leftmost cluster holds mRNAs that generally participate in more interactions than those in the cluster on the right, as can be seen by comparing Tables 4.2a and 4.2b. This also holds for the number of sites these mRNA have with the included miRNAs, again justifying the use of sites as features to predict interaction. Figure 4.7b shows the two initial clusters of mRNAs, plotted with respect to their number of interacting miRNAs and their number of sites. If one knows that most mRNA-miRNA combinations only have a few sites, one can notice that the presence of sites does not guarantee interaction, as can be seen by the big difference between the number of interacting miRNAs and the number of sites.

Clustering of the mRNAs was also performed based on the total number of sites and the number of individual site types, showing comparable results. Two specific mRNAs, *MYT1L* and *PHF6*, always cluster together based on their number of sites. Table 4.2a shows that these are the mRNAs with the highest number of sites, 534 and 483 respectively.

Clustering can also be done to group the miRNAs, but due to the large number of resulting clusters, a dendrogram plot of the clustering is harder to interpret. To see if clustering is relevant, the within group sum of squares after clustering was plotted for k-means clustering ranging from 2 to 200 clusters. Since no "elbow" could be seen, no ideal number of clusters can be determined. Due to the lack of information in the first two principal components (26% of variance), plotting possible clusters with respect to these principal components has no added value.

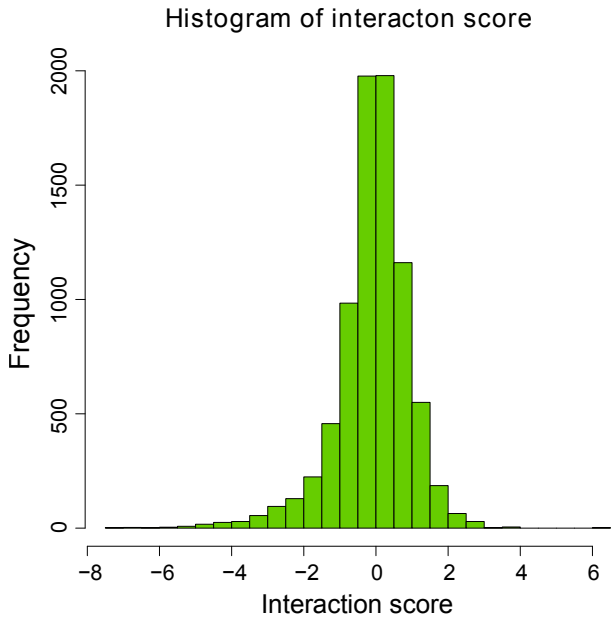


Figure 4.4: The distribution of the interaction score in the mRNA-miRNA dataset resembles a normal distribution but with a slight negative skew and skinny tails

Heat maps Heat maps are very useful for visualising information. Figure 4.8 consists of two heat maps of the mRNA-miRNA dataset. The first shows the interaction score for every mRNA-miRNA combination, while the second shows the number of sites for these combinations. When looking at the first heat map, Figure 4.8a, it can be seen that most mRNA-miRNA combinations do not interact, resulting in green shades on the plot. Interaction scores below the cutoff for interaction, thus meaning the mRNA and miRNA in question do interact, are indicated by shades of yellow, orange and red. Scores close to the cutoff, which do not clearly interact or not, are represented as white. In the second heat map, Figure 4.8b, shades of green show the number of sites for the mRNA-miRNA combinations, where darker tones indicate more sites and white indicates the absence of sites. On can see that multiple mRNA-miRNA combinations have no sites, and for the others the number of sites is usually low. When compar-

Table 4.2: Overview of number of interacting miRNAs and number of sites for the two initial clusters of mRNAs based on their interaction scores.

(a) Table of mRNAs in the green cluster.

mRNA	*PHF6*	*NOTCH1*	*RB1*	*MYB*	*MYT1L*	*ZEB2*	*FBXW7*	*PHOX2B*
nr of interactions	30	22	31	41	39	55	48	23
nr of sites	483	272	273	264	534	246	288	241

(b) Table of mRNAs in the red cluster.

mRNA	*MYCN*	*EZH2*	*MYC*	*BRCA1*	*PALB2*	*ALK*	*HRAS*	*BRCA2*	*TP53*
nr of interactions	31	21	25	35	10	13	6	18	7
nr of sites	203	76	84	217	41	68	46	121	269

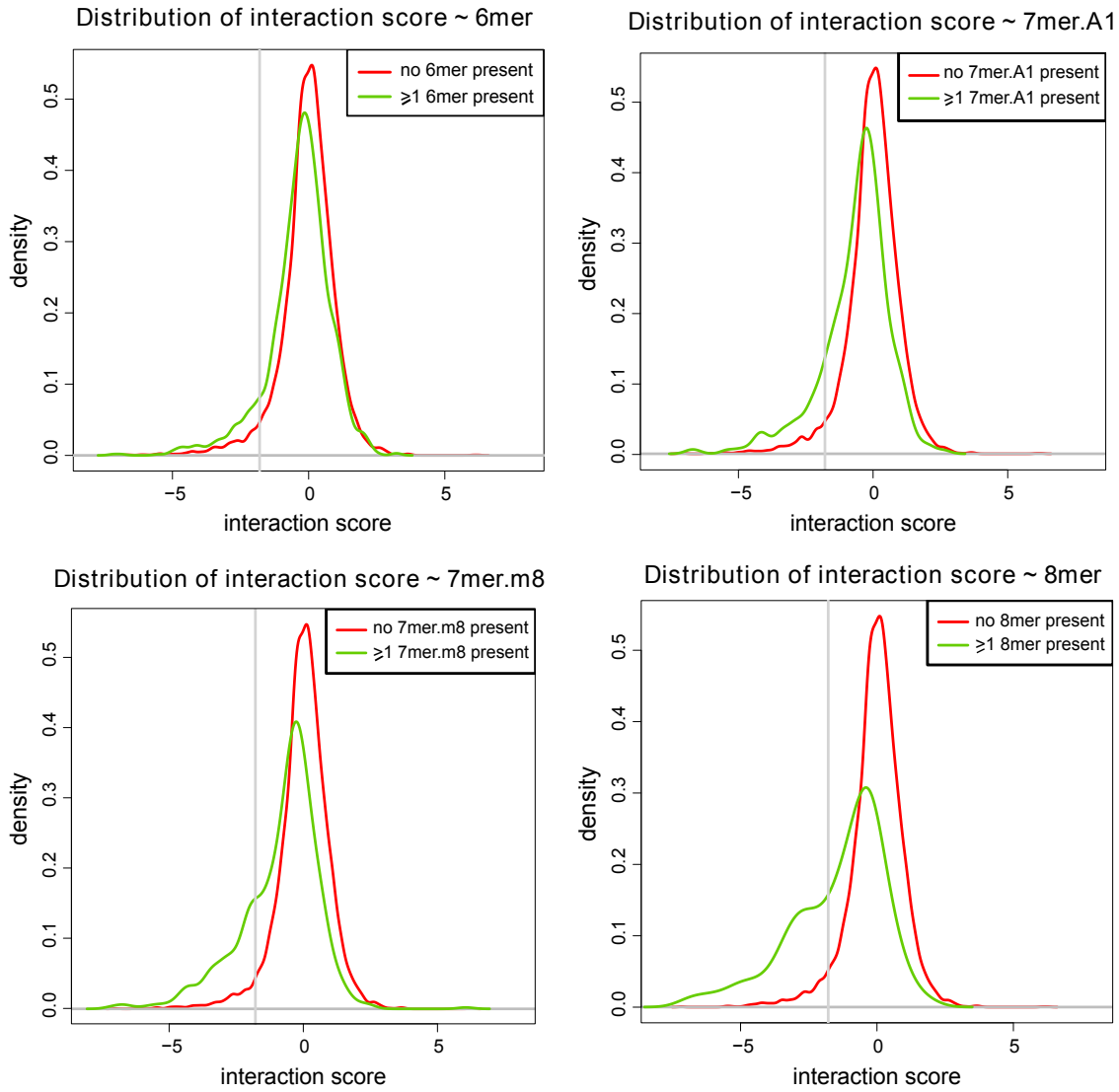


Figure 4.5: Distribution of the interaction score in function of the presence of 6mer, 7mer-A1, 7mer-m8 or 8mer sites. The red curve represents the distribution of the interaction score when the specified site is not present, the green curve if one or more sites of the specified site type are present. The grey vertical line represents the cutoff for interaction at -1.77, mRNA-miRNA combinations with interaction scores left of this line are assumed to interact.

ing the two heat maps, some links can be seen. The mRNA *MYCN*, for example, will interact with the first miRNAs, resulting in the yellow band at the top of the first heat map. In the second heat map, the same position is green, showing that this mRNA also has sites for these miRNAs and indicating that the presence or number of sites can be linked to interaction. This is not a rule however, which can be seen by looking at the mRNA to the left, *MYC*. In this case there is interaction, shown by the yellow band in the first heat map, but no sites are present. One can thus conclude that other factors may influence interaction. The opposite case occurs as well, where a high number of sites does not result in interaction. In this case more information on these sites is necessary to determine whether the presence of sites will lead to interaction.

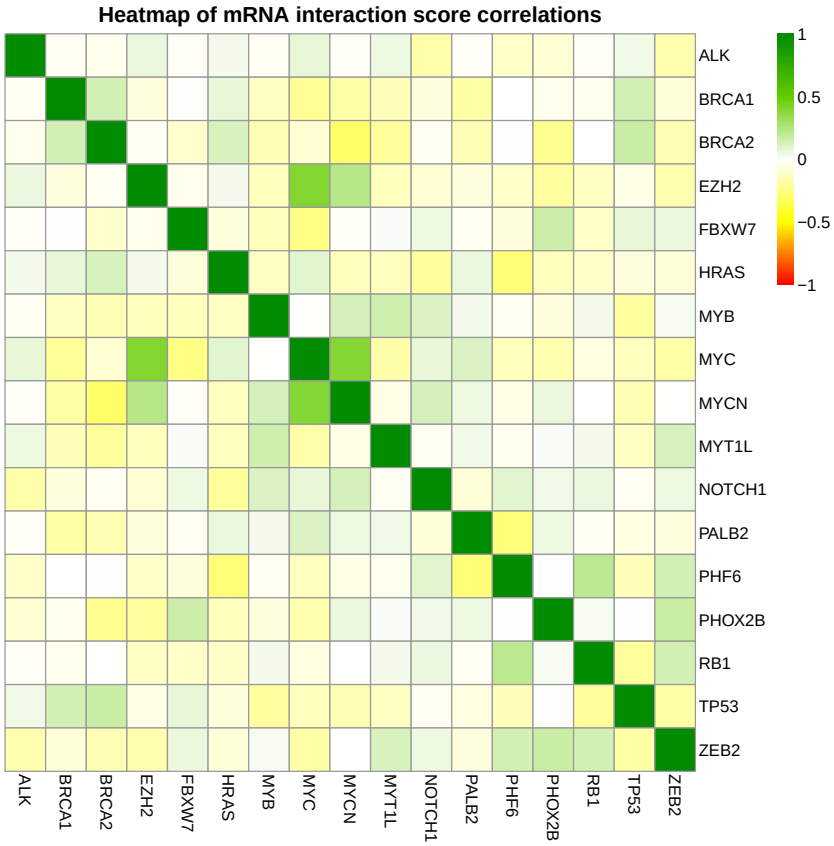
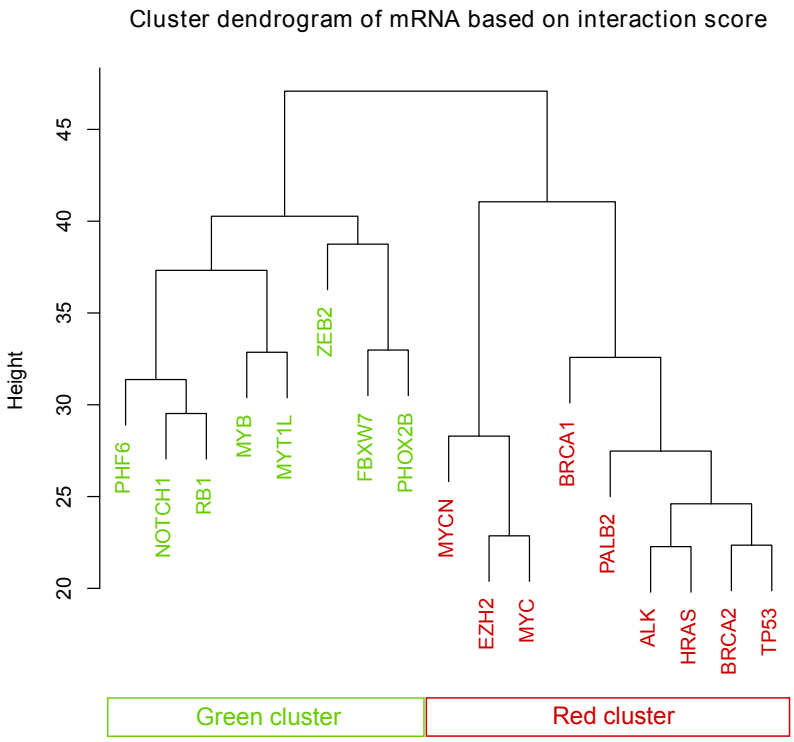
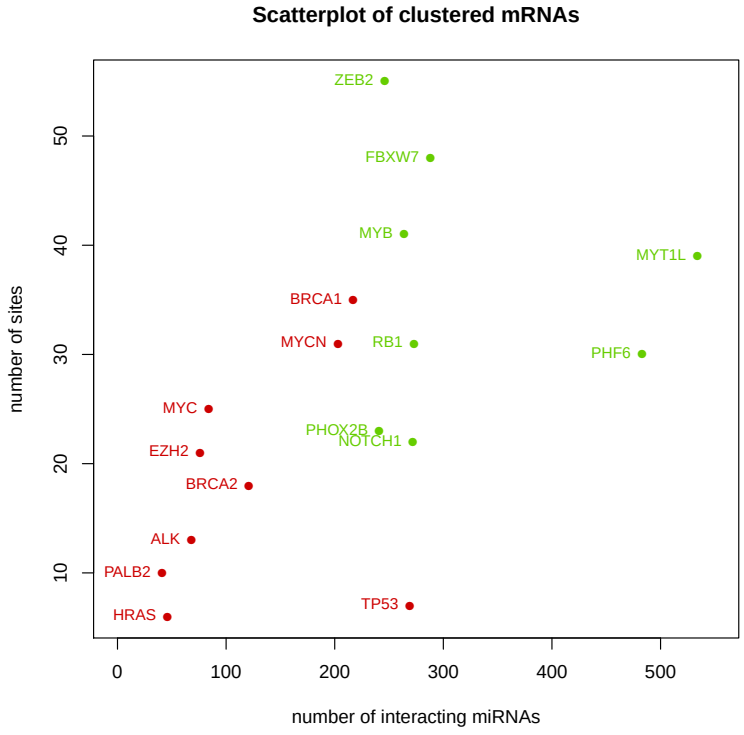


