An Ensemble SVM Model for the Accurate Prediction of Non-Canonical MicroRNA Targets

Asish Ghoshal*
Department of Computer Science
Purdue University
West Lafayette, IN, USA
aghoshal@purdue.com

Ananth Grama
Department of Computer Science
Purdue University
West Lafayette, IN, USA
ayg@purdue.com

Saurabh Bagchi
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN, USA
sbagchi@purdue.edu

Somali Chaterji†
Department of Computer Science
Purdue University
West Lafayette, IN, USA
schaterj@purdue.edu

ABSTRACT

Background MicroRNAs are small non-coding endogenous RNAs that are responsible for post-transcriptional regulation of genes. Given that large numbers of human genes are targeted by microRNAs, understanding the precise mechanism of microRNA action and accurately mapping their targets is of paramount importance; this will uncover the role of microRNAs in development, differentiation, and disease pathogenesis. However, the current state-of-the-art computational methods for microRNA target prediction suffer from high false-positive rates to be useful in practice.

Results In this paper, we develop a suite of models for microRNA target prediction, under the banner *Avishkar*, that have superior prediction performance over the state-of-the-art protocols. Specifically, our final model developed in this paper achieves an average true positive rate of more than 75%, when keeping the false positive rate of 20%, for non-canonical microRNA target sites in humans. This is an improvement of over 150% in the true positive rate for non-canonical sites, over the best competitive protocol. We are able to achieve such superior performance by representing the thermodynamic and sequence profiles of microRNA-mRNA interaction as curves, coming up with a novel metric of seed enrichment to model seed matches as well as all possible non-canonical matches, and learning an ensemble of microRNA family-specific non-linear SVM classifiers. We provide an easy-to-use system, built on top of Apache Spark, for large-scale interactive analysis and prediction of microRNA targets. All operations in our system, namely candidate set generation, feature generation and transformation, training, prediction and computing performance metrics are fully distributed and are scalable.

Availability All source code and sample data is available at `https://bitbucket.org/cellsandmachines/avishkar`. We also provide scalable implementations of kernel SVM using Apache Spark, which can be used to solve large-scale non-linear binary classification problems at `https://bitbucket.org/cellsandmachines/kernelsvmspark`.

Categories and Subject Descriptors

I.2.1 [**Applications and Expert Systems**]: [Medicine and science]; J.3 [**Life and Medical Sciences**]: [Biology and genetics]; G.3 [**Probability and Statistics**]: [Statistical computing, Nonparametric statistics]

General Terms

Algorithms

Keywords

MicroRNA, target prediction, non-canonical matches, mRNA, large-scale, kernel SVM, distributed machine learning, Apache Spark

1. INTRODUCTION

MicroRNAs (miRNAs) are short, approximately 22 nucleotide (nt) long, endogenous, non-coding RNAs that are central to the post-transcriptional regulation of genes [1]. MicroRNAs associate with Argonaute (AGO) proteins, mediating RNA interference (RNAi) by targeting the 3' UTR of the mRNA, or in some cases, other mRNA regions, such as the mRNA's coding sequence (CDS) or its 5' UTR [4]. There are over two thousand miRNAs that have been annotated in humans, displaying many-to-many associations with

*Corresponding author.
†Corresponding author.

BCB '15 Atlanta, GA USA
 ISBN 978-1-4503-3853-0/15/09 ...$15.00
DOI: http://dx.doi.org/10.1145/2808719.2808761.

mRNA targets [7]. Given that miRNAs regulate across the spectrum of *in vivo* biological processes, their aberrant expression or perturbation of their regulatory activities results in disease [18, 12].

Notwithstanding the biological importance of miRNAs, determining their targets with high accuracy and exhaustively is a hard problem, with computational predictions plagued by high false-positive and false-negative rates [27] and laboratory validations being time-consuming and expensive. This complexity of miRNA target prediction can be attributed to the small size of miRNAs, requiring as few as six complementary base pairs for functional miRNA interactions, as well as the diversity of the miRNA interactome [6].

Recent experimental approaches allow for the identification of AGO-miRNA:mRNA ternary complexes using an *in vivo* cross-linking protocol followed by high-throughput sequencing—CLIP-seq, a state-of-the-art experimental method toward developing genome-scale regulatory insights [21]. The technology allows a high-resolution investigation of the occupancy of the RISC-miRNA protein complexes on their complementary mRNA within a small window of resolution. Beyond this window, computational models are required for localizing the binding site, also known as the miRNA recognition element (MRE). Further, while CLIP-seq can identify miRNAs and targets that form a part of the RISC complex, it cannot decipher *which* miRNA forms heteroduplex with *which* targets; although limited advances have been made toward experimentally solving this problem [10]. Several computational methods developed to decipher the specifics of miRNA-mRNA interactions, captured by CLIP-seq [3, 22, 15, 24], have contributed to our understanding of the diverse miRNA targetome. The evolving knowledge base of this targetome has further supported the paradigm switch, wherein it is now widely accepted that the perfect complementarity between the miRNA seed[1] and mRNA 3' UTR is neither necessary nor sufficient for miRNA regulation.

In this paper, we leverage this ability of the CLIP-seq technology to capture endogenous MREs to develop a unified method for understanding the signatures of miRNA-mRNA heteroduplexes focusing on non-canonical matches. This focus is apt because prior work has often neglected this class which accounts for a majority of the target sites—92% of all target sites in humans and 94% in mouse. Specifically, in our system, which we call *Avishkar*, we solve the classification problem of whether a miRNA targets an mRNA region. Toward this end, we use smooth B-spline thermodynamic curves and sequence curves for adenosine-uracil (AU) content, in order to extract enriched interaction features from the experimentally immunoprecipitated regions. We then use a support vector machine (SVM)-based machine learning (ML) system to learn the diverse signatures of this CLIPed (immunoprecipitated), miRNA targetome. We show improved performance (in terms of true positive and false positive rates) over all prior work. The reasons behind this improvement are the use of an extensive set of features, incorporating the spatial nature of the miRNA-mRNA binding process through various thermodynamic and sequence curves and in the process converting noisy data points into smooth curves, converting the categorical feature of seed or non-canonical match into a numerical feature and treating both seed and seedless sites under one unified umbrella.

Our main contributions in this paper are as follows:

1. We develop a simple linear classifier that achieves an average true positive rate (TPR) of 47% at a false positive rate (FPR) of 20%. The simple global linear model achieves similar prediction performance on CLIP-seq data for mouse as well as for inter-species training and prediction. The linear model outperforms the state-of-the-art for both canonical and non-canonical target sites in humans and mouse. Specifically, the Area-Under-the-Curve (AUC) for human canonical and non-canonical target sites is 19.7% and 22.0% better than the state-of-the-art. The AUC improvements for the mouse data set are 15.0% and 22.8% over the state-of-the-art for canonical and non-canonical sites respectively.

