Correlation Between Computed Equilibrium Secondary Structure Free Energy and siRNA Efficiency

Puranjoy Bhattacharjee

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Alexey V. Onufriev
Naren Ramakrishnan
Lenwood S. Heath

08/06/2009
Blacksburg, Virginia

Keywords: RNA interference(RNAi), RNAi efficiency, RNA secondary structure, RNAi equilibrium thermodynamics, Support Vector Machine

Correlation Between Computed Equilibrium Secondary Structure Free Energy and siRNA Efficiency

Puranjoy Bhattachrjee

Abstract

We have explored correlations between the measured efficiency of the RNAi process and several computed signatures that characterize equilibrium secondary structure of the participating mRNA, siRNA, and their complexes. A previously published data set of 609 experimental points was used for the analysis. While virtually no correlation with the computed structural signatures are observed for individual data points, several clear trends emerge when the data is averaged over 10 bins of $N \sim 60$ data points per bin.

The strongest trend is a positive linear ($r^2 = 0.87$) correlation between ln(remaining mRNA) and ΔG_{ms}, the combined free energy cost of unraveling the siRNA and creating the break in the mRNA secondary structure at the complementary target strand region. At the same time, the free energy change ΔG_{total} of the entire process $mRNA + siRNA \rightarrow (mRNA - siRNA)_{complex}$ is not correlated with RNAi efficiency, even after averaging. These general findings appear to be robust to details of the computational protocols. The correlation between computed ΔG_{ms} and experimentally observed RNAi efficiency can be used to enhance the ability of a machine learning algorithm based on a support vector machine (SVM) to predict effective siRNA sequences for a given target mRNA. Specifically, we observe modest, 3 to 7%, but consistent improvement in the positive predictive value (PPV) when the SVM training set is pre- or post-filtered according to a ΔG_{ms} threshold.

Dedication

This work is dedicated to my family, without whose support and encouragement I wouldn't be where I am today.

Acknowledgments

I would like to thank Dr. Alexey Onufriev, my advisor, for his advice and encouragement. His patient handling of a few curveballs my first attempt at research threw at me is invaluable, and something I would like to remember and emulate in my career. I would also like to thank Dr. Naren Ramakrishnan for his advice. This thesis would have been substantially poorer without his inputs. He has been a great mentor and always there to talk about any issue I may have faced. I am grateful to Dr. Lenwood Heath. With a keen eye on the progress of this research and a smile to go along with the advice and encouragement that he has been generous with, he was a great influence in his role as the taskmaster for a greater part of the project. I would also like to thank Dr. Ruth Green who taught a bumbling computer science student the basics of biology and never stopped encouraging during the process. I thank Mr. Ramu Anandakrishnan, Mr. Andrew Fenley, Mr. Charles Baker, and Mr. Mark Lawson for their help and support on numerous occasions, which helped smooth over a number of bumps on the road of graduate research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The RNAi phenomenon

Since the discovery of RNA interference (RNAi) in the nematode worm *Caenorhabditis elegans*,[1] there has been tremendous interest in its mechanism. RNA interference is induced by double stranded RNA (dsRNA).[2] In cells, the dsRNA is cut into 19-23 nucleotide long pieces by Dicer, a ribonuclease-like enzyme.[3] These pieces are called short interfering RNAs (siRNAs). RNA Induced Silencing Complex (RISC) takes the antisense strand of the siRNA, which hybridizes with the complementary sequence in the target mRNA.[4] Upon addition of ATP, the complex is activated and the region complementary to the siRNA is cleaved, thus causing the gene knockdown. Researchers have found numerous uses of RNAi. For example, it is being used to understand the signalling pathways in mammalian cell systems.[5] Similarly, microRNAs and siRNAs have been used to silence genes in plants.[6] Fig. 1.1 shows the process.

RNAi has numerous uses, yet one of its serious drawbacks is that not all siRNAs work equally well at gene silencing. Different siRNAs, complementary to different regions in the same mRNA, can have drastically different silencing efficiencies. This has led to intense research on the RNAi mechanism, with a view to designing better and more effective siRNAs. Some designers focussed on sequence characteristics of the siRNA, e.g, absence of the nucleotide G at position 13. Sequence-centric designs such as these do not account for the possible influence of protein binding, mRNA target region accessibility, structure of the siRNA, and other parameters. The design of siRNAs will improve as the role of these features are understood better. Conversely, demonstrable influence of these features on RNA interference can provide valuable hints regarding the RNAi machinery and improve our understanding of the same.

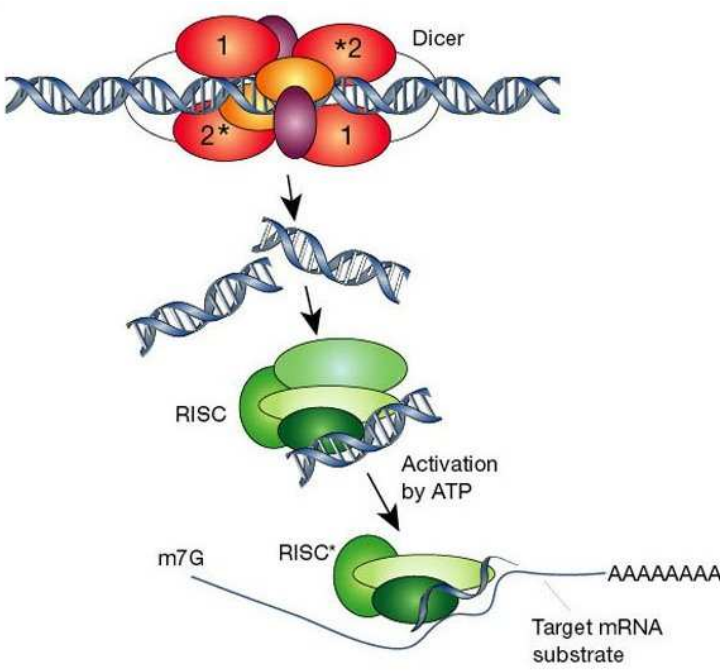


Figure 1.1: General schematics of RNA interference. Dicer cleaves dsRNA into 19-23 nucleotide long siRNAs, which are loaded onto RISC. The siRNA antisense strand then binds to the target mRNA which is subsequently degraded. Adapted by permission from Macmillan Publishers Ltd: Nature Insight Review. Gregory J. Hannon, "RNA interference", Nature, Volume 418, July, 2002. Pages 244–251.[4]

1.2 Motivation

Direct experiments on a short fragment of a single mRNA[7] demonstrated that the efficiency of the RISC complex (RNAi efficiency) may depend critically on the accessibility of the mRNA target region for binding of the antisense siRNA. The mechanistic explanation for the observed dependence was very appealing in its simplicity: if and only if the secondary structure of the mRNA is such that the target region for the siRNA is "open", the siRNA can readily bind leading to successful cleavage of the target mRNA. Later, this picture received further support from Schubert et. al.,[8] who found a high correlation between free energies of local mRNA target structures and silencing efficiency for a set of 9 specially designed mRNAs ranging in length from 955 to 984 nucleotides. However, the high direct correlation between the accessibility of the target region and siRNA efifciency observed earlier for several specific targets was not observed in a systematic study by Lu and Mathews,[9] who used 3084 siRNA-mRNA pairs for the analysis using hybridization thermodynamics. In fact, none of the carefully chosen thermodynamic signatures, such as free energy of formation of the siRNA-mRNA complex or free energy of disruption of the target mRNA region, were found to correlate appreciably with the RNAi efficiency: all such correlations were very weak. This

raised the question: Why have secondary structure signatures of RNAi reaction components coupled via equilibrium thermodynamics not been able to predict the silencing efficiencies of siRNAs? Is it because of deficiencies in the thermodynamic modelling of the RNAi reaction? Or is there a "biological" reason to it, e.g., infusion of energy (an unknown amount of) from ATP hyrdrolysis that throws the thermodynamic model off? Is there a newer way to use "physics" to analyze RNA interference? These are the questions we will investigate in this work. The very thermodynamic signatures that we are interested in are shown in Fig. 1.2.

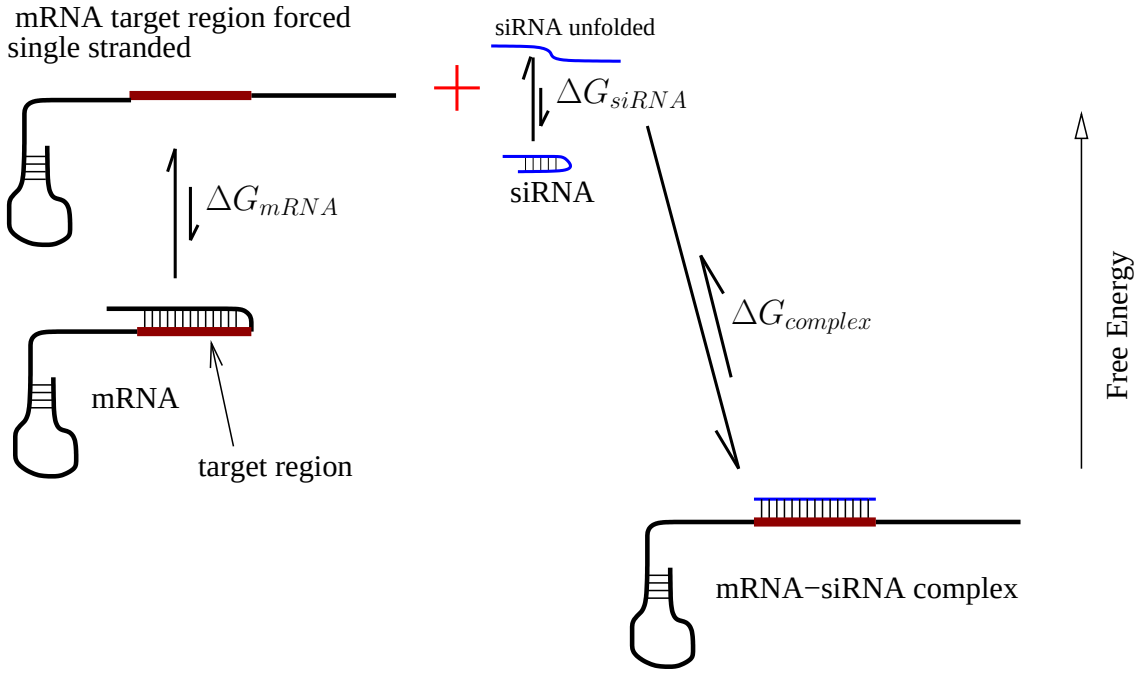


Figure 1.2: Energy diagram of formation of mRNA-siRNA complex. The complex is formed from the constituent mRNA and the siRNA. Here, ΔG_{mRNA} is the free energy required to create a break in the mRNA secondary structure at the target strand region. ΔG_{siRNA} is the free energy of formation of single-stranded siRNA. $\Delta G_{complex}$ is the free energy of formation of the *siRNA-mRNA* complex. Computational details are in the "Methods" section.

1.3 Background

Effective utilization of RNA interference hinges on the ability to select effective siRNAs for a given mRNA target sequence. This involves two related computational challenges: a) given an mRNA sequence, being able to choose the best or a amall set of most efficient siRNAs, and b) given an mRNA and an siRNA, being able to predict the silencing activity of the mRNA-siRNA pair. Various factors are known to influence the efficiency of an siRNA.

Some researchers have focussed on the siRNA sequence and its secondary structure. Patzel et. al.[10] found that siRNAs that produce unstructured guide-RNAs improve interference efficiency. Holen et. al.[11] looked at the effect of mutations in siRNAs on RNAi efficiency and found that wobble mutations at the ends lead to lesser degradation of RNAi efficiency than mutations in the central part of the antisense strand. Liao et. al.[12] studied the effectiveness of siRNAs with high G/C contents and suggested that the accessibility to target sequences influences the efficiency of siRNA. Ameres et. al.[13] found that the $5'$ part of the siRNA effects the stability of the association between RISC and the target mRNA. Heilersig et. al.[14] showed that size and sequence of inverted repeats sequences had an effect on silencing efficiency. Ichihara et. al.[15] showed that the siRNA duplex must be unstable for better silencing activity.

Other researches have looked at the effects of the target mRNA in the process. Schubert et. al.[8] studied correlation between the local RNA target structure and silencing efficiency. Target RNA accessibility has been found to be important.[16,17] Westerhout and Berkhout[18] studied the impact of target mRNA structures and found an accessible 3'-end to be a very important factor for RNAi-mediated inhibition. Gredell et. al.[19] found that mRNA target regions that were unpaired at the 5'-end or the 3'-end were silenced more strongly than target regions unpaired in the center or those paired throughout the target strand. Shao et. al.[20] suggested the effects of target structure on RISC assembly and target recognition. However, there has been a debate whether to focus on the sequence or the structure of the target mRNA strand and the siRNA.[21,22] Russell et. al.[23] have found that silencing could be effected by temperature. Lu and Mathews[9] account for the differences in equilibrium consideration when designing siRNAs and antisense oligodeoxynucleotides.

The ability to investigate the influence of RNA secondary structure on the outcomes of RNAi depend critically on the availability of accuracte secondary structure prediction methods. There are various methods to determine the structure of RNAs and their associated free energies. Mfold[24] and Vienna[25] determine optimal secondary structure by searching for low free energy conformations, the free energy depends on various thermodynamic and auxiliary parameters.[26,27] MC-Fold and MC-Sym[28] use sequence data to infer RNA structure. Harmanci et. al.[29] use probabilistic alignment constraints in their Dynalign code. There are web servers such as RNAbor[30] and RNA2D3D[31] that utilize sequence and secondary structure characteristics to provide more information. RNAbor computes statistics related to δ-neighbors which can be used to study structural neighbors of intermediate, biologically active structures among other things. RNA2D3D computes first-order approximation of a 3-dimansional conformation consistent with sequence and secondary structure information.