Figure 4.6: Heat map of correlation between mRNAs based on their interaction score. Green shades represent positive correlation, red shades negative correlation. The intensity of the colour represents the intensity of correlation, resulting in very soft shades for hardly correlated mRNAs.

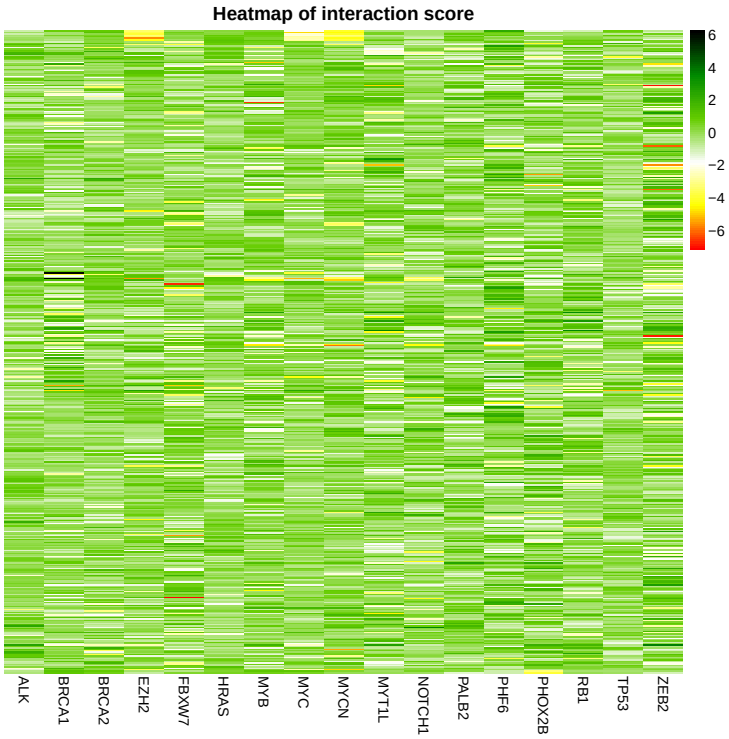


(a) Ward's hierarchical clustering of mRNAs based on interaction score.

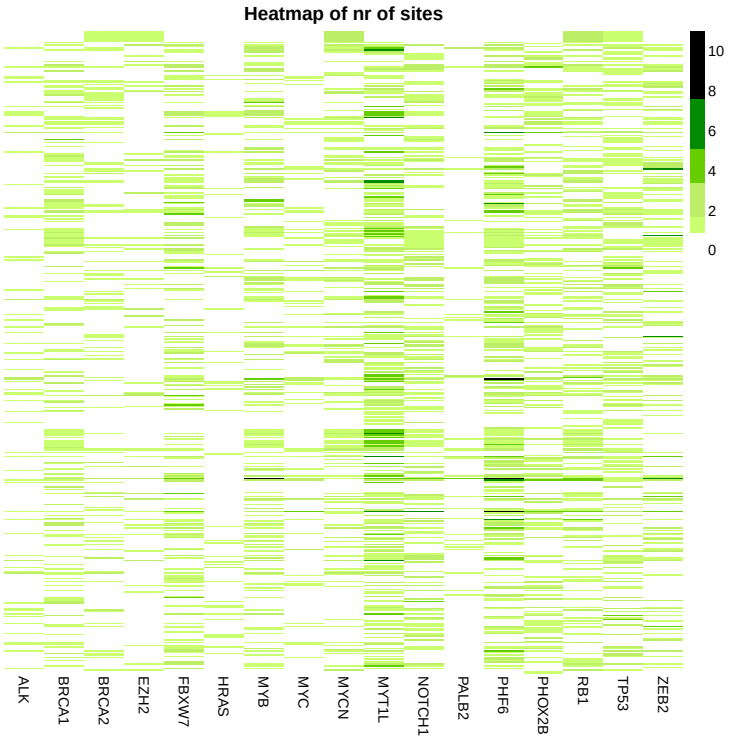


(b) Scatterplot of the clustered mRNAs

Figure 4.7: Clustering of mRNAs. As can be seen in the scatterplot (b), mRNAs present in the red and green cluster in (a) clearly show differences in number of sites and interacting miRNAs. The mRNAs in the green cluster generally show more interacting miRNAs and more sites than the mRNAs in the red cluster.



(a) Heat map of interaction score. Most mRNA-miRNA combinations do not interact and are represented by shades of green. Interacting combinations apprear yellow, orange or red.



(b) Heat map of number of sites. Since the intensity of green represents the number of sites, it can be seen that most mRNA-miRNA combinations have a low number of sites.

Figure 4.8: Heat maps to compare interaction score and number of sites for the same mRNA-miRNA combinations. The presence of sites might coincide with the presence of interaction, but this is not a rule.

Chapter 5

The models and their structure

Now that we have an idea of how the data is distributed, we turn to our real purpose: predicting interaction. Given the structure of the data, with information on two levels as discussed in Chapter 4, this is not straightforward. We will examine different models and gradually increase their complexity and the amount of information used. Regardless of the model type, we will split the data in a training set to build the model, and a test set for which we make predictions to check the performance of the model. In order to make predictions for all mRNA-miRNA combinations we use cross-validation, as explained in Chapter 3.

5.1 Single-model approach

Most of the time, machine learning techniques are applied as stand alone models: they receive the data, build a model and make predictions. However, depending on the complexity of the data and the actual goal of the analysis, practical complications may arise. In our case, predictions have to be made at the level of mRNA-miRNA combinations, whereas most features are site dependent, and thus present at site-level.

5.1.1 Site Count model

The most simple model type considered in this work is the "Site Count" model. The only features considered here are the number of sites, resulting in five features: one for each of the four site types described in Chapter 4 and one for the total number of sites. Note that no additional information on the sites is incorporated, resulting in a model solely based on the level of mRNA-miRNA combinations. Figure 5.1 shows the structure of this model, the usage of data and the flow of information. As discussed in Chapter 3, the data is split in a training and test set. Using the training set, the model is trained, after which this part of the data is no longer used. Subsequently, the test data is given to this model and the interactions of the mRNA-miRNA combinations present in the training set are predicted.

Strong points

- Simplicity in data: no effort is made to incorporate site-level information. Only the number of sites has to be determined by alignment, no additional features have to be

computed.

- Simplicity in structure: only one model has to be trained.
- Fair data usage: all site types are treated equally in terms of amount of information.

Weak points

- Information loss: no information of the site-level is used, ignoring a lot of potentially decisive data.

5.1.2 Extended Site Count model

In order to include more information on site-level, we take a closer look at the data. Four site types are present, each described by six conservation methods, the relative position on the 3'UTR of the mRNA and the percentage of adenine and uracil in the flanking regions. The easiest way to add this information to the data on the mRNA-miRNA combination level is by simply adding the columns with the site-level features to the table with mRNA-miRNA combinations. However, multiple problems arise, making this simple addition impossible. To start with, different sites have different lengths: they are made up of six, seven or eight neighbouring nucleotides. Since the conservations are computed for each nucleotide, different sites also have a different number of features. If one would like to add one type of conservation of an 8mer site to the mRNA-miRNA combination table, one would have to add eight features, each representing the conservation for one nucleotide position in the site. However, if one wants to add the information of a 6mer, which only had six nucleotides, the last two features are undefined. A straightforward solution is to impute these features with the mean of the column, but this is not relevant for this problem, since these nucleotides are simply not part of the 6mer site and their conservation cannot be considered in the same way. So, to avoid this problem one could just add the features of each site separately, resulting in the addition of a lot of features. In this case, 38 features of the 6mer site, 44 features of the 7mer-A1 and 7mer-m8 and 50 features of the 8mer site. In total, 176 features would be added. However, this only makes the problem worse, since we now have even more columns that are irrelevant in case a certain site type is not present. For example: if only a 6mer site is present, only these features are available and relevant, while the other 138 features on site-level are not relevant since none of these sites are present. The complete dataset only holds a few hundred sites of each type, as a consequence most of the 7990 mRNA-miRNA combinations will have multiple irrelevant features. As a matter of fact, only 11 mRNA-miRNA combinations will have no irrelevant features. Since no imputation can be done and most machine learning techniques do not consider records with unavailable entries, only these 11 records can be used to train the model. To make thing worse, the resulting model will only be able to make predictions for mRNA-miRNA combinations with at least one site of each type, which hardly ever occurs. Clearly, this is not acceptable, but there is yet another problem. On mRNA-miRNA combination level, there is only one record for each combination. Even if one could easily incorporate different kinds of sites, what has to be done with multiple sites of the same type? One could try to find the most important site, but there is no way to know which of the sites are responsible for a possible interaction. One could also take the average for each features and include this, but for some features, like the relative position on the mRNA, averaging does not make sense.

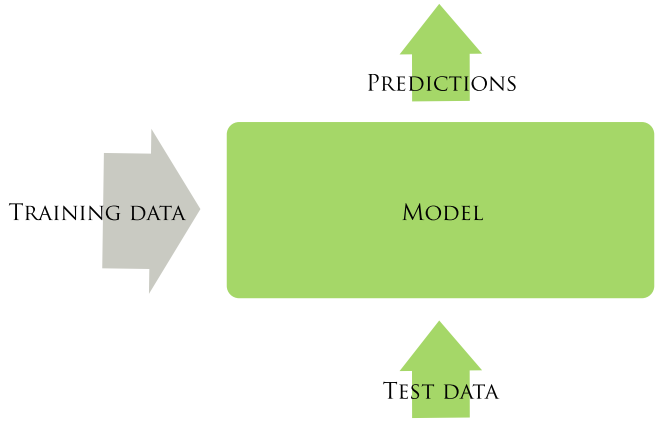


Figure 5.1: Structure and information flow of the "Site Count" model. The model is trained on the training set, after which the latter is no longer used. In a second stage, the test set is provided to the model, which predicts the interaction for each mRNA-miRNA combination present in this test set.

Other researchers have chosen to include only the 7mer-m8 site in their analysis, avoiding problems with different site types, and this only if exactly one site is present, avoiding the issue of multiple sites. This approach is used in MirTarget2 (Wang and El Naqa, 2008), one of the leading methods in predicting mRNA-miRNA interaction, which will be discussed in Section 5.3.1. Site type 7mer-m8 is chosen because of the balance between influence of the site, as can be seen in Figure 4.5, and number of occurrences of it. The site type with maximal influence is the 8mer, but since this is also the least occurring site, this limits the number of mRNA-miRNA combinations considered. For the "Extended Site Count" model, we have chosen to use a similar approach, and only add information to the table holding the information on mRNA-miRNA level if a single 7mer-m8 site is present. The structure of the model is exactly the same as in case of the "Site Count" model, represented in Figure 5.1. The resulting dataset has 7400 records with only the number of sites as features, and 590 with all available information present. If no 7mer-m8 site is present, the site features are set to zero, even though this is not entirely correct. Without this, the algorithm would only be able to use or predict mRNA-miRNA combinations with exactly one 7mer-m8 site, and all others would be considered not to interact.

