Machine Learning Approaches to siRNA Efficacy Prediction

by

Sahar Abubucker

B.E., Madras University, 2000

THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Computer Science

The University of New Mexico

Albuquerque, New Mexico

May, 2005

Dedication

To all the wonderful people I have met.

Acknowledgments

I would like to express my deepest thanks to my advisor, Professor Terran Lane, for his invaluable guidance, support, and encouragement.

Machine Learning Approaches to siRNA Efficacy Prediction

by

Sahar Abubucker

ABSTRACT OF THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Computer Science

The University of New Mexico

Albuquerque, New Mexico

May, 2005

Machine Learning Approaches to siRNA Efficacy Prediction

by

Sahar Abubucker

B.E., Madras University, 2000

M.S., Computer Science, University of New Mexico, 2005

Abstract

RNA interference (RNAi) is being widely used to study gene expression and gene regulation via selective knockdown of gene expression, which is important for functional genomic studies. RNAi refers to the biological process by which short interfering RNA (siRNA) after incorporation into the RNA induced silencing complex (RISC) degrades complementary messenger RNA (mRNA) sequences. This knockdown of mRNA prevents it from producing amino acid sequences that are responsible for gene expression. Recent studies indicate that all siRNAs do not produce the same knockdown effects. Due to the high cost of synthesizing siRNAs and the extensive effort required to test siRNAs, rational siRNA design—a priori prediction of functionality for specific siRNAs—is a critical task for practical RNAi studies. Therefore, a key component of RNAi applications is the selection of *effective siRNA sequences*—ones that degrade at least 80% of the targeted mRNA. The goal of siRNA efficacy prediction is to aid in designing siRNA sequences that are highly efficient in degrading target mRNA sequences. Most of the current algorithms use positional features, energy characteristics, thermodynamic properties and secondary

structure predictions to predict the knockdown efficacy of siRNAs. In this work, frequently occurring patterns in siRNAs are used to make predictions about their efficacy. Time after transfection is shown to be an important factor in siRNA efficacy prediction—a feature that has been ignored in previous efficacy studies. The relevance of these extracted patterns to previous results and their biological significance are analyzed. Random feature sets are generated and the ability of these sets to predict efficacy are studied and their results discussed. Our algorithm does not require any specialized hardware and consistently performs better than other publicly available efficacy prediction algorithms.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

RNA interference (RNAi) is widely studied for its importance in genomic studies and its potential use in therapeutics. It is the mechanism by which messenger RNA (mRNA) sequences are degraded by complementary short interfering RNAs (siRNAs) incorporated into RNA induced silencing complex (RISC). Recent studies indicate that all siRNAs do not produce equal knockdown effects. In vivo experiments to observe siRNA functionality are expensive and time-consuming. Different studies have proposed several characteristics of the siRNA that indicate functionality, including the presence or absence of certain nucleotide in certain positions in the siRNA, thermodynamic properties related to stability and secondary structure. We propose a computational approach to siRNA efficacy prediction that makes use of frequently occurring positional patterns in the siRNA data to discriminate between functional and non-functional siRNAs. Our algorithm performs better than other publicly available efficacy prediction algorithms. We found that the time after transfection is an important feature in determining efficacy. Feature sets consisting of random positional patterns in the data showed reasonable performance leading to the hypothesis that the positional features used for predicting efficacy could be a consequence of limited data.

1.1 RNA Interference

RNAi, also known as co-suppression [24], quelling [37] and post-transcriptional gene silencing [8], plays a part in cellular anti-viral defenses and transposon silencing mechanisms [34]. RNAi refers to the biological process by which mRNA is destroyed or degraded when exposed to complementary siRNA sequences incorporated into RISC. The siRNAs are formed from double stranded RNA (dsRNA) or are synthesized externally and then introduced into the cell.

The RNAi pathway was discovered by Fire and Mello in 1998 [12]. They injected dsRNA—a mixture of both sense and antisense strands—in to *C. elegans* and observed that silencing was much greater than when using either the sense or the antisense strands alone. Further, just a few molecules of dsRNA per cell were found sufficient to completely silence the target gene's expression. It is believed that organisms may have developed the RNAi mechanism to protect themselves from viruses and other exogenous agents that produce long dsRNA [46, 34]. This dsRNA is degraded by the naturally occurring RNAi pathway in the organism. RNAi has been observed in nematodes, plants, fungi, vertebrates and mammals and appears to be present in almost all eukaryotic organisms.

1.1.1 RNAi Pathway

A nucleotide (nt) is a subunit of DNA or RNA and is made up of one of adenine (A), guanine (G), cytosine (C) or uracil (U) (in RNA) or thymine (T) (in DNA), along with a phosphate molecule, and a sugar molecule. The RNA molecule is formed from a sequence of these nucleotides. The complementary nucleotides of A, C, G and U are U, G, C and A respectively. When long dsRNA from an external source is introduced into the cell, it is recognized by Dicer, a member of the RNase III family of dsRNA-specific ribonucleases. Dicer cleaves the dsRNA to produce siRNA duplexes of lengths 19 - 21 nt [1,2]. Each

siRNA strand has a 5’ phosphate group and a 3’ hydroxyl group and has a 2 nt overhang at the 3’ end [46]. The siRNA duplex separates into sense and antisense strands and one of the strands is taken up by a RNA-protein complex, referred to as RISC [28].

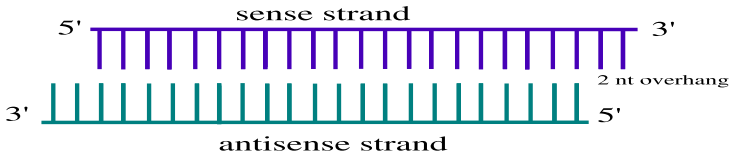

Figure 1.1: Anatomy of an siRNA

Activation of RISC requires an ATP-dependent unwinding of the siRNA duplex. Both the sense and antisense strands of the siRNA are capable of directing RNAi but specificity depends on the anti-sense strand. The active RISC then targets mRNA transcripts that have sequence complementarity with the siRNA sequence. The targeted mRNA sequences are cleaved into smaller fragments which are then degraded. This results in sequence-specific removal of mRNA in targeted genes, which are then not expressed at the protein level. Figure 1.2 graphically illustrates the RNAi pathway initiated by the introduction of dsRNA. The knockdown effects induced by RNAi are usually transient but using vector-based delivery methods, stable RNAi can be induced. As described in section 7.1, RNAi is not immediate and there is a time course associated with the process. RNAi has also been shown to be inheritable in *C. elegans* [14].

In mammals, it was observed that long dsRNAs, with lengths more than 30 nt activate the PKR kinase pathway in the cell, also known as the interferon response. This causes

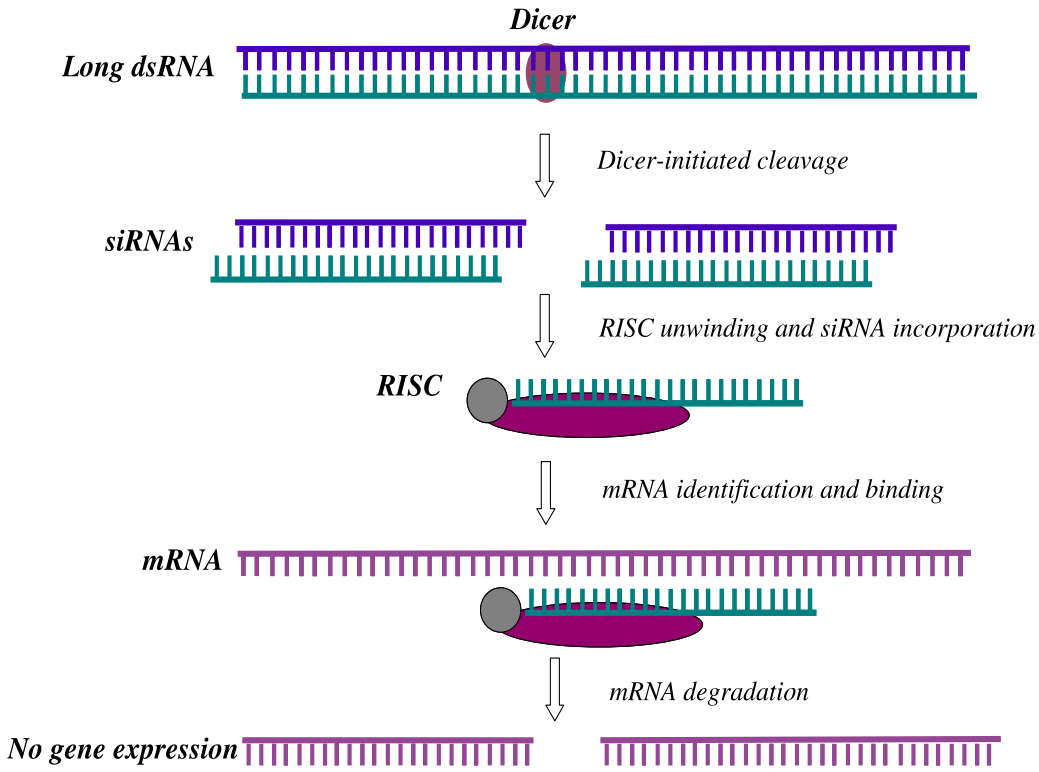

Figure 1.2: RNA Interference Pathway

non-specific degradation of mRNA, and leads to apoptosis or cell death. However, using synthesized siRNAs of lengths 21 - 23 nt [10] does not evoke the interferon response and provides effective silencing by RNAi.

In addition to siRNAs, gene silencing can also be caused by micro RNAs (miRNA). miRNAs are small RNAs, processed from double stranded hairpin structures that are encoded in the genome, and are believed to be involved in gene regulation. Unlike siRNAs, which work by mRNA degradation, miRNA work by suppressing translation of mRNA to protein. miRNAs have been shown to function as siRNAs by binding to perfectly complementary mRNA sequences to cause degradation. On the other hand, siRNAs can act as

mRNAs with 3 - 4 nt mismatches and G-U mismatches [38]. This demonstrates that it is only the degree of complementarity which determines the functionality of the siRNA or miRNA sequence [21, 9]. However, the effects of miRNA-like behavior of siRNAs on efficacy experiments have not been extensively studied so far. This work only deals with the siRNA and so this document will not delve in to the details of miRNA, but it is important to note that they are closely related.

1.2 Importance of RNAi

Gene expression is the process by which the information encoded in a gene is converted into amino acid sequences. When a gene is expressed, DNA is transcribed into mRNA which then acts as a template for the production of proteins. Thus, degradation and regulation of mRNA help govern cellular mRNA and, therefore, protein levels that result from gene expression. Complete genomes are being sequenced for several organisms and there is an increasing need for studying gene behaviors and functions. Changes in phenotype, resulting from RNAi, gives information about the functions of the targeted gene. Therefore, a mechanism like RNAi, which employs existing cell machinery, is highly useful.

RNAi is also becoming increasingly important in developing therapeutic applications for a number of diseases due to its potential for specific targeted silencing [41]. During gene expression, there are three stages where genes causing diseases can be controlled – transcriptional, post-transcriptional, and post-translational intervention. Traditionally, drugs for disease control have been targeted towards proteins, which occurs in the post-translational phase. RNAi targets the protein-producing mRNA and can thereby control disease earlier - in the transcription phase. RNAi has been successfully used to target diseases such as AIDS [29], neurodegenerative diseases [51], cholesterol [42] and cancer [4] on mice with the hope of extending these approaches to treat humans.