With the availability of these tools, researchers tried to predict the sequences of siRNAs suitable for a particular mRNA. Jiang et. al.[32] used a random forest regression model along with database searching to design siRNAs. There are web servers like RNA-Workbench[33] and OligoWalk[34] that give a selection of efficient siRNAs. Gong et. al.[35] surveyed the features associated with high RNAi effectiveness and suggest a set of design rules. Reynolds et. al.[36] proposed an algorithm incorporating eight characteristics associated with siRNA

functionality, as did Amarzguioui and Prydz.[37] Linear models that combine the basic features of siRNA sequences for siRNA efficiency prediction have been proposed.[38] Tools using support vector machines have been used to account for thermodynamic, accessibility, and dinucleotide factors.[39,40]

1.4 Roadmap of the study

In Chapter 2, we present our highly simplified model of the RNAi reaction based on equilibrium thermodynamics. We discuss how the secondary structures of the mRNA and the siRNA would effect the reaction, and what conclusions can be drawn from them. We also discuss support vector machine (SVM) and how they fit in our study.

In Chapter 3, we describe the datasets used for our experiments. We also describe the calculations of free energies in more detail and present our statistical analysis. siRNAs can be classified as functionally efficient and functionally inefficient based on some efficiency threshold. We describe the feature space–the set of attributes related to siRNA and mRNA that determine the efficiency of the siRNA–and the performance metrics for the SVM.

We present our findings and conclusions in Chapter 4 regarding thermodynamic free energy signatures and RNAi efficiency. We perform additional calculations to test the robustness of our model and to determine if the conclusions are robust to computational details. One of the key constraints we have is the length of the mRNA region relevant in our model. We describe our analysis at determining if the conclusions are related to it. The dataset we use is collected from different experiments described in the literature, and we further analyze whether differences in experimental protocol could possibly influence the result or not. Finally, we present our efforts at determining if using minimum free energy conformations for mRNA and siRNA secondary structure as opposed to a combination of all the possible conformations have any influence on the results.

We also describe our computations related to SVM. We divide the dataset into two subsets in different ratios and use the sets to train and test the SVM. Based on the results of these calculations, we choose a suitable ratio for training set and test set and perform further analysis using different combinations of features in an attempt to improve the SVM performance. Finally, we attempt to incorporate our knowledge of the significance of thermodynamic free energies in addition to including them in the feature space to further improve the predictive performance of the SVM, by pre-filtering the training set and post-filtering the test set.

We conclude with a summary of our findings in Chapter 5.

Chapter 2

The Approach

2.1 Model Based on Equilibrim Thermodynamics

Various studies have implicated target site accessibility in determining RNAi efficiency. Target site accessibility is a key factor in determining if the siRNA can attach to the mRNA. In addition, the secondary structure of the siRNA itself may be a factor in determining siRNA efficiency. The logic is that if the siRNA has a very stable secondary structure in the equilibrium, it will require more energy to unravel it, and hence the probability of the siRNA to attach to the target site will decrease.

One way to quantify the accessibility of the target site of the mRNA and the ease of unraveling of siRNA is to look at their thermodynamic Gibbs free energies of formation compared to that of their respective native states. Fig 1.2 shows the reactions that are involved in the formation of the mRNA-siRNA complex free in solution. Here we neglect the possible formation of mRNA-mRNA and siRNA-siRNA dimers.

As shown in Fig. 1.2, the mRNA and siRNA are initially in their folded states in the solution. For the reaction to progress, the secondary structure of the target site of the mRNA has to break. The free energy cost for this is ΔG_{mRNA}. The siRNA also has to unravel before the siRNA can participate in the reaction. The free energy change involved in this is ΔG_{siRNA}. The mRNA-siRNA complex formation results in lowering of the free energy and the free energy of formation of the complex is $\Delta G_{complex}$ [For computational details, see Chapter 3].

Within our model (Fig. 1.2), the overall free energy change involved in the formation of the mRNA–siRNA complex is thus,

$$\Delta G_{total} = \Delta G_{complex} - \Delta G_{mRNA} - \Delta G_{siRNA} \tag{2.1}$$

According to statistical physics, the relative probability of formation of the mRNA-siRNA

complex

$$P \quad = \quad Ae^{-\frac{\Delta G_{total}}{kT}} \tag{2.2}$$

where k is the Boltzmann constant and T is the absolute temperature.

We assume that the more the siRNA-mRNA complex is formed, the more of mRNA is degraded. Suppose now that the RNAi efficiency is proportional to P. Then, the more negative is the sum of ΔG_{mRNA} and ΔG_{siRNA}, the more the formation of the mRNA-siRNA complex will be hindered; on the other hand, a more negative $\Delta G_{complex}$ favors the formation of the complex, and thus presumably favors the RNAi efficiency. A more negative ΔG_{total} would then indicate a higher RNAi efficiency.

Taking natural logarithm on both sides of Equation 2.2, we get

$$ln(P) \quad = \quad -\frac{\Delta G_{total}}{kT} + ln(A) \tag{2.3}$$

The more of mRNA is degraded, the less is the remaining level of mRNA in the experiment.

$$ln(remaining\, mRNA) \quad = \quad -\frac{\Delta G_{total}}{kT} + ln(A) \tag{2.4}$$

With $ln(A)$ and kT=0.59 kcal/mol being constant, we expect to see a linear correlation trend in a plot of $ln(remaining\, mRNA)$ vs ΔG.

This expectation is based on at least two critical assumptions: 1) The thermodynamic equilibrium is reached during the RNAi process. 2) The oversimplified "free in solution" diagram in Fig. 1.2 is at least partially relevant to the *in vivo* RNAi reaction shown in Fig. 1.1.

2.2 Support Vector Machine

Support vector machine (SVM) is a classification technique in machine learning with roots in statistical learning theory.[41] They construct the decision bounday using a subset of the training set, called the **support vectors**. SVM is particularly useful in binary classification in high-dimensional space, and thus fits our problem because we have a number of siRNA and mRNA attributes that contribute to determining the RNAi efficiency for an siRNA and a target mRNA.

An SVM implementation usually performs as a linear classifier for a separable case where it constructs a maximum-margin hyperplane as a decision boundary to cleanly separate the two classes of data points by optimizing for the margin between the decision boundary and the nearest points in the two classes. In non-separable cases, it may be possible to construct

a hyperplane that cleanly separates the classes but has a low-margin, called the problem of over-fitting. To avoid over-fitting and thus reducing the margin for classification error, SVMs can be extended by incorporating `slack variables`[42] into the constraints of the separable maximum-margin optimization problem. It can also be extended to a non-linear classifier by mapping the original attributes in a non-linear separation to a different set of transformed attributes so that the data points are separated by a hyperplane in the transformed attribute space. SVMs use a `kernel trick`[43] to perform calculations in the original attribute space that would be computationally expensive in the transformed attribute space.

Chapter 3

Methods

3.1 Experimental Dataset Used

For our computational experiments, we use a slightly smaller version of the original dataset Shabalina et. al[44] that has been used by others in the field for similar purposes.[45] The original dataset collects 653 results from RNAi experiments reported in literature. The mRNAs reported belong to *Homo sapiens* (human), *Mus musculus* (house mouse), *Streptomyces alboniger* (bacteria), and artificial mRNA sequences. We exclude 44 data points from Harborth et.al.[46] Some of the siRNAs in the Harborth et.al.[46] dataset are used to target more than one mRNA. Also, the siRNA concentrations reported by Harborth et.al. are much lower than those used in the rest of the Shabalina dataset (private communication with Dr. Thomas Tuschl, one of the authors of the paper.). To avoid possible irregularities, we restrict our dataset to the remaining 609 data points of the original Shabalina dataset.

In the subset that we use, the mRNAs vary in size from 556 nucleotides to 11242 nucleotides. They exhibit a range of RNAi activity, with the efficiency ranging from 0% to 100%. Table 3.1 shows the length distribution of the mRNAs.

RNAi efficiency is represented in the dataset as the percentage of mRNA remaining after the interference reaction has taken place. In the dataset, some efficiency values are at 0%, which were set to the lowest non-zero activity value of 0.06% found in the set. Similarly, at the other end of the spectrum of RNAi activity, values which are greater than 100% are set to 100%.

If an mRNA mentioned in the dataset has since been replaced by updated versions in GenBank (http://www.ncbi.nlm.nih.gov/Genbank/), we use the updated version if it retains the target regions complementary to the original siRNAs that have been used in the original reported experiment. All the siRNAs used in the experiments in the dataset are 19 nt long.

mRNA Length	Number of mRNA	Number of siRNA-mRNA pair
0-1000	7	132
1000-2000	7	71
2000-3000	13	126
3000-4000	10	73
4000-5000	2	13
5000-6000	5	143
7000-8000	3	15
8000-9000	2	10
11000-12000	2	26

Table 3.1: Length distribution of mRNAs in the dataset

3.2 Prediction of Secondary Structure and Free Energy

In this section, we describe how the components of free energy in the simplistic thermodynamic model of RNAi reaction as shown in Fig. 1.2 are calculated. We use `Mfold`,[24] a tool based on dynamic programming that predicts equilibrium secondary structures and related free energies of nucleic acids. It produces multiple possible secondary structures and associated free energies. For details about the energies considered for our calculations, see Section 3.3. The various default parameters are

LC–sequence type (default linear)

T–temperature (default 37 deg C, the normal human body temperature)

P–percent of suboptimality to consider for suboptimal structures (default 5)

NA_CONC–Na+ molar concentration (default 1.0)

MG_CONC–Mg++ molar concentration (default 0.0)

W–window parameter (default - set by sequence length, 2 for sequence length less than 100nt, 5 for 200nt, 15 for 800nt, 25 for 8000nt). Mfold calculates more structures with similar energies for a smaller window, and fewer structures with different energies for a larger window.

MAXBP–max base pair distance (default - no limit)

MAX–maximum number of foldings to be computed (default 100)

MAX_LP–maximum bulge/interior loop size (default 30)

MAX_AS–maximum asymmetry of a bulge/interior loop (default 30)

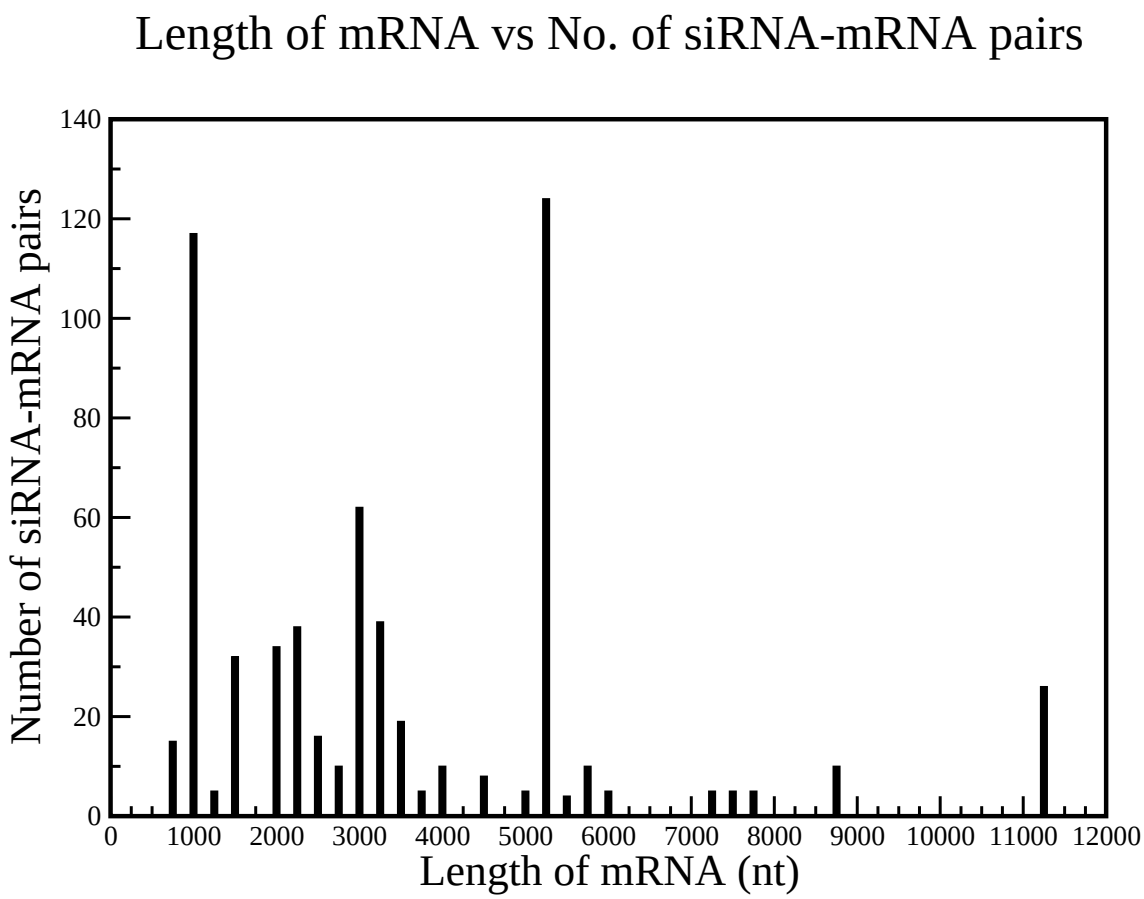


Figure 3.1: Distribution of siRNA-mRNA pairs by mRNA length. The lengths are plotted in bins of width 250 nt.