2. Observing that the linear classifier has a moderate amount of bias, we develop non-linear classifiers for the problem. Inspired by previous categorization of miRNAs into multiple families with structural similarities among family members, we proceed to create an ensemble of family-specific models. This achieves an average TPR of 76% at a FPR of 20% for predicting non-canonical miRNA target sites, compared to 47% for a linear SVM model. The AUC for the ensemble non-linear model is 20% higher than for the simple linear model.

3. Since training non-linear SVMs is computationally expensive, we provide a general-purpose and efficient implementation of a popular algorithm for parallel training of SVMs (from 2004), called Cascade SVMs [8], on top of Apache Spark [32]. We open source our implementation and believe this could be of use to the bioinformatics community that deals with the problem of classification of large data sets with complex separating boundaries.

The rest of the paper is organized as follows. In Section 2, we survey related work. Section 3 describes our method, and in Section 4 we present our results. Finally, we discuss potential limitations of our approach and future directions in Section 5 before presenting our conclusions in Section 6.

2. RELATED WORK

Among non-canonical prediction methods, mirSVR [3] allows for a single GU wobble or a mismatch in the 6-mer seed region. For encoding the seed match pattern, mirSVR uses an 8-bit long vector, with "1" representing a match and "0" a mismatch and then uses this bit-vector as a feature in their Support Vector Regression (SVR) model. Recent methods have expanded the target search to other areas of the gene, such as to the 5' UTR and coding sequence (CDS) [22, 30]. In this bracket, Liu *et al.* generate predictions for

[1]Nucleotide positions 2-7, and sometimes 2-8, from the 5' end of the miRNA is generally referred to as the seed region of the miRNA. A target region is said to have a seed match if there is a continuous pairing of mRNA nucleotides with the seed region of the miRNA, and the target site is called as a canonical or seed-match site. Other target sites that don't have continuous base pairing with the seed region of the miRNA are called non-canonical or seedless sites. In this paper we refer to the pattern of non-canonical alignment of the miRNA seed region with the mRNA nucleotides as the "non-canonical seed-match pattern".

Table 1: CLIP data used for training and prediction in *Avishkar*. Very few positive target sites are located in the 5' UTR region

	# Positive examples (Seed:Seedless)	# Negative examples	# mRNA	# miRNA	# Positive target sites in 3' UTR	CDS	5' UTR
HITS-CLIP (Mouse)	861,208 (6%:94%)	35,608,333	4,059	119	478,138 (≈ 56%)	367,371 (≈ 43%)	15,699 (≈ 1%)
PAR-CLIP (Human)	141,109 (8%:92%)	2,659,748	1,211	35	80,775 (≈ 57%)	55,250(≈ 39%)	5,084 (≈ 4%)

Table 2: Comparison of the average number of candidate target sites per miRNA-mRNA pair considered by various methods. mirSVR [3], PITA [14] and TargetScan[9] only consider the 3' UTR region. Therefore, the average number of candidate target sites is very low for these two methods. It should be noted that very few candidate sites are functional miRNA target sites. However, the larger the initial candidate set size is, the closer the method is to doing a genome-wide search for miRNA targets.

	mirSVR	PITA	TargetScan	STarMir	***Avishkar***
Human	1.256	3.078	NA	56.183	66.081
Mouse	0.56	2.179	0.318	37.418	75.503

non-canonical sites, but they do not take into consideration the type of non-canonical seed-match patterns for the examined sites. Instead, they use thermodynamic and mRNA sequence features (*e.g.,* local AU content) to generate predictions for the seedless sites. In doing so, they miss out on the potential signal from the non-canonical seed-match patterns that our findings indicate as enriched in the identified functional miRNA-mRNA interactions. One possible reason for this, as pointed out by Xu *et al.* [31], is the difficulty of incorporating large numbers of possible patterns of insertions and deletions in the match patterns.

Some computational methods have exclusively relied on thermodynamic features, such as the stability of the miRNA-mRNA heteroduplex and the structural accessibility of the mRNA target region to identify functional miRNA binding sites. For example, Xu *et al.* [31] only use binding energy and site accessibility to predict functional miRNA target sites. Another recent method, MIRZA [15] develops a rigorous biophysical model via parameterizing the alignment between a miRNA and an mRNA segment, interpreted as the binding energy between the two, and optimizing this using CLIP data. While, MIRZA uses a novel model to incorporate canonical and non-canonical matches in a unified manner, it does not take into account mRNA secondary structures, which are known to modulate miRNA binding [14]. Secondary structures of an mRNA can potentially limit the target site accessibility to the docking miRNA-RISC complex and therefore plays an important role in miRNA target recognition. Further, these approaches compute various thermodynamic scores only at the target site region to summarize the thermodynamics of the miRNA-mRNA interaction. Liu *et al.* [22] compute site accessibility in the target region's vicinity in discrete chunks of 5-30 nt, in increments of 5, around the target site region. However, they report only the accessibility computed at the target site to be an important predictor of functional miRNA targets. In *Avishkar*, we remove these shortcomings through the use of spatial curves to profile the miRNA-mRNA interactions at *and in the vicinity* of the target site, in order to extract more signal out of these thermodynamic and sequence features, ΔG and $\Delta\Delta G$. These thermodynamic features are defined as follows: *Gibbs free energy* (ΔG) is used as a measure of the stability of a biophysical system and in the case of an mRNA-miRNA interaction quantifies the stability of the mRNA-miRNA heteroduplex. *Site accessibility* ($\Delta\Delta G$) measures the accessibility of the mRNA target site to the miRNA by computing the "difference between free energy gained from the formation of the miRNA-target duplex (ΔG) and the energetic cost of unpairing the target (ΔG_{open}) to make it accessible to the miRNA" [14].

3. METHODS

In the following paragraphs we describe the data sets used in our system, our overall solution approach and also the various models we developed to predict miRNA targets.

3.1 Data

PAR-CLIP [17] data for the human cell line HEK 293 was downloaded from Gene Expression Omnibus (GEO)[2]. The data contained 190,764 AGO binding sites across 10,159 different mRNAs. As in [17], we used the ten most-abundantly expressed miRNA families, consisting of 44 different miRNAs, in human HEK 293 cells, for our bioinformatic analysis. Since feature computation (described next) for all possible miRNA-mRNA pairs was expensive, we randomly selected around 1,200 mRNAs for analysis.