Strong points

- Simplicity in structure: only one model has to be trained.
- Data usage: it is possible to include some information on the site level, showing the importance of a site if present.

Weak points

- Unfair data usage: Only info on the 7mer-m8 sites is included, resulting in an overem-

phasis on the importance of this site type.

- Information loss: no data on other site types or multiple 7mer-m8 sites is used.

5.2 Stacked model approach

The goal of the "Stacked" model is to use all information provided, in a way which is biologically and technically sound. To do this, the idea of using a single prediction model is abandoned. Since the data naturally has two levels of information, the mRNA-miRNA combination level and the site-level, these will also be present in the stacked model approach. We start on the site-level and split the dataset by site type, resulting in four datasets with one record per site. Since there are no issues with different numbers of features nor with presence or absence of other sites, a straightforward prediction of the interaction of the mRNA-miRNA combination linked to a given site is possible. Note that all sites present for a specific mRNA-miRNA combination are assigned the same interaction score, because it is not clear which of these sites mediate a possible interaction. This results in four models, called the "bottom" models, all predicting a scaled probability of interaction for all sites of their own site type, covering all sites present in the dataset. The next step is to incorporate this information in a "top" model, essentially similar to the "Site Count" model described in Section 5.1.1. The predicted chance of interaction provided by the "bottom" models can be used as a feature for a mRNA-miRNA combination. In this case, differing site lengths is not an issue, since they all have their own model with their own number of features, and only the result is considered on the mRNA-miRNA combination level. We have nevertheless chosen to make a different feature for each site type, since they may have other influences or importances on the mRNA-miRNA level. This results in a model with five features on the number of sites, as in the "Site Count" model, and four features with predictions form the "bottom" models. Sadly, the problem of multiple sites of the same type still exists. In order to avoid selecting a site to incorporate without a decision rule, the minimum, median and maximum predictions per site type are included in the "top" model. The final data table of this model holds five features for the number of sites and 12 features for the predictions of the "bottom" models.

The general structure and information flow of the "Stacked" model is represented in Figure 5.2. It can be seen that the basic structure of Figure 5.1 is used as a building block, and occurs five times, once for each model. Normally, no predictions are made for the training set, since these predictions are too optimistic because the model was optimised to make good predictions for these cases and the actual label was considered. However, in order to train the "top" model we need training data, so the predictions on the training set made by the "bottom" models have to be used. In this case the information flow is as follows: the "bottom" models are trained on their own training set, resulting in four trained "bottom" models. Predictions of the training data are made with these models, and minimum, median and maximum per site and per mRNA-miRNA are added as features of the training set for the "top" model. This "top" model is trained and all training data is discarded. In a second stage, the "bottom" models receive their respective test sets and predictions for these sites are made. The minimum, median and maximum per site and per mRNA-miRNA are added to the test set of the "top" model, which makes the final predictions for the mRNA-miRNA combinations present in the test set.

Strong points

- Data usage: all available information is used.
- Fair data usage: all site types are treated equally in terms of amount of information.

Weak points

- Complexity in structure: two levels, five models have to be trained.
- Information condensation: the "top" model only receives a summary of the actual site features, resulting in possible loss of information.

5.3 Other prediction algorithms

In order to construct our models and check their performance, some other algorithms developed to predict mRNA-miRNA interaction will briefly be discussed in this section. The choice of which algorithms to compare with was based on performance of the algorithms, frequency of use and availability of predictions for our dataset.

5.3.1 MirTarget2

MirTarget2 (Wang and El Naqa, 2008) is a machine learning based prediction algorithm, published in 2008. The main idea of MirTarget2 greatly resembles the "Extended Site Count" model discussed in the previous section. It focusses on mRNA-miRNA combinations with exactly one 7mer-m8 site, all other combinations will be considered non-interacting. The training set was made up of 6 miRNAs and 3027 mRNAs, but since only mRNA-miRNA combinations that have 7mer-m8 sites were considered, this resulted in only 1461 combinations: 454 interactions and 1017 non-interactions. A total of 131 features were used, spanning the same categories as those used in our work: conservation of nucleotides in the seed site, nucleotide composition in the flanking regions and position on the mRNA. MirTarget2 also considered the accessibility of the seed site region, which will only be introduced in a later version of our models. However, the presence of other sites is hardly included: one feature describes whether this site is actually a 8mer site, and also has a match at nucleotide nine of the miRNA, or not. Besides this, only the number of 7mer-A1 sites is included. The most predictive features where selected and included in a SVM model. Although it is not stated which features are included, it is mentioned that the conservation over different species was most decisive.

5.3.2 MiRanda

MiRanda (John et al., 2004), published in 2004, does not use any machine learning techniques, and is thus independent of a training set. It is based solely on the seed site matches between the mRNA and miRNA for which the interaction as to be predicted. A site is said to induce interaction if it passes three thresholds. The first is a matching score threshold, which means that the alignment between mRNA and miRNA has to be good enough in order to count as a match. The second is a free energy of the duplex formation threshold, taking into account that a mRNA-miRNA match and fold have to be stable enough to be possible. The third threshold

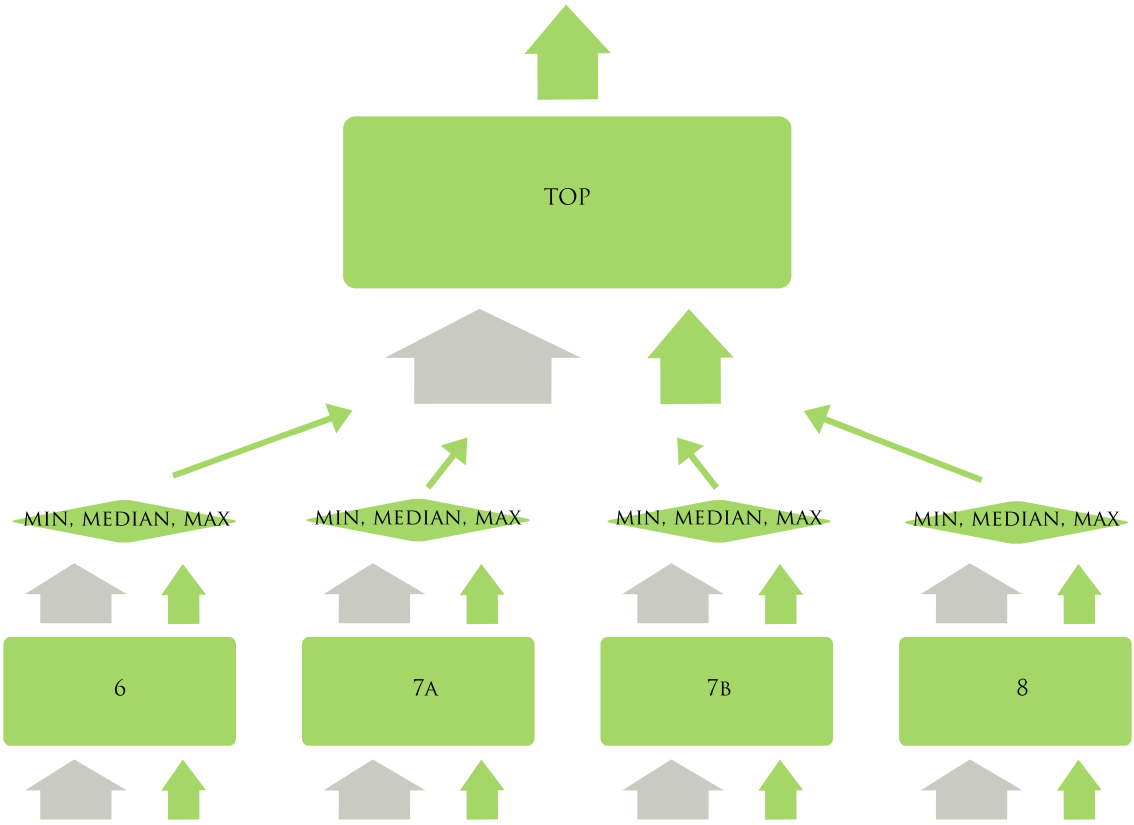


Figure 5.2: Structure and information flow of the "Stacked" model. All training data is represented in grey, test data and trained models are represented in green. The four "bottom" models, each considering one site type, are trained on their own training set. With these models, predictions for this same training set are made. For each mRNA-miRNA combination and site type, minimum, median and maximum of the predictions are computed and added as features to the "top" model training set. After training of this "top" model, all training data is discarded. In a second stage, the "bottom" models receive their respective test sets and predictions for the sites in these test sets are made. In the same way as in case of the training set, the minimum, median and maximum of these predictions are incorporated in the test set of the"top" model, allowing this model to make the final predictions for the mRNA-miRNA combinations present in this test set.

is a conservation threshold, representing the notion that highly conserved seed sites have more chance to be mediating interaction. How these thresholds are set is not explained, but they are probably based on wet-lab experience. Since the code is freely available under an open-source license, one can set one's own parameters if preferred.

5.3.3 PITA

PITA (Kertesz et al., 2007), published in 2007, is a parameter-free thermodynamic model. It does not use machine learning techniques and is based on the accessibility of the seed sites. It considers seed sites similar to the ones we consider, but somewhat less strictly defined. PITA decides whether interaction will take place by calculating an energy score based on the accessibility of the mRNA and the binding energy of the miRNA-mRNA complex. In order to bind, the mRNA must be accessible in the region of the site. Additional, the mRNA-miRNA bond has to be stable enough in order to exist. The energy score considered by PITA is thus

the difference between the energy needed to make a site accessible and the energy gained by binding to a mRNA. All possible sites for one mRNA-miRNA combination are mathematically combined to one final interaction score.

5.4 The settings: set-up and data usage

The four settings described in Chapter 3 can be used by model-builders in order to assess the performance of the model in situations were the user wants to make predictions for mRNAs or miRNAs that were not included in the dataset on which the final model is trained. A general overview of the data usage for each setting is given in Figure 5.3 and will be discussed in the following paragraphs.

Random combinations out The most straightforward interaction to predict is one between a mRNA and a miRNA present in the dataset, for example the interaction of *mRNA-1* with *miRNA-1*. Since we want to make predictions for this combinations, it is included in the test set. As a result, the training set obviously does not contain any information on this mRNA-miRNA combination, but it does contain information on how *mRNA-1* interacts with other miRNAs and *miRNA-1* with other mRNAs. To test this setting, we used 10-fold cross-validation, resulting in 799 mRNA-miRNA combinations included in each fold. Consequently, every run had a training set with 7191 combinations to predict the remaining 799 interactions. After all 10 runs were executed, the resulting 7990 predictions were pooled and performance assessment was done on this collection. Note that the prediction of a mRNA-miRNA combination included in the dataset is not very interesting to a researcher, since it is not useful to predict an interaction for which one already knows the true interaction form wet-lab tests.

New mRNA If a researcher is working with a mRNA not included in the dataset, he or she might want to know how well the model can predict interactions of a new mRNA with the included miRNAs. Since only a limited number of mRNAs are present in the dataset, we used leave-one-out cross-validation to test this setting, which results in 17-fold cross-validation. In each run, predictions are made for one mRNA, based on the data of the remaining 16 mRNAs. As shown in Figure 5.3, this results in 7520 mRNA-miRNA combinations in the training set and 470 combinations in the test set of each cross-validation run. After all 17 runs, all predictions were pooled and the performance was estimated on this set. In Chapter 3 we explained that this might give an over optimistic result, and AUCs have to be calculated for each miRNA. However, these AUCs where very similar to the pooled AUC, indicating that we can use the pooled results from our models.