3.2.1 Terminology and Definitions

mRNA Target Strand

The target strand in the mRNA is the region complementary to the siRNA. The antisense strand of the double-stranded siRNA binds to it, leading to knockdown of the gene (refer Fig. 1.2). It is the same exact length as the siRNA, and therefore varies from 19-23 nucleotides in length.

mRNA Local Region(L)

The mRNA local region is defined as the target region plus a fixed number of nucleotides padding on both the $5'$ and $3'$ ends. The total length of this region is `L`. Generally, we use the same number of padding nucleotides on both ends. However, if the target strand is located

towards one end of an mRNA such that it cannot be centered on the local region, we adjust the paddings on $5'$ and $3'$ ends to ensure that `L` remains the same. If the length of the local region exceeds the mRNA length, we use the whole mRNA as the local region.

mRNA constraint length (cl)

RNAi reaction involves unraveling of the target strand region of the mRNA prior to the siRNA binding to it. Unravelling it might force a few nucleotides on either end to be single-stranded as well. We simulate this by forcing a few nucleotides at either end of the target strand to be single-stranded. The length of the sequence forced single-stranded at either end of the target strand is the `constraint length` (`cl`).

3.2.2 ΔG_{mRNA}

Determining the minimum energy secondary structure of a long RNA sequence requires significant computational expense with the methodology we employ in this work. Therefore, we calculate free energies only for the local region of the mRNA. `Mfold` provides parameters START and STOP which indicate the start and end of the sequence fragment under consideration.

Thus, calculation of free energy of the mRNA break involves the following two steps:

1) Calculation of free energy of the local region `L` of the mRNA.

2) Calculaiton of free energy of the local region of the mRNA with the constraint that the target region and constraint length at $5'$ and $3'$ ends of it are forced to be single stranded (unfolded).

Subtracting (2) from (1) gives the free energy required to unravel the target region. This method has been proposed by Lu and Mathews,[45] however, they force only the target region to be single-stranded as opposed to our method, where we force additional number of nucleotides adjacent to the target region to be single-stranded as well. This number is denoted by `cl`. Fig. 3.2 illustrates this process.

3.2.3 ΔG_{sirna}

The siRNA is introduced into the reaction as a double-strand sense-antisense duplex. However, only the antisense strand of the siRNA hybridizes with the complementary target strand in the mRNA. Therefore, we consider only the antisense strand of the dsRNA for our calculations and by siRNA we refer to the antisense strand henceforth. `Mfold` is used to calculate the free energy of the siRNA.

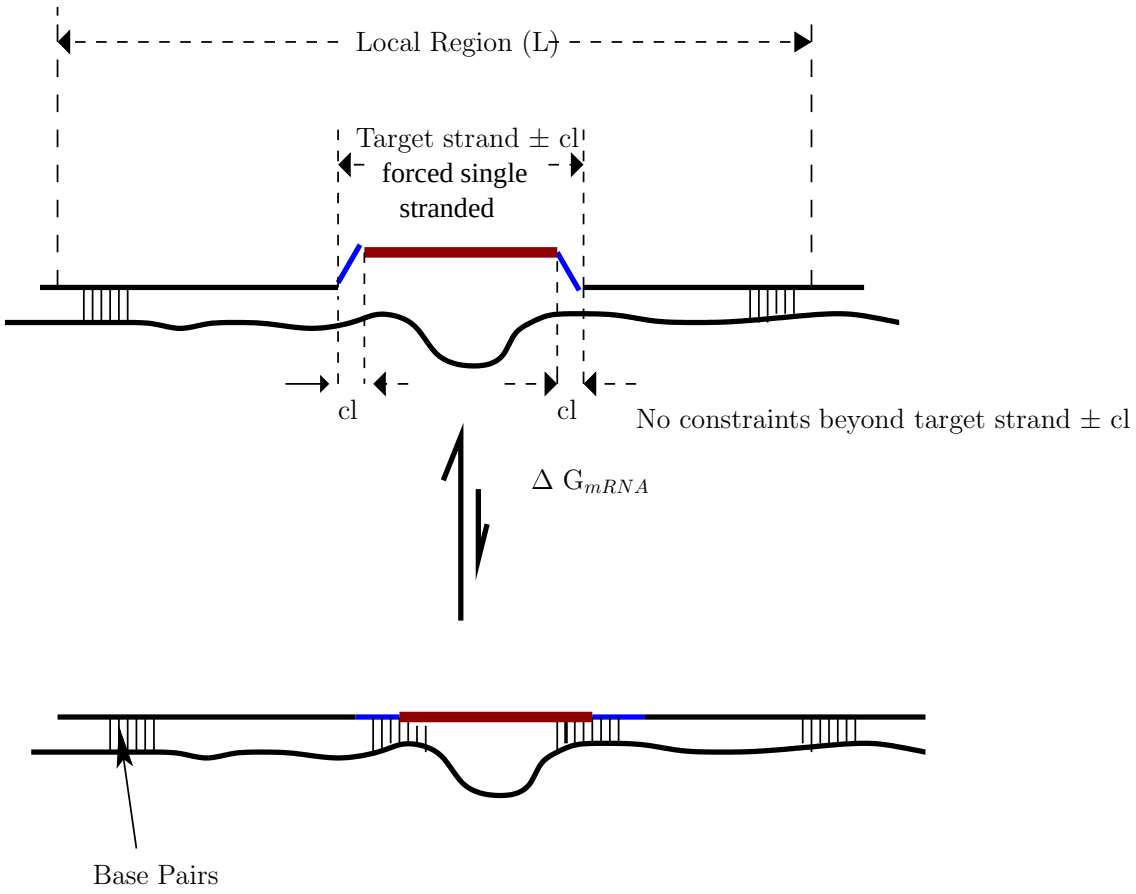


Figure 3.2: Formation of constrained mRNA. `L`–The length of the local region, which is the target strand plus an additional number of padding nucleotides at both ends, that is considered relevant for secondary structure computation. `cL`–Additional nucleotides at both end of the target strand that are forced single-stranded as the target strand unravels.

3.2.4 ΔG_{ms}

ΔG_{ms} is the sum of ΔG_{mRNA} and ΔG_{siRNA}, $\Delta G_{ms} = \Delta G_{mRNA} + \Delta G_{siRNA}$. We define this term because it represents the total free energy requirement on the left hand side of the reaction of complex formation from mRNA and siRNA [see Fig. 1.2]. In the Chapter 4, we will see how ΔG_{ms} is a significant indicator of RNAi efficiency.

3.2.5 $\Delta G_{complex}$ mRNA-siRNA

There is no straightforward way to calculate the free energy of two strands of nucleic acids pairing up. Here we describe how we calculate the free energy of hybridization of the complex.

From the mRNA, we take the target region and two nucleotides in addition from the $5'$ and $3'$ ends. Let us call the nucleotides from the $5'$ end XY, and the two at the $3'$ end PQ. If the length of the siRNA is l_s, we have $l_s + 4$ nucleotides from the mRNA. Then we attach a 275 poly-A string at the $3'$ end, followed by the QP, siRNA, and YX. Thus the whole complex is as follows XY-mRNA-PQ-(AAA...275 times)-QP-siRNA-YX. We fold this sequence with appropriate constraints such that only the mRNA target region and the siRNA are free to pair up, which they will, being exact complements of each other. To the free energy thus calculated, we add another 3.6Kcal/mol to obtain the free energy of the complex.

We arrive at the number 3.6 Kcal/mol as the energy of the poly-A loop by an experiment where we increase the size of the loop, and measure the difference in free energy caused by the loop. It approaches 7.2Kcal/mol asymptotic as the length of the loop increases. At length 275, the free energy difference is 3.6Kcal/mol. Thus, we determine the computational-time/accuracy tradeoff acceptable at a poly-A loop of length 275. This is illustrated in Fig. 3.3.

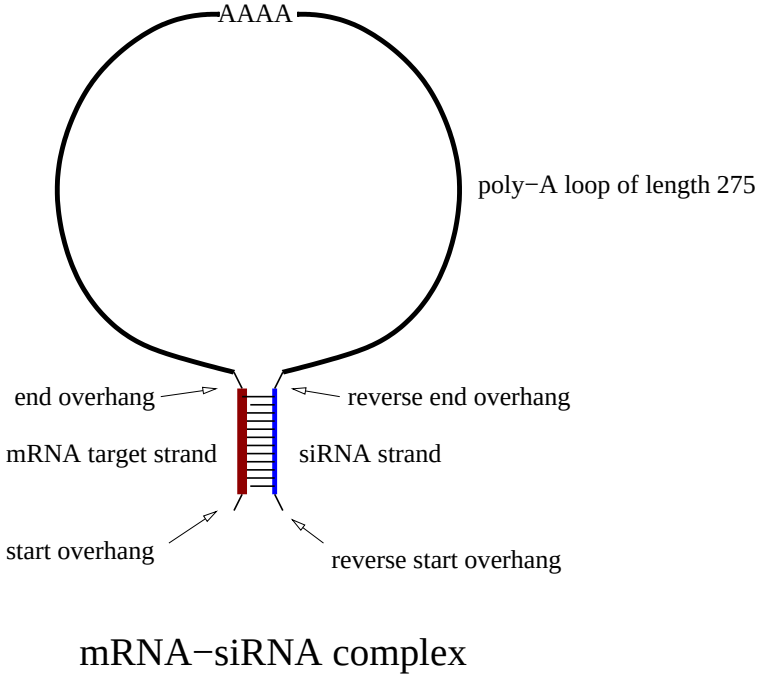


Figure 3.3: Schematic representation for free energy computation of siRNA-mRNA complex. The complex is constructed by combining the target strand, overhangs at either end of the target strand, a poly-A loop of length 275, the siRNA, and the reverse of the overhangs at either end of the siRNA strand.

3.2.6 ΔG_{total}

We obtain ΔG_{total} by subtracting ΔG_{ms} from $\Delta G_{complex}$, $\Delta G_{total} = \Delta G_{complex} - \Delta G_{ms}$. The lower the ΔG_{total} value, the thermodynamically better the siRNA-mRNA target strand combination should be at RNAi activity.

3.3 Statistical Analysis

In this section, we discuss the different representations of thermodynamic free energies that we use to find the correlation between the free energies and the RNAi efficiency.

For the main result in Section 4.1, we consider a single conformation with the minimum free energy out of all the possible conformations. The free energies and the natural logarithm of the corresponding activities, where activity is defined as the percentage of mRNA remaining after RNAi treatment as compared to the control, are grouped into bins. We sort the siRNA-mRNA target strand pairs based on the thermodynamic signature under consideration, e.g., ΔG_{ms}. Sixty-one pairs are put into one bin, to get a total of 10 bins. For each bin, we calculate the average free energy, and the average efficiency, where efficiency is the logarithm of the remaining level of mRNA. We plot these averages to obtain the graphs in Fig. 4.2.

For the various calculations discussed in Section 4.2 that use a subset of the full dataset, we try to bin the siRNA-mRNA target strand pairs there such that we obtain either the same number of bins or the same number of data points in each bin as in Section 4.1.

For the calculations in Section 4.2.3, we calculate the free energy using Boltzmann average. The Bolzmann average free energy is

$$\Delta G_{Boltzmann} = \frac{\sum_i \Delta G_i e^{-\frac{\Delta G_i}{kT}}}{\sum_i e^{-\frac{\Delta G_i}{kT}}}, \tag{3.1}$$

where the sum is over all possible conformations of the sequence. For our calculations, we consider a maximum of 100 conformations having free energy within 100% range of the minimum free energy conformation.

3.4 Support Vector Machine (SVM)

3.4.1 SVM Implementation

We use the `libsvm`[47] implementation of SVM. The implementation comes with a script `easy.py` that finds the proper values of parameters `C` and γ for the radial-basis kernel used by default by this implementation. We use this script to run all the SVM computations and do not manually tune any parameters.

3.4.2 Feature Space

18 of the 28 features used by Lu and Mathews are incorporated in our feature space. They contain the 15 sequence-position specific features and three computed free energy features ΔG_{siRNA}, ΔG_{mRNA}, and $\Delta G_{complex}$. In addition, we add two other energy features ΔG_{total} and ΔG_{ms}. Thus our feature space has 20 features, 15 sequence features and five computed free energy features. The full feature list is show in Table 3.2.

We train the SVM to classify an siRNA as efficient or inefficient (silencing efficiency greater than a threshold. We use 70% as the threshold for our calculations unless otherwise specified). The dataset is divided into a training set and a test set, and the SVM is trained on the training set and tested on the test set. If the number of data points in the training set is in $x:1$ proportion to the number of testing data points, the calculation is said to be $x+1$-*fold cross-validation*. Thus, if the dataset is divided in $3:1$ ratio of training and testing sets, this would be called 4-fold cross-validation. We conduct 2-fold, 4-fold, and 8-fold cross-validation calculations.

3.4.3 Performance Metrics

This section explains how we quantify the performance of the SVM at classification of siRNAs as functionally efficient or inefficient. Those siRNAs that result in silencing efficiency above a threshold (70% unless otherwise specified) in silencing experiments are classified as efficient, and the rest are inefficient. Thus, we pursue a binary classification scheme, and denote `efficient` as class `1`, and `inefficient` as `0`. Let us also denote the siRNAs predicted as `class 1` as `Positive` and the ones as `class 0` as `Negative`. Thus `True Positive(TP)` are the siRNAs which are predicted as positive and are actually positive. `False Positive(FP)` are the ones classified as positive but are actually negative. `True Negative(TN)` are the siRNAs which are classified as negative and are experimentally determined to be negative. `False Negative(FN)` are the ones which are wrongly classified as negative. Using these numbers, we use the following metrics to determine the performance of the SVM.