HITS-CLIP [5] data for mouse brain tissue was downloaded from starBase database [20], which contained 11,117 AGO-CLIP tag clusters, across the mouse genome (mm9) assembly. Following the approach in [5], we used the twenty most abundant miRNA families, containing 119 miRNAs and over 4000 mRNAs for our bioinformatic analysis.

Human and mouse PhastCons [28] conservation scores were dowloaded from the UCSC Genome browser. We used the mouse conservation scores that were generated by alignment of 30 vertebrate genomes to the mouse genome (mm9 assembly), while human conservation scores were generated by alignment of 44 vertebrate genomes to the human genome (hg18 assembly). Table 1 summarizes the data sets used in our model.

3.2 Approach

The general approach for identifying miRNA targets in different mRNA regions has been to first identify a set of candidate target regions across the whole genome using various rules [30, 22, 3, 14]. Labels are then assigned to candidate target sites from experimental data and the labelled data is used to train (or fit) a model. Previous approaches

[2]Series GSE28865 and accession codes GSM714642, GSM714644, and GSM714646.

have relied on enforcing a minimum threshold on the alignment score between the miRNA and an mRNA segement [3], or a minimum threshold on the binding energy (ΔG) of the mRNA-target duplex [30, 22, 14], or constraining the target region to be in certain mRNA regions, such as the 3' UTR region [3]. Previous approaches have also used seed-match rules [3] and evolutionary conservation to filter miRNA target regions. In this paper, we use the least restrictive filters to generate the initial set of candidate target sites. Specifically, we enforce a minimum threshold of -15 $kcal/mol$ on the binding energy (ΔG). To include seed-match sites that may have a binding energy above the threshold of -15 $kcal/mol$, we additionally include candidate target sites having a seed match, *i.e.,* continuous base pairing with nucleotides 2-7 of the miRNA from the 5' end. The average number of candidate target sites per mRNA-miRNA pair for our method is largest among competition (Table 2).

After generating the candidate set of target sites, we label each site as 1 (functional miRNA-mRNA interaction) or 0 (not an interaction), depending on whether or not the candidate site is contained within an IP region, as obtained from the CLIP data sets. *Thus, we deem miRNA-mRNA interactions to be functional if the target site is contained within an immunoprecipitated (IP) region and either the binding energy (ΔG) of the duplex is above a cutoff value or there is a seed match.* The data generated as such serves as the ground truth against which we train and evaluate the performance of our method and also evaluate the performance of competition. We develop four models for miRNA target prediction: a global linear model, a global non-linear model, an ensemble of miRNA family-specific linear models, and an ensemble of miRNA family-specific non-linear models. Out of the four, the global non-linear model did not scale to the size of the data used in this paper. Therefore, we present results for the other three models.

3.3 Features for classifier

Table 3 summarizes the features used for the classifiers in this paper. While, some of the features have been proposed in the miRNA target-prediction literature, our main contributions are the functional features listed in the first six rows and the seed enrichment metric that enables us to accurately predict non-canonical targets and also discover a whole gamut of biologically functional non-canonical seed-match types. We describe those features in subsequent sections.

3.3.1 Using curves to capture spatial interaction

Thermodynamic stability of the miRNA-mRNA duplex and accessibility of the target site nucleotides have been known to be important predictors of miRNA targeting [2]. However, most of the previous methods quantify thermodynamic stability and accessibility using a single number (scalar covariate) computed at the target site, with additional flanking nucleotides [14, 3, 30]. Recently, there has been some evidence that accessibility follows a certain pattern *around* the mRNA target site [31]. **Hence, we come up with thermodynamic and accessibility curves in order to create a profile of stability and accessibility in and around the target site.** We do this by sliding the miRNA, both upstream and downstream of the target site, and computing the quantities ΔG and $\Delta\Delta G$ at various points along the mRNA, in the neighborhood of the target sites. However, these quantities are *estimated* by using computational tools— RNAhybrid [19] and RNAfold [23]— which are themselves based on approximate models. Thus, this approximation can result in inaccurate and noisy values. Thus, we fit smooth curves through these vector observations in order to develop a non-parametric characterization of thermodynamic and accessibility curves. Figure 1 shows an example of such curves. Specifically, if $\mathbf{f}$ is the vector of ΔG or $\Delta\Delta G$ values computed at various points in and around the mRNA target site, then the smooth curve $f(t)$, for a given data point, is obtained as:

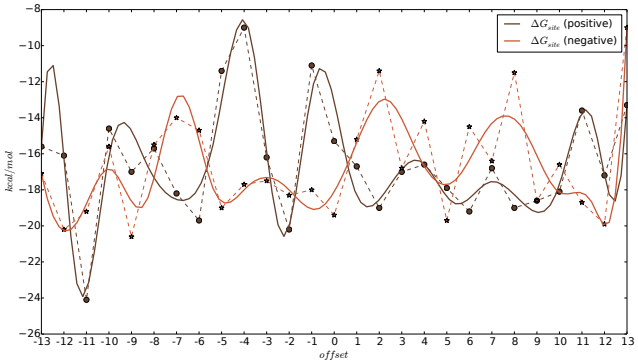

Figure 1: Examples of smooth curves for a positive and negative (non-functional) mRNA-miRNA interaction.

$$f(t) = \sum_{j=1}^{K} c_j \psi_j(t) \tag{1}$$

where $\psi_j(t)$ are the cubic B-spline basis functions. We use the zero value to replace missing values in the vector $\mathbf{f}$, *e.g.,* when the target site is toward the beginning or the end of the mRNA. Now, consider that we have to fit curves for each of the N data points, and thus, we have to estimate the coefficients $c_{i,j}$ for $1 \leq i \leq N$ and $1 \leq j \leq K$ for each curve, by minimizing the least squares error on the discrete observations $\mathbf{f}_i$, as follows:

$$\mathbf{c}_i = (\mathbf{\Psi}^T \mathbf{\Psi})^{-1} (\mathbf{\Psi}^T \mathbf{f}_i) \tag{2}$$

where $\mathbf{\Psi}$ is the $(2W+1) \times K$ matrix of the K basis functions evaluated at $\{t : t \in \mathbb{Z}, 0 \leq t < (2W+1)\}$. We compute these curves at two different resolutions. First, we use a window size of 46 nucleotides that is centered at the mRNA target site region, and second, we use a window size of 9 nucleotides that is centered at the seed-matched region, within the target region. The two types of curves computed as such are referred to as *site curves* and *seed curves*. The rationale underlying computing the seed curves is to incorporate those target sites, where the miRNA seed region (of maximum length 8), alone, is responsible for targeting. In such cases, computing the curves by using a larger window size might miss the finer variations in structural and thermodynamic stability and accessibility. We also extend

[3] [2] enumerate seven different types of seed match patterns, out of which three are canonical seed-match patterns and four are non-canonical or atypical patterns. Canonical and non-canonical seed-match patterns are believed to have different efficacies for miRNA targeting.