1.3 siRNA Efficacy

Recent studies indicate that out of the possible siRNAs that can be synthesized against a particular target, only a fraction of these are successful in causing any degradation [19, 16] and, further, all siRNAs do not result in equal knockdown effects [19]. The efficacy of the siRNAs differed among different target sites in the same target mRNA. Therefore, it is important to select effective siRNA sequences—ones that are highly functional in causing more than a certain percentage of the target mRNA sequence to degrade. Reynolds et al. [35] observed in their siRNA knockdown experiments that properties of the target mRNA did not affect knockdown and that efficacy seems to be solely based on properties of the siRNA. However, other studies [18, 27] have indicated that secondary structure and thermodynamic properties (related to stability) of the siRNA are also important determinants of functionality. There is no consensus on the importance of each of these properties. In most studies, siRNAs causing knockdown of more than 80% of the target mRNA are considered highly efficient but the threshold varies depending on the level of silencing required. The goal of siRNA efficacy prediction is to aid in designing siRNA sequences that are highly efficient against their target mRNA sequences. In this study, we use a knockdown threshold of 80% and siRNAs causing this amount of knockdown or more are considered functional.

1.4 siRNA Specificity

Another factor to be considered in siRNA design, in addition to potency or efficacy, is the specificity of the siRNA. While maximum degradation of target mRNA is required, silencing of non-target mRNA should be avoided. siRNA-mediated gene silencing is generally believed to be highly sequence-specific. In some cases, even a single base mismatch between an siRNA and its mRNA target abolished gene silencing [10]. However, gene ex-

pression profiling in cultured human cells demonstrated silencing of non-targeted genes. Even eleven complementary matches out of the 19 nucleotides of an siRNA was enough to cause silencing [22]. This indicates that siRNAs may cross-react with targets of limited sequence similarity. Therefore, due consideration must be given to the implications arising from siRNA specificity in design algorithms. Qiu et al. [33] have examined the effects of siRNA lengths on off-target error rates. We do not consider specificity issues in this work, but due consideration must be given to the implications arising from siRNA specificity, specifically before using the highly functional siRNAs derived from design algorithms.

1.5 Classification Problem

Given a set of inputs and its corresponding output class, the problem of classification is to determine the output class of an input which has not been seen before by the classifier. In a two-class problem, the output can be of the form +1 or -1 and the classifier predicts in which of the two classes an input belongs. Classification of siRNAs is a two-class problem with output labels—efficient and inefficient—determined by the knockdown threshold. This can be formally stated as follows. Let X be the input data and Y be the class labels. Let Y $\in$ {-1,1} be the class label, where -1 indicates a non-functional siRNA and 1 indicates a functional siRNA. Let X $\in \{a, c, t, g\} \times \{\mathrm{t}\}$, where t refers to the time after transfection. Efficient siRNAs are those causing degradation equal or above the knockdown threshold and inefficient ones are those causing less knockdown than the threshold. The classification problem in siRNA efficacy prediction is to determine the characteristics of the siRNA which determine its efficacy.

1.6 Overview of this work

In this work, an algorithmic framework based on classification using frequently occurring patterns in the siRNA sequences is proposed. The patterns are extracted using the Apriori [1] algorithm and act as features to a classifier. In addition to these Apriori patterns, the time at which the knockdown was measured, and the difference between the number of A/U nt in the sense and antisense strands are included as additional feature. The reasons for using these features are discussed in the following sections. These features are used as input in addition to the siRNAs' efficacies for the support vector machine (SVM) classifier, a well-known machine learning classification method [48]. The SVM learns from input data and predicts functionality of siRNAs whose efficacies are not known. A pictorial view of the Pattern-Classifier algorithm is given in Figure 1.3. Each component of the algorithm will be discussed in detail in the following sections.

The next section contains a brief review of existing techniques. The Apriori algorithm is discussed in detail in Chapter 3. The SVM classifier and its characteristics are described in Chapter 4. Chapter 5 gives an overview of the Receiver Operating Characteristics (ROC) Curve which is used for comparison, while Chapter 6 describes the implementation details. Chapter 7 discusses the results of the Pattern-Classifier approach. The Pattern-Classifier algorithm is compared with efficacy prediction algorithms by : Reynolds et al. [35], Amarzguioui et al. [2],and Sætrom et al. [45, 44] using ROC curves. The results are analyzed and the conclusions are stated. We show that the Pattern-Classifier algorithm has the best performance among other publicly available algorithm. We show that time after transfection is an important factor in efficacy prediction. Finally, we show that random features have reasonably good performance and this leads to interesting implications concerning the prevalent use of positional features for efficacy prediction.

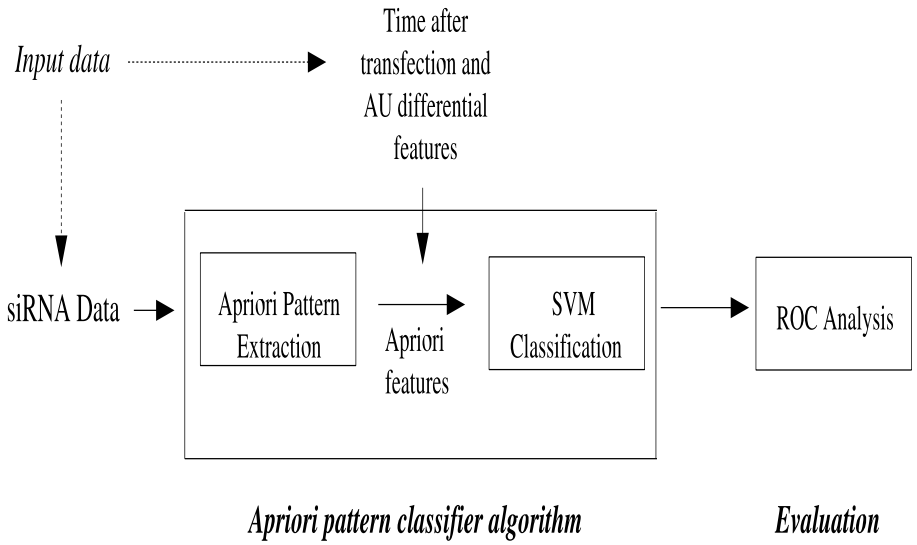

Figure 1.3: Diagrammatic representation of the Apriori pattern-classifier algorithm and the evaluation step

Chapter 2

Review of Existing Techniques

Most studies suggest that positional features (presence or absence of specific nucleotides in certain positions within the siRNA), thermodynamic properties and secondary structures of siRNAs are important in predicting efficacy [2, 10, 35, 27, 18]. However, the positional characteristics suggested by different studies vary and there are also conflicting results concerning the effects of secondary structure on functionality. A study by Amarzguioui et al. [2] and another by Holen et al. [19] did not find any correlation between functionality of the siRNA and secondary structures of the target mRNA. Reynolds et al. [35] observe in their siRNA knockdown experiments that target mRNA properties did not effect knock-down and that efficacy seems to be solely based on properties of the siRNA. But other studies by Luo et al. [27] and Heale et al. [18] suggest that secondary structure of the siRNA plays a role in determining efficacy. Schwarz et al. [39] suggest that RISC takes up either the sense strand or the antisense siRNA strand depending on their stabilities and the strand with lower stability at the 5' end is preferred. Khvorova et al. [25] also found that functional siRNAs are characterized by low base-pairing stability at the 5' end of the anti-sense strand.

The following sections briefly summarize the results of several studies and also the

prediction algorithms emerging from them.

2.1 Tuschl Rules

The earliest techniques for designing effective siRNAs were proposed by Elbashir et al. [10]. They suggested that synthesizing siRNA duplexes of lengths 21 nt—19 nt base-paired sequence with 2 nt 3' overhang at the ends—mediates efficient cleavage of target RNA. The results of their study are summarized below :

- Antisense target cleavage was affected more when changes were made to the 5' end of the sense siRNA than when changes were made to its 3' end, indicating that the 5' end of the siRNA is more important for target RNA cleavage.
- The most effective 2 nt 3' overhangs in the antisense siRNA strand had the following sequence types - NN/UG, NN/UU, NN/TdG and NN/TT, where N stands for any nucleotide and dG is 2'-deoxyguanosine, a 2'-deoxy modification.
- The changes in sequence located between the 3' end and the middle of the antisense siRNA completely abolished target RNA recognition, but mutations near the 5' end of the antisense siRNA exhibited a small degree of silencing.
- Nucleotides in the center of the siRNA, located opposite to the target RNA cleavage site, are important specificity determinants and even single nucleotide changes in these positions reduce RNAi to undetectable levels.

2.2 Reynolds Rules

Reynolds et al. [35] analyzed a set of 180 siRNAs. They divided the siRNAs in to different groups based on their functionality to find properties with high correlation to functionality.

- <F50 - knockdown less than 50%
- >F50 - knockdown of 50% or more
- >F80 - knockdown of 80% or more
- >F95 - knockdown of 95% or more

They described a set of eight rules governing the siRNA sequence that are highly indicative in determining the extent of mRNA knockdown. These rules are listed below :

- GC content between 30% and 52%
- Presence of nucleotide A at positions 3 and 19
- Presence of U at position 10
- Absence of G or C at position 19
- Absence of G at position 13
- Presence of A/U in positions 15 through 19

This algorithm assigns a score based on the number of rules satisfied and siRNAs satisfying 6 or more rules are predicted to be functional.

2.3 Amarzguioui Method

Another study by Amarzguioui et al. [2] follows a similar scoring method but identified a different set of rules. They studied 46 siRNAs, and identified the following features of the 19 nt siRNA that correlates with knockdown of more than 70%.

- Difference in the number of A and U

- Presence of G or C at position 1
- Presence of A at position 6
- Absence of U at position 1
- Absence of G at position 19
- Presence of A/U at position 19

Each rule either adds or subtract a point when it is satisfied. Those siRNAs with a score of 3 or more are considered efficient. In this study, functionality is indicated by a knockdown of 70% or more.

2.4 Stockholm Rules

This prediction algorithm by Chalk et al. [6] incorporates the thermodynamic properties of the siRNA. The rules, called 'Stockholm rules' are summarized below –

- Total hairpin energy < 1
- Antisense 5' end binding energy < 9
- Sense 5' end binding energy in range 5 - 9 exclusive
- GC between 36% and 53%
- Middle (7 - 12) binding energy < 13
- Energy difference < 0
- Energy difference within -1 and 0

Using a scoring scheme that adds 1 for each rule satisfied, and a cutoff score of 6, efficient siRNAs can be detected. They further analyzed the siRNAs using the regression tree technique, but the energy parameters which were found to be statistically significant in their study did not get chosen as important features by this method.

2.5 Ui-Tei Rules

Ui-Tei et al. [47] analyzed 62 targets in mammalian cells and *Drosophila* cells and came up with four features which siRNAs should simultaneously satisfy to cause efficient silencing. These features which efficient siRNA should have are -

- A/U at the 5' end of the antisense strand
- G/C at the 5' end of the sense strand
- At least five A/U bases in the 5' terminal one-third of the antisense strand
- Absence of any GC stretch of more than 9 nt in length

These rules were found applicable to mammalian cells but did not apply to *Drosophila* cells.