Feature	Position on siRNA Sequence	Class Name
ΔG_{mRNA}	N/A	ΔG
ΔG_{siRNA}	N/A	ΔG
ΔG_{ms}	N/A	ΔG
ΔG_{total}	N/A	ΔG
$\Delta G_{complex}$	N/A	ΔG
ΔG_1	1	Seq
ΔG_2	2	Seq
ΔG_{13}	13	Seq
ΔG_{18}	18	Seq
$\Delta\Delta G_{1\_19}$	1,19	Seq
A	19	Seq
C	1	Seq
CC	1	Seq
CG	1	Seq
G	1	Seq
GC	1	Seq
GG	1	Seq
U	1	Seq
U	2	Seq
UU	1	Seq

Table 3.2: Feature space used in SVM calculations. The position on the sequence as calculated from 5' end of the antisense strand of the siRNA, and the class of the feature are listed. If it is calculated using equilibrium secondary structure predictions (`Mfold`), it is classified as ΔG, otherwise it is classified as of class `Seq`. {All}= Set containing all the features, $\{\Delta G\}$ = $\{\Delta G_{mRNA}, \Delta G_{siRNA}, \Delta G_{ms}, \Delta G_{total}, \Delta G_{complex}\}$, and {Seq}={All} - $\{\Delta G\}$

Accuracy

Accuracy is defined as the percentage of correct predictions of the total number of predictions made. Thus,

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \tag{3.2}$$

Positive Predictive Value (PPV)

Positive Predictive value is the percentage of siRNAs predicted efficient that are actually efficient as proved by silencing experiments. Thus,

$$PPV = \frac{TP}{TP + FP} \times 100\% \tag{3.3}$$

Sensitivity

Sensitivity is the percent of efficient siRNAs that are predicted to be efficient. Thus,

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \tag{3.4}$$

Specificity

Specificity is defined as the percent of inefficient siRNAs that are predicted to be inefficient. Thus

$$Specificity = \frac{TN}{TN + FP} \times 100 \tag{3.5}$$

Method Improvement Quotient (IQ)

We present this metric to account for the difficulty in improving the performance of a method as the baseline gets higher. In addition, we want to incorporate the Null model performance in quantifying the improvement of performance due to methods we test. To illustrate, methods based purely on random coin flip will have 50% accuracy (by simply "predicting" one class all the time). Thus, an accuracy of say 55% achieved by a method is not really much of an improvement over the prediction based on the "Null" model, which is the prediction performance of a random draw. On the other hand, an improvement to 99% from 90% provided by a certain method is noteworthy.

Consider PPV as the metric of interest. In the context of a typical RNAi application, it is worthwhile to present a few good siRNA sequences for further testing as opposed to predicting just one "best" candidate for a particular mRNA. We now define the Method Improvement Quotient (IQ) for PPV. Suppose PPV on random prediction is PPV_{null} (this equals the percentage of the efficient class in the whole dataset), and PPV for a method in question is PPV_{method}. Then IQ is defined by

$$IQ = \frac{100 - PPV_{null}}{100 - PPV_{method}}. \tag{3.6}$$

In our dataset, there are 247 out of 609 siRNAs with efficiency greater than the threshold of 70%, yielding a PPV_{null} of 40.56%. This is the value of PPV_{null} we use in future PPV calculations unless otherwise specified.

Chapter 4

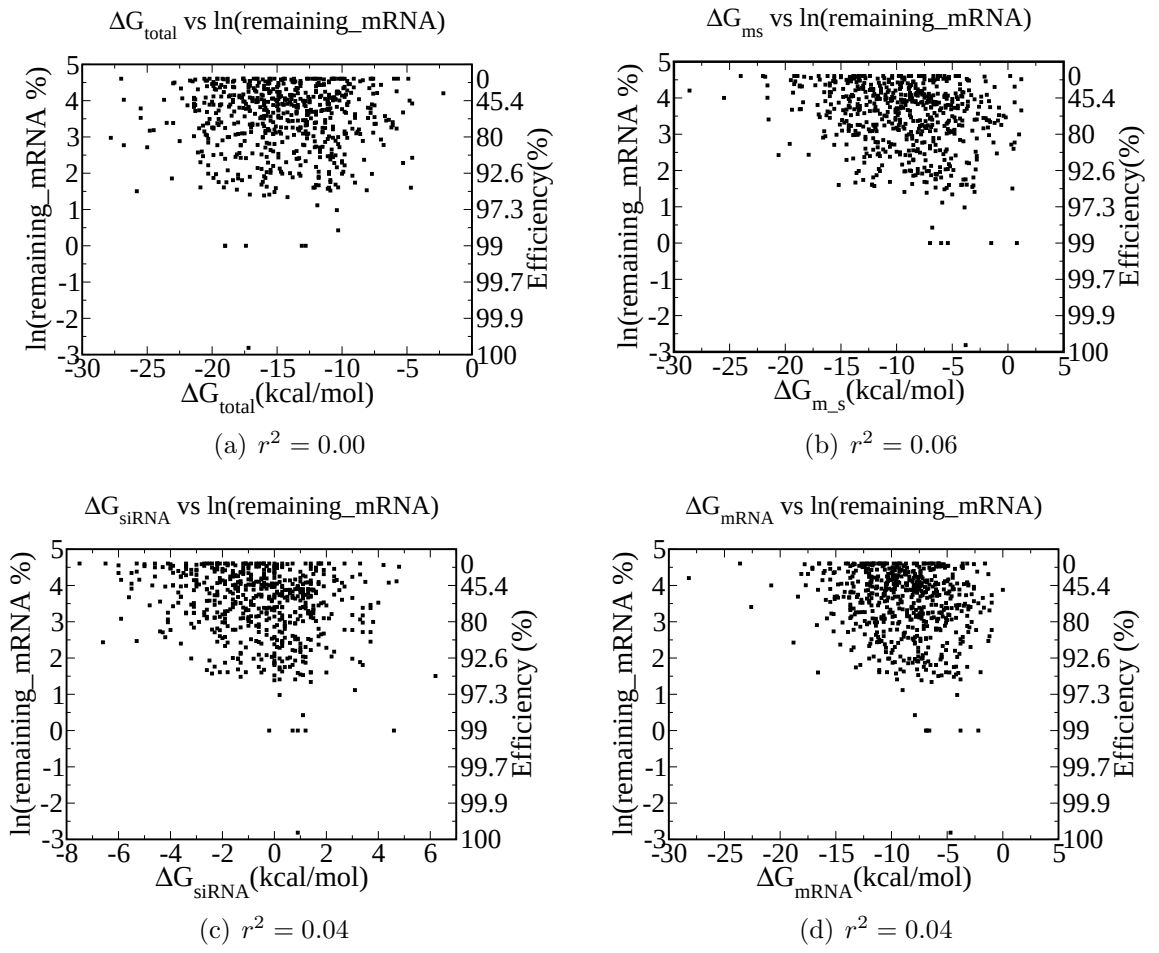
Results and Discussion

4.1 Correlations between RNAi efficiency and secondary structure ΔG

Our key goal is to investigate to what extent the efficiency of the RNAi reaction is governed by the secondary structure of its components. Specifically, we want to see to what extent the simplistic mechanism shown in Fig. 1.2 holds, which implies that the RNAi can be described by equilibrium thermodynamics, e.g., Equation 2.4. To this end, we have computed Gibbs free energies of mRNA, siRNA, and their combinations and compared them with the corresponding RNAi efficiencies for each of the 609 experimental data points used in this study [see Chapter 3 for important details]. The result is shown in Fig. 4.1.

The results are in general agreement with those obtained by Lu and Mathews,[45] in that there is no appreciable correlation between the free energies and the efficiency of the RNAi reaction. It is still possible that there may exist an underlying trend, but it is masked by a large amount of "noise" arising out of various factors such as methodological errors or unaccounted for biological properties of the RNAi mechanism. To discern the possible trend, we average the free energies and corresponding average RNAi efficacies over 10 bins as discussed in Section 3.3 to obtain the main results of this study, as shown in Fig. 4.2. We can clearly see that, on average, there is significant correlation between RNAi efficiency and some of the equilibrium free energies, with the caveat that this correlation is observed only when the values averaged over bins are considered.

From the plots, we observe that there is good correlation of logarithm of the remaining level of mRNA in the RNAi experiment with ΔG_{mRNA} and with ΔG_{siRNA}. The trend of the correlation is consistent with the simplistic thermodynamic model of RNAi reaction presented in Fig. 1.2. This is a non-trivial observation because the model does not account for various possible biological and chemical factors. What is clearly more interesting is the



(a) $r^2 = 0.00$ (b) $r^2 = 0.06$

(c) $r^2 = 0.04$ (d) $r^2 = 0.04$

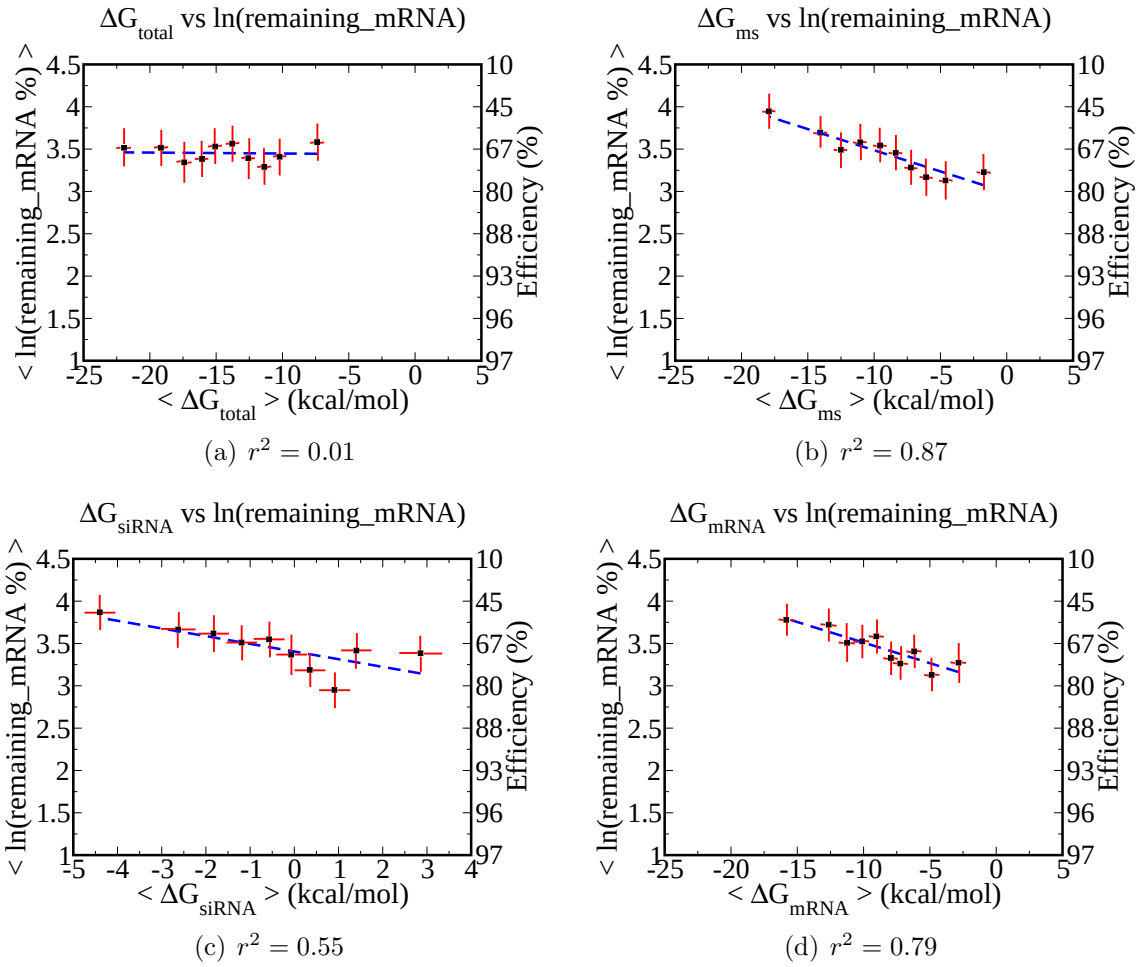
Figure 4.1: RNAi efficiency vs computed free energy components of RNAi reaction. See Fig. 1.2. `L` = 100, `cl` = 2

higher correlation for the sum of ΔG_{mRNA} and ΔG_{siRNA}, ΔG_{ms}. This is in agreement with our model as well, and expected because it takes into account the total free energy cost of getting the reactants to ready-state for the reaction to progress.

However, another observation that is not quite in agreement with our model is the low correlation observed for ΔG_{total}, while one would normally expect otherwise because ΔG_{total} accounts for more factors than ΔG_{ms}. Though we do not have any confirmed explanation for this, we propose the following speculative explanations:

1) The model assumes equilibrium thermodynamics while it is possible that the reaction equilibrium has not been reached yet.

2) The model is deficient in that it does not account for other biological, chemical, and thermodynamic factors. These unaccounted for factors play a significant enough role in the RNAi reaction that without them the model is fundamentally flawed. However, the fact



(a) $r^2 = 0.01$ (b) $r^2 = 0.87$

(c) $r^2 = 0.55$ (d) $r^2 = 0.79$

Figure 4.2: RNAi efficiency vs average computed free energy components of RNAi reaction. The error bars represent the statistical error of the mean. `L` = 100, `cl` = 2

that a clear trend emerges when we average over bins indicates the effect of these factors are probably cancelled out.