Table 3: Summary of features used in our model. The first six are functional covariates (curves) which are obtained by fitting a smooth curve through the vector observations, indicated by bold-faced letters. The rest are scalar covariates. For functional features, the curves are calculated over multiple points—at the target, and separately, 13 upstream and downstream sites.

$\Delta G_{site}(t)$	Thermodynamic binding curve centered at the target site obtained by fitting a smooth curve through the vector observation $\mathbf{\Delta G}^{site}$.
$\Delta G_{seed}(t)$	Finer resolution thermodynamic binding curve centered at the seed match region obtained by fitting a smooth curve through the vector observation $\mathbf{\Delta G}^{seed}$.
$\Delta\Delta G_{site}(t)$	Accessibility curve centered at the target site obtained by fitting a smooth curve through the vector observation $\mathbf{\Delta\Delta G}^{site}$.
$\Delta\Delta G_{seed}(t)$	Finer resolution accessibility curve centered at the seed match region obtained by fitting a smooth curve through the vector observation $\mathbf{\Delta\Delta G}^{seed}$.
$AU_{site}(t)$	Local AU content curve centered at the target site region obtained by fitting a smooth curve through vector observation $\mathbf{AU}^{site}$.
$AU_{seed}(t)$	Finer resolution local AU content curve computed at the seed match region obtained by fitting a smooth curve through vector observation $\mathbf{AU}^{seed}$.
seed enrichment	A scalar feature indicating the extent to which a seed match pattern, both canonical and non-canonical, is enriched in the set of positive miRNA-mRNA interactions on a scale of 0 to 1.
site conservation	The extent to which the mRNA site nucleotides are conserved across different species.
seed conservation	The extent to which the nucleotides in the mRNA site that are paired with the miRNA seed region are conserved across different species. This is only used when there is a canonical seed match.
off seed conservation	Average conservation score of mRNA nucleotides that are not paired with the seed region of the miRNA. This is only used when there is a canonical seed match.
target site length	Length of the mRNA target site
target region	mRNA region where the target site is present, namely, 3' UTR, CDS or 5' UTR
relative position of target site	Relative position of a target site within one of the 3 regions above on a scale of 0 to 1, with 0 indicating the 5' end and 1 indicating the 3' end.

the idea of curves to another feature—content of adenine and uracil nucleotides (AU content)—to come up with site and seed *sequence curves*. All of these curves are used as functional covariates in our classifier. We incorporate these curves as covariates in our model via using the coefficients of their B-spline basis functions as features into the classifier. The number of basis functions control the smoothness of the curves, the optimal value of which is chosen using 10-fold cross-validation. Note that we use the same number of basis functions for all the curves in order to reduce the computational cost of the exploration of the parameter search space during cross-validation, which is already computationally expensive.

3.3.2 Incorporating non-canonical seed matches

Earlier methods for miRNA target prediction have mostly considered canonical seed matches and only a few types of non-canonical seed matches [3, 14, 30]. However, there is increasing evidence that a number of non-canonical seed matches with long bulges are biologically functional [11]. Therefore, in order to incorporate all kinds of non-canonical seed matches in a unified and systematic manner, we came up with the following approach. We first pre-compute the occurrence frequencies of various kinds of seed-match patterns, in the set of positive mRNA-miRNA interactions, by representing the alignment of mRNA nucleotides with miRNA nucleotides 1 to 8 (from the 5' end) as a vector of 1's (match), 2's (mismatch), 3's (alignment to gap), and 4's (GU wobble). We want to distinguish chance occurrences from those that occur above background levels. Let the pattern $\mathbf{a}$ have, say, k occurrences among n positive samples. Then, the probability of the pattern occurring purely due to chance is $0.25^{|\mathbf{a}|}$. Thus, the likelihood of the pattern occurring k times is given by:

$$\alpha = Binomial(k|n, 0.25^{|\mathbf{a}|}). \quad (3)$$

We use $1-\alpha$ as the seed enrichment score for the pattern α. In the feature transformation phase, we replace a particular pattern of seed match by its enrichment score. In doing so, we can represent both canonical and all possible patterns of non-canonical seed matches, using a single numerical feature. We discovered that a whole gamut of non-canonical seed-match patterns with long bulges, that is sequence of 3's, are enriched in both humans and mouse miRNA-mRNA interactions. Further, to speedup feature generation and transformation, we parallelize both operations in our system by distributing the mRNAs across different processors in the cluster and computing, or transforming, the features in parallel. This is accomplished by using Apache Spark [32].

3.4 Improving prediction performance

While our global linear model outperformed state-of-the-art methods for identifying functional canonical and non-

canonical target sites, it still suffered from relatively high FPRs due to high bias of the linear model, as was evident from the high training error (misclassification rate) of the model. Therefore, in order to improve the prediction performance, we decided on learning a more complex model by using non-linear kernels *Avishkar*. However, there were two challenges in using non-linear kernels for our problem. First, the data set size ($\approx 260,000 \times 130$) was too big to use an exact SVM solver for training the model, [4] and second, the problem instance was such that the ratio of number of support vectors to data points, needed to be quite high to get low generalization errors—something that we discovered experimentally. In order to solve the first problem, we started out with SVM training algorithms that scale well to large data sets, at the expense of approximately solving the optimization problem. Toward that end, we used budgeted SVM [29], which is an online algorithm that uses stochastic gradient descent to minimize the SVM loss function. Since, the number of support vectors tends to increase linearly with data set size, the algorithm caps the number of support vectors to a user-specified budget. This ensures that updating the model does not become too expensive as the algorithm sees more and more data points. However, this procedure did not work very well for our problem instance. For example, selecting a low budget of 8,000 resulted in an average misclassification error of around 45%, while selecting 80,000 support vectors resulted in the model not completing in 5 hours. Note that while this might not appear to be too slow, we wanted a model that could finish within at most an hour so that parameter tuning, averaging the results over multiple sub-samples of the negative examples, and cross-validation runs could be done within a reasonable amount of time.