New miRNA If a new miRNA is discovered, researchers might be interested to know if this miRNA will interact with any of the mRNAs included in the dataset. In this light, we want to know how well the model can predict interactions of the included mRNAs with a new miRNA. To test this setting we used 10-fold cross-validation, where each fold considered 47 miRNAs, which results in 47 miRNAs in the test set and the remaining 423 miRNAs in the training set of each run. Figure 5.3 shows that this corresponds to 7191 mRNA-miRNA combinations in the training set and 799 combinations in the test set of each run in the cross-validation. Note

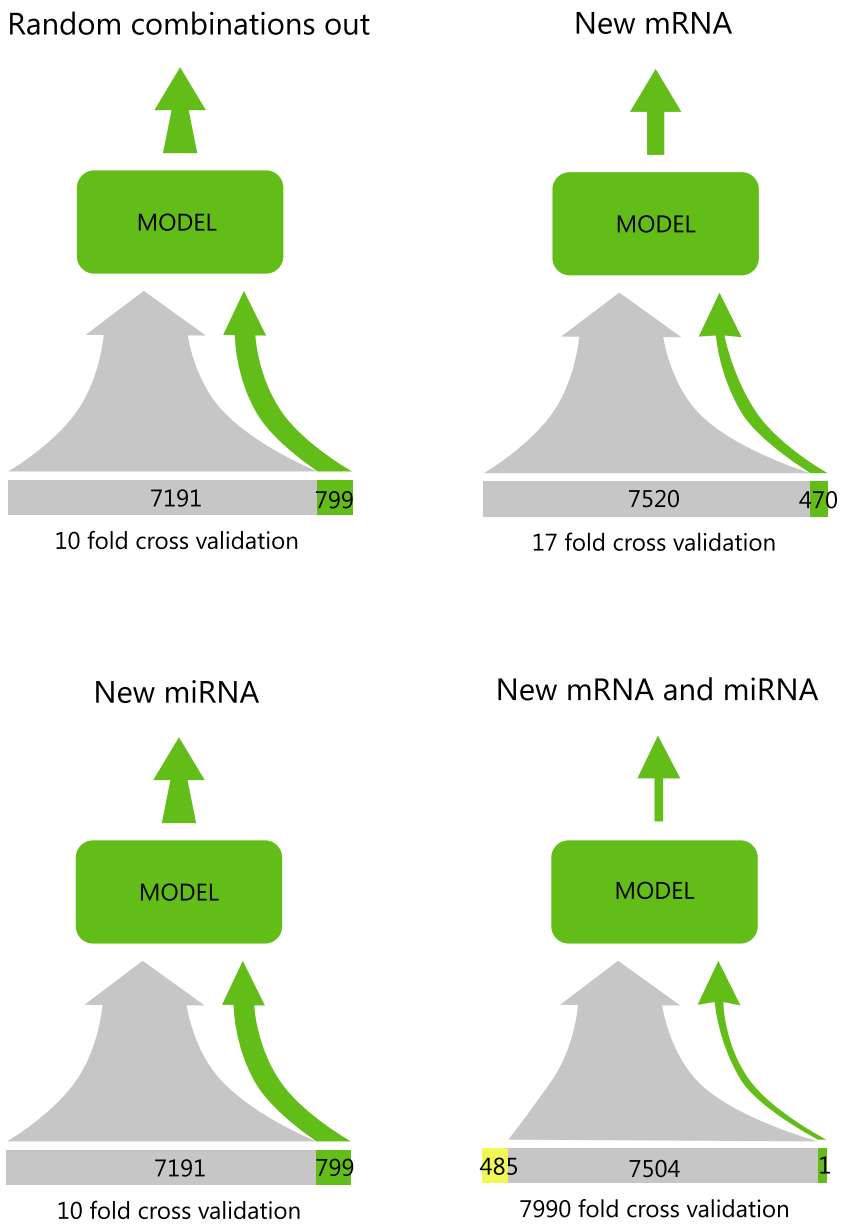


Figure 5.3: Overview of how the different settings were applied in this work. Training data is indicated in grey, test data in green and unused data in yellow. The bar at the bottom represents all the available data, indicating the number of mRNA-miRNA combinations used in the training and test set during each run of the cross-validation. In the "Random combinations out" setting, all mRNA-miRNA combinations are randomly categorised in 10 folds. During each run of the 10-fold cross-validation, nine folds are used as training set and one as test set. In the "New mRNA" setting, leave-one-out cross-validation is applied, resulting in 17 folds, each containing the 470 mRNA-miRNA combinations related to one mRNA. For the "New miRNA" setting, the 470 miRNAs are randomly assigned to 10 folds, resulting in $47 * 17 = 799$ combinations in each fold, on which 10-fold cross-validation is performed. The "New mRNA and new miRNA" setting is tackled with a kind of leave-one-out cross-validation, where each mRNA-miRNA combination in fact has its own fold. As explained in Chapter 3, all records involving the mRNA or miRNA of a mRNA-miRNA combination, indicated in yellow, have to be discarded when training the model for this combination.

that we could also have chosen to use leave-one-out cross-validation, but this was not needed since the training set already contained 423 miRNAs, which is enough for the model to make generalisations. All predictions were pooled and the performance determined on this set. In line with the previous paragraph, the AUC was also computed per mRNA, which again hardly differed from the pooled AUC.

New mRNA and new miRNA As described in Chapter 3, this is the hardest setting for a machine learning algorithm, since it has to make predictions on mRNA-miRNA combinations of which it has never seen any information. The ability to make good generalisations is of utmost importance, but as mentioned in one of the previous paragraphs, this is hard due to the limited number of mRNAs. To test this setting, we also used leave-one-out cross-validation, even though this is very computationally expensive compared to the other setting. In this case, the model has to be built 7990 times, once for each mRNA-miRNA combination in the dataset. However, since both the mRNA and miRNA have to be new to the model after it has been trained, all records involving either of these have to be removed from the training set prior to the building of the model. Figure 5.3 shows that this results in the removal of 485 mRNA-miRNA combinations for every run in the cross-validation. All 7990 predictions were pooled and used to asses the performance.

Chapter 6

Results and discussion

In this chapter, we will analyse the performance of the models described in Chapter 5. We tested the different models with multiple machine learning methods, but only two methods are withheld: logistic regression and random forests. In order to get a realistic idea of the performance of these models, we used the different settings for testing, as discussed in Chapter 3. An overview of the layout of this chapter can be seen in Figure 6.1. We will start by giving a general discussion of the performance of each model in Section 6.1, covering both the use of logistic regression and random forests. This will be followed by a comparison of the performance of the different models in Section 6.2, for which random forests is used. The chapter will be concluded with more detailed discussions of the performance of the "Stacked" model using different methods and different settings in Sections 6.3 and 6.4 respectively.

6.1 General model performance

When one wants to estimate the performance of a model, it has to be clear what is actually expected of this model, which performance measures are appropriate to check this goal and what is considered to be a good performance. We first tried to make a model with the best overall performance, which seems logical when building a model, but one has to keep in mind how the models will be used by researchers. Probably, these models will be used to make predictions for a mRNA or miRNA of interest, after which the most promising combinations can be tested in wet-lab experiments. In other words, researchers might be more interested in a model which performs well on the top scoring predictions, rather than one with a good general performance which happens to make a lot of mistakes in the top predictions. Consequently, this discussion will focus on two performance measures: the AUC, which gives an overall idea of the performance, and the accuracy in the top n predictions. Unless stated otherwise, the AUC and accuracy in the top 10 predictions are calculated on the pooled predictions, giving an idea of the performance over all mRNA-miRNA combinations in general. The only exception is Section 6.2.3, where the average top accuracy over the mRNAs is calculated. This gives the most realistic idea of the accuracy that can be expected when the interactions for a new mRNA and its 10 top scoring miRNAs are tested in web-lab experiments.

Figure 6.2 shows the distributions of the predictions made by the different models considered in this work. The x-axis represents the prediction score, the y-axis the frequency. The blue

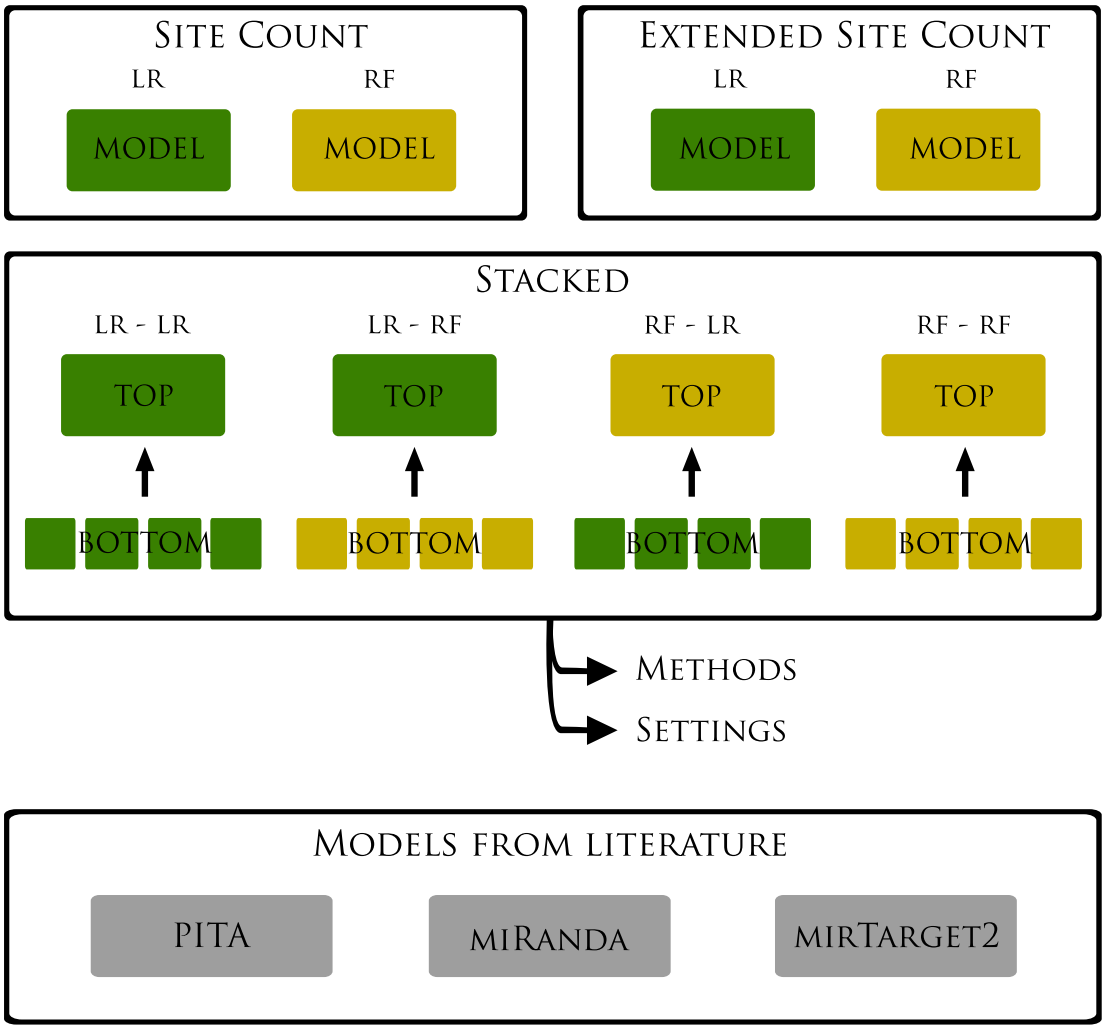
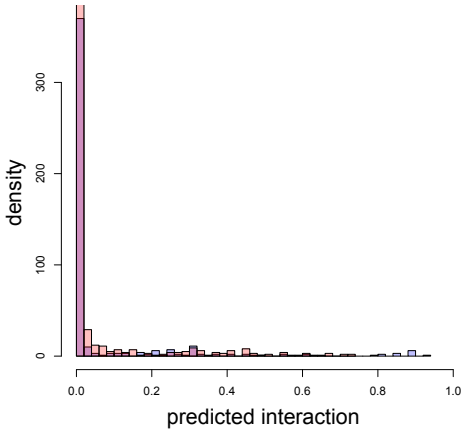
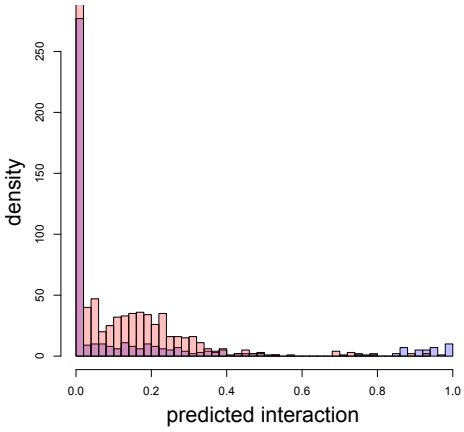


Figure 6.1: Overview of the content of this chapter. The "Site Count" model will be discussed in Section 6.1.1, where both the use of logistic regression (LR) and random forests (RF) will be considered. Section 6.1.2 includes a similar discussion for the "Extended Site Count" model. The "Stacked" model is discussed in Section 6.1.3, giving a general discussion of the four combinations of logistic regression and random forests in the "top" and "bottom" models. For this model, we provide a more detailed discussion on the influence of the methods and the different setting in respectively Section 6.3 and Section 6.4. The performance of the models from literature, i.e. PITA, miRanda and MirTarget2, is discussed in Section 6.1.4. All models, including the ones from literature considered in this work, are compared in Section 6.2.