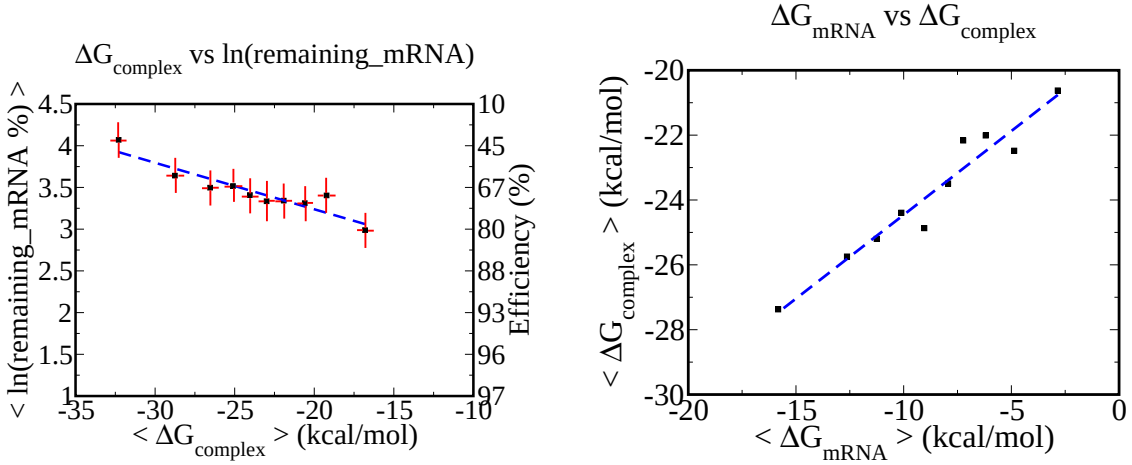
3) The computational protocol we employ for our calculations is flawed. However, we discount this after further analysis regarding the robustness of our model to details of the computational protocol. The results of this analysis are presented in Section 4.2.

4) ΔG_{ms} represents the reaction barrier for RNAi reaction, and consequently the reaction kinetics, while ΔG_{total} corresponds to the reaction equilibrim. Higher correlation for ΔG_{ms} possibly indicates a more pronounced effect of reaction kinetics on RNAi activity compared to that of equilibrium considerations.

5) Some clues may come from the puzzling behavior of $\Delta G_{complex}$. Fig. 4.3(a) shows the plot for average $\Delta G_{complex}$ vs siRNA efficiency. We observe that the trend of the correlation between $\Delta G_{complex}$ is opposite to what we would normally expect from the model in Fig.

1.2. Fig. 4.3(b) shows the plot of $\Delta G_{complex}$ vs ΔG_{mRNA}. We see that there is a very high correlation between the two; and this might result in a poor correlation for $\Delta G_{total} = \Delta G_{complex} - \Delta G_{mRNA} - \Delta G_{siRNA}$ because $\Delta G_{complex}$ and ΔG_{mRNA} cancel each other out.

The counter-intuitive trend observed in Fig. 4.3(a) could be because siRNA activity prefers a less stable siRNA-mRNA complex.[9] In addition, the high correlation in Fig. 4.3(b) could also be responsible for this trend.



(a) $r^2 = 0.86$ (b) $r^2 = 0.94$

Figure 4.3: Behavior of $\Delta G_{complex}$. (a) Correlation between the average experimental efficiency of RNAi and $\Delta G_{complex}$, see Fig. 1.2. The error bars represent the statistical error of the mean. `L` = 100, `cl` = 2. (b) Correlation between average $\Delta G_{complex}$ and ΔG_{mRNA}.

As we have discussed earlier, there is a lot of noise in Fig. 4.1. This could be a result of deficiencies in the thermodynamic model we use. It could also arise from unaccounted for biological, chemical, and thermodynamic factors. Finally, deficiencies in our computational protocol could mean that the results we observe are not consistent. In the next section, we present our efforts to determine the robustness of our model.

4.2 Robustness to Details of the Computational Protocol

The goal in this section is to explore robustness of our main results to details of the computational protocol.

4.2.1 Influence of mRNA length on structure/efficiency correlation

In our model, there are a number of parameters [see Fig. 3.2]. The result of varying some of them are presented in Table 4.1. We show the square of the correlation between average free energies and average silencing efficiency for different `cl` and `L` values when applied on the Shabalina dataset. The free energy values are calculated as described in Chapter 3. Fig. 4.2 presents the values in bold in the table as this set of `cl=2` and `L=100` values gives high correlation for both ΔG_{mRNA} and ΔG_{ms}.

L	cl	$< \Delta G_{total} >$	$< \Delta G_{ms} >$	$< \Delta G_{mRNA} >$	$< \Delta G_{siRNA} >$	$< \Delta G_{complex} >$
800	0	0.04	0.88	0.59	0.55	0.86
800	2	0.03	0.73	0.57	0.55	0.86
800	10	0.06	0.71	0.67	0.55	0.86
200	0	0.14	0.77	0.63	0.55	0.86
200	2	0.02	0.76	0.66	0.55	0.86
200	10	0.11	0.76	0.74	0.55	0.86
100	0	0.09	0.72	0.64	0.55	0.86
100	**2**	**0.00**	**0.87**	**0.79**	**0.55**	**0.86**
100	10	0.11	0.87	0.65	0.55	0.86
90	0	0.08	0.67	0.65	0.55	0.86
90	2	0.02	0.82	0.71	0.55	0.86
90	10	0.02	0.69	0.64	0.55	0.86
50	0	0.02	0.78	0.82	0.55	0.86
50	2	0.00	0.85	0.81	0.55	0.86
50	10	0.03	0.89	0.75	0.55	0.86
Average		0.05	0.78	0.69	0.55	0.86

Table 4.1: Correlation between RNAi efficiency and free energy for different `L` and `cl`. The ΔG values represent the correlation for the corresponding free energy signature whose value depends on the parameters `L` and `cl` in the first two columns.

Another trend that is apparent is that large values of `cl` lead to lower correlation. We think this is because the RNAi machinery does not require extra nucleotides at either end of the target region to be single-stranded. Even if they do get unpaired, it is possible this does not effect the silencing efficiency–hence incorporating them as a factor in our calculation leads to a worsening of the results. Though the correlation is lower, the general trend observed in the table [Plots not shown] for different `L` and `cl` values are still in accordance with our expectations, i.e., lower ΔG_{ms} values indicate more stable constituents of the reaction and hence lower RNAi efficiency, and lower ΔG_{total} values indicate more stable siRNA-mRNA duplex and hence higher RNAi efficiency.

As discussed previously, the noise in the data could have been introduced by deficiencies

in secondary structure prediction. The accuracy of folding tools such as `Mfold` is known to deteriorate for long RNAs. To mitigate this problem, we follow others such as Schubert et. al.,[8] and Lu and Mathews[45] and apply the folding algorithm to a local region of length L that encompasses the 19-nucleotide siRNA target sequence [see Fig. 3.2], and ignore the rest of the sequence.

However, this strategy is not perfect. One concern about using the local mRNA length is that it ignores possible effects of the rest of the mRNA on the secondary structure of the local region. We restrict the folding calculation to the local region on the assumption that the globally folded structure has the same fold in the local region as in the locally folded one. However, it is possible that the structure we obtain via the "local folding" is not the correct one. Thus, the longer the mRNA, the higher the part ignored for a fixed length of local region calculation, and higher the possibility of error in secondary structure consideration.

If we choose a short subset of mRNAs from our dataset (Fig. 3.1), it reduces the ignored sequence, and hence increases the probablity that we obtain the correct secondary structure for the local region. We examine the results of our approach on shorter mRNAs. A higher correlation than observed in Fig. 4.2 would indicate that the errors indeed arise from deficiencies in secondary structure calculation.

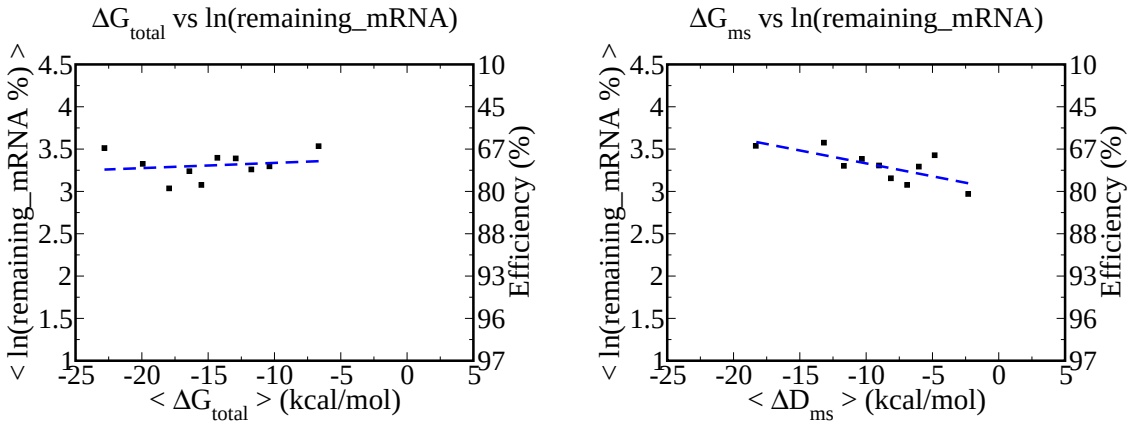
For this purpose, we choose approximately one-fourth (184) of the total number of siRNA-mRNA pairs in the our dataset, which correspond to the shortest mRNAs ranging in length from 570 nt to 1821 nt. We apply our calculations on this subset. Table 4.2 summarizes the results of this calculation. Fig. 4.4 show the results for ΔG_{ms} and ΔG_{total} for `L`=100 and `cl`=2.

L	cl	$< \Delta G_{total} >$	$< \Delta G_{ms} >$	$< \Delta G_{mRNA} >$	$< \Delta G_{siRNA} >$	$< \Delta G_{complex} >$
100	0	0.03	0.35	0.21	0.14	0.23
100	2	0.03	0.53	0.23	0.14	0.23
100	10	0.04	0.33	0.08	0.14	0.23
Full	0	0.11	0.27	0.53	0.14	0.23
Full	2	0.13	0.26	0.25	0.14	0.23
Full	10	0.02	0.29	0.07	0.14	0.23

Table 4.2: RNAi efficiency vs free energy for shortest 184 siRNA-mRNA pairs. The ΔG values represent the correlation for the corresponding free energy signature whose value depends on `L` and `cl` in the first two columns.

We observe no improvements in correlations between structure and efficiency for short mRNAs. It could be because the number of data points is small to have large enough number of bins at the same time with large enough data points in each bin.

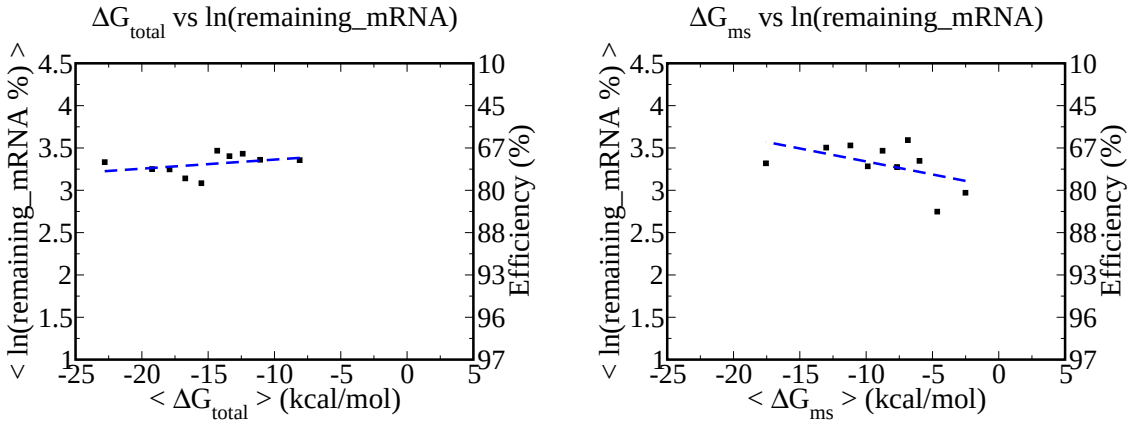
To address the concern that choosing only 100 nucleotides for the local region ignores the possible effects of the rest of the mRNA, we run the calculations for the above 184 mRNA



(a) $r^2 = 0.03$ (b) $r^2 = 0.53$

Figure 4.4: RNAi efficiency vs free energy for shortest 184 siRNA-mRNA pairs. `L` = 100. `cl` = 2

with the full mRNA as the local region. The results are given in Fig. 4.5.



(a) $r^2 = 0.13$ (b) $r^2 = 0.26$

Figure 4.5: RNAi efficiency vs free energy for shortest 184 siRNA-mRNA pairs, `L` = full. `cl`=2

We see that accounting for the full mRNA does not improve the results, compared to that in Fig. 4.4. Although it does not absolutely prove that the methodology is not to blame, it reduces the chance significantly and we believe the noise is introduced most likely by biological factors.

We notice that the correlations are indeed worse than those in Fig. 4.2. Thus, we conclude that deficiencies in secondary structure calculations do exist for large mRNAs. However, over the local region of 100 nucleotides that we use for our calculations, those deficiencies do not contribute significantly towards the noise we observe in Fig. 4.1.