Thus, in order to effectively tackle the two challenges described above, we came up with the following solution. *We first clustered the data points by the miRNA family that the miRNA belongs to and then we trained an ensemble of cluster-specific SVM models.* The advantage of learning miRNA family-specific models is twofold. First, miRNAs belonging to the same family have similar structural or sequence configuration [13]. Therefore, binding patterns of miRNAs within the same family might be similar. This reduces the variability of covariates within clusters, and thus, the number of support vectors required to classify the data. Second, dividing the data set into independent clusters, results in significant computational savings, wherein models for separate clusters can be trained independently and in parallel. This is especially useful for the Cascade SVM approach, wherein the later stages tend to be slow due to reduced parallelism. Since some clusters were fairly large (Table 4), we use the Cascade SVM approach from [8] to parallelize the process of learning the support vectors. We developed (and open sourced) a general-purpose, scalable, and memory efficient implementation of Cascade SVM, on top of Apache Spark [32] that can be used to train kernel SVMs for large problems. In summary, we train kernel SVMs for different clusters in parallel and, within each cluster, we further parallelize SVM training via implementing the Cascade SVM approach, which we describe next.

3.4.1 Cascade SVM

[4]General-purpose, quadratic-programming solvers that are used to optimize the dual formulation of SVMs, scale as cube of the number of training vectors in the data set $\mathcal{O}(N^3)$

The primal objective function for SVM is given by the following equation:

$$L(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \frac{1}{n}\sum_{n=1}^{N} max(0, 1 - y_n \mathbf{w}^T \phi(\mathbf{x}_n)) + \frac{\lambda}{2}||\mathbf{w}||_2^2. \tag{4}$$

In the above equation, data points $(\mathbf{x}_n, y_n)$ that are on the correct side of the decision hyperplane have $y_n \mathbf{w}^T \phi(\mathbf{x}_n) \geq 1$ and are not penalized. λ is the regularization parameter that is used to penalize complex models. The parameter γ controls the radius of influence of a data point, with higher values corresponding to a lower radius of influence, resulting in a larger number of data points becoming support vectors, which in turn results in a more complex model. The dual formulation is given by the following equation, which is obtained by applying the kernel trick.

$$\begin{aligned} \min\ L(\mathbf{a}) &= \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m y_n y_m \mathcal{K}(\mathbf{x}_n, \mathbf{x}_m) - \sum_{n=1}^{N} a_n \\ &= \mathbf{a}^T \mathbf{Q} \mathbf{a} - \mathbf{e}^T \mathbf{a} \\ \text{subject to } 0 \leq a_i \leq C\ \forall\ i \text{ and } &\sum_{i=1}^{N} a_i y_i = 0. \end{aligned} \tag{5}$$

In the above equations, $\mathcal{K}$ is the kernel function, $\phi(\mathbf{x})$ the implicit feature map introduced by the kernel function, $\mathbf{Q}_{n,m} = (1/2) y_n y_m \mathcal{K}(\mathbf{x}_n, \mathbf{x}_m)$, and $\mathbf{e}$ is a vector of 1's. The parameter C controls the number of points that are misclassified, with larger values of C corresponding to fewer mis-classified points, reduced margin, and a more complex model. We use the radial basis function (RBF) kernel in our system, which is given by:

$$\mathcal{K}(\mathbf{x}_n, \mathbf{x}_m) = \exp(-\gamma \|\mathbf{x}_n - \mathbf{x}_m\|). \tag{6}$$

We use the SVM implementation provided by scikit-learn [25], which uses direct optimization of the dual formulation of SVM. As mentioned eariler, we use the Cascade SVM approach from [8] to parallelize the process of learning the support vectors. The basic approach is schematically represented in Figure 2. The training data set is first split into a number of segments. Then, an SVM is trained independently for each of the segments. Since the support vectors in each segment might not be global support vectors, the support vectors from two segments are combined by passing them through another SVM to filter out non-support vectors. This proceeds in a tree-like, hierarchical manner, until only one set of support vectors remain. The support vectors can then be fed back to the first layer and multiple iterations over the cascade of SVMs is guaranteed to take the solution to the global optimum, and often only one iteration over the cascade produces a sufficiently good solution. If $\mathbf{a_1}$ and $\mathbf{a_2}$ are two sets of support vectors from two different SVMs, then the authors in [8] elucidate two ways of combining the support vectors and initializing the objective function. The combined coefficients for the support vectors can either be set to $\mathbf{a}^* = [\mathbf{a}_1^T\ \mathbf{a}_2^T]^T$ or $\mathbf{a}^* = [\mathbf{a}_1^T\ \mathbf{0}]^T$, the first represents the case where the two subsets are identical and the second where the two subsets are orthogonal. We, however, initialize the objective function for the combined SVM to have all coefficients as zero (the default option in scikit-learn). This may make finding solution for the combined SVM a bit slow

Table 4: Enumeration of the ten different miRNA families in the human HEK 293 cell line, representing the most numerous families, their constituent miRNAs, and the number of positive miRNA-mRNA pairings in each family. The total number of samples used in the model building for each cluster is roughly twice the number of positive examples in each family because we sub-sample from the negative samples to keep the number of positive and negative samples approximately equal.

ID	miRNA family	miRNAs	# Positive examples in cluster
1	miR-7	hsa-miR-7-5p	3061
2	miR-25	hsa-miR-25-3p	3577
3	miR-103	hsa-miR-103a-3p	4317
4	miR-15	hsa-miR-15a-5p, hsa-miR-15b-5p	6355
5	miR-10	hsa-miR-10a-5p, hsa-miR-10b-5p	4239
6	miR-106	hsa-miR-106a-5p, hsa-miR-106b-5p	9563
7	miR-19	hsa-miR-19a-3p, hsa-miR-19b-3p, hsa-miR-195-5p	6827
8	miR-320	hsa-miR-320a, hsa-miR-320b, hsa-miR-320c, hsa-miR-320d	24986
9	miR-30	hsa-miR-30b-5p, hsa-miR-30c-5p, hsa-miR-30d-5p, hsa-miR-30e-5p	8559
10	let-7	hsa-let-7a-5p, hsa-let-7b-5p, hsa-let-7d-5p, hsa-let-7e-5p, hsa-let-7f-5p, hsa-let-7g-5p, hsa-let-7i-5p	36290

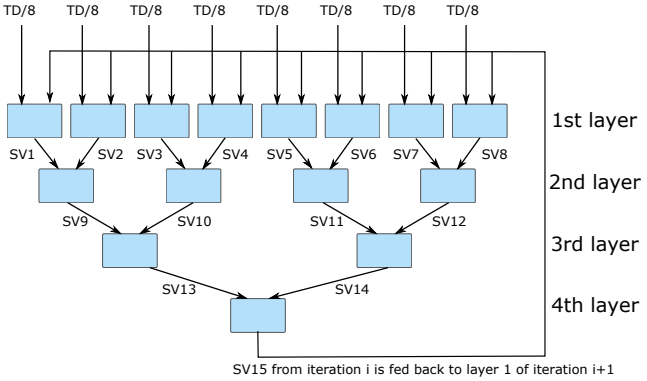

Figure 2: Schematic showing Cascade SVM (adapted from [8]). TD: Training Data, which is split and fed into different model learners that can operate in parallel. The Support Vectors (SVs) from different blocks in each layer are combined to create a larger number of aggregated SVs.

as compared to initialization in [8].