2.6 Hseih Rules

Hseih et al. [20] identify the following features which distinguish effective and ineffective RNAi.

- Target sequences that are in the middle of the coding sequence resulted in significantly less silencing.

- Silencing by duplexes targeting the 3' untranslated region (UTR) is comparable with duplexes targeting the coding sequence.
- Pooling of four or five duplexes per gene results in highly efficient silencing.
- siRNA sequences seen to produce more than 70% knockdown showed preference to G or C in position 11 and T in position 19.

2.7 GPboost Technique

Pal Sætrom et al. [45, 44] use a genetic programming based approach utilizing specialized hardware. Genetic programming techniques operate on a population of syntax trees using operators like subtree swapping, mutation and reproduction . Their method uses a specialized pattern matching chip that evaluates individual expressions much faster than a regular expression matcher. Their algorithm extracts patterns from siRNA data (collected from different studies) using these techniques and learns to differentiate between the functional and non-functional siRNAs.

Chapter 3

The Apriori Algorithm

The Apriori algorithm [1] was developed to discover association rules in a large database of transactions which contains information about the items that are sold in a single transaction. The data that is obtained from this database offers insights about significant shopping patterns and is important for marketing and related applications. An association rule states a relationship between items in a transaction. An example association rule : when customers bought item A and item B, they also bought item C 85% of the time. Transaction databases typically have millions of transactions and are very large. The Apriori algorithm finds implications in the data effectively in a very short time.

There are two thresholds that can be specified in the algorithm.

- Confidence level - the percentage of transactions where the presence of one item of the association implies the other items
- Minimum support - the percentage of transactions where the association rule is satisfied

Both these parameters help control the number of association rules that can be extracted and the level of belief in the association rules.

3.1 Algorithm

The algorithm follows two iterative steps to find associations. It extracts large itemsets which are sets of items that have minimum support. It uses these large itemsets to find association rules with minimum confidence.

Large itemsets are found by the following way (Large itemsets are represented by L and the candidate itemsets are represented by C):

- Initialize L_1 to all large itemsets of length 1
- Initialize k to 2
- While C_k is not empty
  - Find large itemsets of length k, L_k :
    - Generate candidate itemsets C_k by using the large itemset obtained during the previous pass L_{k-1}
    - Search database to count support for itemsets in C_k
    - Add itemsets which have support greater than the minimum support to L_k
  - Form C_{k+1} from L_k
  - Increment k by 1

The algorithm produces rules by finding all non-empty subsets of every large itemset. Association rules are found from each frequent itemset l, in the following way :

- Let b = l - a
- a $\Rightarrow$ b when confidence(a $\Rightarrow$ b) $\geq$ minimum confidence and the confidence(a $\Rightarrow$ b) is given by support (l) / support (a)

This gives a rule of the form a $\Rightarrow$ b when the ratio of support(l) and support(a) is at least the minimum confidence.

The algorithm avoids multiple passes through the data by generating new large itemsets that are subsets of the large itemsets in the previous pass and using only these newly generated itemsets in the next pass. This makes the algorithm efficient for large datasets.

3.2 Apriori patterns from siRNA data

The Apriori algorithm was chosen to extract patterns from the siRNA data, as it captures some of the higher-order interaction in addition to capturing positional characteristics. Each siRNA was transformed in to a transaction by converting each positional feature to an item. The number of items in each transaction is 19, the length of the siRNA sequence. An example Apriori pattern extracted from the siRNA data is U=1, U=15. This indicates that having nt U at position 1 and the nt U at position 15 in the same siRNA is a feature. These extracted patterns are used as input features to the SVM classifier. Each feature is binary where the presence of a feature is indicated by 1 and absence of a feature is indicated by 0.

Chapter 4

Support Vector Machines

SVMs are learning systems that can be applied to labeled data to perform classification or regression. The SVM projects the training data into higher dimensional space through a kernel function. It then identifies a subset of the training data from each output class, called support vectors, and finds a separating hyperplane that has the maximum margin between the training examples and the class boundary. Maximizing this margin results in minimizing the maximum loss [5].

When a new data point needs to be classified, the model learned using the support vectors is used to make the classification. SVMs have good generalization performance, are computationally efficient, and classification is independent of the number of dimensions, making them robust in high dimensions. The SVM algorithm is also capable of distinguishing outliers in the data. SVMs are widely used in applications like pattern recognition [26], recommendation systems [3], text classification [23], and bioinformatics [40]. The SVM uses a nonlinear mapping function ϕ, that maps the data to a higher dimension, where a separating hyperplane can always be found. Each data point x_k is mapped implicitly to $y_k = \phi(x_k)$.

In a two-class classification problem, the discriminant is given by

$$g(y) = a^t \cdot y, \tag{4.1}$$

where a is the augmented weight vector and y is the transformed and augmented data vector.

Let z_k be the class labels - +1 and -1. Therefore, the separating hyperplane will satisfy z_kg(y_k) $\geq$ 1, for k = 1..n. Support vectors are those data points that determine the optimal separating hyperplane and are used in classifying new data points.

In the linear case, the decision function is given by,

$$D(x) = \sum_{i=1}^{N} \omega_i \phi_i(x) + b. \tag{4.2}$$

The decision function in the dual space is represented as a linear combination of basis functions. Here the decision function will be as follows :

$$D(x) = \sum_{k=1}^{p} \alpha_k K(x_k, x) + b, \tag{4.3}$$

where α_k are the parameters and x_k is the training data.

4.1 Kernel Function

The function K is a kernel function that returns the dot product

$$K(x, x') = \sum_{i} \phi_i(x)\phi_i(x') \tag{4.4}$$

The advantage of using the kernel function is that only the dot product needs to be computed. Therefore, the dimensions of the data does not affect computation time. Let D(x) = w.$\phi(x)$+b define a separating hyperplane in ϕ-space. Let margin M be the distance between the class boundary and the training data. The following inequality is fulfilled by the

training data.

$$\frac{y_k D(x_k)}{||w||} \geq M \tag{4.5}$$

The goal is to find a weight vector w that maximizes the value of M.

$$M^* = max_{w,||w||=1} M, \tag{4.6}$$

subject to $y_k D(x_k) \geq M$, for k = 1 .. p. The input data that satisfy $min_k\ y_k D(x_k) = M^*$ are the support vectors.

$$max_{w,||w||=1} min_k y_k D(x_k) \tag{4.7}$$

If the product of the norm of the weight vector w and the margin M is fixed to 1, maximizing the margin M is equivalent to minimizing the norm. This reduces the problem to the following quadratic problem.

$$min_w ||w||^2 \tag{4.8}$$

subject to $y_k D(x_k) \geq 1$, for k = 1 .. p. Therefore, the maximum margin is given by $M^* = \frac{1}{||w^*||}$

4.2 Classification of input features

The SVM takes in the Apriori patterns, time after transfection and the AU differential as input features along with the class labels— +1 for functional siRNAs and -1 for nonfunctional siRNAs. It learns the maximum margin hyperplane and the support vectors for this data. When given new input data without class labels, it predicts the class labels based on the learned model.

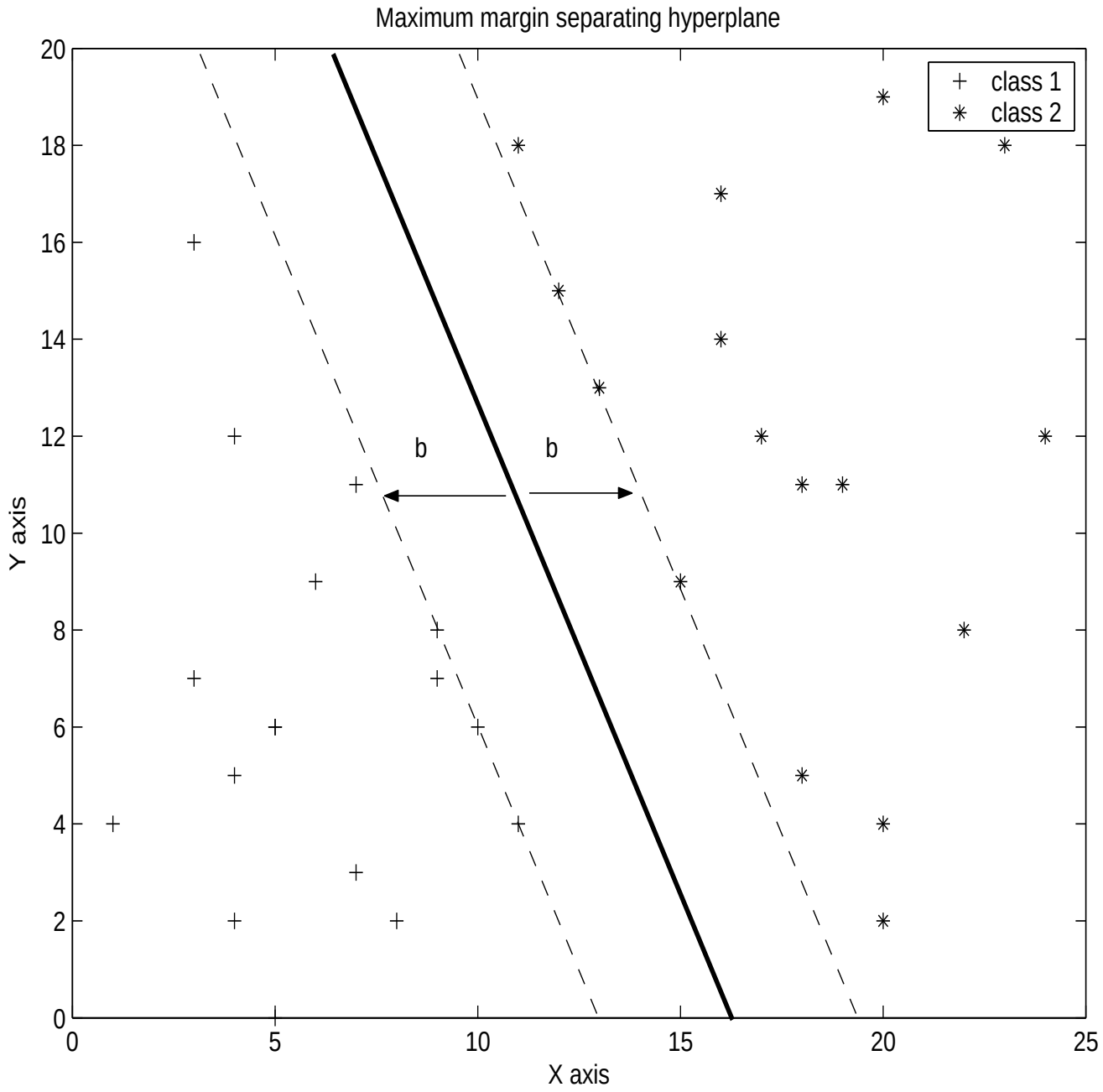

Figure 4.1: Maximum margin hyperplane for a two class problem

Chapter 5

Receiver Operating Curves

Receiver Operating Characteristics (ROC) curves are used to visualize the performance of the classifiers and therefore, useful in analyzing their performance. Area Under the ROC Curve (AUC) is another measure which maps the ROC curve in to a single scalar value. In our work, we use ROC Curves and the AUC values to compare different classifiers that predict siRNA functionality.