4.2.2 Possible Influence of Variation in Experimental Conditions

The dataset we use has been collected from different experiments. Possible differences in the way each experiment is set up and conducted may lead to errors if data from different experiments are merged. Hence, we choose two experiments with the highest number of siRNA-mRNA entries in the dataset, and perform our calculations separately on these two subsets with the 10 bins, the same number of bins as in Fig. 4.2. For comparison, we select the same number of siRNA-mRNA pairs as in the experimental subset of interest at random from the full dataset, and compute the correlations over 10 runs. The results are shown in Table 4.3

Set	$<\Delta G_{total}>$	$<\Delta G_{ms}>$	$<\Delta G_{mRNA}>$	$<\Delta G_{siRNA}>$	$<\Delta G_{complex}>$
Khvorova et. al,[48] 179 pairs	0.13	0.78	0.55	0.28	0.58
Random Selection, 179 pairs	0.1 ± 0.1	0.58 ± 0.12	0.34 ± 0.16	0.39 ± 0.2	0.64 ± 0.16
Hsieh et. al,[49] 103 pairs	0.21	0.26	0.26	0.02	0.02
Random Selection, 103 pairs	0.08 ± 0.05	0.43 ± 0.2	0.31 ± 0.18	0.31 ± 0.21	0.5 ± 0.19

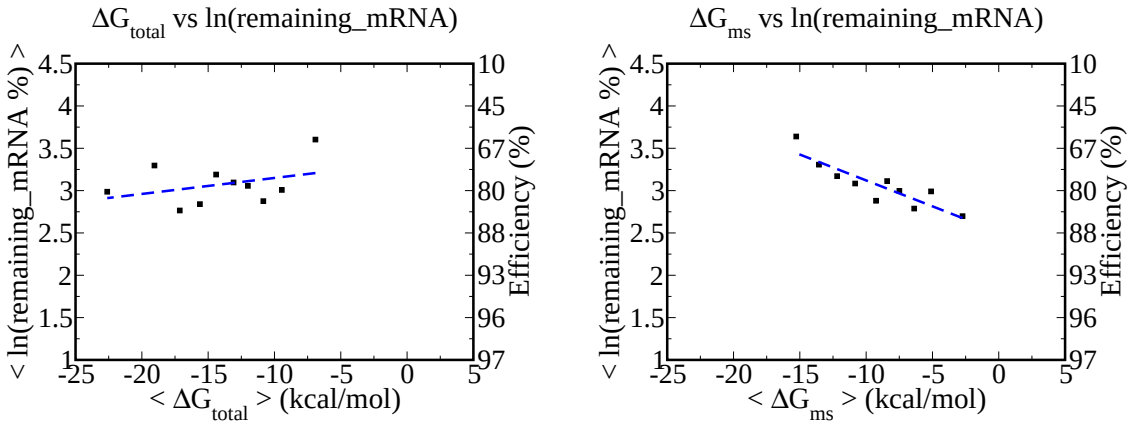
Table 4.3: RNAi efficiency vs free energy for two largest experimental subsets. For comparison, the same number of siRNA-mRNA pairs as in the dataset are randomly selected and the correlations calculated over 10 runs.

The Khvorova et. al[48] dataset is the largest with 179 siRNA-mRNA pairs. It has 2 mRNAs of lengths 851 nt and 5010 nt targeted by 89 and 90 siRNAs respectively. Fig. 4.6 shows the results for this dataset.

We calculate the correlations for the second largest experimental set[49] in our dataset. This dataset comprises 103 siRNA-mRNA pairs for 21 mRNAs. The mRNAs range in length from 829 nt to 11242 nt. Fig. 4.7 shows the results for this dataset.

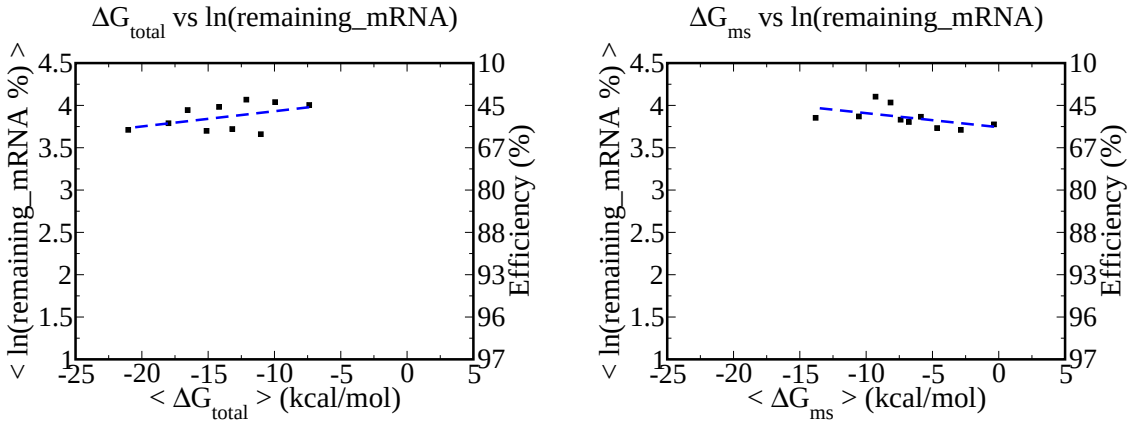
We observe slightly higher correlation for ΔG_{ms}in Fig. 4.6(b), and the others are similar or worse compared to Fig. 4.2. Thus, analysis within a single experimental set is likely to yield better results. But for purposes of analyzing the correlation between RNAi efficiency and computed equilirbrium free energies, our choice of dataset is better because we average over the particular influence of a single experimental set.

The higher correlations in Fig. 4.6 compared to those in Fig. 4.7 could also arise from the difference in the number of mRNAs used in the two experiments. Khvorova et. al[48] use only two mRNAs for the results of 179 siRNA-mRNA pairs in Fig. 4.6, while Hsieh et. al[49] use 21 mRNAs for 103 pairs in Fig. 4.7. This suggests the possibility that data from experiments with a single mRNA is consistent with our model; however combining the



(a) $r^2 = 0.13$ (b) $r^2 = 0.78$

Figure 4.6: RNAi efficiency vs free energy for Khvorova et.al.[48] This experimental subset in our dataset has the highest number of siRNA-mRNA pairs, 179. `L` = 100, `cl` = 2



(a) $r^2 = 0.21$ (b) $r^2 = 0.26$

Figure 4.7: RNAi efficiency vs free energy for Hsieh et. al.[49] This experimental subset in our dataset has the second highest number of siRNA-mRNA pairs, 103. `L` = 100, `cl` = 2

results from different mRNAs could worsen the correlation. In Table 4.4, we present the correlations between ΔG_{ms} and RNAi efficiency for 5 mRNAs with the highest frequency of siRNA-mRNA pairs within a single experimental sub-dataset in our dataset. For comparison, we select the same number of siRNA-mRNA pairs at random from our dataset and calculate the correlation for ΔG_{ms}.

mRNA	No. of targetting siRNAs	Correlation	Random Selection
$U47298_1$	90	0.13	0.07 ± 0.04
M60857	89	0.04	0.06 ± 0.05
J03132	38	0.08	0.07 ± 0.05
$U47298_2$	34	0.3	0.14 ± 0.08
U92436	29	0.00	0.09 ± 0.09

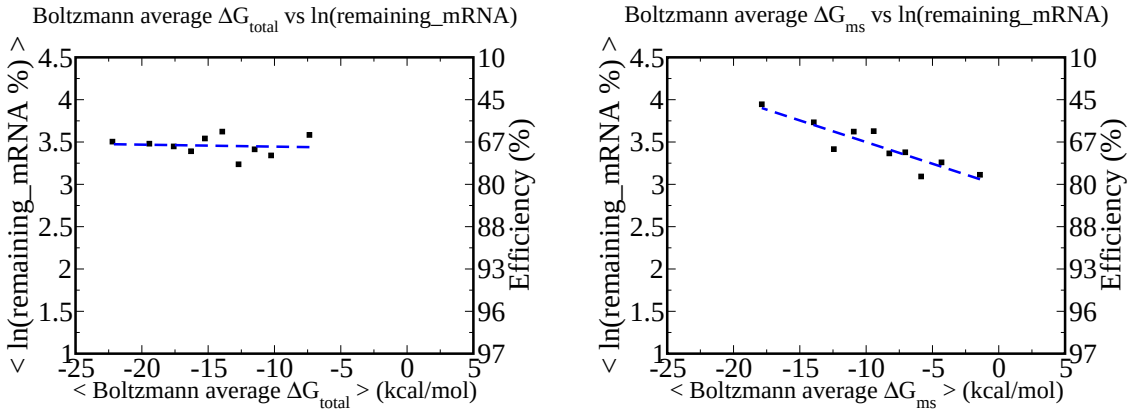
Table 4.4: RNAi efficiency vs ΔG_{ms} for mRNAs with highest number of targetting siRNAs. For comparison, in the column Random Selection, we compute correlations over 10 runs for the same number of siRNA-mRNA pairs selected at random from our dataset. U47298 occurs twice because it is used in two different experiments. The mRNAs are identified by their corresponding GenBank accession numbers.

We do not find any worsening of the correlations for random values compared to those for single mRNAs. It appears the silencing efficiency of an siRNA depends on the target mRNA. It is conceivable that use of multiple mRNAs leads to weaker correlation. However, the significantly higher correlation in our earlier results (Fig. 4.2) which combine results from multiple mRNAs indicate that our conclusions are robust to choice of mRNA.

4.2.3 Use of Single Minimum Energy vs Boltzmann Average Energy

The free energies we have used so far in our calculations are the minimum values, representing the most stable secondary structure as predicted by `Mfold`. However, in reality, the equilibrium comprises multiple secondary structures, their relative abundance in the solution being in exponential proportion to the stability of the structure. Thus, a more accurate calculation would take into account the suboptimal structures as well.[45] Indeed, the best calculation would consider all the possible secondary structures. However, such a calculation is computationally expensive, hence we compromise by taking a Boltzmann weighted average of all the possible secondary structures with a free energy in the range of 100% deviation from the minimum free energy, subject to a maximum of 100 possible structures, Equation 3.1. The results are shown in Fig. 4.8.

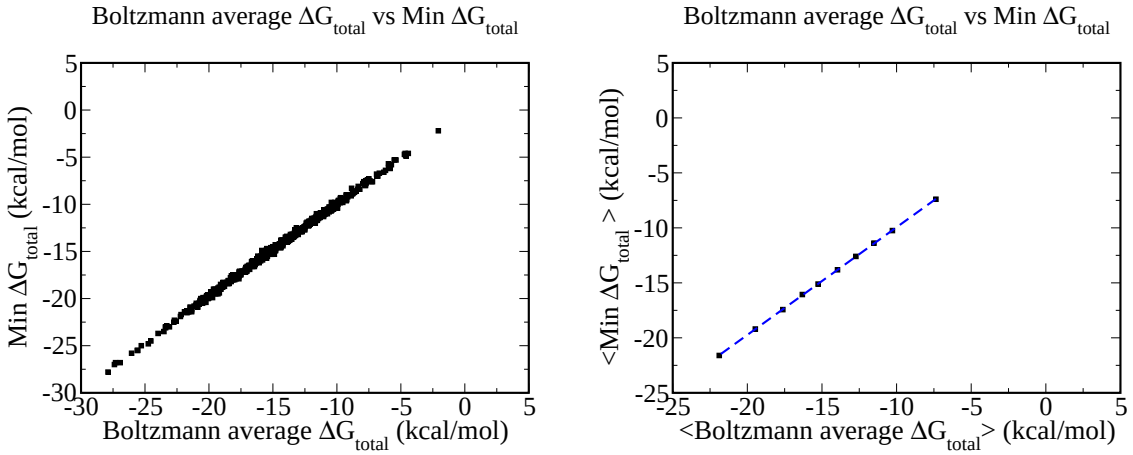
We observe that the results are very close to those observed in Fig. 4.2. This could be because the Boltzmann average free energy calculation does not change the free energy values by much from the minimum free energies. Fig. 4.9 shows the plot for Boltzmann average



(a) $r^2 = 0.01$ (b) $r^2 = 0.82$

Figure 4.8: RNAi efficiency vs Boltzmann average free energy. `L` = 100, `cl` = 2

and minimum ΔG_{total}s. There is not much difference between the free energies calculated using minimum energy values and the ones using average energy values, thus confirming the results of Fig. 4.8.



(a) All points. $r^2 = 0.99$ (b) Averaged over bins of size 61. $r^2 = 0.99$

Figure 4.9: Average ΔG_{total} over set of conformations vs minimum energy conformation.

4.3 SVM Based Predictions

In this section, we describe the results of our computational experiments using a support vector machine (SVM). We performed 2-fold, 4-fold, and 8-fold cross validation calculations on the dataset described in Section 3.4. To determine the standard deviation, each calculation

was conducted 10 times with training and testing subsets chosen at random. We calculate the mean and standard deviation for the following metrics of performance: accuracy, PPV, sensitivity, specificity, and IQ [See Section 3.4.3 for definitions and further details]. In addition, to understand the improvement achieved by using SVM over the null model, the null model results are calculated. The results are given below.

Test/Feature Set	Accuracy(%)	PPV(%)	Sensitivity(%)	Specificity(%)	IQ
Baseline	60	Indeterminate	0	60	N/A
Null model	50	40	20	50	1
2-fold/{All}	64.5 ± 3.3	59.1 ± 4.2	47.7 ± 5.9	76 ± 6.7	1.5
4-fold/{All}	**67.4 ± 4.9**	**61.1 ± 5.1**	**50.8 ± 13.4**	**78.9 ± 6.1**	**1.5**
8-fold/{All}	69.3 ± 5.6	63.9 ± 6.8	56.3 ± 10	78.4 ± 6.2	1.7

Table 4.5: Results of 2-fold, 4-fold, and 8-fold SVM cross validation predictions. The averages reported are calculated on 10 runs. {} indicates that all of the features of the specified class were considered.

We obseve that 8-fold cross validation performs the best and yields the most improvement over the corresponding baseline numbers. However, we notice 4-fold performs only slightly worse compared to 8-fold, and we choose 4-fold cross validation for further analysis. In case we further sub-divide the test set in half according to some criteria, we shall still have enough data points in the test set to yield statistically significant results, as shown for 8-fold cross-validation.