3.5 Performance evaluation

We evaluated three different models for miRNA target prediction. The first model is the global linear model. In this, we train a single SVM for the entire data, without regard to the miRNA families. The second model we evaluated was the ensemble linear model where we trained separate linear SVMs for each miRNA family. The third model was the ensemble non-linear model where we trained a non-linear SVM model for each miRNA family. In each of the cases, we sub-sampled the negative data set to have the same number of positive and negative examples. We computed the average performance of the models over five sub-sample runs. Additionally, we evaluated the performance of the global linear model by training on the mouse data set and predicting on the human data set, and vice versa. We compared the performance of our algorithm *vis-à-vis* the following algorithms, which are representative of the current state-of-the-art for miRNA target prediction—mirSVR [3], PITA [14], TargetScan [9], STarMir [22], and MIRZA [15]. We downloaded pre-computed predictions for all the algorithms, except MIRZA [15], from their respective websites. For MIRZA, we downloaded the tool from [16] and ran the tool locally on our data set. Among all the algorithms considered in this paper, only MIRZA [15] and STarMir [22] generate predictions extensively for non-canonical sites. mirSVR [3] only considers seedless sites in the 3' UTR region of the mRNA that have a single mismatch or a GU wobble in the seed region of the miRNA.

4. RESULTS

Figure 3 shows the performance of the global linear model *vis-à-vis* competition. The global linear model clearly outperforms the competition in terms of prediction performance for both seed and non-canonical sites in humans and mouse. Further, the performance of the global linear model during inter-species validation is close to the ten-fold cross-validation performance, which shows that the model captures miRNA targeting rules across species.

Since, seedless sites account for more than 90% of the AGO-binding regions in CLIP-seq data, we evaluated the performance of the ensemble linear and ensemble non-linear models for seedless sites in humans. Figure 4 shows the average five-fold cross-validation test error (misclassification rate) for different miRNA families. From Figure 4, we note that the mean test error of the non-linear SVMs for each of the miRNA families is less than the corresponding linear models, except for the miR-103 family. The benefit of the non-linear SVM is more pronounced for larger miRNA families, where the size is measured by the number of positive bindings of miRNAs from the given family. For example, for `let-7` and `mir-320`—the two largest families—non-linear SVM shows a 50% and a 69.9% advantage over the linear model. We can infer that the linear model suffers from a high bias and our intuition underlying transition to the more complex non-linear model was to remove this bias. Thus, our results bear out the advantage of our new model. The advantage of the non-linear model when looked at, aggregated over all the miRNA families, is reflected in the ROC curves, where the mean ROC curve for the ensemble non-linear SVMs is much better than the ensemble linear models—the area under the curve for the ensemble non-linear model is almost 20% higher than that for the

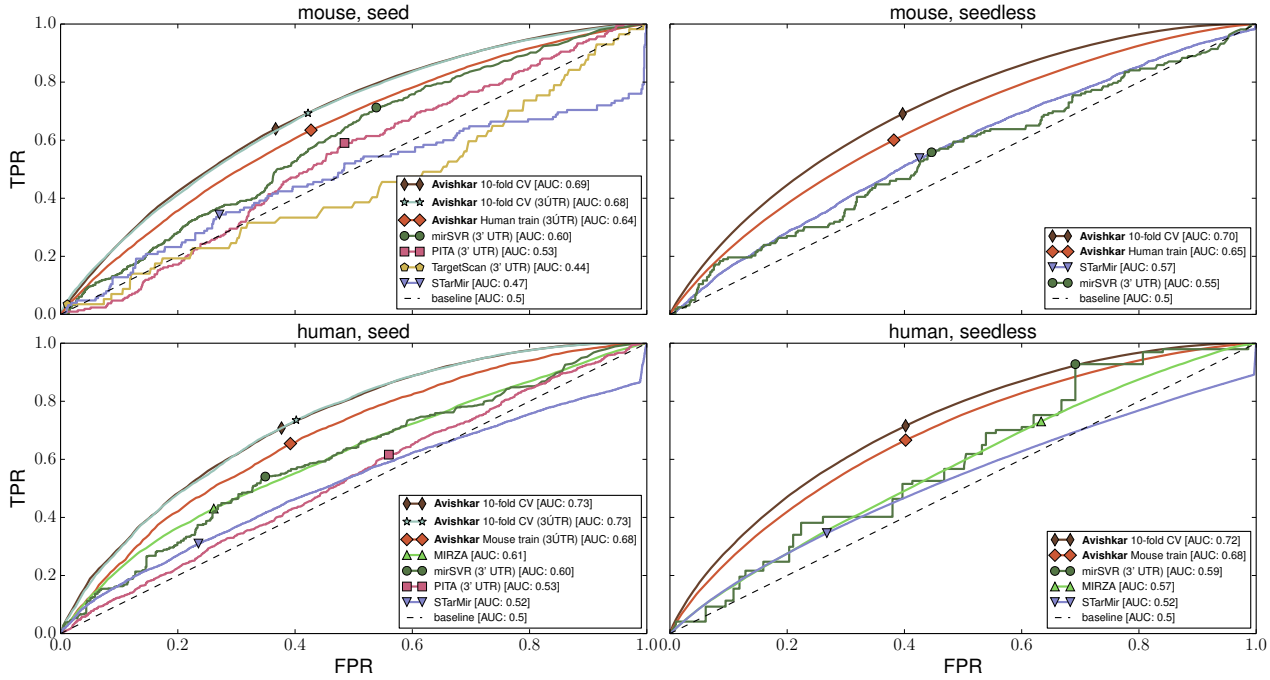

Figure 3: ROC curves for Human (PAR-CLIP) and Mouse (HITS-CLIP). The figures in the first row are for target sites involving canonical seed matches, while the second row shows results for non-canonical, seed-match target sites. The legend "Human train" in the ROC curves for mouse data indicates the model that was trained on human data, while the mouse data was used as test data set. Similarly the legend "Mouse train" in the ROC curves for human data indicates the model that was trained on mouse data, while the human data was used as the test data set. mirSVR [3], PITA [14], and TargetScan [9] only generate predictions for seed-match sites in the 3' UTR region. Note that for seedless sites in humans, although mirSVR appears to perform slightly better than MIRZA, it generates very few seedless target sites, thereby resulting in a very jagged ROC curve. The markers indicate points on the curve where the difference between the TPR and FPR is maximum.