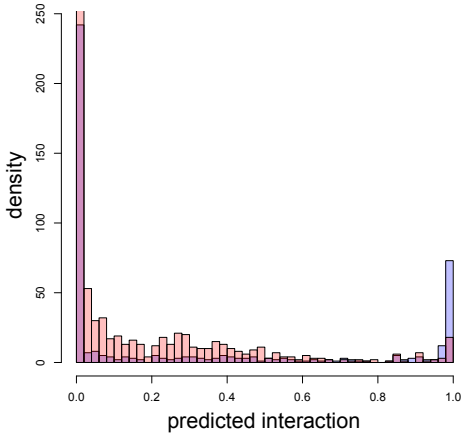
bars represent the true interactions, whereas the red bars represent the non-interacting mRNA-miRNA combinations. When comparing the different models, two silhouettes can be distinguished: a first one with bell-shaped functions and a second one with a high peak at zero and a moderate spread for all other prediction values. Generally, the methods with high accuracy in the top scoring predictions follow the second silhouette, and are machine learning based algorithms.



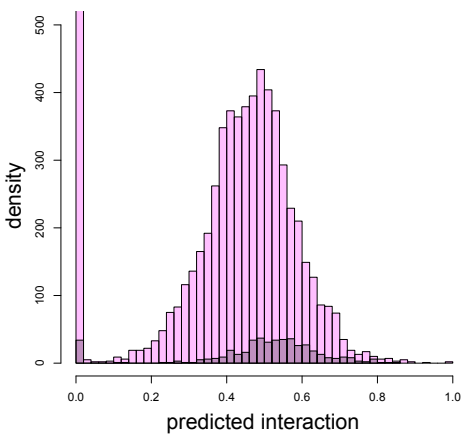
(a) Site Count model



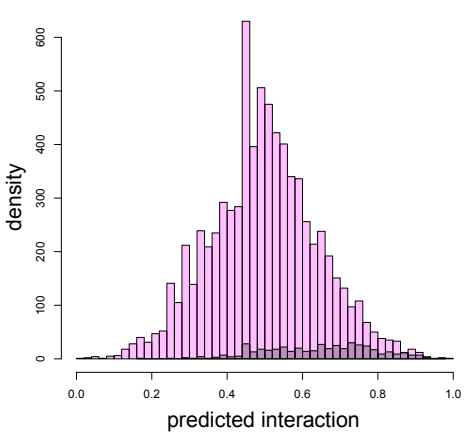
(b) Extended Site Count model



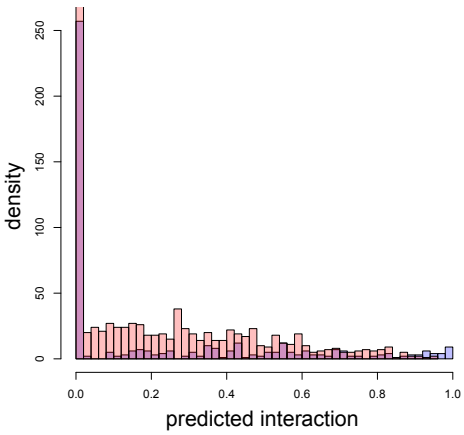
(c) Stacked model



(d) PITA



(e) miRanda



(f) MirTarget2

Figure 6.2: Histograms of the prediction score distributions made by the different models.

6.1.1 Site Count model

This is our model of minimum information, including only the number of sites of each type. Since this model has no notion of the site level features, it might overestimate the importance of the sites that are present and might consider any mRNA-miRNA combination with sites as "likely to interact". However, since the dataset holds mRNA-miRNA combinations that have sites but do not interact, this model mainly focuses on the presence of 8mer sites, which have the highest influence on interaction. When logistic regression is used, this is reflected in the loadings the different features of the model receive: the highest loading is assigned to the number of 8mer sites, followed by the number of 7mer-m8 sites and the total number of sites present. The number of 6mer and 7mer-A1 sites are not withheld in the regularised logistic regression model. Notwithstanding the simplicity of this model, it can reach a nice overall performance with an AUC of 0.79. When the "Site Count" model is built using random forests, its overall performance drops significantly, resulting in an AUC of 0.61. Although the same features receive high importance, the use of this more complex method reduces the overall performance, probably due to overfitting.

However, if we look at the distribution of the predicted scores resulting from this model, only 17 mRNA-miRNA combinations reach a score higher than 0.5 with logistic regression. This is not really a problem, but a lack of spread in the predictions is not a desirable trait. In this light, random forests perform better, having 46 predictions with a score higher than 0.5. Moreover, the accuracy on the top 10 scoring predictions is 0.6 with logistic regression, whereas this is 1 with random forests, making the latter the preferred method if one is interested in the top scoring predictions, even if its AUC is clearly lower. Figure 6.2a shows the score distributions resulting from the predictions made by the "Site Count" model using random forests in the "New miRNA" setting. Although this model makes predictions for all mRNA-miRNA combinations, a lot of those combinations receive the minimal score, representing a very low possibility of interaction. This is good for the non-interacting combinations, but almost 370 combinations are interacting and still receive this minimal score. One can conclude that only using the number of sites from each site type is clearly not sufficient to make accurate predictions for these mRNA-miRNA combinations. However, when looking at the top scoring predictions, this method performs surprisingly well. The accuracy on the top 10 scoring predictions is 1, since the first 13 predictions are true positives, and permanently drops beneath 0.5 when more than the top 40 is considered.

6.1.2 Extended Site Count model

As the name suggests, the 'Extended Site Count" model adds extra information to the features in the "Site Count model", in this case the site features of the 7mer-m8 sites. Since more information is included, this model is given the possibility to assess the importance of a present site and thus to perform better. However, the algorithms decide for themselves which features to include in the model, in case of regularised logistic regression, or to assign high importance to, in case of random forests. When we take a look at the features used in the regularised logistic regression, only the percentage of adenine and uracil in the region flanking a site (AUscore) is added to the features already used in the "Site Count" model. So even though they are present, the model seems unable to extract information from these features in order to enhance its performance, and since they are not useful, the model will just drop the features due

to regularisation. This is probably linked to the way the added features are imputed when no 7mer-m8 site is present: all added features are set to zero. In order to overcome this, we tried to impute with the mean or median, but the results only got worse. Since this model hardly includes more information than the "Site count" model, the AUC is identical: 0.79. When random forests are used, however, also the relative position of the site and the conservations of the fourth nucleotide in the site are added as features. This is interesting, especially since the algorithm seems to have a preference for a specific position. Observations like these show a useful trait of machine learning models: since the algorithm is not explicitly programmed and learns from the data, it can find links humans are not able to find on sight. In this way, researchers can test different features of which the influence is not known and deduce their importances. Here, the two features with striking importance are *phyloP_Mammal.4* and *phyloP_Primate.4*, representing the conservations of the fourth nucleotide within mammals and within primates, calculated with the phyloP algorithm. Including these features results in an AUC of 0.7, which is a 0.09 rise compared to the "Site Count" model.

Regarding the score distributions, the same conclusions can be made as with the "Site Count" model: random forests give a higher accuracy in the top scoring predictions and a clearer spread on the scores. As before, this makes the random forest method the preferred method if one is interested in selecting a limited number of promising mRNA-miRNA combinations. The prediction score histograms of the "Extended Site Count" model with random forests under the "New miRNA" setting are represented in Figure 6.2b. The peaks at zero contain 7049 non-interactions and 276 true interactions, which is a quarter less than in case of the "Site Count" model.

6.1.3 Stacked model

As described in Chapter 5, the "Stacked" model is able to include all available information. Due to the two level structure and five models, interpretation is not as straightforward. We will start by analysing the "bottom" models, since their output will be used as features in the "top" model.

Bottom models In contrast to the previously discussed models, random forests performs better than logistic regression in terms of AUC, reaching respectively 0.77 and 0.64 as mean AUCs of the four "bottom" models. The importance of the features shows that logistic regression includes no features at all in the 6mer and 7mer-m8 models, predicting non-interaction for all these sites. The logistic regression models of the 7mer-A1 and 8mer sites include the AUscore, favouring higher AUscores in sites from interacting mRNA-miRNA combinations. The 7mer-A1 model also includes some phyloP conservations, a few of which even have negative loadings, indicating that lower conservations in these positions might increase the chance on interaction. However, one has to be careful when interpreting these loadings: the sores should always be considered as a whole, since loadings might change dramatically if one feature is left out. In this case, some conservations might have a negative influence on interactions, given that all other features have the loading they are assigned. They might, for example, add a nuance to the high loading assigned to the AUscore. Additionally, the loadings for these ambiguous polyP conservations are so small that it is not clear how much attention should be granted to them. If the models are built with random forests, more features receive high importances. Generally,

the AUscore and most phyloP conservations seem important, with a preference for the fourth, fifth and sixth nucleotide position in the site. The fact that this model is able to deduce information from the same site features that were hardly used in the "Extended Site Count" model shows the advantage of treating each site type separately: all features are easily interpretable for the models, since there is no ambiguity on how to handle or impute the irrelevant features of different site types.

Top model The "top" model includes the number of sites of each type and the minimum, median and maximum predictions of each "bottom" model. The good thing of these features is that it is not wrong to impute them with zero if no sites are present, since they resemble a scaled probability that these sites will cause the interaction, which is zero if no sites are present. All combinations of logistic regression and random forests in "top" and "bottom" models give comparable performances. However, logistic regression in the 'top" and random forests in the 'bottom" models performs slightly better than the others, with an AUC of 0.8. The combination with the lowest AUC is the one were all models are built with random forests, resulting in an AUC of 0.74.

In terms of feature importance, all method combinations prefer the features coming from the "bottom" models over the site count features, although the number of 8mers and the total number of sites are sporadically included. This can be explained by the fact that the features resulting from the "bottom" models also include a big part of the information present in the site count features. If the three features, i.e. minimum, median and maximum, resulting from one "bottom" model are present, there is at least one site of this type. In other words, if these features are not zero in the dataset of the "top" model, the feature holding the number of sites from this site type will at least be one. If minimum, median and maximum are different from one another, there are multiple sites of this type. Since more information can be deduced form the features resulting from the "bottom" models, which also give an indication of the importance of a site, most site count features are no longer used. There is no real preference between the minimum, median or maximum scores and most models include a few of each. The most frequently used features are the minimum and maximum of the 7mer-m8 model, accompanied by the maximum of the 8mer model, which are present in almost all of the "top" models.

6.1.4 Models from literature

This section contains a general score distribution for the three methods from literature considered in this work. A further comparison to our models is made in Section 6.2.

PITA Although the AUC of this algorithm, being 0.71, is quite high, it is not suited for the selection of a small subset of promising mRNA-miRNA combinations. The distribution of the predictions made by the PITA algorithm can be seen in Figure 6.2d. It is striking to see that the distribution of true interactions falls fully within the distribution of non-interacting mRNA-miRNA combinations. This results in an accuracy of 0 for the top 10 predictions, since the first 11 predictions are false positives. Over 1900 mRNA-miRNA combinations of our dataset do not receive a score, since they do not contain sites regarded by PITA, and the algorithm cannot make predictions for these combinations. However, only 34 combinations display true

interaction, which is not a great loss compared to the issue that all top scoring predictions are false positives.

MiRanda When considering the AUC, miRanda performs best compared to the other two methods form literature, reaching an AUC of 0.76. The histogram of the predictions made by miRanda is shown in Figure 6.2e. Like in the case of the PITA algorithm, the plot displays two bell-shaped functions. However, since the means of both distributions differ more than in case of the PITA algorithm, the distribution of the true interactions is not fully included in that of the non-interacting mRNA-miRNA combinations. As a result, the performance on the top scoring predictions is better, reaching an accuracy of 0.5 on the top 10 predictions. The accuracy permanently drops beneath 0.5 when more than the 24 top scoring predictions are considered.

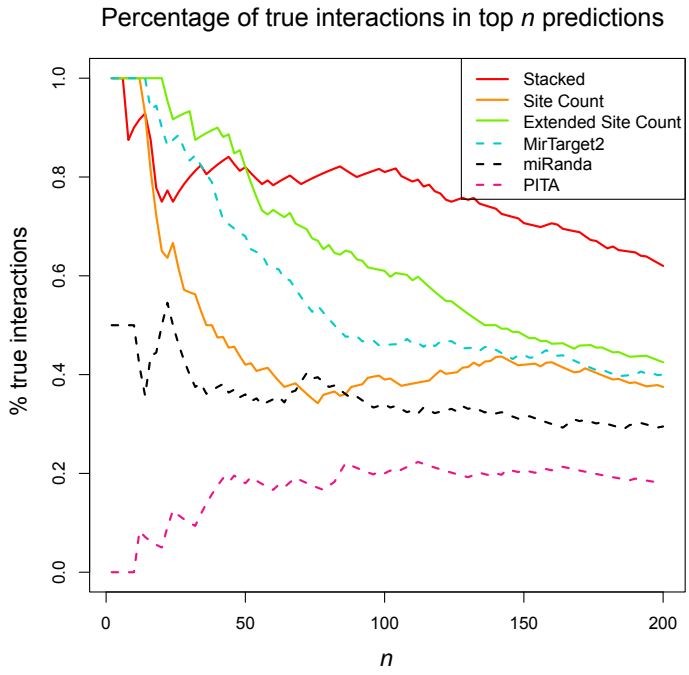
MirTarget2 If one wants to make a selection of promising mRNA-miRNA combinations for wet-lab experiments, this is by far the best algorithm from literature that is considered in this work. However one has to keep in mind that this algorithm will only consider combinations with one 7mer-m8 site. MirTarget2 also displays the lowest AUC of the models form literature, being 0.67, although we have discussed that this might not be the most relevant performance measure for researchers. Figure 6.2f represents the score distributions resulting from this model. It can be seen that the silhouettes of these distributions shows closer resemblance to the ones resulting from our models than to those of PITA or miRanda. The peaks at 0 are due to the fact that no predictions are made for the 7132 mRNA-miRNA combinations without a 7mer-m8 site, which were imputed with the minimum prediction score. Note that this includes more than 250 true interactions, which will never be detected by the algorithm. The accuracy in the top scoring predictions is 1, since the top 14 predictions are all true positives. When more than the top 82 predictions are considered, the accuracy permanently drops beneath 0.5.