We should note here that the Null model is distinct from the commonly used baseline model. In the baseline model, the classifier consistently predicts the most frequent class in the training set. In our case, the classifier would consistently predict "Functionally inefficient", which is `Negative` according to our definitions in Section 3.4.3. This would lead to $TP + FP = 0$, and an indeterminate value for $PPV = TP/(TP + FP)$. Since PPV is our metric of interest in this exercise, we use the Null model for comparison.

4.3.1 Further Computational Analysis with SVM

In order to improve the predictive performance, we tried various combinations of the feature space. Each combination uses 4-fold cross-validation. The results are aggregated over ten trials. We select different combinations of attributes for feature space. Table 4.6 shows the results of these calculations.

From the results in Table 4.6 we observe that the SVM does not perform well with only energy features or with only sequence features. A combination of sequence and energy features perform better than either only sequence or only energy features. However, these are still poorer than using all the available features for 4-fold cross validation as shown in Table 4.5.

Feature Space	Accuracy(%)	PPV(%)	Sensitivity(%)	Specificity(%)	IQ
{All}	67.4 ± 4.9	61.1 ± 5.1	50.8 ± 13.4	78.9 ± 6.1	1.5
{Seq}	64.5 ± 3.4	57.8 ± 9.7	45.5 ± 5.8	77.2 ± 8.4	1.4
$\{\Delta G\}$	56.5 ± 3.5	53.5 ± 20.3 *	11.3 ± 8.6	90.1 ± 10.9	1.3
ΔG_{ms}	56.4 ± 3.8	32.5 ± 11 *	2.6 ± 4.6	95.9 ± 8.1	0.9
ΔG_{total}	59.5 ± 4.5	30.8 ± 21.7 *	0.9 ± 1.5	99.1 ± 1.2	0.9
{Seq} + ΔG_{mRNA}	**66.1 ± 3.4**	**60.2 ± 6.7**	**51.4 ± 4.3**	**76.3 ± 5.6**	**1.5**
{Seq} + ΔG_{siRNA}	66.7 ± 3.8	59.8 ± 6.8	52.2 ± 6.9	76.4 ± 7.5	1.5
{Seq} + ΔG_{ms}	66.5 ± 3.6	58.7 ± 10.3	53.1 ± 3	75.5 ± 6.5	1.4
{Seq} + $\Delta G_{complex}$	64.7 ± 2.3	59.7 ± 5	49.2 ± 6.2	75.8 ± 6.6	1.5
{Seq} + ΔG_{total}	64.9 ± 4	60.8 ± 7.5	49.2 ± 7.1	76.4 ± 7.6	1.5

Table 4.6: 4-fold cross-validation analysis with different feature spaces. Entries marked with (*) have one or more indeterminate values of PPV among the 10 runs; the averages reported are calculated on the remaining valid values. {} indicates that all of the features of the specified class were considered.

What is clear from the results is that energy features do have a role to play in SVM classification.

4.3.2 Pre-filtering by ΔG_{ms}

We consider improving the SVM performance by pre-filtering the dataset. To this end, we present in this section the results of our efforts at SVM classification where we pre-filter the dataset according to ΔG_{ms} and then perform 4-fold cross-validation on the filtered dataset with different feature spaces.

We sort the dataset according to ΔG_{ms}. We know that the higher the ΔG_{ms} value, the easier it is to break the secondary structure of the mRNA and siRNA involved in the RNAi reaction, and hence the more thermodynamically efficient the reaction is. Thus, a higher ΔG_{ms} value indicates a higher RNAi silencing efficiency. Hence, we choose the half with higher ΔG_{ms} value from the sorted dataset.

To better understand the influence of pre-filtering, we conduct cross-validation exercises to establish a baseline. For baseline calculations, we randomly divide the dataset in half, and conduct 4-fold cross-validation analysis on one half. Thus for pre-filtering and baseline calculations, we have the same number of data points for training the SVM and testing it. Table 4.7 shows the results of 4-fold cross-validation on the dataset.

From the table, it is evident that pre-filtering by ΔG_{ms} results in small but consistent improvement of the performance of SVM classification. The improvement ranges from 3 to 7%, depending on the set of features. Note that the PPV calues in Table 4.7 should be

Filter/Feature Space	Accuracy(%)	PPV(%)	Sensitivity(%)	Specificity(%)	IQ
ΔG_{ms}/ {All}	59.74 ± 4.89	60.76 ± 7.56	57.15 ± 8.71	62.71 ± 11.16	1.51
Null/ {All}	64.1 ± 5.0	57.2 ± 8.9	46.6 ± 17.8	76.6 ± 10.6	1.4
ΔG_{ms}/ {Seq}	61.7 ± 4.7	61.1 ± 7.4	64.1 ± 13.7	60.5 ± 9.3	1.5
Null/ {Seq}	62.8 ± 7	60.7 ± 6.8	42.4 ± 16.9	78.5 ± 9.5	1.5
ΔG_{ms}/{All}-ΔG_{ms}	59.7 ± 5	59.5 ± 9.5	62.2 ± 5.5	57.7 ± 11.2	1.5
Null/ {All}-ΔG_{ms}	65.4 ± 5.4	55.7 ± 12.2	43.5 ± 14.9	78.8 ± 5.9	1.3
ΔG_{ms}/ {Seq}+ΔG_{mRNA}	61.6 ± 3.6	59.4 ± 7.0	66.7 ± 6.8	57.2 ± 10.7	1.5
Null/ {Seq}+ΔG_{mRNA}	64.9 ± 5.6	59.6 ± 13.1	43.9 ± 9	78.6 ± 10.6	1.5
ΔG_{ms}/ {Seq}+ΔG_{siRNA}	64.2 ± 3.6	62.4 ± 5.4	± 64.2 ± 13	65.1 ± 7.3	1.6
Null/ {Seq}+ΔG_{siRNA}	62.2 ± 5.8	57.8 ± 11.7	42.1 ± 15.5	77.1 ± 11.2	1.4
ΔG_{ms}/ **{Seq}**+ΔG_{ms}	**63 ± 5.6**	**65 ± 7**	**61 ± 11.7**	**66.2 ± 9**	**1.7**
Null/ {Seq}+ΔG_{ms}	63.7 ± 5.7	58.9 ± 7.9	46.7 ± 14.6	75.8 ± 5	1.4
ΔG_{ms}/ {Seq}+$\Delta G_{complex}$	61.4 ± 3.8	59.7 ± 8.7	63.1 ± 12.9	61.4 ± 11.3	1.5
Null/ {Seq}+$\Delta G_{complex}$	64.6 ± 5.5	59.2 ± 12.5	46.9 ± 8	77.7 ± 8.1	1.5
ΔG_{ms}/ {Seq}+ΔG_{total}	62.4 ± 6.1	59.7 ± 8.7	66.9 ± 7.7	58.4 ± 8.6	1.5
Null/ {Seq}+ΔG_{total}	65.4 ± 6.6	58.8 ± 12.9	48.4 ± 6.2	76.5 ± 10.8	1.4

Table 4.7: SVM performance when the dataset is pre-filtered according to ΔG_{ms}. It is followed by 4-fold cross-validation on the filtered half. For the Null model filter, a random half of the dataset is chosen for subsequent 4-fold cross-validation (to ensure the same number of data points as in ΔG_{ms} pre-filtering). Entries marked with (*) have one or more indeterminate values of PPV among the 10 runs; the averages reported are calculated on the remaining valid values. {} indicates that all of the features of the specified class were considered.

compared either between themselves, e.g., ΔG_{ms}/{Seq}+ΔG_{ms} of 65.0 vs Null/{Seq}+ΔG_{ms} of 58.8 or they should be compared to the corresponding 2-fold numbers. These comparisons keep sizes of the training sets the same or to at least roughly equal. Although fairly modest, the improvement of PPV because of pre-filtering results in a notable increase in ± Q. In addition, we see that adding one of the energy attributes to the sequence attributes in the feature space leads to better classification performance. This observation proves the concept that pre-filtering by incorporating prior knowledge of thermodynamic free energy signatures involved in the RNAi reaction can improve SVM classification performance.

4.3.3 Post-filtering by ΔG_{ms}

In this section, we present the results of our efforts to improve the SVM classification performance by post-filtering the predicted results. To this end, the data points predicted as efficient by the SVM, but having ΔG_{ms} values lower than the threshold (-8.9 kcal/mol) are reclassified as `inefficient`, and the performance of the classification scheme measured. We arrive at the threshold value by extrapolating from Fig. 4.2(b). Table 4.8 shows the results of these calculations. Each case is calculated 10 times.

Feature Space/Filter	Accuracy(%)	PPV(%)	Sensitivity(%)	Specificity(%)	IQ
{All}/ ΔG_{ms}	65.7 ± 4.7	63.8 ± 6.4	33.1 ± 7.8	87.5 ± 3.2	1.6
{Seq}/ ΔG_{ms}	64.3 ± 2.8	61.3 ± 11.5	27.1 ± 5.1	88.8 ± 5.8	1.5
{Seq}+ΔG_{mRNA}/ ΔG_{m_s}	64.5 ± 4.1	62.7 ± 12.1	32.3 ± 6.1	86.6 ± 5	1.6
{Seq}+ΔG_{siRNA}/ ΔG_{m_s}	66.6 ± 4	65.9 ± 10.5	33.3 ± 5.2	88.5 ± 4.5	1.7
{Seq}+ΔG_{ms}/ ΔG_{m_s}	65.9 ± 3.3	62 ± 11	34.9 ± 3	86.1 ± 4.2	1.6
{Seq}+$\Delta G_{complex}$/ ΔG_{m_s}	64 ± 2.6	64.7 ± 6.2	30.2 ± 5	88.1 ± 3.7	1.7
{Seq}+ΔG_{total}/ ΔG_{ms}	**64.6 ± 3.3**	**67.2 ± 10.4**	**32.8 ± 4.6**	**87.8 ± 5.6**	**1.8**

Table 4.8: SVM performance when the output is post-filtered according to ΔG_{ms}. Predicted efficient data points with ΔG_{ms} values greater than a threshold are marked inefficient and the performance of the SVM and post-filter together is determined. Entries marked with (*) have one or more indeterminate values of PPV among the 10 runs; the averages reported are calculated on the remaining valid values. {} indicates that all of the features of the specified class were considered.

We see that post-filtering can improve the result of SVM performance as well. However, it leads to higher standard deviation than in pre-filtering. This caveat should be considered

while designing classification models.

4.3.4 ΔG_{ms} as a Stand-alone Predictor of RNAi efficiency

Feature Space	Accuracy(%)	PPV(%)	Sensitivity(%)	Specificity(%)	IQ
ΔG_{ms}	57.5	48.7	59.8	55.9	1.2

Table 4.9: Result of using ΔG_{ms} as a stand-alone predictor of RNAi efficiency. siRNAs with corresponding ΔG_{ms} values greater than -8.9 kcal/mol are predicted to be functionally efficient.

In this section, we present the result of using ΔG_{ms} as a stand-alone predictor of RNAi efficiency by predicting the siRNA to be functionally efficient if the corresponding ΔG_{ms} value is greater than the threshold of -8.9 kcal/mol. We arrive at this threshold by extrapolating at efficiency of 70% from Fig. 4.2(b). Table 4.9 shows the results of this computation.

There is a small improvement over the Null model by using ΔG_{ms} value as a predictor of siRNA efficiency. However, it is obvious that ΔG_{ms} alone does not have nearly as much predictive power as the larger sets of features used here [See Table 4.5 and Table 4.6].

Chapter 5

Summary

RNA interference (RNAi) has attracted considerable interest in the research community for its potential as a viable, non-invasive, and safe technique of gene silencing. Key to understanding that potential is being able to design highly targeted and efficient siRNAs. To that end, characteristics of both the siRNA and the target mRNA have been explored and design criteria suggested. However, first order thermodynamic principles have proved inadequate in being able to predict reliably the siRNA silencing efficiency for a particular siRNA-mRNA pair. In this work, we present a very simplistic thermodynamic model of the RNAi reaction, and explore its effectiveness.

In our model, the antisense strand of the siRNA unravels and the free energy associated is ΔG_{siRNA}. The free energy required to create the break at the complementary mRNA target region is ΔG_{mRNA}. The free energy of formation of the mRNA–siRNA complex is $\Delta G_{complex}$. Thus, the total free energy involved in the reaction is $\Delta G_{total} = \Delta G_{complex} - \Delta G_{mRNA} - \Delta G_{siRNA}$. We expect the RNAi efficiency to be effected by these thermodynamic free energy signatures, which are calculated using `Mfold`, which considers equilibrium thermodynamic RNA secondary structures to predict free energies of formation of the secondary structures. Our expectation is that the more the energy required to create a break at the target region of the mRNA and to unravel the siRNA, the less efficient the RNAi reaction is; and the more energy released by formation of the siRNA–mRNA complex, the more favorable the reaction and hence the higher the silencing efficiency of the siRNA.

In addition, we use the `libsvm` implementation of support vector machine (SVM) to classify siRNAs as functionally efficient and inefficent using a combination of computed free energy characteristics and sequence characteristics of both the siRNA and the target mRNA. We try to improve the predictive performance of SVM by incorporating prior knowledge of the correlation between thermodynamic free energies and siRNA efficiency by pre- and post-filtering the datasets used to train and test the SVM.

As observed by others, we find no clear correlation between the free energy of the reaction

and silencing activity for individual data points. However, when we take the average of the computed free energies and the silencing efficiency, we find that there is a significant trend in general agreement with our expectations. The highest correlation is observed between the sum of ΔG_{mRNA} and ΔG_{siRNA}, $\Delta G_{ms} = \Delta G_{mRNA} + \Delta G_{siRNA}$. This represents the total free energy cost of unraveling the siRNA and creating the break at the complementary mRNA target strand region.