5.1 ROC Curves

Receiver Operating Characteristics (ROC) curves are a widely used technique that helps in visualizing the performance of the classifiers at different settings of its parameters. When using accuracy as a metric to compare classifiers, it is assumed that the class distribution is not skewed. However, this does not hold true in practice. Comparing two classifiers on the basis of accuracy may not be correct as positive and negative instances can occur in varying ratios. It is also assumed that error costs are equal, but in many domains the cost of a false positive error may not be equal to the cost of a false negative error.

A ROC curve has the following properties which makes it useful for comparing classifiers :

- Independent of class distribution.
- Independent of classification cost.

The ROC curve implicitly represents performance at different misclassification costs and class distributions. These properties makes the ROC curve a good technique for comparing classifiers that have been tested on data with different class distributions and misclassification costs.

For siRNA efficacy prediction, it is desirable to have low false positive rates. For RNAi studies, where functional siRNAs are required, it is important that siRNAs having low efficacy are not predicted to be functional. On the other hand, misclassifying siRNAs with high efficacy rates as nonfunctional is of much lesser consequence. ROC curves are useful in comparing classifiers based on this property by aiding in visualizing the trade-off between true positive and false positive rates.

For a two class distribution, a classifier can classify a data instance into the following four categories -

- False Positive - A negative instance incorrectly classified as positive.
- True Positive - A positive instance correctly classified as positive.
- False Negative - A positive instance incorrectly classified as negative.
- True Negative - A negative instance correctly classified as negative.

These can be represented by a confusion matrix as shown in Table 5.1.

True Instance	Positive	Negative
Positive	True Positive	False positive
Negative	False negative	True negative

Table 5.1: Confusion matrix

The false positive (FP) rate is the ratio between the number of false positives and the total number of negative instances. The true positive (TP) rate or recall is the ratio between the number of true positives and the total number of positive instances. The TP rate or recall is also referred to as the sensitivity of the classifier. The FP rate is also 1 - specificity, where the specificity is the ratio between the true negative and the sum of true negatives and false positives. The classifier's precision and accuracy rates can be obtained from these values. Precision is calculated by $\frac{TP}{TP+FP}$ and the accuracy is calculated by $\frac{TP+TN}{P+N}$.

The ROC curve is a plot of the false positive values against the true positive values. Each ROC point represents a (FP,TP) pair and different points are obtained by varying the classifier's parameters. The interpretation of a ROC curve is that the closer a curve is to the (0,1), the better performance the classifier has. At the point (0,1) the classifier can classify all positive instances correctly and no negative instance is classified as positive. A random classifier will have its ROC curve along the straight line from the point (0,0) to the point (1,1). A classifier whose ROC curve dominates another classifier's ROC curve has better performance. However, it is likely that no classifier's ROC curve is dominant. This implies that different classifiers perform better under different conditions and classifiers are chosen accordingly.

5.2 Area Under Curve

AUC is also another way to compare classifiers using the ROC curve. This method reduces the ROC points to one scalar value which can be used for comparing performance between

two classifiers. The AUC value of a classifier represents the probability that a random positive instance will be ranked positive instead of negative. Higher AUC values indicate better performance. Techniques like vertical averaging [32] and threshold averaging [11] are used to get variances on a classifiers performance.

Chapter 6

Implementation

This section contains details of the implementations of the Apriori algorithm, the SVM classifier and the algorithm used to obtain the ROC points and also the siRNA data that is used in this work.

6.1 siRNA Data

The siRNA data is a collection of siRNAs from different studies [35, 2, 25, 17, 20, 47, 49] and consists of the siRNA antisense sequences, their observed efficacy, the target gene along with the accession number, start and end positions of the siRNA, time after transfection, cell line, concentration of the siRNAs and the technique used for measurement. The targeted cells are either from the mouse or the human genome. This dataset was obtained from Pal Sætrom [45]. The data consists of 581 siRNAs out of which 8 siRNAs occur in two different studies and 14 siRNAs do not have data for time after transfection. All the data analysis in this work is performed on the sense siRNA sequences.

6.2 LibSVM

LIBSVM [7] is a publicly available SVM program written in Java. It contains implementations of the linear, polynomial, radial basis function, and the sigmoid kernels. This implementation of the SVM was used for solving the classification problem.

6.3 ROC algorithm

The ROC points were generated using the method in [11]. This method uses the probabilities of the output classes generated by the LIBSVM algorithm.

6.4 Weka

Weka [50] is a publicly available collection of machine learning algorithms. The weka implementation of the Apriori algorithm was used for extracting patterns from the dataset.

Chapter 7

Results and Discussion

This section contains a detailed description of all our results and conclusions. The siRNA data contains efficacy information for siRNAs taken from different organisms. From the data, we did not find any correlation between the organism and siRNA efficacy and there has not been any indication that siRNA functionality varies across organisms. Each section describes the experiment and the results of that experiment followed by discussion of the results.

7.1 Time after transfection

RNAi is not an immediate process and a certain amount of time has to elapse before all the mRNA is degraded and the previously transcribed protein is exhausted before the effects of RNAi can be observed. The knockdown rate of the target mRNA or protein level is measured after a certain time has elapsed since transfection. This time is typically 24 hours but can be longer, sometimes as long as 3 days. To study the importance of time after transfection in determining the siRNA knockdown rate, it is added as an additional feature in the classification step. The performance of the Apriori Pattern-Classifier algorithm is

studied after adding this feature. Figure 7.1 illustrates the difference in performance in including and excluding the time feature during classification. This feature considerably improves performance of the Pattern-Classifier algorithm. Some studies [15] have noticed that the knockdown rate of the siRNA varies with time. They observe that efficacy slowly increases before reaching a maximal knockdown rate at a certain time. However, previous efficacy prediction algorithms have largely ignored this feature and not considered it in the efficacy prediction process.

The implication of this result is not entirely surprising. The knockdown efficacy of the siRNA varies from the time of transfection until the effects of RNAi cease. In order to predict siRNA efficacy accurately, the time after transfection when the amount of knockdown is measured has to be taken in to account.

7.2 GC content and AU Differential

Heale et al. [18] suggest in a recent study that GC content along with the AU differential—the difference in the number of A and U nt between the 3' end and the 5' end—are capable of predicting functionality with an accuracy of 71%. To study if the GC content and the AU differential are an important indicator of functionality, they are added as input features in addition to the Apriori patterns and the time after transfection. Adding only GC content or both GC content and AU differential to the existing features does not change accuracy. Adding only the AU differential feature improved accuracy marginally from 78% to 81%. Figure 7.2 illustrates the ROC curves obtained with and without AU differential feature. Time after transfection is included as a feature in both cases.

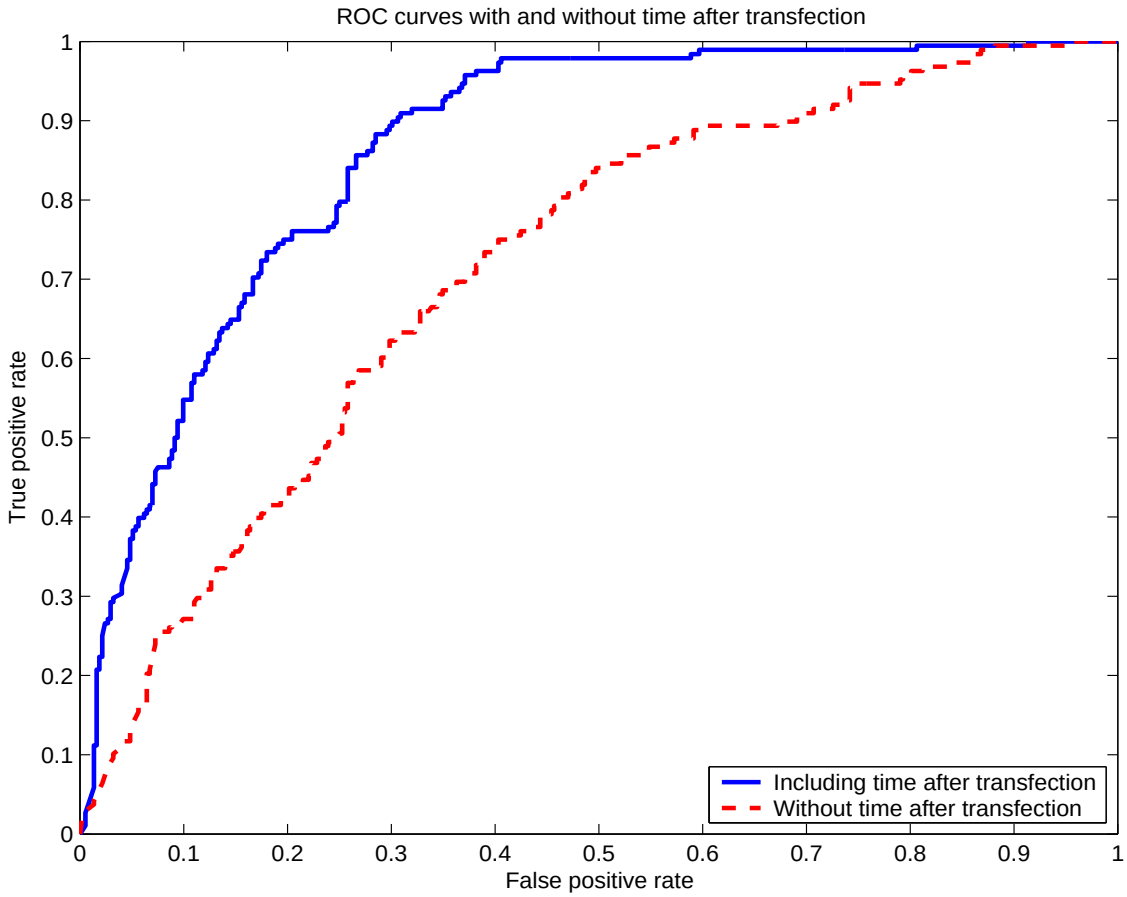

Figure 7.1: ROC Curve for the Pattern - Classifier algorithm for efficacy greater than or equal to 80% with and without the time after transfection

Since adding the AU differential feature improved accuracy, it was included as another feature to the classifier. This property is related to the stability of the siRNA strand as the binding between A and U is weaker than the binding between G and C. Addition of this feature to the Apriori features and time after transfection, gives good classification accuracy.