6.2 Comparing the models

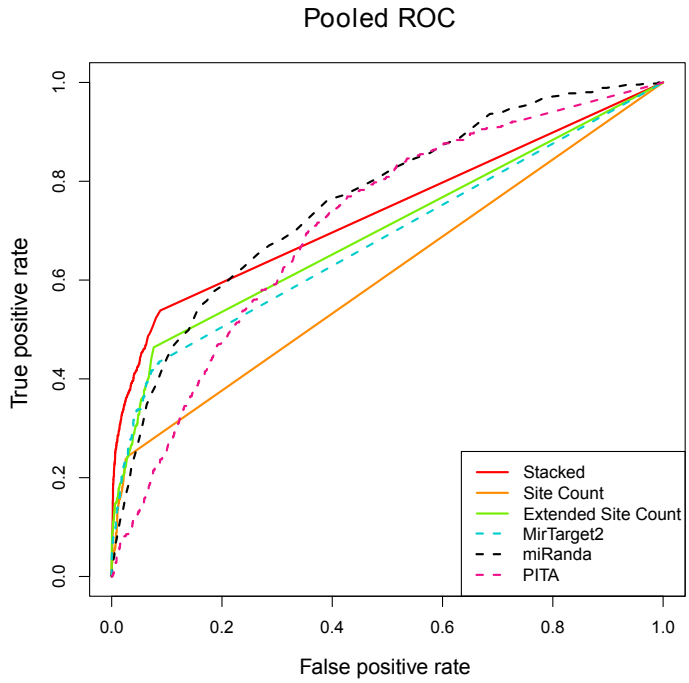
In order to compare the models, we will fix the setting to "New miRNA" and the method to random forests. This is not ideal, since the previous section showed that some models perform better with one method and other models with a second method. Still, it will give us the possibility to compare some aspects of the different models. We will start by a comparison if all data is used, followed by a specific case when all data on 7mer-m8 sites is not considered. Finally, we also look at the performance per mRNA, which might be of interest to researchers working on a specific mRNA.

6.2.1 Comparison on all data

In this main comparison we will consider the model performances on all the available data. Figure 6.3 visualises the performance of the different models developed for this work, also including the three published methods described in Chapter 5. Since the models will probably be used to make a selection of promising mRNA-miRNA combinations to be tested in wet-lab experiments, the performance in the top scoring predictions is of special interest.



(a) Visualisation of the percentage of true positives, the accuracy, on the top 200 predictions of each model.



(b) ROC curves of the different models.

Figure 6.3: Performance plots for comparison of the different models. In this case, our models are built for the "New miRNA" setting and use random forests.

Performance in top scoring predictions Figure 6.3a shows the percentage of true positives in the top n predictions of each model, which also resembles the accuracy in this selection. As can be seen, the "Extended Site Count" model performs best on the top 50 predictions, making this the preferred model if one is mainly interested in the very top. If more predictions are considered, the "Stacked" model has the highest percentage of true positives, making this the best overall model. At first, it may seem quite surprising that the "Extended Site Count" model outperforms the "Stacked" models in the top predictions, since the latter uses all available information, including that integrated in the former. However, the structure of the models is very different and the "top" model in the "Stacked" approach only receives a kind of summary of the site features, whereas the "Extended Site Count" model receives the exact site features. On the long run however, when one considers more than the top 50 predictions or looks at the overall performance, the "Stacked" model outperforms the "Extended Site Count" model, since it receives extra information on all site types instead of just the 7mer-m8 sites. As expected, the "Site Count" model is a weaker version of the "Extended Site Count", since it has exactly the same structure as the latter, but does not include any site features. When comparing the methods from literature, MirTarget2 performs best, but is still slightly weaker than the "Extended Site Count" model. This similarity was expected, since the concept of MirTarget2 greatly resembles that of the "Extended Site Count" model and also focusses on 7mer-m8 sites. However, it does not include the number of sites as extensively as our model does and only makes predictions if 7mer-m8 sites are present. Our model, on the other hand, makes predictions for al mRNA-miRNA combinations. If no 7mer-m8 sites are present, the "Extended Site Count" model will act as the "Site Count" model and make predictions based on the number of 6mer, 7mer-A1 and 8mer sites. miRanda and PITA have a low accuracy in their top predictions for the mRNA-miRNA combinations in our dataset.

Overall performance: ROC Figure 6.3b shows the ROC curves of all models considered in this work. To assess their performance, one can look at a few different things, including the shape of the curve and the area under the curve (AUC). As explained in Chapter 3, ROC curves which are positioned more to the top left corner of the graph represent a better overall performance, resulting in a higher AUC. The slope in the bottom left of the curve represents the performance on the top scoring predictions, resulting in high slopes for high accuracies. Clearly, two shapes of ROC curves can be distinguished: the ones with a nice arc, like most common ROC curves, and the ones with a distinct kink that is connected to the top right corner by a straight line. This arises when a vast number of the predictions have the same, minimal score. In our models, this minimum score is zero, and reflects the predicted impossibility of interaction for these mRNA-miRNA combinations. Since the majority of combinations in our dataset in fact does not interact, this is a good thing. However, all interacting combinations that receive a score of zero will never be expected to interact by our models. In order to reduce the number of true interactions that receive a minimal score, more features might be needed. All our models have this second shape and it can be seen that their AUC is linked to the amount of information included, resulting in the highest AUC for the "Stacked" model since its ROC curve tends most to the top left of the graph. Note that the MirTarget2 algorithm also displays this kink, which is due to the fact that this algorithm only makes predictions for the mRNA-miRNA combinations where 7mer-m8 sites are present. For all other mRNA-miRNA combinations, no predictions are made and the interactions have been categorised as non-interacting by assigning the minimal interaction score. The two other algorithms, PITA and miRanda, which are not

based on a single site type and have a more flexible definition of sites, show a conventional arched ROC shape. However, the slope in the bottom left of their curves is lower than these of the other methods, showing their weaker performance on the top predictions.

6.2.2 Comparison if 7mer-m8 sites are left out

All algorithms seem to use different definitions of what a 7mer-m8 site is: we use it in a very strict sense where 7 nucleotides need to match, but some algorithms, like MirTarget2, allow less strict interpretations. This complicates the comparison between these algorithms, since differences are not only due to the models themselves, but also due to the training data used. MirTarget2 does not make predictions if no 7mer-m8 sites, according to their definition, are present. Still, the accuracy drops only about 20% in the top scoring predictions when the 7mer-m8 sites, according to our definition, are left out. As a result, MirTarget2 is still better than the stacked model on the top 10 scoring predictions. It can be concluded that the way in which the sites are defined has an influence on the performance of the algorithms. However, our "Extended Site Count" model uses the more strict definition and, if all data is used, outperforms MirTarget2 in the top scoring predictions. As a consequence, the more strict definitions might be better if a high accuracy in the top predictions is what is wanted. Note that for MirTarget2 this might not be feasible, since it does not consider the mRNA-miRNA combinations without a 7mer-m8 site, making a broader definition necessary to cover a reasonable amount of combinations. Our approach is to include more strictly defined site types and give them the chance to have different influences, rather than using one broader site type.

6.2.3 Comparison per mRNA

Researchers could be interested to make a selection of 10 miRNAs which might interact with an mRNA of interest. Table 6.1 shows the accuracy in the top 10 predictions for each mRNA made by the different models. For our models, the "New mRNA" setting is used, since this is what we want to analyse. More information on this setting and the difference with other settings can be found in Section 6.4. In order to avoid the need of setting a threshold, we assume that a good model will rank all true interactions in the top, so as long as more than 10 interactios per mRNA are present, the top 10 scoring predictions per mRNA should be true positives. However, the left most column in the table, showing the true number of interactions for every mRNA, indicates that two mRNAs have less than 10 interactions. For these mRNA, the accuracy can never reach 1, altough this is not due to the model. However, since this is also possible for a new mRNA, for which the total number of interactions with the considered miRNAs is unknown, we have chosen to include them in this top 10 accuracy measurement. Clearly, not all top scoring predictions of the models considered are true positives, resulting in accuracies smaller than 1, and some mRNAs seem to be easier to predict than others. From the column means at the bottom of the table, it can be seen that the "Site Count" model has the best mean performance, where on average half of the top 10 scoring predictions per mRNA are true positives. The "Stacked" and "Extended Site Count" model have an average performance comparable to that of MirTarget2, while the others have a poorer performance. The fact that the most simple model gives the best mean performance over the mRNAs is probably due to its ability to make

Table 6.1: Overview of the accuracy in the top 10 predictions per miRNA for each model, using the "New mRNA" setting for our models. The last column shows the total number of true positives of each mRNA in the dataset.

mRNA	SC	ESC	S	PITA	miRanda	MirTarget2	nr of interactions
ALK	1	0.4	0.4	0	0.1	0.3	13
BRCA1	0.3	0.1	0.1	0.1	0.2	0.3	25
BRCA2	0.7	0.1	0.2	0	0.2	0.4	18
EZH2	0.9	0.9	0.7	0.7	0.8	0.6	21
FBXW7	0.8	0.6	0.5	0.5	0.3	0.8	48
HRAS	0.6	0.3	0.3	0.2	0.1	0.2	6
MYB	0.4	0.8	0.8	0.4	0.4	0.4	41
MYC	0.7	0.3	0.5	0.2	0.2	0.3	25
MYCN	0.6	0.4	0.9	0.3	0.3	0.3	31
MYT1L	0.1	0.3	0.4	0.2	0.3	0.6	39
NOTCH1	0.1	0	0.1	0.4	0.5	0.5	22
PALB2	0.9	0.7	0	0	0.1	0.1	10
PHF6	0	0	0.3	0.2	0.2	0.1	30
PHOX2B	0.3	0.1	0.2	0	0.1	0.4	23
RB1	0.4	0.3	0.4	0	0.5	0.3	31
TP53	0.2	0.1	0.1	0.1	0	0.1	7
ZEB2	0.9	0.8	0.5	0.1	0.5	0.7	55
mean	0.524	0.365	0.376	0.200	0.282	0.376	

good generalisations. The other models might systematically assign higher scores to some mRNA and make good predictions for these instances, resulting in a high accuracy on the top of the pooled data. However, when the mean of all mRNAs is considered, the influence of mRNAs which are hard to predict can be seen, resulting in a lower mean accuracy on the top 10 predictions per mRNA.

A more general overview can be seen in Figure 6.4 showing the mean accuracy over the mRNAs for all models, where the considered top for which the mean accuracy is computed ranges form the top 1 to the top 10 predictions per mRNA. The previously made conclusion, that the "Site Count" model performs best for the top 10 predictions per mRNA in the "New mRNA" setting, can also be seen form this plot. However, when less than the top 5 is considered, MirTarget2 displays the best performance. Calculating the accuracy in the top predictions is a relevant test to estimate the performance in a research setting where one does not know the true number of interactions and a set number of promising combinations will be tested. However, it might be interesting to know how many of the true interactions can easily be fount. This can be tested by calculating how many of the n true interactions per mRNA are also in the top n predictions made for this mRNA. In this case there are no issues with mRNAs or miRNAs that show less than ten interactions. The results for this test are represented in Table 6.2, showing that the difference between the performance of the "Site Count" model and the other models is even bigger in this case.

Table 6.2: Average recall in the top n predictions per mRNA, with n the number of true interactions for this mRNA.