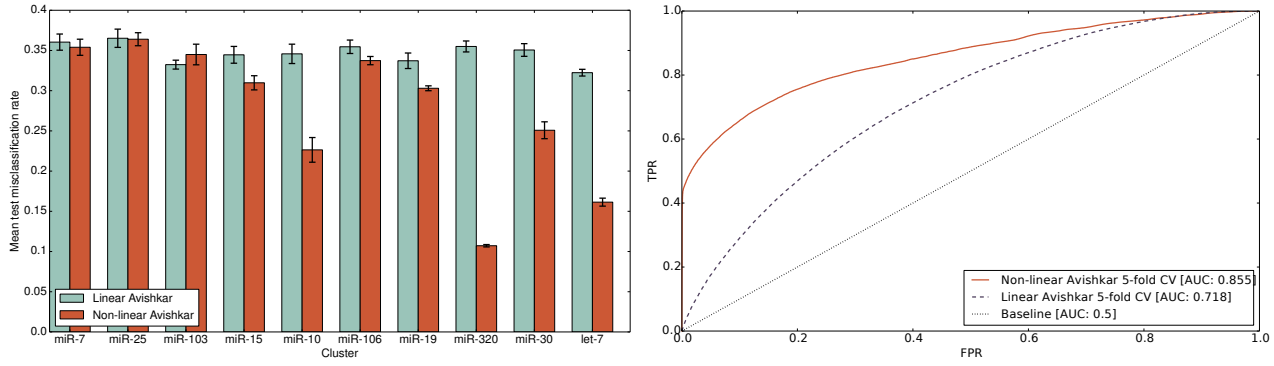

Figure 4: (Left) Mean misclassification error on the test data for the linear and non-linear model. (Right) ROC curves for the ensemble linear model and ensemble non-linear model. The misclassification error, true positive rate and false positive rates were computed using 5-fold stratified crossvalidation for seedless sites in the human data set.

linear model. To generate the ROC curve, we varied the probability threshold for the output of the SVM, after it has been passed through Platt scaling [26] (basically fitting a sigmoid that maps the SVM outputs to posteriori probabilities). One possible operating region is with a FPR of 0.2, for which the linear model has a TPR of 0.469, while the non-linear model has a TPR of 0.756, a 61% improvement. Still, the incidence of the non-negligible FPR indicates that there is scope for improvement of the classifier, possibly by doing further feature engineering.

An analysis of the weights of the global linear model revealed that site curves are important features for predicting non-canonical target sites, while for canonical sites, seed curves are more important. This vindicated our initial assumption of using curves at different resolutions for seed and seedless sites. We also found our seed enrichment metric to be an important feature for predicting non-canonical target sites. For the ensemble non-linear model, we separately tune the parameters C and γ for the different miRNA families since the clusters have varying sizes. Figure 5 shows the mean misclassification rate as a function of the kernel parameter γ, keeping C fixed at 1.0. The optimal value of γ for the larger clusters is 0.1, while for the majority of smaller clusters, it is 0.01. As γ increases, thereby reducing the sphere of influence of each support vector, the models learn a large number of support vectors, and perform poorly due to overfitting. For the ensemble linear model, we fixed the value of the regularization parameter, λ, to 0.01, for all the clusters, since varying λ did not result in significant improvement in test misclassification rate, as shown in Figure 5.

5. DISCUSSION

While, the ensemble non-linear model achieves the best-in-class prediction performance among competition by a big margin, one disadvantage of developing such miRNA family-specific models is that it precludes cross-species training and prediction because some miRNA families are species specific. One way to address this issue is to cluster the data by sequence similarity of miRNAs, or to use some other similarity measure of miRNAs, instead of using rigid miRNA families. Our implementation is generic enough to handle such cases where the user will just have to change the cluster assignment function and then *Avishkar* can efficiently learn an ensemble of classifiers in parallel.

We should also note that the performance of the ensemble linear model is the same as the global linear model. While, we could have improved the performance of the ensemble linear model over the global linear model by using different values for the regularization parameter λ, the improvement would have been marginal because the performance of the cluster-specific linear models do not change significantly with λ. This points to the fact that a linear model intrinsically has high bias for this kind of data (*i.e.,* it is too simple to fit the data well) and specializing the model to the family does not improve matters.

Finally, we note that instead of using the same number of basis functions for all the functional covariates in our model, we could have chosen different numbers of basis functions. The idea is to compute functional principal components for each feature and then select a subset of the principal components that explain 90% of the variance in the feature. This is part of our future plan.

6. CONCLUSION

In this paper we developed a method for accurate prediction of miRNA targets, both globally and in a miRNA family-specific manner. The improvement in accuracy of our system over state-of-the-art methods for predicting non-canonical miRNA target sites was more than 150%, while using the largest number of positive miRNA-mRNA interaction samples among all prior work. We achieved this higher performance by developing an ensemble of miRNA family-specific non-linear models alongside feature engineering of the inputs to our SVM-based learning models. This involved representation of the thermodynamic and sequence features in the form of spatial curves and a novel seed enrichment metric to characterize all patterns of matches, whether canonical or non-canonical, under one unified method. By using miRNA families to create ensemble models, we create predictors that are biologically more meaningful and computationally more tractable. Further, we also overcome the challenge of scaling non-linear SVMs to large data sets by implementing an algorithm for parallel training of kernel SVMs on top of Apache Spark. We believe that with more accurate prediction of miRNA targets, miRNA-based drug discovery will receive a boost since the mechanisms of miRNA action will be more precisely mapped out.

References

[1] D. P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004.

[2] D. P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, 2009.

[3] D. Betel, A. Koppal, P. Agius, C. Sander, and C. Leslie. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome biology*, 11(8):R90, 2010.

[4] R. W. Carthew and E. J. Sontheimer. Origins and mechanisms of miRNAs and siRNAs. *Cell*, 136(4):642–655, 2009.

[5] S. W. Chi, J. B. Zang, A. Mele, and R. B. Darnell. Argonaute hits-clip decodes microRNA–mRNA interaction maps. *Nature*, 460(7254):479–486, 2009.

[6] P. M. Clark, P. Loher, K. Quann, J. Brody, E. R. Londin, and I. Rigoutsos. Argonaute clip-seq reveals miRNA targetome diversity across tissue types. *Scientific reports*, 4, 2014.