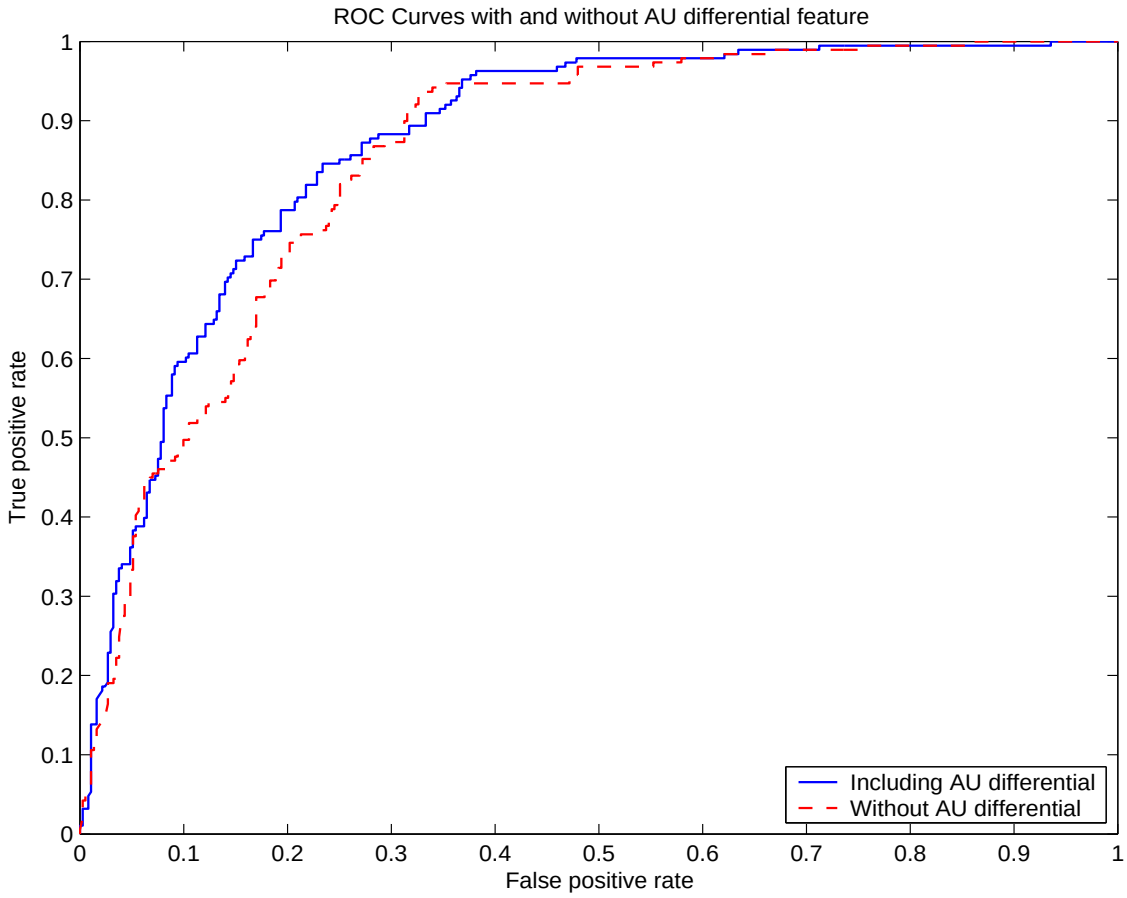

Figure 7.2: ROC Curve for the Pattern-Classifier algorithm for efficacy greater than or equal to 80% with and without A/U differential

7.3 Performance comparison

The performance of the Apriori Pattern-Classifier algorithm was compared with three other algorithms—Reynolds, Amarzguioui and GPboost. These three algorithms were chosen as they performed most consistently and had the best performance among other algorithms [45]. The Reynolds [35] algorithm was the first method to introduce rational siRNA design and has reasonable performance. The Amarzguioui [2] algorithm has good performance across different datasets [45]. The GPboost [44] method uses patterns from

siRNA data like the Apriori Pattern-Classifier but uses a genetic programming technique to evaluate patterns. It also has good performance.

The ROC curves for these algorithms are shown in Figure 7.3 and the accuracy rates and the AUC rates are shown in Table 7.1. The ROC points for GPboost algorithm were obtained from Sætrom [45]. All the four ROC curves are obtained from the same siRNA data. The Apriori pattern-classifier algorithm has the best performance among these algorithms: its ROC curve clearly dominates the curves of the other methods in all portions of the graph. This is also reflected in Table 7.1 where the AUC value of the Apriori Pattern-Classifier is higher than the rest. It has a lower false-positive curve than the rest and this implies that the number of actual non-functional siRNAs that the algorithm classifies as functional is less compared to the other algorithms. This feature is valuable in designing functional siRNAs. It also has better accuracy than any other technique, which is demonstrated in Table 7.1. The accuracy rate is not available for the GPboost algorithm as it is proprietary. This algorithm also does not require any specialized hardware as in the case of GPboost, which uses specialized hardware—a Pattern Matching Chip (PMC) [44].

Algorithm	AUC Value	Accuracy
Reynolds	0.58	62
Amarzguioui	0.55	70
GPboost	0.77	-
Apriori Pattern-Classifier	0.86	77.7

Table 7.1: Area Under ROC Curve values and Accuracy

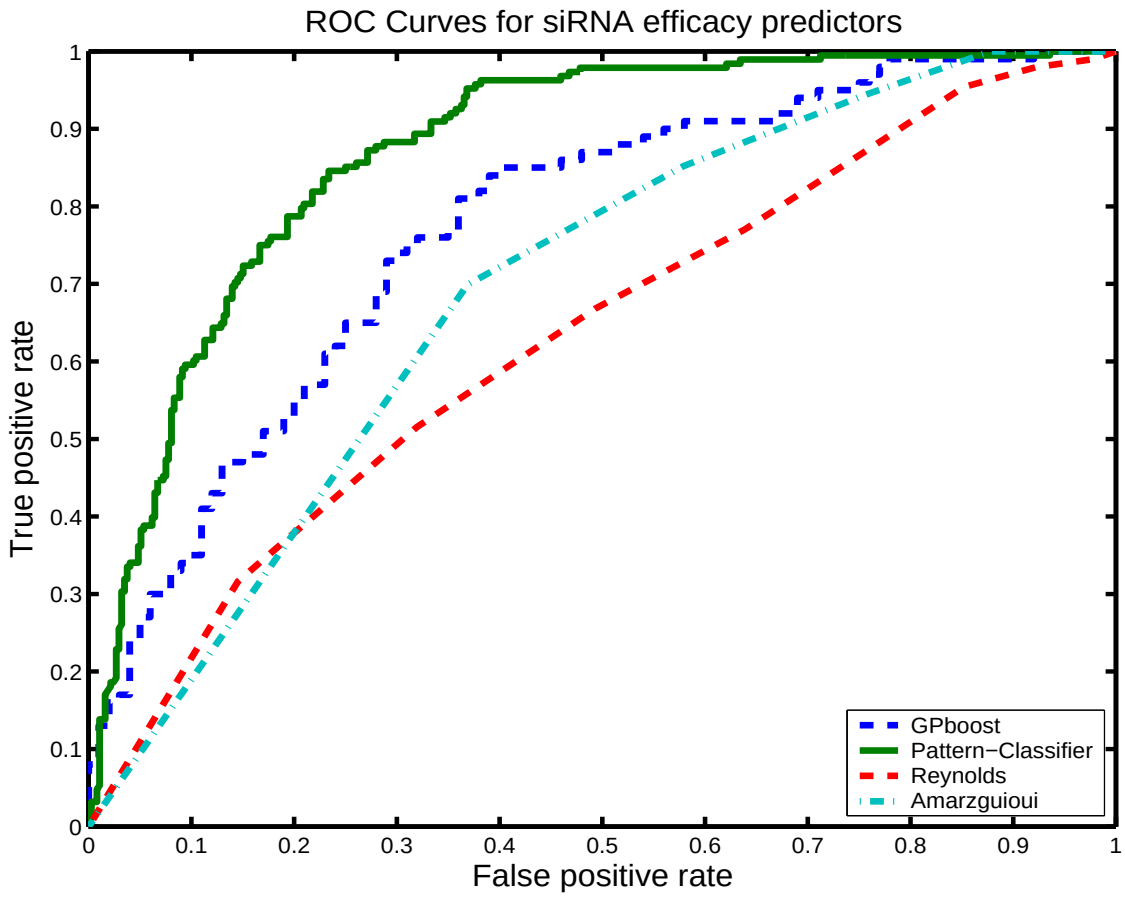

Figure 7.3: ROC Curve for the Reynolds, Amarzguioui, GPboost and the Pattern-Classifier algorithms for efficacy greater than or equal to 80%

7.4 Performance at different efficacy thresholds

The performance comparison graph in Figure 7.3 uses a threshold of 80% for determining functionality. To study the performance of the Apriori Pattern-Classifier algorithm, the resulting ROC curves are analyzed across different thresholds ranging from 10% to 90% knockdown with intervals of 10%. The resulting plot can be seen in Figure 7.4.

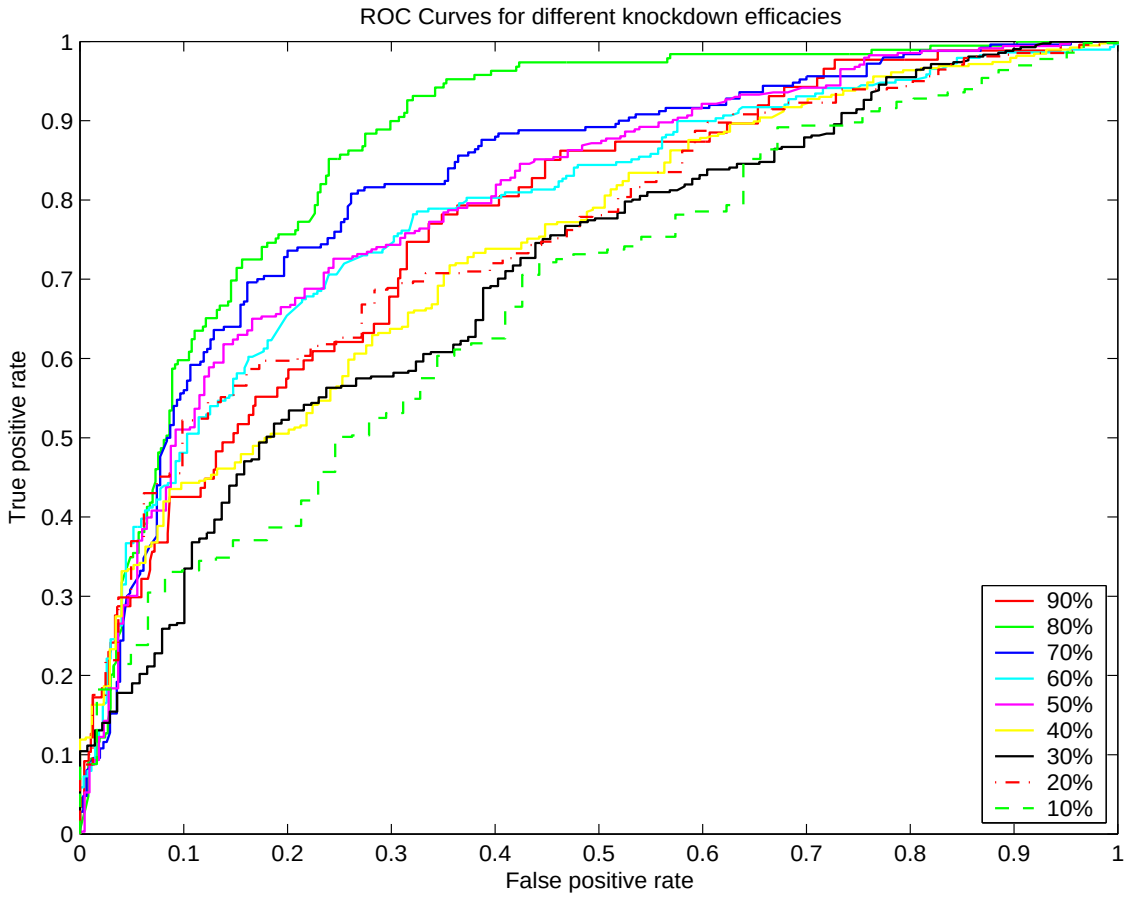

Figure 7.4: ROC Curves for the Pattern-Classifier algorithm at different efficacy thresholds

The classifiers worst performance is at predicting functionality at a knockdown threshold of 10% and it performs best at a threshold of 80%.