Site Count	Extended Site Count	Stacked	PITA	miRanda	MirTarget2
0.6252091	0.3028160	0.2792296	0.1696495	0.2515049	0.3206957

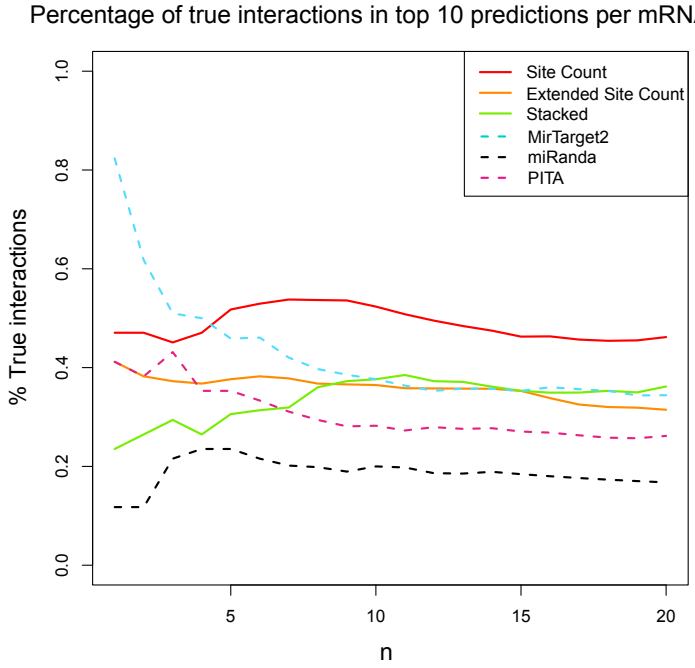
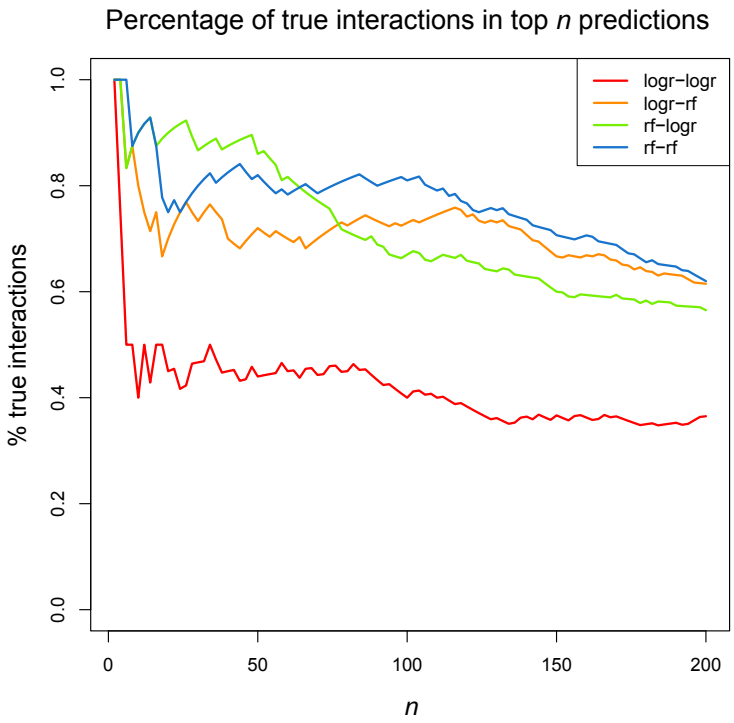


Figure 6.4: Overview of the mean accuracy over the mRNAs for the top n predictions made by the differnt models. When less than the top 5 predictions are considered, MirTarget2 performs best. However, when a top selection of more than 5 miRNAs per mRNA is considered, the "Site Count" model outperforms all others.

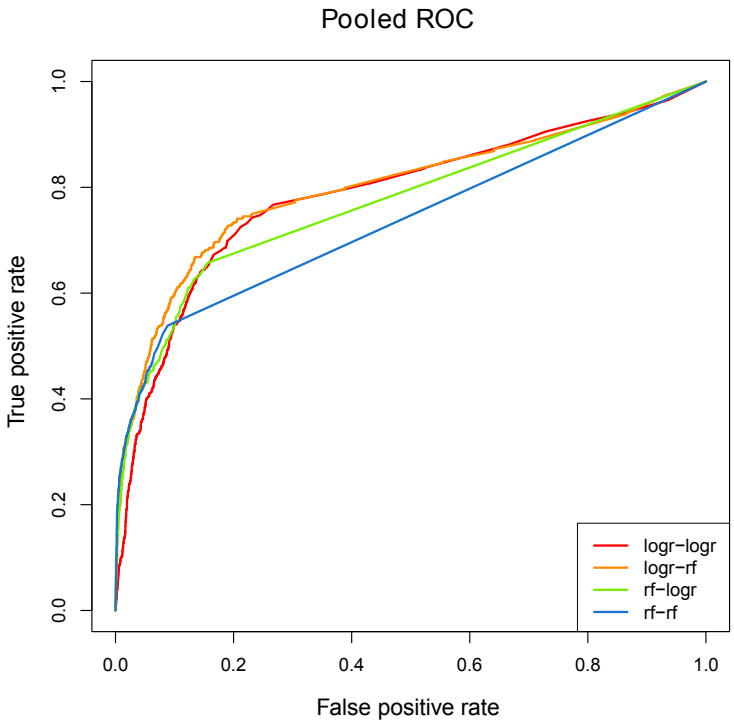
6.3 The methods

In Section 6.1, we have drawn some general conclusions regarding the performance of our models when using logistic regression and random forests. In this section, we will further discuss the influence if these methods on the "Stacked" model in the "New miRNA" setting. An overview of the performances using different combinations of methods for the "top" and "bottom" models can be found in Figure 6.5. The method of the "top" model is mentioned first, followed by the method used in the "bottom" models, where logistic regression and random forests are abbreviated as respectively "logr" and "rf". The use of a random forests model on at least one level clearly improves the performance in the top predictions, as can be seen in Figure 6.5a. It is hard to decide which combination of methods will be optimal, since no combination outperforms all others in the top 200 predictions. However the model that was built exclusively with random forests might be preferred based on this plot, depending on the demands of the researchers who will use the model.

In Figure 6.5b, the ROC curves of the method combinations are represented, given an overall idea of the model performance. As mentioned in Section 6.2, all our models show the ROC silhouette with a distinct kink. When comparing the methods, it can be seen that the ROC curves of models with logistic regression are positioned higher, reflecting the higher AUCs. The kink in the ROC curves of these models also appears higher and more to the right, indicating that respectively more interacting ans non-interacting mRNA-miRNA combinations receive a score differing form the minimum.



(a) Visualisation of the percentage of true positives, the accuracy, on the top 200 predictions of each method.

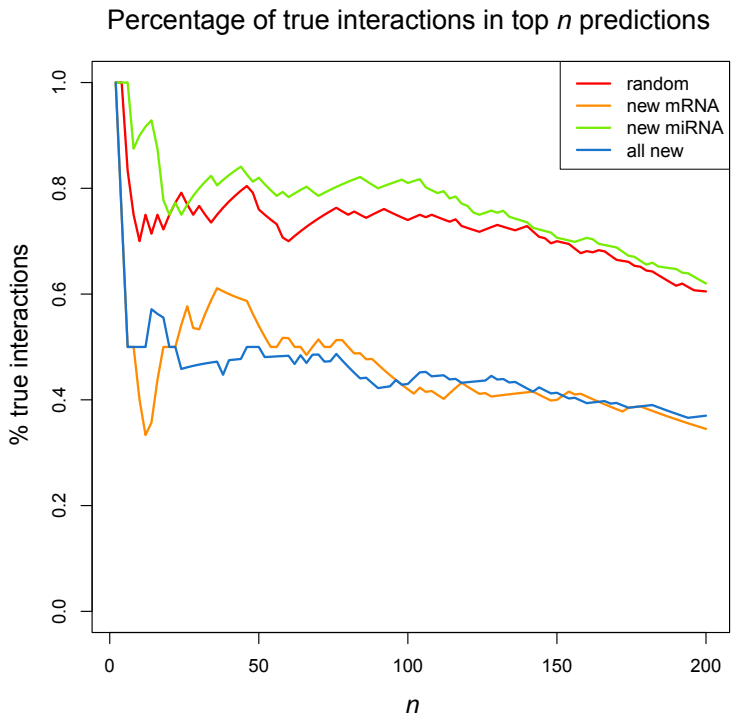


(b) ROC curves of the different methods.

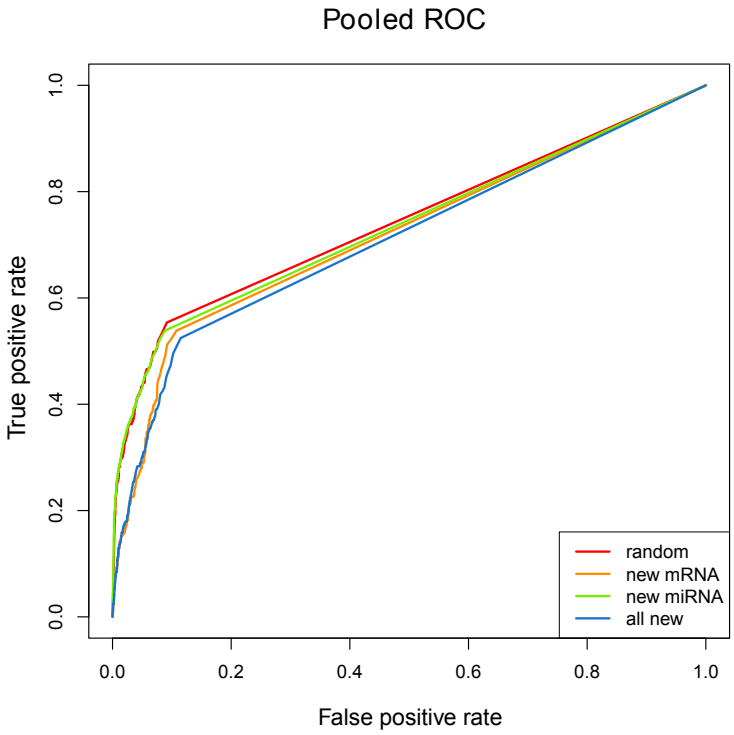
Figure 6.5: Performance plots for comparison of the different methods for the "Stacked" model in the "New miRNA" setting.

6.4 The settings

As mentioned in Chapter 4, the used dataset contains 17 mRNAs and 470 miRNAs. Since there are clearly more miRNAs than mRNAs, this will have its effects on the performance in the different settings. To compare these settings, one has to consider a single model and method by which this model is trained. Here, we have chosen to discuss the setting for the "Stacked" model, trained with random forests. As mentioned before, the easiest setting is "Random combinations out", which also reaches the highest AUC: 0.743. However, the performance in case of the "New miRNA" setting is very similar, with an AUC of 0.737, and their ROC curves can hardly be distinguished from one another in Figure 6.6b. Even though the model does not receive any information on the new miRNAs during training in this setting, it is able to make good predictions on their interactions with the 17 mRNAs included in the dataset. This is due to the high number of miRNAs in the training set, giving the model the possibility to extract a lot of information and enabling it to make good generalisation. If the model would be used in practice, all information in the dataset will be used to train the model, giving it even more miRNAs to learn from. Surprisingly, the "Stacked" model even performs better on the top scoring predictions of new miRNAs than on random combinations out, as can be seen in Figure 6.6a. As expected, the model displays a poorer performance for the "New mRNA" setting. Since only 17 mRNAs are present, the possibility of overfitting on specific traits of a mRNA is high. This makes it hard for the model to generalise, which is exactly what is needed to make good predictions for unseen cases. It reaches an AUC of 0.721, but as can be seen from Figure 6.6a, the accuracy in the top scoring predictions is always about 50%, whereas this is around 80% for the previous to settings. It must be mentioned that an accuracy of 50% can be acceptable, since a random classifier would only reach 5.7%, which is the fraction of interacting mRNA-miRNA combinations in the dataset. In case of the "New mRNA and new miRNA" setting, the performance is comparable to the one of the "New mRNA" setting, resulting in an AUC of 0.711. This shows that the additional novelty of an miRNA hardly has any influence on the performance of the model.



(a) Visualisation of the percentage of true positives, the accuracy, on the top 200 predictions of each setting and three models form literature.



(b) ROC curves of the different settings.

Figure 6.6: Performance plots for the comparison of the different settings. The model used is the "Stacked" model, built with random forests.

Table 6.3: Overview of the results. Some situations were not tested since the proceeding results showed that only poor performances could be expected and the computational demands for these combinations is very high.