We conduct additional analysis to test the robustness of the findings to details of the computational protocol. The length of the local region `L` of influence of the mRNA and `cl`, the length of additional nucleotides that are forced single-stranded during the creation of the break at the mRNA target strand, are varied. We find that the general conclusion of the findings remain the same.

The strategy of using a smaller local region instead of the full mRNA sequence to compute free energy is not perfect because it potentially ignores the effect of the rest of the mRNA. We therefore take approximately a quarter of the data points (184) representing the shortest mRNAs in our dataset. The logic is that local region folding of short mRNAs will lead to less of the effects of the rest of the mRNA being ignored, and we could possibly observe higher correlation. We analyze this theory by varying `L` and `cl` for the shortest 184 data points. However, we do not find any increase in the correlation between the thermodynamic free energies and siRNA efficiency. Although it is possible the lack of correlation in these cases is because of lack of enough data points to average over in each bin, the observation does reduce the possibility of our earlier conclusion being an artifact of the particular choice of parameters.

In order to rule out possible differences arising from variation in experimental conditions among the various RNAi experiments from which the dataset is assembled, we perform our calculations on the two largest experimental subsets in our dataset containing 179 and 103 siRNA-mRNA pairs. In this case as well, we do not observer any increase in the correlation over our main result suggesting that our conclusion remains valid over possible differences in experimental conditions.

Finally, all the calculations till now have been done on the minimum free energy conformation of the secondary structure of the RNA sequences, while, in actual experimental conditions, RNA sequences fold in various conformations, their abundance in exponential proportion to the free energies of formation of the various conformations. We analyze our results by using the Boltzmann weighted average of thermodynamic free energies of formation of the various components in our model of the RNAi reaction. We find the results remain the same, largely because of the high correlation observed between the free energies calculated using Boltzmann average and those calculated using minimum free energy conformation secondary structures. Thus we conclude that the correlations observed between the free energies of the various components of RNAi reaction and the siRNA silencing efficiency is robust to computational details.

We use this knowledge to improve the predictive accuracy of an SVM to classify siRNAs

as functionally efficient and functionally inefficient. We divide the dataset into training set and test set in different ratios and conduct computational experiments with the full set of computed free energy and sequence specific characteristics to find that 4-fold cross validation provides good performance, at the same time with enough test data points to yield statistically significant result in case of further sub-division of the dataset in half according to some criteria. Thus we choose 4-fold cross validation for further analysis. We try various combinations of features in our feature space.

We try to further improve the performance by incorporating prior knowledge of the influence of thermodynamic free energies. Thus, we pre-filter the dataset in half according to ΔG_{ms}, and conduct 4-fold cross-validation on the filtered set. We observe that the SVM performs modestly better on the filtered dataset compared to the unfiltered set. It yields modest but consistent improvement on adding computed free energy characteristics to the set of sequence specific attributes while training and testing the SVM.

We also conduct post-filtering computations where we reclassify the predicted outputs of a 4-fold cross-validation exercise according to a threshold of the corresponding ΔG_{ms} value. Data points predicted to be functionally efficient but with ΔG_{ms} values lower than the threshold are reclassified as functionally inefficient and the performance metrics of the SVM calculated. We observe that post-filtering also yields modest but consistent improvement in prediction accuracy.

Based on these observations, we conclude that the silencing activity achieved by an siRNA-mRNA pair is influenced by the thermodynamics of the reaction. However, a lot of "noise" is introduced into the process. This can result from a number of factors in the reaction that we have not accounted for, e.g., the activity of Dicer and RISC. One of the reasons is that the thermodynamic influence of these factors is not well understood. In addition, if some of the steps of the reaction are achieved by expending energy in the form of ATP, then the basic thermodynamic model will fail to account for them, and reactions thermodynamically "bad" could produce excellent results. Of course, it is also possible our knowledge of the RNAi process remains incomplete, and there are vital factors which are unknown as yet. However, on average, the reactions behave as expected according to thermodynamic principles.

In spite of these *a priori* limitations, we are confident the basic thermodynamic model will help us better understand the interference reaction, and can serve as a useful guide in future endeavors to predict "good" siRNAs for an mRNA.

In the future, we hope to further improve the performance of machine learning techniques by incorporating a more rigorous modelling of the RNAi reaction, and by incorporating a larger feature space of siRNA and mRNA attributes in the feature space. In addition, we will use a more rigorous Boltzmann average calculation. We also propose to use SVM to rank the available set of siRNAs in terms of their silencing efficiency rather than trying to predict their actual efficiency.

Bibliography

[1] Fire, A., Xu, S., Montgomery, M., Kostas, S., Driver, S., and Mello, C. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391:806–811, 1998.

[2] Carthew, R. Gene silencing by double-stranded RNA. Current Opinion in Cell Biology 13:244–248, 2001.

[3] Ji, X. The Mechanism of RNase III Action: How Dicer Dices, volume 320. Springer-Verlag, , 2008.

[4] Hannon, G. J. RNA interference. Nature 418(6894):244–251, July, 2002.

[5] Moffat, J. and Sabatini, D. M. Building mammalian signalling pathways with RNAi screens. Nature Review Molecular Cell Biology 7:177–187, March, 2006.

[6] Ossowski, S., Schwab, R., and Weigel, D. Gene silencing in plants using artificial microRNAs and other small RNAs. The Plant Journal 53:674–690, 2007.

[7] Brown, K. M., Chu, C.-y., and Rana, T. M. Target accessibility dictates the potency of human RISC. Nature Structural and Molecular Biology 12:469–470, 2005.

[8] Schubert, S., Grünweller, A., Erdmann, V. A., and Kurreck, J. Local RNA target structure influences siRNA efficacy: systematic analysis of intentionally designed binding regions. Journal of Molecular Biology 348(4):883–893, May, 2005.

[9] Lu, Z. J. J. and Mathews, D. H. H. Fundamental differences in the equilibrium considerations for siRNA and antisense oligodeoxynucleotide design. Nucleic Acids Research 36(11):3738–3745, May, 2008.

[10] Patzel, V., Rutz, S., Dietrich, I., Köberle, C., Scheffold, A., and Kaufmann, S. H. E. Design of siRNAs producing unstructured guide-RNAs results in improved RNA interference efficiency. Nature Biotechnology 23(11):1440–1444, October, 2005.

[11] Holen, T., Moe, S. E., Sorbo, J. G., Meza, T. J., Ottersen, O. P., and Klungland, A. Tolerated wobble mutations in siRNAs decrease specificity, but can enhance activity *in vivo*. Nucleic Acids Research 33(15):4704–4710, 2005.

[12] Liao, J.-Y. Y., Yin, J. Q., Chen, F., Liu, T.-G. G., and Yue, J.-C. C. A study on the fundamental factors determining the efficacy of siRNAs with high C/G contents. Cellular and Molecular Biology Letters 13(2):283–302, June, 2008.

[13] Ameres, S. L., Martinez, J., and Schroeder, R. Molecular basis for target RNA recognition and cleavage by human RISC. Cell 130(1):101–112, July, 2007.

[14] Heilersig, H., Loonen, A., Bergervoet, M., Wolters, A., and Visser, R. Post-transcriptional gene silencing of GBSSI in potato: Effects of size and sequence of the inverted repeats. Plant Molecular Biology 60:647–662, 2006.

[15] Ichihara, M., Murakumo, Y., Masuda, A., Matsuura, T., Asai, N., Jijiwa, M., Ishida, M., Shinmi, J., Yatsuya, H., Qiao, S., Takahashi, M., and Ohno, K. Thermodynamic instability of siRNA duplex is a prerequisite for dependable prediction of siRNA activities. Nucleic Acids Research 35(18):e123, September, 2007.

[16] Pan, W. H. and Clawson, G. A. Identifying accessible sites in RNA: The first step in designing antisense reagents. Current Medical Chemistry 2006(13):3083–3103, 2006.

[17] Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. The role of site accessibility in microRNA target recognition. Nature Genetics 39(10):1278–1284, October, 2007.

[18] Westerhout, E. M. and Berkhout, B. A systematic analysis of the effect of target RNA structure on RNA interference. Nucleic Acids Research 35(13):4322–4330, 2007.

[19] Gredell, J. A. A., Berger, A. K. K., and Walton, S. P. P. Impact of target mRNA structure on siRNA silencing efficiency: A large-scale study. Biotechnology and Bioengineering 100(4):744–755, February, 2008.

[20] Shao, Y., Chan, C. Y., Maliyekkel, A., Lawrence, C. E., Roninson, I. B., and Ding, Y. Effect of target secondary structure on RNAi efficiency. RNA 13:1631–1640, October, 2007.

[21] Kurreck, J. siRNA Efficiency:Structure or Sequence–That Is the Question. Journal of Biomedicine and Biotechnology 2006:1–7, 2006.

[22] Pei, Y. and Tuschl, T. On the art of identifying effective and specific siRNAs. Nature Methods 3:670–676, September, 2006.

[23] Russell, P., Walsh, E., Chen, W., Goldwich, A., and Tamm, E. R. The effect of temperature on gene silencing by siRNAs: Implications for silencing in the anterior chamber of the eye. Experimental Eye Research 82:1011–1016, 2006.

[24] Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research 31(13):3406–3415, 2003.

[25] Hofacker, I. L. Vienna RNA secondary structure server. Nucleic Acids Research 31(13):3429–3431, July, 2003.

[26] Mathews, D., Sabina, J., Zuker, M., and Turner, D. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of Molecular Biology 288:911–940, 1999.

[27] Zuker, M. and Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research 9:133–148, 1981.

[28] Parisien, M. and Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature 452(7183):51–55, 2008.

[29] Harmanci, A. O., Sharma, G., and Mathews, D. H. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. BMC Bioinformatics 8(1):130, April, 2007.

[30] Freyhult, E., Moulton, V., and Clote, P. RNAbor: A web server for RNA structural neighbors. Nucleic Acids Research 35:W305–309, 2007.

[31] Martinez, H. M., Maizel, J. V., and Shapiro, B. A. RNA2D3D: A program for Generating, Viewing, and Comparing 3-Dimensional Models of RNA. Journal of Biomolecular Structure and Dynamics 25(6):669–683, 2008.

[32] Jiang, P., Wu, H., Da, Y., Sang, F., Wei, J., Sun, X., and Lu, Z. RFRCDB-siRNA: Improved design of siRNAs by random forest regression model coupled with database searching. Computer Methods and Programs in Biomedicine 87:230–238, 2007.

[33] Vařeková, R. S. S., Bradáč, I., Plchút, M., Skrdla, M., Wacenovsky, M., Mahr, H., Mayer, G., Tanner, H., Brugger, H., Withalm, J., Lederer, P., Huber, H., Gierlinger, G., Graf, R., Tafer, H., Hofacker, I., Schuster, P., and Polčík, M. www.rnaworkbench.com: A new program for analyzing RNA interference. Computer Methods and Programs in Biomedicine 90(1):89–94, January, 2008.

[34] Lu, Z. and Mathews, D. OligoWalk: An online siRNA design tool utilizing hybridization thermodynamics. Nucleic Acids Research 36, web server issue:W104–W108, July, 2008.

[35] Gong, W., Ren, Y., Xu, Q., Wang, Y., Lin, D., Zhou, H., and Li, T. Integrated siRNA design based on surveying of features associated with high RNAi effectiveness. BMC Bioinformatics 7:516, November, 2006.

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

[37] Amarzguioui, M. and Prydz, H. An algorithm for selection of functional siRNA sequences. Biochemical and Biophysical Research Communications 316(4):1050–1058, February, 2004.

[38] Vert, J.-P., Foveau, N., Lajaunie, C., and Vandenbrouck, Y. An accurate and interpretable model for siRNA efficacy prediction. BMC Bioinformatics 7:520–536, November, 2006.

[39] Ladunga, I. More complete gene silencing by fewer siRNAs: transparent optimized design and biophysical signature. Nucleic Acids Research 35(2):433–440, January, 2007.

[40] Peek, A. S. Improving model predictions for RNA interference activities that use support vector machine regression by combining and filtering features. BMC Bioinformatics 8:182–201, June, 2007.

[41] Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, 144–152 (ACM, New York, NY, USA, 1992).

[42] Smith, F. W. Pattern classifier design by linear programming. IEEE Transactions on Computers 17(4):367–372, 1968.

[43] Aizerman, A., Braverman, E. M., and Rozoner, L. I. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control 25:821–837, 1964.

[44] Shabalina, S. A., Spiridov, A. N., and Ogurtsov, A. Y. Computational models with thermodynamic and composition features improve siRNA design. BMC Bioinformatics 7(1):65, 2006.

[45] Lu, Z. J. and Mathews, D. H. Efficient siRNA selection using hybridization thermodynamics. Nucleic Acids Research 36(2):640–647, February, 2008.

[46] Harborth, J., Elbashir, S. M., Vandenburgh, K., Manninga, H., Scaringe, S. A., Weber, K., and Tuschl, T. Sequence, chemical, and structural variation of small interfering RNAs and short hairpin RNAs and the effect on mammalian gene silencing. Antisense and Nucleic Acid Drug Development 13:83–105, 2003.

[47] Chang, C. C. and Lin, C. J. LIBSVM: A library for support vector machines, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[48] Khvorova, A., Reynolds, A., and Jayasena, S. Functional siRNAs and miRNAs Exhibit Strand Bias. Cell 115:209–216, October, 2003.

[49] Hsieh, A., Bo, R., Manola, J., Vazques, F., Bare, O., Khvorova, A., Scaringe, S., and Sellers, W. 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.