The reason for the classifier's bad performance at lower thresholds can be attributed to the decreasing number of training instances at these levels. At the 90% threshold, there are very few positive examples and at the 10% threshold, there are very few negative examples. At the 80% threshold, there are an almost equal number of negative and positive examples which leads the algorithm to pick features that can discriminate well between them.

7.5 Apriori Patterns

Apriori patterns	siRNA functionality	Feature	Algorithms
U - 1, U - 15	Nonfunctional	Absence of U at 1	Ui-Tei, Takasaki, Amarzguioui
A - 1, C - 18	Nonfunctional	Absence of A at 1	Ui-Tei, Takasaki, Amarzguioui
C - 7, U - 16	Nonfunctional	Presence of U at 16	Reynolds
C - 14, C - 17	Nonfunctional	None	-
G - 9, C - 13	Nonfunctional	Absence of G at 9	Takasaki
C - 6, A - 18	Nonfunctional	Absence of C at 6	Hsieh
		Presence of A at 18	Reynolds, Amarzguioui
A - 1, G - 19	Nonfunctional	Absence of A at 1	Ui-Tei, Takasaki, Amarzguioui
U - 1, A - 5	Nonfunctional	Absence of U at 1	Ui-Tei, Takasaki, Amarzguioui
G - 18, G - 19	Nonfunctional	Absence of G at 19	Hsieh, Reynolds, Ui-Tei, Takasaki, Amarzguioui
A - 1, C - 6	Nonfunctional	Absence of A at 1	Ui-Tei, Takasaki, Amarzguioui
		Absence of C at 6	Hsieh

Table 7.2: List of patterns that have high confidence. The numbers after the nucleotide indicate the position within the siRNA. The algorithms (Ui-Tei [47], Takasaki [43], Amarzguioui [2], Hseih [20], Reynolds [35]) that included the pattern as one of the rules are listed against the pattern.

Techniques currently used for efficacy prediction place high importance on positional characteristics of the siRNA—on the presence or absence of certain nucleotides at certain positions. To study if the Apriori patterns have any biological relevance, the patterns are compared with positional characteristics suggested by other efficacy studies. The Apriori patterns with the ten highest confidence values, listed below in Table 7.2, are examined. The confidence levels are computed using the bootstrap method introduced by Friedman et al. [13]. If the pattern contains any positional feature suggested by an earlier study, that

study is listed against the Apriori pattern in the table.

Table 7.2 indicates that the Apriori patterns effectively capture some of these known positional effects in addition to other characteristics. The Apriori patterns also capture higher-order interactions between these effects leading to increased prediction accuracy.

Most Apriori patterns are obtained from non-functional siRNAs. There are few patterns among the functional siRNAs that satisfy the required minimum confidence level during the Apriori extraction step. This indicates that nonfunctional siRNAs can be more easily identified than functional siRNAs.

7.6 Significance of Apriori Patterns

We analyzed the patterns resulting from the Apriori algorithm to determine if any of the patterns are more significant than the others in determining efficacy. Using the bootstrap method for computing confidence [13], we discovered the most significant patterns. Performance of the classifier was compared by excluding these significant patterns one at a time. The resulting ROC curves are seen in Figure 7.5.

Performance of the algorithm without including significant patterns (Figure 7.5) reveals that no single pattern or feature is highly significant in determining efficacy. Removing any single patterns did not result in any significant loss in performance. We can conclude from these results that no single pattern has high significance and it is a combination of these features that capture functionality.

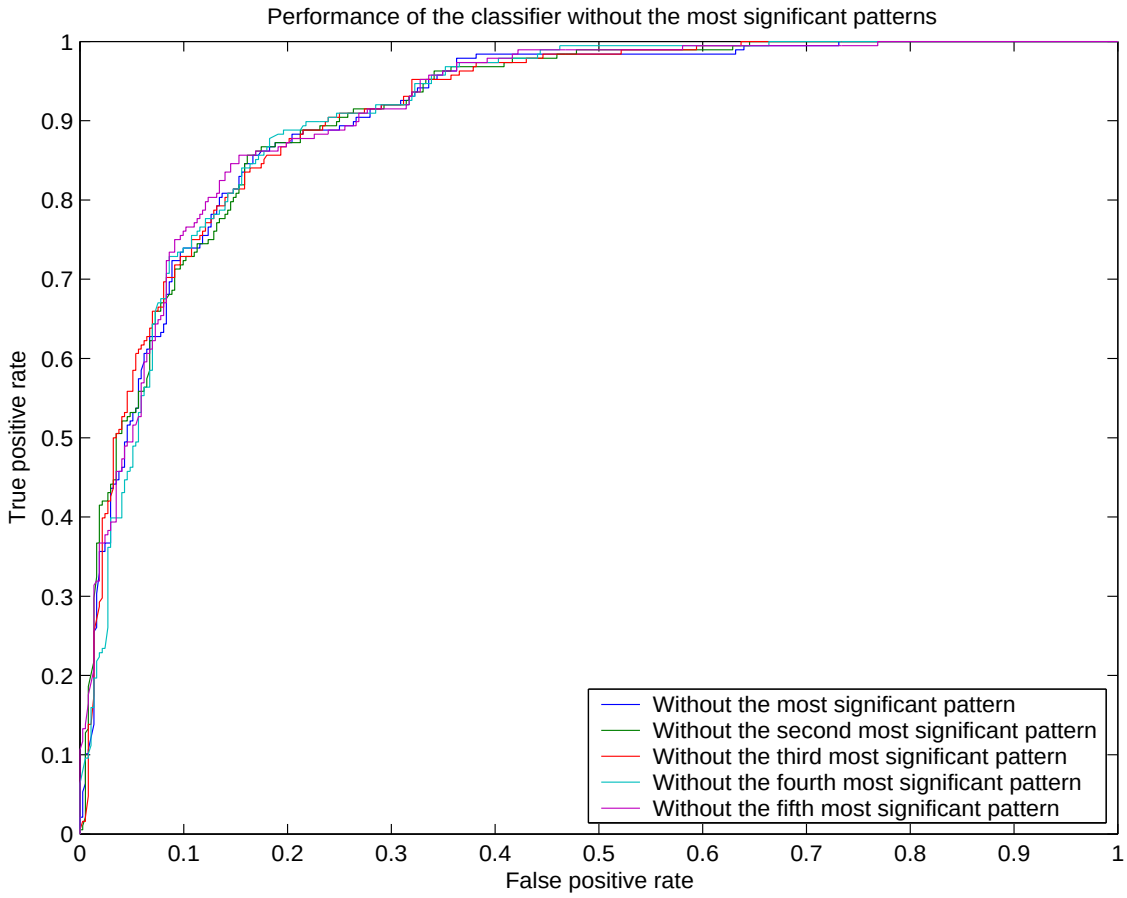

Figure 7.5: ROC Curves for the Pattern-Classifier algorithm excluding Apriori patterns with high confidence values

7.7 Random features

We see that using Apriori features gives good classification of siRNA efficacy. To test if the Apriori patterns are really significant, we compare them with random patterns from the data. The random features are selected randomly but from within the siRNA data. The reason for selecting randomly within the siRNA data is because drawing features entirely randomly results in features that are mostly inactive (not present in any of the siRNAs). DNA and RNA sequences that occur are restricted by biological constraints like a limited

number of stop codon sequences and amino acid coding sequences. This implies that only a portion of the possible combinations that can occur with the 4 nt actually does occur in nature.

The random features are extracted from the data and the time after transfection and the AU differential features are added. The classifier is trained on the input set and the results are obtained using 10 fold cross-validation. The performance of different random feature sets are shown in Figure 7.6. It shows ROC curves for 10 different runs of the algorithm, choosing 100 random features at each run.

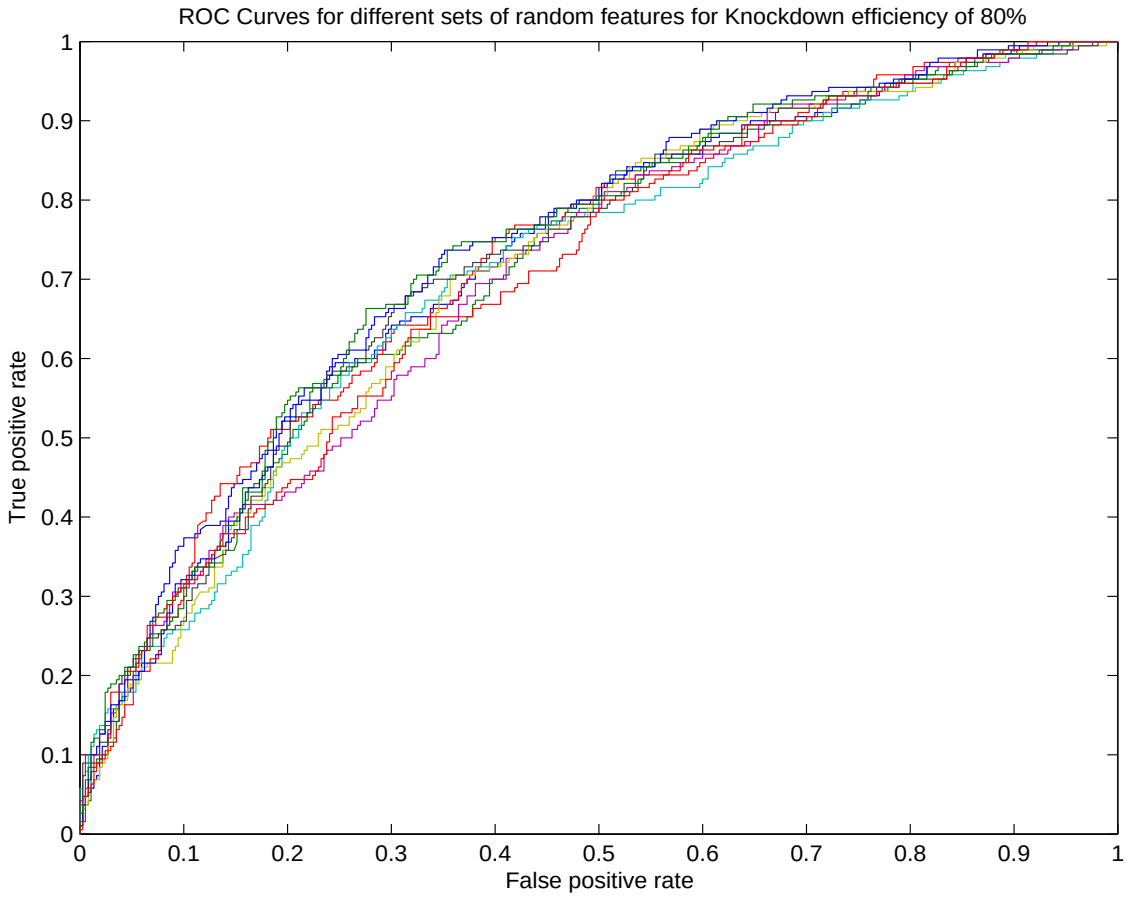

Figure 7.6: ROC Curve for different sets of 100 random features

Though the random feature sets do not perform better than the Apriori features, they have reasonable performance with an average accuracy of 75% averaged over 10 runs. It is a surprising result that random features sets can give this level of accuracy. We also observe that there is no significant difference in the performances of different random features.