(a) Our models

		AUC				top 10 accuracy			
model	method	random	mrna	mirna	both	random	mrna	mirna	both
SC	lr	0.789	0.775	0.784	0.760	0.6	0.4	0.4	0.4
	rf	0.611	0.598	0.608	0.597	1	0.9	1	1
ESC	lr	0.787	0.774	0.785		0.6	0.4	0.5	
	rf	0.694	0.697	0.697	0.677	1	0.4	1	0.4
S	lr-lr	0.787	0.768	0.787		0.6	0.4	0.4	
	lr-rf	0.805	0.776	0.800		0.8	0.4	0.8	
	rf-lr	0.787	0.734	0.773		0.8	0.1	0.9	
	rf-rf	0.743	0.721	0.737	0.711	0.7	0.4	0.9	0.5

(b) Models form literature

model	AUC	top-10 accuracy
PITA	0.710	0
miRanda	0.760	0.5
mirTarget2	0.680	1

Chapter 7

Conclusions

All algorithms considered in this work can be evaluated in different ways, often resulting in different conclusions regarding the performance. In this light, it is important to know exactly what is expected of an algorithm in order to decide which is appropriate. The conclusions for some approaches, illustrated in Figure 7.1, will be summarised.

In Chapter 6, most of the comparisons where made in light of examining the influence of a given setting or method on the models. This gives us the possibility of trying to understand how the models differ, relate to one another and what might be done to improve them. To make this comparison as good as possible, these analyses were done on the pooled data in the "New miRNA" setting. Due to the high number of miRNAs, the data was very suitable for this, and it usually gives a more realistic view on the performance than when both mRNA and miRNA are seen before. Based on these conditions, some general conclusions can be made. First, the number of sites present form each site type gives an indication of mRNA-miRNA interaction. In addition, these site types influence interaction in a different degree, making it important not to focus on one site type. Secondly, when more information on these sites is added to the models, the overall performance improves. However, the way in which this information is added to the model greatly influences the performance. The "Stacked" model might lose information since only a summary of the site info is included in the "top" model, but the "Extended Site Count" model only considers one site type, losing a lot of information. When a general conclusion has to be made, the "Stacked" model is best if one wants to do an extensive lab-test, involving for example 200 mRNA-miRNA combinations, whereas the "Extended Site Count" is most relevant if only 10 combinations will be tested. However, it is not very likely that a researcher will be interested in any combination of mRNAs and miRNAs. When examining a disease which is linked to a specific gene, researchers only want to make predictions for one mRNA. This is not the same machine learning problem as previously described, since a model that performs well in this case has to perform equally well for all mRNAs, even if this means the overall performance is lower. This is the case considered in Section 6.2.3, from which it can be concluded that the "Site Count" model performs best if this is your wet-lab research question. When considering the methods form literature, MirTarget2 performs well when the top scoring predictions are considered, both on the pooled data and per mRNA. However, for each of these cases, the best performing models mentioned above are able to outperform MirTarget2.

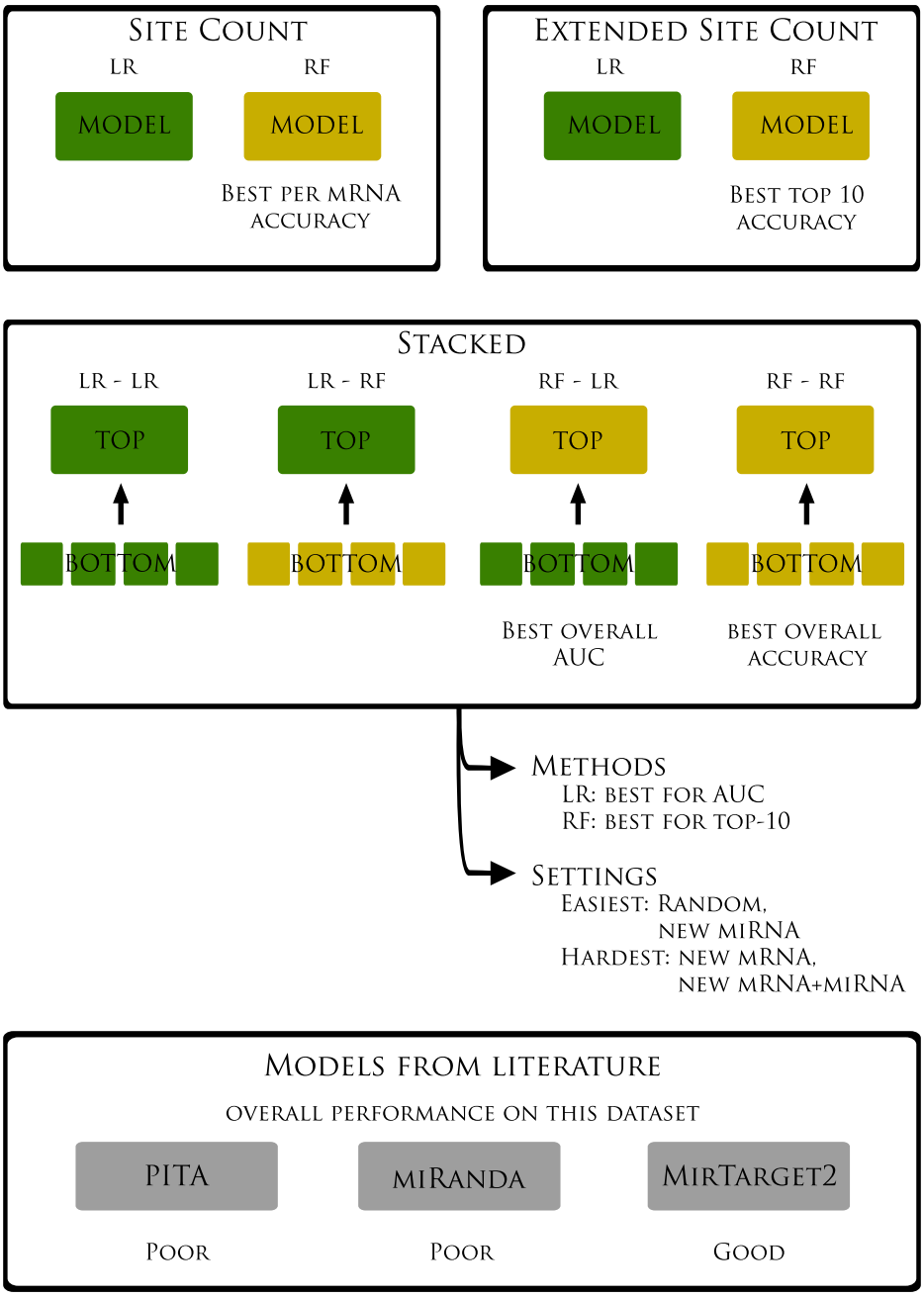


Figure 7.1: Overview of the different models and the conclusion on their performance.

Notwithstanding the complexity of predicting mRNA-miRNA interactions, the use of machine learning algorithms seems very promising. Generally, further optimisation in function of the exact expectations and applications of our model could lead to even more powerful predictors.

Bibliography

Bartel, D. P., Lee, R., and Feinbaum, R. (2004). MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, 116:281–297.

Blaze, J. and Roth, T. L. (2013). Epigenetic mechanisms in learning and memory. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(1):105–115.

Calin, G. A., Cimmino, A., Fabbri, M., Ferracin, M., Wojcik, S. E., Shimizu, M., Taccioli, C., Zanesi, N., Garzon, R., Aqeilan, R. I., Alder, H., Volinia, S., Rassenti, L., Liu, X., Liu, C.-g., Kipps, T. J., Negrini, M., and Croce, C. M. (2008). MiR-15a and miR-16-1 cluster functions in human leukemia. *PNAS*, 105(13):1–6.

Calvo, B., Bengoetxea, E., and Larra, P. (2010). Bioinformatics Methods in Clinical Research. *Methods in Molecular Biology*, 593:25–49.

Cissell, K. A. and Deo, S. K. (2009). Trends in microRNA detection. *Analytical and Bioanalytical Chemistry*, 394(4):1109–16.

Cooper, G. M. (2000). *The Cell: A Molecular Approach*. Sunderland (MA): Sinauer Associates.

Costa, F. F. (2010). Non-coding RNAs: Meet thy masters. *BioEssays*, 32(7):599–608.

Croce, C. M. (2008). Oncogenes and cancer. *The New England Journal of Medicine*, 358:502–511.

Desouza, G. and a.C. Kak (2002). Vision for mobile robot navigation: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):237–267.

Esau, C., Kang, X., Peralta, E., Hanson, E., Marcusson, E. G., Ravichandran, L. V., Sun, Y., Koo, S., Perera, R. J., Jain, R., Dean, N. M., Freier, S. M., Bennett, C. F., Lollo, B., and Griffey, R. (2004). MicroRNA-143 regulates adipocyte differentiation. *The Journal of Biological Chemistry*, 279(50):52361–5.

Farazi, T. a., Spitzer, J. I., Morozov, P., and Tuschl, T. (2011). miRNAs in human cancer. *The Journal of Pathology*, 223(2):102–15.

Fawcett, T. (2004). ROC graphs : notes and practical considerations for researchers. *Machine Learning*, pages 1–38.

Fawcett, T. O. M. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 316:291–316.

Ferrucci, D., Brown, E., Chu-carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., and Prager, J. (2010). Building Watson : an overview of the DeepQA project. *AI Magazine*, pages 59–79.

Goldberg, A. D., Allis, C. D., and Bernstein, E. (2007). Epigenetics: a landscape takes shape. *Cell*, 128(4):635–638.

Gong, H., Liu, C.-M., Liu, D.-P., and Liang, C.-C. (2005). The role of small RNAs in human diseases: potential troublemaker and therapeutic tools. *Medicinal Research Reviews*, 25(3):361–81.

Guzella, T. S. and Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7):10206–10222.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning, Second Edition: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 0002-2009. corr. 3rd edition.

Heneghan, H. M., Miller, N., and Kerin, M. J. (2010). Role of microRNAs in obesity and the metabolic syndrome. *Obesity Reviews*, 11(5):354–361.

Iorio, M. V. and Croce, C. M. (2012). MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Molecular Medicine*, 4(3):143–159.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004). Human microRNA targets. *PLoS Biology*, 2(11):e363.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nature Genetics*, 39(10):1278–84.

Larranaga, P. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112.

Li, X., Zhang, Y., Zhang, Y., Ding, J., Wu, K., and Fan, D. (2010). Survival prediction of gastric cancer by a seven-microRNA signature. *Gut*, 59(5):579–85.

Lyons-Weiler, J., Patel, S., and Bhattacharya, S. (2003). A classification-based machine learning approach for the analysis of genome-wide expression data. *Genome Research*, 13(3):503–12.

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill Education (ISE Editions), 1st edition.

Nayeem, A., Sitkoff, D., and Jr, S. K. (2006). A comparative study of available software for high-accuracy homology modeling : From sequence alignments to structural models. *Protein Science*, pages 808–824.

Pillai, R. S., Bhattacharyya, S. N., and Filipowicz, W. (2007). Repression of protein synthesis by miRNAs: how many mechanisms? *Trends in Cell Biology*, 17(3):118–26.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–21.

Poy, M. N., Spranger, M., and Stoffel, M. (2007). microRNAs and the regulation of glucose and lipid metabolism. *Diabetes, Obesity & Metabolism*, 9 Suppl 2:67–73.

Rajewsky, N. (2006). microRNA target predictions in animals. *Nature Genetics*, 38 Suppl(June):S8–13.

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, pages 206 – 226.

Sato, F., Tsuchiya, S., Meltzer, S. J., and Shimizu, K. (2011). MicroRNAs and epigenetics. *The FEBS journal*, 278(10):1598–1609.

Sauder, J. M., Arthur, J. W., and Dunbrack, R. L. (2000). Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins*, 40(1):6–22.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Si, M.-L., Zhu, S., Wu, H., Lu, Z., Wu, F., and Mo, Y.-Y. (2007). miR-21-mediated tumor growth. *Oncogene*, 26(19):2799–803.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. a., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–50.

Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M., and Mattick, J. S. (2010). Non-coding RNAs: regulators of disease. *Journal of Pathology*, (October 2009):126–139.

Tong, a. W. and Nemunaitis, J. (2008). Modulation of miRNA activity in human cancer: a new paradigm for cancer gene therapy? *Cancer Gene Therapy*, 15(6):341–55.

Verikas, a., Gelzinis, a., and Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2):330–349.

Voinnet, O. (2001). RNA silencing as a plant immune. *Trends in Genetics*, 17(8):449–459.

Waegeman, W., Pahikkala, T., Airola, a., Salakoski, T., Stock, M., and De Baets, B. (2012). A kernel-based framework for learning graded relations from data. *IEEE Transactions on Fuzzy Systems*, 20(6):1090–1101.

Wang, X. and El Naqa, I. M. (2008). Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, 24(3):325–32.

Yanaihara, N., Caplen, N., Bowman, E., Seike, M., Kumamoto, K., Yi, M., Stephens, R. M., Okamoto, A., Yokota, J., Tanaka, T., Calin, G. A., Liu, C.-G., Croce, C. M., and Harris, C. C. (2006). Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell*, 9(3):189–98.

Zotos, P., Roubelakis, M. G., Anagnou, N. P., and Kossida, S. (2012). Overview of microRNA target analysis tools. *Current Bioinformatics*, pages 1–14.