[7] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome research*, 19(1):92–105, 2009.

[8] H. P. Graf, E. Cosatto, L. Bottou, I. Durdanovic, and V. Vapnik. Parallel Support Vector Machines : The Cascade SVM. *In Advances in Neural Information Processing Systems*, pages 521–528, 2005.

[9] A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell*, 27(1):91–105, 2007.

[10] A. Helwak, G. Kudla, T. Dudnakova, and D. Tollervey. Mapping the human miRNA interactome by clash reveals frequent noncanonical binding. *Cell*, 153(3):654–665, 2013. 60% of seed interactions are noncanonical, containing bulged or mismatched nucleotides. Seed matches contains bulges.

[11] A. Helwak and D. Tollervey. Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (clash). *Nature protocols*, 9(3):711–728, 2014.

[12] M. D. Jansson and A. H. Lund. MicroRNA and cancer. *Molecular oncology*, 6(6):590–610, 2012.

[13] T. K. K. Kamanu, A. Radovanovic, J. a. C. Archer, and V. B. Bajic. Exploration of miRNA families for hypotheses generation. *Scientific reports*, 3:2940, 2013.

[14] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal.

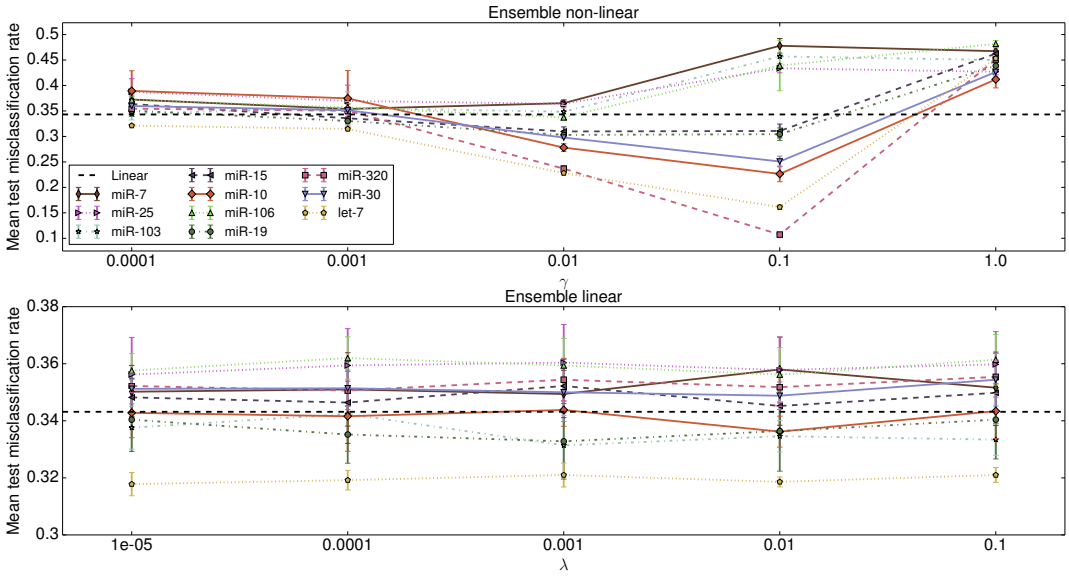

Figure 5: (TOP) Mean five-fold cross-validation misclassification error of the non-linear model for each miRNA family as a function of the RBF kernel parameter, γ, for the ensemble non-linear model, and (BOTTOM), mean five-fold cross-validation error for the linear model as a function of the regularization parameter, λ. The curves were generated by fixing the SVM misclassification penalty parameter C to 1.0, and varying the values for γ, for the ensemble non-linear model. The dashed line is the mean misclassification error of the global linear model.

The role of site accessibility in microRNA target recognition. *Nature genetics*, 39(10):1278–1284, 2007.

[15] M. Khorshid, J. Hausser, M. Zavolan, and E. van Nimwegen. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nature methods*, 10(3):253–255, 2013.

[16] M. Khorshid, J. Hausser, M. Zavolan, and E. van Nimwegen. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. `http://www.clipz.unibas.ch`, 2013. [Online; accessed 01-Mar-2015].

[17] S. Kishore, L. Jaskiewicz, L. Burger, J. Hausser, M. Khorshid, and M. Zavolan. A quantitative analysis of clip methods for identifying binding sites of RNA-binding proteins. *Nature methods*, 8(7):559–564, 2011.

[18] J. Krol, I. Loedige, and W. Filipowicz. The widespread regulation of microRNA biogenesis, function and decay. *Nature Reviews Genetics*, 11(9):597–610, 2010.

[19] J. Krüger and M. Rehmsmeier. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic acids research*, 34(suppl 2):W451–W454, 2006.

[20] J.-H. Li, S. Liu, H. Zhou, L.-H. Qu, and J.-H. Yang. starbase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale clip-seq data. *Nucleic acids research*, page gkt1248, 2013.

[21] D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, et al. Hits-clip yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469, 2008.

[22] C. Liu, B. Mallick, D. Long, W. A. Rennie, A. Wolenc, C. S. Carmack, and Y. Ding. Clip-based prediction of mammalian microRNA binding sites. *Nucleic acids research*, 41(14):e138–e138, 2013.

[23] R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker, et al. ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[24] W. H. Majoros, P. Lekprasert, N. Mukherjee, R. L. Skalsky, D. L. Corcoran, B. R. Cullen, and U. Ohler. MicroRNA target site identification by integrating sequence and binding information. *Nature methods*, 10(7):630–633, 2013.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[26] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advanced in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[27] W. Ritchie, S. Flamant, and J. E. Rasko. Predicting microRNA targets and functions: traps for the unwary. *Nature methods*, 6(6):397–398, 2009.

[28] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, 2005.

[29] Z. Wang, K. Crammer, and S. Vucetic. Breaking the Curse of Kernelization : Budgeted Stochastic Gradient Descent for Large-Scale SVM Training. *The Journal of Machine Learning Research*, 13(1):3103–3131, 2012.

[30] W. Xu, A. San Lucas, Z. Wang, and Y. Liu. Identifying microRNA targets in different gene regions. *BMC Bioinformatics*, 15:1–11, 2014.

[31] W. Xu, Z. Wang, and Y. Liu. The characterization of microRNA-mediated gene regulation as impacted by both target site location and seed match type. *PloS one*, 9(9):e108260, 2014.

[32] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, pages 10–10, Berkeley, CA, USA, 2010. USENIX Association. Apache Spark.