7.8 Different number of random features

To study the effect of the number of random features on the performance of the algorithm, we varied the number of random features. The ROC curves for feature sets containing different number of random features ranging from 10 to 200 are compared and the resulting graph is shown in Figure 7.7

Increasing the number of random features till they reach 50 improves the algorithm's performance. After a threshold of 50, increasing the number of random features does not have any significant impact on the performance of the algorithm.

7.9 Performance of random feature sets

The performance of the Pattern-Classifier using a random feature set is compared with the Apriori patterns and with the GPboost algorithm. Time after transfection was added to the random feature set and the resulting performance was compared to the same random feature set without including time. Figure 7.8 demonstrates the results.

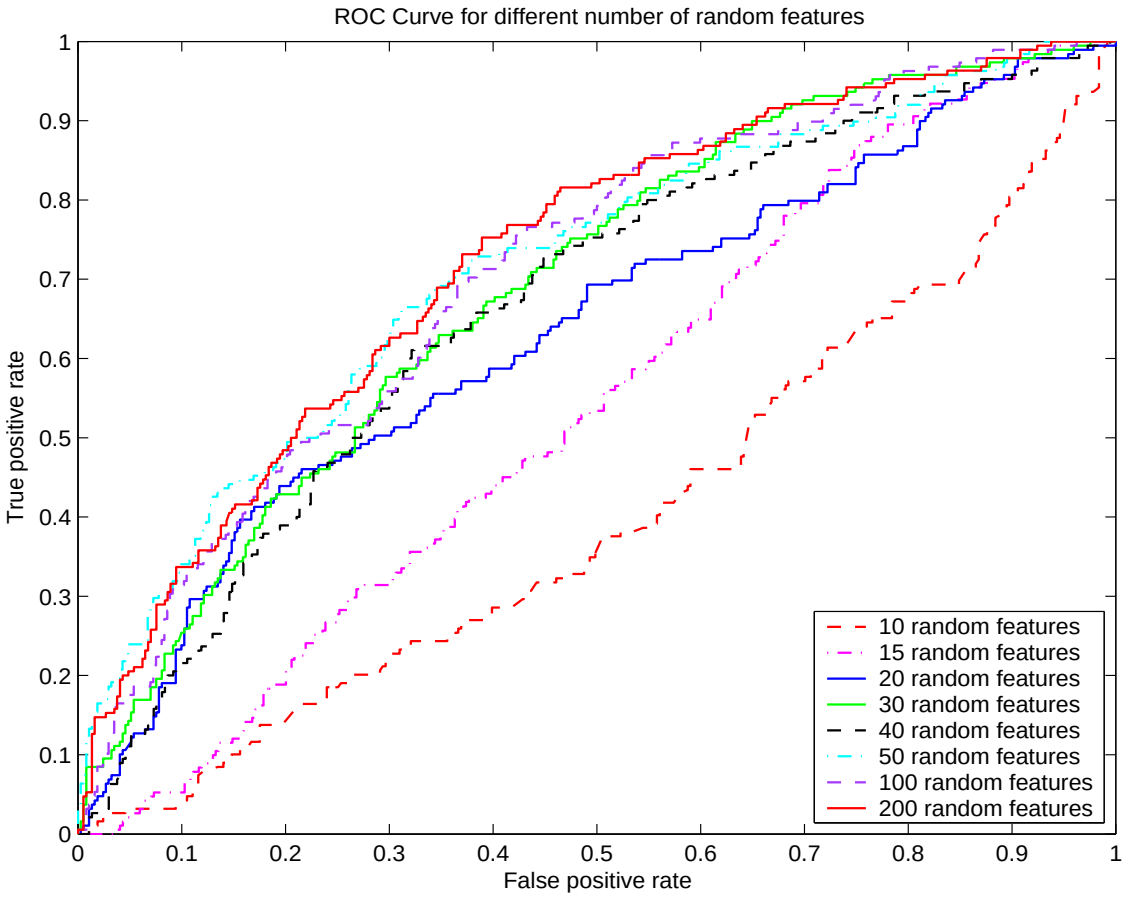

Figure 7.7: ROC Curves for different number of random features

Adding the time feature increases the performance of the random features to be better than the GPboost algorithm but it still has lower performance than the Apriori pattern-classifier algorithm. It is surprising that the random features perform as well as they do. The number of siRNAs with length 19 that can occur are 4^{19}. However, studies on the distribution of nucleotides have shown that only a percentage of this number occur in nature [36, 31] due to biological constraints like coding sequences, preserved motifs, etc. In spite of this, the number of siRNAs available in the database could just be a very small

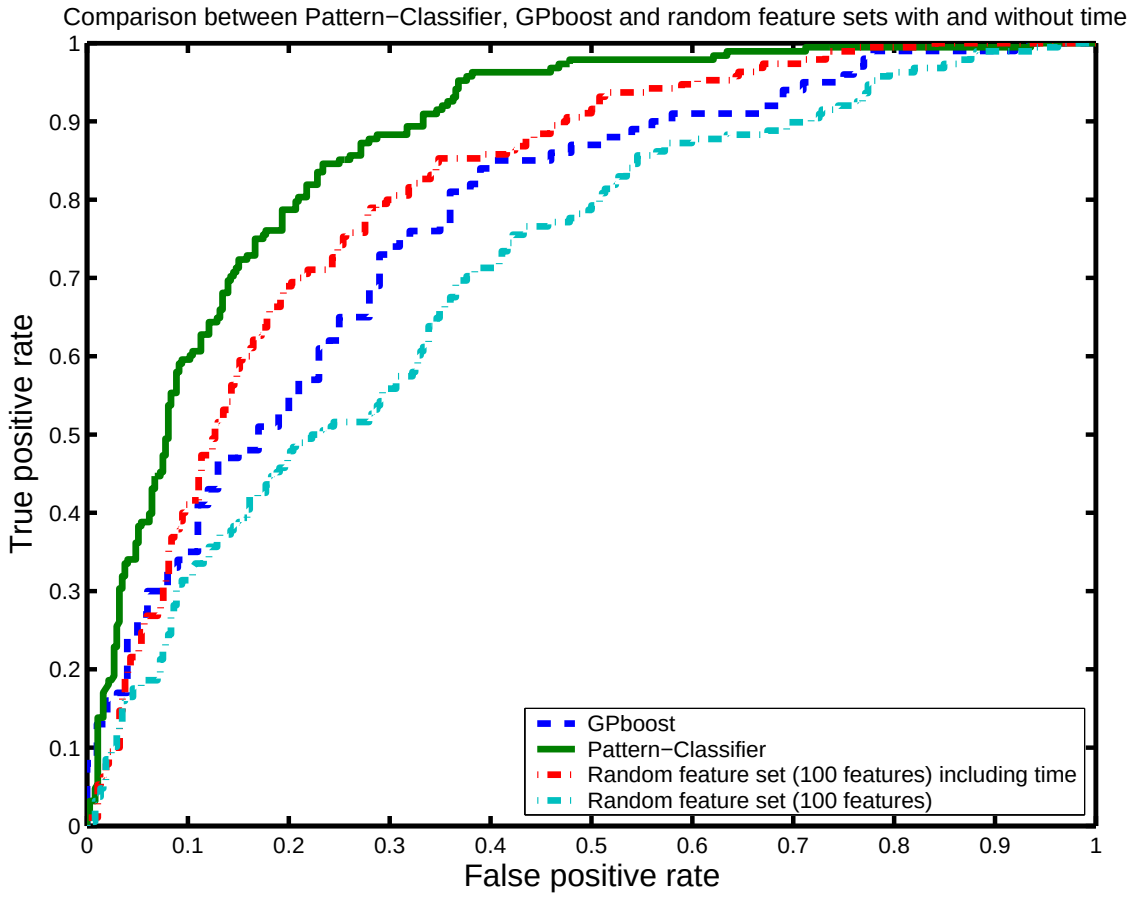

Figure 7.8: ROC Curves to compare the performance of the Pattern-Classifier, GPboost and random feature sets with and without time

sample of the entire siRNA pool. This seems to imply that the positional features that have been found significant in previous studies could be just an artifact of the small pool of siRNA data. The good performance of the Apriori algorithm could also be attributed to this effect. However, the Apriori algorithm has the best performance among all other publicly available algorithms.

7.10 Second order random patterns

Since the Apriori patterns use a combination of more than one positional feature, we examined similar patterns chosen randomly from the available data. Using sets of two features chosen randomly from the data resulted in a decrease in performance. The resulting accuracy was 67% which is much lesser than using single random features, which had an accuracy of 75%. This can be explained by the fact that the probability of two random features which are selected from an siRNA occurring in other siRNAs is very low. This results in a large number of inactive features that results in improper classification. However, the Apriori features work well because they are based on a high confidence level and high support which enables them to capture only the effective patterns accurately.

7.11 Online efficacy prediction tool

We have developed a web-based tool that uses the Apriori Pattern-Classifier algorithm to predict efficacies of siRNAs. This is available online at http://rnai.cs.unm.edu. Given the input sequences and the time after transfection when the knockdown will be measured, the tool will predict if the siRNAs will be effective or not. The knockdown threshold for functionality can also be specified to get predictions for different thresholds. Currently, different experiments use different thresholds depending on the requirement. Cut-offs for functionality range anywhere between 55% knockdown [18] to 90% knockdown [35]. The default threshold value for efficient siRNAs is mRNA degradation equal to or greater than 75%. The default time after transfection is 24 hours. The tool offers efficacy prediction at thresholds of 70%, 75%, 80% and 90%. The output of this tool is a table listing all the input siRNAs and their functionality at the chosen efficacy threshold.

Figure 7.9: Screen shot of the efficacy prediction tool available at http://rnai.cs.unm.edu

Chapter 8

Conclusions and Future Work

8.1 Conclusions

In this work, we use frequently occurring patterns in the siRNAs, in addition to time after transfection and the A/U differential feature, to predict their efficacy. Our algorithm consistently performs better than other publicly available prediction algorithms. It has lower false positive rates than the other algorithms which is desirable in RNAi studies. It does not require any specialized hardware. We show that time after transfection is an important factor in siRNA prediction—a feature that has been ignored in previous efficacy studies. We analyze the relevance of the extracted Apriori patterns and show that they are consistent with previous results. We also show that random feature sets perform quite well and demonstrate that current rules for predicting efficacy are an artifact of limited data and these rules need to be reexamined.

8.2 Future work

Improving siRNA efficacy prediction is an important aspect in effective RNAi studies. Studying the factors determining functionality helps in understanding the mechanism by which siRNAs degrade target mRNAs. This ultimately aids in our knowledge of the RNAi pathway. As more experiments are conducted, more data is becoming available though not all of it is public. A main assumption in this study is that siRNA functionality is determined solely by the sequence itself. However, there are several factors that could potentially affect efficacy though they do not have direct correlation with the siRNA sequence information. Studies have suggested that the secondary structure of the siRNA and the secondary structure of the mRNA might play a role in determining the accessibility of the siRNA to the target site [18, 27] and therefore, its functionality . Currently, techniques exist [30] that predict secondary structures that have high probabilities of occurring, given the sequence. The effect of secondary structure and stability on efficacy needs to be examined more thoroughly.

Not much is known about the mechanism of incorporation of siRNAs into RISC. Analysis of thermodynamic energy profiles of siRNA and miRNA duplexes reveal some correlation with functionality [25]. And it has been suggested that this causes selective incorporation of either the sense or antisense strand. The role of RISC in determining efficacy needs to be studied further.

Another factor that could play a role in determining the amount of knockdown is when the siRNA acts like an miRNA by causing translational degradation. When siRNAs have only partial complementarity to the target, they can still cause degradation. In designing siRNAs, the miRNA pathway needs to be considered as this has the potential to change efficacy.

References

[1] R Agrawal and R Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994.

[2] M Amarzguioui and H Prydz. An algorithm for selection of functional sirna sequences. *Biochem Biophys Res Commun.*, 316(4):1050–1058, 2004.

[3] C Bomhardt. Newsrec, a svm-driven personal recommendation system for news websites. In *IEEE/WIC/ACM International Conference on Web Intelligence*, 2004.

[4] A Borkhardt. Blocking oncogenes in malignant cells by RNA interference — new hope for a highly specific cancer treatment? *Cancer Cell*, 2(3):167–168, September 2002.

[5] BE Boser, I Guyon, and V Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learing Theory*, pages 144–152, 1992.

[6] AM Chalk, C Wahlestedt, and EL Sonnhammer. Improved and automated prediction of effective sirna. *Biochem. Biophys. Res. Commun.*, 319:264–274, 2004.

[7] CC Chang and CJ Lin. Libsvm: a library for support vector machines. 2001.

[8] C Cogoni and G Macino. Post-transcriptional gene silencing across kingdoms. *Genes Dev.*, 10:638–643, 2000.

[9] JG Doench, CP Petersen, and PA Sharp. sirnas can function as mirnas. *Genes and Development*, 17(4):438–442, 2003.

[10] SM Elbashir, W Lendeckel, and T Tuschl. RNA interfernce is mediated by 21- and 22- nucleotide RNAs. *Genes and Development*, 15:188–200, 2001.

[11] T Fawcett. Roc graphs: Notes and practical considerations for data mining researchers. *Tech Report HPL-2003-4*, 2003.

[12] A Fire, S Xu, MK Montgomery, SA Kostas, SE Driver, and CC Mello. Potent and specific genetic interference by double stranded RNA in c. elegans. *Nature*, 391:806–811, 1998.

[13] N Friedman, M Goldszmidt, and A Wyner. On the application of the bootstrap for computing confidence measures on features of induced bayesian networks, 1999.

[14] A Grishok, H Tabara, and CC Mello. Genetic requirements for inheritance of rnai in c. elegans. *Science*, 287(5462):2494–2497, 2000.

[15] S Gupta, RA Schoer, JE Egan, GJ Hannon, and V Mittal. Inducible, reversible, and stable rna interference in mammalian cells. *Proc Natl Acad Sci.*, 101(7):1927–1932, 2004.

[16] M Hamada, T Ohtsuka, R Kawaida, M Koizumi, K Morita, H Furukawa, T Imanishi, M Miyagishi, and K Taira. Effects on rna interference in gene expression (rnai) in cultured mammalian cells of mismatches and the introduction of chemical modifications at the 3'-ends of sirnas. *Antisense Nucleic Acid Drug Dev.*, 12:301–309, 2002.

[17] J Harborth, SM Elbashir, K Vandenburgh, H Manninga, SA Scaringe, K Weber, and T Tuschl. Sequence, chemical, and structural variation of small interfering rnas and short hairpin rnas and the effect on mammalian gene silencing. *Antisense Nucleic Acid Drug Dev.*, 13(2):83–105, 2003.

[18] BSE Heale, HS Soifer, C Bowers, and JJ Rossi. sirna target site secondary structure predictions using local stable substructures. *Nucleic Acids Research*, 33(3), 2005.

[19] T Holen, M Amarzguioui, MT Wiiger, E Babaie, and H Prydz. Positional effects of short interfering rnas targeting the human coagulation trigger tissue factor. *Nucleic Acids Res.*, 30(8):1757–1766, 2002.

[20] AC Hsieh, R Bo, J Manola, F Vazquez, O Bare, A Khvorova, S Scaringe, and WR Sellers. A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens. *Nucleic Acids Research*, 32(3):893–901, February 2004.

[21] G Hutvagner and PD Zamore. A microrna in a multiple-turnover rnai enzyme complex. *Science*, 297:2056–2060, 2002.

[22] AL Jackson, SR Bartz, J Schelter, SV Kobayashi, J Burchard, M Mao, B Li, G Cavet, and PS Linsley. Expression profiling reveals off-target gene regulation by rnai. *Nature Biotechnology*, 21:635–637, 2003.

[23] T Joachims. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[24] RA Jorgensen, PD Cluster, J English, Q Que, and CA Napoli. Chalcone synthase cosuppression phenotypes in petunia flowers: comparison of sense vs. antisense constructs and single-copy vs. complex t-dna sequences. *Plant Mol. Biol.*, 31:957–973, 1996.

[25] A Khvorova, A Reynolds, and SD Jayasena. Functional sirnas and mirnas exhibit strand bias. *Cell*, 115:209–216, 2003.

[26] V Kumar and T Poggio. Learning-based approach to realtime tracking analysis of faces, 1998.

[27] KQ Luo and DC Chang. The gene-silensing efficacy of siRNA is strongly dependent on the local structure of mRNA at the targeted region. *Biochemical and Biophysical Research Communications*, 318:303–310, 2004.

[28] J Martinez, A Patkaniowska, H Urlaub, R Luhrmann, and T Tuschl. Single-stranded antisense sirnas guide target rna cleavage in rnai. *Cell*, 110(5):563–574, 2002.

[29] MA Martinez, A Gutierrez, M Armand-Ugon, J Blanco, M Parera, J Gomez, B Clotet, and JA Este. Suppression of chemokine receptor expression by rna interference allows for inhibition of hiv-1 replication. *AIDS*, 16(18):2385–2390, 2002.

[30] DH Mathews, J Sabina, M Zuker, and DH Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure. *Journal of Molecular Biology*, 288(5):911–940, 1999.

[31] R Nussinov. Doublet frequencies in evolutionary distinct groups. *Nucleic Acids Res.*, 12(3):1749–1763, 1984.

[32] F Provost, T Fawcett, and R Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proc. 15th International Conf. on Machine Learning*, pages 445–453. Morgan Kaufmann, San Francisco, CA, 1998.

[33] S Qiu, CM Adema, and T Lane. A computational study of off-target effects of rna interference. *Nucleic Acids Res.*, 33(6):1834–1847, 2005.

[34] F Ratcliff, BD Harrison, and DC Baulcombe. A similarity between viral defense and gene silencing in plants. *Science*, 276(5318):1558–1560, 1997.

[35] A Reynolds, D Leake, Q Boese, S Scaring, W Marshall, and A Khvorova. Rational siRNA design for RNA interference. *Nature Biotechnology*, 22(3):326–330, 2004.

[36] AC Rogerson. There appear to be conserved constraints on the distribution of nucleotide sequences in cellular genomes. *J Mol Evol.*, 32(1):24–30, 1991.

[37] N Romano and G Macino. Quelling: transient inactivation of gene expression in neurospora crassa by transformation with homologous sequences. *Mol. Microbiol.*, 6:3343–3353, 1992.

[38] S Saxena, ZO Jonsson, and A Dutta. Small rnas with imperfect match to endogenous mrna repress translation. implications for off-target activity of small inhibitory rna in mammalian cells. *Journal of Biological Chemistry*, 278(45):44312–44319, 2003.

[39] DS Schwarz, G Hutvagner, T Du, Z Xu, N Aronin, and PD Zamore. Asymmetry in the assembly of the rnai enzyme complex. *Cell*, 115:199–208, 2003.

[40] Bernhard Schlkopf, Isabelle Guyon, and Jason Weston. Statistical learning and kernel methods in bioinformatics.

[41] H Shi, A Djikeng, T Mark, E Wirtz, C Tschudi, and E Ullu. Genetic interference in trypanosoma brucei by heritable and inducible double-stranded rna. *RNA*, 6(7):1069–1076, 2000.

[42] J Soutschek, A Akinc, B Bramlage, K Charisse, R Constien, M Donoghue, S Elbashir, A Geick, P Hadwiger, J Harborth, M John, V Kesavan, G Lavine, RK Pandey, T Racie, KG Rajeev, I Rohl, I Toudjarska, G Wang, S Wuschko, D Bumcrot, V Koteliansky, S Limmer, M Manoharan, and HP Vornlocher. Therapeutic silencing of an endogenous gene by systemic administration of modified sirnas. *Nature*, 432:173–178, 2004.

[43] S Takasaki, S Kotani, and A Konagaya. An effective method for selecting sirna target sequences in mammalian cells. *Cell Cycle*, 3(6):790–795, 2004.

[44] P. Sætrom. Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming. *Bioinformatics*, 20(9):3641–3650, 2004.

[45] P Sætrom and O Snove Jr. A comparison of siRNA efficacy predictors. *Biochemical and Biophysical REsearch Communications*, 321:247–253, 2004.

[46] T Tuschl. RNA interference and small interfering RNAs. *Chembiochem*, 2(4):239–245, 2001.

[47] K Ui-Tei, Y Naito, F Takahashi, T Haraguchi, H Ohki-Hamazaki, A Juni, R Ueda, and K Saigo. Guidelines for the selection of highly effective sirna sequences for mammalian and chick rna interference. *Nucleic Acids Res.*, 32:936–948, 2004.

[48] VN Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

[49] TA Vickers, S Koo, CF Bennett, ST Crooke, NM Dean, and BF Baker. Efficient reduction of target rnas by small interfering rna and rnase h-dependent antisense agents. a comparative analysis. *J. Biol. Chem.*, 278(9):7108–7118, 2003.

[50] IH Witten and E Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 2000.

[51] H Xia, Q Mao, SL Eliason, SQ Harper, IH Martins, HT Orr, HL Paulson, L Yang, RM Kotin, and BL Davidson. Rnai suppresses polyglutamine-induced neurodegeneration in a model of spinocerebellar ataxia. *Nature Medicine*, 10:816–820, 2004.

Glossary

gene expression	The process by which the information encoded in the gene's DNA is converted into protein.
mRNA	Messenger or mRNA is a copy of the information carried by a gene on the DNA. The role of mRNA is to move the information contained in DNA to the translation machinery.
nucleotide	A subunit of DNA or RNA containing one of the following bases - adenine (A), guanine (G), cytosine (C) and uracil (U) (in RNA) or thymine (T) (in DNA).
RNAi	RNA interference is the biological process by which mRNA is destroyed or degraded when exposed to complementary siRNA sequences.
siRNA	small interfering RNAs, typically 21-23 nucleotides in length.
transcription	The synthesis of a RNA sequence from a DNA sequence.
transfection	The process of introducing foreign DNA in to a cell.
translation	The process by which the genetic code in the mRNA directs the production of proteins.